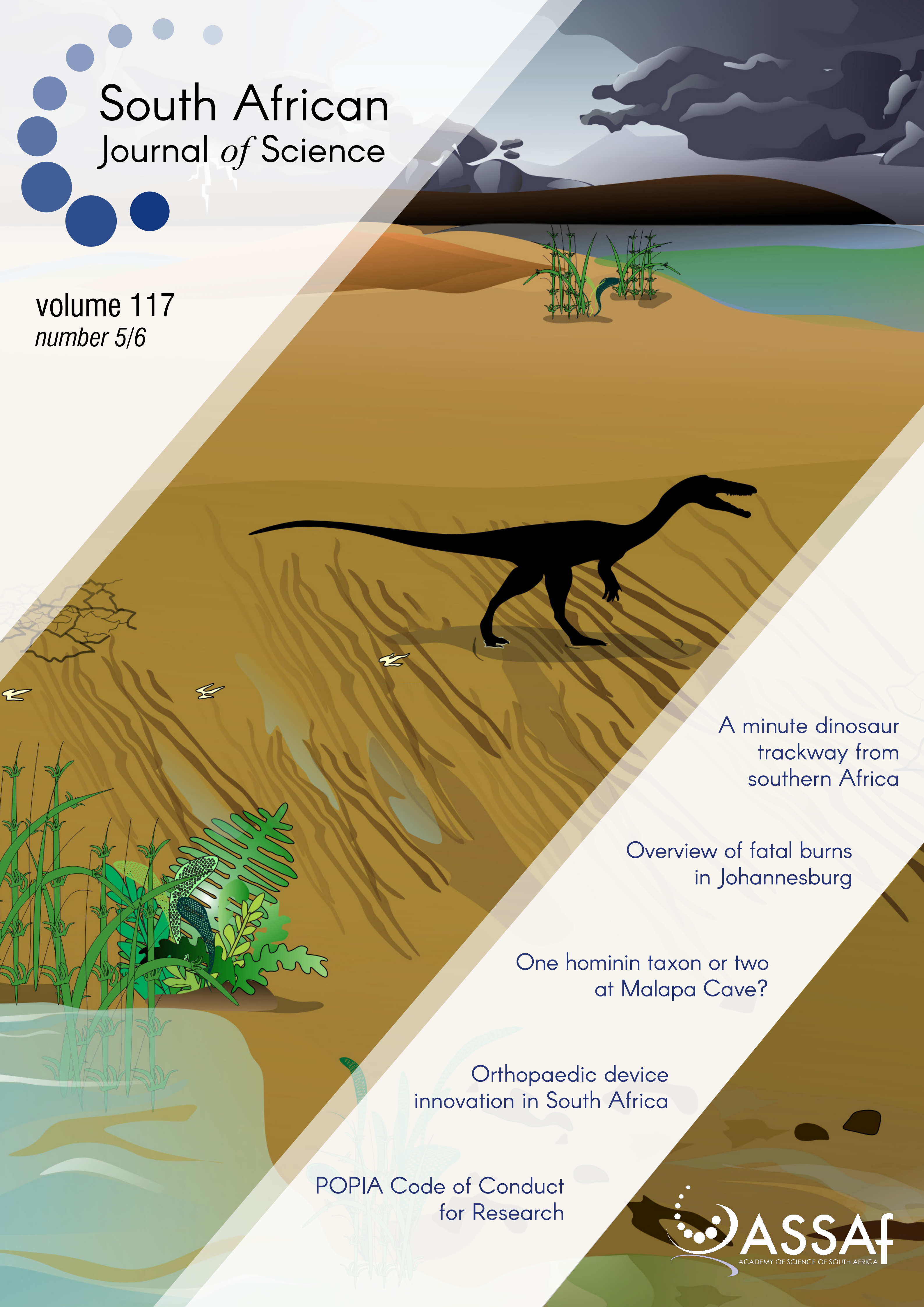
South African Journal *of* Science

volume 117
number 5/6



A minute dinosaur trackway from southern Africa

Overview of fatal burns in Johannesburg

One hominin taxon or two at Malapa Cave?

Orthopaedic device innovation in South Africa

POPIA Code of Conduct for Research

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

Volume 117
Number 5/6
May/June 2021

South African Journal *of* Science

eISSN: 1996-7489

Leader

News

Obituaries

Book Reviews

Scientific Correspondence

Commentaries

Published by
the Academy of Science of South Africa (www.assaf.org.za) with financial assistance from the Department of Science & Innovation.

Design and layout
SUN MeDIA Bloemfontein t/a
SunBonani Media
T: 051 444 2552
E: publish1@sunbonani.co.za

Correspondence and enquiries
sajs@assaf.org.za

Disclaimer
The publisher and editors accept no responsibility for statements made by the authors.

Submissions
Submissions should be made at www.sajs.co.za

Cover caption
Palaeoenvironmental reconstruction of the *Grallator* trackway in the lower Clarens Formation at the Storm Shelter ichnosite in the Eastern Cape, South Africa (modified after the original artwork by Akhil Rampersadh and Emese M. Bordy). Bordy documents this newly discovered Early Jurassic trackway in an article on page 94.

Discussion Document

Review Article

Research Articles

Research Letter

Information, loss, and the complex work of reclamation

On 18 April 2021, fires destroyed irreplaceable archival materials and other property at the University of Cape Town, as well as fynbos and other properties in the same area. Gas canisters at Rhodes Memorial, just above the upper campus of the University, exploded, and there was substantial damage, both on the slopes of Table Mountain and in surrounding areas.

All of us at SAJS extend our condolences to those affected by the fires. The loss of important archival material at any South African university is a loss to the entire scholarly and scientific community. There have been many reports of teams of volunteers doing what they can to salvage materials in the library, along with international efforts to collate digitised versions of documents, often collected by individual scholars. These are praiseworthy and important efforts.

At the same time, there has been serious loss to the environment, as documented by conservation ecologists.[1] In the analysis of Rebelo and Esler[1], the issue of fynbos fires is a 'wicked problem' contributed to in part by the proliferation of alien vegetation and lack of resourcing for environmental management – all issues which many authors and supporters of our Journal are trying hard to resolve. It is not entirely clear who or what started the fires in the first place, but a number of reports have suggested that the blazes started from a small fire used by homeless people, probably for warmth and cooking. Homelessness is an issue, clearly, not just for homeless people, and an area needing attention and evidence-based input. The huge inequalities in our country are bad for everyone, not just the poor, and for the environments and heritages which we share. 'Wicked problems' like destructive fires are in their nature interdisciplinary, requiring solutions through cooperation across a range of scientific disciplines, including social scientific disciplines. It has been said many times before, but it can never be said too many times: we need to work together to address the huge challenges our country, our continent, and, indeed, the world, face.

At the same time that we in the scientific community and as South Africans and Africans contemplate the devastation of the fires and what this means for all of us, the scientific community is engaged in an exercise which will affect us all: the policy, legislative, and consultation processes related to the *Protection of Personal Information Act* (POPIA). It has been a pleasure and a privilege for us at SAJS to work with those who, on behalf of the entire science community, have worked exceptionally hard to produce the POPIA Code of Conduct for Research[2], and to have created dedicated space for commentary and feedback in the form of the 'Discussions on POPIA' series, in which contributions receive priority publication to facilitate discussion and commentary within the tight timeframes needed. POPIA represents an important milestone in legislation affecting research in South Africa. Implementation and compliance will take work and adjustment from us all, and we are grateful that debates about what POPIA is, does, and should be, have started on our pages and will continue. It is important that the science community are engaged with questions of information protection, and that, as academics and researchers we continue to debate and disagree, learning from one another. We hope that more authors will send us commentaries for review, as the realities of POPIA take hold. No document is perfect, no process without flaws, and it is our job to contest constructively and make things better.

If the fires are a story of loss and the POPIA process a story of reclamation of the rights of, amongst others, research participants, it is also important to see the links between these two issues affecting us all. There is no question that the loss of information associated with the fires is bad for research, and that reclamation in a range of ways, from the digital to the environmental, is important. But there are other, less obvious, issues at stake. It has become a cliché to say that history is written by the victors, with the exact origins of that formulation, and related formulations, being contested. But issues of representation and participation in scholarship need our attention. Behind the long POPIA process, going back years, is the question, amongst others, of the rights of those who are researched to be recognised as agents in their own lives and not just as data, the objects of the gaze of the more powerful. POPIA engages with the question of people's rights to consent (or not consent) to information about them or from them to be used. In many of our archives all over the world we have documents and stories which are told without the consent of the central characters, narratives which may and do have many good qualities but which may also contribute to continued othering of people positioned as objects of research. Archival scholars have begun a long recuperative process not just to consider the importance of protecting the archive – however complex and laden with histories of power, inclusion and exclusion that archive may be – but also to think about and research information not properly archived, if archived at all. This lack of archiving, as a number of scholars including Phalafala[3,4] have shown, differentially excludes voices of the less powerful from what is commonly viewed as knowledge in our context, and indeed, globally.

Part of the work of science is that of reclamation and restoration. Maintaining our world is indeed a 'wicked problem', which many people in our science community are working on. Part of the maintenance and development of our world lies not only in protecting people's rights to some control over the information they supply to us, but also in recognising and interrogating information and knowledges which have been suppressed. Science is changing, and there is much to celebrate in this regard, but we still have some way to go in incorporating, drawing from, and engaging critically with the voices of those who have been, and continue to be, excluded. The Cape Town fires started at Rhodes Memorial, and the name Rhodes, of course, has been associated with tumult and upheaval in our scholarship systems. We need also to think about who else needs to be remembered and engaged with, in such a way that respects both their right to privacy and their right to be active participants in the research conversation.

References

1. Rebelo A, Esler KJ. Why the fire on Cape Town's iconic Table Mountain was particularly devastating. The Conversation. 2021 April 20. Available from: https://theconversation.com/why-the-fire-on-cape-towns-iconic-table-mountain-was-particularly-devastating-159390

2. Adams R, Adeleke F, Anderson D, Bawa A, Branson N, Christoffels A, et al. POPIA Code of Conduct for Research. S Afr J Sci. 2021;117(5/6), Art. #10933. https://doi.org/10.17159/sajs.2021/10933

3. Phalafala UP. Polyglot internationalism and the matriarchive: The case of Keorapetse Kgositsile. Interventions. 2020;22(3):346–363. https://doi.org/10.1080/1369801X.2020.1718539

4. Phalafala UP. The matriarchive as life knowledge in Es'kia Mphahlele's African humanism. a/b Auto/Biography Stud. 2020;35(3):729–747. https://doi.org/10.1080/08989575.2020.1762999

HOW TO CITE:
Swartz L. Information, loss, and the complex work of reclamation. S Afr J Sci. 2021;117(5/6), Art. #11164. https://doi.org/10.17159/sajs.2021/11164

AUTHOR:
Crain Soudien[1]

AFFILIATION:
[1]Chairperson of the Humanities Standing Committee, Academy of Science of South Africa, Pretoria, South Africa

CORRESPONDENCE TO:
Crain Soudien

EMAIL:
crain.soudien@uct.ac.za

HOW TO CITE:
Soudien C. The 2020 Academy of Science of South Africa Book Award – We build a culture. S Afr J Sci. 2021;117(5/6), Art. #10945. https://doi.org/10.17159/sajs.2021/10945

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
humanities, social sciences, scholarship

PUBLISHED:
28 May 2021

The 2020 Academy of Science of South Africa Book Award – We build a culture

In 2011, the ASSAf *Consensus Study on the State of the Humanities in South Africa: Status, Prospects and Strategies,* suggested, as part of a suite of ten recommendations for 'improv(ing) the circumstances faced by the Humanities, not only in South Africa, but also around the world', that we, referring obviously to everybody with a stake in the Humanities, including government,

> *(r)eview and refine government funding allocations to the Humanities with substantive earmarked funding in critical areas such as African languages, Philosophy, History and the Creative and Performing Arts. In this context, the advancement of books by the academy and the funding of books by government could significantly enhance the book as a cultural and human asset in both the scholarly and public mind. One possibility would be to link an award for the best Humanities book every year to the annual Alan Paton Award.*[1]

ASSAf itself acted on this recommendation and established a Book Award for the Humanities. Five years later, in 2016, the inaugural ASSAf Book Award was made to Professor Keith Breckenridge for his book, *The Biometric State: The Global Politics of Identification and Surveillance in South Africa, 1850 to the Present.*[2] In 2018, the second book award went to Professor Chabani Manganyi for his book, *Apartheid and the Making of a Black Psychologist*.[3] For 2020, the third award, reported on here, went to Professor Charles van Onselen for his book *The Night Trains.*[4]

The decision by ASSAf to establish this award, in a context of persistent concern about the state, place and role of the Humanities, was both necessary and strategic. Characterising the concern were the findings of the *Consensus Study* which drew attention, inter alia*,* to the post-apartheid government 'systematically benefit(ting) Science, Technology, Engineering and Mathematics' and to what it described as 'the moribund condition' of the Humanities within institutions of higher learning. In the wake of these findings, ASSAf established a Humanities Standing Committee which was given the responsibility for ensuring that the Humanities disciplines remain an important focus of Academy activities and for overseeing and guiding Academy activities in the Humanities. As one of its tasks, the Humanities Standing Committee has been overseeing the institutionalisation and management of the Book Award since 2016.

How the Book Award, and the other awards and events that have come into being following the *Consensus Study* such as the ASSAf Humanities Lecture, the Human Sciences Research Council's Social Science and Humanities Medal (given in 2020 in conjunction with Universities South Africa) and the Humanities and Social Sciences' Annual Book, Creative Collection and Digital Contribution Awards, actually address the concerns raised in the *Consensus Study* remains to be seen. Their impact is not easily demonstrated. They offer, however, distinct opportunities for institutions and the fields of the Humanities and the Social Sciences to be doing that most difficult of things – establishing cultures and traditions that matter.

Cultures and traditions that matter, in a context of building a nation, are deeply important. The ASSAf Book Award is powerfully constitutive for this project, as are the other awards. It helps us see where we are, what we are doing and what we could be doing. Less obviously, it helps, also, in a South African society still finding its bearings, in *shaping* attitudes and understandings about scholarship. In 2015 in an interesting blog entitled 'Absent Amandla: Is South Africa Anti-Intellectual?', Christopher Wheeler[5] commented that

> *'... science simply isn't a big enough part of South Africa's social discourse. We don't value its proven track record and we struggle with its consequences. We see it as something other countries and groups pursue; perhaps even, and this would indeed be terrifying, as a white/European Thing.*

Having been directly involved in the processes for making the award, I have been struck by the intellectually wholesome proposition the Book Award represents in the face of what Wheeler is decrying. All the qualities of scholarship which a society would wish for to cherish and protect Itself against superficiality, facile solutions, coarse demagoguery and, topping it all, the cultivation of immediatist discourses of material gain – intellectual courage, compassion and clarity – have been in abundant evidence.

The process of adjudicating the 2020 Book Award was, for all of us involved, extremely stimulating. We began by reminding ourselves what the purpose of the exercise was – to identify an example of scholarship that was inspiring, well written and educative. We were looking for a monograph, not an edited collection, which exemplified the importance of the book as a valued social and cultural artefact. The call for nominations went out in October 2020. In response, as has been our experience on the two previous occasions, we received many queries, many suggestions and many nominations. We had to make early judgements about textbooks and edited collections which we could not include and in the end worked with 36 submissions from the breadth of the university and science landscape. There were submissions from a range of genres and disciplines and, notably, a range of young and more experienced scholars. Our assessment, and we celebrate the moment, was that our scholars in the Humanities and the Social Sciences had not abandoned the value and enjoyment of writing books.

Sifting through the submissions was not easy, but we were able to arrive at a shortlist. This shortlist consisted of five wonderful books: Jacklyn Cock's[6] *Writing the Ancestral River,* a book about the colonial and apartheid history of the Kowie River, Charles van Onselen's[4] *The Night Trains* about the rail transport that carried Mozambican migrant workers to the Witwatersrand from about the first decade of the 20th century, Alex Broadbent's[7]

Philosophy of Medicine about medicine and the curing of illness, Lazlo Passemiers'[8] *Decolonisation and Regional Geopolitics* which looked at the role and contribution of South Africa in the internal struggles of post-independence Congo, and Bridget Kenny's[9] *Retail Worker Politics, Race and Consumption in South Africa* which tells the story of the politics around workers in the retail sector from about the 1950s to the present. The scholarship on offer provided an important response to questions about the state of the Humanities and science scholarship in South Africa. It was engaged and, importantly, deeply relevant. Standing on its own merits as globally noteworthy, it more pointedly spoke to the politics – past and present – of South Africa itself.

Van Onselen's *The Night Trains* received the unanimous support of the adjudication committee for its lucidity, rigour and sustained significance for understanding the nature of the South Africa we have inherited and currently have. A colleague on the panel said of it that

> (h)*is chapter on madness and masculinity is van Onselen at his best, and this section places him alongside Thompson and, especially, Genovese. As he tells the story of the miners and their apparent 'madness', he provides a searing indictment not only of the systemic, regional operation of racial capitalism and colonialism, but of a callow and callous white South African citizenry. He raises these points in his Studies of the Social and Economic History of the Witwatersrand, published forty years ago, but returns to them with a precision and passion, even, that was not evident in his earlier work.*

Another colleague said, 'This… is an outstanding work, engaged, humane and displaying the qualities of sharpness, insight and balance that exemplifies the work of critical and engaged scholarship at its best.' Writing in the liner notes for the book, James Scott, Sterling Professor of Political Science and Anthropology at Yale, said of the work that it was 'an unsurpassable lesson in the commodification and disposal of human life'.

The event at which Charles van Onselen received the award was itself a significant occasion. He not only spoke to the content of the book but raised questions about the nature of South Africa, sharply drawing attention to two important issues in its narrative of itself: first, the need to reclaim the country's forgotten histories, making the point that *The Night Trains* was an experiment in reclamation. His second point was that the use of technology in South Africa, steam engines in this case, had important parallels elsewhere in the world and most notably in the experience of the Jewish Holocaust in Nazi Germany and the Baghdad Railway which contributed directly to the genocide of the Armenians by the Ottomans. In making these points, he spoke of what he described as 'the shortsighted' ways in which South Africans, both black and white, had appropriated the song 'Shosholoza' – the popular South Africa sports anthem for 'pushing forward' – and in the process were 'reducing complexities to brutal banalities'.

What we now have in this award is an important new opportunity to showcase the best of our scholarship and to assert our point that high-quality Social Science and Humanities research is essential for understanding ourselves, who we are, why we are in the place where we now find ourselves and where we are going. It is the opportunity we forsake at our peril of hearing what the most far-sighted amongst us are telling us what we wish not to hear. It is, as Van Onselen himself said in receiving the award, the riposte to self-serving narrow nationalism.

References

1. Academy of Science of South Africa (ASSAf). Consensus study on the state of the humanities in South Africa: Status, prospects and strategies. Pretoria: ASSAf; 2011. http://hdl.handle.net/20.500.11911/33
2. Breckenridge K. The biometric state: The global politics of identification and surveillance in South Africa. Cambridge: Polity Press; 2014. https://doi.org/10.1017/CBO9781139939546
3. Manganyi C. Apartheid and the making of a black psychologist. Johannesburg: Wits University Press; 2016. https://doi.org/10.18772/12016058622
4. Van Onselen C. The night trains. Johannesburg: Jonathan Ball; 2019.
5. Wheeler C. Absent Amandla: Is South Africa anti-intellectual? Woolgatherist. 2015 September 21 [cited 2021 Apr 23]. Available from: Medium.com/@Woolgatherist/absent-amandla-is-south-africa-anti-intellectual-3cebb66b71ca1
6. Cock J. Writing the ancestral river – A biography of the Kowie River. Johannesburg: Wits University Press; 2018. https://doi.org/10.18772/12018031876
7. Broadbent A. Philosophy of medicine. New York: Oxford University Press; 2019.
8. Passemiers L. Decolonisation and regional geopolitics: South Africa and the 'Congo Crisis'. Abingdon, Oxon: Routledge; 2020. https://doi.org/10.4324/9781351138161
9. Kenny B. Retail worker politics, race and consumption in South Africa. London: Palgrave Macmillan; 2018. https://doi.org/10.1007/978-3-319-69551-8

Check for updates

AUTHOR:
Nico Cloete[1]

AFFILIATION:
[1]DSI-NRF Centre of Excellence in Scientometrics and Science, Technology and Innovation Policy (SciSTIP), Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Nico Cloete

EMAIL:
ncloete@chet.org.za

HOW TO CITE:
Cloete N. Rolf Stumpf (1945–2020): Tough, decisive and compassionate technocratic-intellectual. S Afr J Sci. 2021;117(5/6), Art. #10684. https://doi.org/10.17159/sajs.2021/10684

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

PUBLISHED:
28 May 2021

Rolf Stumpf (1945–2020): Tough, decisive and compassionate technocratic-intellectual

Reflecting on the career of Rolf Stumpf, who passed during late October 2020 from cancer, reminded me of a meeting I had around 1992 with the famous Hungarian sociologist Ivan Szelenyi, who had started working on his very interesting book *Making Capitalism without Capitalists* (Verso, 1998). With the South African transition looming, he was interested, from a Hungarian perspective, in what made certain communists successful in the new capitalist Hungary and he wanted to know if I was interested in doing a similar study in South Africa. The transition became far too busy for me to consider another study, but the central theme of Szelenyi's work was that one of the characteristics of a successful transition was their distinctive technocratic-intellectual elites who were both successful communists and successful capitalists.

As we started thinking about establishing the National Education Policy Investigation project to rethink higher education in the forthcoming post-apartheid era, I was told that I must talk to Rolf Stumpf who was then on the 'other side' in government in the higher education division of the South African Department of Education. From the first moment it was clear that this statistician knew the university sector like nobody that I knew, and that he was also passionate about making it a successful sector in post-apartheid South Africa. Alas, Rolf moved very quickly to the 'new side' when he was asked by then President De Klerk, in consultation with the ANC, to save the Human Sciences Research Council (HSRC) for the new South Africa (1991).

The HSRC showed different sides of Rolf: tough and decisive but compassionate. He retrenched a large number of 'apartheid deadwoods' but told me how difficult it was and how in some cases he had to go to the homes of retrenched employees to explain to their spouses that the employee had not done something wrong at work; he once said with a chuckle, 'But I could not tell the spouse that he had also not done anything at work.' When the HSRC fortress needed painting, Rolf embarked on a consultation process, but after a week of no progress and conflict, he gave a colour chart to his secretary and said 'You decide' – the building is still the same colour, albeit not universally loved.

So when Prof. Bhengu, then Minister of Education, asked some of us for names for possible members of the Nelson Mandela appointed National Commission on Higher Education (NCHE), Rolf was very high on the list. He was a huge resource who knew how the old system worked (he had partially designed the funding system for it) and we trusted him as somebody who was committed to transformation.

And he 'saved' the NCHE: when Prof. Teboho Moja (advisor to Minister Bengu) and I said that we were leaving the NCHE because we could not work in the chaotic offices of the new Education Department, Rolf immediately said 'come to the HSRC', where he provided us not only with generous office space in the purple palace, but also with secretarial and financial services. We would never have finished the NCHE on time without Rolf's support.

We spent long hours with Rolf during the 2 years of the NCHE where we became more than commissioners – we became a close group trying to design a new higher education system. It was during this time that Prof. Moja and I received funding from Ford Foundation to establish the Centre for Higher Education Transformation (CHET), which was launched in the HSRC despite me having written a polemical article a few years earlier arguing for the closure of the HSRC – but that was before Rolf started there. It was also a 'no brainer' for Rolf to be on the first Board of CHET with luminaries like Walter Kamba (Vice Chancellor of the University of Zimbabwe), Brenda Gourley (Vice Chancellor of the University of KwaZulu-Natal) and Mike O'Dowd (Director of Anglo America).

Most of us lost touch with Rolf during his time at Stellenbosch University, and the language struggles, which he lost. But Stellenbosch's loss was the University of Port Elizabeth's gain. He became the Vice Chancellor of the old University of Port Elizabeth (UPE) and was exactly the right person to become the CEO and Vice Chancellor of the three merged institutions (UPE, Port Elizabeth Technikon and the Port Elizabeth campus of Vista University). The unusual title of CEO showed that he was serious about management – which was really his forte in life.

Soon after the merger was announced, Rolf phoned me and said that he had a problem with naming the new institution. The mayor wanted to include Port Elizabeth in the name and Rolf wanted Mandela in the new name, but the Mandela Foundation had informed him that it would need approval through a lengthy and difficult process involving Adv. Bezos.

I then phoned Jakes Gerwel, the Director-General in the Presidency, and he said he liked the idea and I must give him a few days to talk to the man himself. Two days later, Gerwel phoned and said Mr Mandela had agreed and would be honoured, but to accommodate the mayor, why not call it Nelson Mandela Metro University – and if Metro is included, it does not need Mandela Foundation approval – and Adv. Bezos also liked the idea. Another master stroke from the great man himself and a few days later Rolf made the announcement. And as we know, in 2017 the name was changed to Rolf's original idea: Nelson Mandela University.

Nelson Mandela University, along with the University of Johannesburg, which is also a merger of a traditional university with a technikon, is arguably the most successful merger in South Africa and a university that is the pride of the Eastern Cape and will for a very long time have the stamp of Rolf Stumpf on it. It took a very special person to merge three very different institutional cultures. In a tribute, the current Vice Chancellor, Professor Sibongile Muthwa said:

> *He delicately, skilfully and successfully steered our University through what was a complex merger process, leaving us with a viable institution that we can be proud of.*

He remains one of the most respected leaders in the South African higher education sector.

After contributing to fixing the public higher education system, he told Prof. Moja in March 2020 how he was enjoying working with colleagues in private higher education as well. He also tried fixing the Botswanan higher education system when he was asked by their National Council to develop a quality assurance system for their higher education system.

When Rolf became sick, I had a brief conversation with him in which he told me he had just completed a very interesting evaluation of efficiency at Stellenbosch University. To the end Rolf was committed to changing and improving higher education in Africa.

To return to Szelenyi, Rolf Stumpf could certainly be characterised as a technocratic-intellectual who moved seamlessly from one regime to another, but with a difference. Rolf never joined either the National Party or the ANC, but he was committed to democracy with a strong belief that a healthy higher education and science system was an essential ingredient.

Photo: Nelson Mandela University News

AUTHOR:
Ernesta M. Meintjes[1]

AFFILIATION:
[1]Division of Biomedical Engineering, Department of Human Biology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Ernesta Meintjes

EMAIL:
ernesta.meintjes@uct.ac.za

HOW TO CITE:
Meintjes EM. Tania Samantha Douglas (1969–2021): Biomedical engineer, academic and leader. S Afr J Sci. 2021;117(5/6), Art. #11067. https://doi.org/10.17159/sajs.2021/11067

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

PUBLISHED:
28 May 2021

Tania Samantha Douglas (1969–2021): Biomedical engineer, academic and leader

Professor Tania Samantha Douglas, leading biomedical engineer, innovator and academic, died on 20 March 2021.

Tania's untimely death, after an extended battle with cancer, came as a tremendous blow to her colleagues and students at the University of Cape Town (UCT), and to many friends and collaborators further afield. Tania was an internationally recognised scholar, admired by many and consulted broadly for her unique insights, in-depth understanding of the higher education environment in South Africa, and open-mindedness. Always vibrant, she was able to fully engage with issues in an unbiased way, sharing her well-considered thoughts in a friendly and practical way.

Tania was born to Helena (Rita) Harriet (née Volkwyn) and Aubrey Douglas on 11 August 1969 in Pacaltsdorp, a small community on the outskirts of George in the Southern Cape. After obtaining the second highest grade in the country in the matriculation examinations administered by the House of Representatives in 1987, Tania proceeded to read for a BScEng in Electrical and Electronic Engineering at UCT. This was followed by an MS in Biomedical Engineering at Vanderbilt University in Nashville, Tennessee, a PhD in Bioengineering from the University of Strathclyde in Glasgow, and a postdoctoral fellowship in image processing with the Japan Broadcasting Corporation in Tokyo. In 2000, Tania returned to her alma mater to take up a position as lecturer in the Department of Biomedical Engineering.

During her 21 years at UCT, Tania held numerous leadership positions within the department and faculty, including serving as Divisional Head for a period, leading the MRC/UCT Medical Imaging Research Unit for the past decade, and serving as Deputy Dean of Research in the Faculty of Health Sciences. In 2016, Tania was awarded the prestigious South African Research Chair in Biomedical Engineering and Innovation, and in 2018 was Founding Director of the Biomedical Engineering Research Centre at UCT. Tania excelled in all spheres of academia. She headed up a large research group, trained and graduated more than 50 master's and doctoral students, mentored postdoctoral fellows and junior staff, published extensively in leading international journals, and taught and developed courses. Her scholarly contributions were recognised through numerous awards, including research fellowships from the International Institute for Theoretical Physics in Trieste, Italy; the Alexander von Humboldt Foundation in Germany; the Erasmus Mundus programme of the European Union; and Female Academic/Researcher of the Year by the IEEE Women in Engineering South Africa section. In 2019, she was recognised at the South African Women in Science Awards as Distinguished Woman Researcher in Research and Innovation, and in 2018 as a Quartz Africa Innovator. In the past decade, she was elected a Fellow by the South African Academy of Engineering, the International Academy of Medical and Biological Engineering, and the University of Cape Town, and as a member of the Academy of Science of South Africa.

Tania's research focused on major public health problems in South Africa for which she developed novel instruments and computer-assisted techniques. Some of her early work involved developing image-processing techniques to characterise the facial phenotype associated with Fetal Alcohol Syndrome – a condition of which the incidence in certain communities in South Africa is amongst the highest in the world. Tania also made seminal contributions in tuberculosis (TB) diagnosis, including the development of a 'smart microscope' for automated detection of TB bacilli in stained sputum smears, and computer-aided detection of pulmonary pathology in paediatric chest X-rays.

Quoting her friend and mentor, Emeritus Professor Christopher Vaughan, 'Tania Douglas was a true citizen of the world, transcending geography and embracing the environment in which she found herself.' This is clearly reflected by Tania's recent work that strived to combine biomedical engineering with social context to find novel solutions towards improved health. To this end, she developed a new postgraduate programme in Health Innovation that teaches human-centred innovation, with an emphasis on end-user engagement. She believed and advocated that Africa needs to find solutions to its own problems and strove tirelessly to build biomedical engineering capacity across Africa. As part of these efforts, she played a leading role in the establishment of the African Biomedical Consortium, launched and was founding Editor-in-Chief of the open-access electronic journal *Global Health Innovation*, and edited the open-access eBook entitled *Biomedical Engineering for Africa* (University of Cape Town Libraries; 2019).

Since 2014, Tania served as Associate Editor of both the *South African Journal of Science* and *Medical Engineering and Physics*, and in January of this year was appointed as Editor-in-Chief of the latter.

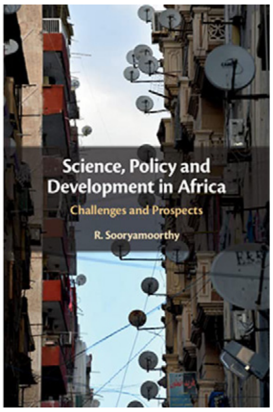
In addition to Tania's many scholarly achievements, she impacted the lives of her students, colleagues and collaborators in a very personal way through her caring and thoughtful nature. She was warm and empathetic, and an inspiring mentor to many. The manner in which she carried her illness demonstrated incredible courage and inner strength. Having touched the lives of so many, Tania leaves a great void. This is expressed beautifully in the words of her friend and head of the Department of Human Biology, Prof. Sharon Prince, who wrote:

> *We will remember Tania for being an amazing woman – brave, humble and brilliant. She lived her life, and carried her illness, with extraordinary grace and dignity. We will remember her for her astute intellect and her quiet humanity to build others in the process. She was talented and gracious, and we will remember her positive attitude and ever-present beautiful smile.*

Heartfelt condolences to her parents, Aubrey and Rita Douglas, and other members of her immediate family.

Science potential for development in Africa not yet fully realised

BOOK TITLE:
Science, policy and development in Africa: Challenges and prospects



AUTHOR:
Radhamany Sooryamoorthy

ISBN:
9781108842037 (hardcover, 352 pp)

PUBLISHER:
Cambridge University Press, Cambridge, UK; ZAR2745

PUBLISHED:
2020

REVIEWER:
Fanie Cloete[1,2]

AFFILIATIONS:
[1]School of Public Management, Governance and Public Policy, University of Johannesburg, Johannesburg, South Africa
[2]School of Public Leadership, Stellenbosch University, Stellenbosch, South Africa

EMAIL:
cloetegs@gmail.com

HOW TO CITE:
Cloete F. Science potential for development in Africa not yet fully realised. S Afr J Sci. 2021;117(5/6), Art. #9390. https://doi.org/10.17159/sajs.2021/9390

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

PUBLISHED:
28 May 2021

Radhamany Sooryamoorthy is Professor of Sociology in the School of Social Sciences at the University of KwaZulu-Natal in South Africa. His field of interest is the contribution of science to development in Africa compared to more developed nations. This book summarises his recent research into this topic.

Structured over seven chapters, Sooryamoorthy maps the evolution, development and implications of scientific practices in both colonial and contemporary Africa. In this process he attempts to answer 10 research questions. They deal inter alia with issues such as (1) the extent to which science was undertaken historically in Africa under colonial rule; (2) the extent to which it is currently practised and prioritised through systematic science, technology and innovation approaches, policies and funding; (3) which African countries do what in this regard and why, and (4) how weaknesses in current practices can be improved to facilitate and support development on the continent (p. 25–26).

In Chapter 1, concepts such as science, technology, innovation and development, as well as their inter-relationships within the African context are explained. Sooryamoorthy then summarises his overall conclusions which point to a direct correlation between the availability of scientific knowledge, development and wealth creation, both in the literature and from his own empirical data. However, he also finds that the generation of knowledge itself (i.e. scientific productivity) is not always aimed at wealth creation (p. 6). Rather, scientific research is stimulated by a need to find solutions for intractable problems and to improve efficiency and effectiveness. Science is, therefore, not necessarily driven by a desire to improve economic growth, but rather to resolve technical and other problems (p. 7).

Sooryamoorthy's main theoretical points of departure include Gibbons et al.[1] and Krishna et al.[2] He also uses Goldemberg's[3] different models to assess the relationship between science and development as illustrated by US and Japanese practices, and Moravcsik[4] with respect to a short-term economic growth focus for scientific research, as opposed to a longer-term development impact. The conceptualisations of Schott[5], Nagtegaal and De Brun[6], and Castells[7], respectively, are furthermore employed with regard to modernisation, dependency, and the unequal and discriminatory impact of scientific research activities on development outcomes in core and peripheral contexts (p. 30–31).

These different theoretical approaches provide interesting and potentially useful conceptual frameworks to assess the impact of scientific research activities on development in both colonial and contemporary African societies. However, despite the many strengths of the book, the author unfortunately does not summarise in his other chapters, or in his final conclusions, the extent to which his findings are congruent with, or contradicting, these theoretical frameworks.

The author undertakes empirical data collection mainly though a systematic bibliometric approach (p. 33, 36). The research findings and conclusions are summarised in Chapter 2, while more detailed qualitative and quantitative data supporting these findings are given in Chapters 3 to 5.

In Chapter 2, the author summarises the main differences in approach between Anglophone and Francophone African countries, including how colonial science research practices focused mainly on finding solutions to urgent, short-term health- and agriculture-related problems and issues in their respective home countries, with little regard to local needs and challenges in the African context (p. 42). This conclusion is fully in line with the general nature and focus of colonial governance in Africa. However, he also details how decolonisation since the 1960s has not immediately resulted in significant changes to the practise of scientific research and he explains this as emanating from the negative impacts of the colonialist legacies, aggravated by various contra-productive political, educational and economic management strategies of successive post-colonial governments (p. 44–48).

Nonetheless, Sooryamoorthy finds that there has been a slow increase and improvement in science research in African countries since independence. However, these improvements remain minimal owing to serious political instability, institutionalised corruption and ideologically competing policy priorities as well as the need to provide basic infrastructure and services. Scientific research in Africa remains driven largely by foreign donors and researchers. This is problematic, because African priorities, and not international donor priorities, are needed (p. 19–20).

A number of tables, lists and analyses which provide substantiation for the above findings are presented in Chapters 3 and 4. Chapter 3 gives extensive, and in the opinion of this reviewer, sometimes unnecessarily detailed, lists, descriptions and summaries of 549 427 research publications (p. 63) in different sectors of African countries between 1945 and 2015. In addition, detailed analyses and assessments of funding allocations for these activities are provided (p. 74). In Chapter 4, the author lists, classifies, compares and assesses the sectoral research areas on which these science publications focussed, as well as the relative strengths and weaknesses of individual African countries in this regard (p. 101).

The logic and relevance of presenting this compendium of descriptive lists of publications, sectoral foci and funding allocations in the minutest detail in the text is not clear. The readability of the book would have been improved if these details had been contained in separate annexures at the end of the text. However, this approach and content are consistent with the author's explicitly stated intention to map the development of historical and contemporary science practices in different countries in Africa, both under colonialism and thereafter.

Chapter 5 deals with the importance of scientific collaboration in Africa, within the discriminatory core–periphery model of scientific impact referred to in Chapter 1. Again, there is detailed comparative historical and contemporary technical data to substantiate his conclusion that '… the colonial past continues to influence research agendas and patterns of collaboration' (p. 150). However, the conclusion is correct that the situation is changing and that there is increasing collaboration among African countries when it comes to scientific research (p. 174).

Chapter 6 assesses the consequences of the finding that systematic science policies and policy systems in Africa are either largely absent or, where they do exist, are not synchronised or integrated with other sectoral policies, technologies and innovations to ensure sustainable development outcomes and impacts (p. 198). Sooryamoorthy concludes that S&T policy development in Africa is still seriously underdeveloped as a result of colonial practices, and he correctly identifies a number of important policy design and implementation improvements that need to be made by African governments to fast-track more successful science impacts in those societies, including more effective and better-informed policy and implementation. At the same time, he acknowledges the complicating nature of policymaking in politically volatile situations or where there is conflict. (p. 217–219, 221). His plea for the African Union to intervene, however, is – in the opinion of this reviewer – probably doomed to fail, because of the inability of the African Union so far to persuade its members to subject national interests to regional, continental or global needs.

In his concluding Chapter 7, the author probably correctly identifies the main science policy constraints in Africa as different subjective national policy priorities as well as objective implementation capacity. They are indeed the main bugbears of failed policy experiments across the world. Improved political commitment to better education, training, evidence-informed policy processes (including local, indigenous knowledge systems), and better management and governance of those activities are important strategic directions for improvement (p. 234, 243). (See Cloete et al.[8,9] for more details on how to achieve this.) In order to achieve these transformations, the debilitating brain drain from Africa to more developed nations also clearly must be turned around (p. 253–256).

Significant improvements in the nature and constructive impacts of sustainable, decolonised, democratic governance in African countries are possible as explained elsewhere.[10,11] The author also correctly identifies better use of Fourth Industrial Revolution technologies in Africa as a prerequisite for success. However, he acknowledges that the current weak state of digital empowerment on the continent is a potentially serious obstruction (p. 267).

The book's final conclusion is that:

> *…considerable change will have to occur if Africa wants to improve its scientific systems and capabilities to serve the interests of the countries and the people on the continent… After more than half a century of independence, it is time that Africa grows the science, technology and innovation that will ultimately take the continent to an advanced level of development.* (p. 287)

This conclusion is essentially correct, but it is probably over-idealistic, given the current African reality of political instability, ideologically bad economic decisions, inadequate and unreliable infrastructure and services as well as weak democratic governance practices. In reality, colonial policy still dominates the political psyches of many African nations, 80 years and four generations since decolonisation. This is a serious indictment – both of the respective colonial powers and of the post-colonial governments.

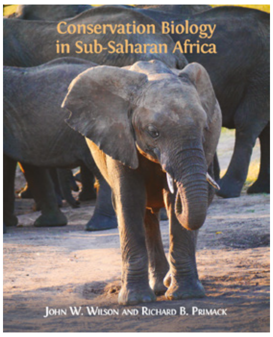
This book suggests a number of practical policy strategies to improve scientific contributions to development in Africa and identifies the main constraints and obstacles that need to be overcome in order to achieve this goal. However, despite the author's acknowledgement of some recent progress, the short- and medium-term outlooks to achieve this goal remain bleak.

The largely descriptive nature of the book, a significant degree of repetition in writing style and the lack of succinct chapter summaries of dense, technical content, complicate the reading experience, but do not overshadow the strengths of the book. With this publication, Sooryamoorthy has made a valuable new technical and reference contribution to the study of science and development in Africa, albeit that the price of the book is steep in South African monetary terms. His contribution, however, would have been even more significant had he undertaken a final concluding assessment of the extent to which the practise of science in colonial and post-colonial Africa relates to the theoretical explanatory and predictive frameworks that he used to structure his approach to the research.

References

1. Gibbons M, Limoges C, Nowotny H, Schwartzman S, Scott P, Trow M. The new production of knowledge: The dynamics of science and research in contemporary societies. London: Sage Publications; 1994.
2. Krishna VV, Waast R, Gaillard J. Globalization and scientific communities in developing countries. In: UNESCO World Science Report 1998. Paris: UNESCO; 1998. p. 273–287.
3. Goldemberg J. What is the role of science in developing countries? Science. 1998;279:1140–1141. https://doi.org/10.1126/science.279.5354.1140
4. Moravcsik MJ. Two perceptions of science development. Res Policy. 1986;15:1–11. https://doi.org/10.1016/0048-7333(86)90018-1
5. Schott T. Ties between center and periphery in the scientific world system: Accumulation of rewards, dominance and self-reliance in the center. J World Syst Res. 1998;4:112–144. https://doi.org/10.5195/jwsr.1998.148
6. Nagtegaal LW, De Brun RE. The French connection and other neo-colonial patterns in the global network of science. Res Eval. 1994;4:119–127. https://doi.org/10.1093/rev/4.2.119
7. Castells M. The university system: Engine of development in the new world economy. In: Ransom A, Khoo S-M, Selvaratnam V, editors. improving higher education in developing countries. Washington DC: The International Bank for Reconstruction and Development / World Bank; 1993. p. 65–80.
8. Cloete F, De Coning C, Wissink H, Rabie B, editors. Improving public policy for good governance. 4th ed. Pretoria: JL Van Schaik Publishers; 2018.
9. Cloete F. The complex dynamics of evidence-informed policy change. Administratio Publica. 2016;25(1):93–120. https://journal.assadpam.net/index.php?journal=assadpam&page=issue&op=viewIssue&path%5B%5D=36&path%5B%5D=28
10. Cloete F. Measuring progress towards sustainable development in Africa. Afr J Public Affairs. 2015;8(3):51–74. https://repository.up.ac.za/bitstream/handle/2263/58167/Cloete_Measuring_2015.pdf?sequence=1&isAllowed=y
11. Cloete F, Auriacombe C. Revisiting decoloniality for more effective research and evaluation. Afr Eval J. 2019;7(1):1–10. https://doi.org/10.4102/aej.v7i1.363

BOOK TITLE:
Conservation biology in sub-Saharan Africa



AUTHORS:
John W. Wilson, Richard B. Primack

ISBN:
9781783747504 (softcover, GBP42.95)
9781783747511 (hardcover, GBP59.95)
9781783747535 (EPUB, GBP5.99)
9781783747528 (free PDF)

PUBLISHER:
OpenBook Publishers, Cambridge, UK

PUBLISHED:
2019

REVIEWER:
Brian W. van Wilgen[1]

AFFILIATION:
[1]Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Stellenbosch, South Africa

EMAIL:
bvanwilgen@sun.ac.za

HOW TO CITE:
Van Wilgen BW. A valuable resource for African conservation students. S Afr J Sci. 2021;117(5/6), Art. #9310. https://doi.org/10.17159/sajs.2021/9310

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

PUBLISHED:
28 May 2021

A valuable resource for African conservation students

Conservation biology is a discipline that strives to document the earth's biological diversity, to investigate how it is influenced by humans, and to find sustainable ways to protect or restore species or ecosystems. These are extremely important aims, given the significance of healthy ecosystems for the survival of life on earth as we know it. It is a field that needs to attract and develop the best talent available if these issues are to be effectively addressed. Conservation biology is also a discipline in crisis. Earth's ecosystems are being irreparably damaged, and species are being driven to extinction, at accelerating rates. Conservation biologists therefore have to operate under considerable pressure and have to make far-reaching decisions against tight deadlines and often with very limited funding. Students and teachers of conservation biology in Africa face additional challenges in that, up to now, textbooks that cover the discipline have been prohibitively expensive; for example, Richard Primack's book *Essentials of Conservation Biology*, and Martha Groom's book *Principles of Conservation Biology* retail at ZAR2300 and ZAR2400, respectively. These books are also typically illustrated with examples of species, ecosystems and practices from outside of Africa. It is against this background that John Wilson and Richard Primack's book is such a welcome contribution. It covers all the principles of the discipline and illustrates them with a wealth of examples from across sub-Saharan Africa. It is well written, profusely illustrated with colour photographs and diagrams, and each chapter concludes with a brief summary, suggested topics for discussion, suggested readings, and a bibliography. In addition, and more importantly, the entire book can be downloaded free of charge, placing it within the reach of students who ordinarily simply would not be able to afford it.

The senior author of this book (John Wilson) is a South African environmental scientist who now works in the USA using remote sensing for wildlife movement mapping, environmental monitoring, ecological restoration, and protected areas management. Richard Primack is a professor of biology at Boston University, with an impressive record of publication in the field of conservation biology, including several textbooks. Additional case studies from 29 African countries have been contributed by 59 additional authors, enhancing the relevance of this book across the continent.

Reading through this book will leave students with no doubt as to the challenges they will face if they choose to pursue a career in conservation. One chapter describes 'the scramble for space' in which landscapes across the continent are being fragmented by burgeoning human populations. This leads to widespread local extinctions as a result of habitats dwindling to isolated pockets that cannot support viable populations of species. Another chapter describes how climate change is already affecting the ability of many species to survive, and how this problem is set to accelerate rapidly over the next few decades. A third chapter addresses the lethal cocktail of pollution, overharvesting, invasive species and disease. The point is made that many of these threats to biodiversity do not lead to immediate and/or direct mortality, but instead have sub-lethal effects, and that responses to these 'silent, insidious and easily-overlooked' threats are often delayed until the problem becomes unmanageable. Readers are also reminded that species and ecosystems are seldom exposed to only one threat. Multiple threats are compounded, and any effective conservation strategy would have to deal with all these threats collectively.

With this background, the book moves on to what people could possibly do to conserve the vast array of species and ecosystems upon which we all depend. Extinction, for example, can be avoided by ensuring that areas of vital habitat remain connected, or by boosting populations with captive-bred individuals. The case of the northern white rhinoceros is provided as an example of how advanced 'assisted reproduction technologies' may even save a species that is now functionally extinct (the last male individual has died, and only two post-reproductive female individuals remain). However, frozen sperm is available, and eggs have been grown from the ovary tissue of deceased female rhinos. Scientists are cautiously optimistic that embryos can successfully be implanted into female southern white rhinos, and that the species could be saved.

It is not just species that are important, but whole ecosystems that need to be conserved. Protected areas are one of the pillars of conservation, but most of the world's protected areas were established without considering the strategic placement that would be needed to ensure their effectiveness. The book sets out the robust principles that are now available to guide the effective placement, size, spacing and configuration of new protected areas, and there has been some progress in this regard – 17% of sub-Saharan Africa's land surface is now included in over 7500 protected areas, many of them recently established. However, this does not mean that they are safe from a myriad of threats, and the principles by which they will have to be maintained through active management and the creation and enforcement of legal instruments are also described.

The book concludes, as it must, with 'an agenda for the future'. This stresses the need for adopting a vision that will take humanity past the immediate crisis, into a sustainable course of economic development. Such a strategy would have to satisfy our present and future needs and move away from the model of unsustainable growth that currently dominates all thinking. All of this is going to require strong leadership from a new breed of individuals – trained conservation biologists who fully understand the consequences of our current path to self-destruction in a rapidly urbanising world that increasingly sees itself as divorced from nature. The authors are to be congratulated on publishing an open-access book brimming with examples relevant to Africa. My fervent hope is that it will assist in some way in the training of a cohort of passionate and persistent leaders among the next generation of conservation biologists. We are going to need them.

Check for updates

AUTHORS:
Falko Buschke[1]
Lischen du Randt[1]
Ntsu Mokhehle[1]
Izak Gouws[1]
Thia Oberholzer[1]
Witness Mamatho[1]
Sivuyisiwe Mapapu[1]
Zimkhitha Mehlomakhulu[1]
Mqondisi Mehlomakhulu[1]
Boipelo Dondolo[1]

AFFILIATION:
[1]Centre for Environmental Management, University of the Free State, Bloemfontein, South Africa

CORRESPONDENCE TO:
Falko Buschke

EMAIL:
falko.buschke@gmail.com

HOW TO CITE:
Buschke F, Du Randt L, Mokhehle N, Gouws I, Oberholzer T, Mamatho W, et al. The legal principles guiding a cohort of early career environmental professionals. S Afr J Sci. 2021;117(5/6), Art. #9298. https://doi.org/10.17159/sajs.2021/9298

ARTICLE INCLUDES:
☐ Peer review
☒ Supplementary material

KEYWORDS:
environmental values, legislation, post-normal science, value plurality

PUBLISHED:
28 May 2021

The legal principles guiding a cohort of early career environmental professionals

Environmental management has the hallmarks of a post-normal science: the stakes are high, decisions are urgent, facts are uncertain and values are often disputed.[1,2] Many modern environmental problems are intractable and cannot be solved using evidence-based tools from narrow scientific disciplines.[3] Instead, managers must juggle competing priorities while negotiating various development and conservation trade-offs. This introduces unavoidable subjectivity in the way we manage nature. Sound judgement becomes as relevant as measurement, analysis and optimisation.

Environmental professionals do not all fit the common stereotype of the green activist. Studies have consistently shown, both nationally[4] and internationally[5,6], how environmental professionals hold diverse values. These values vary from ecocentrism, that is, the belief that humans have a duty to protect nature for its own sake, to anthropocentrism, which is the belief that nature should be managed to improve human well-being. Despite ethical pluralism, one would expect that environmental legislation supplies a common set of rules that apply to all, regardless of personal values.

In South Africa, the *National Environmental Management Act* 107 of 1998 (NEMA) is the overarching law that guides environmental management. Chapter 1 of the Act outlines the National Environmental Management Principles – a set of ideals that underpin environmental regulations nationally. These principles are non-hierarchical, so each should be equally important. But environmental managers are human and would naturally favour principles that resonate with their own core beliefs. We explored how a cohort of early career environmental professionals (the authors of this paper) prioritises the principles set out in NEMA. First, we explored whether there is general consensus on which principles are considered relatively more important. Second, we explored patterns in these principles and determined whether certain preferences tend to be complementary or mutually exclusive.

Prioritising NEMA principles

As a substitute to a face-to-face lecture on environmental legislation necessitated by the national lockdown due to COVID-19, students in the master's programme in environmental management at the University of the Free State were asked by the course instructor (F.B.) to reflect on Chapter 1 of NEMA and identify the three principles that they considered most important. Each student was tasked with writing an essay of no more than 1000 words justifying their selection of their three most important principles. This paper is a synthesis of the students' (coauthors') reflections.

The master's programme is a part-time degree for early career environmental professionals. The nine participants had varying years of experience (1–15 years) in the private, public or academic sectors. Thus, they met the broad definition of 'experts' by having substantive knowledge on environmental management, the normative ability to communicate environmental judgements, and the adaptive ability to apply knowledge under new circumstances.[7] Expert performance is often uncorrelated with the perception of expertise[8], so the fewer years of experience of some respondents should not invalidate their judgements.

Chapter 1 of NEMA is made up of Section 2, with four sub-sections. Section 2(1) refers mainly to the position of NEMA in the South African legal landscape. Section 2(2) refers to the need of environmental management to put people's needs at the forefront, while Section 2(3) outlines how development should be socially, environmentally and economically sustainable. Section 2(4) is made up of 18 sub-sections (*a-r*), which describe the guiding ideals of the Act. Therefore, if we consider Sections 2(2), 2(3) and the subsections of 2(4), respondents had a set of 20 principles from which to select their three most important ones. The elicitation process was, therefore, similar to the IDEA protocol[9] (Investigate, Discuss, Estimate, Aggregate). First, participants selected their three most important principles individually (Investigate). Second, ideas were summarised anonymously by F.B. in a draft manuscript and circulated to participants for comments (Discuss). Third, participants could again make private comments on the collective contributions of the group (Estimate). Lastly, the second round of responses was combined into this final version (Aggregate).

Perspectives on NEMA principles

In general, there was a lack of consensus on the most important principles in NEMA and 14 of the 20 principles were prioritised by at least one respondent (Figure 1). Principle 2(4)*a* was the most frequently selected principle, being chosen by four of the nine respondents. This principle is the longest in the chapter because it describes the considerations of sustainable development, including the mitigation of impacts, the risk-averse precautionary principle, and keeping natural resources within sustainable limits. One respondent justified their choice by explaining that 'it is not always possible to stop development...but if developments take place, measures should ensure that the environment is not totally degraded'. A second respondent wrote 'this principle is important because it compels managers to not only consider the immediate, but also the long-term, impact of development', which was echoed by another: '[this principle] aims to ensure present needs are being achieved without jeopardising future generations'. Thus, these views place the onus on developers to ensure intergenerational equity.

Section 2(4)*f* was prioritised by three participants. This section focuses on the participation of all interested and affected parties in environmental governance, especially disadvantaged and vulnerable individuals. The reasons for prioritising this principle varied from those of social justice ('informal rights and community customs are recognised by law and should be protected'), to operational pragmatism ('marginalised people can delay projects if engagement is superficial and does not truly consider their concerns'). Section 2(4)*p,* which describes the polluter-pays principle, was also selected by three respondents. One respondent justified their choice by explaining that 'those who act irresponsibly must face the consequences and costs of remediation'.

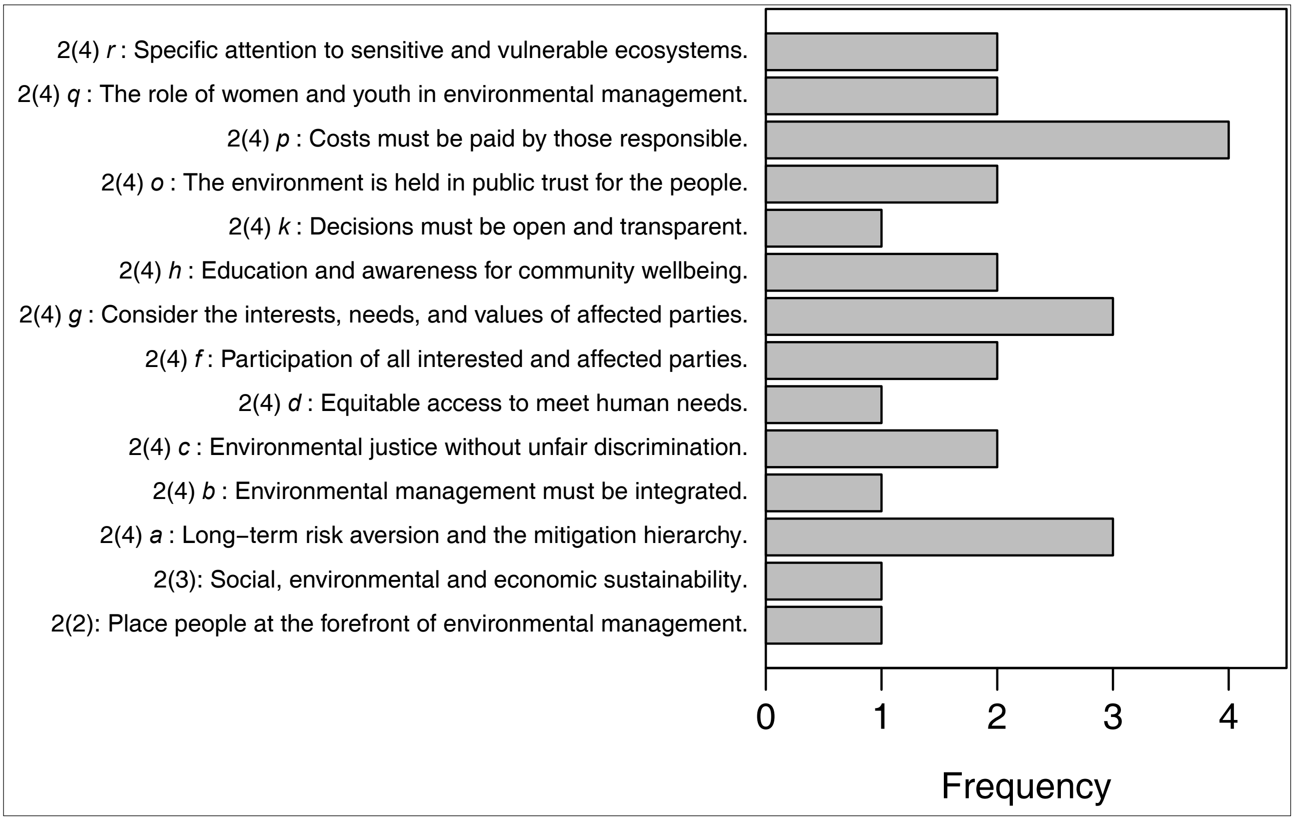


Figure 1: The frequency at which principles from the *National Environmental Management Act* were prioritised by a cohort of early career environmental professionals.

A further six principles were each selected by two respondents (Figure 1). Many of these also focused on the human side of environmental management (Sections 2(2), 2(4)*d*, 2(4)*g*), but they included that environmental management should integrate social, economic and environmental considerations (2(3) and 2(4)*b:* 'this compels managers to consider aspects that are often overlooked by development') and that decisions should be transparent (2(4)*k*: 'lack of transparency can lead to corruption').

A remaining five principles were each prioritised by only one respondent (Figure 1). Intriguingly, one of these is perhaps the most ecocentric principle in NEMA, Section 2(4)*r*, which requires that sensitive, vulnerable, highly dynamic or stressed ecosystems require specific attention by managers and planners. While the respondent highlighted the fragility of ecosystems, their justification also stated that 'degraded land with low ecological significance should be prioritised for development'. This demonstrates how even ecocentric views can be expressed in the context of development.

Of the six principles that were not selected by any respondent, three concerned issues of environmental governance (2(4)*l* – the need for intergovernmental coordination, 2(4)*m* – resolving conflicts of interest amongst organs of state, 2(4)*n* – implementing international commitments to further national interests). The remaining unselected principles referred to ensuring environmental health and safety through the whole project life cycle (2(4)*e*), considering all impacts in decision-making (2(4)*i*), and that workers have the right to refuse work harmful to themselves or the environment (2(4)*j*).

Identifying archetypes of environmental priorities

There was a lack of consensus about which NEMA principles were most important, which is understandable because these principles are supposed to be equal under law. Nevertheless, we explored whether respondents who favoured certain principles would also be more or less likely to favour others. We quantified this using a cluster analysis based on whether prioritised principles tended to be selected together (Figure 2).

Two clear clusters emerged. The first cluster (grey in Figure 2) reflected humans as the focus of environmental management. Thus, this cluster related to *why* we ought to implement environmental management. The anthropocentric vision of this cluster was encapsulated by one respondent who explicitly stated that 'a holistic take-home message is that people are at the centre of any and every form of environmental planning, management or decision-making'. By contrast, the second cluster (black in Figure 2) included principles related to sustainability, transparency, risk-aversion and public participation, and can be interpreted as answering questions about *how* we ought to implement environmental management. This cluster was described by one respondent who wrote that these principles 'enable authorities to make environmentally-centred decisions with the aim of providing an environment that is not harmful to future generations'. One exception to this separation of the *why* and *how* clusters was the general interpretation of the polluter-pays principle, Section 2(4)*p.* This principle could be interpreted operationally (i.e. who is liable for environmental damage?), but it seems as though respondents interpreted it in terms of fairness (a social justice issue) and not in terms of accountability (a legal liability issue). Nevertheless, the existence of two clusters seems at odds with a recent argument that environmental management tools are constrained by the anthropocentric ethical position[10] (although, admittedly, pluralistic environmental values do not preclude narrow environmental implementation).

NEMA: Dividing wedge or unifying foundation?

Here, a relatively small cohort of environmental professionals did not prioritise legal principles the same way. This suggests that unifying environmental professionals under a common creed[12], similar to the Hippocratic Oath for medical practitioners, is unlikely to be successful.

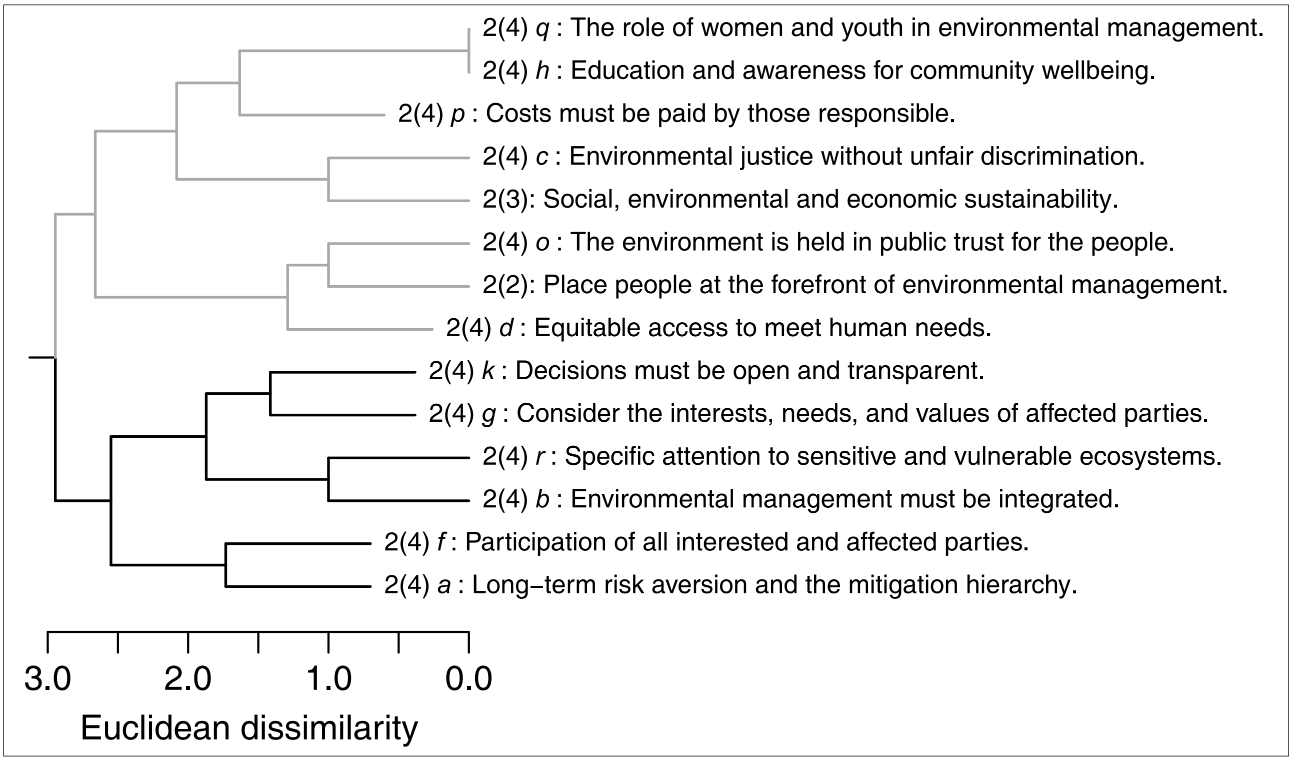


Figure 2: A cluster analysis of principles from the *National Environmental Management Act* that tended to be selected together by early career environmental professionals. Here, minimum Euclidean dissimilarity of 0 implies that the same principles were always selected together, while a maximum Euclidean dissimilarity of 3 denotes principles that were never selected by the same respondents. Cluster analysis was based on Ward's hierarchical agglomerative clustering method on a Euclidean distance matrix in R version 4.0.2.[11]

Pluralistic values can improve the efficacy of environmental management[6,13], but only if it avoids the pitfalls of factionalism. There need not be an ideological battle between those who prioritise human needs and those who prioritise the environment.

At the start of this millennium, Adams and colleagues[14] articulated four independent ways in which nature conservation is related to development aspirations. Their framework is particularly relevant in South Africa, where environmental and development ambitions regularly conflict. First, environmental protection could be seen as completely independent of development. Second, environmental protection is constrained by poverty, so development is a means to more effective conservation. Third, environmental protection is a means to achieve development and poverty alleviation. Fourth, environmental protection and development are mutually dependent and should not be viewed separately. None of these perspectives is superior to the others, but effective and sustainable environmental protection requires awareness of such ethical pluralism.[15] After reading a first draft of this manuscript, one of the contributors noted their surprise at the results: 'I didn't intentionally interpret these principles as human- or environment-focused'. This contributor went on to share a personal experience about differing perspectives: 'We don't all see the world and development through the same eyes. This is something I have experienced in my work life [as an environmental manager], especially when there is a bunch of engineers around'.

Post-normal science allows for differing worldviews. Environmental problems are complex and cannot be solved algorithmically. Instead, solving most environmental problems requires deliberation around scientific evidence while considering values, trade-offs and political feasibility. In post-normal science, authenticity is as important as expertise[1,2] because effective deliberation is only possible when all stakeholders believe they are negotiating in good faith (rather than battling hidden agendas). Sincere dialogue between stakeholders can help define a shared consensus and bring together different worldviews. Based on this, we make two recommendations for the education and career development of resilient environmental professionals. The first is a call for critical introspection around our own values and ethical priorities. An essential form of professional development for environmental managers is understanding our own motivations, worldviews and biases. This should not entail conformation to the perceived norms of the environmental sector, but rather an appreciation that our own set of beliefs is only one of many possibilities. The second recommendation is that we make a concerted effort to acknowledge that others do not necessarily interpret the world in the same way we do. Rather than imposing our own worldviews on others, it might be more effective to pursue common goals even when motivations might differ.[16] If environmental professionals understand the social, political and ethical contexts of their work, they are more likely to realise their vision for a sustainable future that supports human and ecological flourishing.[17]

Acknowledgement

We thank Jane Carruthers, Bettine van Vuuren, and Dirk Roux for their guidance on preparing this contribution.

Competing interests

We have no competing interests to declare.

Data availability

All data are included as a self-contained R-script in the supplementary material.

References

1. Ravetz JR. Post-normal science and the complexity of transitions towards sustainability. Ecol Complex. 2006;3:275–284. https://doi.org/10.1016/j.ecocom.2007.02.001

2. Buschke FT, Botts EA, Sinclair SP. Post-normal conservation science fills the space between research, policy, and implementation. Conserv Sci Prac. 2019;1, e73. https://doi.org/10.1111/csp2.73

3. DeFries R, Nagendra H. Ecosystem management as a wicked problem. Science. 2017;356:265–270. https://doi.org/10.1126/science.aal1950

4. Wilhelm-Rechmann A, Cowling RM, Difford M. Responses of South African land-use planning stakeholders to the New Ecological Paradigm and the Inclusion of Nature in Self scales: Assessment of their potential as components of social assessments for conservation projects. Biol Conserv. 2014;180:206–213. https://doi.org/10.1016/j.biocon.2014.10.012

5. Sandbrook C, Scales IR, Vira B, Adams WM. Value plurality among conservation professionals. Conserv Biol. 2010;25:285–294. https://doi.org/10.1111/j.1523-1739.2010.01592.x

6. Holmes G, Sandbrook C, Fisher JA. Understanding conservationists' perspectives on the new-conservation debate. Conserv Biol. 2017;31:353–363. https://doi.org/10.1111/cobi.12811

7. Martin TG, Burgman MA, Fidler F, Kuhnert PM, Low-Choy S, McBride M, et al. Eliciting expert knowledge in conservation science. Conserv Biol. 2012;26:29–38. https://doi.org/10.1111/j.1523-1739.2011.01806.x

8. Burgman MA, McBride M, Ashton R, Speirs-Bridge A, Flander L, Wintle B, et al. Expert status and performance. PLoS ONE. 2011;6, e22998. https://doi.org/10.1371/journal.pone.0022998

9. Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC. A practical guide to structured expert elicitation using the IDEA protocol. Methods Ecol Evol. 2018;9:169–180. https://doi.org/10.1111/2041-210X.12857

10. Bond A, Pope J, Morrison-Saunders A, Retief F. Taking an environmental ethics perspective to understand what we should expect from EIA in terms of biodiversity protection. Environ Imp Assess Rev. 2021;86, Art. #106508. https://doi.org/10.1016/j.eiar.2020.106508

11. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

12. Child M. The Thoreau Ideal as a unifying thread in the conservation movement. Conserv Biol. 2009;23:241–243. https://doi.org/10.1111/j.1523-1739.2009.01184.x

13. Robinson JG. Ethical pluralism, pragmatism, and sustainability in conservation practice. Biol Conserv. 2011;144:958–965. https://doi.org/10.1016/j.biocon.2010.04.017

14. Adams W, Aveling R, Brockington D, Dickson B, Elliott J, Hutton J, et al. Biodiversity conservation and the eradication of poverty. Science. 2004;306:1146–1149. https://doi.org/10.1126/science.1097920

15. Spash CL. The shallow or the deep ecological economics movement? Ecol Econ. 2013;93:351–362. https://doi.org/10.1016/j.ecolecon.2013.05.016

16. Jarvis RM, Borrelle SB, Forsdick NJ, Pérez-Hämmerle K, Dubois NS, Griffin SR, et al. Navigating spaces between conservation research and practice: Are we making progress? Ecol Sol Evid. 2020;1, e12028. https://doi.org/10.1002/2688-8319.12028

17. Johns D. Adapting human societies to conservation. Conserv Biol. 2010;24:641–643. https://doi.org/10.1111/j.1523-1739.2010.01510.x

AUTHORS:
Vukosi Marivate[1]
Philip Aghoghovwia[2]
Yaseera Ismail[3]
Faheema Mahomed-Asmail[4]
Sandy-Lynn Steenhuisen[5]

AFFILIATIONS:
[1]Data Science for Social Impact Research Group, Department of Computer Science, University of Pretoria, Pretoria, South Africa
[2]Department of English, University of the Free State, Bloemfontein, South Africa
[3]Quantum Research Group, School of Chemistry and Physics, University of KwaZulu-Natal, Durban, South Africa
[4]Department of Speech-Language Pathology and Audiology, University of Pretoria, Pretoria, South Africa
[5]Department of Plant Sciences and Afromontane Research Unit, University of the Free State, Phuthaditjhaba, South Africa

CORRESPONDENCE TO:
Vukosi Marivate

EMAIL:
vukosi.marivate@cs.up.ac.za

HOW TO CITE:
Marivate V, Aghoghovwia P, Ismail Y, Mahomed-Asmail F, Steenhuisen SL. The Fourth Industrial Revolution – what does it mean to our future faculty? S Afr J Sci. 2021;117(5/6), Art. #10702. https://doi.org/10.17159/sajs.2021/10702

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
technological revolution, emerging technologies, societal impact, artificial intelligence, South Africa

PUBLISHED:
28 May 2021

The Fourth Industrial Revolution – what does it mean to our future faculty?

The future of a country like South Africa is predicated upon policies; whether these policies are effective or ineffective is not primarily an issue. The prospects and realities of the Fourth Industrial Revolution (4IR) have proven to be shaping strategic policies across various spheres of national life already, including the national government, academia, civil society and the private sector. Ultimately, as strategic policies begin to take shape and come along, there is a need to pose certain important questions: what direction(s) and against what context(s) is the 4IR being embraced? In this Commentary, authored by young faculty, we discuss and debate some of the strategic recommendations of the South African Presidential Panel of the Fourth Industrial Revolution[1], specifically 'securing and availing data to enable innovation', 'incentivising future industries, platforms and application of 4IR' and 'building 4IR infrastructure'. We look at the historical context of such recommendations, and identify advantageous positions as well as gaps that may need more discussion. We then ask: What does the 4IR truly mean for our future academics and researchers?

History and context

The descriptive 'Fourth Industrial Revolution'[2] is certainly not a unique term because it has precedents – that is, it is preceded by the First, Second, and Third Industrial Revolutions[2]. The term 'revolution' that the numbers and the word 'industrial' qualify is generative. As such, glossing over this word, as if it does not bear out in consequential ways, is not wise. If South Africa is to fully prepare itself to take advantage of the wave of technological and industrial change wrought by this technological revolution, it needs to take seriously the full implications of what is afoot, including what is meant by the term 'revolution'.

A revolution is by its very nature a violent process. It does not really matter whether such a revolution is literal (as in political) or metaphoric (as in industrial). Revolutions are an existential struggle between two protagonists of history. And as with any other struggles or collisions between two forces of history, the process is invariably violent. Even when it is not spectacularly violent in real time, it is profoundly violent in its cumulative impacts in the longue durée[3], resulting in epochal shifts that leave in their wake winners and losers. Indeed, the ultimate winner goes away with everything – it is a winner-takes-all process. By winners, we mean those who embrace or are taken in by the technological changes, and by losers, we mean those whose reality – technological and educational station, cultures, and subjectivity – presents impediments to their availing themselves of the change.

The First Industrial Revolution began soon after 1784 with James Watts' invention of the steam engine that led to mechanised farming in Britain and beyond. The Second Industrial Revolution was a consolidation of the First. It is marked by the invention and smelting of steel, which resulted in the invention of other locomotive engines, including motor vehicles in the late 19th and early 20th centuries. But without harkening back too far into these past industrial epochs (i.e. 1IR and 2IR), the 3IR provides important lessons pertaining to the losers and victims it has produced, even in the face of the much-vaunted industrial and technological progress. The 3IR probably started earlier, but it took hold in the 1980s and 1990s, consisting of computer technologies and advanced modes of mechanical production, including the use of robots, for industrial manufacturing. With the expansion of the Internet and other uses of digital technology, the 3IR democratised computer technology for individual and civil industrial usage beyond the hitherto military and big government preserve. Despite this democratisation of technological progress, digital technology also rendered obsolete other forms of mechanical production, creating in effect degraded and denuded landscapes and redundant people – workers whose life and culture became 'backward' and divorced from the network and circuits of flows created by the emergent technology.

Think for instance of 'the Rust Belt' from the USA to China, areas left behind by the shifts in the forms of industrial production. The job losses as a result of the migration of industries from one locality to the other more suited to the efficiency (and profits) wrought by the existing technology, are not factored in as part of the technological 'progress'. Think also of the mines in the Congo and other locations where minerals – such as cobalt, coltan and other tantalite materials, uranium, platinum and copper – are extracted for the production of microchips and circuit boards for computers and smartphones. The two sites described here constitute the material detritus of digital evolution and culture, indexical of what Naomi Klein[4], writing in a different context, calls 'sacrifice zones' of people and places: discounted as collateral damage, the essential costs of technological progress for 'the greater common good'[5].

The 4IR is a platform that seeks to accelerate, at scale, the existing networks of flow for goods and services and to transfer all modes and markers of being into the virtual, using artificial intelligence (AI) as a catalyst. What would make this revolution different from the previous? As yet, existing technology suggests that the materiality of the 4IR, and therefore its implicit hegemony, remain the same as that of the 3IR. In fact, it is upon the digital technologies of the 3IR that the 4IR is being built. Moreover, the 4IR seeks, among other things, to design technological models that enable businesses and modes of production to do away with people and, by extension, the community. Thus, accelerated development for some, sacrificial zones and people for others. How do we avoid creating losers in the same manner as the previous IRs? Inevitably, what sites and what people is South Africa prepared to sacrifice in order to achieve the 4IR? And once identified, can South Africa be forthright in informing them for them to prepare themselves for the coming revolution?

Reacting and not leading

It is imperative to understand the disruptiveness that the 4IR presents, and the pros and cons that it constitutes to our livelihoods, which will require substantial effort to educate and inform society at all levels. Several of the technologies driving this revolution are blockchain, artificial intelligence (AI), biotechnology, nanotechnology, quantum technology, cloud computing, the Internet of things, 3D printing and autonomous vehicles.[2] It has been identified that the major common thread through many of the different technological pillars of the 4IR is the manipulation of data and information processing. By 2025, it is projected that the world will have 163 trillion gigabytes of data. As connectivity is the underlying pillar of the 4IR, there is a dependence on a communication infrastructure that is trusted and secure. The security risk these technologies pose is identified, and the growing importance of data coupled with underestimation of the cybercrime threat has contributed to the vulnerability of South African businesses.

Software improvement techniques are not the only method to mitigate these vulnerabilities: a full scope of recent advancement in technology needs to be considered. Quantum technology and nanotechnology have the potential to play a vital role in enhancing encryption techniques. Conventional methods of ensuring the security of information are based on the complexity of a mathematical construct. However, with the current increase in powerful resources, the security of information is not guaranteed; more so a breakthrough in mathematics could instantaneously make classical cryptography vulnerable. With the growth of resources, it is imperative that secure methods of transferring information are achieved. 'Quantum information processing and communication' brings together the science of quantum mechanics and information science. The aim of the field is to provide the next generation of information and communication tools in the form of quantum computing and quantum communication providing a secure and trusted network. The potential applications of quantum technology, although it is in its infancy, should be considered to address the needs of the challenges faced with upholding the security of information. As an emerging technology, the potential for innovation and subsequent commercialisation is enormous and needs to be optimally exploited for the benefit of society and to address the socio-economic challenges of the current times.

To ensure that we are leading and not reacting, it is necessary that we tap into the resources and skills available to build an ecosystem of equal opportunity in the 4IR space across all provinces in South Africa, including training the workforce for the skill transformation that is required to benefit from the rapidly changing trends. This should not be limited to artificial and machine learning; instead, there should be an incentive to develop 4IR centres/hubs for research and development across all the provinces.

It has been identified that for South African enterprises to remain globally competitive, there is a necessity to adopt technologies that improve the efficiency of operations. To address these needs, we will require further investment into research and development and the establishment of internship programmes in the industrial sector. Universities can play a vital role in the research and development phase, as this will provide a pipeline of skilled individuals to enter the workforce. The graduate unemployment rate is 10% for those aged 25–34 and 33.5% for those aged 15–24. By encouraging linkages between universities and the private sector, this rate could be curbed. Imperatively this would encourage closing the gap between industry and academia. Industrial stakeholders would need to be encouraged to partner with universities and Technical and Vocational Education and Training colleges, with equal opportunity across all provinces, to drive the change that is essential. To a large extent, this would also require training society to identify 4IR opportunities, as it is not fully understood how these technologies can be embedded into our daily lives.

There should be an incentive for industry to invest in these high-tech projects that could lead to the growth of the country in the 4IR space and place South Africa at an opportune position in various areas of expertise. A substantial influx of investment would be required to advance towards 4IR, which includes encouraging SMMEs and an ethos for entrepreneurship. Ultimately it summates to, are we as a country ready to make this paradigm shift?

Getting the basics right

On a global scale, the 3IR brought about the advent of computers leading to the development of electronics, smartphones, the Internet and automation which is said to have increased productivity, efficiency and worldwide virtual connection. As with the current report[1], 3IR also promised to eliminate income poverty and inequality, as well as increase employment for South Africans. However, this has not come to life, and South Africans still face high unemployment rates and inequality has been vastly exacerbated. If we look at our basic education system, according to the Department of Basic Education's Action Plan report, it concedes that technology-enhanced learning has not advanced in South Africa as predicted.[6] About 48% of schools do not have any digital devices available to them, let alone the skills to utilise these devices.[7] Furthermore, the gap to access digital platforms in the current education system became very clear during the 2020 lockdown in South Africa due to the COVID-19 pandemic. Only 7.7% of households in Africa were estimated to have a computer at home prior to lockdown.[8] Many public school learners were (and still are) restricted to radio or television broadcasts, or printed textbooks and worksheets distributed to them. Furthermore, 68.4% of learners with access still reported that they had difficulty adapting to the online environment. This highlights the general lack of digital literacy among learners and educators; and even where there is digital literacy, there is a lack of access to the tools needed.

At the tertiary education level, a recent survey of undergraduate and postgraduate lecturers conducted by a local university investigating the perceptions and implementation of 4IR in university modules, revealed that some academics were confused by the terminology of 4IR and unsure if digital platforms already employed in modules were part of 3IR or 4IR (Steenhuisen S-L, Department of Plant Sciences, University of the Free State, 2019 November). Lecturers commented on using various software in statistical analyses and perhaps digital control systems in some research methodology, but few could confidently state that they were implementing or developing AI systems, drone technology or likewise in their research endeavours. Many commented that they would implement these technologies if funding and training were available. This clearly shows that even at a tertiary level we are still catching up on the 3IR without access to basic computing skills and hardware meant to have been implemented in 3IR. During the lockdown in South Africa, lecturers were forced to use social communication platforms such as WhatsApp and Telegram to teach full modules to learners without access to university communication channels off-campus, stemming from a lack of devices. While universities competed to buy limited stocks of laptops in the country to loan out to staff and students in 2020, many students expressed that they were uncomfortable receiving digital devices due to fears of safety and expense. Lecturers furthermore needed to train colleagues and students in the basic use of mobile scanning apps and the like in order to operate digitally and share handwritten student assignments. With the current evidence, it is clear that the lag in the development of basic digital literacy skills in learners and lecturers handicaps South Africa's academic cohorts in implementing 4IR without intervention.

Cisco[9] reported that South Africa, as an emerging economy, is less 'digitally ready' than its peer middle-income countries such as Chile and New Zealand. Besides the lack of infrastructure, the general affordability for access to and usage of technology for the general public remains another growing concern. Costs of airtime, data and electricity are increasing at an exponential rate, and the lack of growth in the economy leads to further inaccessibility and availability of tools brought about in 3IR for education and the economy at large to fully embrace 4IR in a way that will achieve goals such as poverty alleviation.

While the report[1] makes it appear that South Africa wants to fit in with the growing interest around the 4IR, and rightly so, significant resources are being spent by policymakers on promoting the 4IR. This is despite the limited information on what it actually means and to what extent it may impact South Africans, given that we are in a different stage of development in comparison to the global developed and developing countries. We need to understand how we can benefit from 4IR and not be harmed further by it as a result of compounding on the current resource constraints our country faces. For example, electricity, which is a much-needed commodity for 3IR and even more so for 4IR, is still a luxury for some and is becoming rather erratically available in the country as a whole.

Technological developments offer many opportunities for sustainable development, but its advancements cannot address the underlying inequalities still at play. In order to participate and lead in 4IR, South African policymakers need to avoid following what is set out by developed countries and instead focus on certain innovations and their uses in a way that can establish our country and its people. This will also assist in carving a feasible and tenable space for ourselves in the global community.

To what end?

Ultimately, as a country, we need to have a clear set of strategic goals that we are working towards. As such, we can place the report[1] within a framework of understanding how it adds to clearer strategic goals. Are we as a country going to build upon new capabilities that will make us as a nation more competitive in the globalised economy? Have we identified our strengths and weaknesses? Will we build on our strengths and address some of our weaknesses? The overarching envisaged future from the report of the Presidential Commission on the 4th Industrial Revolution[1] is that:

> *South Africa will have a globally competitive, inclusive and shared economy with the technological capability and production capacity that is driven by people harnessing the 4IR to propel the country forward towards its social and economic goals, instead of falling behind.*

When reading the report in terms of strategy and competitiveness, one does see that human capital, as well as technological development, is a top priority. The question, though, is to what end? The report correctly highlights that our current industrial strategy, or the National Development Plan, is not specific in terms of what direction it will take. We posit that the 4IR report also is affected by this challenge. If the competitive strategy will be for South Africa to export technology to the rest of the African continent, we need to identify the current strengths and opportunities, and how we can realign our different sectors to meet this goal. If we will be looking at advanced manufacturing, is it for internal consumption (can we compete against imports) or will it be the most competitive country on the continent or in the world on the specific technologies in manufacturing? The report highlights countries such as Germany and China that have benefitted from clear competitive advantage strategies that require an integrated approach to reach.

At the beginning of this Commentary we highlighted the challenges of 'losers' in this revolution, but we should be clear to 'what end' and 'how' these challenges can be used to create opportunities. What are our assets (physical, human, intellectual) as a country that we can build on for global competitiveness? What are our unique strengths, processes and procedures that give us an advantage on the continent, and beyond? Which sectors can we scale to increase both our advantage and solidify our competitiveness?[10] These questions can guide how we come to more defined strategies.

There are emerging local technologies that the country can harness by investing in them to become advanced enough to create the necessary networks for local industrial production and services. For example, we have pockets of advanced manufacturing technologies in aerospace, and an emerging AI ecosystem (with growing research, but also the need for industrial development). We have quantum technology capability that is nascent. We need to move from the 4IR as a buzzword to the reality with mid- to long-term strategies for which the country can aim. This requires a common understanding of the concepts and bringing different sectors of society into this understanding. More importantly, we need to get the basics right.

There are no shortcuts. There may be temptation to pick the low-hanging fruit in taking advantage of what this country may be able to exploit in the short term, but we need to refrain from choosing interventions that are quick and look ostentatious but ultimately have a low impact (if at all). Without thought, in both the public and private sectors, how we set up the basics as a foundation to build upon may end up being a disastrous omission. We need not look far for examples. The COVID-19 pandemic exposed the cracks in many a government's readiness to use data for decision-making.[11] This highlighted the need for better data policies and infrastructure and the implementation thereof. This was not only a challenge in South Africa but in many countries across the globe. For us in South Africa, it might have led to less situational awareness, less information available for decision-making (especially for those outside health departments as COVID-19 affected more than just health) and delays in our collective understanding. This was not because of a need for a high-tech solution, but because of the lack of a strong integrated digital foundation on which to build.

We are at an interesting point in our human development, with the rate of emerging technology disruption at a pace not seen before. We need to evaluate different directions quickly, but, ultimately, we have to make a choice on the directions we take and work to get society behind those choices.

Acknowledgements

We all are members of the first and second cohort of the Department of Higher Education and Training's (DHET) Future Professors Programme (FPP). The views expressed in this Commentary do not represent the views of DHET or the FPP. We acknowledge the editorial comments of Herkulaas MvE Combrink.

Competing interests

We have no competing interests to declare.

References

1. Presidential Commission on the 4th Industrial Revolution. Report of the Presidential Commission on the 4th Industrial Revolution [document on the Internet]. c2020 [cited 2021 Apr 12]. Available from: https://www.gov.za/documents/report-presidential-commission-4th-industrial-revolution-23-oct-2020-0000#

2. Schwab K. The Fourth Industrial Revolution. New York: Currency Books; 2017.

3. Nixon R. Slow violence and the environmentalism of the poor. Cambridge, MA: Harvard University Press; 2011. https://doi.org/10.4159/harvard.9780674061194

4. Klein N. Let them drown: The violence of othering in a warming world. London Review of Books. 2016;38(11):11–14.

5. Roy A. The greater common good. Frontline. 1999 May 22. Available from: https://frontline.thehindu.com/other/article30257333.ece

6. Padayachee K. A snapshot survey of ICT integration in South African schools. S Afr Comput J. 2017;29(2):36–65. https://doi.org/10.18489/sacj.v29i2.463

7. Van Wyk C. Survey of ICT in schools in South Africa. Public expenditure analysis for the Department of Basic Education, Report 8. Stellenbosch: Stellenbosch University; 2012.

8. Alsop T. Share of households in Africa with a computer at home from 2005 to 2019. Statista. 2021 February 18. Available from: https://www.statista.com/statistics/748549/africa-households-with-computer/

9. Cisco. White Paper: Cisco Global Digital Readiness Index 2019 [document on the Internet]. c2020 [cited 2021 Apr 12]. Available from: https://www.cisco.com/c/dam/en_us/about/csr/reports/global-digital-readiness-index.pdf

10. Van den Steen E. Creating and sustaining competitive advantage. Harvard Business School Course Overview Note 714-491, March 2014.

11. Marivate V, Nsoesie E, Combrink HMVE. Africa's responses to COVID-19: An early data science view. In: Milan S, Treré E, Masiero S, editors. COVID-19 from the margins: Pandemic invisibilities, policies and resistance in the Datafied Society. Amsterdam: Institute of Network Cultures; 2021. p. 110–112. Available from: https://networkcultures.org/blog/publication/covid-19-from-the-margins-pandemic-invisibilities-policies-and-resistance-in-the-datafied-society/

Drafting a Code of Conduct for Research under the *Protection of Personal Information Act No. 4 of 2013*

AUTHORS:
Rachel Adams[1]
Susan Veldsman[2]
Michèle Ramsay[3]
Himla Soodyall[2]

AFFILIATIONS:
[1]Human Sciences Research Council, Pretoria, South Africa
[2]Academy of Science of South Africa, Pretoria, South Africa
[3]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Susan Veldsman

EMAIL:
susan@assaf.org.za

HOW TO CITE:
Adams R, Veldsman S, Ramsay M, Soodyall H. Drafting a Code of Conduct for Research under the *Protection of Personal Information Act No. 4 of 2013*. S Afr J Sci. 2021;117(5/6), Art. #10935. https://doi.org/10.17159/sajs.2021/10935

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
POPIA compliance, POPIA, Code of Conduct, research participants' privacy, ASSAf, research community, open science

PUBLISHED:
03 May 2021

On 22 June 2020, President Ramaphosa announced that the *Protection of Personal Information Act No. 4 of 2013* (POPIA) would come into effect on 1 July 2020. A one-year grace period was provided to give organisations time to comply with the provisions of the Act. It will therefore be mandatory as of 1 July 2021, for all sectors in South Africa to comply with POPIA.

POPIA gives effect to the constitutional right to privacy. In so doing, it balances the right to privacy with other rights and interests, including the free flow of information within South Africa and across its borders. POPIA adopts a principle-based approach to the processing of personal information. It sets out eight conditions for the lawful processing of personal information: accountability, processing limitation, purpose specification, further processing limitation, information quality, openness, security safeguards, and data subject participation. These principles apply equally to all sectors that process personal information.

Chapter 7 of POPIA makes provision for the development of Codes of Conduct to provide guidance on the interpretation of POPIA in relation to a particular sector or industry, or class of information. Codes of conduct are particularly important in providing prior authorisation in terms of Section 57 of POPIA for the sector to which the Code applies. Prior authorisation is required for the processing of unique identifiers, such as ID numbers, for any purpose other than that for which they were originally collected, and for use within an information matching programme. In addition, prior authorisation is required for transferring special personal information and the personal information of children to a country outside South Africa that does not have an adequate level of data protection regulation. Further guidance on Chapter 7 of POPIA and the development of Codes of conduct was published by the Information Regulator in February 2021.[1] Once a code is approved by the Information Regulator and comes into force, it is legally binding.

The Academy of Science of South Africa (ASSAf) has begun a process to facilitate the development of a Code of Conduct for Research. In addition to providing prior authorisations for research, as set out above, the Code of Conduct is needed to provide guidance to researchers on how to rationalise the provisions of POPIA in relation to existing laws and standards regulating research. The general norm, in this instance, is that whichever law provides a greater level of protection of rights, and particularly the right to privacy, takes precedence.

This process began in 2020 following a call from South African scientists to consider the development of a POPIA Code of Conduct specifically to guide the use of personal information in research. Two public fora were held in 2020 to discuss: during Open Access Week on 21 October 2020 and at the Science Forum South Africa on 10 December 2020. Two committees – a Steering Committee and a Drafting Committee – were subsequently established by ASSAf to lead the process of developing the Code of Conduct for Research (Table 1).

It is important to note that in 2020, Universities South Africa (USAf), a membership organisation that represents public universities in South Africa, began the process to draft a Code of Conduct to help regulate the processing of personal information within higher education institutions. This Code of Conduct has not yet been submitted to the Information Regulator. Once it has been submitted and approved by the Information Regulator, this Code of Conduct will form another regulatory tool for guiding the research community to comply with POPIA. The USAf Code of Conduct does not set out extended provisions and explanations of POPIA in respect of research activities. Therefore, with respect to the processing of personal information for research purposes, the ASSAf-led Code of Conduct will take precedence over the USAf Code of Conduct.

This Commentary sets out why a Code of Conduct for Research is being developed, its purpose and scope, and why ASSAf is the body that is coordinating its development.

What is a Code of Conduct and why is it needed for research?

POPIA is to be welcomed as it gives greater guidance to researchers regarding the use and protection of personal information for research. This should serve to improve transparency, accountability and oversight of personal information and promote public trust in the use of personal information in research. However, there is uncertainty and need for further guidance on the application of POPIA to research. First, it is unclear how some of the high-level principles will apply in practice to research. Second, POPIA provides certain exceptions from the lawful conditions of processing personal information for research, and further interpretation is required to understand how and where these exceptions would apply in different research contexts. Third, it is important that there is a comprehensive and uniform approach to the regulation of personal information for research across all government departments, academic institutions, research councils and the private sector.

A Code of Conduct is a sectoral or industry-wide regulation issued under POPIA that provides further details on how the Act should be interpreted in relation to that particular sector or class of information. Codes of Conduct must not derogate from, or water down, the provisions of POPIA. Where relevant, Codes can heighten data subjects' rights, and can provide exemptions from the conditions of processing of personal information for all the bodies bound under the Code, in terms of Section 37 of POPIA.

A Code of Conduct must address all of the eight provisions for the processing of personal information in terms of the specific sector or class of information or provide for their functional equivalent where there are existing provisions in law (Section 60 (2) (a)). In addition, Codes of Conduct must provide for appropriate measures for information matching programmes and high-risk information in terms of the sector or class of information in question.

The Code of Conduct for Research is being developed to ensure compliance with POPIA by the research community in South Africa and to promote uniformity in the interpretation and application of the Act. Additionally, the Code will guide information officers, data stewards, research integrity officers, research ethics committees and other research governing structures in their roles with respect to POPIA. With respect to international collaboration in research, this Code will strive to meet international standards of data protection so as to allow for cross-border data sharing in international research projects and to enable compliance with the requirements of POPIA in relation to trans-border information flows of personal information, as well as serve as a mechanism to protect the international flow of personal information.

Of principal importance, the Code will enhance and protect the rights of data subjects (who, in research, we would call ‘research participants’) and build the trust of data subjects and the public in the functioning of the research sector. It will also stimulate transparent processing of personal information to promote cultural change for research bodies in relation to the lawful processing of personal information and serve as a mechanism to hold responsible parties accountable for the processing of personal information.

Lastly, the Code will ensure alignment with other legislation and regulation that governs the conduct of research in South Africa and promote responsible open science in line with the principles and objectives of POPIA and international best practice.

Why the Academy of Science of South Africa?

ASSAf was approached in 2020 by various scientists in South Africa to consider leading the process to develop a Code of Conduct for Research. ASSAf is the official national science academy of South Africa. It is mandated under the *Academy of Science of South Africa Act, 67 of 2001*, as amended by the *Science and Technology Laws Amendment Act, 16 of 2011*. ASSAf’s mission is to use evidence-based science to address challenges in society and to use science for societal benefit. ASSAf currently has nearly 600 members, who consistently have demonstrated academic excellence in various fields such as Agricultural Sciences, Earth Sciences, Economic Sciences, Education, Health/Medical Sciences, Humanities, Life Sciences, Mathematical Sciences, Physical Sciences, Social Sciences and Technological and Engineering Sciences.

POPIA stipulates under Section 61 (1) (b) that a body ‘sufficiently representative of any class of bodies, or of any industry, profession, or vocation as defined in the Code in respect of such class of bodies or of any such industry, profession or vocation’ can develop a Code of Conduct, to be reviewed and approved by the Information Regulator.

Given that POPIA will have a significant impact on research processes in South Africa, ASSAf has engaged widely with representatives of the scientific community to develop a single Code of Conduct for Research. ASSAf has the capacity to represent the scientific community to facilitate evidence-based research and to ensure compliance with regulations that guide research. Given the placement of ASSAf within the National System of Innovation, ASSAf has the ability to provide policy advice on matters relating to science and the governance of science. ASSAf is broadly considered ‘sufficiently representative’ in terms of POPIA and therefore best placed to develop a Code of Conduct for Research through an inclusive and consultative process.

Scope of the Code: To what and whom does it apply?

The full scope of the Code of Conduct is set out in the Discussion Document by Adams et al.[2] The proposed Code will pertain to all research conducted in South Africa or by a responsible party domiciled in South Africa, and which uses (collects, processes or stores) personal information as defined under POPIA as part of the research process.

The Code will pertain to all research activities in South Africa that ordinarily undergo prior and independent ethics review, that follow a recognised scientific methodology or system of analysis, and that aim to publish the research in contribution to the respective field of study.

The Code will further set out where there are other existing laws that pertain to the use of personal information in research and how these are to be reconciled with POPIA. These laws and regulatory instruments include the *National Health Act No. 61 of 2003*[3] and its 2012 regulations, the Department of Health’s[4] ‘Ethics in Health Research: Principles, Processes and Structures’ guidelines and the *Promotion of Access to Information Act, No. 2 of 2000*. In short, whichever law provides a stronger level of protection of the rights with regard to personal information takes precedence. This is set out in Section 3 of POPIA.

During 2019, draft guidelines were published to guide agencies in the developing of the Code of Conduct. New guidelines were published in February 2021. The main differences in the later guidelines are:

1. Requirement of notification to the Information Regulator about an intention of a relevant body (i.e. the body that develops a Code) to develop a code (Section 11);
2. Further details of the paperwork required to show engagement with stakeholders and response to inputs from stakeholders (Section 16.2.2.), including a ‘statement of consultation’;
3. Further details pertaining to the reports to the Information Regulator which relevant bodies must submit about compliance with the Code (Section 25.3); and
4. Removal of the provisions relating to alternative dispute resolutions where parties are aggrieved by the decision of the relevant body regarding a complaint.

Research in South Africa is governed by several existing legal instruments and provisions. The Constitution of the Republic of South Africa[5] provides under the Bill of Rights that ‘[e]veryone has the right to bodily and psychological integrity, which includes the right not to be subjected to medical or scientific experiments without their informed consent’ (Section 12 (2) (c)). In addition, the *National Health Act*[3] requires all research projects that involve human participants to have the express consent of the individual involved and to ‘be conducted in the prescribed manner’ (Section 71). This prescribed manner relates to any regulations which further govern research, which include, particularly, the Department of Health’s guidelines[4] noted above. These Guidelines pertain to ‘research that involves living human participants’ (para 1.1.7) and requires prospective and independent ethics review from a research ethics committee registered with the National Health Research Ethics Council.

In addition, there are standards being developed and issued globally to promote open science. Open science is intended to promote the benefit and advancement of science for all, and requires research data to be made publicly available. Such data would typically be de-identified as far as possible, and the provisions of POPIA would not apply, as POPIA does not apply to de-identified information that cannot be reasonably re-identified. This is consistent with the objectives of POPIA, set out in the Preamble[6], which include that the Act is

> *consonant with the constitutional values of democracy and openness, the need for economic and social progress, within the framework of the information society, requires the removal of unnecessary impediments to the free flow of information, including personal information.*

However, it is important in the development of this Code to take into account international standards, including those relating to open science and data protection law in the European Union and African Union, as, too, is noted in the Preamble to POPIA.[6] This is particularly important given how data protection laws worldwide provide for a provision of 'adequacy' when sharing data with institutions in other countries. This means that cross-border data sharing can only take place where the other jurisdiction has an adequate standard of data protection in place or a data access agreement in place to ensure adequate data protection.

Conclusion

It is anticipated that a Code of Conduct for Research will be submitted by ASSAf to the Information Regulator by early June 2021. This submission will follow wide consultation with researchers, research institutions and other relevant stakeholders. The Discussion Document by Adams et al.[2] sets out the main substantive issues that the Code of Conduct for Research will address, and a public consultation forum is planned for May 2021 in which the ASSAf Steering and Drafting Committees will receive further input from the research community. Through a transparent and consultative process, we hope to develop a Code of Conduct that has lasting value in guiding the research community of South Africa in complying with POPIA.

References

1. Information Regulator (South Africa). Guidelines to develop codes of conduct: Issued under the Protection of Personal Information Act 4 of 2013 (POPIA) [document on the Internet]. c2021 [cited 2021 Apr 20]. Available from: https://www.justice.gov.za/inforeg/docs/InfoRegSA-Guidelines-DevelopCodeOfConduct-22Feb2021.pdf

2. Adams R, Adeleke F, Anderson D, Bawa A, Branson N, Christoffels A, et al. POPIA Code of Conduct for Research. S Afr J Sci. 2021;117(5/6), Art. #10933. https://doi.org/10.17159/sajs.2021/10933

3. National Health Act 61 of 2003, Republic of South Africa. Available from: https://www.gov.za/sites/default/files/gcis_document/201409/a61-03.pd

4. South African Department of Health (DoH). Ethics in health research: Principles, processes and structures. 2nd ed. Pretoria: DoH; 2015. Available from: https://www.sun.ac.za/english/research-innovation/Research-Development/Documents/Integrity%20and%20Ethics/DoH%202015%20Ethics%20in%20Health%20Research%20-%20Principles,%20Processes%20and%20Structures%202nd%20Ed.pdf

5. The Constitution of the Republic of South Africa, 1996. Available from: https://www.justice.gov.za/legislation/constitution/saconstitution-web-eng.pdf

6. Protection of Personal Information Act 4 of 2013, Republic of South Africa. Available from: https://www.gov.za/documents/protection-personal-information-act#

Table 1: Steering and Drafting Committees

Name	Affiliation
Steering Committee members	
Dr Rachel Adams	Human Sciences Research Council
Prof. Ahmed Bawa	Universities South Africa
Prof. Alan Christoffels	University of the Western Cape
Prof. Jantina de Vries	University of Cape Town
Prof. Monique Marks	Durban University of Technology
Dr Mongezi Mdhluli	Medical Research Council
Dr Mapitso Molefe	Council for Scientific and Industrial esearch
Dr Tshilidzi Muthivhi	Department of Health
Prof. Caroline Ncube	University of Cape Town
Prof. Michèle Ramsay (Chair)	University of the Witwatersrand
Prof. Jerome Singh	Stellenbosch University
Drafting Committee members	
Dr Rachel Adams (Chair)	Human Sciences Research Council
Dr Fola Adeleke	University of the Witwatersrand
Dr Dominique Anderson	University of the Western Cape
Dr Nicola Branson	University of Cape Town
Dr Harriet Etheredge	University of the Witwatersrand
Ms Eleni Flack-Davison	University of the Witwatersrand
Prof. Safia Mohammed	University of South Africa
Dr Antonel Olckers	DNABiotec
Prof. Maria Papathanasopoulos	University of the Witwatersrand
Ms Jane Pillay	National Health Laboratory Service
Prof. Tobias Schonwetter	University of Cape Town
Dr Carmen Swanepoel	Stellenbosch University

POPIA Code of Conduct for Research

AUTHORS:
Rachel Adams[1]
Fola Adeleke[2]
Dominique Anderson[3]
Ahmed Bawa[4]
Nicola Branson[5]
Alan Christoffels[3]
Jantina de Vries[6]
Harriet Etheredge[7,8]
Eleni Flack-Davison[9]
Mark Gaffley[10]
Monique Marks[11]
Mongezi Mdhluli[12]
Safia Mahomed[13]
Mapitso Molefe[14]
Tshilidzi Muthivhi[15]
Caroline Ncube[16]
Antonel Olckers[17]
Maria Papathanasopoulos[18,19]
Jane Pillay[20]
Tobias Schonwetter[21]
Jerome Singh[22]
Carmen Swanepoel[23]
Michèle Ramsay[24]

AFFILIATIONS:
[1]Human Sciences Research Council, Pretoria, South Africa
[2]School of Law, University of the Witwatersrand, Johannesburg, South Africa
[3]South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa
[4]Universities South Africa, Pretoria, South Africa
[5]Southern African Labour Development Research Unit, University of Cape Town, Cape Town, South Africa
[6]Department of Medicine, University of Cape Town, Cape Town, South Africa
[7]Wits Donald Gordon Medical Centre, University of the Witwatersrand, Johannesburg, South Africa
[8]Steve Biko Centre for Bioethics, School of Clinical Medicine, University of the Witwatersrand, Johannesburg, South Africa
[9]Research Office, University of the Witwatersrand, Johannesburg, South Africa
[10]Faculty of Law, University of Cape Town, Cape Town, South Africa
[11]Urban Futures Centre, Durban University of Technology, Durban, South Africa
[12]South African Medical Research Council, Cape Town, South Africa
[13]College of Law, University of South Africa, Pretoria, South Africa
[14]Council for Scientific and Industrial Research, Pretoria, South Africa
[15]National Department of Health, Pretoria, South Africa
[16]Department of Commercial Law, University of Cape Town, Cape Town, South Africa
[17]DNABiotec, Pretoria, South Africa
[18]Assistant Dean: Research and Postgraduate Affairs, University of the Witwatersrand, Johannesburg, South Africa
[19]Director: HIV Pathogenesis Research Unit, University of the Witwatersrand, Johannesburg, South Africa

On 1 July 2021, the *Protection of Personal Information Act* (POPIA or the Act), *No. 4 of 2013*, will come into effect. The Act will have implications for all research activities that involve the collection, processing, and storage of personal information. POPIA provides for the development of Codes of Conduct to guide the interpretation of the Act with respect to a particular sector or class of information.[1] Codes of Conduct are particularly important for providing for prior authorisations in terms of Section 57 of POPIA for the sector to which it applies. Prior authorisations are required for using unique identifiers of personal information in data processing activities, and for sharing special personal information or the personal information of children with countries outside of South Africa that do not have adequate data protection laws. In order to understand and functionally interpret the provisions of POPIA for the research community in the Republic of South Africa (South Africa), the Academy of Science of South Africa (ASSAf) is leading a process to develop a Code of Conduct (Code) for research under the Act. A Code can be developed by the Information Regulator or by a public or private body deemed 'sufficiently representative' of the bodies in respect of the particular class of information or sector to which the Code will apply. During 2020, ASSAf was approached by scientists in South Africa to consider the development of a Code for research, and public events were held during Open Access Week in October 2020, and Science Forum South Africa in December 2020, to further discuss the role of ASSAf in this regard. A Commentary published in this issue sets out the full rationale for the development of the Code by ASSAf and details the consultation process to date.[2]

Within the research setting, POPIA regulates the processing of personal information for research purposes, and the flow of data across South Africa's borders to ensure that any limitations on the right to privacy are justified and aimed at protecting other important rights and interests. The new regulatory system that POPIA establishes will function alongside other legislation and regulatory structures governing research in South Africa, as outlined below. The law which takes precedent will be that which provides the most comprehensive protections to the rights of individuals in South Africa.

This paper sets out the key discussion points in relation to the development of the Code. It is intended as a paper that can support further stakeholder consultation and public engagement in the process of developing a Code which meets the needs, and is representative of, the South African research community.

Background to POPIA

POPIA provides for the lawful processing of personal information in South Africa. It sets out the roles for various parties involved in the processing (including collection, use, transfer, matching and storage) of personal information. Briefly, these roles include but are not limited to:

- the 'Responsible Party', which – in this case – is the researcher (Principal Investigator) or research institution responsible for determining why and how the personal information is being processed;
- the 'Operator' – a third party contracted by the responsible party to process personal information on their behalf;
- an 'Information Officer' who is the designated individual within an institution responsible for ensuring compliance with POPIA; and
- the 'Data Subject' who is the person whose information is being processed and, in the case of research, would be the 'study/research participant'.

The Act outlines eight (8) conditions for the lawful processing of personal information, all of which must be fulfilled in order for such processing to be lawful. These conditions are:

1. *Accountability*: the responsible party must ensure that all the conditions for the lawful processing of personal information laid out in POPIA are complied with at the time of the determination of the purpose of processing and during processing (Section 8).
2. *Process limitation*: the responsible party must ensure there is a lawful basis for the processing of personal information; that such processing is necessary for a defined purpose and could not be achieved without processing such personal information; and that the information is collected directly from the data subject and with informed consent (Sections 9–12). The lawful basis must be determined at the outset of the processing and will have an effect on the rights of data subjects. The lawful bases outlined in POPIA are[1]:

POPIA Section 11 (1)

a. *the data subject or a competent person where the data subject is a child consents to the processing;*

b. *processing is necessary to carry out actions for the conclusion or performance of a contract to which the data subject is party;*

c. *processing complies with an obligation imposed by law on the responsible party;*

d. *processing protects a legitimate interest of the data subject;*

e. *processing is necessary for the proper performance of a public law duty by a public body; or*

f. *processing is necessary for pursuing the legitimate interests of the responsible party or of a third party to whom the information is supplied.*

[20]National Health Laboratory Service, Johannesburg, South Africa
[21]Director: Intellectual Property Unit, University of Cape Town, Cape Town, South Africa
[22]Centre for Medical Ethics and Law, Stellenbosch University, Stellenbosch, South Africa
[23]Division of Haematological Pathology, Stellenbosch University, Stellenbosch, South Africa
[24]Director: Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Rachel Adams

EMAIL:
radams@hsrc.ac.za

HOW TO CITE:
Adams R, Adeleke F, Anderson D, Bawa A, Branson N, Christoffels A, et al. POPIA Code of Conduct for Research. S Afr J Sci. 2021;117(5/6), Art. #10933. https://doi.org/10.17159/sajs.2021/10933

ARTICLE INCLUDES:
☐ Peer review
☐ Supplementary material

KEYWORDS:
data protection, protection of personal information, research participants, research ethics, open science

PUBLISHED:
03 May 2021

It is important to note that because consent can be withdrawn at any time, it is not an ideal lawful basis for the processing of personal information for research purposes. In addition, there are circumstances where processing for research purposes can take place without the express consent of the data subject. Bodies which perform research functions would be advised to determine if their processing of personal information for research complies with a public law duty, such as would be the case with a research council founded through an act of Parliament, and/or whether their processing of personal information for research fulfils their legitimate interests. Further guidance in this regard will be provided under the Code.

3. *Purpose specification*: the collection and processing of personal information must be for a defined purpose; records should not be retained longer than is necessary and must be deleted or destroyed after the purpose for collection and processing has been fulfilled. The retention of records containing personal information is allowed for research purposes where there is a specifically defined need to retain such information and where further relevant safeguards are in place (Sections 13–14).

4. *Further processing limitation*: further processing of personal information is permitted where such information is used for research, and only research, purposes (Section 15).

5. *Information quality*: personal information collected and stored must be accurate, up to date, complete and not misleading (Section 16).

6. *Openness*: responsible parties must maintain a record of all processing of personal information. The data subject must be informed regarding: why the information was collected, who collected it and where it is being held, what rights the data subject has to access and delete/correct the data, and if the data will be transferred to a third party and/or internationally during the processing. It is not necessary to inform the data subject of the above if their information is being processed only for research purposes (Sections 17–18).

7. *Security safeguards*: responsible parties must ensure that personal information is kept secure to maintain confidentiality and integrity, and to prevent data breaches. Any security breaches must be reported to the Information Regulator (Sections 19–22).

8. *Data subject participation*: the responsible party must ensure that the data subject is informed of their right to access, correct and delete their personal information and of the manner in which to do so (Sections 23–25).[1]

POPIA provides for a general prohibition on the processing of special personal information. Special personal information includes information relating to the health, political persuasion, race or ethnic origin, or criminal behaviour of the data subject. There is a similar ban on the processing of personal information relating to a child. There are some exceptions to these bans, discussed below.

Existing regulatory framework

Research in South Africa is governed by a number of existing legal instruments and provisions. The Constitution of the Republic of South Africa[3] provides under the Bill of Rights that '[e]veryone has the right to bodily and psychological integrity, which includes the right not to be subjected to medical or scientific experiments without their informed consent'[3]. In addition, the *National Health Act, No. 61 of 2003* requires all research projects that involve human participants to obtain the express consent of the individual involved and to 'be conducted in the prescribed manner'[4]. This prescribed manner relates to any regulations which further govern research, which include, particularly, the South African Department of Health's[5] guidelines on 'Ethics in Health Research Principles, Processes and Structures' of 2015 (hereafter the DoH Guidelines). The DoH Guidelines pertain to 'research that involves living human participants'[5] and require prospective and independent ethics review. While the DoH Guidelines apply to 'health research', this is broadly defined as all research which contributes to the knowledge of:

- biological, clinical, psychological, or social welfare matters including processes as regards humans;
- the causes and effects of, and responses to disease;
- effects of the environment on humans;
- methods to improve healthcare service delivery;
- new pharmaceuticals, medicines, interventions, and devices; and
- new technologies to improve health and health care.[5]

Accordingly, all research projects that involve human participants – including where any personal information is collected, processed, or stored – are required to undergo a prior ethics evaluation from a suitably constituted research ethics committee, preferably registered with the National Health Research Ethics Council.

In addition, best practice guidelines for open science are being promulgated globally and nationally, with research funding agencies including the National Research Foundation now requiring that research data be made publicly available. Open science seeks to promote the benefit and advancement of science for all and open access data would typically be de-identified as far as reasonably possible to prevent direct identification of a data subject. In this circumstance, the provisions of POPIA would not apply, as POPIA does not apply to de-identified information that cannot be reasonably re-identified. This is consistent with the objectives of POPIA, set out in the Preamble, which include that the Act is

> *consonant with the constitutional values of democracy and openness, the need for economic and social progress, within the framework of the information society, requires the removal of unnecessary impediments to the free flow of information, including personal information.*[1]

However, it is important in the development of this Code to consider international standards, including both those relating to open science and data protection law in the African Union and European Union, as, too, is noted in the Preamble to POPIA.[1] This is particularly important given how data protection laws worldwide provide for a provision of 'adequacy' when sharing data with institutions in other countries. This means that cross-border data sharing can only take place where the other jurisdiction has an adequate standard of data protection or a data access agreement in place to ensure adequate data protection.

Scope of the Code

The Code, as it is currently being considered, pertains to research conducted in South Africa, or conducted by a responsible party domiciled in South Africa, and which – as part of the research process – uses (collects, processes or stores) personal information as defined under POPIA. This includes personal information that is used directly, i.e. collected directly from the data subject/research participant or that is used in the process of the research, e.g. research that uses a database which includes personal information.

As such, this Code is relevant to research – whether basic or applied – in any discipline including, but not limited to, natural sciences, engineering and technology, medical and health sciences, social sciences, education, management, economics, theology, law, and the humanities and which:

- follows a recognised scientific methodology or system of analysis, and improves or creates new knowledge, or deepens understanding; and
- would ordinarily undergo prior independent ethics review.

In this regard, this Code applies to both industry and academia and broadly takes research to mean the generation, preservation, augmentation, and improvement of knowledge by means of investigations and methods pertinent to the scientific or disciplinary field[6-8], and which is mindful of the value of knowledge for the betterment of society, including open science.

The proposed Code pertains to research that uses personal information undertaken as part of experimental development research[9], public health surveillance, statistical data collection on the part of state organs, and clinical trials, where such research is intended to be published in contribution to the respective field of knowledge.

The proposed Code will not apply to the following research or research-related activities: market research, political and public opinion polling, audits, quality assurance or programmatic monitoring and evaluation; or other research where the purpose is not directly to contribute to the improvement of knowledge through peer-reviewed publication.

Exclusions, exemptions and exceptions for research under POPIA

POPIA contains exclusions, exemptions and exceptions, some of which pertain to the processing of personal information for research purposes. The South African Law Reform Commission's report on 'Privacy and Data Protection', on which the drafting of POPIA was based, explains that exceptions 'map out the extent of the obligations under the rule – or principle', 'exemptions involves lifting a burdensome obligation from a responsible party while the burden continues to apply to others', and exclusions are 'where certain classes of responsible parties are excluded completely from the coverage of the law'.[10]

While emphasis here will be on outlining research-specific exceptions, it is important to not lose sight of the fact that some processing activities linked to research may also benefit from general or research-specific exclusions and exemptions in the Act.

Exclusions

In respect of research activities, the most pertinent exclusion relates to the processing of 'de-identified' information. This is defined as[1]:

> *"de-identify", in relation to personal information of a data subject, means to delete any information that—*
>
> *(a) identifies the data subject;*
>
> *(b) can be used or manipulated by a reasonably foreseeable method to identify the data subject; or*
>
> *(c) can be linked by a reasonably foreseeable method to other information that identifies the data subject.*

In practice, however, it is not always possible to completely de-identify data and there are certain categories of information which may not be de-identifiable, including genetic information (see section below on Genetic Data). In addition, re-identification can occur through matching or linking data sets. Under the General Data Protection Regulation of the European Union (GDPR), the term 'pseudo-anonymisation' has been used to describe information that can be re-identified. The process of pseudo-anonymisation under the GDPR is described as:

> *the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information provided, that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.*[11]

In addition, there are certain categories of information which may not be entirely de-identifiable, or may be identifying (that is, with the relevant technical and scientific understanding and resources, identifying information is information that immediately identifies an individual, such as a picture of a face), but not identifiable (not able to be identified).

Exemptions

Section 36 of POPIA[1] stipulates that the processing of personal information ultimately does not violate a condition for the lawful processing of personal information under POPIA if the Regulator grants, on a case-to-case basis, an exemption under Section 37 of POPIA, or if the processing of personal information is carried out in accordance with Section 38 of POPIA and is necessary for the fulfilment of a function of a public body.

The Regulator may grant an exemption under Section 37 if the Regulator believes that the processing activities are in the public interest, or in the interest of either the data subject or a third party, provided these interests substantially outweigh the interference with privacy. Notably, according to Section 37 (2), the public interest includes historical, statistical or research activity, as well as processing toward 'the prevention, detection and prosecution of offences'[1].

Exceptions

In addition to the exclusions and exemptions addressed above, POPIA also expressly provides that some of the POPIA conditions for processing personal information (part A) and restrictions for processing of special personal information (part B) as well as personal information of children (part C) do not (fully) apply in the context of research. These research-specific exceptions are captured in Table 1.

Consent

POPIA defines consent as any voluntary, specific and informed expression of will in terms of which permission is given for the processing of personal information. While POPIA emphasises the importance of consent being specific, it also recognises the importance of fostering research, and particularly research that is in the public interest (Section 37). In South Africa, research that involves human participants typically requires informed consent in terms of the *National Health Act* and the Constitution, and should seek prior clearance from a health research ethics committee or, in the case of a non-health discipline, another suitably constituted research ethics committee.

The DoH Guidelines endorse several forms of consent for use in health research in South Africa; these forms of consent are listed in Table 2.

It is clear that with the promulgation of POPIA, and in line with the definitions in the DoH Guidelines, a consent model for research where the data subject consented to the use of their data in future for absolute unknowns, such as in blanket consent (which is notably *not endorsed* by the DoH), would not be permissible. Instead, researchers will need to be

Table 1: Exceptions for research under POPIA[1]

POPIA Condition / obligation / provision etc.	Research-specific exception
Condition 3: Purpose specification Section 14: Record retention: 'records of personal information must not be retained any longer than is necessary for achieving the purpose for which the information was collected or subsequently processed, unless -…'	Section 14 (2): Records of personal information may be retained for periods in excess of those contemplated [in Section 14 (1)] for historical, statistical or research purposes if the responsible party has established appropriate safeguards against the records being used for any other purposes.
Condition 4: Further processing limitation Section 15: Further processing to be compatible with purpose of collection	Section 15 (3): The further processing of personal information is not incompatible with the purpose of collection if— (d) the further processing of the information is necessary to prevent or mitigate a serious and imminent threat to— (i) public health or public safety; or (ii) the life or health of the data subject or another individual; or (e) the information is used for historical, statistical or research purposes and the responsible party ensures that the further processing is carried out solely for such purposes and will not be published in an identifiable form.
Condition 6: Openness Section 18: Notification to data subject when collecting personal information: 'if personal information is collected, the responsible party must take reasonably practicable steps to ensure that the data subject is aware of - [*list of information the data subjects need to know about follows*]	Section 18 (4): It is not necessary for a responsible party to comply with [s18(1)] if— *(f)* the information will— (ii) be used for historical, statistical or **research purposes**.
Processing of special personal information Section 26: General prohibition on processing of special personal information:	Section 27 (1): The prohibition on processing personal information, as referred to in Section 26, does not apply if the— *(a)* processing is for historical, statistical or research purposes to the extent that— (i) the purpose serves a public interest and the processing is necessary for the purpose concerned; or (ii) it appears to be impossible or would involve a disproportionate effort to ask for consent, and sufficient guarantees are provided for to ensure that the processing does not adversely affect the individual privacy of the data subject to a disproportionate extent;
Section 32: Authorisation concerning the data subject's health or sex life (Part B)	Section 32 (5): Personal information concerning inherited characteristics may not be processed in respect of a data subject from whom the information concerned has been obtained, unless— *(b)* the processing is necessary for historical, statistical or research activity.
General prohibition on processing personal information of children (Part C) Section 34: A responsible party may, subject to Section 35, not process personal information concerning a child	Section 35 (1): The prohibition on processing personal information of children, as referred to in Section 34, does not apply if the processing is— (d) for historical, statistical or research purposes to the extent that— (i) the purpose serves a public interest and the processing is necessary for the purpose concerned; or (ii) it appears to be impossible or would involve a disproportionate effort to ask for consent, and sufficient guarantees are provided for to ensure that the processing does not adversely affect the individual privacy of the child to a disproportionate extent;

Table 2: Forms of consent outlined in the Department of Health guidelines (2015)[5]

Form of Consent	Description
Narrow (restrictive) consent	The data subject/research participant provides consent to the use of their specimen or personal information for a single defined use only. The sharing of the data or specimen that is donated by the research participant is not allowed under this form of consent. This form of consent necessitates new consent if further use is deemed as being desirable by the Principal Investigator/responsible party.
Tiered consent	The data subject/research participant provides consent to the use of their data/specimen for the primary study and chooses whether to permit storage for future use, and specimen or data sharing.
Broad consent	The data subject/research participant provides consent to the use of their data/specimen for current research, for storage and for possible future research purposes where the precise nature of future research may not be specifically defined as yet. For broad consent, the nature of the further usage should be described as fully as possible and it should be stipulated that further prior ethics review of the new study will be necessary. Permission may be sought to re-contact the research participant if intended future use is outside the scope of the current consent.[6]
'Blanket' or 'unrestricted' consent	The data subject/research participant consents to their data/specimen being used for any research activity, without any limitations. The guideline is clear that this kind of consent is 'not recommended' as it is difficult to ensure that ethical principles are upheld.[5]

as specific as possible in describing how the personal information from a data subject/research participant would be used in future.

POPIA envisages circumstances where the re-use of data by the same (or a different) responsible party would occur. This is outlined in condition four of the lawful processing of personal information on 'further processing limitations', provided under Section 15 of the Act. Further processing of personal information – which in the research community is understood as the re-use of data – is allowed for research where such processing is solely for research purposes and where the information will not be published in an identifiable form (Section 15 (3) (e)). Further processing is also permissible where the data subject has consented to such further processing, or where the information is already in the public domain. In addition, and as noted in Table 1, POPIA states that further processing for research purposes is permitted if: processing is necessary to 'prevent or mitigate a serious and imminent threat to' public health or public safety or where the processing is necessary to prevent or mitigate a serious threat to 'the life or health of the data subject or another individual'.[1]

In short, further processing is allowed where it is: for research purposes and where the information is not published in an identifiable form (note here the discussion on de-identification below); or where there is consent from the data subject to do so. Therefore, 'broad consent' – as defined under the DoH 2015 Guidelines – is permissible under POPIA if these conditions are fulfilled.

All three of the consent options endorsed for use in health research in South Africa are actively used in the country and would be permissible under POPIA where data subject rights are protected and the responsible parties are as specific as possible in detailing the future use of personal information at the time of consent, including any possible sharing with another responsible party. The important question for implementation of POPIA in research is therefore not *whether* different consent models ought to be used, but *how* they can be used in ways that minimise the risk of harm to the data subject as a result of a loss of privacy. The common conditions that are currently in place in research projects that seek consent for future use, generate a governance framework which seeks to ensure that the re-use of samples and data are ethical. This 'governance framework' includes all the arrangements that determine the re-use of data and samples for future use, and includes: the strength of ethics regulatory oversight; the presence and effective functioning of data use and oversight, for instance, data access committees; community engagement; and mechanisms of data protection which are to prevent unauthorised access to data.

To ensure compliance with POPIA in relation to consent, responsible parties should assess, through an initial risk assessment conducted prior to ethics approval and documented in the data management plan, the balance between the risk of harm resulting from a loss of privacy to the data subject, the strength of the governance framework that regulates the re-use of data, the potential utility of data for future use, and the model of consent adopted. Where the risk to privacy is higher, greater safeguards should be put in place to mitigate against potential harms. Where a research project is determined 'high risk', a full privacy impact assessment should be conducted to determine where further safeguards may be necessary to protect personal information and mitigate against any potential harm to the data subject. See section below on 'High-risk information and risk assessments' in relation to high-risk research.

Data sharing and re-use under POPIA

In addition to the issue of consent discussed above, data sharing and data re-use invokes two notions under POPIA:

1. the use of personal information not collected directly from the data subject, where personal information is shared outside of the original responsible party; and
2. where the information is shared with a body outside South Africa, the sharing of data by a responsible party to a foreign third party.

Under Section 12, POPIA provides that personal information should be collected directly from the data subject. However, POPIA also allows for circumstances where data are not collected directly from the data subject if the data subject has consented (Section 12 (2) (b)), or if the personal information is already in the public domain (Section 12 (2) (a)), or where it is 'not reasonably practicable' (Section 12 (2) (f)). In addition, Section 18 (4) (f) (ii) provides that notification to the data subject when processing their personal information is not necessary when the information is being processed for research. In order to invoke Section 12 (2) (f), the burden of proof would be on the responsible party to show why it was not reasonably practicable to obtain the data directly from the data subject, and this would need to be documented in a risk assessment, outlined above, recorded in the data management plan. This may include assessing what resources were or were not available to the researchers to obtain the data directly from the data subject and the number of data subjects involved.

In addition, in accordance with the principle of data minimality set out in POPIA, data sharing should be encouraged between trusted responsible parties using the data for similar research-based purposes, over the collection of a new batch of personal information from a new set of data subjects.

Consent for processing personal information of a child

Where consent of a child is required for processing personal information, Section 11 of the Act provides that a competent person must provide consent on their behalf. The person consenting must be legally competent to consent to the action or decision of the child. The Act does not, however, distinguish between children of different ages and therefore between different levels of competency and autonomy with respect to the rights of the child.

POPIA prohibits the processing of personal information concerning a child. Exceptions include, amongst other things, where processing has been carried out with the prior consent of a competent person (or where deliberately disclosed by the child with the consent of a competent person). Additionally, processing is permitted for research purposes that serve a public interest and the processing is necessary, or where it would be impossible, or require a disproportionate effort, to obtain consent. Here guarantees must be put in place to show that processing does not adversely affect the privacy of the child.

The DoH Guidelines[5] provide detailed insight into how child consent should be construed which go beyond what is provided under POPIA. Importantly, it is for the child, when of an age to consciously do so, to make the decision to consent and the parent (or competent person) to provide permission. Paramount is that the best interests of the child be considered and upheld. Some additional considerations include whether consent should be re-obtained when the child reaches 18 years, reflecting the child's evolving maturity and capacity to give consent. In addition, there are certain matters where, for reasons of sensitivity, it may be desirable and ethically justifiable for minors to consent independently of a competent person. This is particularly important for research where children may not be willing to participate if their parents must know about the nature of the research in order for permission to be obtained. Finally, appropriate risk standards similar to those used in the DoH Guidelines should be developed for POPIA.[5]

Information and samples

Information as a standalone term is not defined within POPIA. However, personal information under the Act means information relating to an identifiable, living, natural person, and where it is applicable, an identifiable, existing juristic person.

For research purposes under POPIA, this implies that a human biological sample by itself – that is not inherently identifiable and which is collected during the research process – does not fall under POPIA's definition of personal information. [Note, however, that the European Data Protection Board has recently prescribed that genetic data be treated as personal data under the GDPR.[12(para.51)]] Human biological samples would therefore fall under the scope of the *National Health Act 2003*, its Regulations and DoH Guidelines. When information is derived from the sample that is identifiable and relates to a living natural person, that specific information would then be considered personal information and fall under the remit of POPIA. However, the fact that certain biological samples (for example, a fingerprint) are innately identifiable can cause confusion around the exact point at which these samples become personal information due to their potential identifiability. In addition, further concerns may arise regarding the point at which these samples and their associated data may become identifiable, through the sharing of anonymised samples across different sectors. The question around whether potentially identifiable samples constitute personal information is debatable. Yet, POPIA is clear that the personal information must relate to an identifiable, living, natural person. Without any national case law providing clarity and the provisions of the Act open to interpretation, uncertainty and ambiguity remain problematic regarding the exact point at which biological samples become personal information as contemplated under the Act. For the purposes of the Code, human biological samples themselves should fall outside the remit of POPIA until identifiable information relating to a natural living person is derived from the sample.

Genetic data

Genetic and biometric data are not separately defined under POPIA. Genetic data are understood as personal information relating to the genetic characteristics (inherited or acquired, e.g. through mutations in cancer cells) of a person that provides unique information about that person. In cases where genetic 'uniqueness' can be considered biometric data, these data would require a lawful basis for processing, subject to POPIA regulations.

In the case of genetic information, literature has demonstrated the ability to identify an individual from a data subset that relied on linkage to other identifiers such as matching the genetic data against a reference sample, connecting genetic data to non-genetic databases, or generating a profile from genetic data (e.g. ethnicity, eye colour, skin colour) and cross-referencing this with another data set. It should be noted, however, that the risk of identifying an individual through genetic data is highly dependent on the availability of additional identifiers. Several additional challenges therefore arise for the use of genetic data in the context of POPIA and the processing of personal information.

Technologies that generate genetic information are rapidly advancing and the associated costs for generating such data are decreasing, making research which generates and processes genetic data more accessible and affordable. There is substantial and translational benefit to genetic/genomic approaches in research and health, heralding new understanding of disease epidemiology, diagnostics and therapeutics. Going forward, it is essential to ensure that no one is left behind in the genomic revolution and that all can benefit from research that could lead to beneficial innovations such as personalised medicine. For this vision to be actualised, it is imperative that the genomic data that are publicly available are also representative of all people.[13] This means, at least in part, that South African data should continue to be made available both nationally and internationally for analysis and re-use for the advancement of science. This is in line with established standards in open science and genomics research that include many journals and funders requiring research data sets to be made available and researchers sharing their data in the spirit of open research and collaboration.[14]

As indicated above, there is tension in the interpretation of genetic data, namely that whilst even limited genetic information from an individual can be highly identifying, this does not necessarily mean that an individual is identifiable through their genetic data. To identify a person on the basis of genetic information requires linking other information that identifies the person, to their genetic data, as is the case with biometric information, such as fingerprint data: whilst a fingerprint is unique to each person, a fingerprint alone is unlikely to identify the person amongst all other people; some other record needs to exist that links the fingerprint to a person's name before the fingerprint can be used as a source of identification of the person.

Risks to data privacy related to personal information lie in the potential for re-identification and in potential discrimination through the use of genetic data (e.g. racial profiling). For these reasons, genetic data need to be subjected to a higher level of privacy protection when compared to traditional health information. Under Section 32 (5) of POPIA, processing of health information concerning inherited characteristics is permitted if a serious medical interest prevails, or the processing is necessary for historical, statistical or research activity.

The purpose of the processing of genetic data and its future use are important in the context of assessment by ethics committees. De-identification or pseudo-anonymisation of genetic data, as well as appropriate consent approaches, need to be clarified and would require more consideration. It is important that community engagement and individual engagement processes precede informed consent to explain the risks and how they would be mitigated. It may also be useful to consider the addition of a right not to have one's personal data de-identified, as once de-identified, the individual to which the personal information originally related has no rights over that information, such as a right to access or delete it. In which case, it may be prudent to include de-identification of personal information as an express item a data subject must provide consent for.

Additional safeguards which would be relevant to genomic research include provision of detailed information related to data access and use, by the researchers, and the informed consent process for the participant should specifically refer to any potential data sharing (nationally and internationally). A risk assessment should be conducted by the responsible party to determine the likelihood that an individual could be re-identified, and such assessments should be included in ethics review processes. Confidentiality certificates with consent to limit access could also be considered.

High-risk information and risk assessments

POPIA requires that Codes of Conduct provide specific provisions for the processing of personal information considered 'high risk' within the context and scope of the Code.[15] In this context, we take 'risk' here to mean a risk to the rights of the data subject, including but not limited to the right to privacy, as a result of that person's personal information not being adequately protected. Such risks – which often overlap in reality – include:

- Individual identification:
    - risk of loss of privacy; and
    - risk of unconsented identification,
- Stigmatisation: risk of individual stigma (group/community belonging);
- Discrimination and bias;
- Trauma: risk to mental well-being and health (particularly acute for children, vulnerable and marginalised people); and
- Legal prosecution.

Examples of personal information that could be considered high risk include: health data, particularly HIV status; hereditary diseases or other information that could lead to individual stigma; children's information, and information of other vulnerable and marginalised individuals; and behavioural information in relation to a crime, or behaviour deemed deviant or non-normative.

At the outset of a research study, the responsible party/parties must conduct a risk assessment, as noted above. In addition to assessing how high risk the types of personal information being processed are, the risk assessment should also take into account: whether personal information will be transferred outside of South Africa and the extent of the data protection regulations in the country where the personal data will be received; whether unique identifiers will be processed as part of an information matching programme (see below); and whether any operators will be contracted to perform any processing on behalf of the responsible party and what risks such operators may pose (this could include assessing whether the operator has a POPIA compliance policy, or has recently had any data breaches). The risk assessment should be documented under the data management plan, together with the lawful basis for the processing of personal information, and details of the accountable party in terms of POPIA (particularly in the case of a research consortium where there may be more than one responsible party). Where a study is deemed high risk, a full privacy impact assessment should be carried out and vetted by the Information Officer of the research institution(s), as per Section 4 (1) (b) of the POPIA Regulations (No. R. 1383, 14 December 2018).

The processing of high-risk information, as outlined above, requires further safeguards to be in place to balance the potential harms caused by disclosure or breach of confidentiality with the benefits to the improvement of knowledge through research. Additional safeguards provided under POPIA include: data minimisation (ensuring that only the personal information that is essential for testing the research hypothesis or answering the research question is collected), anonymisation of data and data security.

Table 3 sets out the types of information listed under POPIA and their potential risks.

Information matching programmes

POPIA requires Codes of Conduct to develop provisions for how personal information rights will be protected where information matching programmes are in use. POPIA defines information matching programmes (IMP) as:

> *''information matching programme'' means the comparison, whether manually or by means of any electronic or other device, of any document that contains personal information about ten or more data subjects with one or more documents that contain personal information of ten or more data subjects, for the purpose of producing or verifying information that may be used for the purpose of taking any action in regard to an identifiable data subject.*[1]

Information matching, for example through two or more spreadsheets, using code/macros to link sources via an identifier, can be achieved in several ways: (1) non-algorithmic means such as the comparison or combination of data across multiple data sources, or (2) algorithms. When using algorithmic means, information matching can be generally, but not exclusively, performed via machine learning and artificial intelligence (AI).

There are numerous ways in which data sources can be linked, of which a non-exhaustive list of examples is shown below across different data dimensions:

1. Individual identifiers: identity numbers, tax reference numbers, phone numbers.
2. Geographic identifiers: country, town, village, metro.
3. Activity identifiers: employment, hobbies, social media, church, political party membership.

Although many of these data points/sources are in the public domain, triangulation across data sources and data dimensions can allow identification of the data subject. Oftentimes, the matched data are informative for research and also cost effective from a research perspective. It is important to note that information matching for cost-effective purposes may be an important enabler for the research community to minimise the amount of personal data that is collected, sometimes from over-researched communities, in order to comply with the principle of data minimisation.

The challenge in using such matched data in research is ensuring that the data subject's rights are upheld. Research ethics committees play an important role here in ensuring the rights of data subjects are protected during such research activities. However, it would also be required that data subjects provided consent, at the time of collection, to their data being potentially matched with a data set of another responsible party, if not matching data with a data set that is already in the public domain. Where this activity is for research purposes, this is permissible in terms of POPIA.

In other jurisdictions, data protection oversight and regulatory bodies have considered how to protect data subject rights in relation to the use of IMPs, and particularly AI and machine learning based data systems.[16] Some notable points include that the responsible party must implement measures to prevent arbitrary discrimination of an individual. The AI model must therefore be trained with appropriate data, and, where possible, should not prioritise high-risk information, such as related to racial/ethnic origin or political opinion, which may lead to discrimination. Research ethics committees should – in cases where these data are required to answer a specific research question in order to not erode the quality of the IMP – evaluate the data used for this purpose in the context of a risk-based-consent model, as above.

It might be important to set requirements for responsible parties to outline how data are being selected, as well as to provide an outline of how the algorithm was or would be developed and tested. In this case, if a previously developed IMP will be used in the research, this information should be conveyed to the data subject during the informed consent process.

Table 3: Information risk typology

				Potential risk	
				Identification	Harm, exploitation, or stigmatisation
Personal information governed under POPIA	Standard personal information	Any identifying number, symbol, email address, physical address, telephone number, location information, online identifier or other particular assignment to the living person	Potential for identification	🟩	
		Name of an individual (surname and forename individually or together)	Potential for identification and direct privacy invasion	🟩	
		Information on educational, financial, or employment history of an individual	Potential for identification and exploitation	🟩	
		Private correspondence (where this does not contain information listed in the categories below of special personal information relating to the data subject or any other individual. Where the private correspondence contains special personal information it must be handled in terms of the provisions relating to special personal information. Where the private correspondence is of a person who is no longer alive or refers to an individual who is no longer alive, the provisions of POPIA and this Code do not apply)	Potential for identification and harm	🟩	🟩
		Information relating to the gender, sex, pregnancy, marital status, national or social origin, sexual orientation, age, well-being, disability, culture, language and birth of the person	Potential for discrimination		🟩
		Information including personal opinions and views and preferences	Potential for harm or exploitation		🟩
	Special personal information	Biometric information	Potential for re-identification if linked to identifying information	🟩	🟩
		Race or ethnic origin	Potential for discrimination		🟩
		Trade union membership	Potential for harm		🟩
		Political persuasion	Potential for harm		🟩
		Religious or philosophical beliefs	Potential for discrimination or other harm		🟩
		Information relating to the health status of an individual, including information relating to their medical history, disability, physical or mental health	Potential for discrimination, stigmatisation or other harm, particularly in relation to HIV status		🟩
		Information relating to the sex life of an individual	Potential for discrimination, stigmatisation or other harm		🟩
		Information relating to criminal behaviour, in terms of an alleged crime or criminal proceedings, or criminal history	Potential for identification (arrest), discrimination, stigmatisation or other harm	🟩	🟩

Table 4: Examples of data security measures

	Data security measure examples
1.	Policies and procedures for authorised access to personal information, including physical access, computational infrastructure access and network access.
2.	Physical security safeguards, such as locks, barriers and anti-theft systems.
3.	Use of hardware and/or software to protect personal information.
4.	Policies to ensure employee training and review of information access privileges.
5.	Automatic updates of anti-virus or anti-malware software on all person information storage devices.
6.	Encryption of storage and transmission mechanisms (including email) and secure applications for decryption.
7.	The level of security measures should increase when risk is higher.
8.	Policies for access to personal information when working off-site, particularly on less secure networks, logs to trace system activity of a specific user accessing personal information, and to prevent storage of personal information on mobile computing devices.
9.	Policies and procedures to ensure correct disposal of paper and/or electronic personal information, redundancy and backups, as well as disaster recovery safeguards.
10.	Technical safeguards such as firewalls, virus scanners, monitoring operating system logs, version control and encryption methods.

However, if the purpose of the data collection from the data subject is to develop a new IMP, this information will not yet be available, and a general IMP development scenario should be explained to the data subject.

It would also be key to ensure that the purpose for processing personal information in AI systems is clearly established, and indicated, when the data are collected. The purpose of the processing must be fully explained to the data subject such that they can make an informed choice regarding whether to provide consent, given that the responsible party will know the overall processing purpose in the research context, but perhaps not yet the underlying sub-processing purposes that may be revealed during the research study.

In the context of AI, it may be difficult to explain how information is connected within a specific process embedded in a 'black box' or algorithm. A similar challenge is encountered in genomic research where complex concepts have to be conveyed in lay terms to data subjects. Transparent processing requires that processing information be clarified with the data subject during the informed consent process.

The subset of conditions for lawful processing of personal information outlined in POPIA and addressed above, requires the consideration of overarching principles for matching data. These include: ensuring the confidentiality and integrity of personal information through security measures and safeguards, minimising the risk of re-identification of de-identified data; data minimisation; transparency in the processing of personal information; documenting and conveying the purpose specification; and notifying data subjects such that they know where, and by whom, their data are held, and can access and claim their data rights.

Software design should also consider privacy by design principles.[17] Privacy protection can be built into systems as far as possible and ensure data protection is safeguarded in the default settings.

In addition, risk assessment and management plans could be included in the review process for research approval by the research ethics committees where IMPs are being utilised. Risk assessments should evaluate the reasonable likelihood of data subject identification or re-identification with respect to objective factors such as skill required, technology available, and time/cost required. The risks associated with using an algorithm and the impact on the data subject should be recognised and articulated to the data subject during the consent process. In cases where impact assessments have identified categories of data with higher levels of risk, more stringent safeguards must be put in place, where there are the resources to do so.

Security safeguards

Under POPIA, Condition 7 of the lawful processing of personal information requires responsible parties to ensure that personal information collected by the responsible party is kept secure at all times – through appropriate, reasonable, organisational and technical measures – to protect against security breaches.

In order to determine which measures are the most appropriate and reasonable, an organisational risk analysis and privacy impact assessments must first be performed, prior to evaluation of processes to manage and mitigate the risks of a data breach. There are several accepted frameworks for IT security practices and procedures, with the ISO27000 series being the most widely accepted Information Security management standard.[18-20] The US National Institute of Standards and Technology cyber security framework has also been recognised as an important standard for organisations.

Information technology security strategies prevent unauthorised access to data assets of an organisation, to maintain the integrity and confidentiality of sensitive information. It is important to note that these strategies do not solely rely on the hardware and software mechanisms, but include additional security measures such as appropriate policies, procedures, and physical controls. At an organisational level, specific IT policies may be in place and would form part of the 'appropriate, reasonable, organisational and technical measures' stipulated by POPIA.

The use of security measures to protect personal information differs from one research body to another, depending on both organisational requirements and available resources.[21] Examples of data security measures are included in Table 4.

Social media data

POPIA provides that when information is in the public domain it can be used and processed without consent. This would include publicly available personal information on social media platforms, as well as blogs and websites that are open to the public. A useful report in this regard is 'Ethical Guidelines on Social Media' published by the Health Professions Council of South Africa.[22] This is particularly pertinent in South Africa where there is no specific legislation regulating social media.[23]

As consent is not necessary for the processing of personal information from public sources, the lawful basis under POPIA for such processing will not be consent. Unless the researchers were conducting research mandated by a public law, the lawful basis that would be relied on to process publicly available personal information would be the legitimate interest of the responsible party.

A data subject's expectation of privacy when using social media platforms is inversely proportional to the rigour of privacy settings associated with the platform upon which information is shared. Hence, when sharing information on an account that has no privacy settings, and is thus publicly available, the data subject has – in effect – forfeited their right to privacy. When sharing information on an account that has activated certain privacy settings, but where that information can be viewed by millions of people (for instance a platform hosted by a public figure or institution), a data subject has a lower expectation of privacy than when sharing information to a platform that can be seen only by a select few. When information is shared with only one other individual, such as on a *WhatsApp* messenger, the data subject has a high expectation of privacy. The level of POPIA-related safeguards for research applications in social media must be commensurate with the expectation of privacy implied by the data subject when they posted their personal information.

In this case there are three instances of how the data subjects' information could be used for research purposes: (1) the information can be de-identified and thus falls outside the scope of POPIA; (2) where the information is not de-identified; and (3) where the data cannot be de-identified. While the expectation of privacy is diminished, cases involving minors and other vulnerable groups[5] need to be considered in order to ensure their rights are protected and that they are not subject to any harm as a result of the publicly available information.

In addition, the Information Regulator recently released a comment in relation to changes to the terms and conditions of *WhatsApp*, a Facebook company. The Information Regulator stated that in terms of Section 57 of POPIA, the social media platform may not

> *process any contact information of its users for a purpose other than the one for which the number was specifically intended at collection, with the aim of linking that information jointly with information processed by other Facebook companies*

unless it obtains prior authorisation from the Information Regulator to do so. Hence, the intent of the data subject is a significant factor in how the data of the data subject may be dealt with, although in cases involving international platform providers it poses cross-jurisdictional challenges.[24]

Overall, social media data should be de-identified as early as possible in the research process and the principle is to only collect information that is directly relevant to the research. The de-identification process must also include a de-coupling process in which, for example, location and other data that are not relevant to the research hypothesis are disconnected from the data relevant to the research and (1) are not collected by the researcher, or (b) are disconnected from the post prior to processing it for research purposes. It is recommended that the specific requirements for this process should be processed and approved by the relevant research ethics committee.

Cross-border data sharing/information flows

Section 72 of POPIA sets out the conditions for transferring personal information to a foreign jurisdiction. In principle, responsible parties must ensure that the foreign country with which personal information is being shared or transferred to has as high a level of data protection as offered under POPIA. Responsible parties must also ensure that a transfer agreement is in place, which offers the necessary safeguards and protections for transferring personal information. Transfer agreements must be in binding contractual form. This broadly echoes the requirements of the GDPR in relation to cross-border data sharing. In July 2020, the Court of Justice of the European Union decision held that the EU-US Privacy Shield is no longer a valid basis for transferring EU personal information to the USA.[25] The decision, known as *Schrems II*, found that the European Commission's adequacy determination for the EU-US Privacy Shield Framework is invalid due to concerns regarding the necessity and proportionality of the surveillance activities of the US government and the availability of actionable judicial redress for EU data subjects. Second, the decision affirmed the validity of standard contractual clauses, while stating that data exporters and importers must verify, on a case-by-case basis, whether the law in the recipient country ensures adequate protection, similar to what is offered under the GDPR.[25,26] Where no adequate protection is in place (such as where the foreign country to which data are being transferred does not have a functional data protection regulatory system in place), additional safeguards must be provided by the data exporters and importers to guarantee such protection, and built into the transfer agreement. In effect, *Schrems II* made this incumbent on 'data controllers' (what would be 'responsible parties' under POPIA) to ensure not only that other countries with which personal information is shared have a similar level of data protection regulation, but also that safeguards are in place to protect the personal information and rights of the data subject.

Parties to a transfer can offer enhanced legal guarantees that build on those in standard contractual clauses but provide stricter conditions for suspending data flows and deleting data in cases of unauthorised government access. Second, technical measures such as strong encryption methods, as well as organisational measures such as commitments to suspend data transfers to countries that do not respect the rule of law, based on internationally recognised standards, could be adopted. It is the responsibility of the responsible party to ensure that sufficient legal protections are in place when transferring personal information outside of South Africa, and it is encouraged that responsible parties sharing information outside of South Africa take note of the obligations around cross-border data transfer set out by the GDPR and the *Schrems II* decision.

Section 72 of POPIA provides that data can also be shared with a foreign country where the data subject consents to the transfer, or where the transfer would be to the benefit of the data subject, or is necessary for the performance of a contract between the responsible party and the data subject. However, Section 57 (1) (d), indicates that if special personal information, or the personal information of a minor, is to be transferred to a country that does not provide an adequate level of protection under Section 72, then prior authorisation of the Information Regulator is required. Section 57 (3) states that this prior authorisation will not be needed if a Code has come into force for a specific sector. Thus, the Code for Research will need to include provisions to guide researchers in transferring or sharing personal information outside of South Africa, and will need to take into account the developing international best practice in this regard, in order to ensure that South African researchers remain internationally competitive.

Governance of the Code of Conduct

It is recommended by the Information Regulator that the body which develops the Code for Conduct takes responsibility for governing the Code, on the basis that the body has been deemed representative enough of the sector to which the Code will apply.[27] There are two key duties in respect to the governance of the Code, which will fall on ASSAf. First, to report on the sector's compliance with the Code to the Information Regulator on an annual basis. This will include receiving annual statistics from all bodies that fall under the Code with respect to the number and nature of complaints received in relation to the Code.[27]

The second will be for ASSAf to handle complaints in relation to the Code. Given that the scope of research activities that fall under this Code would ordinarily have undergone prior ethics approval, ASSAf would not be the first port of call for handling complaints. Instead, what is being considered is a tiered process whereby a complainant would first approach the relevant research ethics committee which had authorised the research. If the complainant is, at this stage, aggrieved by the outcome decided by the research ethics committee, the complainant would then approach the National Health Research Ethics Council if the research ethics committee in question is registered with the National Health Research Ethics Council, as provided for under the *National Health Act*. If a complaint in relation to the Code was in relation to a research project that had not undergone ethics clearance, then the complaint could be brought directly to ASSAf. However, as ethics review will constitute a key safeguard for ensuring compliance with POPIA and the Code, the complaint against the responsible parties would not be reviewed favourably. At this point, if the matter was still not resolved and related squarely to a violation of this Code, it would be handled by an independent committee established by ASSAf, and in accordance with the guidance on POPIA complaints handling as published by the Information Regulator.[28] The last port of call for complaints, following handling by ASSAf's independent committee, would thereafter be the Information Regulator.

Conclusion

This document has sought to outline the main areas relating to the processing of personal information for research purposes which the Code will address, including: what consent models would be permissible under POPIA; the issues in relation to genetic research and the processing of personal information contained in inherited characteristics; the use of IMPs by researchers; and the use of personal information obtained from social media platforms for research. This is not an exhaustive list of the concerns faced by the research community in respect of the changes that POPIA will bring about. Other issues which the Code will address relate to intellectual property law, including but not limited to patents, as well as the commercialisation of research data. However, with ongoing and wide consultation with the scientific community in South Africa and all relevant stakeholders, it is hoped that the Code will provide guidance in supporting the lawful and responsible use of personal information while conducting scientific research in South Africa.

Acknowledgements

We thank all those who have contributed to this Discussion Document from various sectors and disciplines. We are particularly grateful to three anonymous readers for their input.

Competing interests

There are no competing interests to declare.

Authors' information

The authors, excluding Mark Gaffley who is working as a research assistant on this project, are all members of either the ASSAf-appointed POPIA Code of Conduct Steering Committee or Drafting Committee.

Name	Affiliation	Area of expertise
Steering Committee members		
Dr Rachel Adams	HSRC	Human rights law
Prof. Ahmed Bawa	USAf	Physics
Prof. Alan Christoffels	UWC	Bioinformatics and health genomics
Prof. Jantina de Vries	UCT	Bioethics
Prof. Monique Marks	DUT	Social sciences
Dr Mongezi Mdhluli	MRC	MRC representative
Dr Mapitso Molefe	CSIR	CSIR representative
Dr Tshilidzi Muthivhi	DoH	Department of Health representative
Prof. Caroline Ncube	UCT	Commercial law
Prof. Michèle Ramsay (Chair)	WITS	Human genetics
Prof. Jerome Singh	SUN	Ethics and law
Drafting Committee members		
Dr Rachel Adams (Chair)	HSRC	Human rights law
Dr Fola Adeleke	WITS	Law
Dr Dominique Anderson	UWC	Bioinformatics
Dr Nicola Branson	UCT	Social sciences
Dr Harriet Etheredge	WITS	Bioethics
Ms Eleni Flack-Davison	WITS	Legal advisor
Prof. Safia Mohammed	UNISA	Law
Dr Antonel Olckers	DNABiotec	Human genetics
Prof. Maria Papathanasopoulos	WITS	HIV pathogenesis research
Ms Jane Pillay	NHLS	Immunology
Prof. Tobias Schonwetter	UCT	Intellectual property
Dr Carmen Swanepoel	SUN	Medical biochemistry

References

1. Protection of Personal Information Act 4 of 2013, Republic of South Africa.
2. Adams R, Veldsman S, Ramsay M, Soodyall H. Drafting a Code of Conduct for Research under the *Protection of Personal Information Act No. 4 of 2013*. S Afr J Sci. 2021;117(5/6), Art. #10935. https://doi.org/10.17159/sajs.2021/10935
3. Human Sciences Research Council Act 17 of 2008, Republic of South Africa.
4. The Constitution of the Republic of South Africa Act 108 of 1996, Republic of South Africa.
5. National Health Act 61 of 2003, Republic of South Africa. Available from: https://www.gov.za/sites/default/files/gcis_document/201409/a61-03.pdf
6. South African Department of Health (DoH). Ethics in health research: Principles, processes and structures. 2nd ed. Pretoria: DoH; 2015. Available from: https://www.sun.ac.za/english/research-innovation/Research-Development/Documents/Integrity%20and%20Ethics/DoH%202015%20Ethics%20in%20Health%20Research%20-%20Principles,%20Processes%20and%20Structures%202nd%20Ed.pdf
7. South African Medical Research Council Act 58 of 1991, Republic of South Africa.
8. Agricultural Research Act 86 of 1990, Republic of South Africa.
9. Organisation for Economic Co-operation and Development (OCED). Frascati Manual 2015: Guidelines for collecting and reporting data on research and experimental development, the measurement of scientific, technological and innovation activities. Paris: OECD; 2015.
10. South African Law Reform Commission (SALRC). Project 124: Privacy and data protection report. Pretoria: SALRC; 2009. para 4.4.3.
11. General Data Protection Regulation 2016/679, European Union.
12. European Data Protection Board (EDPB). EDPB Document on response to the request from the European Commission for clarifications on the consistent application of the GDPR, focusing on health research [document on the Internet]. Available from: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_replyec_questionnaireresearch_final.pdf
13. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538:161–164. https://doi.org/10.1038/538161a
14. Powell K. The broken promise that undermines human genome research. Nature. 2021;**590**:198–201. https://doi.org/10.1038/d41586-021-00331-5
15. Information Regulator of South Africa – Department of Justice and Constitutional Development (DoJ). Re: Notice relating to consultations on guidelines to develop codes of conduct in terms of chapter 7 of the Protection of Personal Information Act of 2013 on the 6th November 2019. Johannesburg: DoJ; 2019. para 7.11.

16. The Norwegian Data Protection Authority (Datatilsynet). Artificial intelligence and privacy: Report, January 2018. Oslo: Datatilsynet; 2018.

17. The Norwegian Data Protection Authority (Datatilsynet). Software development with data protection by design and by default. Oslo: Datatilsynet; 2017. Available from: https://www.datatilsynet.no/en/about-privacy/virksomhetenes-plikter/innebygd-personvern/data-protection-by-design-and-by-default/?print=true

18. National Institute of Standards and Technology, U.S. Department of Commerce (NIST). Cybersecurity framework. Gaithersburg, MD: NIST; 2016. Available from: https://www.nist.gov/industry-impacts/cybersecurity-framework

19. Crook G. The implications of IT governance outlined in King IV™. Durban: BDO; 2016. Available from: https://www.bdo.co.za/en-za/insights/2016/report/the-implications-of-it-governance-outlined-in-king-iv

20. Lewinson M. PRINCE2 methodology overview: History, definition & meaning, benefits, certification [webpage on the Internet]. *c*2011 [cited 2021 Apr 26]. Available from: https://mymanagementguide.com/prince2-methodology-overview-history-definition-meaning-benefits-certification/

21. Abiodun OP. Exploring the influence of organizational, environmental, and technological factors on information security policies and compliance at South African higher education institutions: Implications for biomedical research [thesis]. Cape Town: University of the Western Cape; 2020. https://etd.uwc.ac.za/handle/11394/8074

22. Health Professions Council of South Africa (HPCSA). Ethical guidelines for good practice in the health care professions: Ethical guidelines on social media: Booklet 16. Pretoria: HPCSA; 2019. Available from: https://www.hpcsa.co.za/Uploads/Professional_Practice/Conduct%20%26%20Ethics/Ethical2%20Guidelines%20on%20Social%20Media.pdf

23. South African Department of Government Communications and Information Systems (GCIS). Social media policy guidelines: April 2011. Pretoria: GCIS; 2011. Available from: https://www.gcis.gov.za/sites/default/files/docs/resourcecentre/guidelines/social_media_guidelines_final_20_april2011.pdf

24. Information Regulator of South Africa, Department of Justice and Constitutional Development (DoJ). Media statement: Information Regulator SA provides legal analysis on WhatsApp privacy policy. Johannesburg: DoJ; 2021. Available from: https://www.justice.gov.za/inforeg/docs/ms-20210303-Whatsapp.pdf

25. Court of Justice of the European Union (CJEU). The Court of Justice invalidates Decision 2016/1250 on the adequacy of the protection provided by the EU-US Data Protection Shield. Luxembourg: CJEU; 2016. Available from: https://curia.europa.eu/jcms/upload/docs/application/pdf/2020-07/cp200091en.pdf

26. European Data Protection Board (EDPB). Frequently asked questions on the judgment of the Court of Justice of the European Union in Case C-311/18 – Data Protection Commissioner v Facebook Ireland Ltd and Maximillian Schrems [webpage on the Internet]. c2020 [cited 2021 Apr 13]. Available from: https://edpb.europa.eu/our-work-tools/our-documents/ohrajn/frequently-asked-questions-judgment-court-justice-european-union_en

27. Information Regulator of South Africa, Department of Justice and Constitutional Development (DoJ). Guidelines to develop codes of conduct: Issued under the Protection of Personal Information Act 4 of 2013 (POPIA). Johannesburg: DoJ; 2021. Available from: https://www.justice.gov.za/inforeg/docs/InfoRegSA-Guidelines-DevelopCodeOfConduct-22Feb2021.pdf

28. Information Regulator of South Africa, Department of Justice and Constitutional Development (DoJ). Standard for making and dealing with complaints in a code of conduct (prescribed in terms of Section 65 of the Protection of Personal Information Act No 4 of 2013). Johannesburg: DoJ; 2021. Available from: https://www.justice.gov.za/inforeg/docs/InfoRegSA-Standard-CodeOfConduct-Complaints-20210301.pdf

Non-communicable diseases – a catastrophe for South Africa

AUTHORS:
Ebrahim Samodien[1]
Yoonus Abrahams[1,2]
Christo Muller[1,2]
Johan Louw[1,3]
Nireshni Chellan[1,2]

AFFILIATIONS:
[1]Biomedical Research and Innovation Platform, South African Medical Research Council, Cape Town, South Africa
[2]Centre for Cardio-metabolic Research in Africa, Division of Medical Physiology, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa
[3]Department of Biochemistry and Microbiology, University of Zululand, Richards Bay, South Africa

CORRESPONDENCE TO:
Ebrahim Samodien

EMAIL:
Ebrahim.Samodien@mrc.ac.za

DATES:
Received: 15 July 2020
Revised: 06 Nov. 2020
Accepted: 18 Jan. 2021
Published: 28 May 2021

HOW TO CITE:
Samodien E, Abrahams Y, Muller C, Louw J, Chellan N. Non-communicable diseases – a catastrophe for South Africa. S Afr J Sci. 2021;117(5/6), Art. #8638. https://doi.org/10.17159/sajs.2021/8638

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☐ On request from author(s)
☐ Not available
☒ Not applicable

EDITORS:
Pascal Bessong
Sandiswa Mbewana

KEYWORDS:
biology of misfortune, child health, epigenetics, chronic illness, nutrition, social inequality

FUNDING:
South African Medical Research Council

Non-communicable diseases contribute significantly to the disease burden within South Africa. In the most unequal of societies in the world, poverty and socio-economic disparity are amongst the greatest obstacles facing South Africans, impacting heavily on health care. Adverse socio-environmental factors, especially those experienced during early life, can, through neurobiological and epigenetic mechanisms, developmentally programme the outcome of obesity, diabetes, cardiovascular disease and mental health disorders in adulthood. In this narrative review, we describe the social environment experienced by South Africans and discuss the potential contribution of epigenetics to the current and future prevalence of non-communicable diseases. A large part of the population (including 60% of young children) lives in poverty and endures challenging socio-economic environments, due to high unemployment, alcohol and substance abuse, and inter-partner violence. It is imperative that socio-economic factors be considered as risk factors for strategies aimed at reducing or preventing these disorders. If the current situation is left unchecked, the disease incidences could be exacerbated, and be potentially catastrophic for future generations. The consequences can be widespread and can have a direct effect on the future health and economic development of the country. Thus, child and adolescent health requires urgent attention and should be placed at the centre of the healthcare system. Early interventions providing optimum nutrition, a secure environment, together with physical activity and education should be the cornerstones for creating a healthier population for the future.

Significance:

- South Africa already has a high non-communicable diseases burden. Non-communicable diseases – like cardiovascular diseases, cancer, diabetes, respiratory illnesses and mental disorders – are known to be caused by the interaction of socio-environmental factors, physiology, behaviour and genetics. About 60% of South Africa’s children currently live in poverty, with adverse socio-environmental factors known to have a negative effect on development, leading to a plethora of health problems in adulthood.
- The implications for the current situation are widespread: a future population with deteriorated physical and mental health, presenting with co-morbidities that render these individuals more susceptible to infectious diseases. It is important to recognise the potential repercussions on the health prognosis of future generations.
- Endeavours should focus on early interventions that can provide optimum nutrition, education, and physical activity within a safe environment. These interventions can have favourable effects on children’s brain development and genetics, thereby promoting their well-being and increasing their life prospects.

Introduction

Non-communicable diseases (NCDs) are the leading cause of death worldwide, being responsible for 71% of global mortality, with an estimated 41 million people dying from NCDs each year.[1] South Africa is famously known as the home of the ‘Big 5’ animal species, but if the current trajectory continues, it will also be the home of the big 5 NCDs. The major NCDs are cardiovascular diseases (CVDs), cancer, type 2 diabetes mellitus, respiratory illnesses (such as chronic obstructive pulmonary disease) and mental health disorders.[2] NCDs are chronic illnesses that tend to be of prolonged duration and occur as the result of a combination of genetic, physiological, socio-environmental and behavioural factors.[3] It is estimated that NCDs will cost the global economy USD47 trillion over the next two decades, driving millions of individuals into or further into poverty and thereby exacerbating inequalities in quality of life and longevity.[4]

An already poor health prognosis for South Africa has been associated with a quadruple burden of communicable diseases, NCDs, maternal and child health, as well as injury-related disorders.[5,6] The country also experiences a high prevalence of inter-personal violence.[7] Equally important within the South African context is the growing trend of multi-morbidity, with the combination of human immunodeficiency virus (HIV)/NCDs and tuberculosis mycobacterium (TB)/diabetes, increasing the need for resources in the treatment and management of both chronic infectious diseases and NCDs.[8]

Recently, socio-economic status was recognised as a major contributing factor to the development of NCDs, not only in South Africa but worldwide.[9] While the role of socio-determinants of health is not new, especially with regard to NCDs, there is increased interest in understanding the influence of epigenetics in this regard.[10] A better understanding of region-specific risk factors could greatly aid the development of initiatives to reduce disease incidences and lighten the burden thereof. In this narrative review, we discuss the potential contribution of the socio-environment experienced by South Africans towards the current NCDs prevalence. We describe the possible interaction of several socio-environmental factors together with epigenetics, and aim to caution against the perpetual effects thereof, which may result in an even higher disease incidence in future generations. Furthermore, several important considerations which could be useful to mitigate the detrimental epigenetic effects are proposed.

Poverty and socio-economic disparity

South Africa was the most economically unequal society (out of 149 countries) assessed using the Gini index.[11] More than half (55%) of the population experiences poverty[12], with childhood poverty affecting 63% of children.[13] Early-life adversities can have a negative impact on growth and development, with childhood poverty having both short- and long-term consequences.[14] Impoverished children exhibit higher rates of acute and chronic diseases, with worsened physical and mental health in adulthood.[15] Prolonged exposure to early life adversity establishes a developmental 'biology of misfortune', involving neurobiological and epigenetic processes through which one's life course is steered towards diminished health, unrealised potential and reduced longevity.[15] Furthermore, children who do not have access to adequate nutrition (due to malnutrition or over-consumption) are developmentally compromised, exhibit learning disabilities, and are impulsive and prone to erratic and risky behaviour.[16] If poverty and inequality are not adequately addressed, vulnerable children will become adults with a heightened susceptibility to disease. In a country with an already high NCDs burden[2], harsh socio-environmental conditions may contribute to a vicious cycle of unfavourable health prognosis, as is being witnessed in the current generation[12], which could worsen significantly in future generations.

The failure to optimise nutrition, especially during the critical periods of development for vulnerable young children, should be avoided at all costs. There is a definite requirement for efforts directed at improving the national diet. It should be noted, however, that healthier foods are far more expensive than less healthy, nutritionally poor foods.[17] Products like lean meat, fish, fruit and vegetables generally cost more than oil-heavy processed foods which contain more sugar and fat.[18] This makes the promotion of a quality diet difficult, because it is simply unaffordable for many South Africans. International research has shown that the best strategies for changing the dietary environment in favour of healthier foods are those aimed at population level and are accomplished by mass-media nutrition campaigns and transparent food labelling, and, more drastically, through regulation and taxation of unhealthy foods.[18] To this end, the South African government has implemented policies that ensure stricter food labelling, prohibited advertising to children, introduced mandatory salt reduction legislation in 2016 as well as sugar taxation in 2018 (with South Africa being the first African country to do so).[18] While research into the effect of sugar taxation in South Africa is still underway, data from Mexico and Chile have shown taxation to be partially effective, with a reduction in sugary beverage sales.[19,20] However, these policies have a greater impact on poorer households[21], and offer little in reducing socio-economic inequalities in diet-related health.[19,20]

Cultural dynamics and educational influences

South Africa is famously known as the rainbow nation, with a rich ethnic and cultural diversity comprising a variety of population groups including African, European, Indian and others, each embracing varying beliefs and cultural practices. Yet despite this diversity, dietary diversity is ever decreasing. More nutritious traditional foods have largely been replaced by the 'Western diet', which is characterised by the consumption of energy-rich and nutrient-poor processed food, largely from animal origin[22], together with increased consumption of sugar-rich beverages.[18] A concomitant decrease in the consumption of fruit and vegetables, whole grains and fibre has also been observed.[22] This type of food environment has been associated with the rise in chronic illnesses including CVDs, cancer and diabetes.[23]

More awareness about the effects of an unhealthy diet could be useful, especially to younger children. General nutrition scores for the South African population tend to increase with age and peak at 55–64 years of age.[24] Initiatives to improve nutritional knowledge, such as the nutrition education programme, have been implemented, and was shown to improve both the teachers' nutritional knowledge as well as learners' nutrition attitude.[25] However, in the same study, no significant improvements in dietary practice of teachers or learners were found.[25] It is plausible that, even if most of the public are aware of the health risks associated with the so-called 'Western diet', the consumption of heavily processed foods is unavoidable due to economic constraints and will therefore remain widespread. Also, amongst certain ethnic groups in South Africa, the type of food consumed is a measure of economic status.[26] High adiposity is considered a sign of affluence and comfortable living, while being lean may be associated with being sickly or poor (or both), or having contracted TB, HIV or cancer[27], with negative stereotypes and beliefs posing a great impediment to the development and success of healthy eating campaigns.

Indeed, several studies have shown the prevention or partial reversion of NCDs through implementation of lifestyle modification therapies, such as increased levels of physical exercise together with a balanced quality diet.[28,29] Such preventative approaches are heavily under-represented, are certainly not available to most South Africans, and much awareness can be created around them.[28] A school environment represents a controlled system, which could be targeted with efforts aimed at promotion of such interventions. This endeavour could be aided through the implementation of the teaching of crop cultivation within the school curriculum that culminates in a food garden project. The national school nutrition programme has aimed to establish food gardens, and even though the benefits are well known, this pillar of the framework has been described as under-funded and neglected.[30] Such efforts could help feed the children nutrient-rich foods, to support surrounding communities as well as to educate about the health benefits of foods in preventing disease, whilst also being therapeutic. Furthermore, it would enable knowledge transfer to the general public, with regard to health risks associated with unhealthy diets, whilst simultaneously, efforts can be aimed at breaking the negative stereotypes related to obesity/leanness, beginning with young children as the key intervention group.

Socio-environmental factors and epigenetics

Socio-environmental conditions, even before we are born, are extremely important and can fundamentally affect our biological physiology[31], and thereby activate or deactivate specific genes, with experiences of parental hardship in early childhood leading to alterations in chromatin structure, which are detectable a decade and a half later[32]. Epigenetic modifications – which include DNA methylation, micro-RNA circulation, histone modification and chromatin remodelling – have been implicated in the pathophysiology of obesity and several NCDs[33], including diabetes, CVDs, cancers, and neurodegenerative and mental health disorders. Epigenetic markers constitute a biological 'memory' of early life experience, even more so in experiences of misfortune, poverty and stress.[32] Epigenetic alterations can have long-lasting effects, spanning across generations, as observed in the Dutch famine cohort[34] and the seasonal famines of Gambia.[35] The same holds true for children suffering maltreatment, who exhibit long-lasting mental health perturbations and behavioural problems, which persist into adulthood.[36]

It is conceivable that the interaction between epigenetic changes arising from challenging socio-economic conditions are partially responsible for the high prevalence of NCDs in South Africa. The interplay between socio-environmental factors and epigenetics (Figure 1) must be considered and taken seriously when developing strategies to attenuate disease progression. The high genetic diversity together with varying socio-economic factors, although complex, offers a unique milieu of conditions for clinical investigations, and due to the paucity of research, calls for programmes to be initiated for this purpose, not only in South Africa but within Africa too.[37]

Obesity

Obesity has been described as a normal response to an abnormal environment[38]; however, the social and environmental factors contributing to disease aetiology are often underappreciated.[39] While South Africa is the most food secure nation on the African continent, more than half of the population are at risk of hunger.[12,24] Despite high levels of food insecurity and elevated risk of starvation, the country remains one of the top 20 overweight and obese nations in the world.[12]

An estimated 27% of the population is obese, with South African women being amongst those with the highest rates of obesity worldwide, with a prevalence of 42%.[40] This can partially be explained by high levels of physical inactivity amongst women, which is estimated to be 48%.[2]

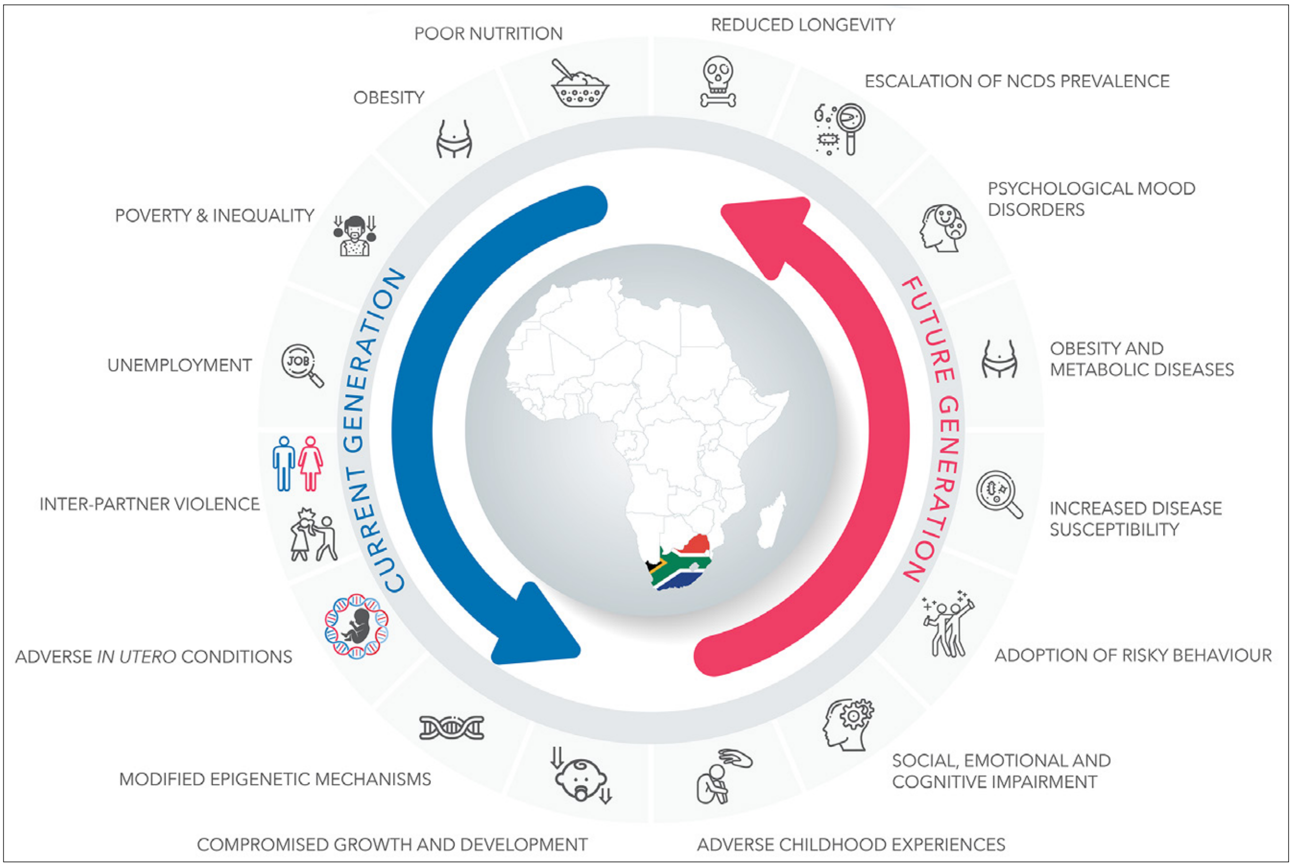

POOR NUTRITION
REDUCED LONGEVITY
OBESITY
ESCALATION OF NCDS PREVALENCE
POVERTY & INEQUALITY
PSYCHOLOGICAL MOOD DISORDERS
CURRENT GENERATION
FUTURE GENERATION
UNEMPLOYMENT
JOB
OBESITY AND METABOLIC DISEASES
INTER-PARTNER VIOLENCE
INCREASED DISEASE SUSCEPTIBILITY
ADVERSE *IN UTERO* CONDITIONS
ADOPTION OF RISKY BEHAVIOUR
MODIFIED EPIGENETIC MECHANISMS
SOCIAL, EMOTIONAL AND COGNITIVE IMPAIRMENT
COMPROMISED GROWTH AND DEVELOPMENT
ADVERSE CHILDHOOD EXPERIENCES

Figure 1: A cycle of non-communicable diseases in South Africa, in which adverse socio-environmental factors affecting a large number of children exponentially increase their susceptibility to diseases and promote the development of chronic illnesses in adulthood – the effects of which can potentially be perpetuated across generations.

Physical inactivity is also prevalent amongst children and youth, with levels deemed insufficient to promote health and prevent NCDs.[41] Obesity in young children in South Africa is also rapidly increasing, which is intriguing as a large proportion of children are at risk of starvation.[12,24] The prevalence of overweight and obesity in 2013 was reported to be 19% and 7%, respectively, for boys, and 26% and 10%, respectively, for girls within urban areas.[40] In 2008, these figures were reported at 11% and 3% for boys, and 29% and 8% for girls within a nationally representative cohort.[42] The country experiences what has been described as the double burden of malnutrition. A large number of children have been found to be developmentally stunted and thereby are at an increased risk of becoming overweight and obese, with a higher likelihood of developing NCDs in adulthood.[43] Not only do obese individuals have an elevated risk of developing metabolic diseases, but offspring born to obese women are increasingly vulnerable to chronic disease development later in life.[44] It has been estimated that more than 90% of type 2 diabetes mellitus, 68% of hypertensive, 45% of ischaemic stroke and 38% of ischaemic heart disease cases in South Africa occur as a result of excess body weight.[45] The sub-Saharan African region is undergoing rapid demographic and epidemiological transition, which is believed to be the driving factor behind the increased risk, prevalence and burden of CVDs, diabetes and neurodegenerative diseases.[46]

It is worthy of mention that obesity is an underlying and re-occurring theme within the development of several major NCDs and thus represents a pivotal preventative and/or therapeutic target. Therefore, strategies aimed at addressing obesity could go a long way in reducing the prevalence of NCDs. In this regard, strategies should target young children, particularly through providing adequate nutrition together with education that facilitates their growth and development. Additionally, specific policies that address physical inactivity amongst children and young adults in both rural and urban settings across diverse socio-economic status are also required.[47] It is notable to mention that the prescription of exercise is a specialist therapy and the importance of having biokineticists in the public sector in the fight to reduce NCDs should be recognised.[48]

Stress, anxiety and depression

The *in utero* environment is increasingly being recognised as important in determining our future health prognosis.[32] An adverse *in utero* environment can contribute to altered epigenetic and gene expression profiles as well as compromised growth and development. Individuals enduring such challenging environments may suffer emotional and cognitive impairment, with an increased likelihood of adopting risky behaviours (Figure 1). Additionally, such circumstances can also contribute to an increased susceptibility for developing metabolic diseases as well as psychological mood disorders.

Childcare often subjects parents to anxiety, stress and depression. Both antenatal and postnatal depression affects an estimated one-third of all mothers. This is indeed worsened by a lower socio-economic status with such conditions affecting 39% and 47% of pregnant mothers in informal settlements and rural areas, respectively.[49,50] Maternal depression during infancy has been associated with dysregulation of the child's biological stress response.[51] Furthermore, alcohol and substance abuse during pregnancy is also rife, with South Africa having one of the highest occurrences of foetal alcohol syndrome in the world.[52] Another alarming matter is the high prevalence of intimate partner violence, with approximately 20% of pregnant women being affected.[53] Half of female homicides in South Africa are perpetrated by intimate partners[54,55], with violence against women being a significant problem that profoundly affects the physical and mental well-being of the individuals involved.[55]

Early-life adversity and struggles are linked to pro-inflammatory shifts in cytokine expression and increased CVDs risk[56], whilst also increasing an individual's vulnerability to developing depressive disorder in adulthood.[57]

In South Africa, we are witnessing an ever-increasing rate of depression, and perhaps due to the stigma associated with mental health disorders, many incidences are unreported.[58] Depression induced by HIV-stressors is also prevalent.[59] It is estimated that one in three South Africans will experience a depressive episode at least once in their lifetime.[60] Interestingly, substance and alcohol abuse are also significant public health problems in South Africa, which are inadvertently linked with increased violence and injury.[61,62] Furthermore, there is a considerable loss of life due to self-harm, with 70% of individuals who had attempted suicide shown to be suffering from a mental health disorder.[63] A high lifetime prevalence of substance abuse is also rife within the country, and with these disorders known to have an early age of onset, it provides an important indication in regard to which demographic to target when planning mental health initiatives and services.[64]

Food for thought

Unhealthy diets high in fat and sugar negatively affect the brain[65] and contribute to 1 in 5 deaths worldwide.[23] Diet-induced hypothalamic inflammation is one of the first symptoms to occur in the development of obesity and metabolic diseases.[65] Increased neuronal Inflammation is also a commonality for several neurodegenerative diseases like Parkinson's and Alzheimer's, as well as in psychological mood disorders such as depression.[66] With diet being intertwined with emotions, cognition and behaviour[67], close attention should be paid to nutrition in order to prevent the induction of metabolic and inflammatory perturbations.

There are dietary regimens, particularly those high in polyphenols, which confer beneficial health effects. Dietary polyphenols are plant compounds found in tea, chocolates, herbs and spices, fruit, vegetables and nuts. Several polyphenols have been shown to be able to attenuate metabolic disease pathologies partially through preventing oxidative stress and inflammation in the brain.[68] These plant compounds are regarded as exercise mimetics and have shown synergistic effects in combination therapies.[69] Interestingly, like exercise, polyphenols hold the potential to positively modulate the epigenetic machinery and thereby restore normal gene expression.[70,71] Furthermore, South Africa sits on a botanical 'goldmine' of indigenous medicinal plants that exhibit anti-obesity, anti-cancer, anti-diabetic as well as anti-ageing properties amongst others[72], and more efforts are required in order to develop such natural therapeutics.

Finally, while the benefits of breastfeeding for both mother and child have long been known,[73] South Africa, like most countries, is still not doing enough to support mothers to breastfeed, despite the immense economic implications.[74] For mothers, breastfeeding decreases stress and promotes positive affect, while improving maternal compassion and care.[75] According to the World Health Organization, all babies should ideally be breastfed exclusively from birth up to 6 months of age. Breastfeeding has positive effects on epigenetics[76] and is critical for the establishment of optimal reference intake values for specific nutrients during lactation. This in turn creates a personalised pattern of nutrition, programming a healthy phenotype in early childhood that will continue into adulthood.[77] Interestingly, there are reports that extended breastfeeding has been positively associated with increased childhood consumption of vegetables, even amongst obesity-prone young children.[78,79] Just as the consumption of vegetables and exercise is important for boosting physical and mental wellness, breastfeeding has been associated with enhanced cognitive performance and socio-affective responses in children, promoting positive affect and social behaviour, while relieving stress and anti-social behaviour.[75] An early investment into a child's health, education, development, security and well-being, provides benefits that compound during their lifetime, and increases prospects for their future and for that of their children and, thus, society as a whole.[80]

Conclusion

South Africa has a high prevalence of NCDs namely, obesity, diabetes, CVDs, cancer and mental health disorders. A large part of the population (including many young children) lives in poverty and under challenging socio-economic environments due to high unemployment, alcohol and substance abuse, and inter-partner violence, amongst others. It is plausible that adverse socio-environmental conditions together with modified epigenetic mechanisms are responsible for amplified disease susceptibility and diminished health outcomes, as is witnessed in the current generation, and if left unchecked can persist to worsen the situation in future generations. The plight of young poverty-stricken children requires urgent attention and should be prioritised and placed at the centre of the country's sustainable developmental goals. It is imperative in South Africa, as well as in countries experiencing similar socio-economic challenges, that children's health and well-being is improved in order to circumvent an impending catastrophe.

Acknowledgements

The work herein was made possible through funding by the South African Medical Research Council's Biomedical Research and Innovation Platform and the Division of Research Capacity Development under the Intra-mural Post-doctoral Programme. The content hereof is the sole responsibility of the authors and does not necessarily represent the official views of the South African Medical Research Council or the funders. We acknowledge the support of colleagues at the Biomedical Research and Innovation Platform, and Corporate and Marketing Communications, South African Medical Research Council.

Competing interests

We declare that there are no competing interests.

Authors' contributions

E.S. conceptualised and produced the original draft. All authors contributed to the paper, with N.C., C.M. and J.L. providing overall guidance. E.S. and Y.A. finalised the manuscript based on comments and feedback from other authors.

References

1. World Health Organization (WHO). Global status report on noncommunicable diseases 2014. Geneva: WHO; 2014.

2. World Health Organization (WHO). Noncommunicable diseases country profiles 2018. Geneva: WHO; 2018.

3. Silvaggi F, Leonardi M, Guastafierro E, Quintas R, Toppo C, Foucaud J, et al. Chronic diseases & employment: An overview of existing training tools for employers. Int J Environ Res Public Health. 2019;16:718. https://doi.org/10.3390/ijerph16050718

4. Allen LN, Feigl AB. What's in a name? A call to reframe non-communicable diseases. Lancet Glob Health. 2017;5:129–130. https://doi.org/10.1016/S2214-109X(17)30001-3

5. Mayosi BM, Flisher AJ, Lalloo UG, Sitas F, Tollman SM, Bradshaw D. The burden of non-communicable diseases in South Africa. Lancet. 2009;374:934–947. https://doi.org/10.1016/S0140-6736(09)61087-4

6. Pillay-van Wyk V, Msemburi W, Laubscher R, Dorrington RE, Groenewald P, Glass T, et al. Mortality trends and differentials in South Africa from 1997 to 2012: Second National Burden of Disease Study. Lancet Glob Health. 2016;4:642–653. https://doi.org/10.1016/S2214-109X(16)30113-9

7. Norman R, Schneider M, Bradshaw D, Jewkes R, Abrahams N, Matzopoulos R, et al. Interpersonal violence: An important risk factor for disease and injury in South Africa. Popul Health Metr. 2010;8:32. https://doi.org/10.1186/1478-7954-8-32

8. Berkowitz N, Okorie A, Goliath R, Levitt N, Wilkinson RJ, Oni T. The prevalence and determinants of active tuberculosis among diabetes patients in Cape Town, South Africa, a high HIV/TB burden setting. Diabetes Res Clin Pract. 2018;138:16–25. https://doi.org/10.1016/j.diabres.2018.01.018

9. Stringhini S, Carmeli C, Jokela M, Avendaño M, Muennig P, Guida F, et al. Socioeconomic status and the 25 × 25 risk factors as determinants of premature mortality: A multicohort study and meta-analysis of 1·7 million men and women. Lancet. 2017;389:1229–1237. https://doi.org/10.1016/S0140-6736(16)32380-7

10. Notterman DA, Mitchell C. Epigenetics and understanding the impact of social determinants of health. Pediatr Clin North Am. 2015;62:1227–1240. https://doi.org/10.1016/j.pcl.2015.05.012

11. Sulla V, Zikhali P. Overcoming poverty and inequality in South Africa: An assessment of drivers , constraints and opportunities (English). Washington DC : World Bank Group; 2018.

12. Hall K, Richter L, Mokomane Z, Lake L. Families and the state collaboration and contestation [document on the Internet]. c2018 [cited 2021 Apr 09]. Available from: http://webcms.uct.ac.za/sites/default/files/image_tool/images/367/South African Child Gauge 2018 - Nov 20.pdf

13. Hall K, Sambu W, Berry L, Giese S, Almeleh C, Rosa S. South African early childhood review 2016. Cape Town: Children's Institute, University of Cape Town and Ilifa Labantwana; 2016 [cited 2021 Apr 09]. Available from: http://bettercarenetwork.org/sites/default/files/South%20African%20Early%20Childhood%20Review%202016.pdf

14. Sun J, Patel F, Rose-Jacobs R, Frank DA, Black MM, Chilton M. Mothers' adverse childhood experiences and their young children's development. Am J Prev Med. 2017;53:882–891. https://doi.org/10.1016/j.amepre.2017.07.015

15. Boyce TW. A biology of misfortune. Focus. 2012;29:1–6.

16. De Lannoy A, Swartz S, Lake L, Smith C. Broad overview of the South African Child Gauge 2015 [document on the Internet].c2015 [cited 2021 Apr 09] Available from: http://www.ci.uct.ac.za/sites/default/files/image_tool/images/367/Child_Gauge/South_African_Child_Gauge_2015/ChildGauge2015-lowres.pdf

17. Temple NJ, Steyn NP, Fourie J, De Villiers A. Price and availability of healthy food: A study in rural South Africa. Nutrition. 2011;27:55–58. https://doi.org/10.1016/j.nut.2009.12.004

18. Igumbor EU, Sanders D, Puoane TR, Tsolekile L, Schwarz C, Purdy C, et al. 'Big food,' the consumer food environment, health, and the policy response in South Africa. PLoS Med. 2012;9(7), e1001253. https://doi.org/10.1371/journal.pmed.1001253

19. Colchero MA, Rivera-Dommarco J, Popkin BM, Ng SW. In Mexico, evidence of sustained consumer response two years after implementing a sugar-sweetened beverage tax. Health Aff. 2017;36:564–571. https://doi.org/10.1377/hlthaff.2016.1231

20. Nakamura R, Mirelman AJ, Cuadrado C, Silva-Illanes N, Dunstan J, Suhrcke M. Evaluating the 2014 sugar-sweetened beverage tax in Chile: An observational study in urban areas. PLoS Med. 2018;15, e1002596. https://doi.org/10.1371/journal.pmed.1002596

21. Okop KJ, Lambert EV, Alaba O, Levitt NS, Luke A, Dugas L, et al. Sugar-sweetened beverage intake and relative weight gain among South African adults living in resource-poor communities: Longitudinal data from the STOP-SA study. Int J Obes. 2019;43:603–614. https://doi.org/10.1038/s41366-018-0216-9

22. Steyn NP, Bradshaw D, Norman R, Joubert JD, Schneider M. Dietary changes and the health transition in South Africa: Implications for health policy. FAO Food Nutrition Paper. c2006 [cited 2021 Apr 09] Available from: http://www.fao.org/3/a0442e/a0442e00.pdf

23. GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. Lancet. 2019;393:1958–1972. https://doi.org/10.1016/S0140-6736(19)30041-8

24. Shisana O, Labadarios D, Rehle T, Simbayi L, Zuma K, Dhansay A, et al. South African National Health and Nutrition Examination Survey (SANHANES-1). Cape Town: HSRC Press; 2014.

25. Kupolati MD, MacIntyre UE, Gericke GJ, Becker P. A contextual nutrition education program improves nutrition knowledge and attitudes of South African teachers and learners. Front Public Health. 2019;7:258. https://doi.org/10.3389/fpubh.2019.00258

26. Venter FC, Walsh CM, Slabber M, Bester CJ. Body size perception of African women (25-44 years) in Mangaung. J Family Ecol Consumer Sci. 2009;37:12–23. https://doi.org/10.4314/jfecs.v37i1.48942

27. Okop KJ, Mukumbang FC, Mathole T, Levitt N, Puoane T. Perceptions of body size, obesity threat and the willingness to lose weight among black South African adults: A qualitative study. BMC Public Health. 2016;16:1–13. https://doi.org/10.1186/s12889-016-3028-7

28. Ezzati M, Riboli E. Can noncommunicable diseases be prevented? Lessons from studies of populations and individuals. Science. 2012;337:1482–1487. https://doi.org/10.1126/science.1227001

29. Arena R, Berra K, Kaminsky L, Whitsel L, Berra K, Lavie CJ, et al. Healthy lifestyle interventions to combat noncommunicable – A novel nonhierarchical connectivity model for key stakeholders: A policy statement from the American Heart Association, European Society of Cardiology, European Association for Cardiovascu. Eur Heart J. 2015;36:2097–2109. https://doi.org/10.1093/eurheartj/ehv207

30. Devereux S, Hochfeld T, Karriem A, Mensah C, Morahanye M, Msimango T, et al. School feeding in South Africa: What we know, what we don't know, what we need to know, what we need to do [document on the Internet]. c2018 [cited 2021 Apr 09]. Available from: https://foodsecurity.ac.za/wp-content/uploads/2018/06/CoE-FS-WP4-School-Feeding-in-South-Africa-11-jun-18.pdf

31. Hertzman C. Putting the concept of biological embedding in historical perspective. Proc Natl Acad Sci USA. 2012;109(suppl):17160–17167. https://doi.org/10.1073/pnas.1202203109

32. Essex MJ, Boyce WT, Hertzman C, Lam LL, Armstrong JM. Epigenetic vestiges of early developmental adversity: Childhood stress exposure and DNA methylation in adolescence. Child Dev. 2014;84:58–75. https://doi.org/10.1111/j.1467-8624.2011.01641.x

33. Kubota T. Epigenetic alterations induced by environmental stress associated with metabolic and neurodevelopmental disorders. Environ Epigenet. 2016;2, dvw017. https://doi.org/10.1093/eep/dvw017

34. Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci USA. 2008;105:17046–17049. https://doi.org/10.1073/pnas.0806560105

35. Waterland RA, Kellermayer R, Laritsky E, Rayco-Solon P, Harris RA, Travisano M, et al. Season of conception in rural Gambia affects DNA methylation at putative human metastable epialleles. PLoS Genet. 2010;6, e1001252. https://doi.org/10.1371/journal.pgen.1001252

36. Gilbert R, Widom CS, Browne K, Fergusson D, Webb E, Janson S. Burden and consequences of child maltreatment in high-income countries. Lancet. 2009;373:68–81. https://doi.org/10.1016/S0140-6736(08)61706-7

37. Hobbs A, Ramsay M. Epigenetics and the burden of noncommunicable disease: A paucity of research in Africa. Epigenomics 2015;7:627–639. https://doi.org/10.2217/epi.15.17

38. Hoffman SJ, Røttingen J-A. A framework convention on obesity control? Lancet. 2011;378:2068. https://doi.org/10.1016/S0140-6736(11)61894-1

39. Lee A, Cardel M, Donahoo WT. Social and environmental factors influencing obesity. In: Feingold KR, Anawalt B, Boyce A, Chrousos G, de Herder WW, Dhatariya K, et al., editors. Endotext [Internet]. South Dartmouth (MA): MDText.com, Inc; 2000.

40. Ng M. Global, regional and national prevalence of overweight and obesity in children and adults 1980-2013: A systematic analysis. Lancet. 2014;384:766–781.

41. Van Biljon A, McKune AJ, DuBose KD, Kolanisi U, Semple SJ. Physical activity levels in urban-based South African learners: A cross-sectional study of 7 348 participants. S Afr Med J. 2018;108:126–131. https://doi.org/10.7196/SAMJ.2018.v108i2.12766

42. Reddy SP, Resnicow K, James S, Funani IN, Kambaran NS, Omardien RG, et al. Rapid increases in overweight and obesity among South African adolescents: Comparison of data from the South African National Youth Risk Behaviour Survey in 2002 and 2008. Am J Public Health. 2012;102:262–268. https://doi.org/10.2105/AJPH.2011.300222

43. Tydeman-Edwards R, Van Rooyen FC, Walsh CM. Obesity, undernutrition and the double burden of malnutrition in the urban and rural southern Free State, South Africa. Heliyon. 2018;4, e00983. https://doi.org/10.1016/j.heliyon.2018.e00983

44. Glastras SJ, Chen H, Pollock CA, Saad S. Maternal obesity increases the risk of metabolic disease and impacts renal health in offspring. Biosci Rep. 2018;38(2), BSR20180050. https://doi.org/10.1042/BSR20180050

45. Joubert J, Norman R, Bradshaw D, Goedecke JH, Steyn NP, Puoane T et al. Estimating the burden of disease attributed to excess weight in SA. S Afr Med J. 2007;97(8 Pt 2):683–690.

46. NCD Risk Factor Collaboration (NCD-RisC) – Africa Working Group. Trends in obesity and diabetes across Africa from 1980 to 2014: An analysis of pooled population-based studies. Int J Epidemiol. 2017;46:1421–1432.

47. Malambo P, Kengne AP, Lambert EV, De Villiers A, Puoane T. Prevalence and socio-demographic correlates of physical activity levels among South African adults in Cape Town and Mount Frere communities in 2008-2009. Arch Public Health. 2016;74:54. https://doi.org/10.1186/s13690-016-0167-3

48. Evans RW, Smith T, McWade D, Angouras N, Van Aarde RF, Kay P, et al. The need for biokineticists in the South African public health care system. S Afr J Sport Med. 2016;28:85–86. https://doi.org/10.17159/2078-516X/2016/v28i3a1310

49. Hartley M, Tomlinson M, Greco E, Comulada WS, Stewart J, Le Roux I, et al. Depressed mood in pregnancy: Prevalence and correlates in two Cape Town peri-urban settlements. Reprod Health. 2011;8:9. https://doi.org/10.1186/1742-4755-8-9

50. Rochat TJ, Tomlinson M, Newell ML, Stein A. Detection of antenatal depression in rural HIV-affected populations with short and ultrashort versions of the Edinburgh Postnatal Depression Scale (EPDS). Arch Womens Ment Health. 2013;16:401–410. https://doi.org/10.1007/s00737-013-0353-z

51. Essex MJ, Klein MH, Cho E, Kalin NH. Maternal stress beginning in infancy may sensitize children to later stress exposure: Effects on cortisol and behavior. Biol Psychiatry. 2002;52:776–784. https://doi.org/10.1016/S0006-3223(02)01553-6

52. Olivier L, Curfs LMG, Viljoen DL. Fetal alcohol spectrum disorders: Prevalence rates in South Africa. S Afr Med J. 2016;106(6 suppl):S103–S106. https://doi.org/10.7196/SAMJ.2016.v106i6.11009

53. Groves AK, Moodley D, McNaughton-Reyes L, Martin SL, Foshee V, Maman S. Prevalence, rates and correlates of intimate partner violence among South African women during pregnancy and the postpartum period. Matern Child Health J. 2015;19:487–495. https://doi.org/10.1007/s10995-014-1528-6

54. Abrahams N, Mathews S, Martin LJ, Lombard C, Jewkes R. Intimate partner femicide in South Africa in 1999 and 2009. PLoS Med. 2013;10(4), e1001412. https://doi.org/10.1371/journal.pmed.1001412

55. Lopes C. Intimate partner violence: A helpful guide to legal and psychosocial support services. S Afr Med J. 2016;106:966. https://doi.org/10.7196/SAMJ.2016.v106i10.11409

56. Miller AH, Maletic V, Raison CL. NIH Public access. Psychiatry Interpers Biol Process. 2009;65:732–741. https://doi.org/10.1016/j.biopsych.2008.11.029

57. Plant DT, Pariante CM, Sharp D, Pawlby S. Maternal depression during pregnancy and offspring depression in adulthood: Role of child maltreatment. Br J Psychiatry. 2015;207:213–220. https://doi.org/10.1192/bjp.bp.114.156620

58. Shilubane HN, Ruiter RAC, Van den Borne B, Sewpaul R, James S, Reddy PS. Suicide and related health risk behaviours among school learners in South Africa: Results from the 2002 and 2008 national youth risk behaviour surveys. BMC Public Health. 2013;13:926. https://doi.org/10.1186/1471-2458-13-926

59. Van Coppenhagen B, Duvenage HS. Prevalence of depression in people living with HIV and AIDS at the Kalafong Provincial Tertiary Hospital Antiretroviral Clinic. S Afr J Psychiatry. 2019;25, art. #1175. https://doi.org/10.4102/sajpsychiatry.v25i0.1175

60. Jack H, Wagner RG, Petersen I, Thom R, Newton CR, Stein A, et al. Closing the mental health treatment gap in South Africa: A review of costs and cost-effectiveness. Glob Health Action. 2014;7:23431. https://doi.org/10.3402/gha.v7.23431

61. Trangenstein PJ, Morojele NK, Lombard C, Jernigan DH, Parry CDH. Heavy drinking and contextual risk factors among adults in South Africa: Findings from the International Alcohol Control study. Subst Abuse Treat Prev Policy. 2018;13:43. https://doi.org/10.1186/s13011-018-0182-1

62. Peltzer K, Phaswana-Mafuya N. Drug use among youth and adults in a population-based survey in South Africa. S Afr J Psychiatry. 2018;24:1139. https://doi.org/10.4102/sajpsychiatry.v24i0.1139

63. Khasakhala L, Sorsdahl KR, Harder VS, Williams DR, Stein DJ, Ndetei DM. Lifetime mental disorders and suicidal behaviour in South Africa. Afr J Psychiatry. 2011;14:134–139. https://doi.org/10.4314/ajpsy.v14i2.5

64. Stein DJ, Seedat S, Herman A, Moomal H, Heeringa SG, Kessler RC, et al. Lifetime prevalence of psychiatric disorders in South Africa. Br J Psychiatry. 2008;192:112–117. https://doi.org/10.1192/bjp.bp.106.029280

65. Guillemot-Legris O, Muccioli GG. Obesity-induced neuroinflammation: Beyond the hypothalamus. Trends Neurosci. 2017;40:237–253. https://doi.org/10.1016/j.tins.2017.02.005

66. Gomez-Pinilla F, Nguyen TTJ. Natural mood foods: The actions of polyphenols against psychiatric and cognitive disorders. Nutr Neurosci. 2012;15:127–133. https://doi.org/10.1179/1476830511Y.0000000035

67. Ahima RS, Antwi DA. Brain regulation of appetite and satiety. Endocrinol Metab Clin North Am. 2008;37:811–823. https://doi.org/10.1016/j.ecl.2008.08.005

68. Samodien E, Johnson R, Pheiffer C, Mabasa L, Erasmus M, Louw J, et al. Diet-induced hypothalamic dysfunction and metabolic disease, and the therapeutic potential of polyphenols. Mol Metab. 2019;27:1–10. https://doi.org/10.1016/j.molmet.2019.06.022

69. Lambert K, Hokayem M, Thomas C, Fabre O, Cassan C, Bourret A, et al. Combination of nutritional polyphenols supplementation with exercise training counteracts insulin resistance and improves endurance in high-fat diet-induced obese rats. Sci Rep. 2018;8:2885. https://doi.org/10.1038/s41598-018-21287-z

70. Fang M, Chen D, Yang CS. Dietary polyphenols may affect DNA methylation. J Nutr. 2007;137:223S–228S. https://doi.org/10.1093/jn/137.1.223S

71. Voisin S, Eynon N, Yan X, Bishop DJ. Exercise training and DNA methylation in humans. Acta Physiol. 2015;213:39–59. https://doi.org/10.1111/apha.12414

72. Van Wyk B-E. The potential of South African plants in the development of new medicinal products. S Afr J Bot. 2011;77:812–829. https://doi.org/10.1016/j.sajb.2011.08.011

73. Kramer MS, Kakuma R. Optimal duration of exclusive breastfeeding. Cochrane database Syst Rev. 2012; CD003517. https://doi.org/10.1002/14651858.CD003517.pub2

74. Walters DD, Phan LTH, Mathisen R. The cost of not breastfeeding: Global results from a new tool. Health Policy Plan. 2019;34(6):407–417. https://doi.org/10.1093/heapol/czz050

75. Krol KM, Grossmann T. Psychologische Effekte des Stillens auf Kinder und Mütter [Psychological effects of breastfeeding on children and mothers]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz. 2018;61:977–985. German. https://doi.org/10.1007/s00103-018-2769-0

76. Verduci E, Banderali G, Barberi S, Radaelli G, Lops A, Betti F, et al. Epigenetic effects of human breast milk. Nutrients. 2014;6:1711–1724. https://doi.org/10.3390/nu6041711

77. Palou M, Picó C, Palou A. Leptin as a breast milk component for the prevention of obesity. Nutr Rev. 2018;76:875–892. https://doi.org/10.1093/nutrit/nuy046

78. Soldateli B, Vigo A, Giugliani ERJ. Effect of pattern and duration of breastfeeding on the consumption of fruits and vegetables among preschool children. PLoS ONE. 2016;11, e0148357. https://doi.org/10.1371/journal.pone.0148357

79. Specht IO, Rohde JF, Olsen NJ, Heitmann BL. Duration of exclusive breastfeeding may be related to eating behaviour and dietary intake in obesity prone normal weight young children. PLoS ONE. 2018;13, e0200388. https://doi.org/10.1371/journal.pone.0200388

80. Clark H, Coll-Seck AM, Banerjee A, Peterson S, Dalglish SL, Ameratunga S, et al. A future for the world's children? A WHO-UNICEF-Lancet Commission. Lancet. 2020;395:605–658. https://doi.org/10.1016/S0140-6736(19)32540-1

AUTHORS:
Craig A. Keyes[1]
Khumo L. Liphoko[1]

AFFILIATION:
[1]Department of Forensic Medicine and Pathology, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Craig Keyes

EMAIL:
craig.keyes@wits.ac.za

DATES:
Received: 06 May 2020
Revised: 11 Dec. 2020
Accepted: 13 Dec. 2020
Published: 28 May 2021

HOW TO CITE:
Keyes CA, Liphoko KL. A 5-year overview of fatal thermal and electrical burns in Johannesburg, South Africa. S Afr J Sci. 2021;117(5/6), Art. #8288. https://doi.org/10.17159/sajs.2021/8288

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☒ On request from author(s)
☐ Not available
☐ Not applicable

EDITORS:
Pascal Bessong
Sandiswa Mbewana

KEYWORDS:
forensic science, unnatural death, mortality rates, fatal burn incidence

FUNDING:
None

A 5-year overview of fatal thermal and electrical burns in Johannesburg, South Africa

Thermal and electrical burn injuries affect millions of people globally each year. South Africa is a developing country where fatal fires are common. Despite the pervasiveness of fatal thermal and electrical burns in South Africa, there is a paucity of information regarding the frequency of such fatal burns in the City of Johannesburg. We aimed to describe the demographics and frequency of fatal thermal and electrical burn cases received at the Johannesburg Forensic Pathology Services Medico-Legal Mortuary for medico-legal death investigations, and compare the burn mortality rates between Johannesburg and those reported in Cape Town, Pretoria, and Mpumalanga. This study was a 5-year (2010–2014) retrospective review of 185 forensic medico-legal case files of fatal burns (156 thermal burn cases and 29 electrical burn cases) received at the Johannesburg Forensic Pathology Services Medico-Legal Laboratory. The demographics at the greatest risk of fatal burns in Johannesburg, South Africa are black (2.11 per 100 000) and male (1.03 per 100 000) individuals, and those in the 30–39-year age group (3.6 per 100 000). Children aged 0–9 years had a high mortality rate due to thermal burns (3.44 per 100 000). The most common type of fatal burns is thermal in nature, as a result of flames (1.62 per 100 000). Electrical burns are relatively rare (0.3 per 100 000). Burns are prevalent in the winter months, most notably in August. Burn wounds are common on the head, chest, back, and abdomen. Johannesburg has an overall lower burn mortality rate and incidence frequency than Cape Town, Pretoria, and Mpumalanga. This study is the first to report on fatal burns in Johannesburg, South Africa.

Significance:

- A description is provided of fatal thermal and electrical burns of individuals whose deaths were investigated at the Johannesburg Forensic Pathology Services Medico-Legal Mortuary.
- The results highlight the demographic groups at risk of fatal burns in Johannesburg, South Africa.
- Johannesburg has a lower fatal burn incidence and mortality rate than Cape Town, Pretoria, and Mpumalanga; however, young people, particularly male individuals, are at greatest risk.

Introduction

In South Africa, it is the legal mandate of the Forensic Pathology Services (FPS) to perform post-mortem examinations in all cases of unnatural death.[1] Approximately 60 000–80 000 unnatural deaths occur annually in South Africa.[2] South African legislation, such as the *Inquest Act (Act 58 of 1959)*, requires that all unnatural deaths have a medico-legal autopsy examination performed.[3] An unnatural death is defined by South African law as:

a. Any death due to physical or chemical influence, direct or indirect, and/or related complications.

b. Any death, including those deaths which would normally be considered to be a death due to natural causes, which in the opinion of the medical practitioner, has been the result of an act of commission or omission which may be criminal in nature.

c. Any death as contemplated in the *Health Professions Act 56 of 1974*: Section 56 (as amended). The death of a person undergoing, or as a result of, a procedure of a therapeutic, diagnostic or palliative nature, or of which any aspect of such a procedure has been a contributory cause, shall not be deemed to be a death from natural causes.

d. A sudden and unexpected, or unexplained, death or where the cause of death is not apparent.[1,3,4]

In the Gauteng Province, there are 11 FPS medico-legal mortuaries which are divided into two clusters: the Northern Cluster (consisting of three medico-legal mortuaries: Pretoria, Ga-Rankuwa, and Bronkhorstspruit) and the Southern Cluster (consisting of eight medico-legal mortuaries: Johannesburg, Germiston, Diepkloof, Roodepoort, Sebokeng, Carletonville, Springs, and Heidelberg). Approximately 17.5% of all unnatural deaths in South Africa are received by the Gauteng Southern Cluster.[2] Over the 2006–2018 period, the Gauteng Southern Cluster received 155 338 cases, averaging 11 949 cases per annum.

The City of Johannesburg municipality has a population of 4.9 million people, making it South Africa's most populous city.[5] As a result, the Johannesburg FPS mortuary, which services a large portion of the City of Johannesburg municipality, is one of the busiest FPS mortuaries in South Africa. The Johannesburg FPS performed a total of 36 043 post-mortem examinations between the years 2006 and 2018, averaging 2773 cases per annum.

Deaths precipitated by thermal and electrical burns are classified as unnatural deaths and a post-mortem examination must be performed by the FPS. Thermal and electrical burn injuries affect millions of people globally each year.[6] Severe burns rank as the fourth highest incidence of injuries requiring medical attention.[7] It is estimated that 300 000 people die annually due to flame- or fire-related burns globally.[6] Lower-middle- to low-income countries in the Middle East, Asia and Africa[6,8], such as Pakistan[9], Iraq[10], Ethiopia[11], and Kenya[12], make up 90% of global burn fatalities.

Thermal burns include burns due to flames, scalding, and contact with a hot surface. In developing countries, such as South Africa, fatal and severe thermal injuries caused by flames frequently occur in cities that are densely populated by people of lower income and are residing in informal residential areas.[9] The causes of the fires have been due to the usage of affordable domestic appliances and methods such as kerosene stoves and open fires, which are used for cooking and heating water.[8,13,14] Informal settlement fires are a common occurrence in South Africa specifically and Africa generally. Ombati et al.[12] found that informal settlement fires are the second highest contributors to burn injuries, compared to the small percentage of burn injuries that occur in formal residential areas.

Scalding burns result from contact of the skin and other tissues with hot liquids.[15] The American Burn Association reported that scalding burns were common in young children below the age of 5 years.[15] Most scalding burns are a result of hot foods and spilled liquids and they tend to cause minor burns that are typically not fatal.[15] However, the effect of scalding burns in the elderly is greater, and even small scalding burns can have fatal consequences.[16]

Severe injuries due to electrical burns are also common in developing nations. The theft of electrical cables is a common problem in South Africa and may result in fatal electrocution.[2] Additionally, illegal electrical connections are common in informal settlements and can have deadly consiquences.[17] Most electrical fatalities are unintentional, and result from poor maintenance of electrical equipment and carelessness.[2]

The epidemiology and prevalence of surviving victims of thermal and electrical burns, treated in South African clinics, burns units, and hospitals, are well documented.[18-26] However, there is a scarcity of published data on fatal burns in South Africa. A few regional studies have been conducted in the Western Cape, Mpumalanga, and Gauteng[13,17,27], with one published study in Durban on suicide by self-immolation[28].

The studies for Cape Town[13] and Mpumalanga[27] used the National Injury Mortality Surveillance System (NIMSS) as their primary source of data and not specific case files or post-mortem reports. There are limitations in using the NIMSS data, especially the limited detail provided. For example, NIMSS combines burn deaths caused by flames, scalding, and contact to a hot object under a single category, which makes comparisons between the three causes of burns difficult. Ideally, data should be collected from FPS case file documents which provide greater detail on each fatal case. The study by Morobadi et al.[17] is the only published South African study, so far, to have reported on fatal thermal burns from data collected directly from FPS case files and post-mortem reports.

The mortality rate of deaths caused by electrical burns has largely been under-researched in South Africa. Only one study, by Blumenthal[2], has reported the incidence of fatal electrical deaths in Gauteng (using NIMSS data).

Despite the pervasiveness of fatal thermal and electrical burns in South Africa, there is a paucity of information regarding the frequency of such fatal burns in Johannesburg. Johannesburg is the largest city in South Africa and the capital of Gauteng – the most densely populated province in South Africa. The Johannesburg FPS is also one of the busiest forensic mortuaries in South Africa. Therefore, in this study, we aimed to describe the incidence and mortality rate of fatal thermal and electrical burns received at the Johannesburg Forensic Pathology Services Medico-Legal Laboratory for medico-legal death investigations. A comparison of the thermal and electrical burn mortality rates for Johannesburg, Cape Town, Pretoria, and Mpumalanga is also provided.

Methods

Data were collected retrospectively from the case files of the Johannesburg Forensic Pathology Services Medico-Legal Mortuary. The inclusion criterion was any case for which the autopsy report stated that thermal or electrical burn injuries were the primary cause of death or contributory to the cause of death. Data were collected for the 5-year period 2010–2014. This amounted to 185 cases of individuals who died as a result of flames, scalding, explosions, electrocution, or lightning strikes. The information that was collected included demographic details, temporal information, geographical location, and the anatomical regions which exhibited burn injuries. The anatomical regions were divided into: head and neck, thorax, back, abdomen (including the lower back), pelvis, upper limbs, and lower limbs.[29]

A descriptive analysis of the cases was performed. The incidence and mortality rates were stratified by age, sex, and population group (as outlined in the FPS and police documentation). This was reported for the thermal burn cases, electrical burn cases, and all burn cases (thermal and electrical combined). The mortality rates (expressed per 100 000 per year) were calculated using the 2011 Census[30] data for the Johannesburg FPS catchment area (estimated population size of 1 922 249) which includes the following areas: Johannesburg, Randburg, Sandton, Midrand, Alexandra, and Diepsloot. The population estimates (total population for the mortuary's catchment area, and relative population size for each age group, sex, and population group) provided by the 2011 Census[30] were used as denominators when calculating the burn mortality rates.

Ethical clearance for this study was provided by the Human Research Ethics Committee (Medical) (reference: M150315) of the University of the Witwatersrand and the principles of the Declaration of Helsinki were adhered to. The data set is available on request from authors.

Results

Prevalence of fatal burns

The Johannesburg FPS received a total of 185 thermal and electrical burn fatalities between the years 2010 and 2014, with an annual average of 37 cases. This constituted 1.5% of all unnatural deaths (N=12 591) and a mortality rate of 1.92 (all mortality rates reported are per 100 000). Thermal burns constituted 1.2% of all unnatural deaths (n=156) with a mortality rate of 1.62. Flame-related incidents were the most prevalent cause of thermal burns (n=123; mortality rate = 1.62) followed by scalding (n=23; mortality rate=0.24), and explosions (n=7; mortality rate = 0.07) (Table 1).

Electrical burns were infrequent and only constituted 0.2% of all unnatural deaths, with a mortality rate of 0.30. Electrical burns were predominantly electrocutions (n=28, mortality rate = 0.29); with one case of a fatal lightning strike (mortality rate = 0.01) (Table 1).

Demographics

For all burn types (thermal, electrical, and combined), the mortality rates were highest in black individuals (thermal = 2.11; electrical = 0.37; combined = 2.48) and in male individuals (thermal = 1.03; electrical = 0.23; combined = 1.25). The 30–39-year age group had the highest mortality rate for all burn types (thermal = 3.60; electrical = 0.36; combined = 4.70) (Table 1). The age group with the second highest mortality rate was 20–29 years for both electrical burns (0.24) and the combined thermal and electrical burn cases (3.56); and 0–9 years for thermal burns (3.44) (Table 1).

Temporal frequency

There was an overall decrease in the total number of fatal burns from 2010 (n=58) to 2014 (n=16); however, there was a spike in the number of fatalities in 2013 (n=46) (Figure 1). Most fatal burn cases occurred in the cooler months (July–October), especially in July and August. Unusually, a relatively large number of deaths repeatedly occurred in the warmer month of February (with a particularly high number in the year 2013) (Figure 2). This pattern was common to both thermal and electrical burns (Figure 2).

Table 1: Population and sex breakdown of the incidence of burn mortalities (per 100 000) in Johannesburg from 2010 to 2014

Age range (years)		0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	Total	% Unnatural deaths	Annual average	Rate (per 100 000)
Thermal burns		**25**	**6**	**35**	**36**	**14**	**21**	**5**	**9**	**5**	**156**	**1.2**	**31.2**	**1.62**
Cause of burn	**Flame**	16	6	29	33	12	13	4	7	3	**123**	**1.0**	**24.6**	**1.28**
	Scalding	8	0	2	1	2	5	1	2	2	**23**	**0.2**	**4.6**	**0.24**
	Explosion	0	0	3	2	0	2	0	0	0	**7**	**0.1**	**1.4**	**0.07**
	Not recorded	1	0	1	0	0	1	0	0	0	**3**	**0.02**	**0.6**	**0.03**
Sex	**Male**	12	2	25	26	7	16	4	5	2	**99**	**0.8**	**19.8**	**1.03**
	Female	13	4	10	10	7	5	1	4	3	**57**	**0.5**	**11.4**	**0.60**
Population group	**Black**	21	6	33	33	11	15	3	1	1	**124**	**1.0**	**24.8**	**2.11**
	White	0	0	1	2	3	3	2	7	3	**21**	**0.2**	**4.2**	**1.00**
	Coloured	0	0	1	0	0	1	0	0	0	**2**	**0.02**	**0.4**	**0.23**
	Asian	4	0	0	1	0	2	0	1	1	**9**	**0.1**	**1.8**	**1.39**
	Other	0	0	0	0	0	0	0	0	0	**0**	**0.0**	**0**	**0.00**
Thermal burns mortality rate		**3.44**	**1.07**	**2.90**	**3.60**	**2.39**	**(>50 years: 0.38)**				**1.62**	–	–	–
Electrical burns		**0**	**0**	**8**	**11**	**1**	**9**	**0**	**0**	**0**	**29**	**0.2**	**5.8**	**0.30**
Cause of burn	**Electrocution**	0	0	8	10	1	9	0	0	0	**28**	**0.2**	**5.6**	**0.29**
	Lightning	0	0	0	1	0	0	0	0	0	**1**	**0.01**	**0.2**	**0.01**
Sex	**Male**	0	0	5	8	0	9	0	0	0	**22**	**0.2**	**4.4**	**0.23**
	Female	0	0	3	3	1	0	0	0	0	**7**	**0.1**	**1.4**	**0.07**
Population group	**Black**	0	0	7	8	1	6	0	0	0	**22**	**0.2**	**4.4**	**0.37**
	White	0	0	1	3	0	3	0	0	0	**7**	**0.1**	**1.4**	**0.33**
	Coloured	0	0	0	0	0	0	0	0	0	**0**	**0.0**	**0**	**0.00**
	Asian	0	0	0	0	0	0	0	0	0	**0**	**0.0**	**0**	**0.00**
	Other	0	0	0	0	0	0	0	0	0	**0**	**0.0**	**0**	**0.00**
Electrical burns mortality rate		**0.00**	**0.19**	**0.24**	**0.36**	**0.41**	**(>50 years: 0.16)**				**0.30**	–	–	–
Thermal and electrical combined		**25**	**6**	**43**	**47**	**15**	**30**	**5**	**9**	**5**	**185**	**1.5**	**37**	**1.92**
Sex	**Male**	12	2	30	34	7	25	4	5	2	**121**	**1.0**	**24.2**	**1.25**
	Female	13	4	13	13	8	5	1	4	3	**64**	**0.5**	**12.8**	**0.67**
Population group	**Black**	21	6	40	41	12	21	3	1	1	**146**	**1.2**	**29.2**	**2.48**
	White	0	0	2	5	3	6	2	7	3	**28**	**0.2**	**5.6**	**1.33**
	Coloured	0	0	1	0	0	1	0	0	0	**2**	**0.02**	**0.4**	**0.23**
	Asian	4	0	0	1	0	2	0	1	1	**9**	**0.1**	**1.8**	**1.39**
	Other	0	0	0	0	0	0	0	0	0	**0**	**0.0**	**0**	**0.00**
Thermal and electrical burn mortality rate		**3.44**	**1.07**	**3.56**	**4.70**	**2.56**	**(>50 years: 0.54)**				**1.92**	–	–	–

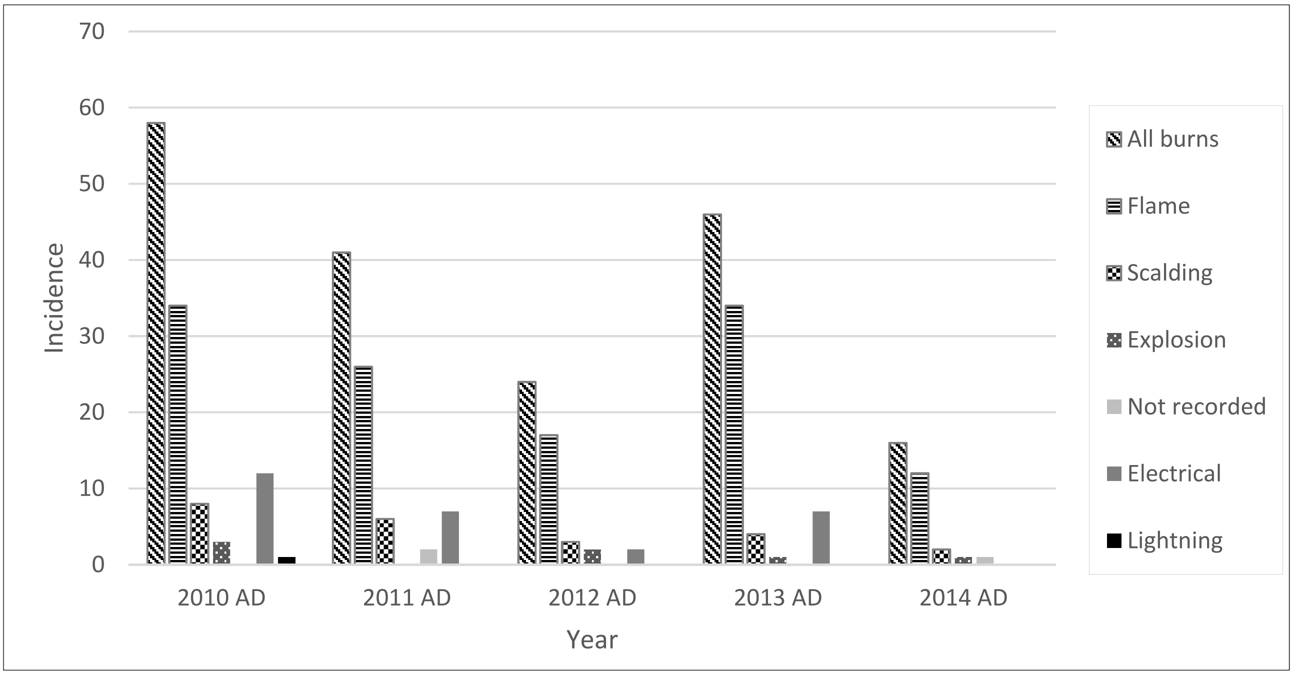


Figure 1: Yearly comparison of fatal thermal and electrical burn incidences in Johannesburg from 2010 to 2014.

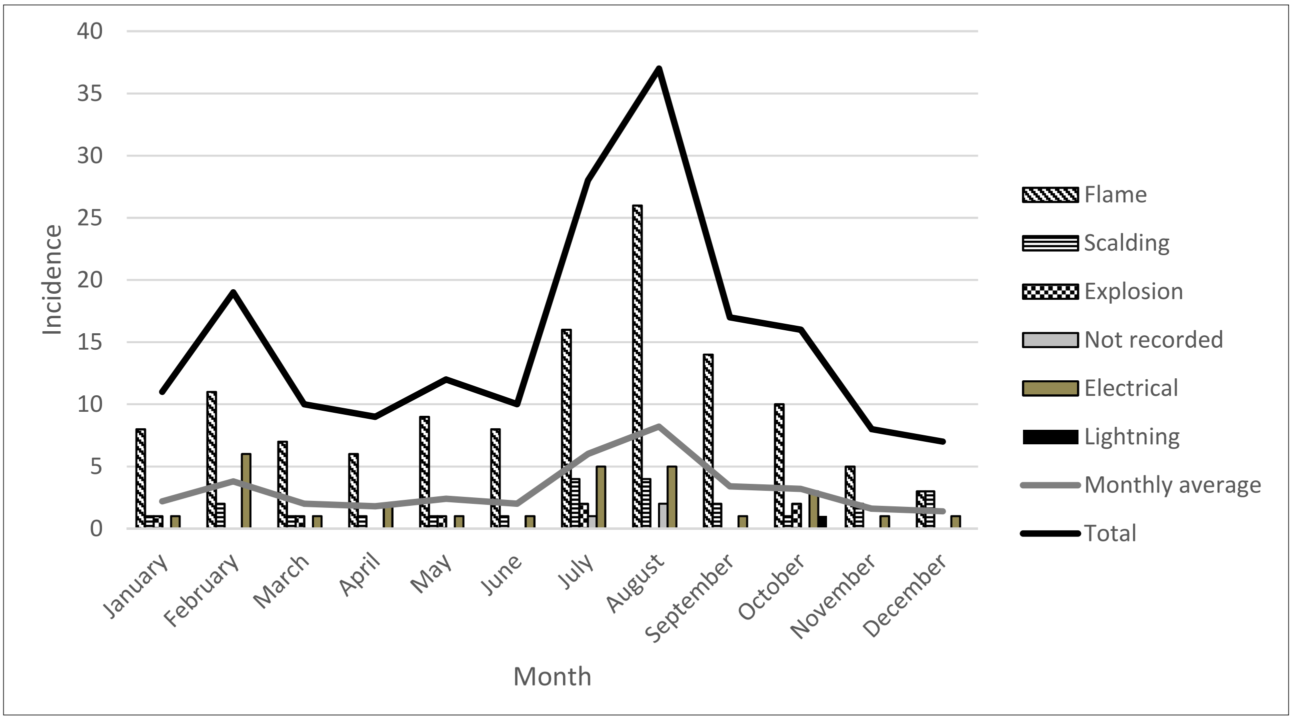


Figure 2: Monthly comparison of fatal thermal and electrical burn incidences in Johannesburg from 2010 to 2014.

Location

The settlement typology and death scene were unknown in many cases because this information was not adequately recorded (in 39.4% and 30.8% of cases, respectively). The incidence of fatal burns was most common in residential settings (52% for both thermal burns and electrical burns). Fatal thermal burns typically occurred in suburbs and informal settlements (54% accumulative), particularly in houses and informal dwellings (56.6% accumulative) (Table 2). Fatal electrical burns were higher in the city and suburbs, in houses and places of employment (Table 2).

Description of burn severity

Overall, burn injuries due to flames were common (≥60%) on all body regions, except for the feet (Figure 3). Scalding burns were common on the thorax, back, and abdomen (Figure 3). Burns that resulted due to explosions were commonly observed on the head, neck, thorax, back, abdomen, arms, and legs (Figure 3).

There was no clear pattern to the anatomical location of burns caused by electrocution; however, the electrical burns were observed on the chest in 57.1% of cases and on the right arm in 50% of cases. The single lightning case showed burn injuries noted on the head, neck, thorax, lower back, and left foot.

Table 2: Burn mortality death scenes in Johannesburg described by location and site

Location	Thermal					Electrical			Thermal and electrical
	Flame	Scalding	Explosion	Not recorded	Total	Electrocution	Lightning	Total of electrical	Total
Settlement typology									
Not recorded	42	12	5	1	60	11	0	11	71
Suburb	33	8	1	0	42	6	1	7	49
Informal settlement	42	1	0	0	43	4	0	4	47
City	3	2	0	0	5	7	0	7	12
Other	3	0	1	2	6	0	0	0	6
Death scene									
House	36	14	0	0	50	7	0	7	57
Not recorded	34	5	1	3	43	8	0	8	51
Informal dwelling	38	1	0	0	39	1	0	1	40
Open area	8	0	3	0	11	5	1	6	17
Place of employment	2	2	3	0	7	6	0	6	13
Other	5	1	0	0	6	1	0	1	7

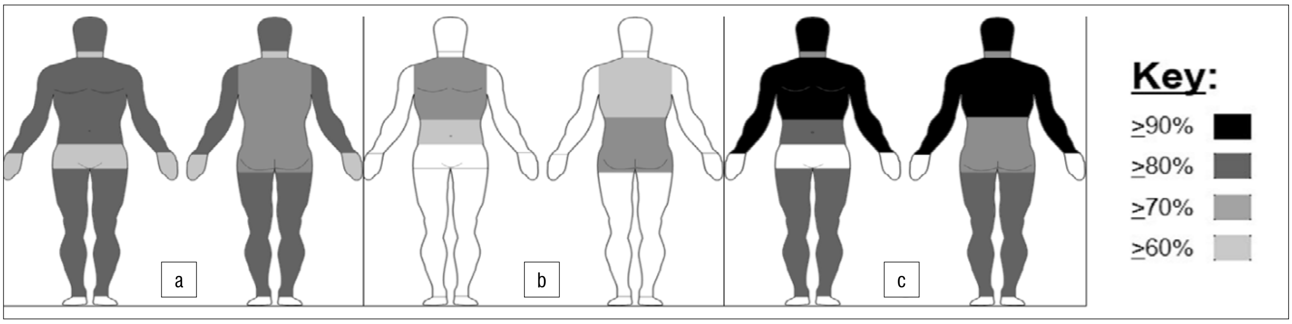


Figure 3: Body regions that exhibited thermal burn injuries in >60% of fatal burn cases caused by (a) flames, (b) scalding, and (c) explosion; on the anterior (left) and posterior (right) body surfaces.

Overall, the feet were the least commonly burned body region (flames = 54.5%, scalding = 39.1%, explosion = 14.3%, electrocution = 7.1%) (Figure 3). In only 27% of all burn cases was the body thermally altered severely enough for a pugilistic posture to be adopted.

Discussion

Mortality rate and cause of burns

Fatal burns constituted only 1.5% of unnatural deaths received by the Johannesburg FPS. Thermal burns were more common and resulted in a higher mortality rate (1.2 per 100 000 per year) than did electrical burns (0.3 per 100 000 per year). Flames were the most common cause of fatal burns and the scene was often in a residential dwelling (formal and informal settlements). This scenario is common in South Africa as it has been observed in studies performed in other South African locations.[13,17,25] The high number of fatal thermal and flame-related burns in informal residential areas is due, in part, to unsafe cooking appliances such as petroleum-based stoves.[14] These stoves are cheap and portable. Furthermore, candles are often used as a light source. These unsafe appliances, combined with the flammability of the materials used to construct informal residences, non-compliance with fire safety standards[31], and the close proximity and density of the dwelling structures allow for fires to spread quickly.

Fatal electrocutions are relatively rare. In the USA, fatal electrocutions occur at a rate of 0.43 per 100 000 per year.[32] Comparatively, Johannesburg experienced a similarly low mortality rate of 0.3 per 100 000 per year (only 28 cases in 5 years). Fatal electrocutions were observed only in individuals within the 20–59-year age bracket and occurred most often in city centres and suburban areas. A likely cause of these electrocutions is cable theft. There has been an ongoing escalation in cable theft in urban centres in South Africa as individuals attempt to retrieve the copper wires to sell. These thefts have led to the accidental electrocution and deaths of the culprits and the engineers who attempt to repair or maintain vandalised substations.[33] This reason, however, is speculative as little information is provided in such cases and the claims of cable theft are often pronounced only by the media.[2]

Demographics

Overall, the Johannesburg burns mortality rate (thermal and electrical) was substantially higher in black men aged 20–39 years, than in any other racial, sex, and age group. Although the catchment area of the Johannesburg FPS is 61.2% black, 50.2% male, and 23% falls into the 20–39-year age group (the largest age group by number of individuals), the mortality rate for each group was disproportionally higher than their relative cohorts.

Regions with a low socio-economic status often present with larger numbers of burn cases than regions with higher socio-economic status. Each of these regions is often predominately composed of particular population or ethnic groups, due to historical inequalities. Differences in burn mortality rates among racial and ethnic groups is a phenomenon that is observed globally and is related to socio-economic challenges. Socio-economic status has a greater impact on a group's susceptibility to burns than do cultural or educational factors.[15] It is therefore unsurprising that the black population in the Johannesburg region had a higher mortality rate, due to the catchment area's predominantly low socio-economic status and the effects thereof on the population in the region.

Generally, the mortality rate of male individuals tends to be higher than that of female individuals.[34] There are numerous theories that have been proposed for why male individuals are more likely to be injured than female individuals, such as that they are socialised differently[35,36] and tend to engage in activities that involve riskier behaviours and act more impulsively than female individuals[37,38]. These behavioural observations are general but have been observed in both young and adult individuals.[15]

The higher mortality rate of male individuals, particularly in the 20–39-year age range, is also not unexpected as the fire mortality rate has also been found to be predominant in men in this age group in other studies in the USA and Australia.[15,39,40] This is likely attributable to the employment and domestic activities of individuals in this age group.

In Johannesburg, young children aged 0–9 years had a particularly high mortality rate due to thermal burns. This is a trend that corresponds to global trends. Globally, fire-related death rates occur most commonly in children under the age of 4 years.[15] The death rate begins to increase again from the age of 15 years, which is related to the age-related changes in behaviours associated with risk-taking, exposure to hazards, and employment.[7] Children in Africa are at a particularly higher risk for burns compared to the rest of the world. African infants have a three-times higher incidence rate of fire-related burns than the world average.[41] A large number of factors influence the higher likelihood of young children being burned in Africa. Some of these factors may include the lower literacy, age, and education level of mothers, the limited availability and means to access healthcare services, combined cooking and living areas in a dwelling, the lack of smoke alarms and access to water in a residential space, and poor emergency response services.[42] The increased mortality rate in young children can also be attributed to their stage of development which results in an increase in their motor skills and activity rates. This increases the chance of their coming into contact with harmful objects that could result in their being burned.[15]

Temporal variations

The global incidence of burns presented at emergency departments and hospitals has been declining since 1982, as has the number of fatal burns.[15,43] The linear decline in the mortality rate due to burns is a general, global trend; however, annual and national variations are to be expected. This was highlighted in the overall decrease in the number of fatal burn cases in Johannesburg between the years 2010 and 2014, with a spike in 2013. Such variations to the general trend could be due to individual factors that impacted that year. Unfortunately, we cannot provide any reasonable explanations for the spikes in the year 2013 or the general increase in cases seen in the month of February. No probable causes related to climatic changes, population dynamics, or mortuary intake variances could account for the spikes in the respective mortality rates. It is possible that factors related to social, political, and environmental variations could be the cause[15]; however, the exact causes are currently not apparent. Ultimately, the decreasing burn mortality rate seen globally and in Johannesburg is a promising trend that is likely the result of increasing urbanisation and development.

Most fatal burn cases occurred in the winter months of July and August. The cooler temperatures result in people using unsafe heating appliances and open flames to warm themselves, which can be dangerous, especially to the very young and elderly. The winter season in Johannesburg is characterised as very dry with no precipitation. Large open fields of long dry grass, called veld, are common in urban areas. The dry grass and low moisture content often result in veld fires which are a safety hazard in urban areas.

Anatomical regions

The anatomical locations of burns can vary greatly because their causes are contextual. Thermal burns in general were common to the thorax, back, and abdomen. These were also the only body regions to be injured by scalding burns. Burn injuries due to flames exhibited the greatest spread over all body regions (excluding the feet). All thermal burns due to explosions displayed burns to the head, neck, back, and both arms. Fatal electrical burns exhibited very few wounds; however, burns were observed on the right arm in 50% of cases and the chest in 57.1% of cases.

Comparison with other South African studies

Previous studies have reported on fatal burns in Cape Town, Pretoria, and Mpumalanga, and a comparison with the Johannesburg data warrants further investigation. In Johannesburg, fatal burns constituted 1.5% (n=185) of the total number of unnatural deaths (N=1259), which included thermal burns (n=156; 1.2%) and electrical burns (n=29; 0.2%). We collected data from one mortuary, for a 5-year period (2010–2014). In Cape Town, fatal thermal burns constituted 4.6% (n=256) of the total number of unnatural deaths (N=5534). Fatal electrical burns were not reported on.[13] The Cape town data were collected from two mortuaries, over 4 years (2010–2013).[13] In Pretoria, fatal thermal burns constituted 3% (n=291) of the total number of unnatural deaths (N=9558). Fatal electrical burns were not reported on.[17] The Pretoria data were collected from one mortuary, over 5 years (2011–2015).[17] In Mpumalanga, a combined total of 304 cases of thermal and electrical burn mortalities occurred over a 2-year period (2007–2008) at 18 mortuaries.[27] The number of fatal thermal burns and electrical burns were not reported individually and the total number of unnatural deaths was not reported.[27]

Generally, Johannesburg had a lower mortality rate and lower incidence of fatal burns (overall, overall males, and overall females) than did Cape Town[13], Pretoria[17], and Mpumalanga[27] (Tables 3–5). However, Johannesburg had a higher mortality rate in female individuals than did Cape Town[13] (Table 3). Johannesburg had a generally higher mortality rate in the younger age groups than did Cape Town[13] and Mpumalanga[27]; particularly the 0–24-year age group (Tables 3 and 5). Johannesburg also had a higher mortality rate in female individuals in the 0–38-year age range than did Cape Town[13] and Mpumalanga[27] (Tables 3 and 5). Johannesburg, Cape Town[13], and Mpumalanga[27] all had higher mortality rates in the later months of the year, compared to the earlier months (Tables 3 and 5). In particular, Johannesburg had higher mortality rates in June to August, compared to June to November in Cape Town[13] (Table 3). Johannesburg and Mpumalanga[27] both had higher mortality rates in August to October (Table 5).

Overall, Johannesburg had a lower mortality rate due to thermal and electrical burns compared to those reported in Cape Town[13], Pretoria[17], and Mpumalanga[27]. Although this finding is surprising, given the high population density in the region, it is a promising result. This is likely because the catchment area of the Johannesburg FPS covers a largely urbanised and developed region made up of formal dwellings with access to electricity. However, there were certain groups that had higher burn mortality rates in comparison to that in other areas in South Africa. This included young individuals (0–24 years) and female individuals aged 0–38 years. These groups typically experience a disproportionally higher burn incidence rate (as has been previously discussed). The reasons for their higher mortality rates in Johannesburg compared to the rest of South Africa is unknown and should be further investigated.

Limitations and recommendations

Limitations to the study included case file documents that were incomplete or lacking detail. In this study, we did not investigate the pathology of the burns and the causes of death, which can be further explored in future studies. Due to a large national backlog in toxicology testing in South Africa (an up to 7–10-year delay in test results), toxicology, blood alcohol concentration and blood carboxyhaemoglobin level were not explored in the present study.

Table 3: Comparison of age, sex, population group, and seasonal burn mortality rates between Johannesburg and Cape Town (adapted from Van Niekerk et al.[13])

Age (years)	Overall		Male		Female	
	Cape Town	Johannesburg	Cape Town	Johannesburg	Cape Town	Johannesburg
0–15	3.6	12.7	4.3	11.0	2.9	15.4
16–24	7.4	7.7	9.6	11.9	5.2	3.5
25–38	12.8	11.3	19.0	21.8	6.6	7.6
39–50	10.9	13.6	15.5	11.9	6.3	15.3
51+	5.2	3.2	6.4	4.4	4.1	2.0
Total	7.9	7.3	10.9	9.3	4.9	5.2

Season	Cape Town rate	Johannesburg rate
Summer (December–February)	1.9	1.5
Autumn (March–May)	1.7	1.4
Winter (June–August)	2.3	3.0
Spring (September–November)	2.4	1.4

Table 4: Comparison of year, population group, and sex burn mortality incidences between Johannesburg and Pretoria (adapted from Morobadi et al.[17])

Year	Pretoria			Johannesburg		
	Number of fatalities	Number of burn fatalities	% of burn fatalities	Number of fatalities	Number of burn fatalities	% of burn fatalities
2010	–	–	–	2459	45	1.8
2011	2037	69	3.4	2422	34	1.4
2012	1919	69	3.6	2339	22	0.9
2013	1817	54	3	2653	39	1.1
2014	1862	56	3	2718	16	0.6
2015	1923	43	2.2	–	–	–
Total	9558	291	3	12591	156	1.2
Population group	**Number of burn fatalities**		**%**	**Number of burn fatalities**		**%**
Black	249		85.6	124		79.5
White	30		10.3	21		13.5
Indian/Asian	2		0.7	9		5.8
Coloured	5		1.7	2		1.3
Unknown	5		1.7	0		0.0
Total	291		100	**156**		100
Sex	**Number of burn fatalities**		**%**	**Number of burn fatalities**		**%**
Male	214		73.5	99		63.5
Female	75		25.8	57		36.5
Unknown	2		0.7	0		0.0
Total	291		100	156		100

Table 5: Comparison of burn incidence mortality rates (per 100 000) between Johannesburg and Mpumalanga (adapted from Blom et al.[27])

Age (years)	Overall		Male		Female	
	Mpumalanga	Johannesburg	Mpumalanga	Johannesburg	Mpumalanga	Johannesburg
0–4	4.5	4.9	5.4	4.1	3.7	5.8
5–14	1.1	1.2	1.7	1.1	0.5	1.4
15–24	1.7	2.2	2.3	3.1	1.0	1.4
25–34	5.0	4.7	8.1	6.3	2.0	2.9
35–44	6.2	3.5	9.4	4.7	3.7	2.2
45–54	5.8	–	9.8	–	2.3	–
55+	8.2	–	9.9	–	6.9	–
Crude all ages	3.8	1.9	5.3	2.5	2.4	1.3
Location	**Mpumalanga**	**Johannesburg**	**Season**		**Mpumalanga**	**Johannesburg**
Where people live	55.6%	52.4%	Summer (Nov–Jan)		18.9%	14.1%
Other	11.5%	20.0%	Autumn (Feb–Apr)		22.2%	20.5%
Unknown	32.9%	27.6%	Winter (May–Jul)		25.6%	27.0%
			Spring (Aug–Oct)		33.3%	38.4%

A comparison of the burn mortality rate between different manners of death (accident, homicide, suicide) was not included in this study because the manner of death is decided by the courts and not the Forensic Pathology Services. This information is not explicitly present in the medico-legal case file documents. The classification of burn levels (such as first-, second- and third- degree burns) was not presented as this is not typically reported in the autopsy reports.

Conclusion

A 5-year retrospective review of forensic medico-legal case files indicates that the following demographic groups are at greatest risk of fatal burns in Johannesburg, South Africa: black individuals, male individuals, and individuals in the 30–39-year age group. Thermal burns, as a result of flames, are the most common type of burn and are prevalent in the winter months, most notably in August. Electrical burns are relatively rare. Burn wounds are most common to the head, chest, back, and abdomen. Johannesburg has a lower burn mortality rate than other regions in South Africa. This study is the first to report on fatal burns in Johannesburg, South Africa.

Acknowledgements

We acknowledge the Johannesburg Forensic Pathology Services Medico-Legal Laboratory for providing access to the case files and Mr Jaco Louw for providing the data on the case load of the Gauteng Forensic Pathology Service.

Competing interests

We declare that there are no competing interests.

Authors’ contributions

C.A.K.: Conceptualisation; methodology; data analysis; validation; writing – the initial draft; student supervision. K.L.L.: Conceptualisation; methodology; data collection; writing – the initial draft.

References

1. National Health Act No. 61 of 2003, South Africa.
2. Blumenthal R. A retrospective descriptive study of electrocution deaths in Gauteng, South Africa: 2001-2004. Burns. 2009;35(6):888–894. https://doi.org/10.1016/j.burns.2009.01.009
3. The Inquests Act No. 58 of 1959, South Africa.
4. National Health Act No. 61 of 2003, South Africa. Regulations: Rendering of Forensic Pathology Service.
5. Statistics South Africa (StatsSA). Community survey 2016, statistical release P0301. Pretoria: StatsSA; 2016.
6. Peck M, Molnar J, Swart D. A global plan for burn prevention and care. Bull World Health Organ. 2009;87:802–803. https://doi.org/10.2471/BLT.08.059733
7. World Health Organization (WHO). The global burden of disease: 2004 update. Geneva: World Health Organization; 2008.
8. Lerer LB. Homicide-associated burning in Cape Town, South Africa. Am J Forensic Med Pathol. 1994;15(4):344–347. https://doi.org/10.1097/00000433-199412000-00012
9. Al Ibran E, Mirza FH, Memon AA, Farooq MZ, Hassan M. Mortality associated with burn injury – a cross sectional study from Karachi, Pakistan. BMC Res Notes. 2013;6(1):1–3. https://doi.org/10.1186/1756-0500-6-545
10. Qader AR. Burn mortality in Iraq. Burns. 2012;38(5):772–775. https://doi.org/10.1016/j.burns.2011.12.016
11. Courtright P, Haile D, Kohls E. The epidemiology of burns in rural Ethiopia. J Epidemiol Commun Health. 1993;47(1):19–22. https://doi.org/10.1136/jech.47.1.19
12. Ombati AN, Ndaguatha PL, Wanjeri JK. Risk factors for kerosene stove explosion burns seen at Kenyatta National Hospital in Kenya. Burns. 2013;39(3):501–506. https://doi.org/10.1016/j.burns.2012.07.008
13. Van Niekerk A, Laubscher R, Laflamme L. Demographic and circumstantial accounts of burn mortality in Cape Town, South Africa, 2001-2004: An observational register based study. BMC Public Health. 2009;9(1), Art. #374. https://doi.org/10.1186/1471-2458-9-374
14. Ahuja RB, Dash JK, Shrivastava P. A comparative analysis of liquefied petroleum gas (LPG) and kerosene related burns. Burns. 2011;37(8):1403–1410. https://doi.org/10.1016/j.burns.2011.03.014
15. Peck MD. Epidemiology of burns throughout the world. Part I: Distribution and risk factors. Burns. 2011;37(7):1087–1100. https://doi.org/10.1016/j.burns.2011.06.005
16. Alden N, Bessey P, Rabbitts A, Hyden P, Yurt R. Tap water scalds among seniors and the elderly: Socioeconomics and implications for prevention. J Burn Care Res. 2006;27:S97. https://doi.org/10.1097/01253092-200603001-00098

17. Morobadi K, Blumenthal R, Saayman G. Thermal fatalities in Pretoria: A 5-year retrospective review. Burns. 2019;45(7):1707–1714. https://doi.org/10.1016/j.burns.2019.05.007

18. Steenkamp W, Botha N, Van der Merwe A. The prevalence of alcohol dependence in burned adult patients. Burns. 1994;20(6):522–525. https://doi.org/10.1016/0305-4179(94)90012-4

19. Van Niekerk A, Rode H, Laflamme L. Incidence and patterns of childhood burn injuries in the Western Cape, South Africa. Burns. 2004;30(4):341–347. https://doi.org/10.1016/j.burns.2003.12.014

20. Van Niekerk A, Reimers A, Laflamme L. Area characteristics and determinants of hospitalised childhood burn injury: A study in the city of Cape Town. Public Health. 2006;120(2):115–124. https://doi.org/10.1016/j.puhe.2005.08.015

21. Van Niekerk A, Seedat M, Menckel E, Laflamme L. Caregiver experiences, contextualizations and understandings of the burn injury to their child. Accounts from low-income settings in South Africa. Child Care Health Develop. 2007;33(3):236–245. https://doi.org/10.1111/j.1365-2214.2006.00724.x

22. Eyal A, Kemp M, Luvhengo T. A 10-year audit of burns at Kalafong Hospital. Burns. 2007;33(3):393–395. https://doi.org/10.1016/j.burns.2006.07.011

23. Allorto N, Oosthuizen G, Clarke D, Muckart D. The spectrum and outcome of burns at a regional hospital in South Africa. Burns. 2009;35(7):1004–1008. https://doi.org/10.1016/j.burns.2009.01.004

24. Maritz D, Wallis L, Van Der Merwe E, Nel D. The aetiology of adult burns in the Western Cape, South Africa. Burns. 2012;38(1):120–127. https://doi.org/10.1016/j.burns.2010.12.007

25. Blom L, Klingberg A, Laflamme L, Wallis L, Hasselberg M. Gender differences in burns: A study from emergency centres in the Western Cape, South Africa. Burns. 2016;42(7):1600–1608. https://doi.org/10.1016/j.burns.2016.05.003

26. Cloake T, Haigh T, Cheshire J, Walker D. The impact of patient demographics and comorbidities upon burns admitted to Tygerberg Hospital Burns Unit, Western Cape, South Africa. Burns. 2017;43(2):411–416. https://doi.org/10.1016/j.burns.2016.08.031

27. Blom L, Van Niekerk A, Laflamme L. Epidemiology of fatal burns in rural South Africa: A mortuary register-based study from Mpumalanga Province. Burns. 2011;37(8):1394–1402. https://doi.org/10.1016/j.burns.2011.07.014

28. Sukhai A, Harris C, Moorad R, Dada MA. Suicide by self-immolation in Durban, South Africa: A five-year retrospective review. Am J Forensic Med Pathol. 2002;23(3):295–298. https://doi.org/10.1097/00000433-200209000-00020

29. Drake R, Vogl A, Mitchell A. Gray's anatomy for students. 2nd ed. Philadelphia, PA: Churchill Livingstone/Elsevier; 2009.

30. Statistics South Africa (StatsSA). Census 2011 statistical release. Pretoria: StatsSA; 2012.

31. Ono R, Da Silva S. An analysis of fire safety in residential buildings through fire statistics. Fire Safety Sci. 2000;6:219–230. https://doi.org/10.3801/IAFSS.FSS.6-219

32. Taylor AJ, McGwin G, Valent F, Rue L. Fatal occupational electrocutions in the United States. Injury Prevention. 2002;8(4):306–312. https://doi.org/10.1136/ip.8.4.306

33. Dzansi D, Rambe P, Mathe L. Cable theft and vandalism by employees of South Africa's electricity utility companies: A theoretical explanation and research agenda. J Social Sci. 2014;39(2):179–190. https://doi.org/10.1080/09718923.2014.11893281

34. Peck MD. Epidemiology of burns throughout the world. Part II: Intentional burns in adults. Burns. 2012;38(5):630–637. https://doi.org/10.1016/j.burns.2011.12.028

35. Fagot BI. The influence of sex of child on parental reactions to toddler children. Child Develop. 1978;49(2):459–465. https://doi.org/10.2307/1128711

36. Block JH. Differential premises arising from differential socialization of the sexes: Some conjectures. Child Develop. 1983;54(6):1335–1354. https://doi.org/10.2307/1129799

37. Eaton WO, Yu AP. Are sex differences in child motor activity level a function of sex differences in maturational status? Child Develop. 1989;60(4):1005–1011. https://doi.org/10.2307/1131040

38. Rosen BN, Peterson L. Gender differences in children's outdoor play injuries: A review and an integration. Clin Psychol Rev. 1990;10(2):187–205. https://doi.org/10.1016/0272-7358(90)90057-H

39. Forjuoh SN. The mechanisms, intensity of treatment, and outcomes of hospitalized burns: issues for prevention. J Burn Care Rehab. 1998;19(5):456–460. https://doi.org/10.1097/00004630-199809000-00019

40. Begg S, Vos T, Barker B, Stevenson C, Stanley L, Lopez AD. The burden of disease and injury in Australia 2003. PHE 82. Canberra: Australian Institute of Health and Welfare; 2007.

41. Hyder AA, Kashyap K, Fishman S, Wali S. Review of childhood burn injuries in sub-Saharan Africa: A forgotten public health challenge: Literature review. Afr Safety Promotion. 2004;2(2):43–58. https://doi.org/10.4314/asp.v2i2.31610

42. Peden M, Oyegbite K, Ozanne-Smith J, Hyder AA, Branche C, Rahman A, et al. World report on child injury prevention. Geneva: World Health Organization; 2009.

43. Ahuja RB, Bhattacharya S, Rai A. Changing trends of an endemic trauma. Burns. 2009;35(5):650–656. https://doi.org/10.1016/j.burns.2009.01.008

Orthopaedic device innovation in South Africa: A study of patenting activity

AUTHORS:
Faatiema Salie[1]
Kylie de Jager[1]
Tania S. Douglas[1†]
[†]Deceased 20 March 2021

AFFILIATION:
[1]Division of Biomedical Engineering, Department of Human Biology, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Faatiema Salie

EMAIL:
slxfaa002@myuct.ac.za

DATES:
Received: 28 May 2020
Revised: 04 Oct. 2020
Accepted: 11 Nov. 2020
Published: 28 May 2021

HOW TO CITE:
Salie F, De Jager K, Douglas TS. Orthopaedic device innovation in South Africa: A study of patenting activity. S Afr J Sci. 2021;117(5/6), Art. #8399. https://doi.org/10.17159/sajs.2021/8399

ARTICLE INCLUDES:
☒ Peer review
☒ Supplementary material

DATA AVAILABILITY:
☒ Open data set
☐ All data included
☐ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Jenni Case

KEYWORDS:
social network analysis, patent, knowledge development, knowledge exchange, innovation system

FUNDING:
South African National Research Foundation (grant no. 82624), South African Department of Science and Innovation/National Research Foundation (grant no. 98788)

We assessed knowledge development and exchange among actors who patent orthopaedic devices in South Africa over the period 2000–2015. A social network analysis was performed on bibliometric data using co-inventorship on patents as an indicator of collaboration between different organisations, with a focus on the spatial and sectoral contexts. Network metrics and innovation system indices are used to describe knowledge development and exchange. The results show that university, healthcare and industry organisations have primarily been responsible for increased patenting over time. The key actors were a set of industry actors – a national actor and its US partner – who have patented many devices jointly. National universities were found to make a small contribution, and science councils were found to be absent, despite the efforts in the changing innovation landscape to encourage publicly financed research organisations to protect their intellectual property. The collaboration networks were found to be sparse and disjointed, with many actors – largely from the private healthcare sector – patenting in isolation.

Significance:

- The considerable number of patents filed by private sector clinicians in orthopaedic device innovation in their personal capacity is highlighted.
- Few patents emanate from national universities, and science council actors are largely absent, despite the *Intellectual Property Rights from Publicly Financed Research and Development Act* to protect intellectual property emanating from public research organisations.
- Patenting networks are more fragmented than are scientific publication networks.

Introduction

Medical devices have a key role to play in addressing South Africa's burden of disease. South Africa's medical device industry is made up of hundreds of small players yet is dominated by a handful of large multinational corporations, with approximately 90% of all products being imported.[1] Of all the domestic medical device companies, 68% are solely distributing imported devices and 26% manufacture devices locally.[2] Approximately 90% of the local manufacturers also act as distributors of imported devices, suggesting that it may be difficult to act as a local manufacturer (only) in the current South African market.[2] Very little is known about the value of local manufacturing and the products of manufacturers.[2] Of the top ten most imported groups of medical devices into South Africa for 2013, four groups were classified as orthopaedic devices, with a combined value of ZAR1.8 billion. Of the top ten most exported medical devices, only one group was classified as orthopaedic devices, with a value of ZAR50 million. These figures highlight the substantial value of orthopaedic devices imported into South Africa and show some local activity in the development of such devices by the domestic market. While manufacture exclusively for the local medical device market would not be represented in these figures, local manufacturing activity appears to be limited. Addressing this imbalance would require identifying opportunities that might exist to expand orthopaedic device innovation in South Africa, which in turn requires an understanding of the current activity in this area.

One way of assessing innovation activity is through the analysis of patents. Because patents must show novelty, they have been used as an economic indicator of the rate and direction of technological progress and innovation.[3,4] The number of patents filed by an organisation or country has been adopted as a measure of the amount of technological knowledge produced.[5,6]

Fleming and Marx[7] illustrated that collaborations recorded in patent data captured personal and professional ties between inventors. Co-inventors may have collaborated intensively over extended periods of time towards novel inventions.[8] Patents, therefore, serve as tools indicative of a collaborative event[9] and could be exploited to map social ties between inventors[10]. Several authors have used co-inventorship of patents as a proxy for collaboration. Balconi et al.[9] linked co-inventors of patents in a network and derived implications for knowledge exchange, with a focus on technology transfer resulting from university–industry interactions. Patra and Muchie[11] mapped the entrepreneurial and collaborative activity of South African universities using joint patents as an indicator of collaboration.

Literature applying social network analysis to data from scientific publications, patents, or a combination of both, is abundant. In the latter, the focus is often on the translation of basic science to commercialising technologies. Within this body of literature, work on medical device development is limited. One such study is that of Murray[12], who investigated the communication between science and technology networks in cartilage tissue engineering. Additionally, the literature on medical device innovation is scant. One such study is by MacPherson[13] who investigated the impact of academic linkages on the innovation performance of medical device manufacturers in New York City. A direct link was found between innovation propensity and the existence of both formal and informal academic linkages, with radical innovations more associated with academic linkages than incremental innovations. In-house research and development, however, was found to be the strongest factor for product development than any other; positive correlates for innovation were also found with investment in university partnerships, collaboration with other industry actors, proximity to university resources, and patent counts.

The scientific base for orthopaedic device innovation in South Africa was previously investigated in its spatial and sectoral contexts for the period 2000–2015.[14] That study applied a technological innovation system (TIS) framework[15] and explored scientific knowledge development and exchange among actors. Social network analysis on bibliometric data of scientific publications, where co-authorship was used as an indicator of collaboration between organisations, showed that scientific knowledge production increased over time; this knowledge production was due largely to activity by national university and national healthcare actors. Scientific collaboration networks were sparse, indicating barriers to knowledge exchange among actors in the network. One of the limitations of that study is that collaboration networks derived from scientific publications favour sectors that publish scientific output, i.e. universities and academic healthcare facilities. Outputs such as patents may better represent collaboration trends in the industry sector.

In this study, we further investigate knowledge development and exchange in the orthopaedic device TIS in South Africa by focusing on its technological knowledge base. The actors in the orthopaedic device innovation system are identified using bibliometric data from patents and are related using co-inventorship as a proxy for collaboration. Using social network analysis techniques, we quantify relationships between actors and characterise knowledge exchange in the networks. Our aim was to answer the following questions:

1. Who are the actors actively patenting within the orthopaedic device innovation system in South Africa?
2. From which sectors and countries do the actors come, and what is the nature of their inter-sectoral and international collaboration?
3. What overlap, if any, exists between the scientific and technological domains of orthopaedic device innovation in South Africa?

Methods

Data sources

A definition developed for an orthopaedic medical device[14] was adopted for this study to develop a search phrase (Appendix 1) used in TotalPatent from LexisNexis to elicit patents that demonstrate orthopaedic device innovation. The search was performed for patents with a priority date between 1 January 2000 and 31 December 2015.

Each patent retained in the data set had to show evidence of orthopaedic device development, with the primary inclusion criterion of at least one inventor listing a South African address. On patents, inventors are only required to provide a full name and an address, thus patent data do not link the inventors to their organisational affiliations. To overcome the biographical shortcomings of patents, a series of steps was followed to determine the affiliation of the inventor at the patent priority date and to extrapolate inventor data to organisational data:

1. The inventor name was cross-checked against names of authors in the scientific publication data set of Salie et al.[14] If the inventor was also an author of a scientific publication published at the time of patenting, the affiliation listed on the scientific publication was used. This approach is similar to that of Tijssen[16] who matched inventor names to authors of scientific publications in the Science Citation Index database and had success in finding affiliations for inventors who were active in science-based technical areas. ‘At the time of patenting’ refers to a 1-year window with the priority date at its centre.
2. A Google Scholar search of the inventor was performed. If the inventor had a Google Scholar profile listing their publications, the publications were screened to retrieve all affiliations within 1 year of the priority date of the patent.
3. A LinkedIn search of the inventor was performed. If the inventor had a LinkedIn profile, the ‘Experience’ section of the inventor’s profile was viewed to establish their affiliation at the priority date of the patent.
4. A Google search of the inventor was performed. Google searches often resulted in links to social and academic networking platforms, including Facebook and ResearchGate. In the case of clinicians, often the clinician’s practice would be found. In these instances, inventors were contacted (via email, social and academic networking messaging platforms, or phone calls) to confirm their affiliations at the priority date of the patent. While clinicians in South Africa may not be employed by private hospital groups, in this social network analysis, inventors were affiliated to these organisations if they practised therefrom.
5. Where multiple affiliations were retrieved per inventor in the above steps, all affiliations for that inventor were captured. This might result in links between multiple organisations due to the activity of one inventor.

A patent is applied for and granted at different times. Hinze and Schmoch[17] suggest that the patent priority date be used in patent analysis as this date is most closely related to the time of invention. The patent family must also be considered in patent analysis. A patent family contains a set of patent documents that refer to the same technical topic.[5] In this study, patent families were considered as a unit and inventor affiliations at the priority date recorded.

Drawing actor collaboration networks

Collaboration networks were generated using UCINet 6 (Version 6.573)[18] and NetDraw (Version 2.152)[19]. Each node in the network is an organisation to which an inventor is affiliated; co-inventorship, at the organisational level, is indicated by an edge between nodes.

Each node was assigned to one of the following four sectors:

1. Healthcare, which includes hospitals, clinics and specialised healthcare facilities.
2. University, which includes higher education organisations such as universities, universities of technology, colleges, etc.
3. Science council, which includes research organisations other than universities.
4. Industry, which includes individuals and organisations whose goal is to take products to market, usually for profit.

Each organisation appearing on a patent has been counted only once, regardless of the number of inventors affiliated with that organisation. Where one inventor was affiliated with several organisations, a link was created between these organisations.

The networks have been drawn in overlapping 5-year moving windows, from 2000–2004 to 2011–2015; there are a total of 12 time frames in this period. This 5-year window period was adopted from Eslami et al.[20], who assumed the lifespan of network links based on co-authorship to be 5 years, given that information exchange takes place for some time during a collaboration.

Network metrics

The degree centrality[21] is calculated as the number of ties between a given node and other nodes in the network, including self-reflecting ties. Degree centrality serves as an indicator of how active the node is. In this study, normalised degree centrality is reported, as in Equation 1, where the node’s degree, $u(y)$, is divided by the maximum possible degree in the network, $u_{\max}$.

$$|D(y)| = \frac{u(y)}{u_{\max}} \ldots \qquad \text{Equation 1}$$

Betweenness centrality[22] of a node is a measure of how influential that node is in transmitting information across the network. It indicates how often a node (y) lies on the shortest path ($\delta_{xz}(y)$) between the paired combination of all other nodes (x *and* z) in the network, divided by the total number of node pairs, as shown in Equation 2.[22]

$$B(y) = \Sigma_{x \neq y \neq z} \frac{\delta_{xz}(y)}{\delta_{xz}} \ldots \qquad \text{Equation 2}$$

The normalised betweenness centrality is the node's betweenness centrality divided by the maximum possible betweenness of the network, and is reported as a percentage[18], as in Equation 3:

$$|B(y)| = \frac{B(y)}{B_{max}} \dots \qquad \text{Equation 3}$$

Nationalisation index

Binz et al.[23] present metrics and typologies to analyse networks spatially. In this study, their nationalisation index, which measures the dominance of national over international ties, was calculated. It is based on the external–internal (E-I) index by Krackhardt and Stern[24] and is defined as the ratio of links among actors inside one country to links with actors outside that country. The nationalisation index, *N*, is given by Equation 4, where the number of ties among South African actors, L_i, is compared to the number of ties South African actors have with actors in other countries, L_e.

$$N = \frac{\Sigma L\mathrm{i} - \Sigma Le}{\Sigma L\mathrm{i} + \Sigma Le} \dots \qquad \text{Equation 4}$$

If most actors are cooperating in a national context, the index would be positive and tend towards 1. If national and international cooperation are equally present, the index would be close to 0. If international cooperation is dominant, the index would be negative, and tend towards -1.

Sectorisation index

The sectorisation index[14] is an adaptation of the Binz et al.[23] nationalisation index. The sectorisation index compares the number of collaborations between South African actors within the same sector (i.e. universities, healthcare, industry or science councils), s_i, to that with South African actors outside the sector, s_e. This metric, shown in Equation 5, is calculated separately for each sector.

$$S = \frac{\Sigma s\mathrm{i} - \Sigma se}{\Sigma si + \Sigma se} \dots \qquad \text{Equation 5}$$

This index measures the dominance of intra-sectoral collaboration over inter-sectoral collaboration. If most actors are participating in intra-sectoral collaboration, the index would be positive and tend towards 1. If intra- and inter-sectoral collaboration are equally present, the index would be close to 0. If inter-sectoral collaboration is preferred, the index would be negative, and tend towards -1.

Results

The TotalPatent search yielded 1926 results. Patents were manually examined to extract those related to orthopaedic devices which had a priority date between 1 January 2000 and 31 December 2015, and at least one inventor with a South African address. A total of 73 patents met all these criteria and were retained.

Inventor affiliation data for 11 of the 73 patents could not be established. For some inventors, no Internet presence was found. For some patents, affiliation data for some of the inventors, but not all, were obtained. These patents were excluded from the data set. The results reported in this study are from 62 patents filed between 2000 and 2015. A total of 57 organisational actors were identified; Table 1 presents a spatial and sectoral breakdown of the actors. Of these actors, 35(61%) are South African. The organisations are represented in fairly equal numbers from the university, healthcare and industry sectors. National and international university actors are represented in similar numbers. The number of national healthcare and industry actors present is twice that of their international counterparts.

Table 1: Spatial and sectoral breakdown of inventors who patent in orthopaedic device development in South Africa

Sector	National	International	Total
University	7	9	16
Healthcare	14	7	21
Industry	14	5	19
Science council	1	0	1
Total	**35**	**22**	**57**

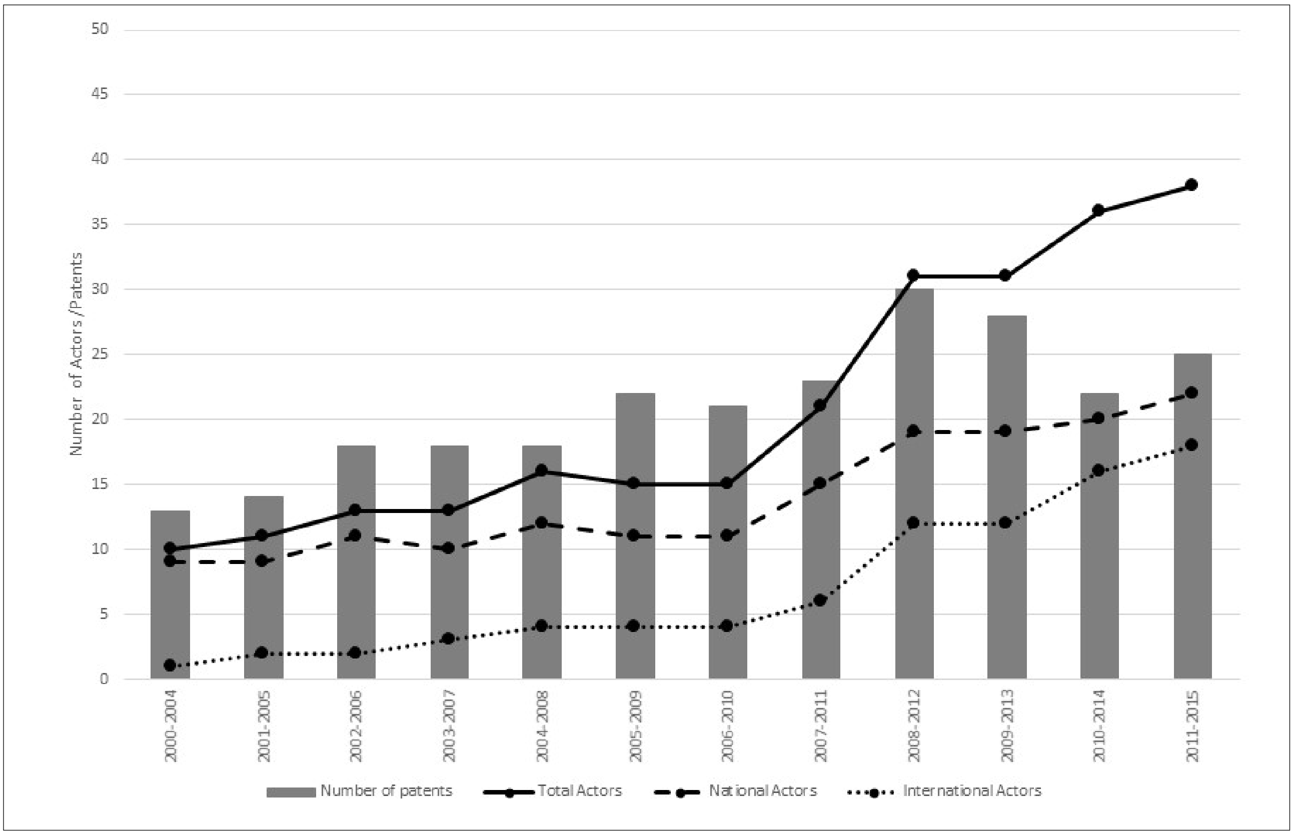


Figure 1: Number of patents and number of national and international actors affiliated with patent inventors.

For the period 2000–2015, 12 overlapping time frames in 5-year moving windows were assessed, starting from 2000–2004 (first time frame) and ending at 2011–2015 (12th time frame). For each time frame, the number of national and international actors and the patents for that period were counted (see Figure 1). There is a gradual increase in the number of patents produced by actors as time progresses, peaking in the ninth (2008–2012) time frame. The total number of actors increases over time; the number of national actors is always greater than the number of international actors. There is a relatively large increase in the number of international actors from the eighth to the ninth time frame (2008–2012).

Selected time frames of the orthopaedic device innovation network are presented in Figure 2. Each actor is represented by a node in the network. Full names of the actors along with their abbreviations are presented in Appendix 2. The actor networks for all 12 time frames are available in the supplementary material to this article.

In the 2000–2004 time frame (Figure 2a), 13 patents were produced by inventors affiliated to 10 organisations. In this time frame, only one international organisation is present – Spinal Motion Inc. (SMI). SMI was, at the patent priority date, an US partner to domestic company Southern Medical (SM). The network component to which SMI and SM belong was largely involved in the development of spinal disc arthroplasty. The other components in the first time frame represent patents arising from individual organisations without collaborators. In the 2005–2009 time frame (Figure 2b), 22 patents were produced by inventors affiliated with 15 organisations. The SM/SMI component has evolved – some previous actors have disappeared, and new actors appear. The patents of this component comprise inventions in spinal fusion devices and disc and lower-arm arthroplasty. The rest of the network comprises either single-node or two-node groups, in which the inventors were either from a single organisation or from two different organisations, or a single inventor was affiliated with two organisations. An example is the UCT/GSH component, where the (single) inventor is affiliated to both the University of Cape Town (UCT) and one of its academic hospitals, Groote Schuur Hospital (GSH). The 2011–2015 time frame (Figure 2c) comprises 25 patents from inventors affiliated with 38 organisations. CMO (Custom Med Orthopaedics), which appears as a single node, has multiple patents for orthopaedic instrumentation. Three components in this network have greater international than national presence – inventor(s) from Saspine are collaborating with inventors affiliated with international healthcare and international industry organisations; inventor(s) from Tshwane University of Technology (TUT) are collaborating with inventors affiliated with international universities; and inventors from Stellenbosch University (SUN) and Stellenbosch MediClinic (SMC) are collaborating with inventors affiliated with international universities, healthcare facilities and industry organisations. While UCT appears to have a central role in this network, it is largely in that position because of the dual affiliations of its inventors.

Across all time frames, there are inventors who chose to patent in isolation. Apart from the Council for Scientific and Industrial Research (CSIR), these inventors are largely from the national healthcare and national industry sectors. These isolated inventors from the national healthcare sector are almost exclusively affiliated with private healthcare facilities, including Netcare Jakaranda Hospital (JH), Life Wilgeleugen Hospital (LWH), Netcare Pinehaven Hospital (PH), Netcare Parklands

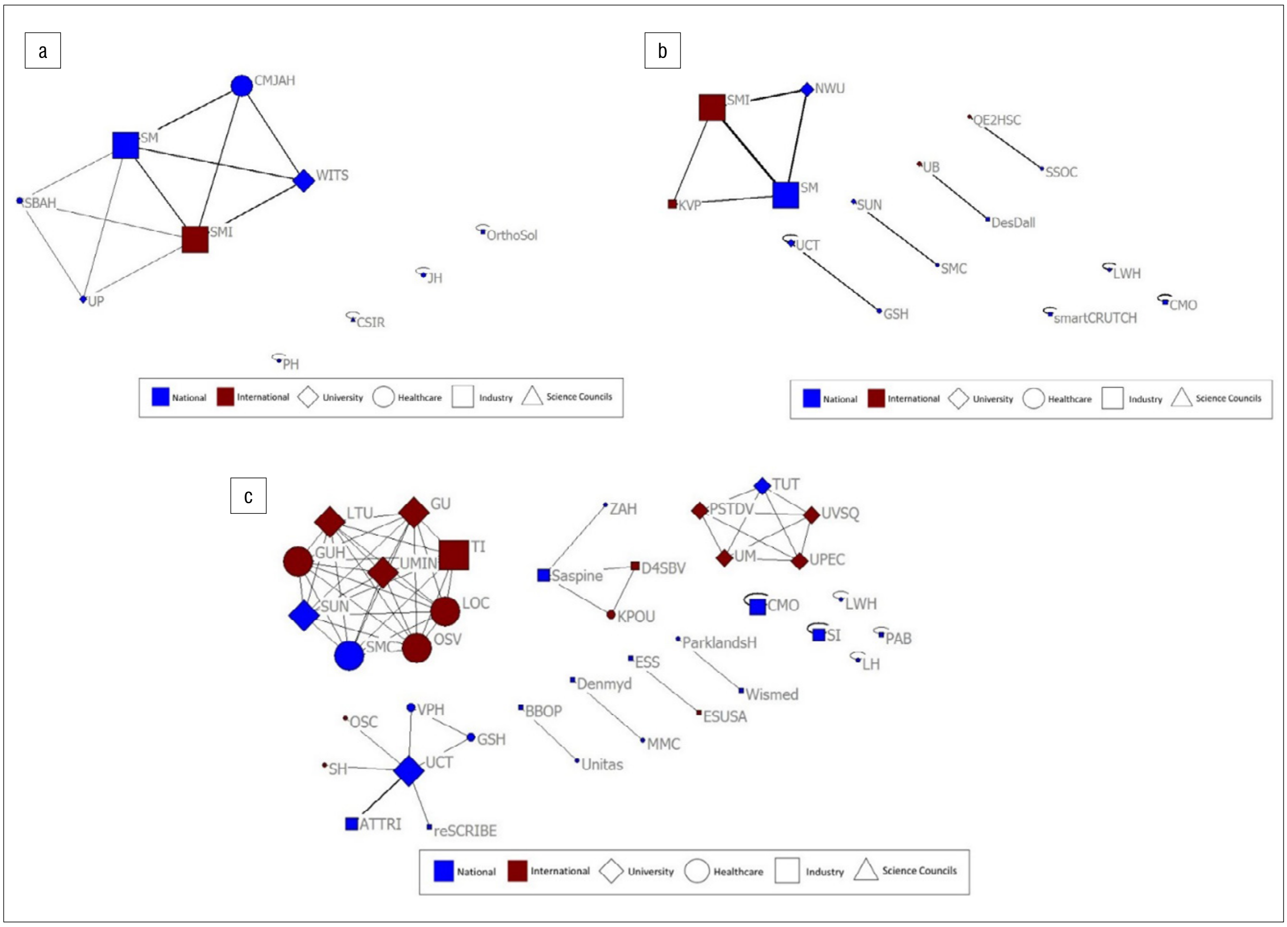


Figure 2: Orthopaedic device innovation network of South Africa based on patent data for (a) 2000–2004, (b) 2005–2009 and (c) 2011–2015. Nodes are sized according to weighted degree centrality. Thickness of edges is weighted to tie strength, i.e. the number of co-invented patents between the nodes. The edges are undirected, as co-inventorship is a reciprocal relationship. A component in the network comprises linked nodes and edges. Where the patent was co-invented by individuals of the same organisation, the ‘collaboration’ is represented by a self-reflecting tie.

Hospital (ParklandsH), Netcare Unitas Hospital (Unitas) and Zuid-Afrikaans Hospital (ZAH). The network contains many actors who have patented only once. On the other end of the spectrum, there are a few actors who have patented many inventions. SM/SMI contributed to 21 of the 62 patents (34%); their parent company, Southern Implants (SI), contributed to a further three patents. CMO contributed to six patents. This is consistent with Balconi et al.[9], who found that very few inventors produce a high number of patents, most producing just one.

Figure 3 presents the evolution of high degree centrality actors of the patent network over the 12 time frames. If an actor's degree was among the top three in any time frame, its degree centrality over all 12 time frames is reported. In some cases, pairs/groups of actors are presented because in some instances the nodes are solely due to the dual affiliations of the inventors, as in the case of WITS/CMJAH and UP/SBAH, or in other cases, inventors have worked together on multiple patents and some have dual affiliations. This is the case for SM/SMI. SM/SMI are high degree actors for the first eight time frames, with their degree centrality decreasing over time. WITS/CMJAH and UP/SBAH are high degree actors in the first five time frames. Their degree centrality values are identical and decreasing. These pairs are only present in the first five time frames and result from university-affiliated clinicians who have co-invented patents with the SM/SMI pair. Nine actors – UMIN, TI, SUN, SMC, OSV, LTU, LOC, GU and GUH – are high degree actors in the last four time frames. They collaborated on a 2013 patent for a set of femoral implants for knee prosthesis and form the largest number of co-inventors listed on a patent in the data set.

Figure 4 presents the evolution of actors having high betweenness centrality. In the entire period, only four actors have betweenness centrality. This includes the SM/SMI pair, UCT and Saspine. The potential of these actors to influence the network is limited to the component in which they operate; the networks remain fragmented across all time frames, limiting knowledge flow.

The nationalisation index of the patent network is presented in Figure 5. The nationalisation index is positive for the first five time frames, but it is ever decreasing. As the network grows, the collaborations become more internationalised, with a negative index between the sixth (2005–2009) and the 11th (2010–2014) time frames. Beyond the fifth time frame, the index increases to zero, and then eases into the positive, suggesting that collaborations become nationalised again. However, the index is very close to zero beyond the seventh time frame, suggesting that the collaborating actors do not show any preference between national and international collaborations. The sectorisation index is presented in Figure 6. Overall, collaborating actors from the national university and industry sectors are largely involved in inter-sectoral collaboration with other national organisations. Across all time frames, the national university and industry actors do not participate in intra-sectoral collaboration with other national actors. This is also the case for the national healthcare sector, except in the very last time frame, where there is a link between two national healthcare actors, GSH and VPH.

Figure 7 illustrates the countries of origin of the actors from different sectors. The actors are from nine different countries, namely France, the United States of America, the United Kingdom, Belgium, Australia, India, Canada, Germany, and the Netherlands.

Discussion

In this study, we have investigated technological knowledge development in the orthopaedic devices TIS in South Africa using patent bibliometric data, identified the actors who are patenting orthopaedic devices, and explored knowledge exchange dynamics by drawing collaboration networks using co-inventorship as a proxy for collaboration. The university, healthcare and industry actors were found to be present in almost equal amounts overall; however, closer examination (Figure 7) reveals that at a geographic level there is greater distinction between actors from different sectors.

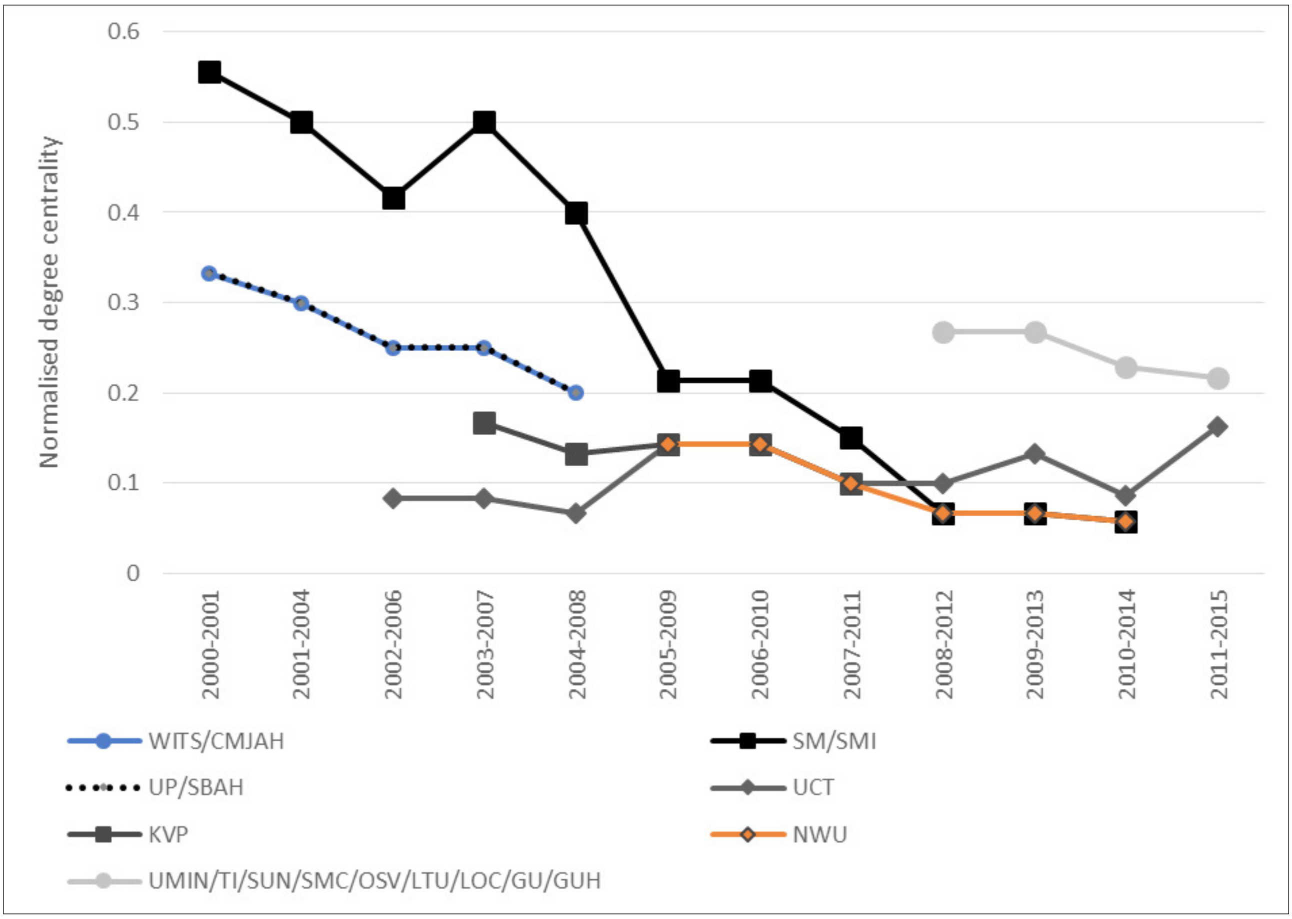


Figure 3: Actors with a high degree centrality in the patent networks.

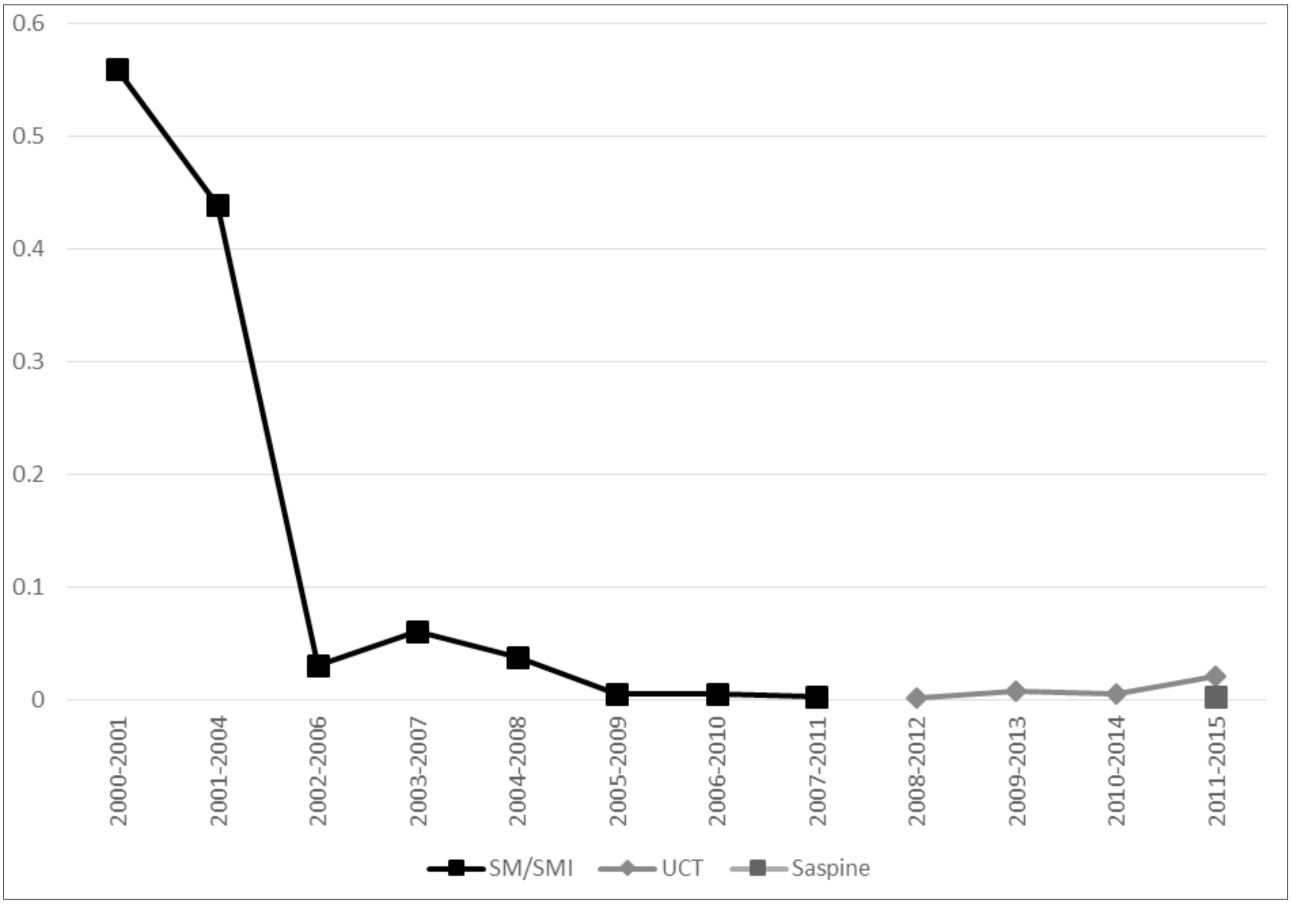


Figure 4: Actors with a high betweenness centrality in the patent networks.

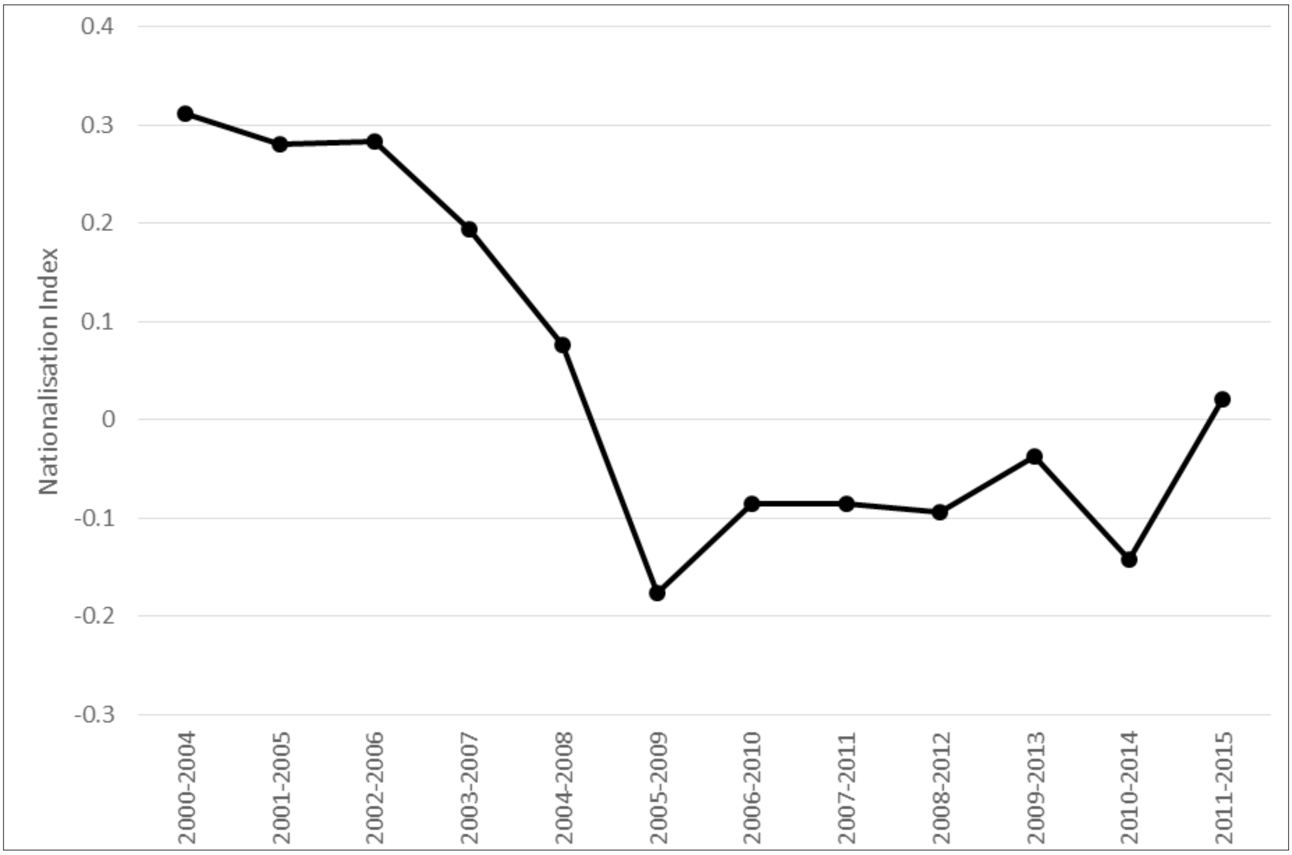


Figure 5: The nationalisation index of the patent network across all 12 time frames.

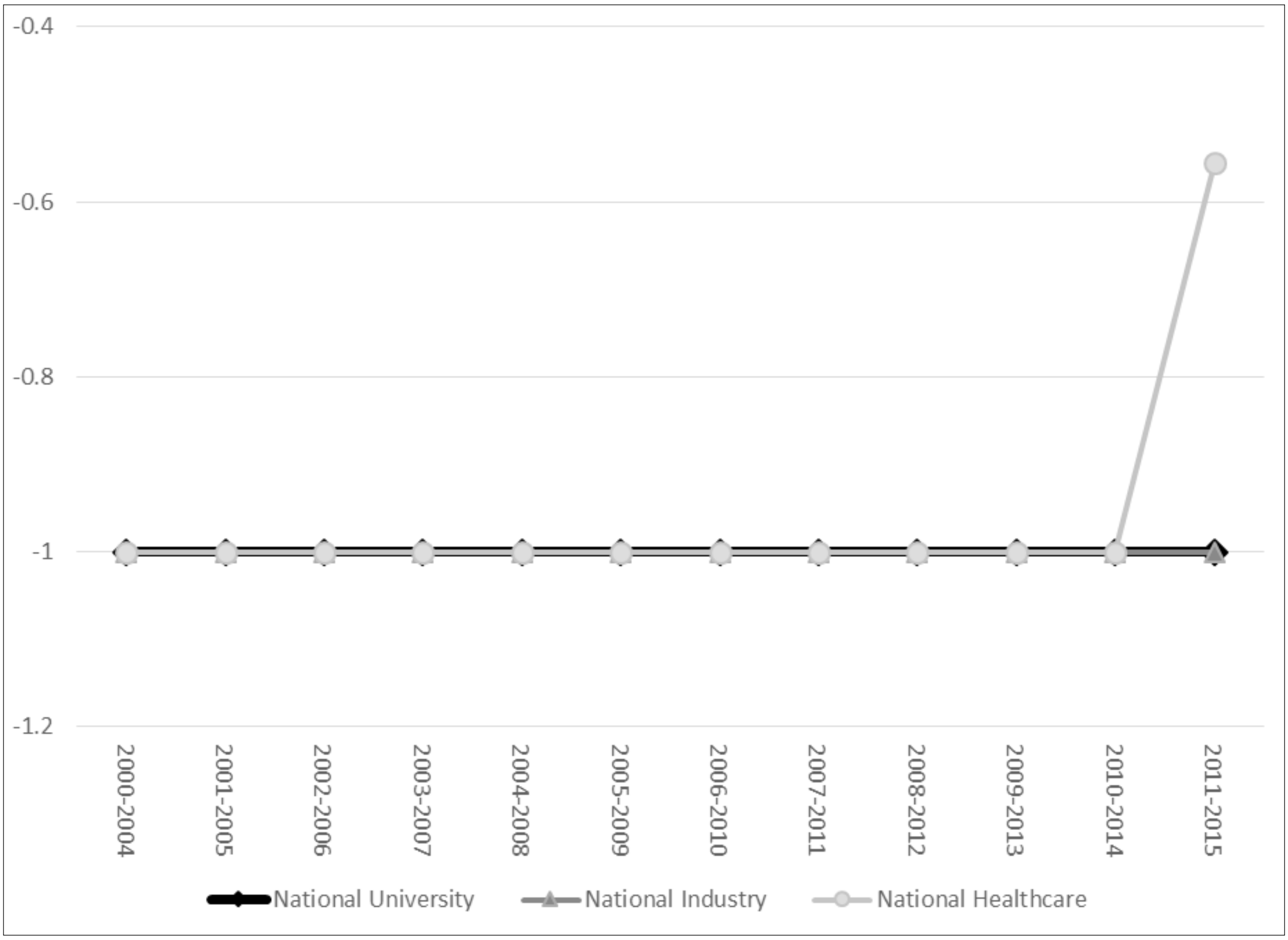


Figure 6: Sectorisation indices of the patent networks across all 12 time frames.

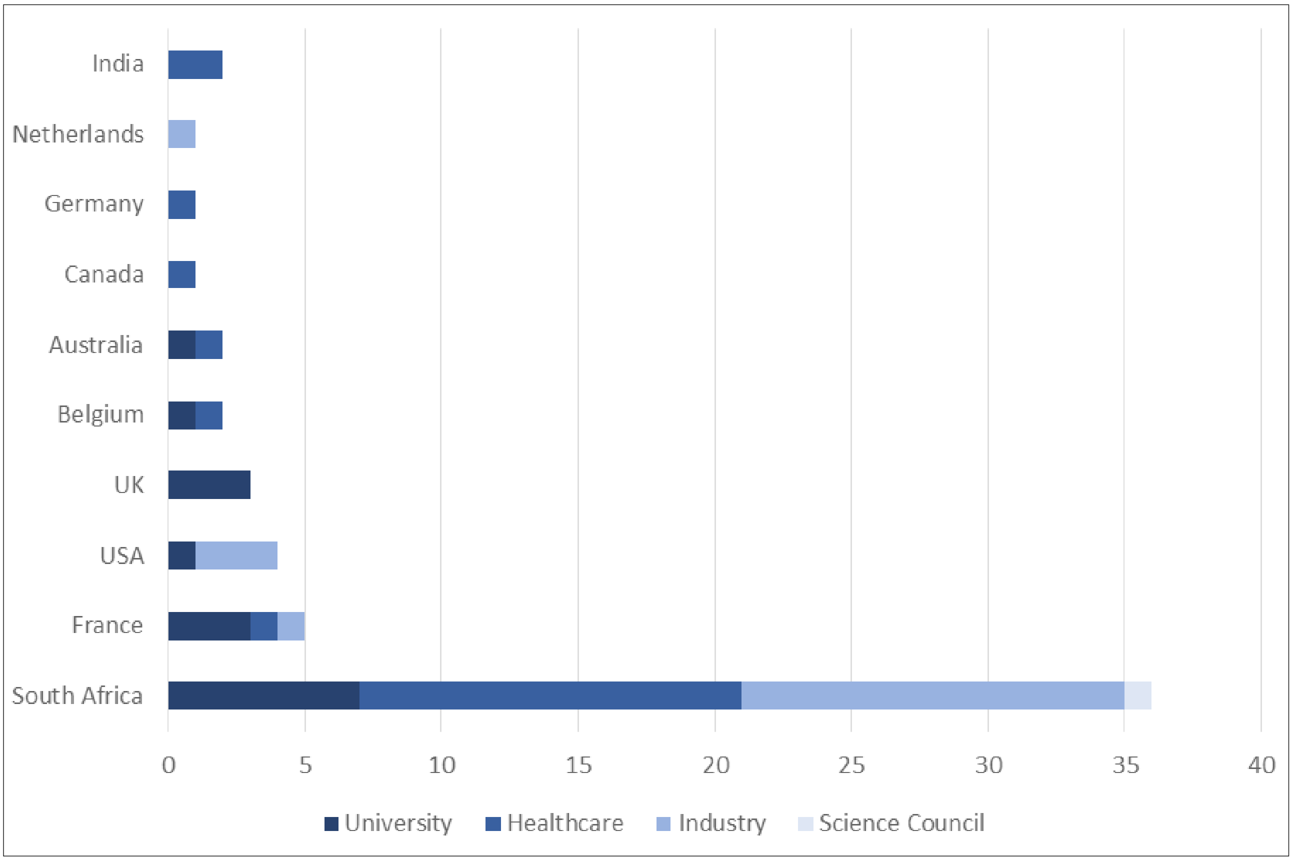


Figure 7: A sectoral breakdown illustrating the location of different actors.

Only one science council actor, the CSIR, appears early in the network, contributing to a single patent. Seven national university actors appear in the patent networks. However, the number of inventors affiliated with national universities in any given time frame is very low – the highest number being four. Patenting from national universities prior to the *Intellectual Property Rights from Publicly Financed Research and Development Act* was low, with a total of only 24 patents emanating from national universities in 2008.[25] By 2015, universities accounted for 14% of the South African portfolio of patents.[11] In contrast, national university actors have produced many scientific publications related to orthopaedic device development.[14] Academics may not be incentivised to patent as they are incentivised to produce scientific publications. Academics would be listed as inventors on university patents and earn royalties therefrom. Incentives available to produce scientific publications may bring more immediate gain, in the form of promotion and the publishing subsidy in the Research Output Policy of the Department of Higher Education and Training (DHET). The DHET has changed the incentive structure, effective in the 2021–22 financial year, to give equal weighting to patents and scientific publications as recognised research outputs for which universities may earn subsidies.[26]

During our period of study, another change occurred in the institutional landscape for innovation in South Africa in the form of the *Intellectual Property Rights from Publicly Financed Research and Development Act, Act 51 of 2008*. This Act led to the establishment of the National Intellectual Property Management Office (NIPMO) and the Intellectual Property Fund in South Africa. The Act provides for more efficient utilisation of intellectual property resulting from publicly financed research and development, and for the establishment of technology transfer offices (TTOs) at publicly financed research organisations (e.g. universities and science councils).

Public organisations like universities, academic hospitals and science councils are bound by the Act, while the Act does not apply to private sector actors from the healthcare and industry sectors. In a study be Ncube et al.[27], one university warned that provisions in the Act would result in a loss of industry-contracted research due to requirements for permissions from NIPMO and the uncompetitive nature of the full-cost model, while another university considered the full-cost model a necessary approach to university–industry collaboration. According to the full-cost model, if a private entity or organisation covers the full cost (both direct and indirect) of research and development in collaboration with a public research organisation, the project is not considered to be publicly financed, and the provisions of the Act no longer apply.[28]

The Act and its Regulations have been criticised for their approach to IP protection, which may present obstacles to scholarly publishing.[29] Rapid publication of new research relating to potentially patentable inventions may be delayed to prevent compromising the novelty requirement for patentability; routine delays may affect scientific publication rates, making South African academics less competitive in open knowledge exchange.[27] The effects of the Act appear (as yet) to not have increased knowledge production in the innovation system. It may, however, have changed actor behaviour. As an example, university-affiliated clinicians (from UP/SBAH and WITS/CMJAH) who collaborate in the SM/SMI components, had their pre-2008 patents assigned to the SM/SMI conglomerate. These patents may not have been affected by initiatives undertaken by universities to promote patenting. The inventors who have gone through national university channels to patent, i.e. where the university was the assignee on the patent, include those from SUN, VUT, TUT and UCT. All these patents have been filed since 2008. These examples suggest that the full effects of the Act may yet become visible in the South African university environment, with university-affiliated inventors responding to incentives to patent, and more patents, which may have been assigned to industry collaborators in earlier times, being assigned to universities. While conflicts between patenting and scholarly publishing are anticipated, Patra and Muchie[11] showed that those national universities who are actively publishing, are also actively patenting. However, patenting was at a lesser degree, with South African universities accounting for 90% of scientific publications emanating from South Africa but for only 14% of South Africa's patent portfolio. In addition, inventors make up a small proportion of academics[30] and a large proportion of university research is not performed with commercialisation in mind[25,31].

Each of the national university actors has formalised structures for the protection of IP emanating from university research. In the UCT component, the UCT/GSH link is due to university-affiliated clinicians who have patented in their own capacity, even though the clinician was affiliated with both organisations. This presents an opportunity for university TTOs to identify inventions by university employees, create an entrepreneurial culture among university-affiliated clinicians (and the university at large), and encourage employees to disclose inventions to the university and pursue university IP processes. Owen-Smith and Powell[32] found that a crucial first step in the university environment is creating an entrepreneurial culture among academics and convincing academics to disclose their potentially valuable innovations to the TTOs. In South African universities, successful technology transfer efforts lie in proactive engagement by the TTOs with inventors[25] and effective and trustworthy relationships between TTO staff and inventors[33].

The number of clinicians from the private healthcare sector in the orthopaedic device innovation system who patent in their own capacity is noteworthy. Of the 14 national healthcare actors present in the network, 10 are from the private healthcare sector. The private healthcare sector inventors largely operate on their own, or with one other actor, and contribute to the large number of components in the network (SMC is the exception). These inventors are largely from the three big private healthcare providers in South Africa, i.e. Netcare Group, Life Healthcare and Mediclinic. Medical practitioners are not employed by the private healthcare groups in South Africa.[34] Yet opportunities may exist for private healthcare groups to facilitate and serve as partners in innovations by clinicians affiliated with the private healthcare group.

In the patent networks, 14 national industry actors (11 not previously identified in the scientific publication network) and five international industry actors (four not previously identified in the scientific publication network) were identified. These national industry actors are largely Small, Medium and Micro Enterprises; two of these are university spin-outs, both from UCT. Of the international industry actors, Tornier Inc. (TI), now merged with Wright Medical Group to form Wright Medical Group N.V.[35], is the only large corporation.

The sectorisation index shows that, of those actors who do co-invent, national intra-sectoral collaboration is absent. Intra-sectoral collaboration takes place between national and international actors in the patent network, but not within South Africa. The presence of intra-sectoral collaboration at an international scale, and not a national scale, suggests some benefit offered by international ties within the same sector over national ties. These could include access to foreign markets, access to different patient demographics, access to specialised infrastructure and resources, and access to patenting expertise. Patra and Muchie[11] found that approximately 20% of South African university patents were collaborative. Approximately 40% of these collaborative patents were with national science council actors, and approximately 17% were with international multinational companies. Joint patents with national industry actors were very low. Because knowledge is a localised phenomenon, knowledge exchange among the same set of co-located actors loses value over time as information and the recombination thereof fades[36]; collaboration with outside partners becomes important to prevent over-embeddedness[37]. As Breschi and Lenzi[37] caution, patent data capture only a subset of links relevant to knowledge exchange, although the network of collaborators is the most immediate and influential environment from which inventors draw ideas and information.

A total of 23 organisations appear in both the scientific publication[14] and the patent networks – 8 from the healthcare sector, 1 science council, 11 universities and 3 from the industry sector. Many, but not all, organisations that appear in both networks, have the author and inventor as the same person. This suggests some translation of scientific knowledge to commercial applications, which is demonstrated in the presence of some patent–paper pairs in the networks, i.e. the same idea described in different ways, resulting in a patent and journal paper.[12] The paper would usually describe experimental results, whereas the patent would define utility and claims of inventiveness. Examples of such pairs from the university sector include a paper and a patent of the TUT component on a mechatronic system for assisting an individual to attain a standing position. In the scientific network, some of these TUT component inventors co-author a series of papers on the biomechanics of knee and ankle trajectories. The inventors of a patent in this network by the SUN/SMC pair on the method for designing a knee prosthesis, are co-authors of several papers on the development and testing of patient-specific knee implants. There are several such examples, and it is evident that the orthopaedic device innovation network arises from the interaction between the science and technology spheres. Murray[12] showed that, for cartilage tissue engineering, a few key scientists produced both publications and patents, and that industry actors had limited participation in scientific publications. Interviews conducted for Murray's[12] study, highlighted overlap between scientific publication and patent networks, not captured in bibliometric data. Reasons for this overlap included the involvement of key scientists in technology development, forming start-up companies, consulting, mentoring and providing informal advice. Through these activities, scientists become active participants in both the scientific and the technical community. Nonetheless, our patent network analysis has identified actors, especially those in the industry sector, who were not identified in the publication network[14], including key contributors to the technological knowledge base.

Limitations

Of the 73 patents identified, 11 were omitted from the analysis as a result of incomplete inventor affiliation data. This means that many organisational actors may not have been identified. The second limitation of this study is that the organisational affiliations associated with each inventor relied on the methodology described. The multiple affiliations of the inventor captured at the priority date may not have been associated with work carried out at all of those organisations; this may result in the network presenting collaborative ties where there were none, or where there were no formal ties. In the absence of a more robust system of establishing affiliation from patent bibliographic data, the methodology presented here is a start to understanding these collaborative activities at an organisational level.

Patenting only reveals a subset of activity towards technological development.[10] Supportive investigations such as case studies would reveal other collaborative activity, might clarify organisational affiliations of actors, and could contribute to a more comprehensive picture of knowledge development and exchange toward technological development for orthopaedic device innovation.

Conclusion

The goal of any innovation system is to develop, apply and diffuse new technological knowledge. In this study, the organisations contributing to the technological knowledge base through patenting of orthopaedic devices in South Africa, and the sectors to which they belong, have been identified and the nature of the relations among them have been characterised. While patenting in the TIS has increased over time, knowledge exchange among actors in the network is limited.

The patent network complements the scientific publication network described by Salie et al.[14] Notable differences between the two networks is the small patenting contribution of national university actors, who were the main contributors of scientific knowledge production, as well as the significant contribution made by private healthcare actors patenting in isolation. The results presented in this study would be enhanced by further exploration of the knowledge functions to capture the factors that promote and hinder knowledge exchange in the TIS and the ties that are not expressed through patent co-inventorship. One such avenue for investigation may be the strength of inter-sectoral ties as expressed, for example, through licensing of university inventions to industry (both local and international). Another avenue to explore is the institutional support for healthcare actors, who have proven to be worthy contributors to both scientific and technological knowledge production in the TIS but are outliers in the innovation system.

Acknowledgements

This work was supported by the Community Engagement Programme of the South African National Research Foundation (NRF; grant no. 82624) and the South African Research Chairs Initiative of the Department of Science and Innovation and the NRF (grant no. 98788).

Competing interests

We declare that there are no competing interests.

Authors' contributions

All authors contributed to conceptualisation of the project. T.S.D. led and managed the project. F.S. and K.d.J. developed the methodology. F.S. collected and analysed the data and wrote the initial draft. All authors contributed to the writing of the final manuscript.

References

1. Deloitte. Research to guide the development of strategy for the medical devices sector of South Africa – Prepared in Partnership with the dti [document on the Internet]. c2014 [cited 2020 Feb 09]. Available from: http://www.samed.org.za/DynamicData/LibraryDownloads/60.pdf

2. Trade & Industrial Policy Strategies. The Johannesburg Health Cluster [document on the Internet]. c2018 [cited 2019 May 30]. Available from: http://www.tips.org.za/research-archive/trade-and-industry/item/3530-the-johannesburg-health-cluster

3. Andersen B. Paradigms and trajectories of technological opportunities 1890-1990. In: Moed H, Glanzel W, Schmoch U, editors. Handbook of quantitative science and techology research – The use of publication and patents statistics in studies of S&T systems. Dordrecht: Kluwer Academic Publishers; 2004. p. 133–162.

4. Park G, Park Y. On the measurement of patent stock as knowledge indicators. Technol Forecast Soc Change. 2006;73:793–812. https://doi.org/10.1016/j.techfore.2005.09.006

5. Organisation for Economic Co-operation and Development (OECD). Special issue on new science and technology indicators [document on the Internet]. c2001 [cited 2020 May 15]. Available from: https://www.oecd.org/sti/37124998.pdf

6. Lach S. Patents and productivity growth at the industry level: A first look. Econ Lett. 1995;49:101–108. https://doi.org/10.1016/0165-1765(94)00618-C

7. Fleming L, Marx M. Managing creativity in small worlds. Calif Manage Rev. 2006;48(4):6–27. https://doi.org/10.2307%2F41166358

8. Singh J. Collaborative networks as determinants of knowledge diffusion patterns. Manage Sci. 2005;51(5):756–770. https://doi.org/10.1287/mnsc.1040.0349

9. Balconi M, Breschi S, Lissoni F. Networks of inventors and the role of academia: An exploration of Italian patent data. Res Policy. 2004;33:127–145. https://doi.org/10.1016/S0048-7333(03)00108-2

10. Breschi S, Lissoni F. Knowledge networks from patent data: Methodological issues and research targets. In: Moed H, Glanzel W, Schmoch U, editors. Handbook of quantitative science and techology research – The use of publication and patents statistics in studies of S&T systems. Dordrecht: Kluwer Academic Publishers; 2004. p. 613–643.

11. Patra SK, Muchie M. Research and innovation in South African universities: From the triple helix's perspective. Scientometrics. 2018;116:51–76. https://doi.org/10.1007/s11192-018-2764-0

12. Murray F. Innovation as co-evolution of scientific and technological networks: Exploring tissue engineering. Res Policy. 2002;31:1389–1403. https://doi.org/10.1016/S0048-7333(02)00070-7

13. MacPherson A. The contribution of academic-industry interaction to product innovation: The case of New York State's medical devices sector. Pap Reg Sci. 2002;81:121–129. https://doi.org/10.1007/s101100100102

14. Salie F, De Jager K, Dreher C, Douglas T. The scientific base for orthopaedic device development in South Africa: Spatial and sectoral evolution of knowledge development. Scientometrics. 2019;119(1):31–54. https://doi.org/10.1007/s11192-019-03041-y

15. Hekkert MP, Suurs RAA, Negro SO, Kuhlmann S, Smits REHM. Functions of innovation systems: A new approach for analysing technological change. Technol Forecast Soc Change. 2007;74:413–432. https://doi:10.1016/j.techfore.2006.03.002

16. Tijssen R. Measuring and evaluating science-technology connections and interactions. In: Moed H, Glanzel W, Schmoch U, editors. Handbook of quantitative science and techology research – The use of publication and patents statistics in studies of S&T systems. Dordrecht: Kluwer Academic Publishers; 2004. p. 694–715.

17. Hinze S, Schmoch U. Opening the black box: Analytical approaches and their impact on the outcome of statistical patent analyses. In: Moed H, Glanzel W, Schmoch U, editors. Handbook of quantitative science and techology research – The use of publication and patents statistics in studies of S&T systems. Dordrecht: Kluwer Academic Publishers; 2004. p. 215–235.

18. Borgatti S, Everett M, Freeman L. Ucinet for Windows: Software for social network analysis. Harvard, MA: Analytic Technologies; 2002.

19. Borgatti S. Netdraw software for network visualisation. Lexington, KY: Analytic Technologies; 2002.

20. Eslami H, Ebadi A, Schiffauerova A. Effect of collaboration network structure on knowledge creation and technoogical performance: The case of biotechnology in Canada. Scientometrics. 2013;97(1):99–119. https://doi.org/10.1007/s11192-013-1069-6

21. Hanneman R, Riddle M. Introduction to social network methods. Riverside, CA: University of California; 2005.

22. Batool K, Niazi M. Towards a methodology for validation of centrality measures in complex networks. PLoS ONE. 2014;9(4):1–14. https://doi.org/10.1371/journal.pone.0090283

23. Binz C, Truffer B, Coenen L. Why space matters in technological innovation systems – Mapping global knowledge dynamics of membrane bioreactor technology. Res Policy. 2014;43:138–155. https://doi.org/10.1016/j.respol.2013.07.002

24. Krackhardt D, Stern R. Informal networks and organizational crises: An experimental simulation. Soc Psychol Q. 1988;51(2):123–140. https://doi.org/10.2307/2786835

25. Alessandrini M, Klose K, Pepper MS. University entrepreneurship in South Africa: Developments in technology transfer practices. Innovation. 2013;15(2):205–214. http://doi.org/10.5172/impp.2013.15.2.205

26. South African Department of Higher Education and Training. Ministerial statement on university funding: 2019/20 – 2020/2021 [document on the Internet]. c2018 [cited 2019 Jun 14]. Available from: http://www.dhet.gov.za/SiteAssets/18%2012%2007%20Ministerial%20Statement.pdf

27. Ncube C, Abrahams L, Akinsanmi T. Effects of the South African IP regime on generating value from publicly funded research: An exploratory study of two universities. In: De Beer J, Armstrong C, Oguamanam C, Schonwetter T, editors. Innovation and intellectual property: Collaborative dynamics in Africa. Cape Town: UCT Press; 2013. p. 282–315.

28. NIPMO. NIPMO interpretation note 13: Everything you need to know about full cost [document on the Internet]. c2019 [cited 2020 Oct 04]. Available from: https://nipmo.dst.gov.za/uploads/files/NIN13_Fullcost_6-Aug-19.pdf

29. Gray E. National environmental scan of South African scholarly publishing [document on the Internet]. c2009 [cited 2020 Oct 04]. Available from: https://open.uct.ac.za/bitstream/handle/11427/29095/Position_Paper_2_National_environmental_scan_of_So.pdf?sequence=1&isAllowed=y

30. Breschi S, Lissoni F, Montobbio F. University patenting and scientific productivity: A quantitative study of Italian academic inventors. Eur Man Rev. 2008;5:91–109. https://doi.org/10.1057/emr.2008.9

31. Simelane T. The innovation landscape of South Africa under new intellectual property management policy. Innovation Summit Journal. 2013:40-45.

32. Owen-Smith J, Powell W. To patent or not: Faculty decisions and institutional success at technology transfer. J Technol Transf. 2001;26:99–114. https://doi.org/10.1023/A:1007892413701

33. Sibanda M. Intellectual property, commercialisation and institutional arrangements at South African public research institutions. In: The economics of intellectual property in South Africa. Geneva: WIPO; 2009; p. 113–145.

34. Health Professions Council of South Africa. Policy document on business practices [document on the Internet]. c2016 [cited 2020 Feb 18]. Available from: https://www.hpcsa.co.za/Uploads/PSB_2019/Policy_on_Business_Practices_final%20-%202016.pdf

35. Haelio. Wright Medical and Tornier announce merger [webpage on the Internet]. c2014 [cited 2020 Oct 04]. Available from: https://www.healio.com/news/orthopedics/20141028/wright-medical-and-tornier-announce-merger

36. Boschma R, Frenken K. The spatial evolution of innovation networks. A proximity perspective. In: Boschma R, Martin R, editors. Handbook of evolutionary geography. Cheltenham: Edward Elgar Publishing Ltd; 2010. p. 120–135.

37. Breschi S, Lenzi C. The role of external linkages and gatekeepers for the renewal and expansion of US cities' knowledge base, 1990-2004. Reg Stud. 2015;49(5):782–797. http://dx.doi.org/10.1080/00343404.2014.954534

Appendix 1: Search term used to extract patent data from LexisNexis TotalPatent

((biomechanical OR bone OR joint OR muscle OR tendon OR ligament OR muscul* OR skelet*) AND (replacement OR arthroplast* OR device OR tool OR instrument OR apparatus OR implement OR implant OR prosthe* OR orthotic OR orthoses OR machine OR appliance OR software OR material OR design* OR develop* OR concept*) AND (Inventor-Res(South Africa) OR Inventor-Nat(South Africa) OR Assignee(South Africa) OR APC(South Africa) OR Applicant-Nat(South Africa) OR Applicant-Res(South Africa))) and DATE(>2000-01-01)

Appendix 2: Actors of the collaboration networks

Abbreviation	Full name	Location
	Healthcare sector	
CMJAH	Charlotte Maxeke Johannesburg Academic Hospital	National
GUH	Ghent University Hospital	International
GSH	Groote Schuur Hospital	National
JH	Jakaranda Hospital	National
KPOU	Klinik und Poliklinik fur Orthopadie und Unfallchirurgie	International
LH	Livingstone Hospital	National
LWH	Life Wilgeleugen Hospital	National
LOC	Lyon Ortho Clinic	International
MMC	Morningside MediClinic	National
OSC	Ortho One Sports Clinic	International
OSV	OrthoSport Victoria	International

Appendix 2 continued

Abbreviation	Full name	Location
	Healthcare sector	
ParklandsH	Parklands Hospital	National
PH	Pinehaven Hospital	National
QE2HSC	Queen Elizabeth 2nd Health Sciences Centre	International
SBAH	Steve Biko Academic Hospital	National
SH	Sunshine Hospital	International
SSOC	Sports Science Orthopaedic Clinic	National
SMC	Stellenbosch MediClinic	National
Unitas	Unitas Hospital	National
VPH	Vincent Palotti Hospital	National
ZAH	Zuid Afrikaans Hospital	National
	Industry sector	
ATTRI	ATTRI	National
BBOP	Bradley Beckerleg Orthotic & Prosthetics	National
CMO	Custom Med Orthopaedics (Pty) Ltd	National
DenMyd	Denmyd Medical Equipment	National
DesDall	Desmond Dall	National
D4SBV	Design4Spine BV	International
ESS	Elite Surgical Supplies	National
ESUSA	Elite Surgical USA	International
KVP	Kearny Venture Partners	International
OrthoSol	Ortho-Sol Pty Ltd	National
PAB	Pressure Air Biofeedback CC	National
reSCRIBE	reSCRIBE	National
Saspine	Saspine	National
SmartCrutch	smartCRUTCH	National
SI	Southern Implants (Pty) Ltd	National
SM	Southern Medical (Pty) Ltd	National
SMI	Spinal Motion Inc	International
TI	Tornier Inc	International
Wismed	Wismed	National
	Science council sector	
CSIR	**Council for Scientific and Industrial Research**	**National**
	University sector	
GU	Ghent University	International
LTU	La Trobe University	International
LU	Loughborough University	International
NWU	North-West University	National
PSTDV	Pole Scientifique Et Technologique de Volzy	International
SUN	Stellenbosch University	National
TUT	Tshwane University of Technology	National
UVSQ	Universite de Versailles Saint-Quentin-en-Evelines	International
UB	University of Bath	International
UCT	University of Cape Town	National
UM	University of Manchester	International
UMIN	University of Minnesota	International
UP	University of Pretoria	National
UPEC	Universite Paris East Central	International
WITS	University of Witwatersrand	National
VUT	Vaal University of Technology	National

Insights from training a blind student in biological sciences

AUTHORS:
Rethabile Tekane[1]*
Marietjie Potgieter[1]

AFFILIATION:
[1]Department of Chemistry, University of Pretoria, Pretoria, South Africa
*Currently: ENGAGE Programme: Engineering, Built Environment and Information Technology, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Marietjie Potgieter

EMAIL:
marietjie.potgieter@up.ac.za

DATES:
Received: 07 July 2020
Revised: 09 Oct. 2020
Accepted: 07 Dec. 2020
Published: 28 May 2021

HOW TO CITE:
Tekane R, Potgieter M. Insights from training a blind student in biological sciences. S Afr J Sci. 2021;117(5/6), Art. #8607. https://doi.org/10.17159/sajs.2021/8607

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☒ All data included
☐ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Jenni Case

KEYWORDS:
blind students in science, accessibility of science, curricular adaptation, tertiary science education, STEM

FUNDING:
South African Department of Higher Education and Training

Higher education institutions have a constitutional obligation to provide reasonable accommodation to students with disabilities. Although the teaching and learning of students with blindness and low vision in STEM disciplines are well documented abroad, to date, there are no published studies in South Africa on successful teaching and learning strategies for students with blindness and low vision in STEM fields, specifically in science disciplines. Therefore, in this paper, we report on how teaching, learning, and assessment were adapted to make science disciplines accessible to John, a blind student enrolled in a biological sciences degree at a research-intensive university in South Africa. Several factors contributed towards the successful completion of John's bachelor's degree. These factors include the availability of tutors who committed a large amount of time to help John understand content presented in lectures, tutorials, and practical sessions; a well-resourced and effective Disability Unit; lecturers who ensured that John was well accommodated in lectures, tutorials, and practical sessions; and, finally, John's commitment and dedication towards learning.

Significance:

- This is the first study to report on successful teaching and learning strategies for a blind student in the natural sciences in the South African context.
- The study provides a guide that scholars, educators, university managers and policymakers can use to ensure that mathematics and science subjects are accessible to blind students and that teaching strategies allow them to perform to their potential.

Introduction

Research has shown that there is an underrepresentation of students with blindness and low vision (BLV) in STEM fields in many countries.[1-3] The few students with BLV who enrol in STEM disciplines are usually frustrated and lose interest in science-related pursuits mainly because they always have to depend on their sighted peers to conduct laboratory activities, report observations, and interpret or understand visual material.[4] Furthermore, students with BLV tend to perform poorly in comparison to their sighted peers.[2] The latter may be because science teachers are not trained to teach students with BLV[5], and STEM subjects mainly depend on visual representations to explain complex concepts or processes[2]. The teaching and learning of students with BLV in science disciplines have been well documented abroad.[6-11] However, to date, there are no published studies in the South African context related to the enrolment of BLV students, specifically in science disciplines, and successful strategies for teaching students with BLV in this domain. The challenge of accommodating students with BLV in science disciplines is compounded where resources are constrained and funds for specialised support are limited. Therefore, in this paper, we report on general support and discipline-specific adjustments that were made to teaching, learning facilitation and assessment to make STEM disciplines accessible to a blind student to ensure that he could perform to his full potential. We expect that our experience will make a significant contribution to current knowledge about the training of blind students in STEM disciplines in South African higher education institutions.

Literature review

The new Constitution of post-apartheid South Africa entrenches equal rights and freedom from discrimination of any kind for all its citizens as a foundational principle for building a new society. The Constitution stipulates that 'Every person shall have the right to basic education and equal access to educational institutions'[12]. Several policy documents were developed by the South African Department of Education after the dawn of democracy to facilitate access and participation of disabled students at all levels of the education system. These policies include the *Education White Paper on The Transformation of the Higher Education System*[13], the *National Plan for Higher Education*[14] and the *Education White Paper 6: Special needs education: Building an inclusive education and training system*[15]. These policy documents subscribe to the social model of disability that sees the problem as located not in the individual, but in the system or culture that fails to meet the needs of these individuals.[16] This stance countered the prevailing philosophy which reduced disability to a medically defined impairment that required the disabled person to adapt to fit into the system. South Africa was one of the early signatories of the influential United Nations Convention on the Rights of Persons with Disabilities and its Optional Protocol in 2007, and is thereby obligated to ensure 'the development by persons with disabilities of their personality, talents and creativity, as well as their mental and physical abilities, to their fullest potential', and that they are able to 'access general tertiary education … without discrimination and on an equal basis with others'.[17] This means that the provision of high quality and equal education for persons with disabilities is not only a moral concern, it is also a constitutional obligation.[18]

Despite the progressive nature of these policies and their good intentions, the literature abounds with studies indicating that the goals of a truly inclusive education system have not been realised.[18-23] Research has shown that not many people with disabilities consider enrolling at institutions of higher education and the few that do enrol face challenges in terms of physical and curricular access, inadequate support, negative attitudes and crippling perceptions.[21,24] Access to tertiary institutions is still available only to the few students with BLV who were fortunate to have attended one of a handful of schools that were able to prepare them for tertiary studies and even in these schools they had limited subject choices.[18,23] The majority of schools for learners with special needs are so poorly resourced that the South African government is implicated as being 'complicit in exclusion', denying these learners their basic human rights in terms of quality education and equal opportunities. Furthermore, Donohue and Bornman[19] estimated that up to 70% of children of school-going age with disabilities do not attend school at all. These findings highlight the urgency of ensuring that, at the very least, the learning environment in tertiary institutions in South Africa meets the needs of students with disabilities. The majority of studies pertaining to disabled students in tertiary institutions concentrate on students' experiences, as will be discussed below.[24-30] There is therefore a need for education management information to assist decision-makers when it comes to accommodating students with BLV, especially in the natural sciences.

Students with disabilities require special convenient access to buildings and facilities on university campuses to make learning possible. However, research has shown that access continues to be a major problem that limits the students' mobility, hinders their learning and may even endanger their lives. In their study, Phukubje and Ngoepe[24] evaluated the accessibility of library services for disabled students at the University of Limpopo. Results revealed that only one librarian was assigned to manage library services for the disabled; as a result, the librarian was not able to individually train each student on how to search the catalogues and shelves for books, how to reference, and how to photocopy and use the printer. Similarly, Ntombela and Soobrayen[25] learnt that most of the visually impaired students at the University of KwaZulu-Natal's Edgewood campus, did not receive mobility training due to understaffing at the Disability Unit (DU). In a study conducted by Engelbrecht and de Beer[26], disabled students complained about architectural constraints which hindered their mobility and access to services. Such constraints included steep ramps which sometimes had potholes, and heavy building doors which the disabled students could not open without the assistance of other students. In another study, Losinsky and colleagues[27] reported that wheelchair-bound students were unable to get to class on time because of the short break between classes.

Scholars have also evaluated the various support mechanisms provided to students with disabilities by the specialised student support units at tertiary institutions. DUs are meant to provide both logistical and academic support services.[28,29] Logistical support services include assisting the students with campus challenges and communicating students' needs to their lecturers. Academic-related services include providing Braille and tape-recorded readings, sign language interpreters, alternative assessments, and assistive technology such as Job Access With Speech (JAWS) software. Students viewed DUs as an irreplaceable source of academic information; they appreciated their contribution to orientate them in their new environment, and to make them feel welcome, comfortable and part of the university.[29]

According to Mutanga[20], lecturers' support of students with disabilities is crucial for the students' academic achievement. However, students have reported mixed experiences of lecturers' support. Some lecturers were amenable to curriculum flexibility, provided alternative styles of teaching and assessment, and responded favourably to requests from either the students or the DUs.[28,30] However, other students experienced an indifference from lecturers[31] and an unwillingness to adjust teaching methods to accommodate disabled students because lecturers did not consider disability support as their responsibility.[28] Students pointed out that some lecturers lacked understanding of disability support needs; however, good communication often resolved the issues at hand.[30]

Assistive technology, such as JAWS, is an important support mechanism as it enhances access to learning, specifically for students with BLV. JAWS is a computer screen reader programme that delivers text-to-speech output or a refreshable Braille display. Although many students indicated that assistive technology was beneficial for their learning[30], JAWS can also restrict the learning of students because it is unable to read graphical material, mathematical and scientific symbols[32], and it is not compatible with African languages such as Zulu[25].

Apart from these studies on the needs and experiences of students with BLV at South African tertiary institutions, there are no reports specifically related to the involvement of students with BLV in a science faculty. There are international studies, however, that report on strategies for teaching students with BLV in science disciplines. A brief summary of the findings of international studies is presented next.

Various adaptable low-cost audible instruments and tactile tools have been developed to increase accessibility to experiments[8,9,33] and promote independence[34] in laboratories, and to enable BLV students to visualise organic chemistry[35]. Examples of such instruments and tools include talking calculators, thermometers and balances[9], colour identifiers, and handheld submersible audible light sensors. Tactile molecular models have been used to aid students to visualise mechanisms and predict products of chemical reactions. Adaptive teaching aids such as magnetic boards, letters and numbers have been used in teaching to assist students to write and balance chemical reactions, and draw Lewis dot structures.[11] Harshman et al.[7] pointed out that chemistry instructors ought to experiment with different teaching strategies as it is not always easy to know which strategy will work best for teaching students with BLV. Guidelines have been developed for physics instructors to adapt class sessions, curricular materials, tutorials and demonstrations to make physics accessible to students with BLV.[36-39] Similarly, in mathematics, effective methods for delivering instruction to students with BLV have been published.[38,39] There are several literature reports on hands-on summer enrichment programmes[33,40] to promote BLV students' interest in STEM education. Lastly, research pertaining to teachers' experiences of teaching science to students with BLV has revealed learning traits portrayed by these students – for example, that students with BLV learnt more effectively when working collaboratively with sighted peers.[41] As science is highly visual, teachers found it challenging to teach science because students with BLV could not visualise abstract concepts without the use of tactile teaching aids. Other scholars reported a lack of confidence of teachers to effectively teach students with disabilities[42]; hence there was need for continuous professional development in 'effective strategies for teaching students with disabilities'[5].

Tertiary institutions are obligated to achieve inclusivity and equity in terms of access to education; however, they need to draw on experiences within the fraternity to guide them regarding the special needs of students with disabilities. This study seeks to address the lack of information on the accommodation of disabled students in science faculties in South Africa. We report on the curricular adjustments that were made to support the teaching and learning of a blind student, called John (pseudonym), who studied biological sciences at our university. John had limited vision at birth, but was functionally blind from the age of about 12.

Background

John passed his National Senior Certificate and achieved seven distinctions with an average of 84% for the final examination. John's Grade 12 subjects were English, Afrikaans, Mathematics, Physical Sciences, Life Sciences, Computer Applications Technology, and Life Orientation. John applied for admission to the Faculty of Natural and Agricultural Sciences at the University of Pretoria with the objective to major in biological sciences. The Faculty had experience in training a blind student to the level of PhD in statistics, but not in biology. The decision to admit him to biological sciences was not taken lightly. Discussions were held between John, his parents and the Faculty to match interest with training demands of specific disciplines and future career possibilities. The latter was supported by John's psychometric test results which confirmed his cognitive ability and his interest and suitability to pursue a scientific investigative career. Special consideration was given to safety (it would not be possible to offer chemistry beyond the first year), the inexperience of the student and staff to deal with the challenges of the situation (his progress would be assessed after the first academic year) and future career prospects for a blind graduate in biology. The DU agreed to provide the following services: to support academic staff in meeting the needs of John; to monitor the situation and alert the Faculty if problems arose; to meet with lecturers before the start of the semester to advise and assist in preparation; to take responsibility for the conversion of study materials, test and exam papers into a format accessible by John; to provide mobility training and supervision during formal assessments; to provide office space for John's tutoring and self-study; and to carry most of the costs of tutor support and of the conversion of study materials to accessible formats. The Faculty agreed to admit John to his degree programme of choice with the proviso that the first academic year would be spread over two calendar years to allow for adjustment, refinement of procedures and exposure to a wide range of disciplines. John was offered a choice of two curriculum packages; both were enriched in the first academic year with mathematics, computer science and/or informatics, to cater for future specialisation in bioinformatics or biological mathematics. This would give John the opportunity to proceed with a mathematics-intensive programme rather than biological sciences in the second academic year if he wanted to do so. In the interest of safety, the practical components of first-year chemistry modules would be replaced by assignments, but John would not be able to enrol for any higher level courses in chemistry. Before admission was formalised, the Faculty also sought written commitment from all discipline departments that would be involved in his training to ensure that they accepted the responsibility for the provision of an enabling environment that would be conducive to his success.

In his first academic year, spread over two calendar years, John completed one semester course of each of the following disciplines: mathematics, physics for biology, biometry, a foundational course in molecular and cell biology, microbiology, genetics, botany and zoology. He also completed two semesters of chemistry and achieved distinctions in all but one first-year course. John decided to spread his second academic year over two calendar years as well; his curriculum included two semesters of microbiology, biochemistry, human physiology and genetics, all passed with distinction. He added two first-year courses in Sepedi but dropped second-year mathematics and biometry courses because of the logistical challenges posed by those disciplines and his lack of interest in them. He completed his third year in one calendar year and graduated *cum laude* in BSc Human Genetics. He also graduated *cum laude* in BScHons in Bioinformatics at the end of the following year and subsequently enrolled for an MSc in Bioinformatics.

Adjustments made to learning support, teaching, lab training and assessment

In this section we report on general support as well as discipline-specific adjustments that were made to accommodate John. We collected data from several sources, namely the DU, library, lecturers and tutors, and from the student himself. Three interviews were conducted with John at different stages during his undergraduate and postgraduate studies, and questionnaires were sent to all his lecturers and tutors to request information on their approaches and experiences ($N = 61$; response rate 53%). Data collected from the interviews and questionnaires were analysed for emerging themes and for information specific to each discipline. The findings were triangulated with information obtained from the DU, the library and the student. The trustworthiness of the findings was confirmed through member checking. Neither of the authors of this study was involved in any way in John's training. At the time of the study, the first author was a postdoctoral researcher and the second author was responsible for education management in the faculty. Ethics approval was granted by the NAS Research Ethics Committee at the University of Pretoria (NAS121/2019). Informed consent was obtained from the participants.

Table 1 provides an overview of findings that were specific or unique to each discipline. It is followed by a discussion of general themes that emerged from the feedback received.

Specialised learning support

The DU provided logistical support such as mobility training, shared office space where the student could work or be tutored, and assistance to address any campus-related challenges the student might have experienced. The DU also assisted with converting course materials to either Braille or a format that is compatible with JAWS. The choice of format depended on the type of course materials submitted by the lecturers. If the course content was mainly text, it was converted to a format that is compatible with JAWS. However, if the content was mainly visual, it was converted to Braille in the form of tactile pictures and sketches. The DU had one printer for printing of text in Braille and acquired a second printer for converting pictures to tactile representations. At the beginning of each semester, the DU held meetings with course coordinators to inform them about the support services offered, that is, what the DU could and could not do; to alert lecturers about the potential need for adaptation of the curriculum; and to plan for upcoming challenges. John completed the majority of his formal assessments electronically, for which the DU provided the venue and the necessary equipment. Because John's tutors were allowed to be present during tests and exams to explain visual material, the DU arranged for invigilation to make sure that test and exam conditions were adhered to. The DU also provided funding for the appointment of personal tutors who assisted John with the learning of content, especially visual material.

The pivotal role of dedicated tutors

Personal tutors were appointed for all the courses in which John was enrolled. At the start of a course, John met with lecturers and requested a personal tutor. The tutors were usually postgraduate students who were doing their MSc or PhD and were thus knowledgeable in the subject areas. The tutors acted as the interface between the student and the lecturer – an arrangement which relieved the pressure on both the lecturer and the student. Each tutor met with John on a weekly basis to go through the lecture material and explain visuals that were discussed in lectures. They typically sat next to John during class tutorials and assisted him by describing any visuals included in tutorial questions. They also assisted during summative assessments and often served as invigilators as well.

Lectures

John attended normal lecture sessions and made sound recordings of the class for use later if required. He sat in the front row where the lecturers could see him. This alerted the lecturers to his presence and often prompted them to make a special effort to speak more slowly and clearly to ensure that John followed what was being taught. They provided detailed verbal descriptions of visual content to help John form mental models of what was being discussed. In cases where lecturers used clickers to record class attendance, John would simply press any button on the response device. However, if clicker exercises counted for marks, another arrangement was made beforehand: either the lecturer or the tutor sitting next to him would read the questions and the answer options so that he could respond appropriately. Mathematics posed the biggest challenge because the lecturer typically worked on the board solving problems which John could not follow. This meant that he relied heavily on his mathematics tutor to provide detailed explanations of the content after class.

Laboratory training

Discipline departments adopted a range of approaches depending on the nature of practical training, the extent of practical work included in the curriculum, and safety considerations. In courses such as chemistry, the student was not allowed in the lab due to safety considerations. The practical sessions were replaced with a research assignment on related content. In courses such as microbiology and physiology, John worked with either a tutor or a lab technician who performed and explained the tasks, and made observations and measurements which John processed for his experimental report. Similarly, in biochemistry, the tutor familiarised John before the practical session with the equipment and the experiments that would be performed. He obtained real experimental data from another student to incorporate in his lab report. In courses like physics for which practicals were undertaken in groups, John was included in a group with sighted students. This arrangement was not beneficial for his learning. Fellow students could not assist him because of their inexperience and time constraints for the task to be completed. In bioinformatics, the lecturer set up a shared terminal from his teaching computer to the student's laptop, which allowed John to 'read' everything the lecturer was typing during online tutorials. Finally, in botany, the lecturer explained the structural features of plants to him personally to ensure that he understood the concepts.

Table 1: Discipline-specific arrangements made to accommodate blind student John

Courses	Curriculum component	Number of lecturers (tutors) who participated*	Discipline-specific arrangements/challenges
Chemistry	First year, both semesters	1 (–)	No training in laboratory. Practicals were replaced by assignments on the chemical industry. The second semester course consisted of organic chemistry and physical chemistry. The building of models in organic chemistry consumed a lot of time: therefore, John wrote the exam sections on separate days.
Physics	First year, Semester 1	1 (1)	Lab sessions: Group work with sighted students did not support his learning.
Mathematics	First year, both semesters	– (1)	Assessments were provided in LaTeX format. John needed the LaTeX source for the textbook to be able to 'read' mathematical expressions and equations.
Molecular and Cell Biology (first course in Biology)	First year, Semester 1	3 (1)	Key visuals from the textbook were identified and submitted in advance for conversion to tactile form. Practical sessions: John was supported by the Disability Unit assistant to make the measurements and observations and to fill out the lab report. Lecturers spoke more slowly, provided full details of visuals displayed on lecture slides, and read aloud clicker questions and their answer options.
Zoology	First year, Semester 2	1 (–)	Lecture notes were sent to the Disability Unit for translation to Braille. The library was unable to obtain the electronic textbook. Practical sessions: John attended more than one of several repeat sessions. He listened to the presentations, but could not study the organisms by viewing them through a microscope or looking at preserved specimens.
Statistics	First year, Semester 2 Honours module (3 weeks)	1 1	Lecture notes and assessments were provided in LaTeX format. Practicals involved coding in statistical language and interpreting the results. Honours: Practical training involved coding with *R* software; however, without the use of graphical presentations. The lecturer taught him how to use summary statistics to deduce specific features normally presented in graphs.
Plant Science	First year, Semester 2	1 (–)	Practical sessions were used to demonstrate differences in the morphology and anatomy of plants: the lecturer taught John personally to make sure that he understood the concepts.
Computer Science	First year, both semesters	2 (–)	Lecturer made an effort to provide extensive verbal descriptions of visual representations in class. A dedicated teaching assistant was appointed to assist John and assess his practical assignments.
Microbiology	First year, Semester 2	1 (–)	Tutor built small models to enable touch, for example to demonstrate the shape of microbes growing on agar plates. A demonstrator performed the experiment on the student's behalf and explained the visual results.
Genetics (major)	First year Second year Third year	2 (1) 1 (2) 2 (1, same tutor for all third-year modules)	First year: Tutor built models with clay and glitter glue for tactile learning e.g. of the chromosome. Tutors learnt through trial and error how to describe visual content so that the student could understand.
Biochemistry	Second year	3 (–)	Practicals: John visited the lab beforehand with the tutor, received experimental results from other students and prepared his own lab reports afterwards.
Human Physiology (major)	Second year Third year	1 (1) 1 (1, same tutor for all third-year modules)	3D prints were made to support learning; however, time demand and cost limited the use of this option. Tutor carefully explained tables, figures and diagrams during one-on-one sessions. For practicals that required physical activities, an alternative essay form of testing was used to test the same content.
Bioinformatics	Honours level	1 (1)	All practicals were computer based. The student shared a terminal with the lecturer so that he could access everything directly from his laptop.

**Number in brackets represents the number of tutors who participated in the study*

Tests and exams

John completed all assessments on his laptop; therefore, assignments, test and exam papers were sent to the DU beforehand so that they could check if the papers were compatible with JAWS. John was allowed the standard amount of extra time granted to students with special needs, except for mathematics for which the time was uncapped due to the tedious process associated with mathematics (see below). Other exceptions were chemistry, physics, and population genetics, because mathematical calculations are used extensively in these modules, and biochemistry, where John had to provide a description of complex chemical structures instead of chemical drawings. All official tests and exams were written at the DU and were invigilated by the tutors who would assist John by either explaining any visuals included or would, under John's directions, draw the required visuals. Official UP invigilators were always present.

The special case of mathematics and statistics

Mathematics and statistics present unique challenges to BLV students because they are abstract, information dense and rich in symbolic expressions with a strict code for presentation. In general, mathematics lecturers followed the traditional style of developing theorems and demonstrating problem solving on the board or on a device that projected on the board, which meant that a blind student could not follow the teaching. Also, the JAWS software program cannot 'read' mathematical expressions in electronic textbooks. The way around this conundrum is for the student to obtain the LaTeX source of the textbooks in order to decipher the mathematical expressions and equations. Similarly, lecturers had to set their test and exam papers in LaTeX format to make it accessible. LaTeX is a high-quality typesetting system that allows for the creation of technical and scientific documentation with precise control over layout and formatting. Assessment in mathematics and statistics was a cumbersome process because John had to read the question in LaTeX format, type the problem on his Braille machine, work it out and then type his answer with its stepwise development in LaTeX for the lecturer to read and evaluate. This, according to John, was tedious and time consuming.

Textbooks

Librarians assisted John to obtain electronic copies of the various textbooks that he needed for his courses. In order to comply with copyright requirements, John had to buy the textbook and present proof of payment to a designated librarian. The librarian included the proof of purchase in their application for an electronic copy of the textbook. Upon receipt of the electronic copy from the publisher, the librarian would copy it to a CD in a format compatible with JAWS. This electronic copy is not the same as an eBook; it is a PDF version of the book which is not available on a virtual platform, as are eBooks. In cases in which the publisher could not be located or electronic copies of the textbooks were not available, the DU scanned sections of the textbooks to an electronic text format (MS Word or PDF) using Optical Character Recognition software (ABBYY FineReader). It has since become much easier to acquire electronic textbooks than at the time when John was an undergraduate student, because many electronic textbooks are now freely available for purchase.

Student reflection on his experience

In general, John was satisfied with the assistance that he received from the Faculty, DU, lecturers and tutors. He graciously acknowledged the efforts of everyone who tried to assist him. In hindsight, as a master's student, he realised the need to include more statistics in his undergraduate curriculum. The BSc Honours programme included a short course offered over 3 weeks on basic statistical knowledge for research in biological sciences, but the format was not conducive to learning for a blind student. The intense block-week presentation did not allow enough time for John to immerse himself in the dense mathematical notations and concepts of the discipline.

John provided the following advice to blind students planning to study science:

- Be prepared: Mobility training provided by the DU at the beginning of each semester is important as it teaches you to find your way around campus and to get to lecture venues. It is also essential to contact the lecturer and make arrangements well in advance, because some processes, such as getting a personal tutor and obtaining textbooks, may take longer than expected.
- Attend classes and make voice recordings for later use, if necessary.
- Motivation is key to being successful in a science discipline: it is important to study what you are interested in and what you will enjoy, otherwise the effort required will be too much.

Discussion and conclusions

More than 7 years have passed since we embarked on a journey to train a blind student in biology. The apprehensive start stands in stark contrast to the celebrations when John graduated *cum laude* 2 years in succession with his first and second degrees. As we reflect on the journey, several critical issues stand out in terms of the demands of the science domain and the team effort required for a blind student to navigate it successfully.

STEM subjects are highly visual[1], which presented lecturers and tutors with a significant challenge to determine how John could be supported to build mental models of sub/micro- or molecular level phenomena, despite a severe handicap on observation (macro-level) and restricted access to the symbolic level of representation. To overcome this handicap, lecturers and tutors had to invest more effort to substitute sensory input from sight with sensory input from touch and sound. The tutors provided verbal explanations of visual representations, built models using everyday materials or molecular model kits, and used tactile artefacts produced by the DU. Mental images are believed to be 'constructed from different sources of sensory information, including sound and touch, that interact with the brain's network of spatial subsystems and visual areas'[1]. These substitutions were clearly beneficial for John's learning, as evidenced by his excellent academic performance.

The successful training of a blind student in science requires a team of dedicated individuals in which the student is the lead player. The demands of such a pursuit require that student interest, cognitive ability, emotional make-up and personality are perfectly aligned. In our case, John's motivation, work ethic and persistence were paramount to his success. Thus, before the admission of a blind student to a science faculty, it is essential to seek professional advice on whether the student's psychometric profile matches their study and career choice and then to engage with the prospective student in an advisory capacity to ensure that they have a reasonable chance of success in their chosen field of study. Secondly, our findings confirm that tutors played a major role in this success story. Curricular demand and large class sizes at the undergraduate level rule out the active involvement of the lecturer in the provision of support to a blind student. This necessitates delegation of the task to dedicated tutors. Tutors spent a large amount of time helping John to understand subject content presented in lectures, tutorials and practical sessions, and visuals included in assessments. Tutors advocated for the needs of the student which informed the lecturers of appropriate special arrangements and relieved the student of the responsibility and discomfort of having to repeatedly explain themselves.[22] Tutoring a blind student is a specialist assignment – one that sighted peers cannot reasonably be expected to do. It requires someone with a solid knowledge of the discipline and an empathic nature. Another essential role player is an effective and well-resourced DU without which the training of a blind student would not be possible. In our case, the DU supported the project through generous tutor funding; guidance for the lecturers and tutors; essential resources such as Braille, 3D printing and JAWS; and academic and non-academic support for John. The Unit provided space for tutoring and study and ensured the integrity of assessments conducted on its premises. Lastly, John was also supported by lecturers who sought to accommodate his needs as far as possible in lectures, tutorials, and practical sessions. All these role players, with the exception of the DU, acted out of goodwill and were largely unaware of John's rights and their legal obligation to provide reasonable accommodation of his needs. While their goodwill is commended, it is not sustainable or scalable. Our experience highlights the need for raising disability awareness in the sector if progress is to be made to improve access and success of disabled students.

John is nearing completion of his master's degree, which testifies to the quality of his undergraduate and honours training. Our 'experiment' has demonstrated quite convincingly that the high demands of STEM disciplines do not render them inaccessible to blind students. However, we do acknowledge that John was an exceptional student and his success cannot be interpreted as evidence that our measures to accommodate him would be sufficient to ensure the success of other students with BLV. Being exceptionally gifted, both mentally and emotionally, should not be a prerequisite for a blind person to succeed in STEM education. The South African education system must urgently address the broad pattern of social exclusion of BLV students that has been prevalent until now. This success story represents a small step towards the goal of greater equity in STEM education.

Acknowledgements

We are grateful for the critical feedback provided by the Education Consultant of the Faculty, Dr Ina Louw. We thank Juan Erwee from the Disability Unit for generous support for the student and the researchers, and the student for his willingness to share his learning experiences with the researchers. We are also grateful for the participation of the lecturers, tutors, lab managers, and librarians. This research was partially funded from a DHET grant for Scholarship of Teaching and Learning.

Competing interests

We declare that there are no competing interests.

Authors' contributions

M.P. was responsible for the conceptualisation of the project, validation of the findings, writing of the manuscript, project leadership, and funding acquisition. R.T. was responsible for development of the methodology, data collection, data analysis, writing of the manuscript, and project management.

References

1. Rule CR, Stefanich GP, Boody RM, Peiffer B. Impact of adaptive materials on teachers and their students with visual impairments in secondary science and mathematics classes. Int J Sci Educ. 2011;33(6):865–887. https://doi.org/10.1080/09500693.2010.506619

2. Rosenblum LP, Herzberg TS. Braille and tactile graphics: Youths with visual impairments share their experiences. J Vis Impair Blind. 2015;109(3):173–184. https://doi.org/10.1177/0145482X1510900302

3. Lewis AL. Bodner GM. Chemical reactions: What understanding do students with blindness develop? Chem Educ Res Pract. 2013;14:625–636. https://doi.org/10.1039/C3RP00109A

4. Supalo CA, Humphrey JR, Malouk TE, Wohlers D, Carlsen WS. Examining the use of adaptive technologies to increase the hands-on participation of students with blindness or low vision in secondary-school chemistry and physics. Chem Educ Res Pract. 2016;17:1174–1189. https://doi.org/10.1039/C6RP00141F

5. Irving MM, Nti M, Johnson W. Meeting the needs of the special learner in science. Int J Spec Educ. 2007;22(30):109–118.

6. McDonald C, Rodrigues S. Sighted and visually impaired students' perspectives of illustrations, diagrams and drawings in school science. Wellcome Open Res 2016;1(8):1–14. https://doi.org/10.12688/wellcomeopenres.9968.1

7. Harshman J, Lowery-Bretz S, Yezierski Y. Seeing Chemistry through the eyes of the blind: a case study examining multiple gas law representations. J Chem Educ. 2013;90:710–716. https://doi.org/10.1021/ed3005903

8. Supalo CA. Teaching chemistry and other sciences to blind and low-vision students through hands-on learning experiences in high school science laboratories [PhD dissertation]. Pennsylvania State University; 2010.

9. Supalo CA, Mallouk TE, Rankel L, Amorosi C, Graybill CM. Low-cost laboratory adaptations for precollege students who are blind or visually impaired. J Chem Educ. 2008;85(2):243–247. https://doi.org/10.1021/ed085p243

10. Pence LE, Workman HJ, Riecke P. Effective laboratory experiences for students with disabilities: The role of a student laboratory assistant. J Chem Educ. 2003;80(3):295–298. https://doi.org/10.1021/ed080p295

11. Boyd-Kimball D. Adaptive instructional aids for teaching a blind student in a nonmajors college chemistry course. J Chem Educ. 2012;89(11):1395–1399. https://doi.org/10.1021/ed1000153

12. Republic of South Africa. The Constitution of the Republic of South Africa. Pretoria: Government Printing Works; 1996. Available from: https://www.gov.za/documents/constitution-republic-south-africa-1996

13. South African Department of Education (DoE). Programme for the transformation on Higher Education: Education White Paper 3 [webpage on the Internet]. No date [cited 2020 Oct 05]. Available from: https://www.gov.za/documents/programme-transformation-higher-education-education-white-paper-3-0

14. Asmal K. National plan for higher education: Ministry of Education [document on the Internet]. c2001 [cited 2020 Oct 05]. Available from: http://www.ru.ac.za/media/rhodesuniversity/content/institutionalplanning/documents/National_Plan_for_Higher_Education_in_South_Africa_2001.pdf

15. South African Department of Education (DoE). Education White Paper 6: Special Needs Education [webpage on the Internet]. c2001 [cited 2020 Oct 05]. Available from: https://wcedonline.westerncape.gov.za/Specialised-ed/documents/WP6.pdf

16. McEwan C, Butler R. Disability and development: Different models, different places. Geogr Compass. 2007;1(3):448–466. https://doi.org/10.1111/j.1749-8198.2007.00023.x

17. United Nations Department of Economic and Social Affairs: Disability. Convention on the Rights of Persons with Disabilities: Article 24 – Education [webpage on the Internet]. No date [cited 2020 Oct 06]. Available from: https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/article-24-education.html

18. Fish-Hodgson T, Khumalo S. Left in the dark: Failure to provide access to quality education to blind and partially sighted learners in South Africa. A Section 27 report [document on the Internet]. c2015 [cited 2020 Oct 05]. Available from: http://section27.org.za/wp-content/uploads/2016/07/S27-left-in-the-dark-2015-accessible.pdf

19. Donohue D, Bornman J. The challenges of realising inclusive education in South Africa. S Afr J Educ. 2014;34(2):10–14. https://doi.org/10.15700/201412071114

20. Mutanga O. Students with disabilities' experience in South African higher education – a synthesis of literature. S Afr J High Educ. 2017;31(1):135–154. http://dx.doi.org/10.20853/31-1-1596

21. McKinney E, Lourens H, Swartz L. Disability and higher education in South Africa: Political responses and embodied experiences. In: Pattman R, Carolissen R, editors. Transformation in higher education in South Africa. Stellenbosch: Sun PReSS; 2018. p. 293–318. https://doi.org/10.18820/9781928480075

22. Lourens H, Swartz L. 'Every now and then you slip up and then you are in trouble': The responsibility on visually impaired university students to access reasonable accommodations in South Africa. Intl J Disabil Dev Educ. 2020;67(3):320–335. https://doi.org/10.1080/1034912X.2019.1587152

23. Martinez E. "Complicit in exclusion" – South Africa's failure to guarantee an inclusive education for children with disabilities [webpage on the Internet]. c2015 [cited 2020 Oct 05]. Available from: https://www.right-to-education.org/resource/complicit-exclusion-south-africa-s-failure-guarantee-inclusive-education-children

24. Phukubje J, Ngoepe M. Convenience and accessibility of library services to students with disabilities at the University of Limpopo in South Africa. JOLIS. 2016;49(2):180–190. https://doi.org/10.1177/0961000616654959

25. Ntombela S, Soobrayen R. Access challenges for students with disabilities at the University of KwaZulu-Natal: A situational analysis of the Edgewood Campus. J Soc Sci. 2013;37(2):149–155. https://doi.org/10.1080/09718923.2013.11893213

26. Engelbrecht L, De Beer JJ. Access constraints experienced by physically disabled students at a South African higher education institution. Afr Educ Rev. 2014;11(4):544–562. https://doi.org/10.1080/18146627.2014.935003

27. Losinsky LO, Levi T, Saffey K, Jelsma J. An investigation into the physical accessibility to wheelchair bound students in an institution of higher learning in South Africa. Disabil Rehabil. 2003;25:305–308. https://doi.org/10.1080/0963828021000043743

28. Matshedisho KR. Experiences of disabled students in South Africa: Extending the thinking behind disability support. S Afr J High Educ. 2010;24(5):730–744.

29. FOTIM Foundation of Tertiary Institutions of the Northern Metropolis. Disability in higher education project report 2011 [document on the Internet]. c2011 [cited 2020 Oct 09]. Available from: https://www.uct.ac.za/usr/disability/reports/annual_report_10_11.pdf

30. Greyling E. Students with disabilities' experiences of support and barriers to their development at Stellenbosch University [master's thesis]. Stellenbosch: Stellenbosch University; 2008.

31. Van Jaarsveldt DE, Ndeya-Ndereya CN. It's not my problem: Exploring lecturers' distancing behaviour towards students with disabilities. Disabil Soc. 2015;30(2):199–212. https://doi.org/10.1080/09687599.2014.994701

32. Mokiwa SA, Phasha TN. Using ICT at an open distance learning (ODL) institution in South Africa: The learning experiences of students with visual impairments. Afr Educ Rev. 2012;9(1):136–151. https://doi.org/10.1080/18146627.2012.755286

33. Supalo CA, Hill AA, Larrick CG. Summer enrichment programs to foster interest in stem education for students with blindness or low vision. J Chem Educ. 2014;91:1257–1260. https://doi.org/10.1021/ed400585v

34. Supalo C. Independent laboratory access for the blind [webpage on the Internet]. c2007 [cited 2020 Oct 09]. Available from: https://nfb.org/sites/default/files/images/nfb/publications/bm/bm07/bm0705/bm070506.htm

35. Poon T, Ovadia RJ. Using tactile learning aids for students with visual impairments in a first-semester organic chemistry course. J Chem Educ. 2008;85(2):240 – 242. https://doi.org/10.1021/ed085p240

36. Parry M, Brazier M, Fischbach E. Teaching college physics to a blind student. Phys Teach. 1997;35(8):470–474. https://doi.org/10.1119/1.2344770

37. Windelborn AF. Doing physics blind. Phys Teach. 1999;37(6):366–368. https://doi.org/10.1119/1.880322

38. Brawand A, Johnson N. Effective methods for delivering mathematics instruction to students with visual impairments. J Blindness Innov Res. 2016;6(1), Art. #060101. http://dx.doi.org/10.5241/6-86

39. Karshmer AI, Bledsoe C. Access to mathematics by blind students. In: Miesenberger K, Klaus J, Zagler W, editors. Computers helping people with special needs. Berlin: Springer; 2002. p. 471–476. https://doi.org/10.1007/3-540-45491-8_90

40. Neppel K, Oliver-Hoyo MT, Queen C, Reed N. A closer look at acid-base olfactory titrations. J Chem Educ. 2005;82(4):607–610. https://doi.org/10.1021/ed082p607

41. Sahin M, Yorek N. Teaching science to visually impaired students: A small-scale qualitative study. US-China Educ Rev. 2009;6(4):19–26.

42. Avramidis E, Bayliss P, Burden R. A survey into mainstream teachers' attitudes towards the inclusion of children with special educational needs in the ordinary school in one local education authority. Educ Psychol. 2000;20(2):191–211. https://doi.org/10.1080/713663717

Application of Taguchi method and artificial neural network model for the prediction of reductive leaching of cobalt(III) from oxidised low-grade ores

AUTHORS:
K. Brest Kasongo[1]
Henock-Michel Mwanat[2]

AFFILIATIONS:
[1]Department of Metallurgy, University of Johannesburg, Johannesburg, South Africa
[2]Department of Metallurgy and Materials, Faculty of Engineering, University of Lubumbashi, Lubumbashi, Democratic Republic of the Congo

CORRESPONDENCE TO:
Brest Kasongo

EMAIL:
brestkasongo@gmail.com

DATES:
Received: 06 Aug. 2020
Revised: 20 Nov. 2020
Accepted: 25 Nov. 2020
Published: 28 May 2021

HOW TO CITE:
Kasongo KB, Mwanat HM. Application of Taguchi method and artificial neural network model for the prediction of reductive leaching of cobalt(III) from oxidised low-grade ores. S Afr J Sci. 2021;117(5/6), Art. #8743. https://doi.org/10.17159/sajs.2021/8743

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☒ All data included
☐ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Priscilla Baker

KEYWORDS:
cobalt reducing leaching, Taguchi method, artificial neural network back-propagation, optimisation

FUNDING:
None

The leaching process of cobalt using a wide range of experimental variables is described. The treated cobalt samples were from the Kalumbwe Mine in the south of the Democratic Republic of Congo. In this study, a predictive model of cobalt recovery using both the Taguchi statistical method and an artificial neural network (ANN) algorithm was proposed. The Taguchi method utilising a L_{25} (5^5) orthogonal array and an ANN multi-layer, feed-forward, back-propagation learning algorithm were adopted to optimise the process parameters (acid concentration, leaching time, temperature, percentage solid, and sodium metabisulfite concentration) responsible for the high recovery of cobalt by reducing sulfuric acid leaching. The ANN was built with a neuron in the output layer corresponding to the cobalt leaching recovery, 10 hidden layers, and 5 input variables. The validation of the ANN model was performed with the results of the Taguchi method. The optimised trained neural network depicts the testing data and validation data with R^2 equal to 1 and 0.5676, respectively.

Significance:

- We statistically investigated the main factors (acid concentration, leaching time, temperature, percentage solid, and sodium metabisulfite concentration) that affect the cobalt(III) leaching performance using both the Taguchi method and artificial neural network model. This allowed us to ascertain that it is indeed possible to leach cobalt(III) from oxide ores and to identify the optimum leaching conditions.

Introduction

Cobalt's utility in green energy and modern industry makes it very important and the price of cobalt is growing rapidly due to its high demand. Cobalt is a key component in rechargeable batteries and other consumer electronic products and its demand is expected to expand further with the increased use of electric vehicles.[1-4] However, the cobalt content in the earth's crust is scarce (only 0.001%).[4] The Democratic Republic of Congo deposits represent an important resource for cobalt ore[5,6] and several deposits are currently in development[2,3,7,8]. The most common cobalts from oxidised ore found in economic deposits include absolane (CoO) and heterogenite (CoOOH (Co_2O_3)). In heterogenite minerals, the cobalt is present in both bivalent and trivalent states. The dissolution of cobalt oxide ores may be accomplished in sulfuric acid media but cobalt in a trivalent state leaches only in the presence of reducing agents such as sulfur dioxide (SO_2)[9], sodium metabisulfite ($Na_2S_2O_5$) known as SMBS, metallic copper powder, and ferrous ions[2,8,10,11]. Under reducing acid leaching, the dissolution rate of cobalt was reported to be faster than that under standard acid leaching.[12] In contrast, using SO_2 could engender environmental issues due to gaseous emissions.[13] The use of SO_2 derivatives such as sodium sulfite (Na_2SO_3) or sodium metabisulfite ($Na_2S_2O_5$) reduces or eliminates the environmental risks.[13]

The Co(III) reduction can be represented by the electrochemical reaction shown by Equation 1[8,14]:

$$Co_2O_3 + 6H^+ + 2e^- \rightarrow 2Co^{2+} + 3H_2O$$
Equation 1

More generally, the reaction mechanism for reducing Co(III) to Co(II) by using reducing agents such as sulfur dioxide and sodium metabisulfite is still not well understood. Several authors have postulated that the iron contained in the ore is responsible for this reduction. On the other hand, some authors believe that the action of SO_2 is responsible. As for the reaction mechanism for the reduction of Co(III) to Co(II) (case of leaching of heterogenite) with sodium metabisulfite ($Na_2S_2O_5$), the possible reactions are:

$$Na_2S_2O_5 + H_2SO_4 \rightarrow Na_2SO_4 + 2SO_2 + H_2O$$
Equation 2

$$SO_2 + H_2O \rightarrow H_2SO_3$$
Equation 3

$$H_2SO_3 \rightarrow HSO_3^- + H^+$$
Equation 4

$$Co_2O_3 + HSO_3^- \rightarrow 2CoO + HSO_4^-$$
Equation 5

$$HSO_4^- \rightarrow H^+ + SO_4^{2-}$$
Equation 6

$$CoO + 2H^+ + SO_4^{2-} \rightarrow CoSO_4 + H_2O$$
Equation 7

The leaching mechanism might involve a direct attack of absolane by sulfuric acid, according to Equation 8:

$$CoO + H_2SO_4 \rightarrow CoSO_4 + H_2O \qquad \text{Equation 8}$$

Cobalt recovery can be achieved by solvent extraction, cementation, selective precipitation, ion exchange, and electrowinning.[4,15-18] The choice of method depends on the concentration of impurities, relative capital costs for disposal, and related operational preferences.[17] Due to environmental issues related to SO_2 emissions, in this work, sodium metabisulfite has been used as a reducing agent.

Parameter optimisation is used to make the process more efficient. Many optimisation methods are described in the literature, such as genetic algorithm[19,20], differential evolution, simplex linear programming[21], and experimental designs, especially the Taguchi method[22,23]. The Taguchi method contributes to study the effects of factors and the optimisation of the leaching yield.[23-26] Reductions in the number of running tests and the financial cost, as well as the gain in time, constitute the recognised benefit of the Taguchi method in optimising the parameters and/or predicting a given response.[27] In this approach, the experimental matrix is designed, and corresponding responses of the system are identified. The artificial neural network (ANN) is an efficient and attractive tool that can complete the Taguchi method.

Recently, several studies have been conducted on the applicability of ANNs as a predictive model algorithm for cobalt recovery in comparison with the particle swarm optimisation algorithm[28] and germanium recovery in comparison with the genetic algorithm[29]. Some studies relate to the prediction of chemical desulfurisation of Tabas coal and the prediction of leaching recovery for Al_2O_3 with ANNs.[30,31] An ANN has been used to estimate nitrate concentration in groundwater[32] and the concentration of major ions in rivers[33]. Hoseinian et al.[34] developed the ANN model for predicting column leaching recovery of copper by considering four leaching parameters as inputs to the model, namely, column height, particle size, acid flow rate, and leaching time.

In this study, the Taguchi method and ANN model were applied for process optimisation. The aim was to determine the optimal conditions of reducing leaching of cobalt ores in the batch by using the Taguchi method and to predict the cobalt recovery by using both the Taguchi method and ANNs.

Materials and methods

Materials

Ore sample and experimental procedure

The raw material was an oxidised copper-cobalt-bearing mineral. Samples of the ore were collected from Kalumbwe Mine (operated By Kalumbwe Myunga Mining). This mine is located 60 km from Kolwezi Town in Lualaba Province, in the south of the Democratic Republic of Congo. An ore sample of 5 kg was carefully extracted. After milling, 80% of cumulative passing, i.e. P_{80}, had a particle size less than 150 μm. The average cobalt contained in the samples was about 0.8%, essentially in the oxide form. A chemical analysis was performed using atomic absorption spectroscopy. The main elements are shown in (Table 1).

Table 1: Chemical analysis of the ore sample

Element	Cu	Co	Fe	Mn
Concentration (wt. %)	2.48	0.8	3.8	0.19

The acid-reductive leaching experiments were performed in glass beakers that were carefully cleaned with distilled water. According to the operational conditions of acid concentration, different leaching solutions were prepared by mixing sulfuric acid (98%) with distilled water. The sodium metabisulfite (≤98%) was used to maintain the leaching media reductive. The whole reagents were analytical grade. The assembly was placed on a hot plate equipped with a mechanical stirring device. The temperature was monitored using a thermometer that was placed permanently in the solution. After each leaching experiment, the pregnant solution was separated using a vacuum pump with a membrane filter and the cobalt concentrations were determined by atomic absorption spectroscopy.

Mineralogical characterisation

Scanning electron microscopy equipped with energy dispersive X-ray spectrometry was used to determine the minerals phase of the ore sample; the results are given in Table 2.

Table 2: Minerals phase of the ore sample

Minerals	**Composition (wt. %)**
$CuCO_3.Cu(OH)_2.Cu_2S$	0.08
$CuFeS_2$	0.10
Cu	0.01
$CuCO_3.Cu(OH)_2$	4.86
$2CuCO_3.Cu(OH)_2$	0.34
$Cu_3[AlSi_3O_2](OH)_2$	0.40
CuO	0.68
FeS	0.36
Talc	3.94
$Fe_2O_3.H_2O$	2.59
Fe_3O_2	2.10
MnOx	0.27
SiO_2	44.86
$(CaMg(CO_3)_2)$	7.25
$Co(Mg.Co)(CO_3)_2$	1.34
$(MgFeCo)(CO_3)_2$	0.24
Mica	8.89
CuCox	0.03
$Co_2O_3.H_2O$	0.53
MgO	13.10
MontMorillonite	7.12
Apatite, rutile, barite	0.19

Methods

Taguchi method

The Taguchi approach involves device design, design of parameters, and design of tolerances to achieve a robust process and the best quality product.[35] Five parameters – namely, acid concentration, leaching time, temperature, percentage solid, and SMBS concentration – were selected and varied in five different levels as shown in Table 3.

Table 3: Different parameters and their levels for the Taguchi method

Leaching parameters	**Levels**				
Acid concentration (g/L)	20	40	60	80	100
Leaching time (min)	60	90	120	150	180
Temperature (°C)	25	35	45	55	65
Percentage solid (%)	10	15	20	25	30
SMBS concentration (g/L)	2	4	6	8	10

SMBS, sodium metabisulfite ($Na_2S_2O_5$)

A total of 25 leaching batch experiments were conducted according to the selected parameters and their levels, in which cobalt recovery was identified as a response. This measure was calculated using Equation 9:

$$r_{Co} = \frac{C_{Co_l} V_l}{W_{Co_i}} \text{ x } 100 \qquad \text{Equation 9}$$

where r_{Co} is the cobalt recovery (%), C_{Co_l} is the concentration of cobalt contained in the pregnant solution (g/L), V_l is the volume of the leaching solution (mL) and W_{Co_i} is the weight of cobalt in the material (g). The value of the experimental performance can be predicted using Equation 10:

$$Y_{opt} = \frac{T}{n} + (A_i - \frac{T}{n}) + (B_j - \frac{T}{n}) + \ldots + (M_n - \frac{T}{n}) \qquad \text{Equation 10}$$

where Y_{opt} is the optimal value of responses, n is the total number of tests, T is the sum of all the test responses, and A_i, $B_j \ldots M_n$ are the response averages of level *i, j, ... n,* respectively.

In the Taguchi approach, the response of each experiment and the corresponding variation were analysed by using the factor signal/noise ratio (S/N). The highest value of the functional metric S/N determined by Equation 11 represents the high performance of the response in the considered criterion of optimisation.

$$\frac{S}{N} = -10log\ (\frac{1}{n})\ \Sigma_i^n \frac{1}{y_i^2} \qquad \text{Equation 11}$$

where y_i is the signal (cobalt recovery) measured in each experiment averaged over *n* repetitions.

The data of S/N report the rank and delta (D) values to identify the parameters that have the greatest effect on the cobalt recovery as a response variable. The delta value is estimated as the difference between the highest and lowest value of S/N for a given operating factor. The rank is the tool helper that allows identification of the factor that has the largest effect.The factor with the largest delta value affects mostly the response. The numerical sorting of the rank values determines the order of importance of the factors.[36,37]

Artificial neural network model

Artificial neural networks were developed in the 1940s for applications in science and engineering.[30] ANNs are common techniques for machine learning which simulates the learning mechanism in biological organisms.[38] Therefore, an ANN consists of several basic components called neurons. The latter are interconnected by weighted links that can be modified using ANN training data to solve a specific problem. Normally, neurons are organised in layers so that those in the same layer behave similarly.[39,40] Network architecture refers to the arrangement of neurons into layers and the connection patterns within and between layers. In general, neurons are not linked inside the same layers.

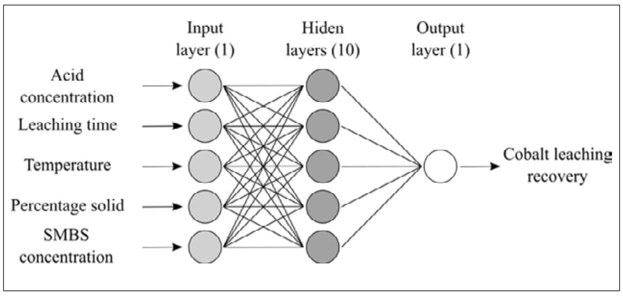
The feed-forward network is a popular ANN architecture which only connects neurons to the output layer. Back-propagation is a method of modifying the weighted connections between neurons using the Widrow–Hoff learning method to reduce the error between predicted data and input data.[40] These configuration procedures of an ANN model template include the following steps[30]:

1. Data collection
2. Train and test set determination
3. Data conversion into the ANN inputs
4. Determining, training, and testing the network topology
5. Repeating the steps *n* times if it is required to determine the optimal model
6. Application of the optimal ANN model

Due to the complexity of extraction mining worldwide, computer models are an important tool for reducing production costs. Recently, in the cobalt industry, analytical techniques were introduced to improve both the process and the results obtained through the leaching process.[39]

The back-propagation algorithm was used for network training, which is a statistical technique using supervised learning, not always converging to the absolute minimum, and has a low convergence rate. The connection weights of ANN by the back-propagation algorithm are modified only from the local angle, and the entire learning process is not examined for the global perspective. So, it can be stopped at a minimum local level.[34] Training of learning works due to the changes in connection weights, based on the calculated errors of the observed values, starting from the output, and progressing to the input.

The three-layer, feed-forward back-propagation ANN was constructed with five neurons in the input layer for five input variables, ten neurons were chosen in the hidden layer, and one neuron was used in the output layer corresponding to cobalt recovery as shown in Figure 1. The acid concentration, leaching time, temperature, SMBS concentration, and percentage solid were variables of the network.



SMBS, sodium metabisulfite ($Na_2S_2O_5$)

Figure 1: Three-layer architecture of the artificial neural network back-propagation training for prediction of cobalt leaching efficiency.

Both input and output data (before feeding to the networks) are standardised in the range of 0.1 and 0.9 (Equation 10)[29,41] to reduce the influence of outliers and to facilitate network learning[42].

$$p = 0.1 + 0.8 \text{ x } \frac{p_i - p_{min}}{p_{max} - p_{min}}$$

where p_{min} and p_{max} for all the feeding data vectors are respectively the minima and maximum values of the i^{th} node in the input layer ($1 \leq i \leq n$).

The outputs, after the simulation step, are converted back into an unnormalised condition by Equation 13:

$$p_i = \frac{1}{0.8} \text{ x } [(p-0.1) \text{x} (p_{max} - p_{min})] + p_{min} \qquad \text{Equation 13}$$

where p_i is the normalised parameter, p_{min} is the minimum of the actual parameters, p_{max} is the maximum of the actual parameters and *p* is the unnormalised predicted parameter. The multilayer feed-forward with the Marquardt algorithm was implemented for the training set. The tangent sigmoid function was used as an activation function. The mathematical expression of the sigmoid function is indicated in Equation 14:

$$y = \Sigma_{j=1}^{9} L_j \Sigma_{i=1}^{9} w_{ij}^{i} * x_i \qquad \text{Equation 14}$$

where x_i represents the i^{th} input value to the neuron, w_{ij}^{i} represents the i^{th} weight associated with the neuron *i* of the layer *j*, and L_j represents the constant of the j^{th} layer.

The ANN model was used to train the networks, according to the design of the experiment showed in Table 4. Of the 25 sets of data collected, 20 sets (80%) were randomly selected to train the network and 5 sets (20%) were used to validate its correctness. These five input variables and one output variable constitute the general model as shown in Figure 2, in which the number of hidden layer neurons is greater than one.

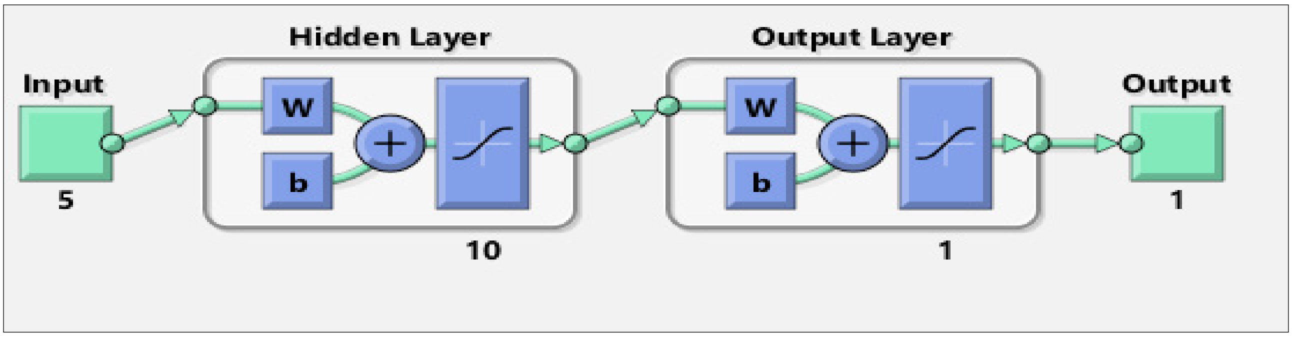


Figure 2: Topology of the neural network architecture in the artificial neural network model.

The coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) were used as the performance criteria of the ANN model. R^2 is a measure of the variability of the data reproduced by the model and the observations. MAE and RMSE indicate residual errors.[42] The values of R^2, RMSE, and MAE were respectively calculated using Equations 15 to 17:

$$R^2 = 1 - \frac{\Sigma^n_{i=1}(\hat{Y} - Y)^2}{\Sigma^n_{i=1}(\hat{Y})^2} \qquad \text{Equation 15}$$

$$MAE = \frac{1}{n}\Sigma^n_{i=1}(\hat{Y} - Y)^2 \qquad \text{Equation 16}$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}(\hat{Y} - Y)^2} \qquad \text{Equation 17}$$

where n is the number of observations, $\hat{Y}$ is the measured value of cobalt recovery, and Y is the estimated value of cobalt recovery by the model.

These criteria were used to evaluate ANN performance in the leaching process of cobalt recovery because the input and output data were quantitative variables as these criteria are common in model performance evaluation. To generate the results with ANN, the use of software such as MATLAB® computing environment or another advanced calculation program is required, due to the high complexity of such a modeling technique.[39]

Results and discussion

Taguchi method

The objective was to determine the optimum conditions at which the yield of cobalt is maximised. The results of the experiments as per the L_{25} (5^5) array and the corresponding S/N ratios are given in Table 4.

The average S/N ratio of factors for each variation level is given in Table 5.

These results allowed us to determine the optimum levels of the factors according to the S/N ratio. The factor levels that maximise the S/N ratio were indicated as the optimal parameters. These values are plotted in Figure 3.

The highest average S/N ratios of cobalt recovery for all runs were found at Level 2, corresponding to 38.22 for the acid concentration, 38.16 for the leaching time and 38.05 for the percentage solid, and at Level 5, corresponding to 38.33 for the temperature and 38.63 for the SMBS concentration (Table 5). Thereby, the optimal combination of factor levels for the maximum recovery of cobalt trivalent from oxide low-grade ore was: 40 g/L, 90 min, 65 °C, 15%, and 10 g, respectively, for acid concentration, leaching time, temperature, solid percentage, and SMBS concentration. Table 5 includes ranks based and delta statistics (D) which compare the relative magnitude of parameter effects. Based on the delta values, the ranks were assigned. From the results showed in Table 5, the metabisulfite weight had the highest value of delta, which means that the metabisulfite weight had a large effect on the leaching process of cobalt(III). The ranking of factors in order of importance was: SMBS concentration > acid concentration > temperature > leaching time > solid percentage. Figure 3 generated by Statistica Enterprise® software gives the levels of factors that optimise the yield of cobalt according to the design of experiments as shown in Table 4.

Under these optimum conditions for cobalt leaching, the predictive model corresponding to Equation 9 has given a leaching yield of cobalt equal to 98.71%, while the experimental test carried out for confirmation under the optimal conditions gave a cobalt recovery of 97.43%. This minimal difference between the theoretical and experimental values demonstrates the robustness of the Taguchi method and the minimisation of the noise factors around the studied response.

Artificial neural network

Throughout the neighbouring layers, neurons are completely attached to each node. During modelling, no concept of bias was used; an impulse concept was used to help achieve better convergence during iterations. The system ran for 60 000 iterations; with each iteration, an error is propagated backward between the expected value and the actual value through the hidden layers from the output layer to the input until the error is within a reasonable limit.[30] In the training step, the ANN model showed good performance with R^2, RMSE and MAE (%) values of 1, 0.0022, and 0.01, respectively. In the testing step, the values for a good adjustment were 0.5676, 0.0081 and 0.81 for R^2, RMSE and MAE (%), respectively (Table 6). The results are plotted in Figure 4. It was observed that the ANN model could be used to predict cobalt leaching satisfactorily.

The ANN was constructed with experimental data from the Taguchi method. Figure 5 shows the comparison of the response of ANN in the training process and the measured data. Figure 5 also shows that the measured cobalt recoveries are close to the estimated recoveries by ANN in the training process.

Conclusion

The Taguchi method and ANN algorithm were implemented to predict the cobalt leaching rate from cobalt-bearing ore using sulfuric acid and sodium metabisulfite mixture in reducing leaching conditions. Using the Taguchi L_{25}(5^5) orthogonal design of experiment and considering the acid concentration, leaching time, temperature, solid percentage, and sodium metabisulfite concentration as controllable parameters, the optimised conditions for the leaching of cobalt were calculated as 100 g/L for acid concentration, 60 min for leaching time, 65 °C for temperature, 15% for solid percentage, and 10 g for sodium metabisulfite concentration. The cobalt leaching yield was 98.71%. In the ANN, the parameters mentioned above were considered as inputs and cobalt leaching rate as the output. In these networks, a multi-layer ANN back-propagation algorithm with {5-10-1-1} was trained by using the Levenberg–Marquardt algorithm to predict the cobalt recovery. The R^2 values were 1 and 0.56761, RMSE values were 0.0022 and 0.0081, and the MAE (%) values were 0.01 and 0.13, respectively, for the training and testing sets for cobalt recovery according to the ANN algorithm. After the validation of the ANN algorithm, the training of normalised optimal conditions obtained by the Taguchi method gave a leaching yield of 86.82% cobalt, whereas with the Taguchi method, the leaching yield was 98.71%. This gap may be explained by the parameters chosen in the architecture for the training model (the number of the hidden layers, iterations, and the algorithm).

Table 4: Experiment matrix and results of the leaching tests

Experiment matrix according to design of experiment Taguchi L_{25} (5^5)						Cobalt yield	S/N (for yield)
Test	Acid concentration (g/L)	Leaching time (min)	Temperature (°C)	Percentage solid (%)	SMBS concentration (g/L)	Co (%)	
1	20	60	25	10	2	**64.41**	**36.18**
2	20	90	35	15	4	**76.67**	**37.69**
3	20	120	45	20	6	**83.97**	**38.48**
4	20	150	55	25	8	**75.70**	**37.58**
5	20	180	65	30	10	**87.19**	**38.81**
6	40	60	35	20	8	**75.88**	**37.60**
7	40	90	45	25	10	**93.10**	**39.38**
8	40	120	55	30	2	**77.34**	**37.77**
9	40	150	65	10	4	**87.87**	**38.88**
10	40	180	25	15	6	**74.79**	**37.48**
11	60	60	45	30	4	**71.16**	**37.05**
12	60	90	55	10	6	**83.62**	**38.45**
13	60	120	65	15	8	**85.52**	**38.64**
14	60	150	25	20	10	**86.03**	**38.69**
15	60	180	35	25	2	**68.93**	**36.77**
16	80	60	55	15	10	**89.73**	**39.06**
17	80	90	65	20	2	**79.56**	**38.01**
18	80	120	25	25	4	**74.33**	**37.42**
19	80	150	35	30	6	**79.32**	**37.99**
20	80	180	45	10	8	**80.84**	**38.15**
21	100	60	65	25	6	**73.50**	**37.33**
22	100	90	25	30	8	**73.10**	**37.28**
23	100	120	35	10	10	**72.47**	**37.20**
24	100	150	45	15	2	**74.09**	**37.40**
25	100	180	55	20	4	**66.34**	**36.43**

Table 5: Response for signal/noise ratio

Level	Acid concentration (g/L)	Leaching time	Temperature (°C)	Percentage solid (%)	SMBS concentration (g/L)
1	37.75	37.44	37.41	37.77	37.22
2	38.22	38.16	37.45	38.05	37.49
3	37.92	37.9	38.09	37.85	37.94
4	38.13	38.11	37.86	37.7	37.85
5	37.13	37.53	38.33	37.78	38.63
Delta (D)	1.09	0.72	0.92	0.36	1.4
Rank	2	4	3	5	1

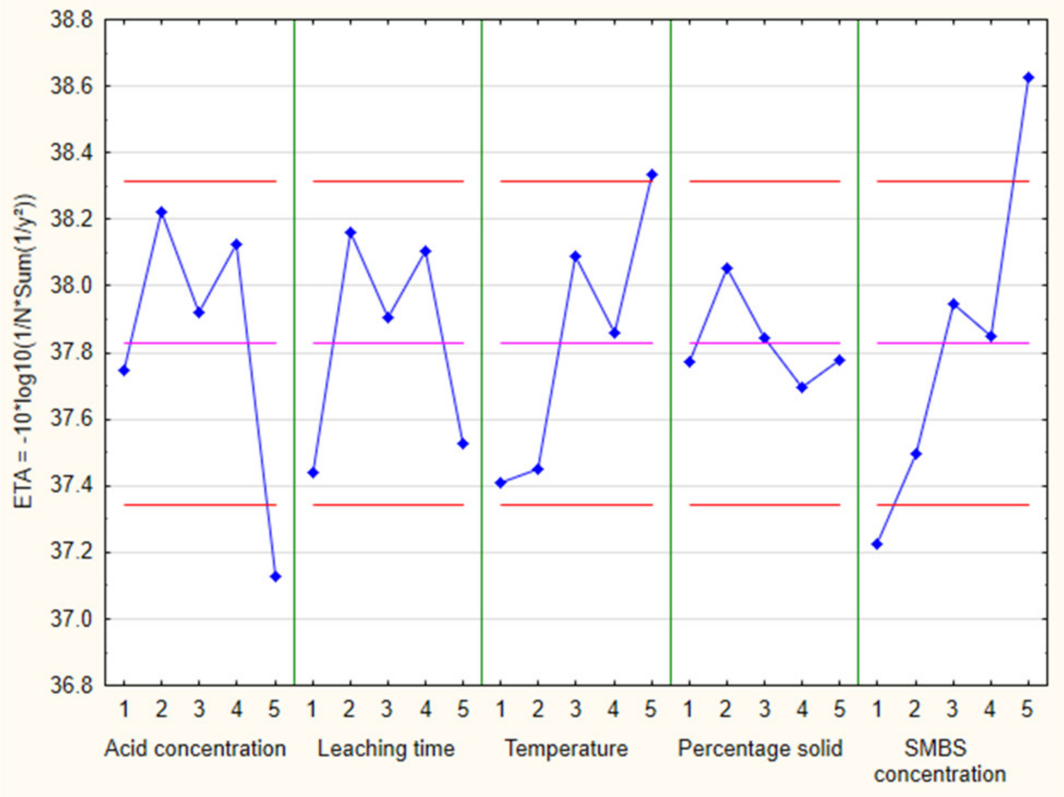


Figure 3: Effects of controllable factors associated with their levels on the statistical performance (signal-to-noise ratio) for the leaching of cobalt.

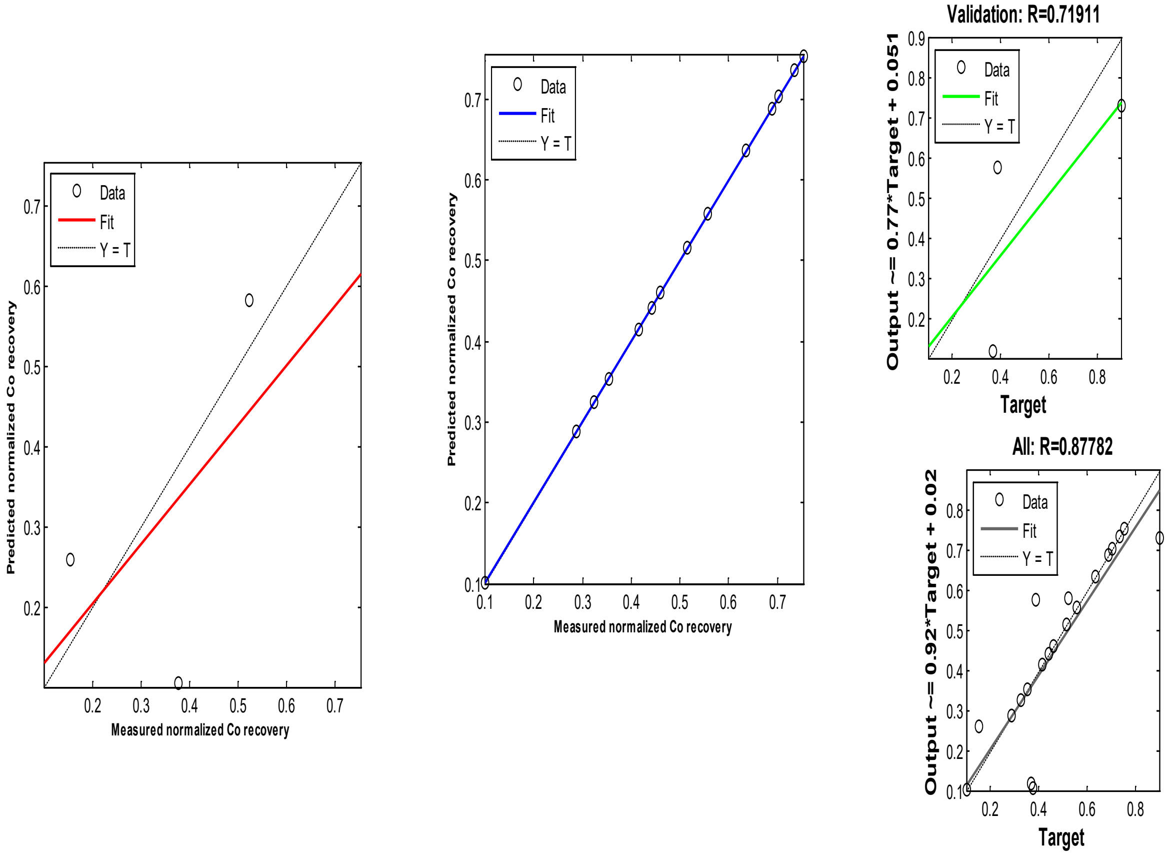


Figure 4: Predicted cobalt recovery by the artificial neural network model in the training process vs actual measurement.

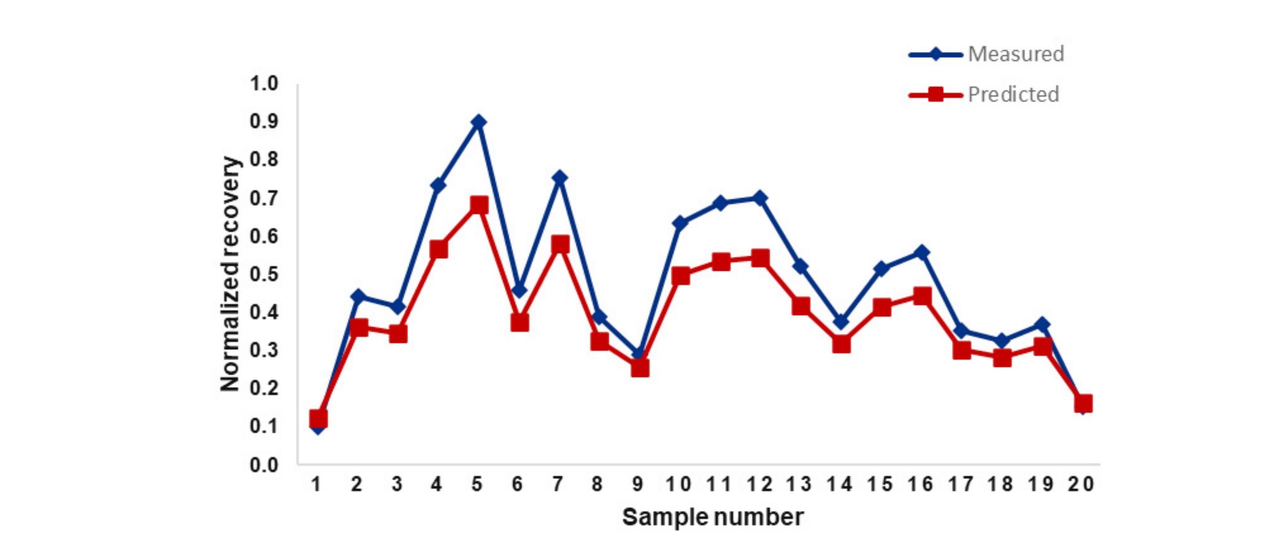


Figure 5: Comparison of measured cobalt recovery with that estimated by the artificial neural network model in the training process.

Table 6: Performance criteria of the predictive ANN model

Criteria	Value	
	Testing	Training
R^2	0.5676	1
Root mean square error	0.0081	0.0022
Mean absolute error (%)	0.1300	0.0100

The results show that the proposed model can be used to predict the cobalt recovery, with a reasonable error, according to the parameters affecting the recovery of cobalt.

Acknowledgements

We thank MetLab Solutions for its generous support.

Competing interests

We have no competing interests to declare.

Authors' contributions

K.B. was responsible for the conceptualisation of the article; processed, analysed and validated the data; and wrote the first draft and revisions. M.H. identified the appropriate methodology; prepared samples, undertook the experiments and data collection; processed, analysed and validated the data; and provided leadership and advice critical for the successful completion of the work.

References

1. Santoro L, Tshipeng S, Pirard E, Bouzahzah H, Kaniki A, Herrington R. Mineralogical reconciliation of cobalt recovery from the acid leaching of oxide ores from five deposits in Katanga (DRC). Miner Eng. 2019;137:277–289. https://doi.org/10.1016/j.mineng.2019.02.011
2. Apua MC, Bafubiandi AFM. Dissolution of oxidised Co–Cu ores using hydrochloric acid in the presence of ferrous chloride. Hydrometallurgy. 2011;108(3–4):233–236. http://dx.doi.org/10.1016/j.hydromet.2011.04.012
3. Crundwell FK, Moats MS, Robinson TG, Davenport WG, editors. Extractive metallurgy of nickel, cobalt and platinum-group metals. Oxford: Elsevier; 2011. https://doi.org/10.1016/B978-0-08-096809-4.10040-1
4. Song S, Sun W, Wang L, Liu R, Han H. Recovery of cobalt and zinc from the leaching solution of zinc smelting slag. J Environ Chem Eng. 2019;7(1):102777. https://doi.org/10.1016/j.jece.2018.11.022
5. Liu W, Rao S, Wang W, Yang T, Yang L, Chen L, et al. Selective leaching of cobalt and iron from cobalt white alloy in sulfuric acid solution with catalyst. Int J Miner Process. 2015;141:8–14. http://dx.doi.org/10.1016/j.minpro.2015.06.002
6. Ferron CJ. Sulfur dioxide : A versatile reagent for the processing of cobaltic oxide minerals. Aqueous Process. 2008;60(10): 50–55. https://doi.org/10.1007/s11837-008-0136-6
7. Shengo ML, Kime MB, Mambwe MP, Nyembo TK. A review of the beneficiation of copper-cobalt-bearing minerals in the Democratic Republic of Congo. J Sustain Min. 2019;18(4):226–246. https://doi.org/10.1016/j.jsm.2019.08.001
8. Zeka L, Lambert F, Frenay J, Gaydardzhiev S, Ilungandala A. Possibilities for Co(III) dissolution from an oxidized ore through simultaneous bioleaching of pyrite. Miner Eng. 2015;75:54–62. http://dx.doi.org/10.1016/j.mineng.2014.12.023
9. Park KH, Kim HI, Das RP. Selective acid leaching of nickel and cobalt from precipitated manganese hydroxide in the presence of chlorine dioxide. Hydrometallurgy. 2005;78:271–277. https://doi.org/10.1016/j.hydromet.2005.05.001
10. Mwema MD, Mpoyo M, Kafumbila K. Use of sulphur dioxide as reducing agent in cobalt leaching at Shituru hydrometallurgical plant. J South Afr Inst Min Metall. 2002;102(1):1–4.
11. Kime MB, Kanowa EK. Valorization of low-grade copper-cobalt ore from the Mukondo mine by heap leaching and solvent extraction. CIM J. 2017;8(4):1–8. https://doi.org/10.15834/cimj.2017.25
12. Sadegh M, Dhawan N, Birinci M, Moradkhani D. Reductive leaching of cobalt from zinc plant purification residues. Hydrometallurgy. 2011;106(1–2):51–57. http://dx.doi.org/10.1016/j.hydromet.2010.11.017

13. Tshibanda P, Kime M, Edouard M, Richard B, Arthur T. Agitation and column leaching studies of oxidised copper-cobalt ores under reducing conditions. Miner Eng. 2017;111:47–54. http://dx.doi.org/10.1016/j.mineng.2017.06.001

14. Pourbaix M. Atlas of electrochemical equilibria in aqueous solutions. J Electroanal Chem Interfacial Electrochem. 1963;13(4):471. https://doi.org/10.1016/0022-0728(67)80059-7

15. Chong S, Hawker W, Vaughan J. Selective reductive leaching of oxidised cobalt containing residue. Miner Eng. 2013;54:82–87. http://dx.doi.org/10.1016/j.mineng.2013.04.004

16. Kongolo K, Mwema MD, Banza AN, Gock E. Cobalt and zinc recovery from copper sulphate solution by solvent extraction. Miner Eng. 2003;16(12):1371–1374. https://doi.org/10.1016/j.mineng.2003.09.001

17. Swartz B, Donegan S, Amos SR. Processing considerations for cobalt recovery from Congolese copperbelt ores. Hydrometall Conf. 2009;385–400.

18. Pradhan N, Singh P, Tripathy BC, Dasq SC. Electrowinning of cobalt from acidic sulphate solutions – Effect of chloride ions. 2001;14(7):775–783. https://doi.org/10.1016/S0892-6875(01)00072-3

19. Bhatti MS, Kapoor D, Kalia RK, Reddy AS, Thukral AK. RSM and ANN modeling for electrocoagulation of copper from simulated wastewater: Multi objective optimization using genetic algorithm approach. Desalination. 2011;274(1–3):74–80. http://dx.doi.org/10.1016/j.desal.2011.01.083

20. Pettersson F, Biswas A, Sen PK, Saxén H, Chakraborti N. Analyzing leaching data for low-grade manganese ore using neural nets and multiobjective genetic algorithms. Mater Manuf Process. 2009;24(3):320–330. https://doi.org/10.1080/10426910802679386

21. Karterakis SM, Karatzas GP, Nikolos IK, Papadopoulou MP. Application of linear programming and differential evolutionary optimization methodologies for the solution of coastal subsurface water management problems subject to environmental criteria. J Hydrol. 2007;342(3–4):270–282. https://doi.org/10.1016/j.jhydrol.2007.05.027

22. Mbuya BI, Kime MB, Tshimombo AMD. Comparative study of approaches based on the Taguchi and ANOVA for optimising the leaching of copper–cobalt flotation tailings. 2017;512–521. https://doi.org/10.1080/00986445.2017.1278588

23. Khoshnevisan A, Yoozbashizadeh H. Determination of optimal conditions for pressure oxidative leaching of sarcheshmeh molybdenite concentrate using Taguchi method. J Min Metall Sect B Metall. 2012;48(1):89–99. https://doi.org/10.2298/JMMB110308003K

24. Ilyas S, Bhatti HN, Bhatti IA, Sheikh MA, Ghauri MA. Bioleaching of metal ions from low grade sulphide ore: Process optimization by using orthogonal experimental array design. Afr J Biotechnol. 2010;9(19):2801–2810.

25. Safarzadeh MS, Moradkhani D, Ilkhchi MO, Golshan NH. Determination of the optimum conditions for the leaching of Cd-Ni residues from electrolytic zinc plant using statistical design of experiments. Sep Purif Technol. 2008;58(3):367–376. https://doi.org/10.1016/j.seppur.2007.05.016

26. Guo ZH, Pan FK, Xiao XY, Zhang L, Jiang KQ. Optimization of brine leaching of metals from hydrometallurgical residue. Trans Nonferrous Met Soc China. 2010;20(10):2000–2005. http://dx.doi.org/10.1016/S1003-6326(09)60408-8

27. Phadke MS. Quality engineering using robust design. Hoboken, NJ: Prentice Hall PTR; 1995.

28. Ebrahimzade H, Khayati GR, Schaffie M. PSO–ANN-based prediction of cobalt leaching rate from waste lithium-ion batteries. J Mater Cycles Waste Manag. 2020;22(1):228–239. https://doi.org/10.1007/s10163-019-00933-2

29. Akkurt S, Ozdemir S, Tayfur G, Akkurt S, Ozdemir S, Tayfur G. Genetic algorithm – artificial neural network model for the prediction of germanium recovery from zinc plant residues. Miner Process Extr Metall. 2016;9553(May):129–134.

30. Jorjani E, Chelgani SC, Mesroghli S. Application of artificial neural networks to predict chemical desulfurization of Tabas coal. 2008;87:2727–2734. https://doi.org/10.1016/j.fuel.2008.01.029

31. Chelgani SC, Jorjani E. Artificial neural network prediction of Al_2O_3 leaching recovery in the Bayer process — Jajarm alumina plant (Iran). Hydrometallurgy. 2009;97(1–2):105–110. http://dx.doi.org/10.1016/j.hydromet.2009.01.008

32. Wagh V, Panaskar D, Muley A, Mukate S, Gaikwad S. Neural network modelling for nitrate concentration in groundwater of Kadava River basin, Nashik, Maharashtra, India. Groundw Sustain Dev. 2018;7:436–445. https://doi.org/10.1016/j.gsd.2017.12.012

33. Nhantumbo C, Carvalho F, Uvo C, Larsson R, Larson M. Applicability of a processes-based model and artificial neural networks to estimate the concentration of major ions in rivers. J Geochemical Explor. 2018;193:32–40. https://doi.org/10.1016/j.gexplo.2018.07.003

34. Hoseinian FS, Abdollahzade A, Mohamadi SS, Hashemzadeh M. Recovery prediction of copper oxide ore column leaching by hybrid neural genetic algorithm. Trans Nonferrous Met Soc China. 2017;27(3):686–693. http://dx.doi.org/10.1016/S1003-6326(17)60076-1

35. Mondal S, Paul B, Kumar V, Singh DK, Chakravartty JK. Parametric optimization for leaching of cobalt from Sukinda ore of lateritic origin – A Taguchi approach. Sep Purif Technol. 2015;156:827–834. http://dx.doi.org/10.1016/j.seppur.2015.11.007

36. Moralı U, Demiral H, Şensöz S. Optimization of activated carbon production from sunflower seed extracted meal: Taguchi design of experiment approach and analysis of variance. J Clean Prod. 2018;189:602–611. https://doi.org/10.1016/j.jclepro.2018.04.084

37. Khanna N, Davim JP. Design-of-experiments application in machining titanium alloys for aerospace structural components. Meas J Int Meas Confed. 2015;61:280–290. http://dx.doi.org/10.1016/j.measurement.2014.10.059

38. Aggarwal CC. Neural networks and deep learning. Cham: Springer; 2018. https://doi.org/10.1007/978-3-319-94463-0

39. Leiva C, Flores V, Salgado F, Poblete D, Acuña C. Applying softcomputing for copper recovery in leaching process. Sci Program. 2017;2017, Art. #6459582, 6 pages. https://doi.org/10.1155/2017/6459582

40. Al-Thyabat S. On the optimization of froth flotation by the use of an artificial neural network. J China Univ Min Technol. 2008;18(3):418–426. https://doi.org/10.1016/S1006-1266(08)60087-5

41. Karri RR, Sahu JN. Process optimization and adsorption modeling using activated carbon derived from palm oil kernel shell for Zn(II) disposal from the aqueous environment using differential evolution embedded neural network. J Mol Liq. 2018;265:592–602. https://doi.org/10.1016/j.molliq.2018.06.040

42. Silva TS, de Freitas Souza M, Maria da Silva Teófilo T, Silva dos Santos M, Formiga Porto MA, Martins Souza CM, et al. Use of neural networks to estimate the sorption and desorption coefficients of herbicides: A case study of diuron, hexazinone, and sulfometuron-methyl in Brazil. Chemosphere. 2019;236:1–15. https://doi.org/10.1016/j.chemosphere.2019.07.064

Check for updates

Source apportionment of ambient $PM_{10-2.5}$ and $PM_{2.5}$ for the Vaal Triangle, South Africa

AUTHORS:
Luckson Muyemeki[1]
Roelof Burger[1]
Stuart J. Piketh[1]
Brigitte Language[1]
Johan P. Beukes[2]
Pieter G. van Zyl[2]

AFFILIATIONS:
[1]Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa
[2]Chemical Resource Beneficiation, North-West University, Potchefstroom, South Africa

CORRESPONDENCE TO:
Luckson Muyemeki

EMAIL:
lucksonmuyemeki@gmail.com

DATES:
Received: 10 July 2020
Revised: 21 Nov. 2020
Accepted: 22 Nov. 2020
Published: 28 May 2021

HOW TO CITE:
Muyemeki L, Burger R, Piketh SJ, Language B, Beukes JP, Van Zyl PG. Source apportionment of ambient $PM_{10-2.5}$ and $PM_{2.5}$ for the Vaal Triangle, South Africa. S Afr J Sci. 2021;117(5/6), Art. #8617. https://doi.org/10.17159/sajs.2021/8617

ARTICLE INCLUDES:
☒ Peer review
☒ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☐ On request from author(s)
☒ Not available
☐ Not applicable

EDITOR:
Yali Woyessa

KEYWORDS:
$PM_{10-2.5}$, $PM_{2.5}$, chemical composition, trajectory cluster analysis, source apportionment

FUNDING:
South African Department of Environment, Forestry and Fisheries

The Vaal Triangle Airshed Priority Area (VTAPA), like other priority areas in South Africa, has an air pollution problem. Understanding the sources contributing to air pollution in this priority area will assist in the selection and implementation of appropriate control strategies. For this study, aerosol samples in the coarse ($PM_{10-2.5}$) and fine ($PM_{2.5}$) fraction were collected at four sites in the VTAPA during summer/autumn, winter, and spring. The contributing sources were identified and characterised based on the elemental and ionic compositions obtained through X-ray fluorescence and ion chromatography analysis. The highest seasonal median concentrations of $PM_{10-2.5}$ (116 μg/m^3) and $PM_{2.5}$ (88 μg/m^3) were observed in Sharpeville during the winter. The lowest median concentrations of $PM_{10-2.5}$ (25 μg/m^3) and $PM_{2.5}$ (18 μg/m^3) were detected in Zamdela during the summer/autumn period. At all sites, there was a high abundance of crustal elements in $PM_{10-2.5}$ and a dominance of coal and biomass combustion-related elements in $PM_{2.5}$. The Positive Matrix Factorisation receptor model identified dust-related and secondary aerosols as the major contributing sources of $PM_{10-2.5}$. $PM_{2.5}$ contributions were predominantly from coal burning for Sebokeng and Sharpeville and from industry, wood and biomass burning, and secondary aerosols for Kliprivier and Zamdela. The results of this study identify the main sources contributing to particulate air pollution in the VTAPA and provide local authorities with valuable information for decision-making.

Significance:

- Dust, industry, domestic coal burning, vehicles, and wood and biomass combustion are the key sources of particulate air pollution in the VTAPA that need to be prioritised by decision-makers.
- Although Sebokeng and Sharpeville are located within the vicinity of industries, domestic coal burning has a greater contribution to particulate loading at these sites.
- Results from this study will assist in the design of local municipality air quality management plans for the VTAPA.

Introduction

Over the past decades, South Africa has experienced strong economic growth, industrial expansion, and rapid urbanisation. This has led to the emergence of cities characterised by high population densities and high industrial and traffic activities. Air pollution is a serious environmental problem in these urban areas and has attracted widespread attention from the public as a result of its negative effects on humans.[1] Pollution from particulate matter (PM) is of primary concern in South Africa.[2] Exposure to PM is the fourth leading human health risk factor and is linked to over 5 million premature deaths all over the world.[3] Exposure to PM, especially $PM_{2.5}$, over long periods is dangerous to humans as inhaled particles will penetrate deep into the lungs and increase the risk of morbidity and premature mortality due to cardiopulmonary diseases and lung cancer.[4-6] Effective strategies are urgently needed to improve air quality and address the health risks associated with PM. Acquiring reliable and comprehensive information on the main sources of PM is the first key step required to achieve this.[7]

Source apportionment is an air quality management tool that can provide statistical information about source contributions which is important in the formulation of mitigation strategies for PM.[8-10] Attempts have been made in South Africa to apportion PM sources and their contributions. Engelbrecht et al.[11] used the Chemical Mass Balance model to compare PM source contributions from residential coal and low-smoke fuels used in the township of Qalabotjha. The Chemical Mass Balance model was also applied to identify the PM sources contributing to air pollution in Kwadela township.[12] Recently, Tshehla and Djolov[13] used the Positive Matrix Factorisation (PMF) receptor model to apportion PM sources in an industrialised rural area in the Limpopo Province. In the case of South Africa, where local source profiles are still lacking, the PMF model is a suitable alternative for the Chemical Mass Balance model as it does not require source profile data. The main sources of particulate (PM_{10} and $PM_{2.5}$) pollution identified from these source apportionment studies in South Africa include industries, residential solid fuel burning, vehicles, dust, and biomass burning.[13] However, despite attempts to identify air pollution sources in South Africa, studies reporting on PM sources are still lacking.[14] A thorough understanding of the different compositions and contributions of PM is required as it will assist air quality planners in assigning precedence to key pollutant sources.[7]

In 2006 the Vaal Triangle, a highly industrialised region, was classified as an air pollution priority area due to public health concerns over the elevated levels of air pollution faced in this region.[15] The major local sources found in the Vaal Triangle Airshed Priority Area (VTAPA) include industries, residential burning, vehicles, waste, and windblown dust.[16] These sources occur within close proximity to one another. In 2009, an Air Quality Management Plan detailing possible intervention strategies for the VTAPA was published.[17] The first 5-year cycle review of this Air Quality Management Plan in 2013 revealed that, despite efforts made, air pollutant concentrations were still above national ambient air quality standards.[18] This was due to inadequate implementation of air quality controls.[18]

The second 5-year cycle of the VTAPA Air Quality Management Plan is currently in review. Target air quality limits still have not been met as daily and annual average PM concentrations still remain above the national standards.[15] A source apportionment study is therefore required to establish an understanding of the current sources contributing to PM and identify opportunities for further emission reductions.

In this study, therefore, we sought to achieve the following objectives: (1) to explore the temporal and spatial variations of PM in the VTAPA; (2) to determine the elemental and ionic compositions of PM; and (3) to identify and apportion the main sources contributing to PM pollution.

Materials and methods

Sampling sites

The VTAPA is situated on the high central inland plateau of South Africa with terrain elevations ranging between 1300 m and 1900 m above sea level. The VTAPA stretches from the southern part of the Gauteng Province to the northern section of the Free State Province. The land use in this region includes commercial, industrial, residential, and low-intensity agricultural activities, all situated within close vicinity to one another. Four sites in the VTAPA were selected for this study. These sites (Figure 1) were selected based on a baseline assessment that identified these sites as ambient PM hotspot zones.[19] Sebokeng (26.5879S, 27.8410E), Sharpeville (26.6810S, 27.8677E) and Zamdela (26.8449S, 27.8551E) monitoring sites are situated inside densely populated low-income settlements, while Kliprivier (26.4203S, 28.0849E) site is in a low-density area.

Sampling strategy

Sampling was performed simultaneously at all of the sites for the summer/autumn (2 February – 9 March 2018 and 12 March –21 March 2018), winter (20 June – 6 July 2018), and spring (13 September – 21 September 2018 and 25 September – 3 October 2018) periods. Dichotomous low volume samplers (MicroPNS Type Dichoto LVS16, Umwelttechnik MCZ GmbH, Bad Nauheim, Germany) with split-flow rates of 1.7 L/min (for fine particles) and 15 L/min (for coarse particles) were employed for the simultaneous and sequential collection of particulate matter in the fine ($PM_{2.5}$) and coarse ($PM_{10-2.5}$) fraction on 47 mm PTFE Teflon filter membranes (2 μm pore size). Prior to being weighed, the filters were inspected for defections and then preconditioned in a stable environment for 24 h so as to allow for their weights to stabilise. The filters were then weighed three times before and after sampling using an XP26 DeltaRange Microbalance (Mettler-Toledo AG, Greifensee, and CH). The averaged mass difference (μg) was used together with the total volume of air sampled (m^3) to calculate the mass concentrations of the particles collected on each filter. These filters were stored in individual Petri slide dishes.

Two consecutive continuous 12-h samples for each size fraction were collected daily to enable comparisons between day (10:00 – 22:00) and night concentrations (22:00 – 10:00).[20-22] In total, 768 filters were sampled for the entire campaign. Laboratory blanks were used to determine the impact of laboratory procedure on the measured filter mass concentrations. Field blanks were utilised to determine the effect of sample handling and the filter itself on measured mass concentrations.

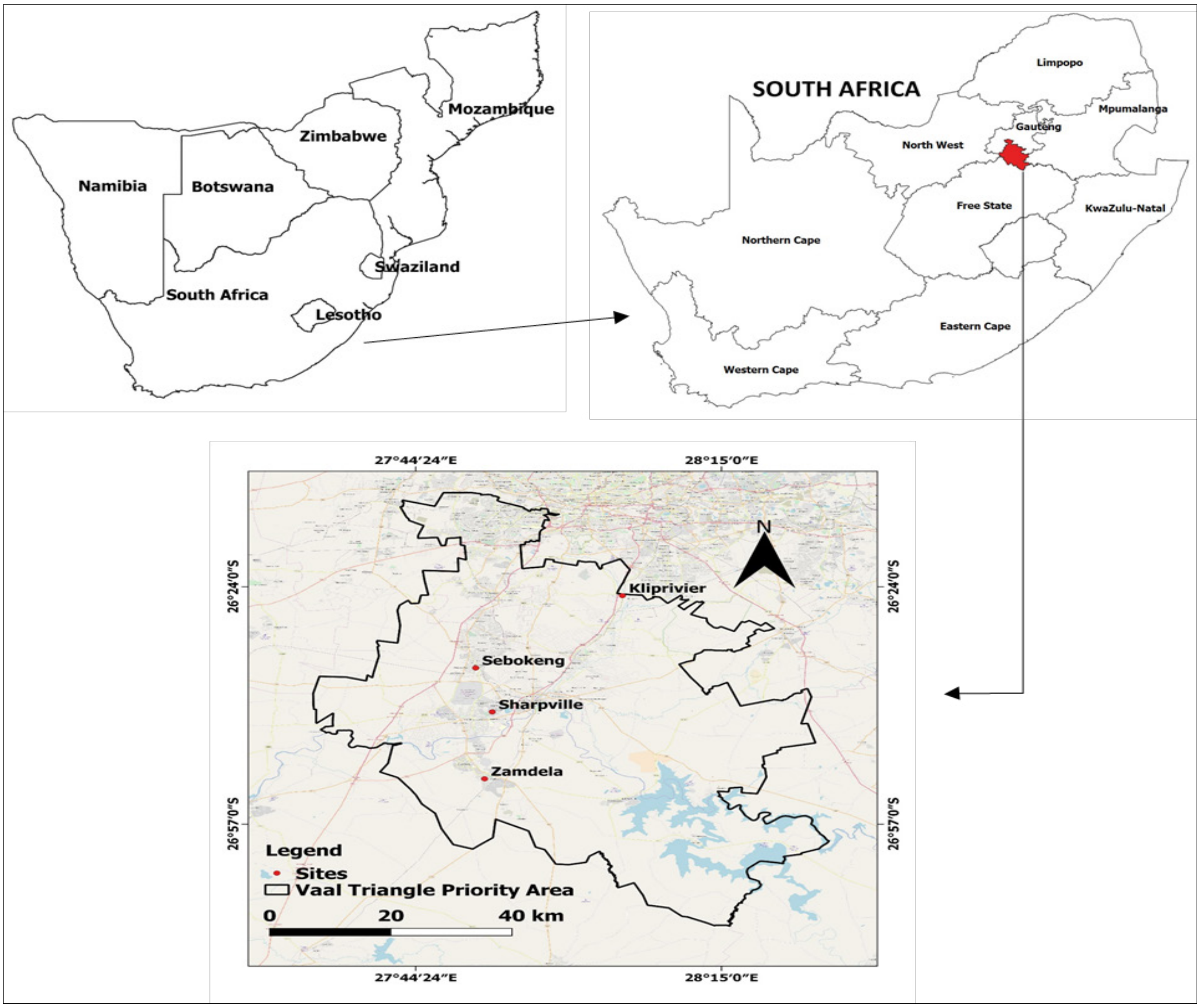


Figure 1: Study area map showing the location of the Vaal Triangle Airshed Priority Area (VTAPA) in South Africa and the four sampling sites within the VTAPA.

Chemical analyses

Elements

Trace elements on the Teflon filters were chemically analysed using X-ray fluorescence, which is a non-destructive procedure that allows for the analysis of filters without being subjected to any pre-treatment process. X-ray fluorescence involves the interplay between X-ray photons and the elements found in the PM species leading to the discharge of electrons, which will result in the release of X-rays that are unique for the individual element.[23] A wavelength dispersive X-ray fluorescence spectrometer was used for this analysis. The exposed Teflon samples were placed in filter holders and put into 47-mm stainless steel sample cups. These samples were then analysed by being exposed to an excitation condition in which X-rays produced from the spectrometer interact with atoms in the filters.[23] The following elements were detected using the spectrometer: Na, Mg, Al, Si, P, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Ni, Cu, Zn, and Pb. The measured concentrations of these elements were corrected using blanks.

Ions

Water-soluble ionic species on the Teflon filters were analysed using ion chromatography. A Dionex ICS-3000 system consisting of two flow lines was used for ion chromatography analysis.[24] One flow line was used for the detection of anion species and the other flow line to detect cation species. Before chemical analysis commenced, the filters were leached in 10 mL deionised water in an ultrasonic bath for 30 min. Five standards, ranging from 20 ppb to 500 ppb, were prepared using certified stock solutions obtained from Industrial Analytical (Johannesburg, South Africa). Filter samples were then analysed for the following water-soluble ionic species: F^-, Cl^-, SO_4^{2-}, NO_3^-, CH_3COO^-, $HCOO^-$, $C_2O_4^{2-}$, Na^+, NH_4^+, K^+, Mg_2^+, and Ca_2^+. In order to avoid contamination, this procedure was conducted in a stabilised room. The measured ion concentrations were corrected using blanks.

Meteorological data

Meteorological data from weather stations near the sampling sites were obtained for each site from the South African Weather Services (https://saaqis.environment.gov.za/). The meteorological variables used in this study include temperature (Temp), relative humidity (RH), wind speed (WS), and wind direction (WD). Wind roses (Supplementary figure 1) for each site and sampling season were generated using the Open Air package in R.

Positive Matrix Factorisation model analysis

$PM_{10\text{-}2.5}$ and $PM_{2.5}$ source contributions to ambient air particulate concentrations in the VTAPA were quantified using the US Environmental Protection Agency (EPA) PMF model. The PMF model is a multivariate factor analysis tool that deconstructs the matrix of speciated sample data into two matrices: factor contributions and factor profiles.[25] This is a well-tested receptor model that has been applied globally.[26-28] For this study, the PMF (Version 5.0) was performed to obtain quantitative source profiles and mass contributions. The PMF model equation can be expressed as follows:

$$X_{ij} = \Sigma_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \qquad \text{Equation 1}$$

where X_{ij} is the concentration of species *j* measured on sample *I*; *p* is the number of factors contributing to the samples; f_{kj} is the concentration of species *j* in factor profile *k*; g_{ik} is the relative contribution of factor *k* to sample *I*, and e_{ij} is the error of the PMF model for the species *j* measured on sample *I*.

In order to run PMF, the model requires sample chemical species concentration values and uncertainty estimates for each species. Uncertainty estimates were calculated by dividing the limit of quantification from the mass concentration for each species so as to obtain a fractional value. This can be expressed as:

$$F = \frac{LOQ}{M} \qquad \text{Equation 2}$$

where *F* is the fractional value, *LOQ* is the limit of quantification per species per exposed filter (μg/m³), and *M* is the species mass concentration (μg/m³).

An uncertainty is then assigned to each species based on the specific range into which the fractional value of a particular species falls.

The PMF model was run multiple times for all sites using elemental and ionic composition data for $PM_{10\text{-}2.5}$ and $PM_{2.5}$. The species used in the model were chosen according to the signal-to-noise (S/N) criterion. Species with S/N values greater than 2 were classified as 'strong', while those within the 0.2–1.9 range were categorised as 'weak'. Species with S/N values less than 0.2 were defined as 'bad' variables and were removed from the analysis. The optimal number of factors for each site was selected based on (1) knowledge of sources affecting the study area, (2) distributions of the scaled residuals and (3) the Qtrue/Qrobust ratio.[29] Species with symmetrically distributed scaled residuals within a range of -3 to +3 are indicative of a good model fit. The Qtrue/Qrobust ratio is useful in determining the influence of outliers on the model. A ratio above 1.5 indicates that outliers may have a disproportionate effect on the model and will need to be down weighted (Supplementary table 1).[30,31] The number of factors chosen for each site is shown in Table 1. A source type was assigned to each factor based on known representative indicator chemical species and source profiles obtained from the US EPA SPECIATE database (https://www.epa.gov/air-emissions-modeling/speciate-2).[11,32]

Table 1: Overview of the number of factors selected for Kliprivier, Sebokeng, Sharpeville, and Zamdela

	Kliprivier	Sebokeng	Sharpeville	Zamdela
$PM_{10\text{-}2.5}$	7	7	7	6
$PM_{2.5}$	5	7	5	5

Air mass origin

Back trajectory analysis was used to identify the transport pathways of air masses reaching the observation sites in the VTAPA. Five-day back trajectories were computed daily throughout each sampling period (summer/autumn, winter, and spring) as they can capture the pattern of pollutant transport from source regions to the study site.[21,33] Each trajectory was run every hour for 12 h so as to capture day and night-time air mass pathways. These trajectories were generated in the PC version of the HYSPLIT (Hybrid Single-Particle Lagrangian Integrated Trajectory) model using the Global Data Assimilation System's meteorological data set (spatial resolution of 1^o x 1^o), developed by the US National Center for Environmental Prediction.[34] A starting height of 500 m above ground level was chosen as it corresponds to trajectories near the ground.[21] The meteorological data were obtained from the US National Oceanic and Atmospheric Administration ftp server (ftp://arlftp.arlhq.noaa.gov/pub/archives/gdas1). The Sharpeville site was used as the reference point for the backward trajectories.

The trajectory cluster analysis tool in the HYSPLIT model was then used to group individual trajectories of similar air mass origins into clusters. Cluster analysis allows for air masses to be examined over time, whilst reducing the uncertainty effects related to long trajectories.[35] This tool employs an algorithm that utilises the latitudes and longitudes of hourly endpoints as input variables.[36] The number of clusters retained for this study was determined by the percentage change in the total spatial variance. Individual trajectories within each cluster were then averaged to produce cluster-mean trajectories.

Statistical analyses

Statistical tests were carried out to deduce whether $PM_{10\text{-}2.5}$ and $PM_{2.5}$ concentrations at each site differed significantly across seasons. Based on the Shapiro–Wilk's test, $PM_{10\text{-}2.5}$ and $PM_{2.5}$ concentrations were found to be not normally distributed. Therefore, the Kruskal–Wallis test was used for the variance analysis.

Results and discussion

Meteorological conditions

Table 2 presents the mean values of the meteorological conditions during the three sampling periods. Wind speed was relatively higher in summer/autumn and spring, and lower in winter. At all sites, relative humidity decreased significantly from summer/autumn to spring. Seasonal variations showed that temperature values at all sites were comparatively higher in spring and summer/autumn and lower in winter. The highest temperatures were experienced in February.

Spatial and temporal variations of $PM_{10\text{-}2.5}$ and $PM_{2.5}$ mass concentration

Statistical analysis of the temporal pattern of $PM_{10\text{-}2.5}$ concentrations (Table 3) at each sample site revealed significant seasonal differences for both day and night. It can be observed that the highest seasonal median values for $PM_{10\text{-}2.5}$ were experienced in Sharpeville during the winter season for both the day (95 μg/m^3) and night (116 μg/m^3) periods (Figure 2). Day and night $PM_{10\text{-}2.5}$ concentrations were significantly higher in spring and winter than in the summer/autumn period. There were no significant seasonal differences in $PM_{2.5}$ concentrations at each sample site. The highest seasonal $PM_{2.5}$ median values were observed in winter, with Sebokeng's maximum concentration (68 μg/m^3) occurring during the daytime and Sharpeville's peak (88 μg/m^3) during the night. The lowest seasonal median values for $PM_{10\text{-}2.5}$ (25 μg/m^3) and $PM_{2.5}$ (18 μg/m^3) were experienced in Zamdela during the summer/autumn period.

$PM_{10\text{-}2.5}$ and $PM_{2.5}$ chemical composition

The elemental and ionic contents of $PM_{10\text{-}2.5}$ and $PM_{2.5}$ for summer/autumn, winter, and spring at the four sampling sites are shown in Figures 3, 4, and 5, respectively. Statistical summaries of the elemental and ionic species for each site are given in Supplementary tables 2–13.

Elements

The elements Si, Mg, Al, Ca, Na, S, and Fe contributed the most towards $PM_{10\text{-}2.5}$ concentrations at all sites, both during the day and night for all seasons. These elements were highest in spring, which is a season associated with strong winds in the VTAPA (Supplementary figure 1). The abundance of Si, Mg, Al, Ca, Na, and Fe in $PM_{10\text{-}2.5}$ indicate that dust is dominant at these sites. There is also a strong presence of S in $PM_{2.5}$ for both day and night at all sites during all three seasons, implying that coal combustion could be an important contributor to atmospheric PM. There is a fairly high abundance of K and Zn in the $PM_{2.5}$ during the days and nights of winter and spring. These elements could have been emitted as a result of wood and biomass burning. Fe, Cr, and Ni were dominant in $PM_{2.5}$ during the day and night for Kliprivier (summer/autumn) and Zamdela (winter and spring). Fe, Cr, and Ni were also significant contributors of $PM_{2.5}$ during the summer/autumn nights at Kliprivier and Sebokeng. In Kliprivier, this could be a result of emissions from commercial heavy-duty vehicles operating on public roads from 20:00 to 06:00. The high concentrations of Fe, Cr, and Ni in Sebokeng could result from night-time operations at ArcelorMittal metallurgical industry.

Ions

Ionic compositions for $PM_{10\text{-}2.5}$ and $PM_{2.5}$ revealed that SO_4^{2-}, NH_4^+, and F^- are all dominant species for summer/autumn, winter and spring at all sites. The occurrence of these ionic species could be a result of coal combustion from industries, and to a lesser extent from residential solid fuel burning. The strong presence of NO_3^- in $PM_{10\text{-}2.5}$ and $PM_{2.5}$ during the day and night in winter could suggest an industrial origin. SO_4^{2-}, NH_4^+, F^-, and NO_3^- concentrations are highest in winter mainly as a result of increased coal combustion. High Na^+, Mg^{2+}, and Ca^{2+} abundances were observed for $PM_{10\text{-}2.5}$, suggesting the possibility of marine and crustal origin sources. Wood and biomass burning is also an important source, as indicated by the abundance of K^+ in $PM_{2.5}$, especially in winter where it is highest due to the need for space heating.

Apportionment of sources identified by PMF

The potential sources that were identified using the PMF model are industry, coal burning, wood and biomass burning, waste burning, dust-related, vehicles, secondary aerosols, and aged sea salt. Figure 6 presents the source apportionment results for the four sampled sites and shows the variations in contributions based on the three sampling periods.

$PM_{10\text{-}2.5}$ and $PM_{2.5}$ sources

The industry source is typically characterised by strong contributions from Zn, Fe, Pb, Ni, Cr, Mn, and V. These elements are usually associated with smelters and metallurgical industries.[37] The metal element V is mainly associated with heavy fuel oil combustion.[38] Coal, coking coal, and heavy fuel oil are the main energy sources that drive industries in the VTAPA.[17] Coal burning is an important source identified through PMF. This source is generally associated with burning in low-income households and industries. The coal-burning source is highly loaded with Cl^-. This ion is mainly from ammonium chloride (NH_4Cl), which occurs as a result of the rapid reaction between HCl and NH_4^+ in the atmosphere.[39] With coal being the primary energy source in South Africa, coal burning can be regarded as the largest potential source of HCl.

Wood and biomass burning are characterised by a high content of K^+ and minor contributions of SO_4^{2-} and NO_3^-. K^+ is widely recognised as an indicator of biomass burning as it is released during the plant combustion process.[31] For the low-income settlements in the VTAPA, K^+ is more representative of wood combustion as wood is an important energy source for cooking and space heating in these settlements.[15] Biomass burning occurs in the VTAPA through the burning of open spaces used for agricultural activities.[17]

Refuse collection in the low-income settlements of South Africa is infrequent and has resulted in the pile-up of solid waste into heaps. As a measure to reduce these heaps, residents have resorted to burning waste.[15] The waste burning source identified through PMF contained high values of NH_4^+ and small contributions from Cl^- and K^+. The occurrence of Cl^- could be as a result of the presence of salt-containing foodstuffs and chlorine-based materials in domestic waste disposals. Residents in low-income settlements often do not sort their waste before disposal, which can result in domestic waste being mixed with garden waste, thus likely explaining the presence of K^+ in the waste burning source.

The dust-related source was identified in the coarse fraction and was characterised by crustal elements which included Ca, Mg, Si, Al, Fe, Ti, and Mn. This source could have been generated locally through resuspension of soil and construction works as well as through regional transportation of dust aerosols.[40] The elements found in the dust-related source could also be associated with resuspended dust resulting from motor vehicle entrainment on unpaved roads in low-income settlements. Other metals such as Cr and V were also present in the dust-related source and could be as a result of soil contamination from industrial emissions.[41] NO_3^-, Pb, Zn, Mn, and Fe were characteristic of the vehicle source. The Zn, Mn, and Fe metal elements found in this source are associated with both petrol- and diesel-fueled vehicles.[42] These elements are also associated with brake, tear, and engine wear.[43] Zn is a common additive found in lubricating oils and can be emitted through combustion by diesel engines.[38] Fe is generally found in catalysts used for petrol fuel combustion.[44]

The secondary aerosol source mainly consisted of SO_4^{2-}, NO_3^- and NH_4^+, formed through the chemical transformation of SO_2, $NO_{x,}$ and NH_3 pollutants originating from other direct sources. The presence of secondary aerosols in the VTAPA could also be a result of long-range transportation. The aged sea salt source was characterised by high loadings of Na and low Cl levels. The lack of Cl in the aged sea salt source is a result of Cl displacement in sea salt particles by acidic pollutants (H_2SO_4 and HNO_3) leading to the formation of sulfate and nitrate salts.[45] The long distances travelled by air masses (transporting sea salt) from the sea to the study site could also have resulted in the loss of Cl along their trajectories.

Table 2: Average values of temperature (Temp, °C), relative humidity (RH, %) and wind speed (Ws, m/s) during the summer/autumn, winter and summer campaigns

Site	Variable	Summer/autumn		Winter		Spring	
		February	March	June	July	September	October
Kliprivier	Temp	20.18	18.77	8.89	8.76	17.5	19.53
	RH	63.18	63.54	50.99	51.39	33.27	38.44
	Ws	1.73	1.79	1.61	1.32	2.75	2.21
Sebokeng	Temp	20.89	20.04	12.12	11.17	19.2	20.49
	RH	62	60.37	40.95	44.53	26.94	36.41
	Ws	2.74	2.72	2.14	2.27	3.55	3.14
Sharpeville	Temp	20.83	19.86	11.07	10.54	18.88	20.11
	RH	62.56	62.3	48.23	48.45	28.28	39.16
	Ws	2.48	2.39	1.97	1.94	3.11	2.61
Zamdela	Temp	20.69	19.46	11.08	10.03	18.14	20.09
	RH	63.01	63.07	45.65	48.9	30.88	37.18
	Ws	2.53	2.4	1.75	1.83	–[a]	2.64

[a] *Wind speed data are not available for Zamdela for the month of September.*

Table 3: Seasonal difference of day and night PM$_{10-2.5}$ and PM$_{2.5}$ at each sampling site

Site	PM$_{10-2.5}$ day		PM$_{10-2.5}$ night		PM$_{2.5}$ day		PM$_{2.5}$ night	
	Chi-square	*p*-value	Chi-square	*p*-value	Chi-square	*p*-value	Chi-square	*p*-value
Kliprivier	6.487	0.03903	23.541	<0.001	3.159	0.206	4.145	0.126
Sebokeng	16.503	<0.001	10.379	0.006	5.125	0.077	5.656	0.059
Sharpeville	16.319	<0.001	18.697	<0.001	3.565	0.168	2.297	0.317
Zamdela	9.804	0.007	15.789	<0.001	1.317	0.518	0.955	0.620

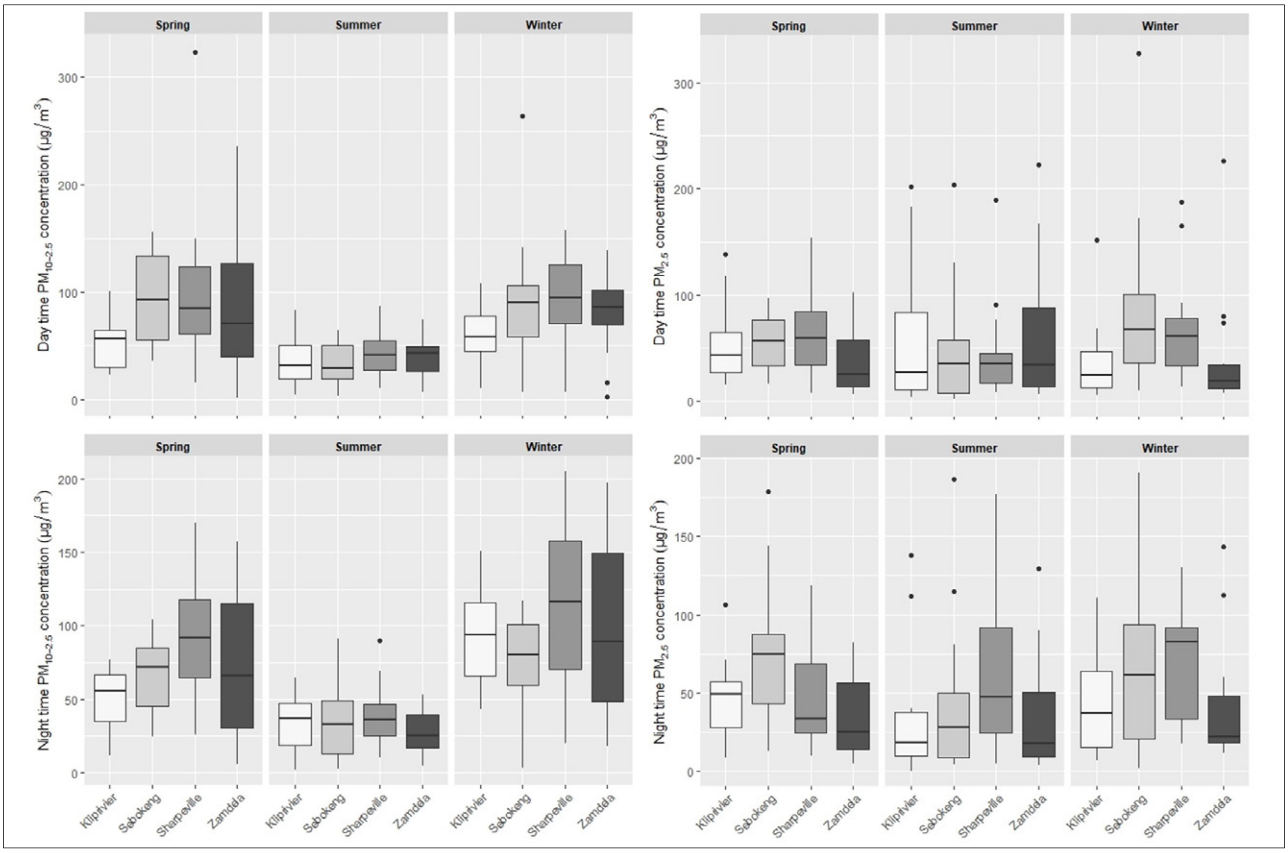


Figure 2: Day and night-time seasonal range of PM$_{10-2.5}$ and PM$_{2.5}$ mass concentration at all the sampling sites (16 observations per site).

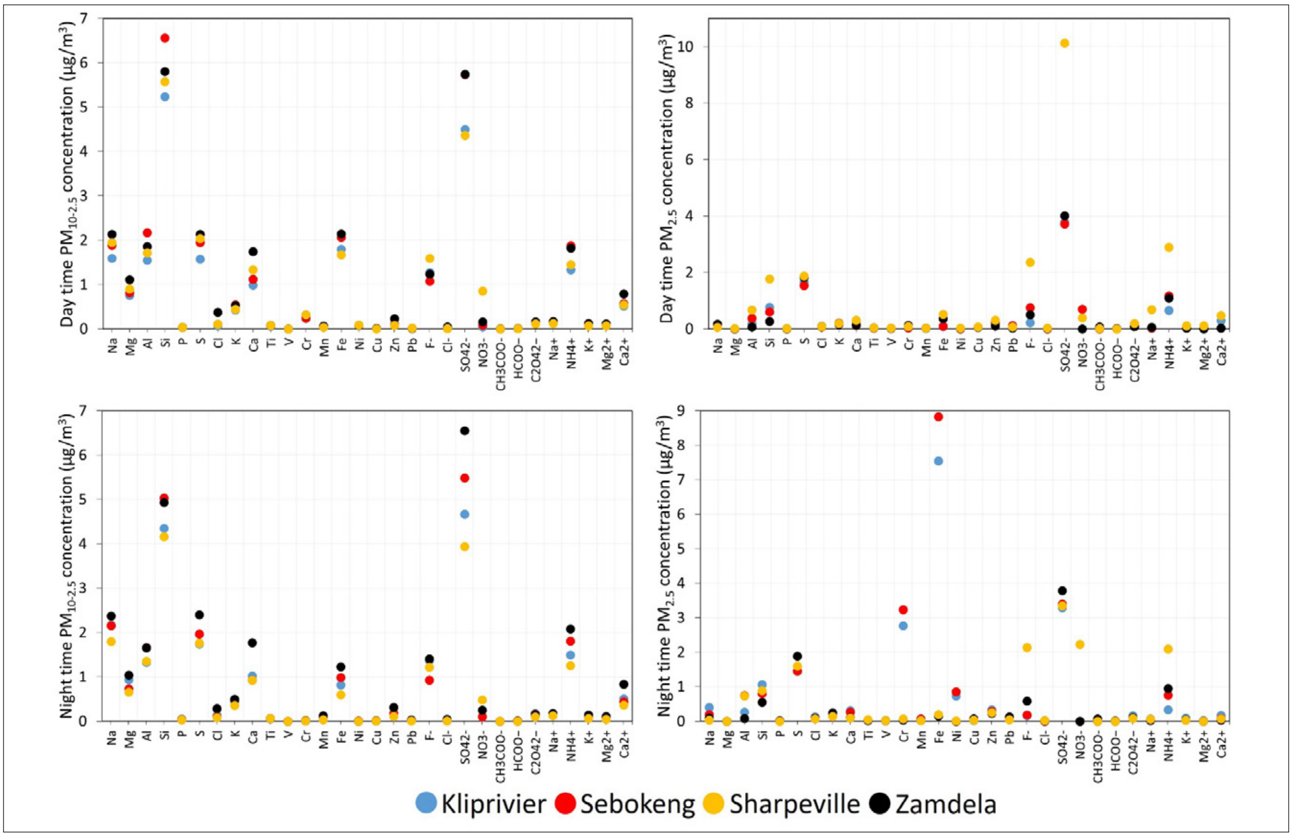


Figure 3: Summer day (top left and right) and night-time (bottom left and right) average elemental and ionic composition of $PM_{10\text{-}2.5}$ and $PM_{2.5}$ at all the sampling sites.

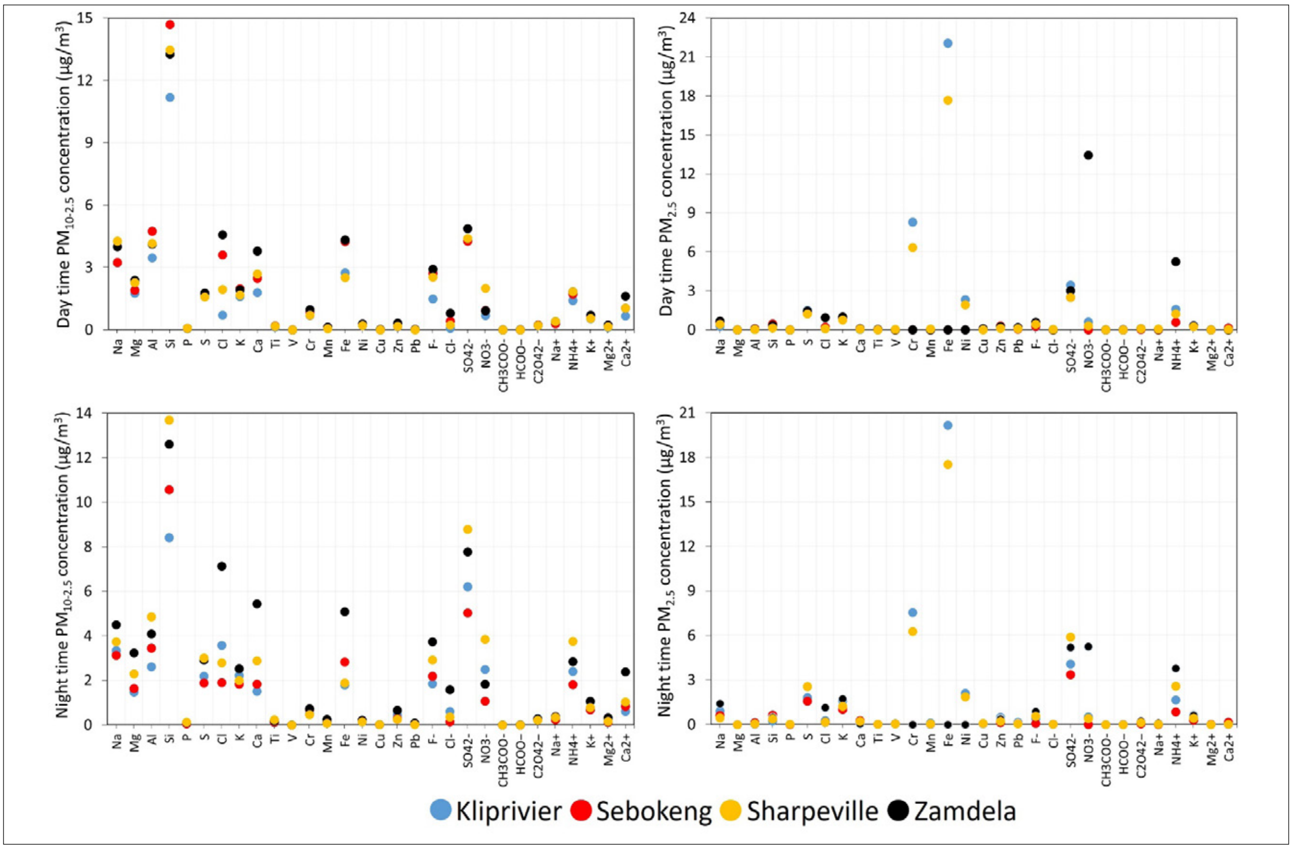


Figure 4: Winter day (top left and right) and night-time (bottom left and right) average elemental and ionic composition of $PM_{10\text{-}2.5}$ and $PM_{2.5}$ at all the sampling sites.

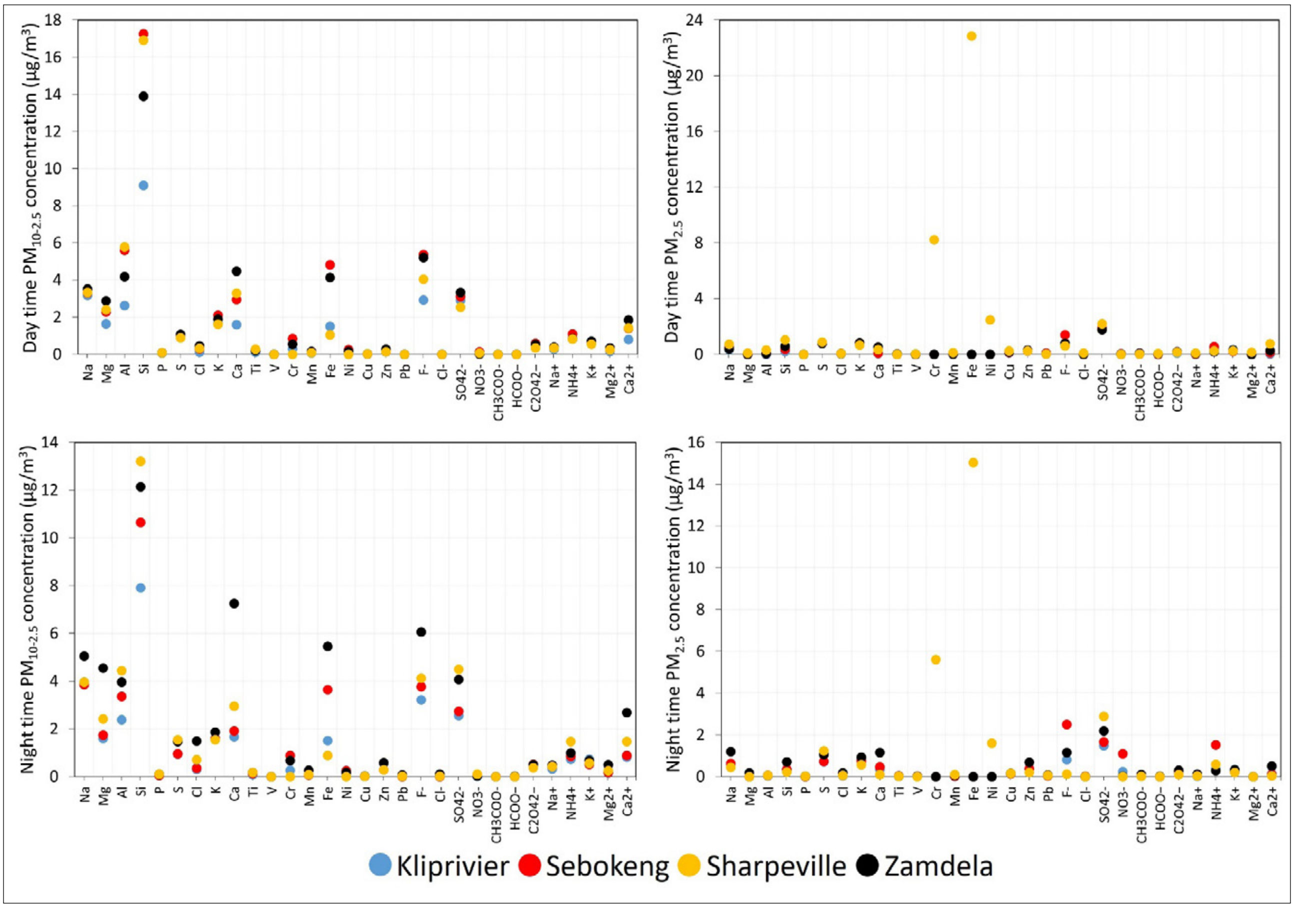


Figure 5: Spring day (top left and right) and night-time (bottom left and right) average elemental and ionic composition of $PM_{10-2.5}$ and $PM_{2.5}$ at all the sampling sites.

Seasonal contributions

Source apportionment results reveal that for $PM_{10-2.5}$, dust-related is a major source at Kliprivier (32–52%), Sebokeng (31–68%), Sharpeville (34–49%), and Zamdela (19–65%). Dust-related contributions show relatively higher concentrations in summer/autumn and spring, and lower concentrations in winter. Secondary aerosols have an important contribution at Kliprivier (12–32%), Sebokeng (11–14%), Sharpeville (12–35%), and Zamdela (29–32%). The seasonal variations showed that contributions of secondary aerosols were relatively higher in spring and summer/autumn and lower in winter. For summer/autumn this could mainly be as a result of regional transportation from the industrial region of Mpumalanga, whilst for spring, secondary aerosols could be from the intensive agricultural region of the Free State Province.[40,46]

Coal combustion and vehicles are sources prominent in the coarse fraction. Coal combustion accounts for 4% to 19%, 6% to 19%, 8% to 14%, and 7% to 16% of $PM_{10-2.5}$ mass concentrations in Kliprivier, Sebokeng, Sharpeville, and Zamdela, respectively. Vehicles account for 11% to 20%, and 16% to 25% of $PM_{10-2.5}$ mass concentrations in Kliprivier and Sebokeng, respectively. Vehicles contributed to 10% of $PM_{10-2.5}$ mass concentrations for both summer/autumn and winter at Sharpeville. In winter, vehicles contributed 14% of $PM_{10-2.5}$ mass concentrations at Zamdela.

Coal burning, secondary aerosols, wood and biomass burning, and industries are the key PM sources in the fine fraction. Coal burning is the main source of $PM_{2.5}$ air pollution in Sebokeng and Sharpeville, contributing over 60% for all three seasons with the highest concentrations being experienced in winter. These results are expected as domestic fuel combustion in low-income settlements is higher during winter due to the high demand for space heating.[47] Secondary aerosols are a key $PM_{2.5}$ source in Zamdela with contributions ranging from 24% to 67%. These secondary aerosols are likely to have an industrial origin as Sasol Chemical Industries Complex is located within the vicinity of Zamdela. Secondary aerosols are also an important $PM_{2.5}$ source at Kliprivier. The contribution from secondary aerosols for all three seasons in Kliprivier varied from 17% to 22%. The presence of secondary aerosols at Kliprivier is likely to be from coal-fired power stations. This site is impacted by pollution originating outside the designated boundaries of the VTAPA. Industries account for 5% to 11%, 7% to 14%, 10% to 12%, and 18% to 35% of $PM_{2.5}$ mass concentrations in Kliprivier, Sebokeng, Sharpeville, and Zamdela, respectively.

Wood and biomass burning is an important source identified in both fractions, accounting for 15% to 25% and 6% to 14% of $PM_{10-2.5}$ mass concentrations in Sharpeville and Zamdela, respectively. The higher contributions in spring for both Sharpeville and Zamdela are consistent with the biomass burning patterns in South Africa, with biomass burning occurring during late winter and early spring.[48] Wood and biomass burning accounts for 26% and 17% of $PM_{10-2.5}$ mass concentrations in Kliprivier and Sebokeng, respectively. In the fine fraction, wood and biomass burning accounts for 72% to 84%, 2% to 13%, 4% to 6%, and 32% to 49% of PM mass concentrations in Kliprivier, Sebokeng, Sharpeville, and Zamdela, respectively. For Kliprivier, Sebokeng, and Sharpeville, the concentrations of the wood and biomass burning source were highest in spring as extensive biomass burning activities take place during August and September.[40] Regional transportation also plays a significant role during the same period as biomass burning emissions originating from Zambia, Angola, Mozambique, and Zimbabwe are transported to South Africa.[48] In Zamdela, the concentrations of wood and biomass burning in the fine fraction were highest in winter and this could be due to the extensive use of wood by households for space heating. Wood is the main solid fuel source for cooking and space heating in Zamdela.[49] Waste burning is an important source of PM in the fine fraction at Zamdela during the summer/autumn period. This source accounted for 20% of PM in the fine fraction.

Seasonality of air masses

Table 4 gives a summary of the cluster means and their associated trajectories for each sampling period. As shown in Figure 7, there are three types of air masses associated with the summer/autumn period. The first type of air mass originates from Mozambique and along its pathway passes through mining and industrial areas in the Mpumalanga region. This air mass accounts for the majority (55%) of trajectories arriving at the study site. The high concentrations of $PM_{2.5}$ observed in Kliprivier, Sharpeville, and Zamdela during the summer/autumn period might be influenced by this air mass. The second (36%) and third (9%) type of air mass originate from the Indian and south Atlantic Ocean, respectively, and both pass through the Mpumalanga region along their trajectories. These air masses are potential contributors of aged sea salt and secondary aerosols in the VTAPA. In winter, three major air masses were identified. Air mass 1, which accounts for 75% of the trajectories, originated within northern South Africa and passed through Botswana and the mining areas of the North West Province via a short pathway, suggesting contributions from both local and regional pollutant sources. This air mass could also account for the high concentrations of $PM_{10-2.5}$ observed at all sites during winter. Air masses 2 and 3 (accounting for 9% and 16% of trajectories, respectively) arrived from the southwest direction, from the south Atlantic Ocean.

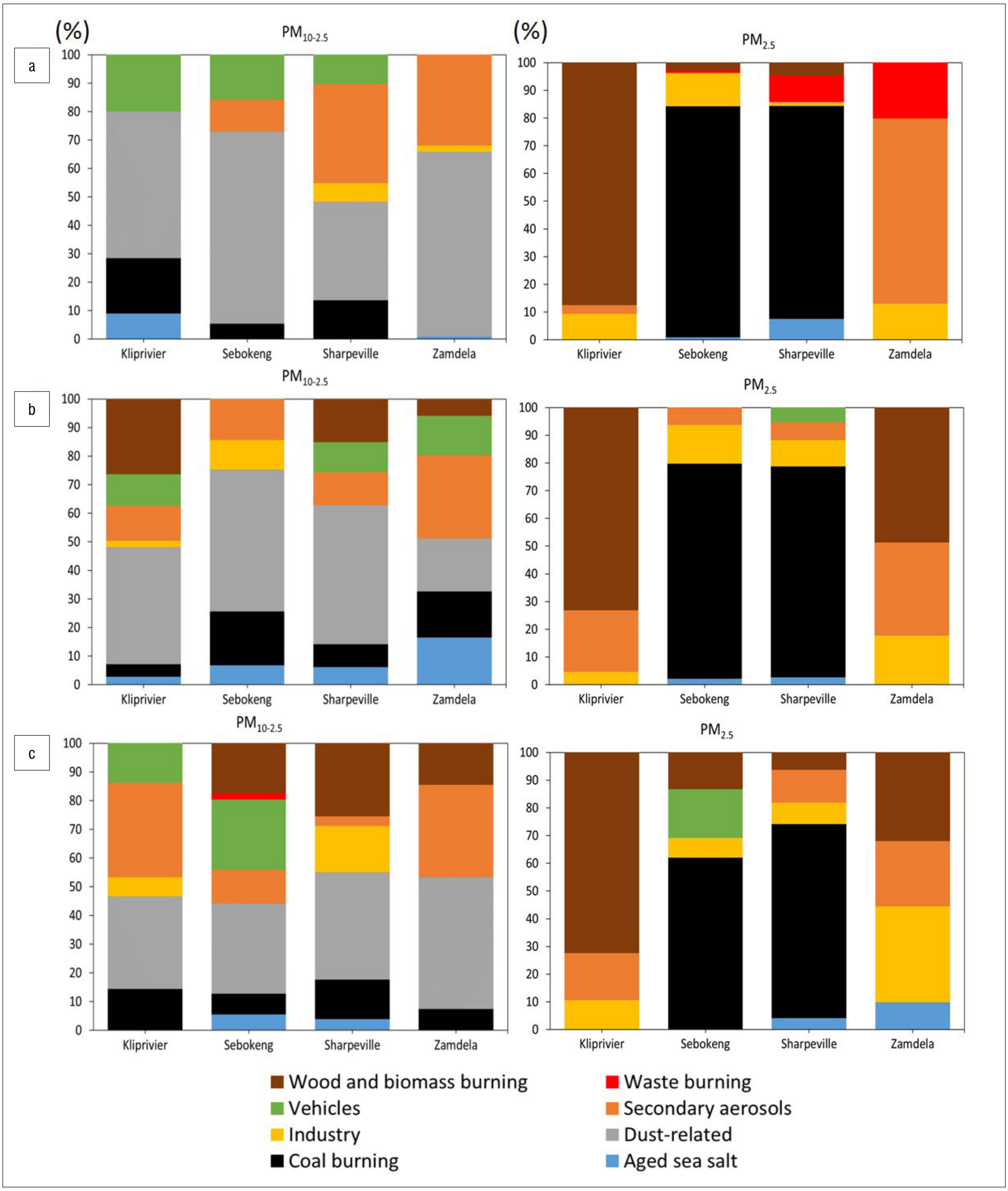


Figure 6: Source contributions for $PM_{10-2.5}$ and $PM_{2.5}$ at all sites for (a) summer/autumn (b) winter and (c) spring.

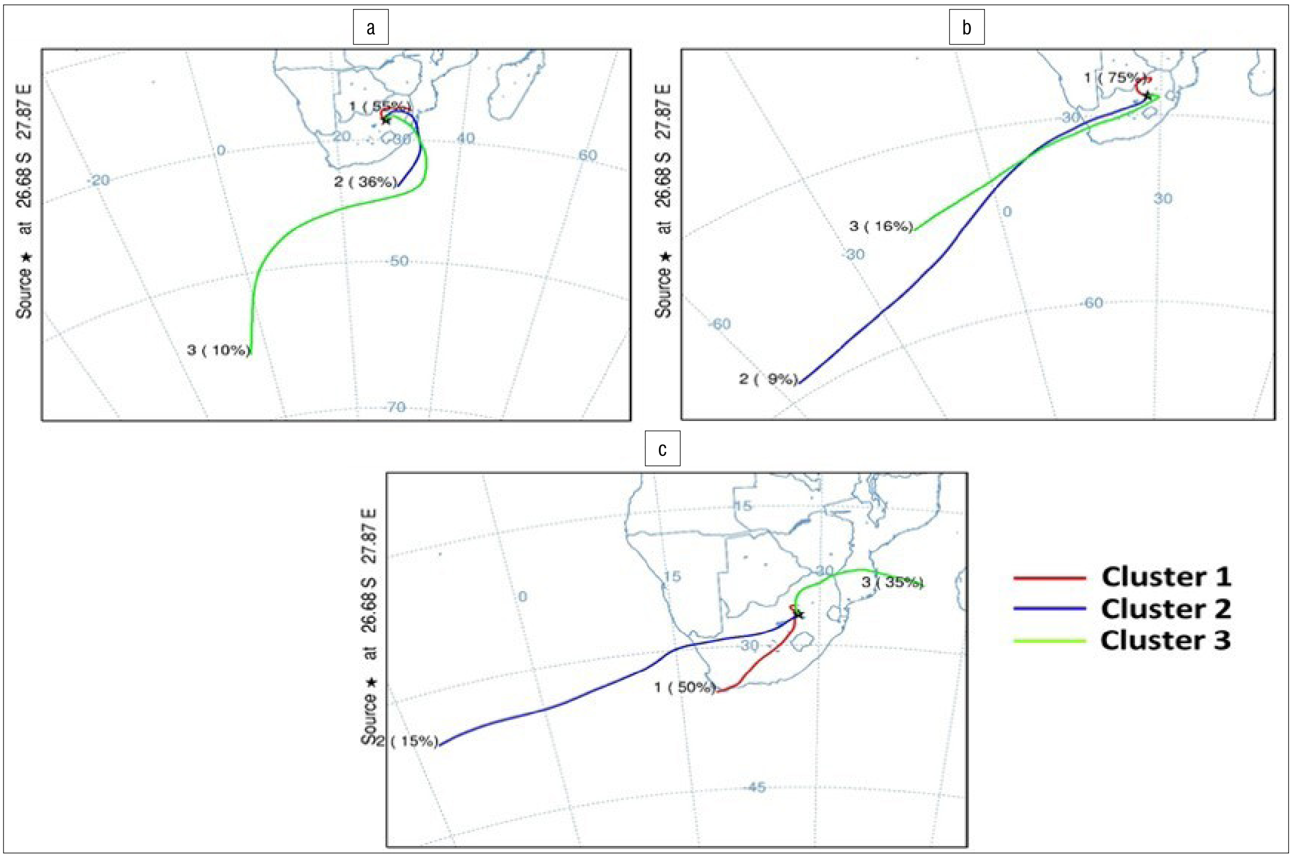


Figure 7: Backward trajectory cluster means for the (a) summer/autumn, (b) winter and (c) spring sampling periods.

Table 4: Cluster means and their trajectories for summer/autumn, winter and spring

Season	Cluster number	Number of trajectories
Summer/autumn	1	23
Summer/autumn	2	15
Summer/autumn	3	4
Winter	1	24
Winter	2	3
Winter	3	5
Spring	1	20
Spring	2	6
Spring	3	14

In spring, the study site is influenced by three major air masses. The first air mass, which accounts for 50% of the total trajectories, originates from the Indian Ocean and travels in a northeasterly direction, passing through the agricultural region of the Free State before arriving at the study site. This air mass could have contributed to the high concentrations of $PM_{10-2.5}$ and $PM_{2.5}$ observed at all the sample sites during the spring period. The second air mass (which accounts for 15% of trajectories) begins in the south Atlantic Ocean, crosses through the Northern Cape, and then passes over the mining areas of the North West Province, making it a potential contributor of aged sea salt and secondary aerosols in the VTAPA. The third air mass – with 35% of trajectories – originates from the Indian Ocean, crosses into Mozambique, and passes through the mining region of Limpopo along its pathway, suggesting contributions from both local and regional pollutant sources.

Conclusion

$PM_{10-2.5}$ and $PM_{2.5}$ aerosol samples were collected for three seasons at four sites in the VTAPA industrial/urban region and were chemically analysed. Elemental and ionic compositions for these samples show an abundance of crustal elements in $PM_{10-2.5}$ and a predominance of coal and biomass combustion-related elements in $PM_{2.5}$ at all sites. Eight sources of $PM_{10-2.5}$ and $PM_{2.5}$ were resolved and identified using the PMF model and include industry, coal burning, wood and biomass burning, waste burning, dust-related, vehicles, secondary aerosols, and aged sea salt. In the coarse fraction, dust-related and secondary aerosols were the major contributing sources. In the fine fraction, secondary aerosols, coal burning, industry and wood and biomass burning were the main sources of PM.

The present study has demonstrated the importance of source apportionment as a tool in the management of air quality management in the townships of the VTAPA. For Kliprivier, appropriate abatement strategies should focus on reducing emissions from dust, wood and biomass burning, and vehicles. The main emission sources to target in Sebokeng are dust, vehicles, and domestic coal burning. In Sharpeville, the focus should be on reducing emissions from domestic coal burning, dust, industry, and vehicles. Abatement strategies in Zamdela should focus on industry, wood and biomass burning, and dust emission sources. Reducing the strength of these sources will benefit residents in the VTAPA by lowering PM exposure and improving air quality.

Acknowledgements

This research was supported by a public/private partnership which included the South African Department of Environmental Affairs (now the Department of Environment, Forestry and Fisheries) and industry stakeholders in the VTAPA, and was coordinated and implemented by the North-West University Climatology Research Group. The views expressed are those of the authors and not necessarily the funding department. We thank Joe Malahlela for his assistance in the field.

Competing interests

We declare that there are no competing interests.

Authors' contributions

The majority of the work was conducted by L.M. who was responsible for the investigation, data analysis and writing (original draft and conceptualisation). S.J.P. was responsible for funding acquisition, supervision and writing (reviewing and editing). R.P.B. was responsible for data curation and writing (reviewing and editing). B.L. was responsible for analysing the data and writing (reviewing and editing). J.P.B. was responsible for analysing the data. P.G.v.Z. was responsible for analysing the data. All authors agreed to the submission of the manuscript.

References

1. Amegah AK, Agyei-Mensah S. Urban air pollution in sub-Saharan Africa: Time for action. Environ Pollut. 2017;220:738–743. https://doi.org/10.1016/j.envpol.2016.09.042
2. Altieri KE, Keen SL. Public health benefits of reducing exposure to ambient fine particulate matter in South Africa. Sci Tot Environ. 2019;684:610–620. https://doi.org/10.1016/j.scitotenv.2019.05.355
3. Bhanarkar AD, Purohit P, Rafaj P, Amann M, Bertok I, Cofala J, et al. Managing future air quality in megacities: Co-benefit assessment for Delhi. Atmos Environ. 2018;186:158–177. https://doi.org/10.1016/j.atmosenv.2018.05.026
4. Norman R, Cairncross E, Witi J, Bradshaw D, South African Comparative Risk Assessment Collaborating Group. Estimating the burden of disease attributable to urban outdoor air pollution in South Africa in 2000. S Afr Med J. 2007;97(7):782–790.
5. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: A review of the effects of particulate matter air pollution on human health. J Med Toxicol. 2012;8(2):166–175. https://doi.org/10.1007/s13181-011-0203-1
6. Feng S, Gao D, Liao F, Zhou F, Wang X. The health effects of ambient $PM_{2.5}$ and potential mechanisms. Ecotoxicol Environ Saf. 2016;128:67–74. https://doi.org/10.1016/j.ecoenv.2016.01.030
7. Thunis P, Clappier A, Tarrason L, Cuvelier C, Monteiro A, Pisoni E, et al. Source apportionment to support air quality planning: Strengths and weaknesses of existing approaches. Environ Int. 2019;130:1–13. https://doi.org/10.1016/j.envint.2019.05.019
8. Gupta AK, Karar K, Srivastava A. Chemical mass balance source apportionment of PM_{10} and TSP in residential and industrial sites of an urban region of Kolkata, India. J Hazard Mater. 2007;142(1–2):279–287. https://doi.org/10.1016/j.jhazmat.2006.08.013
9. Hopke PK. The use of source apportionment for air quality management and health assessments. J Toxicol Environ Health A. 2008;71(9–10):555–563. https://doi.org/10.1080/15287390801997500
10. Zhu Y, Huang L, Li J, Ying Q, Zhang H, Liu X, et al. Sources of particulate matter in China: Insights from source apportionment studies published in 1987-2017. Environ Int. 2018;115:343–357. https://doi.org/10.1016/j.envint.2018.03.037
11. Engelbrecht JP, Swanepoel L, Chow JC, Watson JG, Egami RT. The comparison of source contributions from residential coal and low-smoke fuels, using CMB modeling, in South Africa. Environ Sci Policy. 2002;5(2):157–167. https://doi.org/10.1016/S1462-9011(02)00029-1
12. Van den Berg B. Source apportionment of ambient particulate matter in Kwadela, Mpumalanga [MSc dissertation]. Potchefstroom: North-West University; 2015.
13. Tshehla C, Djolov G. Source profiling, source apportionment and cluster transport analysis to identify the sources of PM and the origin of air masses to an industrialised rural area in Limpopo. Clean Air J. 2018;28(2):54–66. https://doi.org/10.17159/2410-972x/2018/v28n2a18
14. Mathuthu M, Dudu VP, Manjoro M. Source apportionment of air particulates in South Africa: A review. Atmos Clim Sci. 2019;9:100–113.
15. South African Department of Environmental Affairs (DEA). The second generation Vaal Triangle Airshed Priority Area Air Quality Management Plan : Draft baseline assessment report. Pretoria: DEA; 2019.
16. South African Department of Environmental Affairs (DEA). The benefits and costs of air quality management. Pretoria: DEA; 2018.
17. South African Department of Environmental Affairs and Tourism (DEAT). Vaal Triangle Air-Shed Priority Area Air Quality Management Plan. Vol. GN32263, Government Gazette. Pretoria: DEAT; 2009. p. 2–239.
18. South African Department of Environmental Affairs (DEA). The Medium Term Review of the 2009 Vaal Triangle Airshed Priority Area: Air Quality Management Plan. Pretoria: DEA; 2013.
19. Thomas RG. An air quality baseline assessment for the Vaal airshed in South Africa [MSc dissertation]. Pretoria: University of Pretoria; 2008. https://repository.up.ac.za/handle/2263/28444
20. Jalava PI, Wang Q, Kuuspalo K, Ruusunen J, Hao L, Fang D, et al. Day and night variation in chemical composition and toxicological responses of size segregated urban air PM samples in a high air pollution situation. Atmos Environ. 2015;120:427–437. https://doi.org/10.1016/j.atmosenv.2015.08.089
21. Chandra S, Kulshrestha MJ, Singh R, Singh N. Chemical characteristics of trace metals in PM_{10} and their concentrated weighted trajectory analysis at Central Delhi, India. J Environ Sci (China). 2017;55:184–96. https://doi.org/10.1016/j.jes.2016.06.028
22. Hao Y, Deng S, Yang Y, Song W, Tong H, Qiu Z. Chemical composition of particulate matter from traffic emissions in a road tunnel in Xi'an, China. Aerosol Air Qual Res. 2019;19(2):234–246. https://doi.org/10.4209/aaqr.2018.04.0131
23. Research Triangle Institute. Standard operating procedure for the X-Ray fluorescence analysis of particulate matter deposits on teflon filters [document on the Internet]. c2009 [cited 2018 May 20]. Available from: www3.epa.gov/ttnamti1/files/ambient/pm25/spec/pmxrfsop.pdf
24. Conradie EH, Van Zyl PG, Pienaar JJ, Beukes JP, Galy-Lacaux C, Venter AD, et al. The chemical composition and fluxes of atmospheric wet deposition at four sites in South Africa. Atmos Environ. 2016;146:113–131. https://doi.org/10.1016/j.atmosenv.2016.07.033
25. Paatero P, Eberly S, Brown SG, Norris GA. Methods for estimating uncertainty in factor analytic solutions. Atmos Measure Tech. 2014;7(3):781–797. https://doi.org/10.5194/amt-7-781-2014
26. Bove MC, Brotto P, Calzolai G, Cassola F, Cavalli F, Fermo P, et al. PM_{10} source apportionment applying PMF and chemical tracer analysis to ship-borne measurements in the Western Mediterranean. Atmos Environ. 2016;125:140–151. https://doi.org/10.1016/j.atmosenv.2015.11.009
27. Chuang MT, Chen YC, Lee C Te, Cheng CH, Tsai YJ, Chang SY, et al. Apportionment of the sources of high fine particulate matter concentration events in a developing aerotropolis in Taoyuan, Taiwan. Environ Pollut. 2016;214:273–281. https://doi.org/10.1016/j.envpol.2016.04.045
28. Crilley LR, Lucarelli F, Bloss WJ, Harrison RM, Beddows DC, Calzolai G, et al. Source apportionment of fine and coarse particles at a roadside and urban background site in London during the 2012 summer ClearfLo campaign. Environ Pollut. 2017;220:766–778. https://doi.org/10.1016/j.envpol.2016.06.002
29. Vossler T, Černikovský L, Novák J, Williams R. Source apportionment with uncertainty estimates of fine particulate matter in Ostrava, Czech Republic using Positive Matrix Factorization. Atmos Pollut Res. 2016;7(3):503–512. https://doi.org/10.1016/j.apr.2015.12.004
30. Gupta I, Salunkhe A, Kumar R. Source apportionment of PM_{10} by positive matrix factorization in urban area of Mumbai, India. Sci World J. 2012;2012:1–13. https://doi.org/10.1100/2012/585791
31. Weber S, Salameh D, Albinet A, Alleman LY, Waked A, Besombes JL, et al. Comparison of PM_{10} sources profiles at 15 french sites using a harmonized constrained positive matrix factorization approach. Atmosphere. 2019;10(6):1–22. https://doi.org/10.3390/atmos10060310
32. Simon H, Beck L, Bhave PV, Divita F, Hsu Y, Luecken D, et al. The development and uses of EPA's SPECIATE database. Atmos Pollut Res. 2010;1(4):196–206. https://doi.org/10.5094/APR.2010.026
33. Nyanganyura D, Makarau A, Mathuthu M, Meixner FX. A five-day back trajectory climatology for Rukomechi research station (northern Zimbabwe) and the impact of large-scale atmospheric flows on concentrations of airborne coarse and fine particulate mass. S Afr J Sci. 2008;104(1–2):43–52.

34. Stein AF, Draxler RR, Rolph GD, Stunder BJB, Cohen MD, Ngan F. Noaa's hysplit atmospheric transport and dispersion modeling system. Bull Am Meteorol Soc. 2015;96(12):2059-77. https://doi.org/10.1175/BAMS-D-14-00110.1

35. Donnelly AA, Broderick BM, Misstear BD. The effect of long-range air mass transport pathways on PM_{10} and NO_2 concentrations at urban and rural background sites in Ireland: Quantification using clustering techniques. J Environ Sci Health A. 2015;50(7):647–658. https://doi.org/10.1080/10934529.2015.1011955

36. Draxler R, Stunder B, Rolph G, Stein A, Taylor A. HYSPLIT5 user's guide version 4 – Last revision: April 2020 [document on the Internet]. c2020 [cited 2020 Jul 07. Available from: https://www.arl.noaa.gov/documents/reports/hysplit_user_guide.pdf

37. Dall'Osto M, Querol X, Alastuey A, O'Dowd C, Harrison RM, Wenger J, et al. On the spatial distribution and evolution of ultrafine particles in Barcelona. Atmos Chem Phys. 2013;13(2):741–759. https://doi.org/10.5194/acp-13-741-2013

38. Yu L, Wang G, Zhang R, Zhang L, Song Y, Wu B, et al. Characterization and source apportionment of $PM_{2.5}$ in an urban environment in Beijing. Aerosol Air Qual Res. 2013;13(2):574–583. https://doi.org/10.4209/aaqr.2012.07.0192

39. Chen WN, Chen YC, Kuo CY, Chou CH, Cheng CH, Huang CC, et al. The real-time method of assessing the contribution of individual sources on visibility degradation in Taichung. Sci Tot Environ. 2014;497–498(110):219–228. https://doi.org/10.1016/j.scitotenv.2014.07.120

40. Tesfaye M, Sivakumar V, Botai J, Mengistu Tsidu G. Aerosol climatology over South Africa based on 10 years of Multiangle Imaging Spectroradiometer (MISR) data. J Geophys Res Atmos. 2011;116(20):1–17. https://doi.org/10.1029/2011JD016023

41. Okonkwo JO, Awofolu OR, Moja SJ, Forbes PCB, Senwo ZN. Total petroleum hydrocarbons and trace metals in street dusts from Tshwane Metropolitan Area, South Africa. J Environ Sci Health A. 2006;41(12):2789–2798. https://doi.org/10.1080/10934520600966920

42. Squizzato S, Masiol M, Rich DQ, Hopke PK. A long-term source apportionment of $PM_{2.5}$ in New York State during 2005-2016. Atmos Environ. 2018;192:35–47. https://doi.org/10.1016/j.atmosenv.2018.08.044

43. Park M Bin, Lee TJ, Lee ES, Kim DS. Enhancing source identification of hourly $PM_{2.5}$ data in Seoul based on a dataset segmentation scheme by positive matrix factorization (PMF). Atmos Pollut Res. 2019;10(4):1042–1059. https://doi.org/10.1016/j.apr.2019.01.013

44. Zhao W, Hopke PK. Source apportionment for ambient particles in the San Gorgonio wilderness. Atmos Environ. 2004;38(35):5901–5910. https://doi.org/10.1016/j.atmosenv.2004.07.011

45. Laskin A, Moffet RC, Gilles MK, Fast JD, Zaveri RA, Wang B, et al. Tropospheric chemistry of internally mixed sea salt and organic particles: Surprising reactivity of NaCl with weak organic acids. J Geophys Res Atmos. 2012;117(15):1–12. https://doi.org/10.1029/2012JD017743

46. Kruger AC, Pillay DL, Van Staden M. Indicative hazard profile for strong winds in South Africa. S Afr J Sci. 2016;112(1–2), Art. #2015-0094. https://doi.org/10.17159/sajs.2016/20150094

47. Adesina JA, Piketh SJ, Qhekwana M, Burger R, Language B, Mkhatshwa G. Contrasting indoor and ambient particulate matter concentrations and thermal comfort in coal and non-coal burning households at South Africa Highveld. Sci Tot Environ. 2020;699. https://doi.org/10.1016/j.scitotenv.2019.134403

48. Hersey SP, Garland RM, Crosbie E, Shingler T, Sorooshian A, Piketh S, et al. An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data. Atmos Chem Phys. 2015;15(8):4259–4278. https://doi.org/10.5194/acp-15-4259-2015

49. Statistics South Africa (Stats SA). Community survey 2016 statistical release-P0301. Pretoria: Stats SA; 2016.

AUTHORS:
Yoel Rak[1,2]
Eli Geffen[3]
William Hylander[4]
Avishag Ginzburg[1]
Ella Been[1,5]

AFFILIATIONS:
[1]Department of Anatomy and Anthropology, Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel
[2]Institute of Human Origins and School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona, USA
[3]Department of Zoology, Tel Aviv University, Tel Aviv, Israel
[4]Department of Evolutionary Anthropology, Duke University, Durham, North Carolina, USA
[5]Sports Therapy Department, Ono Academic College, Kiryat Ono, Israel

CORRESPONDENCE TO:
Ella Been

EMAIL:
beenella1@gmail.com

DATES:
Received: 08 Aug. 2020
Revised: 04 Nov. 2020
Accepted: 09 Nov. 2020
Published: 28 May 2021

HOW TO CITE:
Rak Y, Geffen E, Hylander W, Ginzburg A, Been E. One hominin taxon or two at Malapa Cave? Implications for the origins of *Homo*. S Afr J Sci. 2021;117(5/6), Art. #8747. https://doi.org/10.17159/sajs.2021/8747

ARTICLE INCLUDES:
☒ Peer review
☒ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☒ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Margaret Avery

KEYWORDS:
Australopithecus sediba, *Homo* spp., Malapa hominins, hominin phylogeny, South Africa

FUNDING:
None

One hominin taxon or two at Malapa Cave? Implications for the origins of *Homo*

A report on the skeletons of two individuals from the Malapa cave site in South Africa attributes them both to a new hominin species, *Australopithecus sediba*. However, our analysis of the specimens' mandibles indicates that *Australopithecus sediba* is not a '*Homo*-like australopith', a transitional species between *Australopithecus africanus* and *Homo.* According to our results, the specimens represent two separate genera: *Australopithecus* and *Homo*. These genera are known to have jointly occupied sites, as seen in several early South African caves, so one cannot rule out the possibility that Malapa also contains remains of the two taxa. Our results lead us to additionally conclude that *all* the *Australopithecus* species on which the relevant mandibular anatomy is preserved (not only the 'robust' australopiths but also the 'gracile' – more generalised – ones) are too specialised to constitute an evolutionary ancestor of *Homo sapiens*. Furthermore, given that the Malapa site contains representatives of two hominin branches, one of which appears to be *Homo*, we must seek evidence of our origins much earlier than the date assigned to Malapa, approximately 2 million years before present. Support for this claim can be found in Ethiopian fossils attributed to the genus *Homo* and dated at 2.4 and 2.8 million years before present.

Significance:

- The proposed hominin species *Australopithecus sediba*, from the Malapa Cave in South Africa, seems to actually consist of two species, each of which represents a different hominin genus: *Homo* and *Australopithecus*. If, indeed, this is the case, *Homo* must have originated prior to the Malapa remains, contrary to the scenario suggested in the original report on *Au. sediba*.

Introduction

The proposal of a new hominin species, *Australopithecus sediba*, announced and described by Berger[1], Berger et al.[2], and de Ruiter et al.[3], is based primarily on the analysis of two partial skeletons, MH1 and MH2. The taxon is claimed to exhibit many features that suggest that it represents an intermediate species between *Australopithecus africanus* and *Homo*. This assertion was recently reiterated in a special issue of *PaleoAnthropology* dedicated to *Australopithecus sediba*.[4,5] However, a careful assessment of the mandibular remains leads us to conclude that the proposed *Au. sediba* species actually encompasses two species representing separate genera – *Australopithecus* and *Homo* – and as such cannot play a role in the origin of the latter. The discovery of two hominin species at one site is not unheard of in South Africa.

The two mandibles from Malapa plainly exhibit different patterns of ramal morphology: MH1 resembles australopith morphology, and MH2 displays the generalised morphology exhibited by *Homo sapiens* and other *Homo* species.

The morphology of the ascending ramus of the mandible in hominins has been found to be a diagnostic character[6] (note that Wolpoff and Frayer[7] claim that the upper part of the ramus is not diagnostic enough to distinguish between *H. sapiens* and *H. neanderthalensis*, but they cannot refute our argument because they have not applied our method to their sample); as such, the ramal morphology clearly distinguishes between *Australopithecus* and *H. sapiens*[8]. In the latter, the condylar and coronoid processes are relatively slender in a lateral view, they are similar in size, and they are separated by a broad, scooped out mandibular (sigmoid) notch, whose deepest point lies about halfway between the tips of the two processes (Figure 1). This configuration lends the notch a somewhat symmetrical appearance. In *Australopithecus*, on the other hand, the coronoid process is tall and broad, occupying about three-fourths of the ramal breadth. The process's superior end is rather flat, with a hook-like profile, and overhangs the relatively small mandibular notch, which is shallow in relation to the mandibular condyle. As a result, the outline of the notch is confined and asymmetric.

Similarly, *H. sapiens* and australopith rami seem to differ vis-à-vis the preangular notch. In *H. sapiens*, as in many other primates (i.e. the generalised configuration), the concave anterior margin of the ramus forms this notch, which is also present in MH2 (Figure 1). In the australopiths, as in MH1, the anterior margin of the ramus usually slopes diagonally in a straight line until it meets the mandibular body. Some exceptions to this dichotomy can be noted – for example, the presence of the preangular notch on MLD 40, Sts 52, Sts 7 and SKW 5, despite their assignment as *Au. africanus*. These exceptions somewhat diminish the diagnostic power of the preangular notch.

Because *H. sapiens* shares its ramal morphology with many other primates (for example, chimpanzees, orangutans, vervets and colobines), that morphology is clearly the primitive one, whereas the *Australopithecus* ramal configuration is derived – a synapomorphic character that combines *Au. robustus*, *Au. africanus* and *Au. afarensis* (and possibly other australopiths, such as *Au. anamensis* and *Au. boisei*, neither of which has a ramus that is sufficiently preserved to permit study) into what seems to be a monophyletic group. To suggest that the derived configuration, that of *Australopithecus*, evolved into the modern human configuration violates the principle of parsimony.

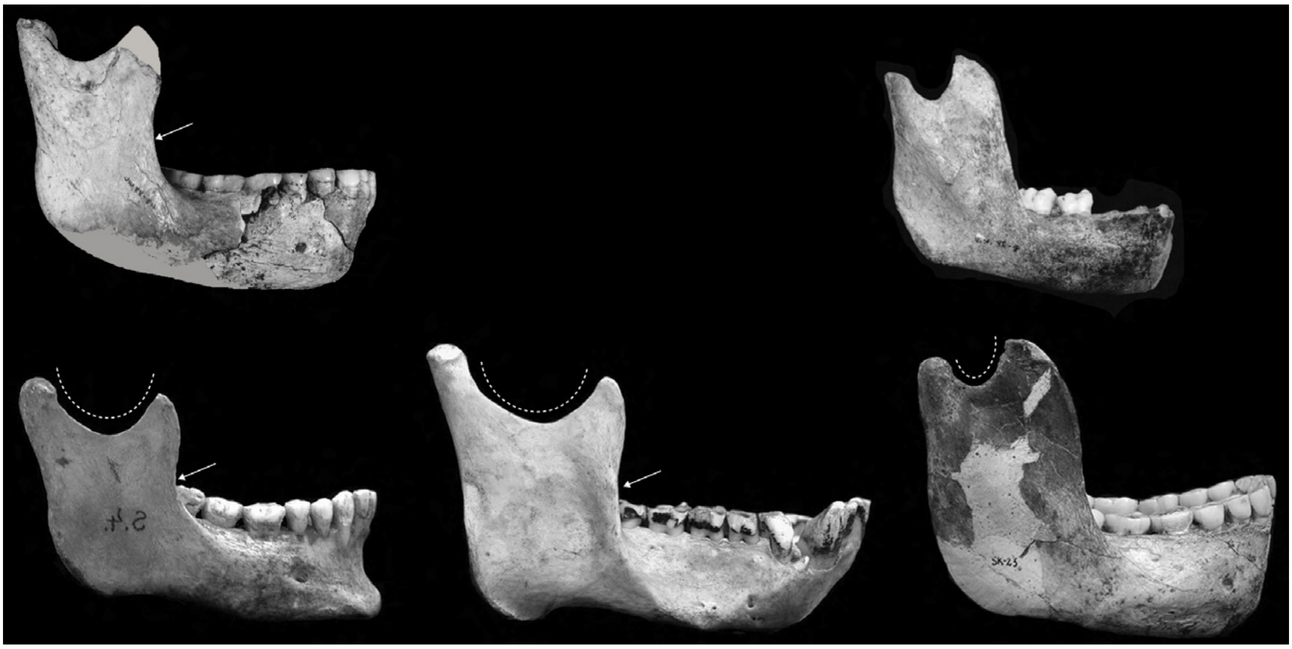


Figure 1: Ramal morphology of five hominoid specimens (not shown to scale). From upper right, clockwise: MH1, *Australopithecus robustus* (SK 23), orangutan, *Homo sapiens* and MH2. Note the hook-like shape of the coronoid process and the confined, narrow, and asymmetric mandibular notch in the *Australopithecus robustus* mandible and the notch's similarity to that of the MH1 mandible. The upper part of the ramus in all the other mandibles exhibits the generalised configuration. Also note the absence of a preangular notch on the anterior margin of the specialised ramus and the presence of the notch, indicated by white arrows, on the generalised ramus. The photographs of MH1 and MH2 were adapted from Berger et al.[2] and de Ruiter et al.[3] with permission. Note the parts that we added (reconstructed) on MH2.

Although we are convinced that the discrepancies that we have observed in ramal morphology stem from profound biomechanical differences, elucidation of the functionality at play (of the derived configuration) is a major project and beyond the scope of this study. Because the morphological differences are manifested in very young individuals[9,10], as described later, one can be certain that these morphologies are embedded in the genome and not generated by some activity during an individual's lifetime. In any case, the functional issue has no bearing on the taxonomic question treated here.

In this paper, our aim is not to determine which species of *Australopithecus* or *Homo* the Malapa mandibles belong to, but to determine how the two mandibles differ and what those differences mean. To accomplish these goals, we show that the differences are beyond what is expected in a trait's normal range of distribution in a given population. Our null hypothesis is that the two mandibles of *Au. sediba* represent a single taxon (as claimed, for example, by Berger[1]; Berger et al.[2]; de Ruiter et al.[3]; de Ruiter et al.[5]; Ritzman et al.[9]; and Williams et al.[4]). The alternative hypothesis is that the mandibles of *Au. sediba* represent a mix of taxa; in this scenario, a statistical analysis would classify one mandible with the *Australopithecus* cluster (but not provide any species assignment), and the other mandible with the generalised cluster (bearing a shared morphology). Indeed, our evidence supports this alternative hypothesis.

Materials and methods

Our sample includes 115 mandibles from mature extant primates, both male and female (Supplementary table 1): 41 modern humans, 58 chimpanzees (29 each of *Pan paniscus* and *Pan troglodytes*, grouped into one class following the results of previous analyses[8]), and 16 orangutans (*Pongo pygmaeus*). The *H. sapiens* specimens emanate from geographically varied regions: Australia (Aboriginal peoples), India, the Levant, and northern Canada (Inuit). Regarding the size of the modern *H. sapiens* sample, see the Results section. Fossil hominins in the sample consist of four rami from mature *Australopithecus* individuals (A.L. 288-1, SK 23, MAK-VP 1/83 and SK 34) and two rami from *Australopithecus* juveniles (SK 63 and A.L. 333-43). The juvenile specimens help increase the sample and were added after it became apparent that no ontogenetic change occurs in ramal morphology[9,10] (Figures 2–4). Another young individual, DIK11, from the Ethiopian Dikika site, exhibits the same ramal morphology, as seen on a photograph of the specimen (no cast has been available to us as yet). In addition, one *Ardipithecus ramidus* ramus, specimen GWM5sw/P56[11] (Figure 3), was included as an unknown. Five *Homo* fossils (three *H. erectus* specimens from Choukoutien, restored by Franz Weidenreich[12]; KNM-WT 15000 – *H. ergaster*; and ATD696, a mandible from Gran Dolina, Spain) were also analysed, although they proved to be of limited value (see Discussion).

Gorillas were excluded from our analysis. It was demonstrated in a 2007 study that the ramal morphology of gorillas is similar – although not identical – to that of *Australopithecus*.[8] As noted in that study:

> *given a phylogeny in which chimpanzees and modern humans are sister groups, parsimony dictates that we view the similarity in ramal morphology between* Australopithecus afarensis [in fact, all the australopiths that provide ramal evidence] *and gorillas as a homoplastic character, a character that appears independently and as such has no phylogenetic value.*[8(p.6570)]

The similarity between the gorilla ramus and that of *Au. robustus* may well stem from the very tall ramus in both groups.

Regarding reconstruction, MH1 requires none. In MH2, the tip of the coronoid process is damaged; nevertheless, its reconstruction is straightforward, as seen in Figure 5. The three dotted white lines on the superimposed images were added by us. The lines demonstrate that there is no way to reconstruct the coronoid process in MH2 to resemble the robust configuration.

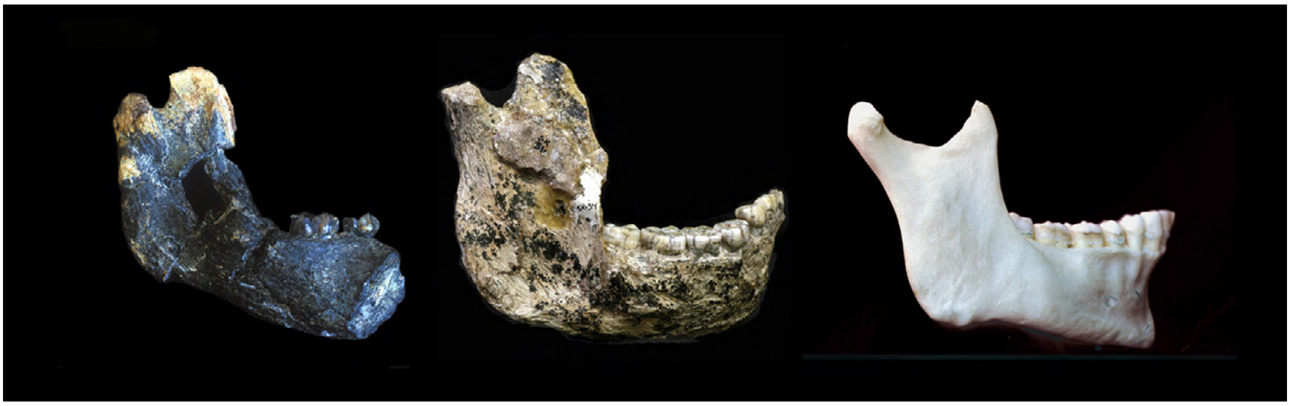


Figure 2: Comparison of ramal morphology in three specimens, left to right (not shown to scale): A.L. 33343 (infant), SK 34 (mature individual), and modern *Homo sapiens* (mature individual). Note that the two australopith rami are virtually identical in shape despite their difference in individual age, and their shape differs from that of the generalised (i.e. shared) configuration, which is seen in *H. sapiens*.

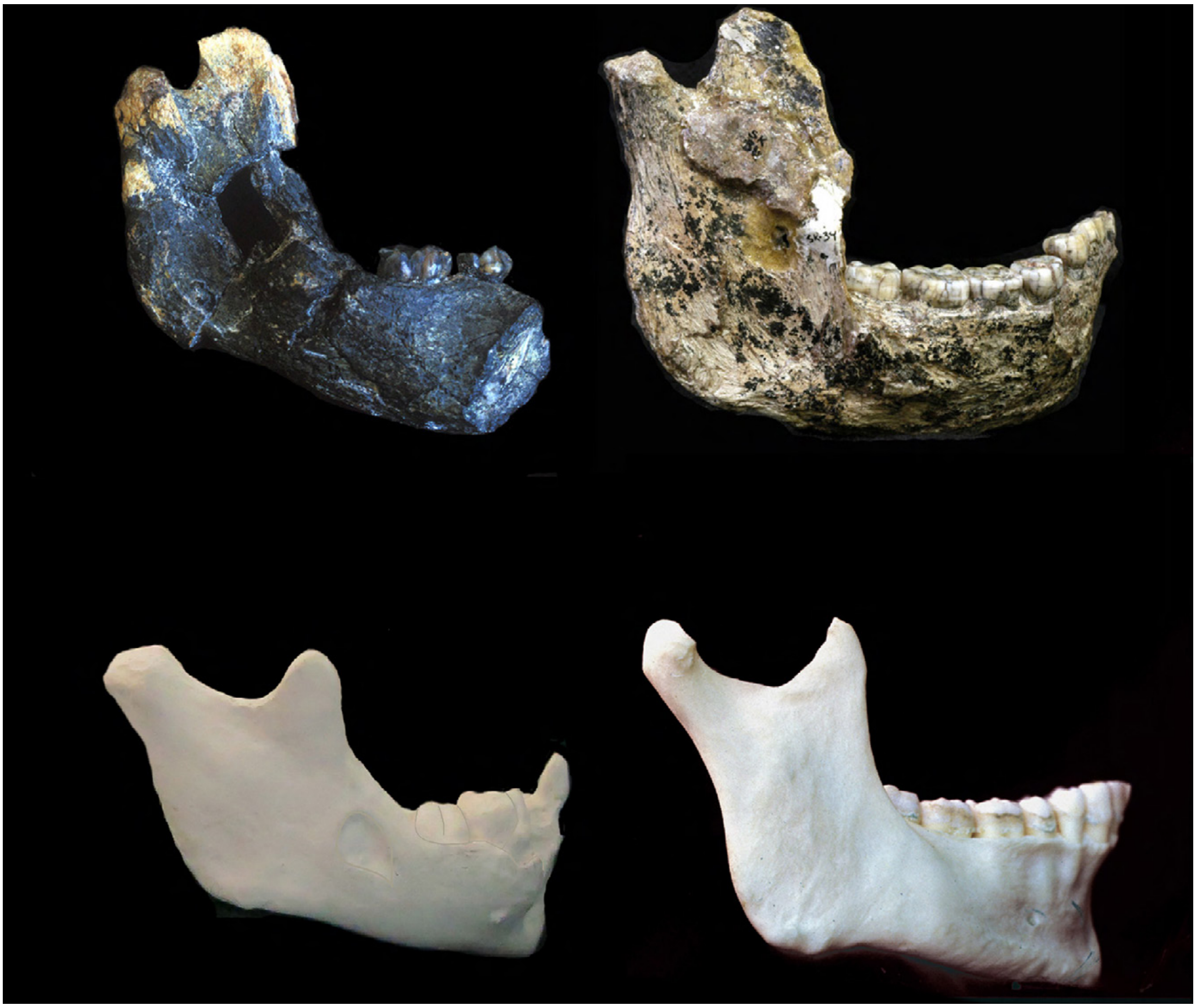


Figure 3: Comparison of ramal morphology in four mandibles (not shown to scale). Upper: juvenile *Australopithecus* specimen A.L. 33343 (left) and adult *Australopithecus* specimen SK 34 (right). Lower: juvenile *Homo sapiens* mandible (left) and adult *H. sapiens* mandible (right). Note that the upper part of the ramal morphology is the same in the juvenile specimen and its corresponding mature specimen in both pairs. In the *Australopithecus* mandibles, the coronoid process is taller than the condylar process; the mandibular notch between them is confined and asymmetric; and its deepest point is very close to the condylar process. This configuration is quite different from that of *H. sapiens*, in which the two processes are the same size in the juvenile and the adult; the mandibular notch is wide; and the deepest point of the notch is midway between the condylar and coronoid processes.

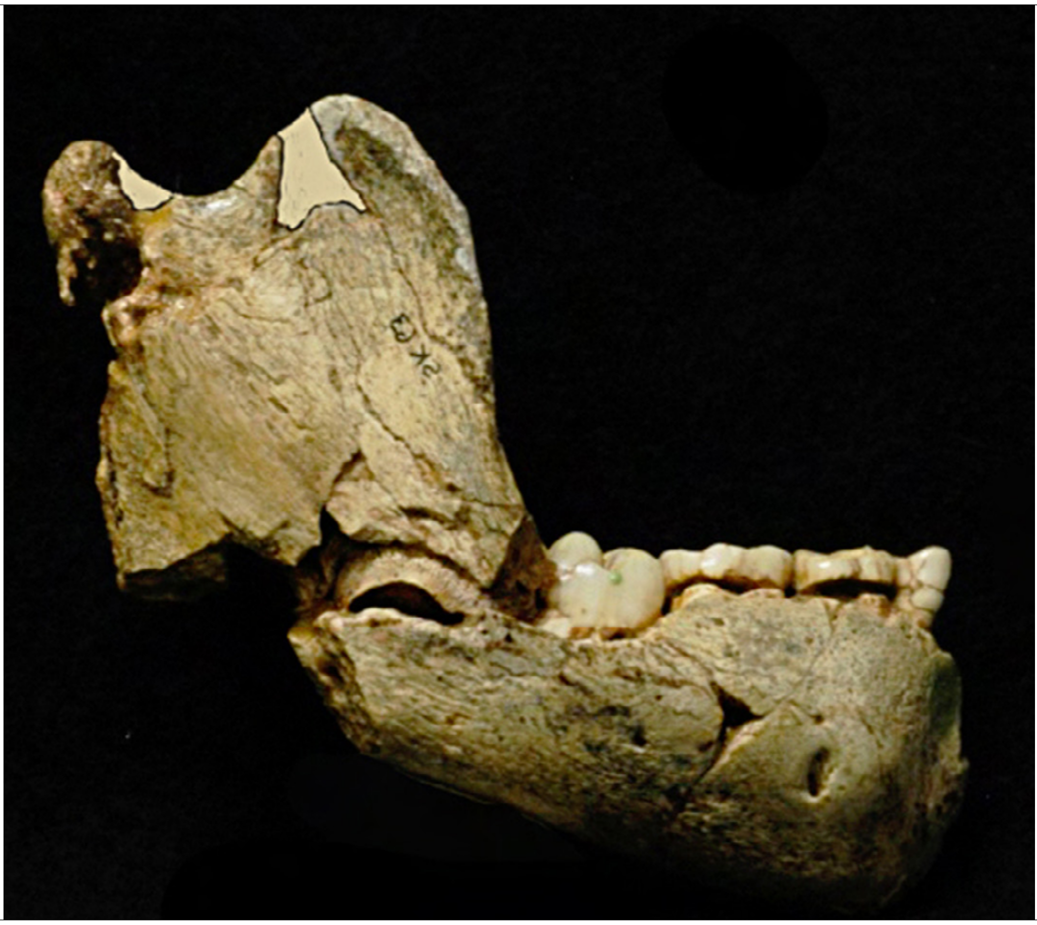


Figure 4: A juvenile *Au. robustus* specimen, SK 63 (flipped), exhibiting the ramal configuration typical of *Australopithecus*.

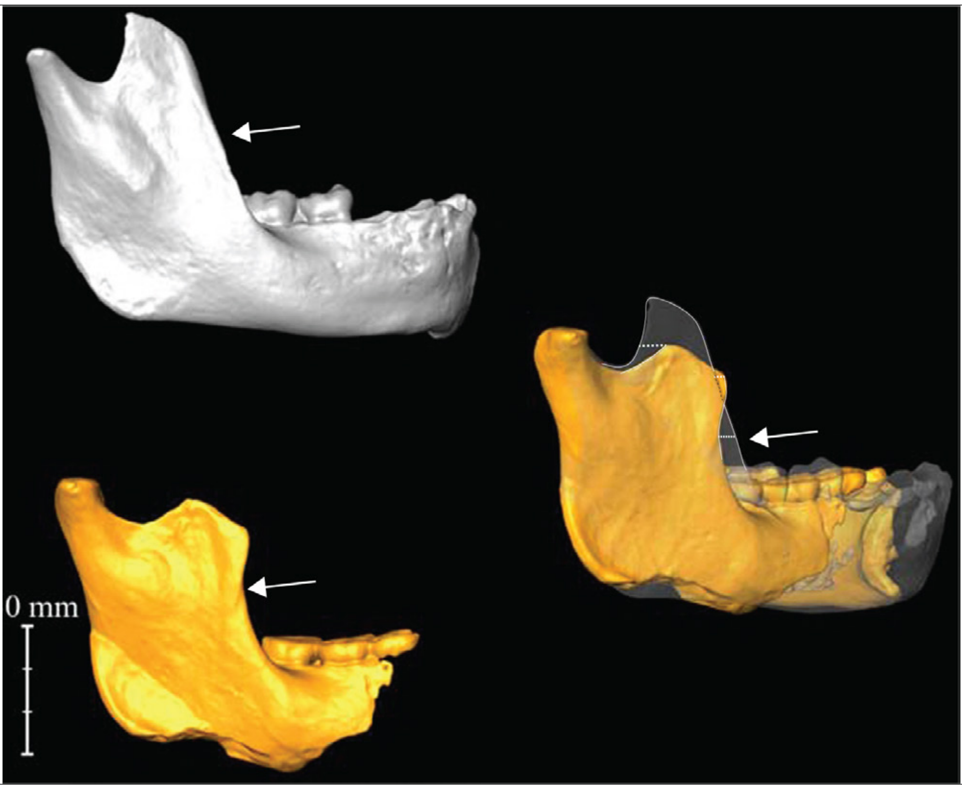


Figure 5: A portion of Berger et al.'s figure S2[2] showing MH1 (upper) and MH2 (lower). Berger et al.[2] have superimposed MH1 on MH2 (right), resulting in a vivid illustration of the morphologies that we claim distinguish between *Australopithecus* and *Homo*. For best viewing, enlarge the image. Note the three dotted white lines that we added to the right-hand image. These three lines indicate the differing morphologies of the upper part of the ramus. The white arrows point to the preangular notch or its absence.

We quantified the upper ramal contours of the specimens through a simple method described by Rak et al.[8(p.6571)]:

> *To convey the anatomical differences in the upper ramal contour, we adopted a method... which consisted of capturing a digital image of the mandibular ramus with the camera centred at the vertical level of the mandibular notch and held perpendicular to the lateral surface of the ramus. ... We traced the digital image of each ramus from the tip of the condylar process to the anterior margin of the ramus. ...*
>
> *...We stretched the contour proportionally on the vertical and horizontal axes by dragging the contour's lower right corner until it occupied the entire width of the area of the fixed coordinates in the background template. This part of the procedure eliminated differences in size in the analysis* [leaving shape only]. *The posterior margin* [of the ramus] *was aligned with the vertical line at 0, and the anterior margin was aligned at T. The posterior ramal margin in the entire sample exhibits a slight concavity between the posterior end of the condyle and the insertion site of the posterior fibres of the masseter and medial pterygoid muscles; using these two posteriorly protruding structures, we were able to orient the posterior margin on a vertical line throughout the sample. The intersection of the ramal contour with each of the vertical lines, A through T, yielded 20 numeric variables for each ramus.*[8(p.6571)]

We define variable T as the maximum *horizontal* distance between the condyle and two-thirds of the anterior ramal margin's height. In this way, we accentuate the most diagnostic part of the ramal outline (A–T). Note that the use of the point defining T (or any other point on the ramus) does not affect the height measurement of the coronoid process in the mandibles under study.

The intersection of each contour with a vertical line and a horizontal line (i.e. coordinates) is assigned a value representing the distance of the intersection point from the zero horizontal line (for example, 10, 20 or 30) (Figure 6). These are the numerical values used for the statistics. Note that as long as all the contours are on the same grid, units of measure are irrelevant, as is the distance between the lines (provided that it is constant).

We chose the same orientation for the posterior margin of all the rami in our sample because that orientation seems to be *fixed* in relation to the base of the skull, the Frankfurt horizontal, and the zygomatic arch (indicating functional significance), as demonstrated in Figure 7. Alternatively, positioning all the mandibles with a horizontal orientation of the occlusal plane or of the base of the mandibular body would introduce variation in the shape of the mandibular notch.

The 20 (AT) variables served as independent variables in a general discriminant analysis to classify two unknown fossil specimens (MH1 and MH2). We used Jump version 15 software for all analyses. General discriminant analysis applies the general linear model approach to discriminant analysis and can use both continuous and categorical independent variables. Our reference classes consisted of *Australopithecus*, *Pan*, *Pongo pygmaeus* and *H. sapiens* mandibles. The prior probability of classification was set as equal. The key assumption in discriminant analysis is that the variables used are not completely redundant.

To reduce dimensionality and eliminate the dependence between the variables, we used two independent approaches. First was the best-subset approach. Of 1 048 556 possible models, we inspected the 100 models that accounted for most of the variation (i.e. that exhibited the lowest misclassification rate) (Supplementary figure 1). Out of those models, we selected the one with the least number of parameters and used its functions to predict the state of unknowns. This approach considerably reduces the number of variables in the analysis and keeps the power of classification nearly the same. Thus, the retained variables are those that are most multidimensionally informative for the classification.

Our second approach was a principal component analysis to accommodate the effects of collinearity among the variables. All 20 variables were collapsed into principal components. Five components – those with an eigenvalue greater than or equal to 1 (after varimax rotation) – were retained. We used the components' scores as independent variables in the general discriminant analysis, and the resulting discriminant functions served to classify the unknown fossil specimens. For cross-validation, we applied the leave-one-out procedure. Through the two approaches just described, we classified the unknown fossil specimens. We also reran the analysis under a two-class model: *Australopithecus* and taxa with a generalised ramus (*Pan, Pongo* and *Homo*).

Results

Even in the absence of the coronoid process on MH2, the differences between it and MH1 are readily visible, as was shown in 2010 by Berger et al.[2] themselves (reproduced and modified here in Figure 5). The unreconstructed outline of the mandibular (sigmoid) notch in MH2 diverges quite clearly from the comparable area in MH1. When the two specimens are adjusted to the same scale (Figure 6), the deepest part of the notch in MH2 is situated much more anteriorly than in MH1 and descends much farther relative to the zero point, i.e. the mandibular condyle, as in the generalised configuration. The statistical analysis tells us that the difference between the height of the two coronoid processes, reconstructed or not, is of less importance than the outline of the mandibular notch itself and has little effect on the results.

The best-subset approach yielded a classification success ranging from 94.6% to 92.5%. Most of these models share variables A, G, H, I, P, R and T (Figure 8). The smallest subset model consists of eight effects (A, F, G, I, O, P, R and T; Figure 8), which correctly classify 93.3% of the cases (the leave-one-out cross-validation classification success is 87.4%). According to the posterior probabilities ($p(k_i)$) from the smallest subset model, MH2 falls in the generalised group (assigned as most likely an orangutan, with $p(k_i) = 0.76$), whereas MH1 is assigned as most likely *Australopithecus* ($p(k_i) = 1.00$). Finally, when only *Australopithecus* and the generalised ramus group are considered, the two best models consist of a single variable (I or J*)* that correctly classifies all cases (100%) (Figure 6). This variable corresponds to the deepest point of the notch in the generalised group; compare with the position of the homologous point in the specialised group (Figure 6). According to this model, MH1 is assigned to the *Australopithecus* cluster ($p(k_i) = 0.98$) and MH2 to the generalised ramus group ($p(k_i) = 1.00$). The ramus of the *Ar. ramidus* mandible[11] (Figure 3) falls in the generalised cluster, with a probability of 0.98.

Note that the size of the *H. sapiens* sample (41 individuals) is not what counts; rather, the statistical analysis regards the entire generalised sample, consisting of 115 individuals, as one group, because the real issue is whether the Sediba mandibles fall in the generalised cluster, the specialised one, or both.

In the principal component approach, the four factors with an eigenvalue greater than or equal to 1 together accounted for 91.7% of the variation in the data. (The first four principal components accounted for 50%, 17%, 17% and 8% of the variance, respectively, totalling 92%.) The eigenvalues of the factors were 9.9 (49.7%), 3.4 (17.2%), 3.4 (17.2%) and 1.5 (7.6%). All four factors were significantly different from each other (Bartlett test, $p < 0.0001$ for each of the factors).

The general discriminant analysis correctly classifies 74.2% of cases (with the leave-one-out crossvalidation classification success at 69.7%). Posterior probabilities of this model assign MH1 as *Australopithecus* ($p(k_i) = 1.0$) and MH2 as most likely an orangutan ($p(k_i) = 0.755$). The latter is in contrast to a probability of 0 as *Australopithecus*. *Ar. ramidus* is classified as most likely a chimpanzee ($p(k_i) = 0.49$) or orangutan ($p(k_i) = 0.49$). Finally, when only *Australopithecus* and the generalised ramus group are considered, MH2 and *Ar. ramidus* are assigned to the generalised ramus group ($p(k_i) = 1.00$ in both cases) and MH1 is classified as *Australopithecus* ($p(k_i) = 1.00$).

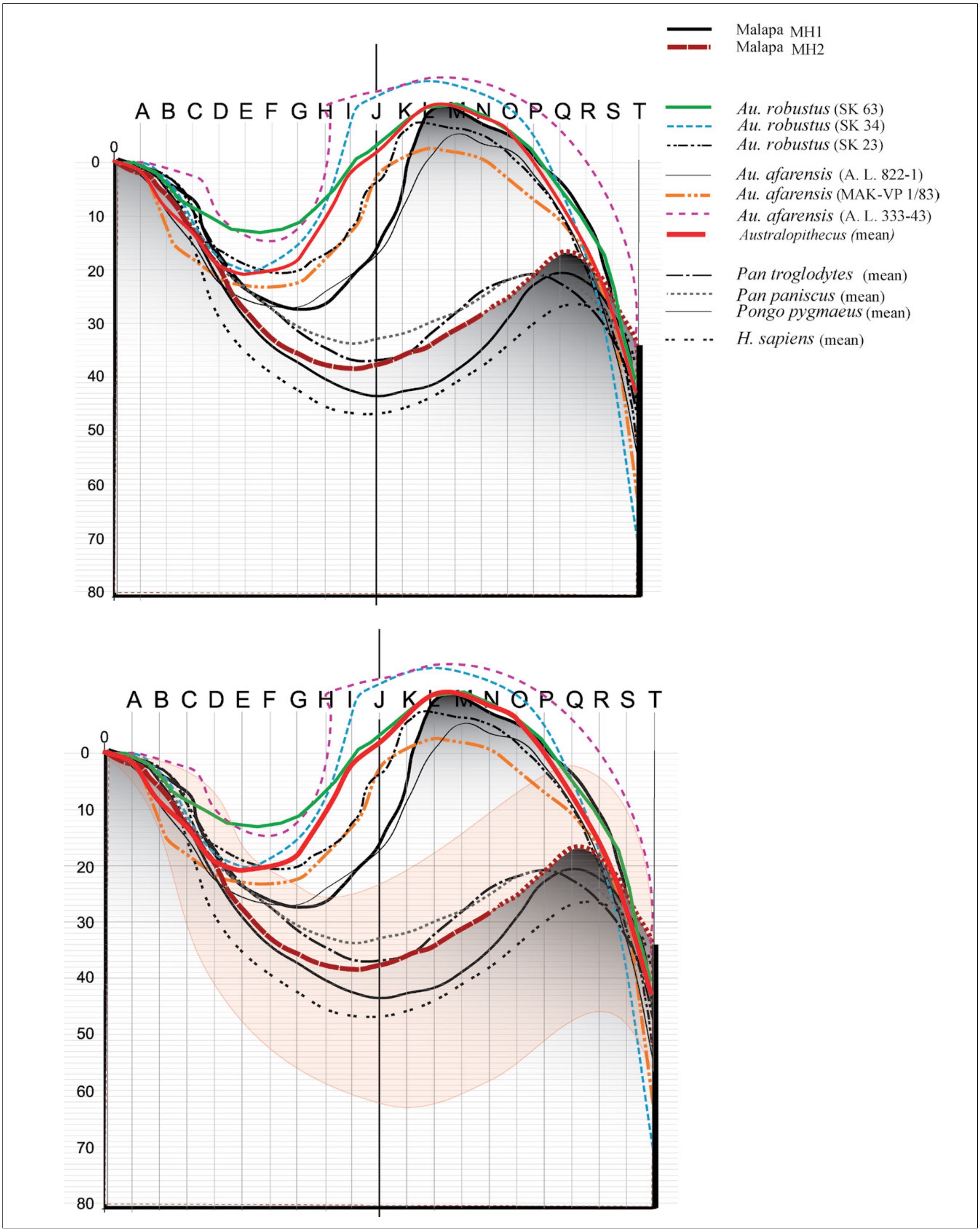


Figure 6: Ramal outlines of fossil specimens and mean ramal outlines of extant hominoid groups, stretched proportionally to fit the distance zero to T. The outlines form two distinct assemblages. Upper graph: the grey-shaded area in the upper portion of the graph represents the MH1 ramus. The lower grey-shaded area represents the MH2 ramus, delineated by the MH2 outline itself (thick, dashed maroon line). The vertical line J represents a variable that is alone sufficient to distinguish between the *Australopithecus* and generalised assemblages. Note that the variable J does not intersect the dotted maroon line, which represents the reconstructed coronoid tip in MH2; i.e. the reconstruction has no influence on the results. Lower graph: this graph is the same as the upper one, with the addition of the range of variation in the *Homo sapiens* sample (light-red shaded area). Note that the means of the extant hominoid sample fall within the range of *H. sapiens*.

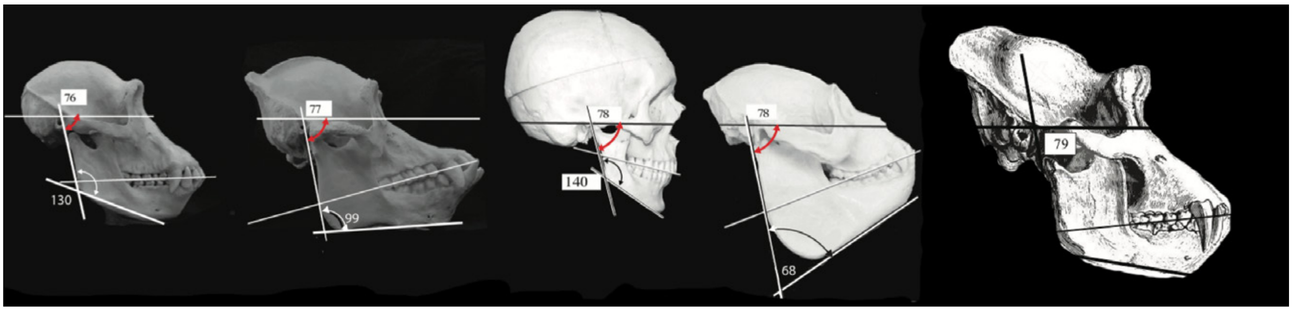


Figure 7: The angle, in degrees, between the posterior margin of the mandibular ramus and the Frankfurt horizontal (which coincides more or less with the zygomatic arch), demonstrating the rationale for using a fixed orientation of the mandibular ramus. From left to right (not shown to scale): chimpanzee, orangutan, *Homo sapiens*, howler monkey and gorilla. Note the similar angle, at 78°, in even the most extreme cases – *H. sapiens* and the howler monkey. In contrast, note the variation in the orientation of the occlusal plane and the inferior margin of the mandibular body.

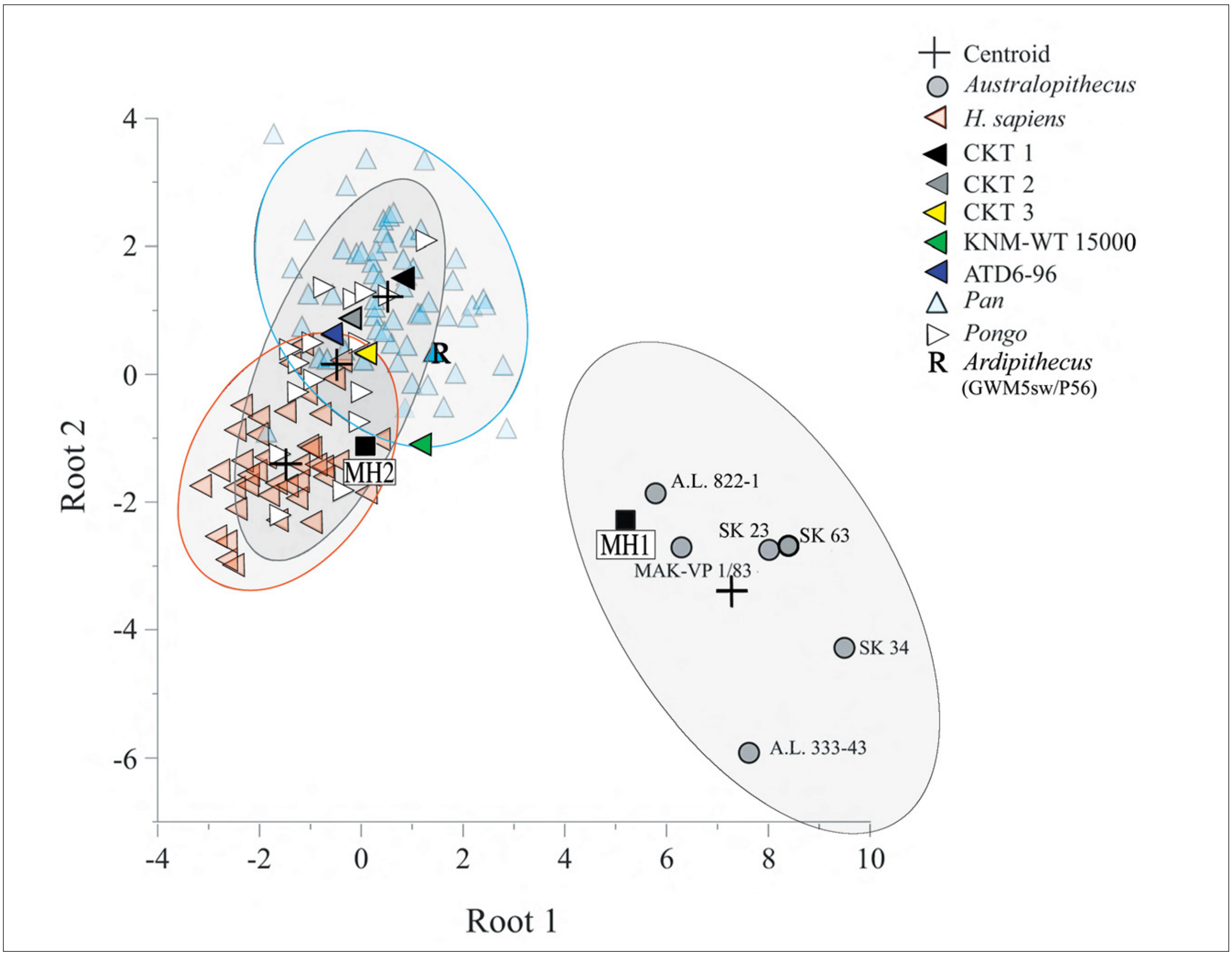


Figure 8: Discriminant function plot including CKT, Choukoutien; KNM-WT 15000, *Homo ergaster*; ATD696, a mandible from Gran Dolina, Spain. Roots 1 and 2 account for 59% and 34% of the variation, respectively. MH1 falls within the *Australopithecus* cluster (with a probability of 0.98), as do the infants A.L. 33343 and SK 63. MH2 falls within the generalised ramus group (with a probability of 1.0), as does *Ardipithecus ramidus* (with a probability of 0.98). The ellipses represent a confidence level of 95%.

The morphological overlap between the comparative taxa is high only in the groups displaying the generalised morphology (not surprisingly, given that 'generalised', by definition, is shared). On the other hand, there is no overlap whatsoever between the generalised group and the derived one (Figure 6 and Figure 8). Note that no attempt has been made to assign MH1 and MH2 to particular species (because one of these configurations is synapomorphic and the other is generalised). The fact that MH2 is classified as most likely an orangutan is of little relevance, nor does it come as a surprise. What counts is that MH2's generalised configuration puts it in the generalised cluster.

Discussion

The data, as seen in both the distribution of the actual contours (Figure 6) and the plot of the discriminant analysis (Figure 8), clearly demonstrate that the MH2 mandible falls in the group that exhibits the generalised configuration, a group that includes *H. sapiens*. The MH1 mandible, on the other hand, is clearly clustered with the australopiths.

Although we were limited to specimens that are complete enough to be included in our analysis, other, more fragmentary, specimens of *Australopithecus* (A.L. 333100, A.L. 333w15, A.L. 333n1, A.L. 333108, A.L. 4381g, A.L. 288–1i and DNH 8) exhibit what is undoubtedly the derived configuration of the ramus. Although not a single ramus of *Au. africanus* is complete enough to be included in the analyses, one can clearly see the derived morphology on a forgotten fragment, Sts 7, that is still embedded in matrix (Supplementary figure 2). (Note that Kimbel and Rak's study[13] of the face of MH1 led them to conclude that the specimen is *Au. africanus*.)

Not surprisingly, the fossil *Homo* specimens that were included in the sample fall in the same cluster as the generalised hominoids (Figure 8). Nevertheless, the *Homo* fossils are of little value to the analysis because in order to serve as an outgroup, they must be assigned to a branch that predates the emergence of the so-called *Au. sediba* species (i.e. fossils that are nested between the *H. sapiens* branch and the *Australopithecus* clade are of no use in this context) – a scenario that we find hard to accept. *Ar. ramidus* is the only hominin that is helpful in this respect; indeed, like chimpanzees and orangutans, *Ar. ramidus* displays a generalised configuration of the ramus (Figure 8).

A recent study[9] examines a claim that has been presented in several forums[14-16] and that we offer here in detail: that two taxa are present in the *Au. sediba* hypodigm. In their analysis, Ritzman et al. state that[9(p.54)]:

> *while the difference between MH1 and MH2 is large relative to within-species comparisons, it does not generally fall outside of the confidence intervals for extant intraspecific variation. However, the MH1–MH2 distance also does not plot outside and below the between-species confidence intervals. Based on these results, as well as the contextual and depositional evidence, we conclude that MH1 and MH2 represent a single species and that the relatively large degree of variation in this species is due to neither ontogeny nor sexual dimorphism.*

Ritzman et al. do, however, acknowledge that 'the possibility that it [*Au. sediba*] samples two taxa cannot be completely refuted'[9(p.62)].

The reason that Ritzman et al.[9] cannot clearly distinguish between the gorilla cluster and that of modern humans, for example, nor between MH1 and MH2, is rather simple: in their study's method, a large percentage of the variables (semi-landmarks) are identical in humans, gorillas, and all the other groups in their sample because of the straight anterior outline of the rami in all. In other words, only a small percentage of the semi-landmarks are of diagnostic value and worth comparing. The Ritzman et al.[9] analysis is thus incompatible with our analysis, which takes into consideration only the relevant, more diagnostic, part of the anatomy.

Furthermore, the absence of an australopith sample in Ritzman et al.'s[9] study seriously detracts from their conclusions. Indeed, the inclusion of australopiths as a distinct known group in our statistical analysis demonstrates clearly that MH1, when treated as unknown, falls in the australopith cluster, whereas MH2, when treated as unknown, falls in the generalised group, as noted earlier (Figure 8). (An additional factor affecting the analysis by Ritzman et al.[9] is their inclusion of gorillas, due to the different goals of their study.)

The discovery of a hypodigm represented by only two individuals whose morphologies fall at the extreme opposite ends of the range of a population with a normal distribution is highly unlikely. Is it possible that, in keeping with the common primate pattern, the specialised mandible (MH1) represents a male, exhibiting more specialised cranial anatomy, and MH2, with its generalised mandible, represents a female? This scenario is also unlikely, given that primate mandibles usually do not exhibit sexual dimorphism to such magnitude in characters other than size (see also Ritzman et al.[9]). Even species with pronounced sexual dimorphism, such as the gorilla, demonstrate no sexual dimorphism in ramal shape. The specialised cluster in our study includes specimens that are clearly female (for example, A.L. 822), as well as young individuals, which, although expected to display the generalised anatomy, are nevertheless characterised by a specialised morphology. Most important – even if MH1 and MH2 do represent one highly dimorphic, transitional species – the presence of a specialised mandible (as commonly defined) in that hypodigm would be sufficient to invalidate Berger et al.'s[3] original claim that the proposed species is part of our ancestry. Hence, the credibility of the MH2 reconstruction is of no relevance to the phylogenetic issue once we recognise that the complete MH1 mandible is specialised. (See also Du and Alemseged[17].)

Regarding the question of the *Homo* species at play, note that the main concern is the distinction between *Australopithecus* at large and *Homo*. Hence, we do not deal with nomenclature on a species level; it is sufficient to demonstrate that one of the Malapa mandibles belongs to the genus *Homo* and the other to the genus *Australopithecus* – that is, that representatives of both genera existed at Malapa.

The presence at one site of two hominin species of two genera ought not to be surprising. The coexistence of *Homo* and *Australopithecus* at South African cave sites has already been documented. In the mid-20th century, a *Homo* mandible with small teeth was discovered in Swartkrans Cave, which has yielded specimens that are mostly *Au. robustus*.[18] Later, Clarke and Howell[19] recognised that the Swartkrans specimens SK 80 and SK 47 actually constitute one *Homo* specimen, SK 847. In the nearby Sterkfontein site, the presence of both *Homo* and *Au. africanus* was acknowledged with the discovery of the *Homo* specimen StW 53[20], although admittedly, the latter specimen may be several hundred thousand years younger than the *Au. africanus* specimens from Sterkfontein. The StW 53 skull includes a neglected fragment of a badly preserved ramus, whose morphology supports the claim by Hughes and Tobias[20] that the specimen represents *Homo*. However, Sts 19, a controversial specimen that emanates from the Sterkfontein type site itself along with numerous *Au. africanus* specimens, has been identified by some researchers as *Homo* rather than *Australopithecus*.[21-24] The Sterkfontein group of *Homo* specimens probably includes the immature individual StW 151 as well.[25] Another South African site, Drimolen, provides additional evidence of the coexistence of these two taxa.[26,27] For a meticulous discussion of the chronology of the hominin-bearing layers in South Africa, see Herries et al.[28] and Herries and Shaw[29].

Conclusions

The phylogenetic scenario that Berger[1] and Berger et al.[2] propose, in which *Au. sediba* is a link in our evolutionary chain, ought to be ruled out because one of the Malapa mandibles is too derived to be positioned in the human lineage. Furthermore, one need not dismiss *in limine* the presence of *Homo* at sites contemporary with Malapa or even earlier, or question their geological age, as does Berger in his discussion of a 2.4-million-year-old *Homo* specimen par excellence, A.L. 666, from the Afar in Ethiopia.[1] Malapa itself clearly contains a generalised specimen, and *Homo*, given the size and shape of the fossil, is the only candidate. The split between *Homo* and the robust clade must have occurred earlier than the occupation of Malapa.

Hence, *Au. sediba* is not a species that 'shares more derived features with early *Homo* than any other australopith species', as claimed by Berger et al.[2] In fact, their *Au. sediba* seems to represent a mixture of two hominin taxa, leading Berger et al. to refer to the new species as a transitional one (Figure 5). Moreover, by viewing *Au. sediba* as an ideal link between *Au. africanus* and *H. habilis*, they ignore all the synapomorphic features that the former already shares with the robust clade and to which attention was drawn many years ago (for example, by Aguirre[30], Johanson and White[31] and Rak[32]) (Figure 9).

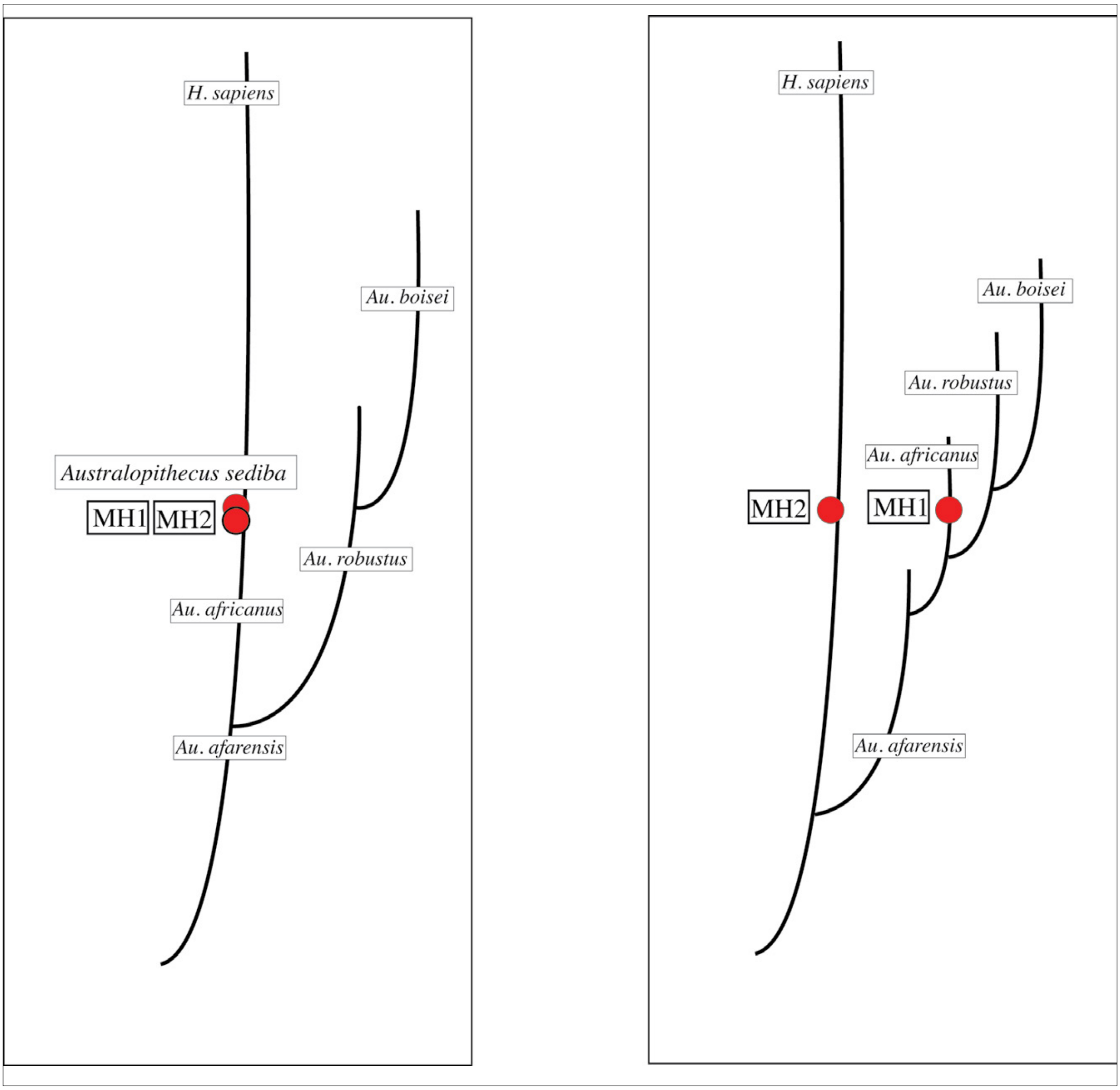


Figure 9: Two proposed phylogenetic trees. The left tree depicts Berger et al.'s 2010 proposal[2], in which *Australopithecus sediba* is a link in the chain between *Au. africanus* and *Homo sapiens*. The right tree depicts our proposal, in which MH1 lies on the robust clade and MH2 is on the *Homo* lineage. In addition, an analysis of the MH1 *skull* indicates that there is no reason to exclude it from the *Au. africanus* hypodigm.[23] The *Au. afarensis* mandibular ramus is too derived to allow us to place it on the *Homo* lineage. We have placed *Au. africanus* on the robust lineage, based on facial synapomorphies that it shares with the robust australopiths.[32]

All the australopiths on which the relevant ramal morphology is preserved (*Au. afarensis*; *Au. africanus*, including the *Australopithecus* specimen at Malapa; and certainly *Au. robustus*) are actually too derived to play the role of a *H. sapiens* ancestor. Given that Malapa already contains representatives of two hominin branches, one of which appears to be *Homo*, we must seek the latter's origin in geological layers that are earlier than those at Malapa, which are dated at approximately 2 million years before present.[33] Support for such a scenario can be found in earlier Ethiopian fossils attributed to the genus *Homo*: A.L. 666, dated at 2.4 million years[34], and LD 3501, dated at 2.8 million years[35].

Acknowledgements

We thank J. Moggi-Cecchi, WH Kimbel and Erella Hovers for their comments on the manuscript. We also thank Lee Berger for his hospitality and for graciously providing Y.R. with access to the original fossils. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

Competing interests

We declare that there are no competing interests.

Authors' contribution

Y.R.: Study conception and design, acquisition of data, analysis and interpretation of data, drafting of manuscript. W.H.: Study conception and design, critical revision. E.B.: Study conception and design, critical revision. A.G.: Acquisition of data, analysis and interpretation of data. E.G.: Analysis and interpretation of data, drafting of manuscript.

References

1. Berger LR. *Australopithecus sediba* and the earliest origins of the genus *Homo*. J Anthropol Sci. 2012;90:117–131. https://doi.org/10.4436/jass.90009

2. Berger LR, De Ruiter DJ, Churchill SE, Schmid P, Carlson KJ, Dirks PHGM, et al. *Australopithecus sediba*: A new species of *Homo*-like australopith from South Africa. Science. 2010;328:195–204. https://doi.org/10.1126/science.1184944

3. De Ruiter DJ, DeWitt TJ, Carlson KB, Brophy JK, Schroeder L, Ackermann RR, et al. Mandibular remains support taxonomic validity of *Australopithecus sediba*. Science. 2013;340:12329971–12329974. https://doi.org/10.1126/science.1232997

4. Williams SA, DeSilva JM, De Ruiter DJ. Malapa at 10: Introduction to the special issue on *Australopithecus sediba*. PaleoAnthropology. 2018:49–55. https://doi.org/10.4207/PA.2018.ART111

5. De Ruiter DJ, Carlson KB, Brophy JK, Churchill SE, Carlson KJ, Berger LR. The skull of *Australopithecus sediba*. PaleoAnthropology. 2018:56–155. https://doi.org/10.4207/PA.2018.ART112

6. Rak Y, Ginzburg A, Geffen E. Does *Homo neanderthalensis* play a role in modern human ancestry? Am J Phys Anthropol. 2002;119:199–204. https://doi.org/10.1002/ajpa.10131

7. Wolpoff MH, Frayer DW. Unique ramus anatomy for Neandertals? Am J Phys Anthropol. 2005;128:245–251. https://doi.org/10.1002/ajpa.10432

8. Rak Y, Ginzburg A, Geffen E. Gorilla-like anatomy on *Australopithecus afarensis* mandibles suggests *Au. afarensis* link to robust australopiths. Proc Natl Acad Sci USA. 2007;104:6568–6572. https://doi.org/10.1073/pnas.0606454104

9. Ritzman TB, Terhune CE, Gunz P, Robinson CA. Mandibular ramus shape of *Australopithecus sediba* suggests a single variable species. J Hum Evol. 2016;100:54–64. https://doi.org/10.1016/j.jhevol.2016.09.002

10. Terhune CE, Robinson CA, Ritzman TB. Ontogenetic variation in the mandibular ramus of great apes and humans. J Morphol. 2014;275:661–677. https://doi.org/10.1002/jmor.20246

11. Semaw S, Simpson SW, Quade J, Renne PR, Butler RF, McIntosh WC, et al. Early Pliocene hominids from Gona, Ethiopia. Nature. 2005;433:301–305. https://doi.org/10.1038/nature03177

12. Weidenreich F. The mandibles of *Sinanthropus pekinensis*: A comparative study. Palaeontology. 1936;Sinica D 7, 1e132.

13. Kimbel WH, Rak Y. *Australopithecus sediba* and the emergence of *Homo*: Questionable evidence from the cranium of the juvenile holotype MH 1. J Hum Evol. 2017;107:94–106. https://doi.org/10.1016/j.jhevol.2017.03.011

14. Been E, Rak Y. The lumbar spine of *Australopithecus sediba* indicates two hominid taxa. In: Paleoanthropology Society Meeting Abstracts; 2014 April 8–9; Calgary, Canada. PaleoAnthropology. 2014:A2. https://doi.org/10.4207/PA.2014.ABS12

15. Rak Y, Been E. Two hominid taxa at Malapa: The mandibular evidence. In: Paleoanthropology Society Meeting Abstracts; 2014 April 8–9; Calgary, Canada. PaleoAnthropology. 2014:A20. https://doi.org/10.4207/PA.2014.ABS12

16. Rak Y, Been E. What do we really know about the origin of humans? (Abstract). In: Proceedings of the 6th Annual Meeting of the European Society for the Study of Human Evolution (PESHE); 2016 September 14–17; Madrid, Spain. p. 199.

17. Du A, Alemseged Z. Temporal evidence shows *Australopithecus sediba* is unlikely to be the ancestor of *Homo*. Sci Adv. 2019;5(5), eaav9038. https://doi.org/10.1126/sciadv.aav9038

18. Robinson JT. *Telanthropus* and its phylogenetic significance. Am J Phys Anthropol. 1953;11:445–502. https://doi.org/10.1002/ajpa.1330110402

19. Clarke RJ, Howell FC. Affinities of the Swartkrans 847 hominid cranium. Am J Phys Anthropol. 1972;37:319–335. https://doi.org/10.1002/ajpa.1330370302

20. Hughes AR, Tobias PV. A fossil skull probably of the genus *Homo* from Sterkfontein, Transvaal. Nature. 1977;265:310–312. https://doi.org/10.1038/265310a0

21. Clarke RJ. The cranium of the Swartkrans hominid, SK 847, and its relevance to human origins [PhD dissertation]. Johannesburg: University of the Witwatersrand; 1977.

22. Dean MC, Wood BA. Basicranial anatomy of Plio-Pleistocene hominids from East and South Africa. Am J Phys Anthropol .1982;59:157–174. https://doi.org/10.1002/ajpa.1330590206

23. Kimbel WH, Rak Y. The importance of species taxa in paleoanthropology and an argument for the phylogenetic concept of the species category. In: Kimbel WH, Martin LB, editors. Species, species concepts, and primate evolution. New York: Plenum; 1993. p. 461–484. https://doi.org/10.1007/978-1-4899-3745-2_18

24. Kimbel, WH, Rak Y, Johanson DC. The skull of *Australopithecus afarensis*. Oxford: Oxford University Press; 2004. https://doi.org/10.1093/oso/9780195157062.001.0001

25. Moggi-Cecchi J, Tobias PV, Beynon AD. The mixed dentition and associated skull fragments of a juvenile fossil hominid from Sterkfontein, South Africa. Am J Phys Anthropol. 1998;106:425–465. https://doi.org/10.1002/(SICI)1096-8644(199808)106:4<425::AID-AJPA2>3.0.CO;2-I

26. Keyser AW, Menter CG, Moggi-Cecchi J, Pickering TR, Berger LR. Drimolen: A new hominid-bearing site in Gauteng, South Africa. S Afr J Sci. 2000;96:193–197.

27. Moggi-Cecchi J, Menter CG, Boccone S, Keyser A. Early hominin dental remains from the Plio-Pleistocene site of Drimolen, South Africa. J Hum Evol. 2010;58:374–405. https://doi.org/10.1016/j.jhevol.2010.01.006

28. Herries IR, Curnoe D, Adams JW. A multi-disciplinary seriation of early *Homo* and *Paranthropus* bearing palaeocaves in southern Africa. Quat Int. 2009;202:14–28. https://doi.org/10.1016/j.quaint.2008.05.017

29. Herries IR, Shaw J. Palaeomagnetic analysis of the Sterkfontein palaeocave deposits: Implications for the age of the hominin fossils and stone tool industries. J Hum Evol. 2011;523–539. https://doi.org/10.1016/j.jhevol.2010.09.001

30. Aguirre E. Identificación de ‘Paranthropus’ en Makapansgat [Identification of ‘Paranthropus’ in Makapansgat]. Crónica del XI Congreso Nacional de Arqueología (Mérida). 1969;98–124. Spanish.

31. Johanson DC, White TD. A systematic assessment of early South African hominids. Science. 1979;201:321–330. https://doi.org/10.1126/science.104384

32. Rak Y. The australopithecine face. New York: Academic Press; 1983. https://doi.org/10.1016/B978-0-12-576280-9.50006-7

33. Pickering R, Dirks PHGM, Jinnah Z, De Ruiter DJ, Churchill SE, Herries AIR, et al. *Australopithecus sediba* at 1.977 Ma and implications for the origins of the genus *Homo*. Science. 2011;333:1421–1423. https://doi.org/10.1126/science.1203697

34. Kimbel WH, Johanson DC, Rak Y. Systematic assessment of a maxilla of *Homo* from Hadar, Ethiopia. Am J Phys Anthropol. 1997;103:235–262. https://doi.org/10.1002/(SICI)1096-8644(199706)103:2<235::AID-AJPA8>3.0.CO;2-S

35. Villmoare B, Kimbel WH, Seyoum C, Campisano CJ, DiMaggio EN, Rowan J, et al. Early *Homo* at 2.8 Ma from Ledi Geraru, Afar, Ethiopia. Science. 2015;347:1352–1355. https://doi.org/10.1126/science.aaa1343

36. Walker A, Leakey R, editors. The Nariokotome *Homo erectus* skeleton. Berlin: Springer; 1993. https://doi.org/10.1007/978-3-662-10382-1

Darting towards Storm Shelter: A minute dinosaur trackway from southern Africa

AUTHOR:
Emese M. Bordy[1]

AFFILIATION:
[1]Department of Geological Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Emese Bordy

EMAIL:
emese.bordy@uct.ac.za

DATES:
Received: 01 Nov. 2020
Revised: 09 Dec. 2020
Accepted: 11 Dec. 2020
Published: 28 May 2021

HOW TO CITE:
Bordy EM. Darting towards Storm Shelter: A minute dinosaur trackway from southern Africa. S Afr J Sci. 2021;117(5/6), Art. #9145. https://doi.org/10.17159/sajs.2021/9145

ARTICLE INCLUDES:
☒ Peer review
☒ Supplementary material

DATA AVAILABILITY:
☒ Open data set
☐ All data included
☐ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Margaret Avery

KEYWORDS:
tridactyl footprints, *Grallator*-like tracks, small theropods, southern Gondwana, Jurassic Clarens Formation

FUNDING:
South African National Research Foundation (grant no. 113394] and DST-NRF Centre of Excellence in Palaeosciences (grant no. CoE PAL 2020)

Theropod dinosaurs are considered the main terrestrial carnivores in the Jurassic and Cretaceous. Their rise to dominance has been linked to, among others, body size changes in their early history, especially across the Triassic–Jurassic boundary. However, to qualitatively assess such temporal trends, robust skeletal and trace fossil data sets are needed globally. The richly fossiliferous southern African continental rock record in the main Karoo Basin offers an unparalleled perspective for such investigations. Herein, by documenting a newly discovered Early Jurassic trackway of very small, functionally tridactyl tracks near Storm Shelter (Eastern Cape) in South Africa, the track record can be expanded. Based on ichnological measurements at the ichnosite and digital 3D models, the footprint dimensions (length, width, splay), locomotor parameters (step length, stride, speed), and body size estimates of the trackmaker are presented. In comparison to other similar tracks, these footprints are not only the smallest *Grallator*-like tracks in the Clarens Formation, but also the most elongated dinosaur footprints in southern Africa to date. The tracks also show that the small-bodied bipedal trackmaker dashed across the wet sediment surface at an estimated running speed of ~12.5 km/h. During the dash, either as a predator or as a prey, the trackmaker's small feet sunk hallux-deep into the sediment. The tracking surface is overgrown by fossilised microbial mats, which likely enhanced the footprint preservation. Based on track morphometrics and the regional dinosaur skeletal record, the trackmakers are attributed to *Megapnosaurus rhodesiensis* (formerly *Syntarsus rhodesiensis*), a small-to-medium-sized, early theropod common in southern Africa.

Significance:

- A newly discovered Early Jurassic theropod trackway in South Africa contains not only the smallest tracks in the Clarens Formation, but also the most elongated dinosaur footprints in southern Africa to date.
- The tracks show that the small bipedal trackmaker dashed across the wet sediment surface at an estimated running speed of ~12.5 km per hour.
- During the run, the trackmaker's feet sunk so deeply into the sediment that even the forwards-directed halluces were impressed.

Introduction

Mounting evidence for body size changes in theropods[1-3], a class of carnivorous dinosaurs, during the dawn of the dinosaurs is increasingly placing the southern African fossil record into the focus of global palaeontological studies. Late Triassic to Early Jurassic dinosaur footprints, for which southern Gondwana, especially southern Africa, is an excellent archive, are particularly useful proxies for such biogeological investigations. In particular, the Lower Jurassic of southern Africa (i.e. the upper Stormberg Group; Figure 1) is key in this regard as: (1) it was deposited in a semi-arid continental ecosystem of rivers, lakes and deserts; (2) it is richly fossiliferous, and (3) it provides clues on how life recovered after the global biodiversity crisis event that occurred ~200 million years ago.[4,5]

This study reports on the smallest dinosaur footprints, forming a single trackway, in the Clarens Formation (uppermost Stormberg Group; Figure 1) of South Africa and shows that these tracks are also, to date, the most elongated dinosaur footprints from the Early Jurassic of southern Africa. These tiny, slender tracks in the lower Clarens Formation are part of the Lower Jurassic footprint assemblage, which in southern Africa is contained in the upper Elliot Formation (Hettangian–Sinemurian) and in the comformably overlying Clarens Formation (Sinemurian–Pliensbachian).[4] These two rock units also contain the two largest (footprint lengths: 55–57 cm)[6-9] and the smallest (footprint length: ~6.5 cm)[10] tridactyl tracks on record thus far, all reported from Lesotho. Because the current tracks are part of the ichnofauna of the lower Clarens Formation, which is likely Sinemurian[4], they enrich the Early Jurassic track record locally and globally, and thus can contribute to the evaluation of the temporal patterns in the footprint record, which, in turn, can add to the ongoing debate on dinosaur body size changes across the Triassic–Jurassic boundary[1-3,11].

Geological background

The dinosaur footprints investigated in this study were discovered north of Maclear in the Eastern Cape of South Africa (Figure 1) by a local resident, Mrs Adele Moore, and her field party in 2014 along a footpath leading to the Storm Shelter archaeological site.[13] Found at the same altitude as the rock art site, the very fine-grained sandstone slab, containing the single trackway of five well-impressed dinosaur footprints, is part of the lowermost Clarens Formation (Figure 1). Here, at Storm Shelter, and throughout the main Karoo Basin (South Africa and Lesotho), the basal Clarens Formation, which is likely Sinemurian[4,14] in age, is characterised by tabular, fine- to medium-grained sandstone beds and subordinate mudstones. Suites of sedimentary structures (e.g. horizontal lamination, cross-bedding, ripple cross-lamination, desiccation cracks; Supplementary figure 1) in these rocks point to deposition in decelerating powerful traction currents and stagnant water.[15,16] Overall, these rocks indicate that the ancient wet aeolian ecosystem was prone to intermittent flash flooding and drying episodes.[15,16] While being less fossiliferous than the conformably underlying Hettangian to Sinemurian upper Elliot Formation, the palaeontological record of the Clarens Formation is diverse, both in vertebrate skeletal remains and trace fossils[4,6,14,17,18], particularly in footprints, and most notably in dinosaur tracks, as recently summarised by Mukaddam et al.[19]

Methods

The footprints investigated in this study were subjected to morphometric measurements as shown in Figure 2. Moreover, for comparative purposes, four additional footprints collected from Lesotho, two *Grallator*-like (LES 283 and 288) and two *Anomoepus*-like (LES 111 and 112) tracks from the Clarens and upper Elliot Formations, respectively, were also briefly analysed for this study. These four Lesotho tracks were collected by Dr Paul Ellenberger pre-1970, and are now housed in the Ellenberger Collection at the University of Montpellier (France).

The ichnological morphometric measurements and the standard proxies derived from them, which estimate the trackmaker's hip height, gait, body length and speed, are summarised in Table 1 and detailed in Supplementary tables 1–5 (with references). The relevant *modus operandi* is explained elsewhere[5,8,9,10,19] and in references therein. The track measurements were taken physically in the field using a calliper. After photographing each track with a Nikon D5100 digital camera (focal length 50 mm), individual photogrammetric 3D models and false-colour depth maps were generated with Agisoft Metashape Professional (version 1.6.4) and CloudCompare (v.2.11.0) software packages, respectively. Following the standard protocol of Falkingham et al.[20], the photogrammetric 3D models and relevant raw data are publicly available online via this open access data repository link: https://doi.org/10.6084/m9.figshare.13007240

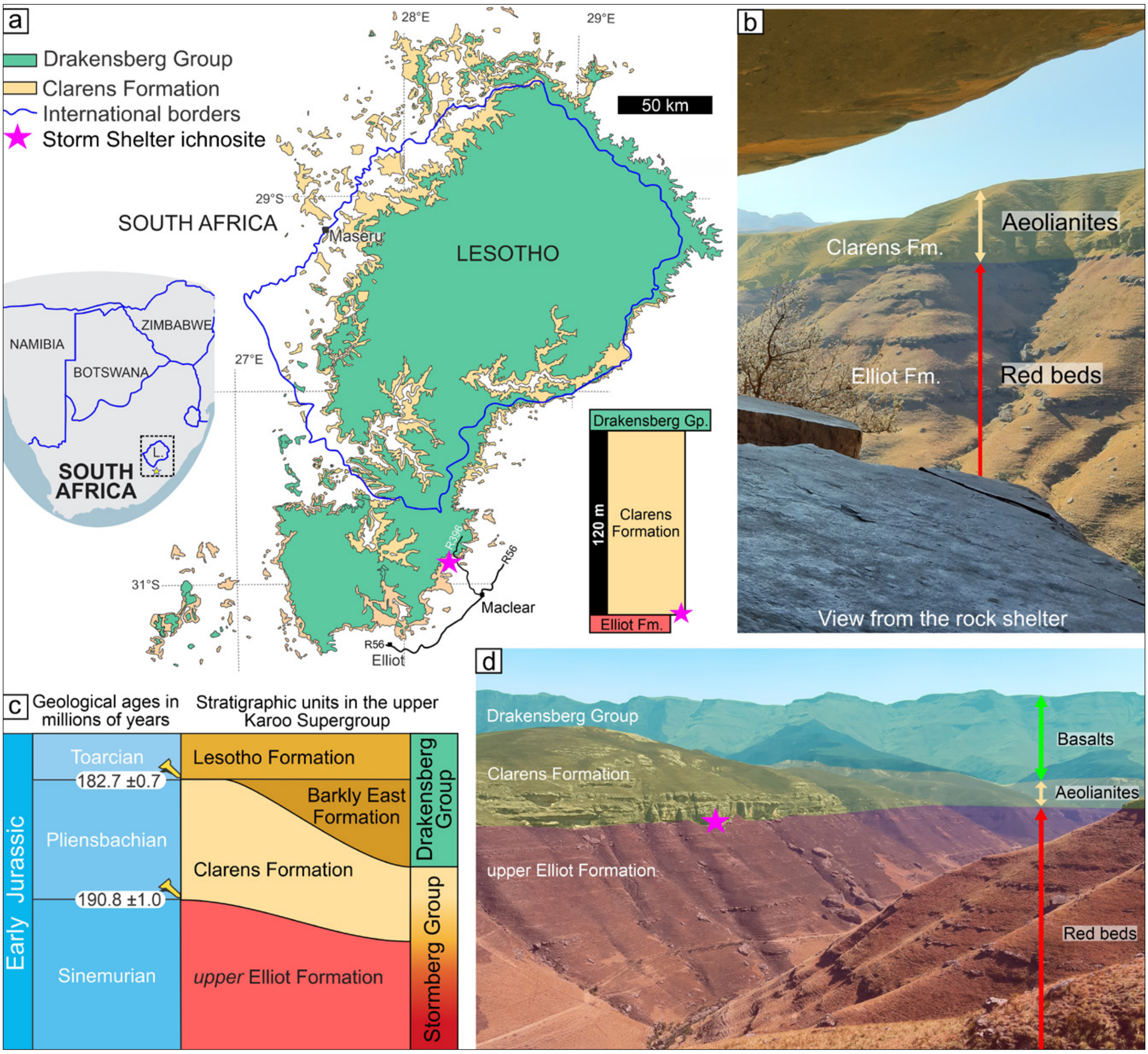
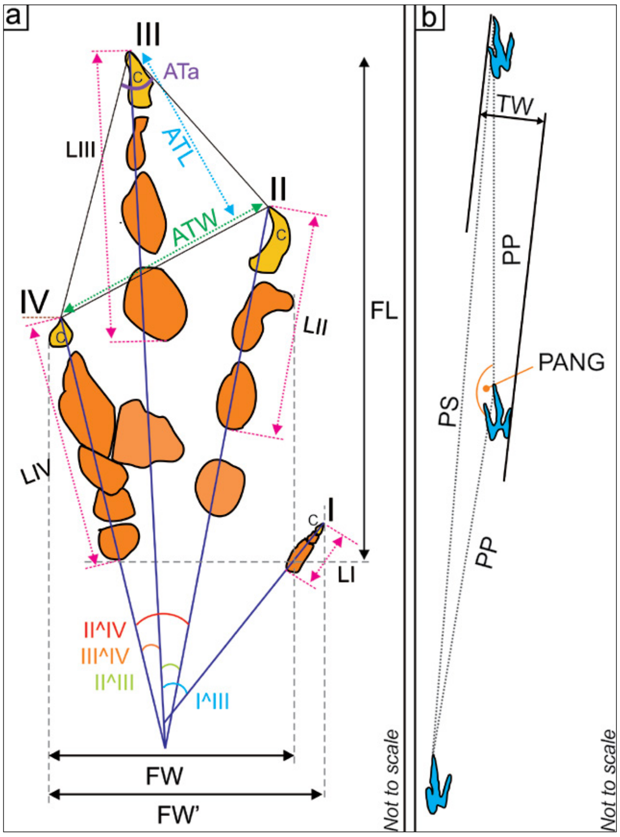


Figure 1: Geological context of the Storm Shelter ichnosite. (a) Simplified geological map of the upper part of the Karoo Supergroup in the Republic of South Africa and Lesotho, showing the position of the study locality northwest of Maclear (Eastern Cape). Inset log shows the local thickness of the Clarens Formation and the relative position of the ichnosite within it. (b–d) Stratigraphic details of the ichnosite within the upper Karoo Supergroup. Map derived by combining data from Johnson and Wolmarans[12] and own mapping.



Roman numerals I–IV denote digits; L, length of digit; C, claw trace; FL, foot length; FW, foot width; FW', foot width including digit I; ATW, anterior triangle width; ATL, anterior triangle length; Ata, anterior triangle angle; interdigital angles between digits I ^ III, II ^ III, III ^ IV and II ^ IV; TW, trackway width; PP, pace; PS, stride; PANG, pace angulation

Figure 2: Key morphological features of the tracks and their measurements obtained for the ichnological analyses: (a) track measurements and (b) trackway measurements.

Results

The Storm Shelter ichnosite preserves a single trackway of five, very small, narrow, elongated and strongly mesaxonic footprints (Figure 3, Table 1, Supplementary table 1) with the following main parameters (average values): length 7.5 cm, width 3.6 cm, length/width ratio >2.1. The ratios of the anterior triangle[21] length to width and the anterior triangle length to track width are 1 and 0.9, respectively. Manus tracks were not observed.

These elongate, digitigrade tracks preserve evidence for claw marks at the tips of the three well-impressed digits and one lightly impressed digit, indicative of a hindfoot that was functionally tridactyl and featured a well-developed hallux (digit I). The slender digit impressions only sporadically show discrete, oval digital pads, most likely due to extensive sediment collapse within the tracks, which would also suggest that these are likely penetrative tracks[22,23] (i.e. the foot penetrated the substrate much more deeply than is apparent from the track outlines on the tracking surface). The penetrative nature of the tracks is further supported by the very narrow distal portion of digit III. Relative to digits II and IV, digit III is better impressed and extends beyond digits II and IV, which are subequal. Both the apex of the anterior triangle and the divarication between digits II and IV form acute angles of 54° and 29°, respectively. On average, digit I is ∼1 cm long and its long axis encloses an acute angle of 54° with the long axis of digit III. The variability of the position of the digit I impression relative to the other digits (Figure 3, Table 1) indicates that the hallux did not have any functional use for locomotion or stability.

The single trackway that these five pes tracks form is narrow (∼6.1 cm), and relative to the footprint length of 7.5 cm, preserves long strides of nearly 1 m (Figure 4, Table 1, Supplementary table 2). The long axis of digit III is essentially parallel to the trackway axis. The average pace is 47.6 cm with an angulation of 172°, and the track width to trackway width ratio is 59%, which define a narrow-gauge trackway. Along the trackway, the morphological quality of the tracks is uniform, although well-developed wrinkle structures, a type of microbially induced sedimentary structure, are present both inside and outside the tracks on the tracking surface (Figure 4 and Supplementary figure 1), and interfere with the outlines of some of the tracks (e.g. see the posterior of track #1 in Figure 3). The microbially induced sedimentary structures and the digit I impression demonstrate that the tracks are not underprints, but rather primary, penetrative footprints impressed directly and deeply on the tracking surface.

Because of the incomplete/missing digital pad impressions, some ill-defined imprint walls, and superimposition of microbial structures, all five Storm Shelter tracks rank low, around 1.5 on the recently proposed[24,25], four-point (0-1-2-3) track-grading scale. Because of the suboptimal morphological quality of the digital pad impressions, only a very tentative phalangeal formula of 1(?2)-2-3(?4)-4-0 (not counting the claw traces) is inferred from the collective assessment of the five tracks.

Using standard ichnological formulae[26,27] (detailed in Supplementary tables 3 and 4), the trackmaker's hip height, body length, body mass, gait and speed were estimated as summarised in Table 1. Based on these parameters, which are essentially derived from the average foot length and stride values, the trackmaker's hip height was ∼32 cm, its body length was ∼120 cm and it weighed ∼3–4 kg (Table 1, Supplementary table 3). The body size of this digitigrade, obligate bipedal trackmaker thus is comparable to small-sized, long-legged, ground-dwelling extant birds, which fall into a size class between a very tall domestic rooster and a bustard or a wild turkey (hip height 0.39 m and body mass of ∼5–7 kg; cf.[28]). The track data are insufficient to determine whether the small trackmaker was a juvenile or a small-sized but fully grown adult.

The trackmaker's allometric and morphometric gait values are 3.2 and 2.9, respectively, indicating a constant 'running' gait[26,27] across the sediment surface (Table 1, Supplementary tables 3 and 4). The allometric and morphometric speed estimations show that the trackmaker ran at an average speed of 3.45 m/s (i.e. ∼12.42 km/h). The relatively rapid locomotion of the trackmaker is also suggested by the narrow trackway, as during fast movement, running bipedal animals place their feet closer to the midline (cf.[29]). The stride/footprint length ratio for the Storm Shelter trackway is 12.9 and thus is only slightly greater than the median value[30] for trackways in this size category (i.e. foot length of 5–10 cm) from the early Mesozoic and attributed to theropods. This suggests that the trackmaker was moving in a typical manner for a small theropod.

Discussion

The morphological characteristics describing the Storm Shelter footprint proportions (Table 1) closely resemble those of *Grallator*-like tracks; however, due to the moderate to low quality of these tracks, their ichnotaxonomic treatment is not warranted here. *Grallator* is an ichnogenus that refers to globally occurring tracks of bipedal theropod dinosaurs of early Mesozoic age. *Grallator* tracks are small (<15 cm long), elongate with a footprint length/width ratio of ≥2, narrow digit divarication angles (10–30°) of the slender, pointy toes and a great anterior projection of digit III (i.e. mesaxonic tracks; e.g.[21,26,31,32]). Moreover, *Grallator* tracks are digitigrade, often preserve claw traces and occasionally also the impression of digit I[28], and thus the ichnogenus refers to functionally tridactyl tracks.

In the early Mesozoic ichnological record of the main Karoo Basin of southern Africa, tracks referable to *Grallator* are not rare.[6-10,33-37] Being only ∼1 cm longer than the smallest *Grallator* tracks in southern Africa (described[10] from the lower part of the upper Elliot Formation in Lesotho), the Storm Shelter tracks, with their average footprint length of 7.5 cm, are among the smallest *Grallator*-like tracks in the Lower Jurassic of this region, and certainly the smallest of all dinosaur tracks in the Clarens Formation to date.

Table 1: Summary of the morphometric measurements and derived parameters of the Storm Shelter tracks. See Supplementary tables 1–4 for full morphometric measurements and calculations. Length measurements are in centimetres.

Track #		**#1**	**#2**	**#3**	**#4**	**#5**	***Average***
Footprint length (FL)		7.47	7.85	7.05	7.47	7.76	7.5
Footprint width (FW)		3.06	3.71	3.38	4.37	3.33	3.6
Footprint width' (FW')		3.21	4.39	3.81	3.35	3.33	3.6
FL/FW ratio		2.4	2.1	2.1	1.7	2.3	2.1
AT angle (degrees)		52	64	57	45	49	54
ATL		3.04	2.79	2.84	3.92	3.44	3.2
ATW		3.09	3.69	3.24	3.23	3.13	3.3
ATL/ATW ratio		1.0	0.8	0.9	1.2	1.1	1.0
ATL/FW ratio		1.0	0.8	0.8	0.9	1.0	0.9
II ^ IV (degrees)		26	34	25	32	30	29
II ^ III (degrees)		16	14	14	12	20	15
III ^ IV (degrees)		11	20	11	21	10	15
I ^ III (degrees)		66	62	45	42	53	54
No. of toes		4	4	4	4	4	4
LI		1.07	1.08	0.87	1.42	1.12	1.1
LII		3.11	4.26	3.16	2.80	3.04	3.3
LIII		4.08	4.51	4.07	5.46	4.22	4.5
LIV		3.77	4.29	3.48	4.05	3.74	3.9
Pace		–	–	–	–	–	47.17
Stride		–	–	–	–	–	96.75
Pace angulation (degrees)		–	–	–	–	–	174.0
Trackway width (TW)		–	–	–	–	–	6.1
Trackway ratio (FW/TWx100)		–	–	–	–	–	59
Allometric	Hip height	–	–	–	–	–	30.2
	Gait	–	–	–	–	–	3.2
	Speed (m/s)	–	–	–	–	–	3.6
Morphometric	Hip height	–	–	–	–	–	33.5
	Gait	–	–	–	–	–	2.9
	Speed (m/s)	–	–	–	–	–	3.3
Body length		–	–	–	–	–	120.6
Body mass (kg)	Allometric	–	–	–	–	–	2.91
	Morphometric	–	–	–	–	–	3.98

AT, anterior triangle; ATL, anterior triangle length; ATW, anterior triangle width; ATL/FW, anterior triangle length to footprint width ratio; II ^ IV, total divarication between digits II and IV; (II ^ III, divarication between digits II and III; III ^ IV, divarication between digits III and IV; I ^ III, divarication between digits I and III; no. of toes, number of visible digit marks on the footprint; LI–LIV: length of digits I, II, III and IV, respectively, inclusive of the claw trace

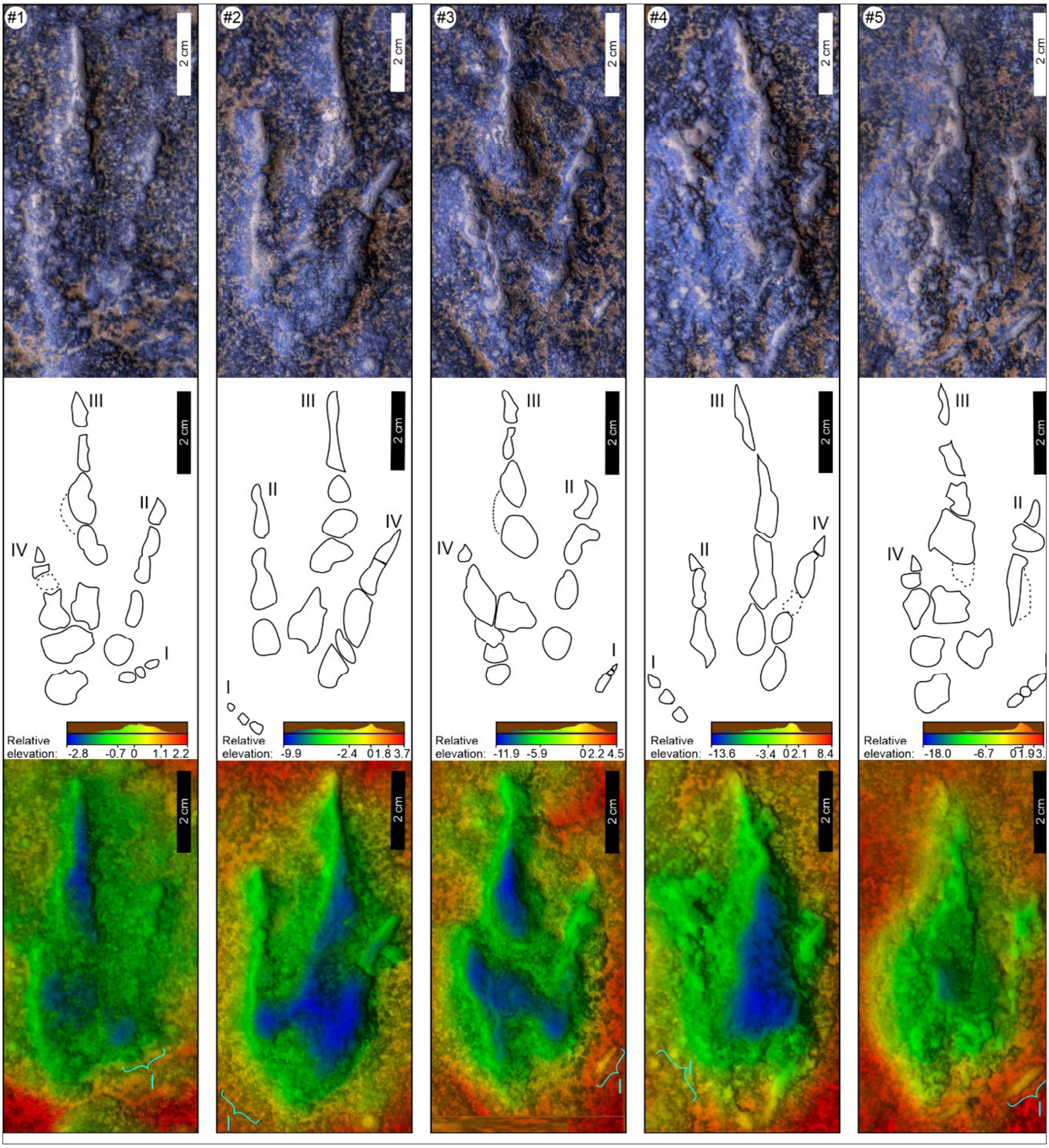

#1
#2
#3
#4
#5
2 cm
III
II
IV
I
Relative elevation: -2.8 -0.7 0 1.1 2.2
Relative elevation: -9.9 -2.4 0 1.8 3.7
Relative elevation: -11.9 -5.9 0 2.2 4.5
Relative elevation: -13.6 -3.4 0 2.1 8.4
Relative elevation: -18.0 -6.7 0 1.9 3.

Figure 3: Five *Grallator* tracks from the lower Clarens Formation at the Storm Shelter ichnosite. Each track (#1–#5) is illustrated with three different images: *top* – orthophoto derived from photogrammetric model; *middle* – interpretive outline drawing; *bottom* – false-colour depth map (dark red is the highest point; blue is the lowest). See Supplementary table 1 for track morphometric measurements and https://doi.org/10.6084/m9.figshare.13007240 for the photogrammetric models. Roman numerals I–IV denote digits; for clarity, the length of digit I is marked with curly brackets on the false-colour depth maps.

In the Clarens Formation, only four tracks that are morphologically similar to *Grallator* were mentioned by Ellenberger[6] from Lesotho: three tracks of similar size (~10 cm long) from zone B5 (his figures 127–129) and one from zone B6 (no illustration). In spite of efforts to relocate these tracks in the field, only the latter track (*G. molapoi* – Leribe-Molapo) as well as *G. (Paragrallator) matsiengensis* (Matsieng, his figure 128) are available for direct study as specimens LES 288 (holotype) and LES 283 (the only replica of the now lost holotype), respectively, in the Ellenberger Collection at the University of Montpellier (France). Because these ichnotaxa are partially preserved, severely obscured by extramorphological features and lacking information regarding the foot anatomy of the trackmaker, they can be considered *nomina nuda,* but being the only tracks identified as '*Grallator*' from the Clarens Formation, the digital 3D models of these tracks are documented here (Figure 5, Supplementary table 5) for sake of completeness. Morphometric measurements show that these '*Grallator*' tracks are >33% longer (average length: 11.3 cm) than the Storm Shelter tracks. The only other footprints in the Clarens Formation that have been likened to *Grallator* are large, ~40-cm-long tridactyl tracks reported[33] from the eastern Free State (Uniondale) in South Africa. However, these large, broad tracks, due to their key track morphological parameters (e.g. length-to-width ratio of 1.2; total divarication of 71°), are not considered here to be attributable to *Grallator* (see [8,21,32,35] and references therein for a more detailed discussion on the ichnotaxonomic considerations of these tracks).

Very small tridactyl tracks from Lesotho have been reported[5,6,10] from the upper Elliot Formation (Hettangian–Sinemurian, zone B1 of Ellenberger). However, with the exception of the minute *Grallator* tracks[10] in the lower part of the upper Elliot Formation at Lephoto in the Roma valley, compared to which the Storm Shelter tracks are similar in length but more elongated, all other very small tridactyl tracks (footprint length <9 cm) from the Elliot Formation seem to be consistent with ichnological parameters of either *Trisauropodiscus*-like or *Anomoepus*-like tracks (Figure 6, Supplementary table 1; also see[5,10]). Compared to these older, likely pre-Sinemurian[4], tridactyl tracks (Figure 6), the Storm Shelter tracks are typically more elongated with narrow digit divarication angles (10–30°) and a greater anterior projection of digit III (i.e. more mesaxonic; see Supplementary table 1). All in all, these comparisons show that, thus far, the Storm Shelter tracks are among the most elongated (i.e. high length/width ratio) *Grallator*-like tracks reported from the Lower Jurassic of southern Africa, and the smallest in the Clarens Formation.

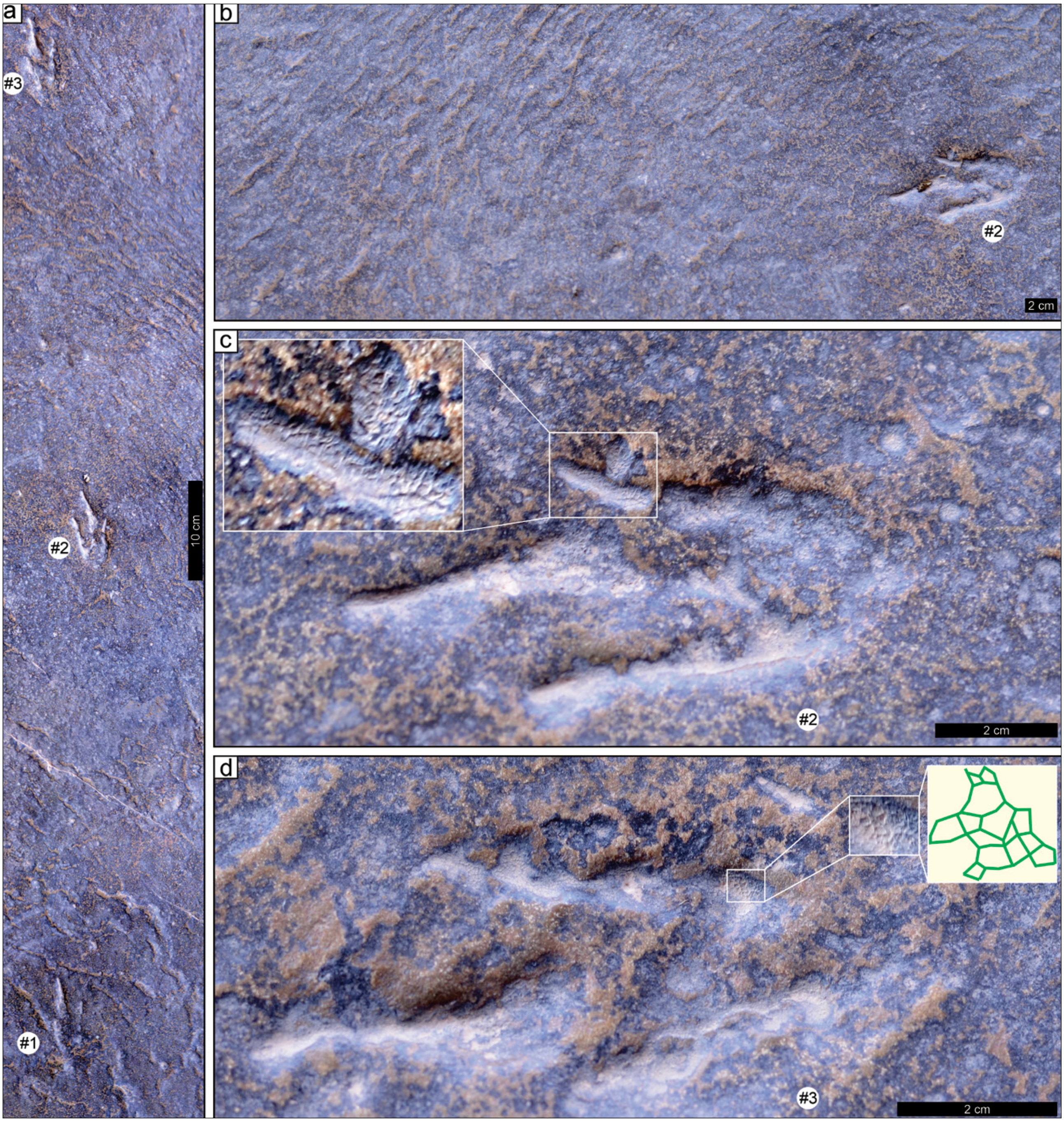


Figure 4: Details of the *Grallator* trackway in the lower Clarens Formation at the Storm Shelter ichnosite. (a) Orthophoto of the first three tracks within the trackway (image derived from photogrammetric model) showing random, parallel as well as curving ridges on the surface, which are interpreted as wrinkle structures – a type of microbially induced sedimentary structure. (b) Details of the tracking surface around track #2. (c and d) Close-up images of track #2 and #3 showing the millimetre-scale polygonal pattern inside small depressions both within and next to the tracks. The small-scale, polygonal pattern is interpreted as dried-out, microbially bound surfaces that draped the entire tracking surface soon after the tracks were formed. See Supplementary table 2 for trackway morphometric measurements, https://doi.org/10.6084/m9.figshare.13007240 for the trackway photogrammetric model and Supplementary figure 1 for further details on the sedimentological context of the ichnosite.

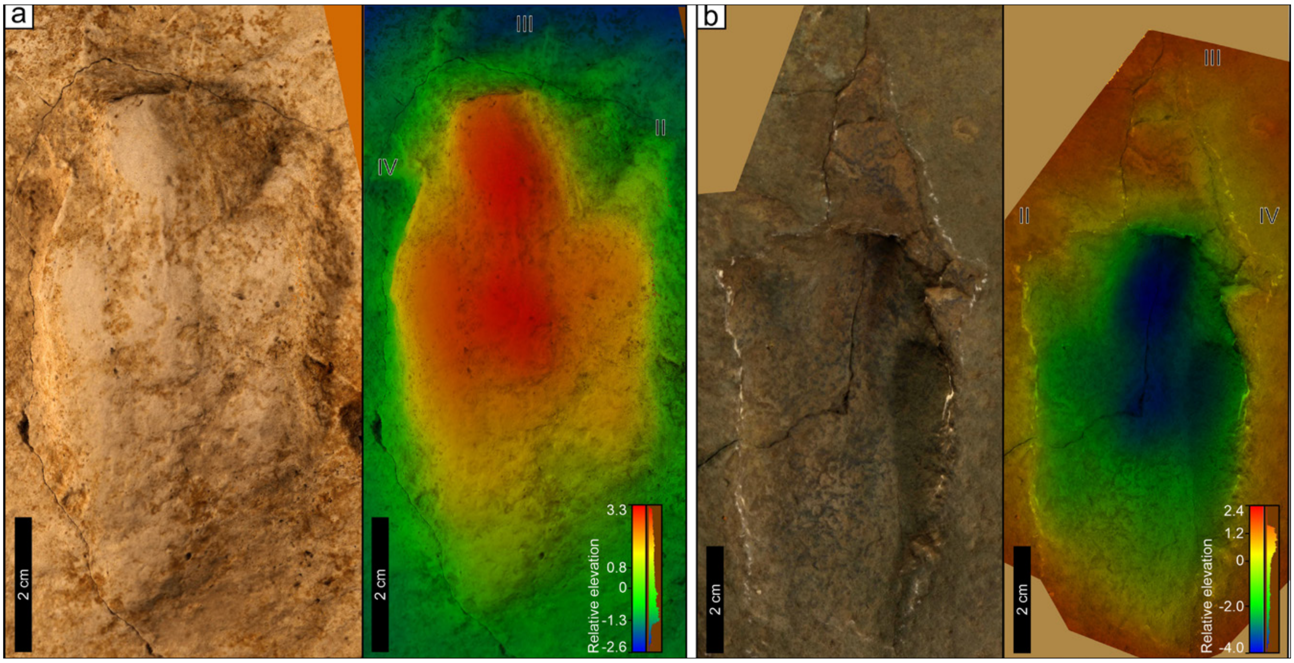


Figure 5: *Grallator* tracks from the Clarens Formation in Lesotho. Left image: orthophoto derived from photogrammetric modelling; right image: false-colour depth model generated using CloudCompare version 2.11.3 (dark red is the highest point, dark blue the lowest). (a) *Grallator (Paragrallator) matsiengensis* ([6,7] figure 128, LES 283) from Matsieng (Zone B/5). This broken, plaster-of-Paris cast is the only replica of the lost holotype. (b) *Grallator molapoi* ([6,7] LES 288, holotype) from Leribe-Molapo (Zone B/6). Both specimens are housed in the Ellenberger Collection at the University of Montpellier (France). See Supplementary table 5 for their morphometric measurements and https://doi.org/10.6084/m9.figshare.13007240 for their photogrammetric models. Roman numerals I–IV denote digits.

Although *Grallator* tracks rarely preserve[28] impressions of digit I, remarkably, all *Grallator*-like tracks at Storm Shelter are associated with hallux (digit I) traces (Figure 3). Moreover, unlike in typical *Grallator* tracks[26,31,39], the impression of the hallux is not directed medially or to the rear, but forward. Figure 3 shows how these ~1-cm-long impressions point in the forward direction, at ~54° (average) away from the long axis of digit III. The hallux impressions are not only important for identifying the trackmaker (see below) but also for determining that the trackmaker's feet sunk hallux-deep into the wet, soft sediment while the animal ran across the surface. A pliable, squidgy tracking surface also explains why digits II, III and IV in these tracks lack well-defined digital pad impressions, imprint walls, etc. (Figure 3), and thus limit the tracks' ichnotaxonomic assessment and conclusive attribution to any specific trackmaker. The wrinkle structures (Figure 4, Supplementary figure 1), which are linked to fossilised microbial mats and are abundant on this tracking surface, were possibly present already at the time of trackmaking and thus enhanced the footprint generation and overall preservation. Microbially induced sedimentary structures are known to be associated with dinosaur tracks in the southern African[10,34,35] and global[40-42] ichnofossil record. Likely, these sediment-binding biofilms developed as localised algal blooms flourished in shallow, stagnant pools of water that were generated in ephemeral downpours and that evaporated over time (Figure 7). Although the host Clarens Formation at Storm Shelter does preserve desiccation cracks (Supplementary figure 1), there are none preserved on the tracking surface itself, which possibly remained wet, or at least moist, until it was buried by the next layer of flash flood generated sediment (Figure 7).

Grallator is commonly attributed to early theropod dinosaurs that were obligatory bipeds with long, thin toes, including well-developed halluces (e.g.[21,26,31,32]). Based on the morphometric parameters of the Storm Shelter tracks (Figure 3, Table 1, Supplementary table 1) and the regional Early Jurassic bone fossil record, a likely trackmaker candidate is *Megapnosaurus rhodesiensis* (formerly *Syntarsus rhodesiensis* and *Coelophysis rhodesiensis*[39,40]). This early theropod is a coelophysoid dinosaur (Figure 7) that is common in the same stratigraphic interval (i.e. Lower Jurassic) of southern Africa, especially in Zimbabwe (e.g.[43-45]). A similar coelophysoid dinosaur, *Coelophysis bauri,* is often assumed to have made the early Mesozoic *Grallator* tracks in North America (e.g.[26,31,32,46-48]). Both of these early theropods were small-to-medium-sized, lightly built, agile, carnivorous bipeds and are often found in richly fossiliferous strata where entire animal populations are jumbled together as partially intact or complete skeletons that also include articulated foot bones.[43] Foot reconstructions show that these early theropods had four pedal digits and a pes phalangeal formula (inclusive of unguals) of 2:3:4:5:0.[43] Reconstructions also show that digit I (the hallux; Figure 7b) did not touch the ground during the steeply digitigrade locomotion of the animal (Figure 7a; e.g.[43,48,49]). Moreover, the well-formed hallux and associated foot bones of *Megapnosaurus rhodesiensis* preserve anatomical features that indicate that the hallux was forwards directed[26,43], alongside digit II (Figure 7b). Raath[43(p.93)] remarked that the hallux 'retained a specialised function in life. Its use as a grooming accessory seems quite feasible'. The congruence in the morphological features of this foot skeleton and the tracks (including the non-reversed hallux configuration) is used here to suggest that the Storm Shelter footprints likely belong to *Megapnosaurus rhodesiensis*. It is noteworthy that in other early Mesozoic *Grallator* tracks[26,31,39,48], the orientation of the hallux impression is reversed (posterior directed) in contrast to the original anatomical configuration. This backward-directed hallux mark resulted from the way the forward-directed hallux interacted with the sediment during the track-generating process as demonstrated by Gatesy et al.[48] In the Storm Shelter tracks, however, the hallux impression closely corresponds to the typical anatomical (forward) orientation of the digit I in coelophysoid dinosaurs[43,48], and this makes the Storm Shelter tracks unique among *Grallator*-like tracks. It is probable that this anterior-directed hallux impression resulted from the fortuitous combination of the microbially influenced substrate consistency that allowed the penetrative tracks to form, the trackmaker's tiny body proportions and the fast speed of motion.

Theropods were the main terrestrial carnivores in the post-Triassic part of the Mesozoic. Their smaller bodied varieties, like the trackmaker of the Storm Shelter tracks, most probably took on the dual role of predator and prey, and thus had good reason to leave behind tracks indicative of a running gait. The ability to run and occasionally sprint at speeds up to 12.5 km/h, even in the smallest individuals, must have been a great advantage for these agile, highly successful predators[43] that had to adapt to an increasingly harsh, desert ecosystem, prone to flash flooding and dry spells in the Early Jurassic of southern Africa[4] (Figure 7).

The minute, elongated Storm Shelter tracks together with the region's smallest[10] and two largest[6-9] tridactyl tracks are taken as evidence, not only of the abundance but also the size diversity of the Early Jurassic theropod dinosaurs in southern Africa. Given the regional abundance and diversity of the footprint record, the dinosaur tracks of southern Africa, in all shapes and sizes, remain an important proxy for meaningfully assessing concepts on macroevolutionary changes in dinosaur body size during the early Mesozoic.[1-3,50] The true potential of this rich ichnological record is only achievable if the collected, but largely undescribed materials in various museum collections, as well as new discoveries like this one at the Storm Shelter, are quantified and integrated with the global early Mesozoic ichnological and osteological fossil records.

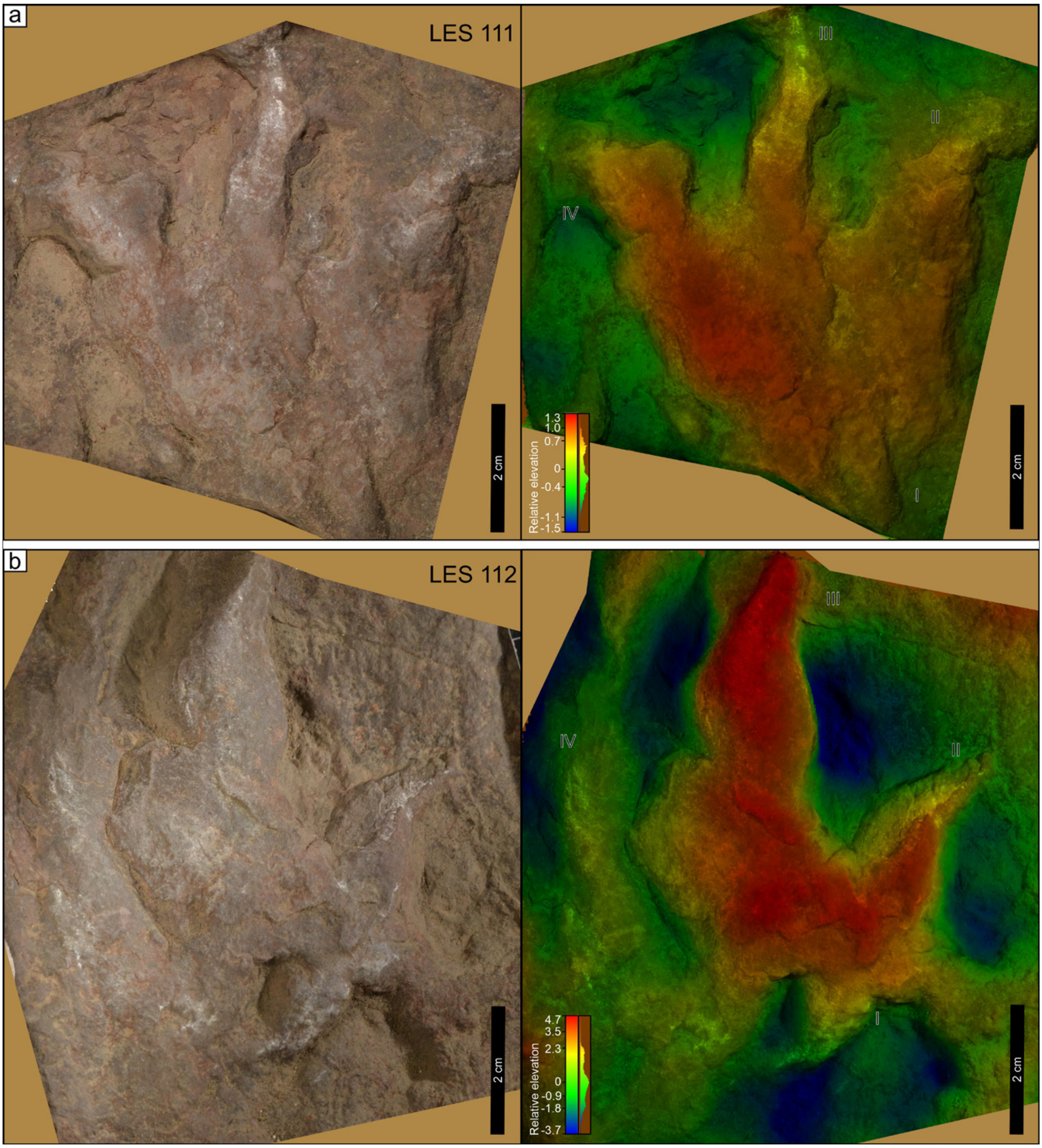
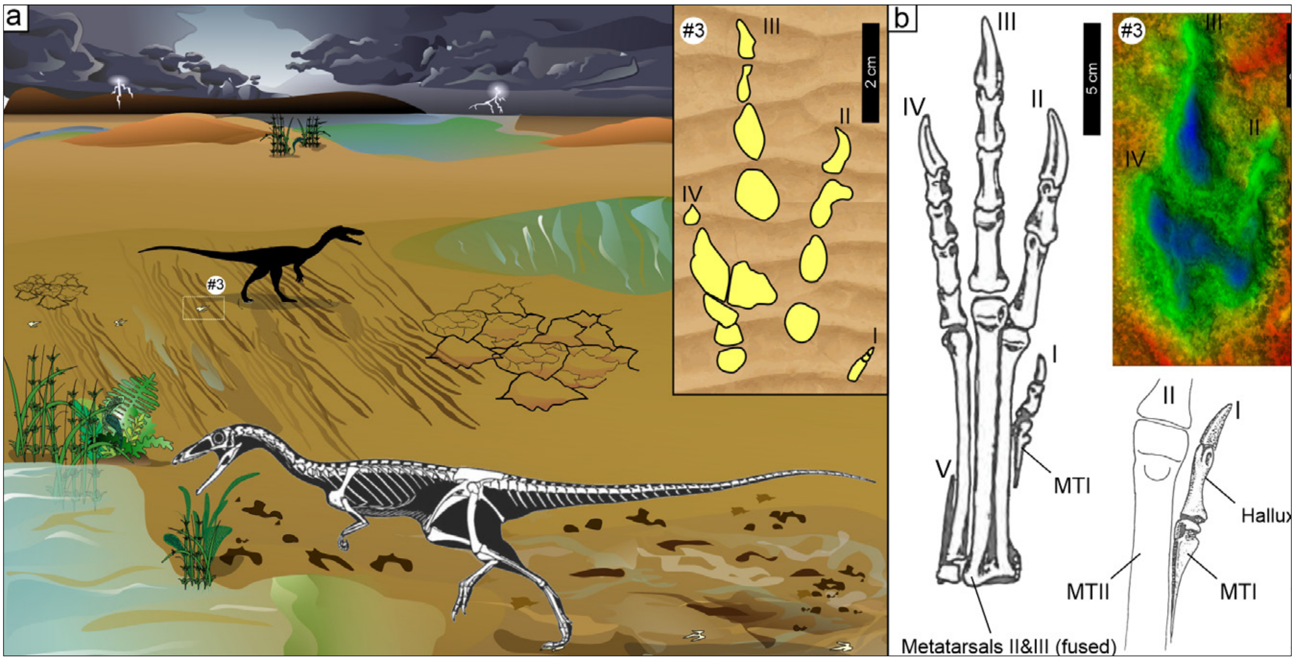


Figure 6: Two examples of small, ~7.7-cm-long tridactyl tracks from the Lower Jurassic of southern Africa. Left image: orthophoto derived from photogrammetric modelling; right image: false-colour depth model generated using CloudCompare version 2.11.3 (dark red is the highest point, dark blue the lowest). (a and b) *Masitisisauropus palmipes* (LES 111 and LES 112, respectively[38]) from the lower part of the upper Elliot Formation at Mokanametsong in southwest Lesotho (Quthing district). Compared to these tracks, *Grallator* tracks are narrower with more slender toe impressions. Both specimens are housed in the Ellenberger Collection at the University of Montpellier (France). See Supplementary table 5 for their morphometric measurements and https://doi.org/10.6084/m9.figshare.13007240 for their photogrammetric models. Roman numerals I–IV denote digits.



MT, metatarsal; Roman numerals I–V denote digits and corresponding MTs

Figure 7: Interpretation of the *Grallator* trackway in the lower Clarens Formation at the Storm Shelter ichnosite. (a) Palaeoenvironmental reconstruction (modified from original artwork by Akhil Rampersadh and Emese M. Bordy). Dinosaur outlines are adapted from Paul[47] (foreground) and Rampersadh et al.[36] (background). Inset shows the interpretative outline of track #3. (b) Skeletal reconstructions of the left foot and hallux of *Megapnosaurus rhodesiensis* are adopted from Raath[43 (his figure 20)]. Inset shows the relative position of the hallux trace in this left footprint of the *Grallator* trackmaker (false-colour depth map of track #3).

Acknowledgements

I acknowledge Mrs Adele Moore, not only for the discovery of the site in 2014 (in the company of Ms Julie Moore and Mr Neels van der Westhuizen), but also for granting and arranging access to the site and providing companionship during my field work in 2020. I also acknowledge curator Ms Suzanne Jiquel for access to the Ellenberger Collection at the Universitè de Montpellier (France). I am thankful to Fabien Knoll and Raquel López-Antoñanzas, for many insightful discussions on the Early Jurassic palaeontology of southern Africa and for being wonderful hosts during my study visit in Montpellier. Jonah Choiniere broadened my mind about the evolution of theropod halluces. Finally, I thank Robert Nagy and Lilla Bordy-Nagy for their field assistance, feedback and overall support during this family (COVID-19 lockdown) project. This work was supported by the South African National Research Foundation (Competitive Programme for Rated Researches grant number 113394) and DST-NRF Centre of Excellence in Palaeosciences (grant number CoE PAL 2020–2021). The funders had no role in study design, data collection or any other aspect of the research. I received no financial benefit from this research.

Competing interests

I declare that there are no competing interests.

Data availability

The ichnological photogrammetric data that support the findings of this study are openly available in Figshare at https://doi.org/10.6084/m9.figshare.13007240, and include photographs used in the photogrammetric models, and the cleaned and aligned 3D models of the tracks figured in this paper (Figures 3, 5, 6).

References

1. Benson RB, Campione NE, Carrano MT, Mannion PD, Sullivan C, Upchurch P et al. Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. PLoS Biol. 2014;12(5), e1001853. https://doi.org/10.1371/journal.pbio.1001853

2. Griffin CT, Nesbitt SJ. Does the maximum body size of theropods increase across the Triassic–Jurassic boundary? Integrating ontogeny, phylogeny, and body size. Anat Rec. 2020;303(4):1158–1169. https://doi.org/10.1002/ar.24130

3. Marsh AD, Rowe TB. A comprehensive anatomical and phylogenetic evaluation of *Dilophosaurus wetherilli* (Dinosauria, Theropoda) with descriptions of new specimens from the Kayenta Formation of northern Arizona. J Paleontol. 2020;94(S78):1–3. https://doi.org/10.1017/jpa.2020.14

4. Bordy EM, Abrahams M, Sharman GR, Viglietti PA, Benson RB, McPhee BW, et al. A chronostratigraphic framework for the upper Stormberg Group: Implications for the Triassic-Jurassic boundary in southern Africa. Earth-Sci Rev. 2020;203, Art. #103120. https://doi.org/10.1016/j.earscirev.2020.103120

5. Bordy EM, Rampersadh A, Abrahams M, Lockley MG, Head HV. Tracking the Pliensbachian–Toarcian Karoo firewalkers: Trackways of quadruped and biped dinosaurs and mammaliaforms. PLoS ONE. 2020;15(1), e0226847. https://doi.org/10.1371/journal.pone.0226847

6. Ellenberger P. Les niveaux paléontologiques de première apparition des mammifères primordiaux en Afrique du Sud et leur ichnologie: établissement de zones stratigraphiques détaillées dans le Stormberg du Lesotho (Afrique du Sud) (Trias supérieur à Jurassique) [The palaeontological levels of first appearance of primordial mammals in South Africa and their ichnology: Establishment of detailed stratigraphic zones in the Stormberg of Lesotho (South Africa) (Upper Triassic to Jurassic)]. In: Haughton SH, editor. Proceedings of the 2nd IUGS Symposium on Gondwana Stratigraphy and Palaeontology. Pretoria: Council for Scientific and Industrial Research; 1970. p. 343–370. French.

7. Ellenberger P. Contribution à la classification des Piste de vertébrés du Trias: les types du Stormberg d'Afrique du Sud (I) [Contribution to the classification of the Triassic Vertebrate Track: The Stormberg types of South Africa (I)]. Montpellier: Palaeovertebrata, Mémoire Extraordinaire; 1972. French.

8. Sciscio L, Bordy EM, Abrahams M, Knoll F, McPhee BW. The first megatheropod tracks from the Lower Jurassic upper Elliot Formation, Karoo Basin, Lesotho. PLoS ONE. 2017;12(10), e0185941. https://doi.org/10.1371/journal.pone.0185941

9. Abrahams M, Sciscio L, Reid M, Bordy EM. Large tridactyl dinosaur tracks from the Early Jurassic of southern Gondwana–uppermost Elliot Formation, Upper Moyeni, Lesotho. Ann Soc Geol Pol. 2020;90(1):1–26. https://doi.org/10.14241/asgp.2020.0

10. Abrahams M, Bordy EM, Sciscio L, Knoll F. Scampering, trotting, walking tridactyl bipedal dinosaurs in southern Africa: Ichnological account of a Lower Jurassic palaeosurface (upper Elliot Formation, Roma Valley) in Lesotho. Hist Biol. 2017;29(7):958–975. https://doi.org/10.1080/08912963.2016.1267164

11. Lucas SG, Tanner LH. The missing mass extinction at the Triassic–Jurassic boundary. In: Tanner LH, editor. The Late Triassic World: Earth in a time of transition. Cham: Springer International Publishing; 2018. p. 721–825.

12. Johnson M, Wolmarans L. Simplified geological map of the Republic of South Africa and Kingdoms of Lesotho and Swaziland. Pretoria: Council for Geoscience; 2008. https://www.geoscience.org.za/images/DownloadableMaterial/RSA_Geology.pdf

13. Blundell G, Lewis-Williams D. Storm Shelter: An important new rock art find in South Africa. S Afr J Sci. 2001;97(1–2):43–46. https://hdl.handle.net/10520/EJC97281

14. Knoll F. The tetrapod fauna of the Upper Elliot and Clarens formations in the main Karoo Basin (South Africa and Lesotho). Bull Soc Géol Fr. 2005;176(1):81–91. https://doi.org/10.2113/176.1.81

15. Beukes NJ. Stratigraphy and sedimentology of the Cave Sandstone stage, Karoo System. In: Haughton SH, editor. Proceedings of the 2nd IUGS Symposium on Gondwana Stratigraphy and Palaeontology. Pretoria: Council for Scientific and Industrial Research; 1970. p. 321–341.

16. Bordy EM, Head HV. Lithostratigraphy of the Clarens Formation (Stormberg Group, Karoo Supergroup), South Africa. S Afr J Geol. 2018;121(1):119–130. https://doi.org/10.25131/sajg.121.0009

17. Kitching JW, Raath MA. Fossils from the Elliot and Clarens Formations (Karoo Sequence) of the northeastern Cape, Orange Free State and Lesotho, and a suggested biozonation based on tetrapods. Palaeontol Afr. 1984;25:111–125.

18. Viglietti PA, McPhee BW, Bordy EM, Sciscio L, Barrett PM, Benson RBJ, et al. Biostratigraphy of the *Massospondylus* Assemblage Zone (Stormberg Group, Karoo Supergroup), South Africa. S Afr J Geol. 2020;123(2):249–262. https://doi.org/10.25131/sajg.123.0018

19. Mukaddam R, Bordy EM, Lockley MG, Chapelle KEJ. Reviving Kalosauropus, an Early Jurassic sauropodomorph track from southern Africa (Lesotho). Hist Biol. 2020; Art. #1834542. http://dx.doi.org/10.1080/08912963.2020.1834542

20. Falkingham PL, Bates KT, Avanzini M, Bennett M, Bordy EM, Breithaupt BH, et al. A standard protocol for documenting modern and fossil ichnological data. Palaeontology. 2018;61(4):469–480. https://doi.org/10.1111/pala.12373

21. Lockley MG. New perspectives on morphological variation in tridactyl footprints: Clues to widespread convergence in developmental dynamics. Geol Q. 2009;53(4):415–432.

22. Gatesy SM, Falkingham PL. Hitchcock's *Leptodactyli*, penetrative tracks, and dinosaur footprint diversity. J Vertebr Paleontol. 2020;40(3), e1781142. https://doi.org/10.1080/02724634.2020.1781142

23. Falkingham PL, Turner ML, Gatesy SM. Constructing and testing hypotheses of dinosaur foot motions from fossil tracks using digitization and simulation. Palaeontology. 2020;63(6):865–880. https://doi.org/10.1111/pala.12502

24. Belvedere M, Farlow JO. A numerical scale for quantifying the quality of preservation of vertebrate tracks. Dinosaur tracks: the next steps. In: Falkingham PL, Marty D, Richter A, editors. Dinosaur tracks: The next steps. Bloomington, IN: Indiana University Press; 2016. p. 92–98.

25. Marchetti L, Belvedere M, Voigt S, Klein H, Castanera D, Díaz-Martínez I, et al. Defining the morphological quality of fossil footprints. Problems and principles of preservation in tetrapod ichnology with examples from the Palaeozoic to the present. Earth-Sci Rev. 2019;193:109–145. https://doi.org/10.1016/j.earscirev.2019.04.008

26. Thulborn T. Dinosaur tracks. London: Chapman and Hall; 1990.

27. Weems RE. Locomotor speeds and patterns of running behavior in non-maniraptoriform theropod dinosaurs. Bull N M Mus Nat Hist Sci. 2006;37:379–389.

28. Gatesy SM, Biewener AA. Bipedal locomotion: Effects of speed, size and limb posture in birds and humans. J Zool. 1991;224(1):127–147. https://doi.org/10.1111/j.1469-7998.1991.tb04794.x

29. Bishop PJ, Clemente CJ, Weems RE, Graham DF, Lamas LP, Hutchinson JR, et al. Using step width to compare locomotor biomechanics between extinct, non-avian theropod dinosaurs and modern obligate bipeds. J R Soc Interface. 2017;14(132), Art. #20170276. https://doi.org/10.1098/rsif.2017.0276

30. Farlow JO, Coroian D, Currie PJ. Noah's ravens: Interpreting the makers of tridactyl dinosaur footprints. Bloomington, IN: Indiana University Press; 2018.

31. Olsen PE, Smith JB, McDonald NG. Type material of the type species of the classic theropod footprint genera *Eubrontes*, *Anchisauripus*, and *Grallator* (Early Jurassic, Hartford and Deerfield basins, Connecticut and Massachusetts, USA). J Vertebr Paleontol. 1998;18(3):586–601. https://doi.org/10.1080/02724634.1998.10011086

32. Lucas SG, Klein HE, Lockley MG, Spielmann JA, Gierlinski GD, Hunt AP, et al. Triassic–Jurassic stratigraphic distribution of the theropod footprint ichnogenus Eubrontes. Bull N M Mus Nat Hist Sci. 2006;37:86–93.

33. Raath MA, Yates AM. Preliminary report of a large theropod dinosaur trackway in Clarens Formation sandstone (Early Jurassic) in the Paul Roux district, northeastern Free State, South Africa. Palaeontol Afr. 2005;41:101–104.

34. Wilson JA, Marsicano CA, Smith RM. Dynamic locomotor capabilities revealed by early dinosaur trackmakers from Southern Africa. PLoS ONE. 2009;4(10), e7331. https://doi.org/10.1371/journal.pone.0007331

35. Sciscio L, Bordy EM, Reid M, Abrahams M. Sedimentology and ichnology of the Mafube dinosaur track site (Lower Jurassic, eastern Free State, South Africa): A report on footprint preservation and palaeoenvironment. PeerJ. 2016;4, e2285. https://doi.org/10.7717/peerj.2285

36. Rampersadh A, Bordy EM, Sciscio L, Abrahams M. Dinosaur behaviour in an Early Jurassic palaeoecosystem – uppermost Elliot Formation, Ha Nohana, Lesotho. Ann Soc Geol Pol. 2018;88:163–179. https://doi.org/10.14241/asgp.2018.010

37. Abrahams M, Bordy EM, Knoll F. Hidden for one hundred years: A diverse theropod ichnoassemblage and cross-sectional tracks from the historic Early Jurassic Tsikoane ichnosite (Clarens Formation, northern Lesotho, southern Africa). Hist Biol. 2020:1–6. https://doi.org/10.1080/08912963.2020.1810681

38. Ellenberger P. Contribution à la classification des pistes de vertébrés du Trias: les types du Stormberg d'Afrique du Sud (II) (Hème partie: le Stormberg Superieur — I. Le biome de la zone B/1 ou niveau de Moyeni: ses biocénoses) [Contribution to the classification of the tracks of Triassic vertebrates: The types of the Stormberg of South Africa (II) (Hème part: the Stormberg Superior - I. The biome of zone B / 1 or level of Moyeni: its biocenoses). Montpellier: Palaeovertebrata, Mémoire Extraordinaire; 1974. French.

39. Weems RE. Evidence for bipedal prosauropods as the likely *Eubrontes* track-makers. Ichnos. 2019;26(3):187–215. https://doi.org/10.1080/10420940.2018.1532902

40. Razzolini NL, Belvedere M, Marty D, Paratte G, Lovis C, Cattin M, et al. *Megalosauripus transjuranicus* ichnosp. nov. A new Late Jurassic theropod ichnotaxon from NW Switzerland and implications for tridactyl dinosaur ichnology and ichnotaxomy. PLoS ONE. 2017;12(7), e0180289. https://doi.org/10.1371/journal.pone.0180289

41. Castanera D, Belvedere M, Marty D, Paratte G, Lapaire-Cattin M, Lovis C, et al. A walk in the maze: Variation in Late Jurassic tridactyl dinosaur tracks from the Swiss Jura Mountains (NW Switzerland). PeerJ. 2018;6, e4579. https://doi.org/10.7717/peerj.4579

42. Marty D, Belvedere M, Razzolini NL, Lockley MG, Paratte G, Cattin M, et al. The tracks of giant theropods (*Jurabrontes curtedulensis* ichnogen. & ichnosp. nov.) from the Late Jurassic of NW Switzerland: Palaeoecological & palaeogeographical implications. Hist Biol. 2018;30(7):928–956. https://doi.org/10.1080/08912963.2017.1324438

43. Raath MA. The anatomy of the Triassic theropod *Syntarsus rhodesiensis* (Saurischia, Podokesauridae) and a consideration of its biology [unpublished PhD thesis]. Grahamstown: Rhodes University; 1978. http://hdl.handle.net/10962/d1002051

44. Bristowe A, Raath MA. A juvenile coelophysoid skull from the Early Jurassic of Zimbabwe, and the synonymy of *Coelophysis* and *Syntarsus*. Palaeontol Afr. 2004;40:31–41.

45. Barta DE, Nesbitt SJ, Norell MA. The evolution of the manus of early theropod dinosaurs is characterized by high inter- and intraspecific variation. J Anat. 2018;232(1):80–104. https://doi.org/10.1111/joa.12719

46. Colbert EH. The Triassic dinosaur *Coelophysis*. Bulletin no. 57. Flagstaff, AZ: Museum of Northern Arizona; 1989.

47. Paul GS. Are *Syntarsus* and the Whitaker quarry theropod the same genus? Bull N M Mus Nat Hist Sci. 1993;3:397–402.

48. Gatesy SM, Middleton KM, Jenkins Jr FA, Shubin NH. Three-dimensional preservation of foot movements in Triassic theropod dinosaurs. Nature. 1999;399(6732):141–144. https://doi.org/10.1038/20167

49. Hattori S. Evolution of the hallux in non-avian theropod dinosaurs. J Vertebr Paleontol. 2016;36(4), e1116995. https://doi.org/10.1080/02724634.2016.1116995

50. Cashmore DD, Butler RJ. Skeletal completeness of the non-avian theropod dinosaur fossil record. Palaeontology. 2019;62(6):951–981. https://doi.org/10.1111/pala.12436

Development and analysis of a long-term soil moisture data set in three different agroclimatic zones of South Africa

AUTHORS:
Lindumusa Myeni[1,2]
Mokhele E. Moeletsi[1,3]
Alistar D. Clulow[2]

AFFILIATIONS:
[1]Agricultural Research Council – Natural Resources and Agricultural Engineering, Pretoria, South Africa
[2]Agrometeorology, School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[3]Risks and Vulnerability Assessment Centre, University of Limpopo, Polokwane, South Africa

CORRESPONDENCE TO:
Lindumusa Myeni

EMAIL:
lindomyeni@gmail.com

DATES:
Received: 22 Jan. 2020
Revised: 02 Sep. 2020
Accepted: 22 Oct. 2020
Published: 28 May 2021

HOW TO CITE:
Myeni L, Moeletsi ME, Clulow AD. Development and analysis of a long-term soil moisture data set in three different agroclimatic zones of South Africa. S Afr J Sci. 2021;117(5/6), Art. #7845. https://doi.org/10.17159/sajs.2021/7845

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☒ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Yali Woyessa

KEYWORDS:
modelling, point-scale, variability, water balance

FUNDING:
EU H2020 Research and Innovation Programme (grant no. 727201), Agricultural Research Council

Understanding the potential impacts of climate variability/change on soil moisture is essential for the development of informed adaptation strategies. However, long-term in-situ soil moisture measurements are sparse in most countries. The objectives of this study were to develop and analyse the temporal variability of a long-term soil moisture data set in South Africa. In this study, a water balance model was used to reconstruct long-term soil moisture data sets from 1980 through 2018, in three sites that represent the diverse agroclimatic conditions of South Africa. Additionally, long-term changes and variability of soil moisture were examined to investigate the potential impacts of climate variability on soil moisture. The results of the Mann–Kendall test showed a non-significant decreasing trend of soil moisture for inland stations at a rate between -0.001 and -0.02 mm per annum. In contrast, a statistically significant (at 5% level of significance) increasing trend of soil moisture for a coastal station at a rate of 0.1131 mm per annum was observed. The findings suggest that the Bainsvlei and Bronkhorstspruit stations located in the inland region are gradually becoming drier as a result of decreasing rainfall and increasing air temperature. In contrast, the Mandeni station located in the coastal region is becoming wetter as a result of increasing rainfall, despite the increase in air temperature. The findings indicate that climate variability is likely to change the soil moisture content, although the influence will vary with region and climatic conditions. Therefore, understanding the factors that affect soil moisture variability at the local scale is critical for the development of informed and effective adaptation strategies.

Significance:

- Long-term modelled estimates were used to investigate the potential impacts of climate variability on soil moisture in three different agroclimatic conditions of South Africa.
- Results show that inland regions are gradually becoming drier as a result of decreasing trends of rainfall and increasing air temperatures while coastal regions are becoming wetter as a result of increasing trends of rainfall.
- This study indicates that climate variability is likely to change soil moisture, although various regions will be affected differently.
- The development of informed adaptation strategies at the local scale is critical to cope effectively with climate variability.

Introduction

Soil moisture plays a critical role in the partitioning of energy fluxes between the land and the atmosphere through its influence on soil reflectivity, emissivity and thermal capacity.[1-3] Soil moisture also plays a critical role in the partitioning of rainfall into different components of the water balance, such as runoff, drainage and soil evaporation through its influence on infiltration rate.[2] Therefore, soil moisture is a key parameter controlling the exchange of carbon, water and energy fluxes between the land and the atmosphere ecosystems.[3-5] Moreover, soil moisture is a key variable that regulates local, regional and global climates through its influence on near-surface air temperatures and feedbacks of rainfall.[5-8] Consequently, soil moisture was identified by the Global Climate Observing System initiative as an essential climate variable.[9]

Soil moisture is a critical parameter in the forecasting and assessment of weather-induced extreme events such as heatwaves, droughts and floods, which are likely to increase in both frequency and intensity as a consequence of the projected climate change in southern Africa.[10-12] Analysis of the trends and variability of the long-term soil moisture data set could be used to detect changes in the water cycle associated with climate change and thus could support climate change modelling and forecasting.[3,4,13-18] Therefore, the long-term soil moisture data set is critical for sustainable agricultural productivity, and efficient management and sustainable use of natural resources within the context of climate change adaptation.[1,16,19,20]

Despite the critical role of soil moisture in weather and climate systems, long-term and representative in-situ soil moisture measurements are sparse in most countries.[3,15,21,22] The scarcity of long-term records of in-situ soil moisture data sets could be attributed to financial constraints that limit the establishment and maintenance of expensive monitoring networks.[13,23] Mittelbach et al.[24] argued that the scarcity of long-term in-situ soil moisture measurements is due to the delayed recognition of the critical role of soil moisture in weather forecasting and climate modelling. In recent years, huge efforts have been undertaken to establish specific soil moisture monitoring networks in some countries to investigate long-term variability in soil moisture and to validate remotely sensed as well as hydrologically modelled soil moisture estimates.[13,15,23-25]

Remotely sensed and hydrologically modelled soil moisture estimates are often used to provide comprehensive soil moisture data sets for weather and climate research studies as a result of the lack of long-term and representative soil moisture measurements.[3,13-18,22,26] Despite the high spatial resolution at a lower cost of remote sensing products, most of the available satellites can only sense very shallow soil depth (2–7 cm) and they have a very poor quality under dense vegetation and mountainous environments.[2,15,20,26,27] On the other hand, long records of weather data of parameters such as air temperature and rainfall are often readily available at good quality in some countries.[13,15,17] Therefore, the use of historical weather data to estimate soil moisture is an alternative and appropriate approach for obtaining long-term soil moisture information.[6,13,17,19,28]

Models have been successfully used to extend and analyse long-term soil moisture data sets within the context of climate change in various countries.[13,15,17,28] However, very few, if any, studies have been conducted to develop and analyse long-term soil moisture data sets under the climatic conditions of South Africa, which was described by Davis and Vincent[11] as the hotspot for climate change. Given the variability of the climatological, biogeographical, pedological and lithological characteristics across South Africa, an understanding of long-term trends and variability of soil moisture is expected to reveal potential impacts of climate change on soil moisture in this region.

Myeni[29] developed and validated a simplified soil moisture model with minimal data input requirements in Bainsvlei, Bronkhorstspruit and Mandeni sites, representing different agroclimatic conditions of South Africa. The findings of Myeni[29] showed that daily soil moisture content can be estimated well from climate data and minimal soil physical properties using a multi-layered soil moisture model, with root mean square error values less than 7.3 mm. These findings gave confidence that this developed model could be reliably used for reconstructing long-term soil moisture data sets with daily temporal resolution under different agroclimatic conditions of South Africa.

In South Africa, most of the in-situ soil moisture measurements have been collected only since 2014, while co-located weather stations have been reporting standard meteorological data since the beginning of the millennium, and in some cases, for some decades prior.[30] We aimed to reconstruct long-term soil moisture data sets from 1980 to 2018 (39 years) using a soil moisture model developed by Myeni[29], at three selected sites that represent different agroclimatic conditions in South Africa. Furthermore, we aimed to address the following pertinent questions: Has the soil moisture changed significantly during the recent last 39 years (1980–2018) in three sites under contrasting agroclimatic conditions of South Africa? And could climate variability and change explain the observed changes in soil moisture at these sites?

Study site description

The study was conducted at three well-calibrated automatic weather stations, situated at Bainsvlei, Bronkhorstspruit and Mandeni, which represent three different agroclimatic zones found in South Africa (Figure 1, Table 1). Distributions of mean monthly rainfall and air temperature (T_{air}) at the three locations are presented in Figure 2. Detailed information about these stations and the measurement descriptions have been reported by Myeni[29].

Methods and materials

Model description

The multi-layered soil moisture model of Myeni[29] was used in this study. In this model, the user divides the profile into layers based on the observed vertical variability in soil physical properties. The daily water balance for the upper layer (i) is calculated as:

$$\theta_{(t)i} = \theta_{(t\text{-}1),i} + P_{(t)} - ET_{(t),i} - R_{(t)} - D_{(t),\mathrm{i}} \qquad \text{Equation 1}$$

where $\theta_{(t),i}$ is the volumetric soil moisture content of the upper layer (mm), $\theta_{(t\text{-}1),i}$ is the volumetric soil moisture content of the upper layer on the previous day (mm), P(t) is the precipitation (mm), ET(t_i is the actual evapotranspiration from the upper layer (mm),R(t) is the total surface runoff (mm) and D(t),$_i$ is the deep drainage from the topsoil layer (mm).

Daily water balance for the bottom layer (i+1) is calculated as:

$$\theta_{(t)\ i+1} = \theta_{(t\text{-}1),\ \mathrm{i}+1} + D_{(t),\mathrm{i}} - ET_{(t),\ i+1} - D_{(t),\ i+1} \qquad \text{Equation 2}$$

where $\theta_{(t),i}$ is the volumetric soil moisture content of the layer i+1 (mm), $\theta_{(t\text{-}1),i+1}$ is the volumetric soil moisture content at layer i + 1 on the previous day (mm) and $D_{(t),i+1}$ is the volume of water exceeding the field capacity of soil layer i + 1.

The model assumes no bare surface evaporation or interception losses as the land cover should always be short grass at standard weather station sites.[3,15,21,22] Furthermore, the model assumes that runoff occurs only when precipitation exceeds the infiltration capacity of the topsoil layer and water in excess of the field capacity storage of the top layer will drain to the bottom layer. The model requires soil water retentivity properties such as wilting point, field capacity and saturation of each soil layer. The model also requires measurements or estimates of reference evapotranspiration (ET_0) in addition to rainfall as climate inputs to estimate daily soil moisture storage at point scale. The detailed model description is given in Myeni[29].

Data collection and processing

Climate data

The daily measurements of solar irradiance (R_S in MJ/m^2), minimum air temperature ($T_{\text{air min}}$ in °C), maximum air temperature ($T_{\text{air max}}$ in °C), minimum relative humidity (RH_{min} in %), maximum relative humidity (RH_{max} in %) and wind speed (U in m/s) for the period between 1979 and 2018 for each station were extracted from the databank of the Agricultural Research Council of South Africa. The choice of this data set was based on the availability of the complete data set which is of sufficient duration to track trends as a result of climate variability as recommended by Burn and Elnur[33]. Retrieved data underwent a data quality control process to identify erroneous, suspicious and implausible data, for example, daily rainfall values greater than 200 mm or less than zero, T_{min} greater than T_{max}; R_s values less than zero or greater than 35 MJ/m^2; relative humidity values less than zero or RH_{min} greater than RH_{max}; and U values less than zero or greater than 10 m/s^{-1}. Furthermore, erroneous, suspicious and impossible values were patched using good quality data from nearby weather stations (within a radius of 100 km) to obtain complete long-term data sets of good quality. An inverse distance weighting method was used to estimate missing or erroneous daily rainfall and RH from neighbouring station data based on the recommendations of Moeletsi et al.[30] The multiple regression method was used to estimate missing or erroneous $T_{\text{air min}}$, $T_{\text{air max}}$ and U values from neighbouring station data based on the recommendations of Shabalala et al.[34] The Hargreaves–Samani equation was used to estimate missing or erroneous daily R_s from measurements of $T_{\text{air min}}$ and $T_{\text{air max}}$ based on the recommendations of Abraha and Savage[35].

Soil characteristics

The number of layers per profile and thickness of each layer were defined based on soil physical properties (Table 2).[29]

Reconstruction of long-term soil moisture data sets

To initialise a soil moisture model, a rainy day between October and December of the year 1979, with a daily rainfall above 25 mm and a total rainfall of three preceding days exceeding 30 mm was identified for each station, assuming soil moisture at field capacity. This is a reasonable assumption as soils are generally wet during this rainy season in these stations. To reconstruct long-term soil moisture data sets, the model was run starting on the identified date using historical climate data and soil properties of each station, with initial soil moisture at field capacity. The estimates of the year 1979 were then discarded, only the remaining 39 years' (1980–2018) estimates were used for analysis purposes. A similar approach was used by DeLiberty and Legates[36] to reconstruct soil moisture data sets from the historical climate data sets using the water balance approach. Estimates of soil moisture storage of each layer were summed into total soil moisture content stored in a profile of 60 cm at each station. Daily soil moisture estimates were then averaged to produce monthly estimates, which were used for analysis purposes.

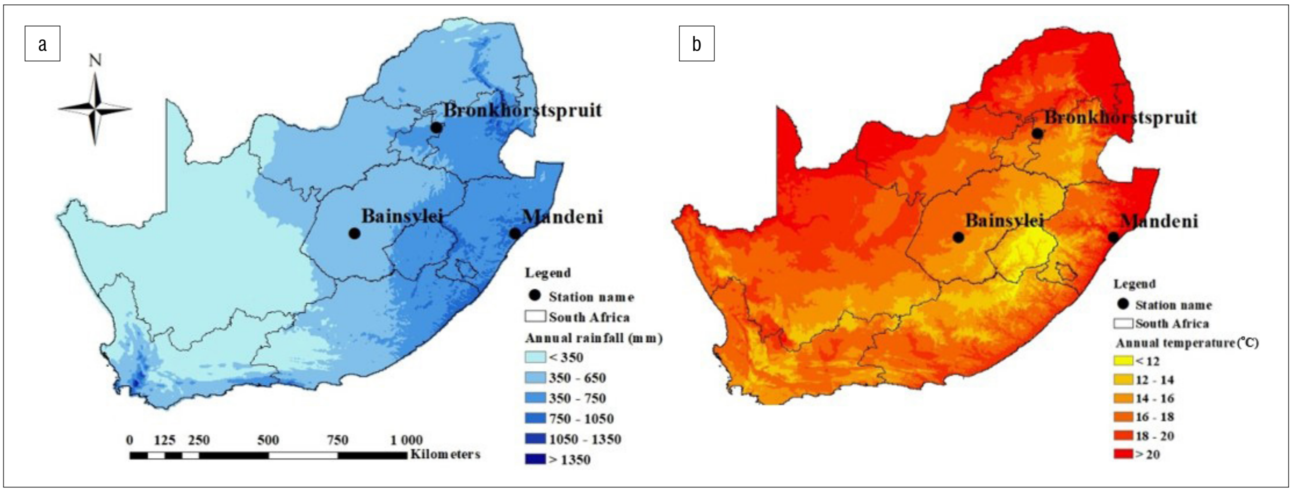


Figure 1: Long-term (a) mean annual rainfall and (b) mean annual air temperature at the soil moisture measurement stations used for model evaluation within South Africa.[31]

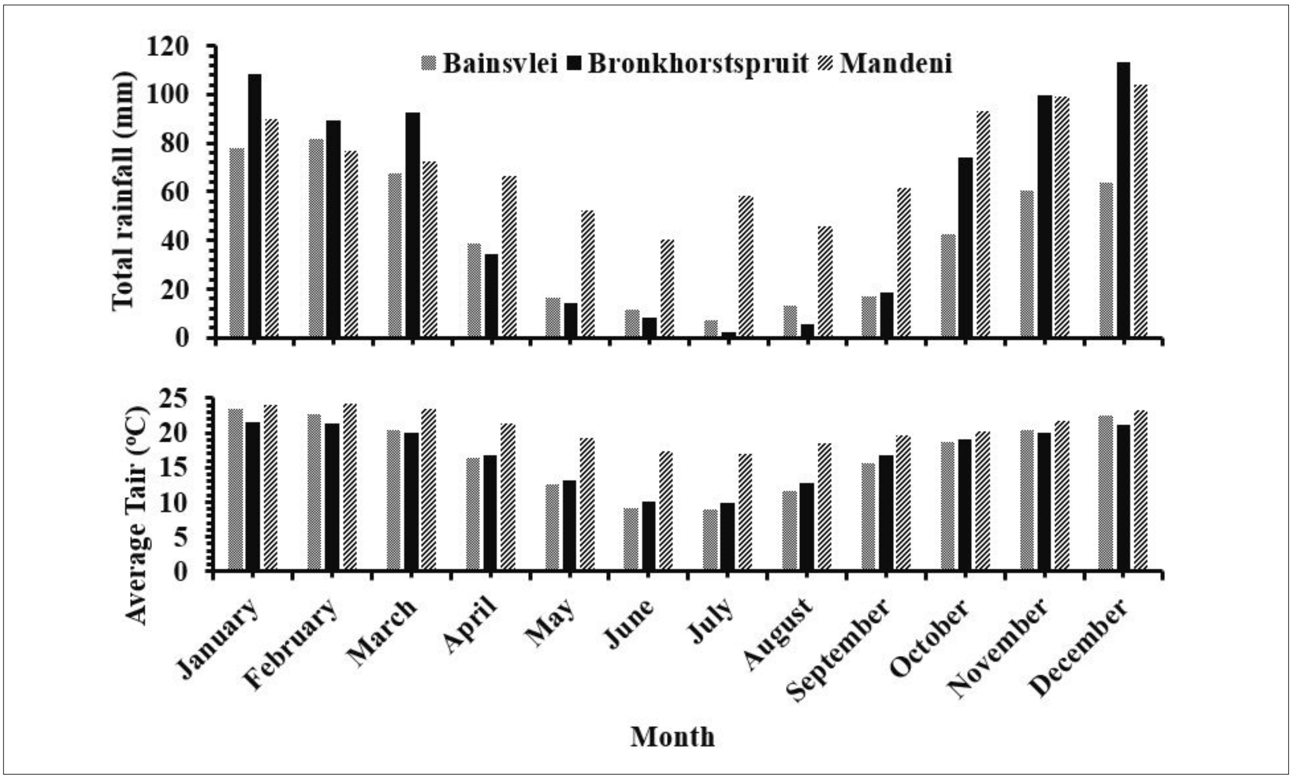


Figure 2: Distribution of monthly rainfall and air temperature (T_{air}) at three weather stations.

Table 1: Characteristics of the three sites in this study

Station name	Latitude (S)	Longitude (E)	Elevation (m)	MAP (mm)	T_{air}(°C)	Climate conditions
Bainsvlei	-29.146	26.146	1290	550	17	Arid, steppe and cold arid
Bronkhorstspruit	-25.702	28.799	1500	677	16	Warm temperate, dry winter and warm summer
Mandeni	-29.156	31.344	107	910	25	Warm temperate, fully humid and hot summer

MAP, mean annual precipitation; T_{air}, mean annual air temperature

Note: The description of climatic conditions was based on the Köppen–Geiger climate classification of Conradie[32].

Table 2: Characteristics of soil layers at all stations

Station	Soil layer	Thickness (cm)	Textural class	θ_{wp}(m^3 m^{-3})	θ_{fc}(m^3 m^{-3})	θ_{sat}(m^3 m^3)
Bainsvlei	1	0-40	Sand	0.05	0.14	0.23
	2	40-60	Sandy loam	0.10	0.24	0.30
Bronkhorstspruit	1	0-15	Sand	0.04	0.14	0.24
	2	15-40	Loamy sand	0.07	0.19	0.28
	3	40-60	Sandy loam	0.09	0.22	0.34
Mandeni	1	0-60	Sand	0.02	0.08	0.11

θ_{wp}, θ_{fc} and θ_{sat} are soil moisture content at the wilting point, field capacity and saturation points, respectively.

Data analyses

The Mann–Kendall and Theil–Sen slope non-parametric statistical methods were used to detect the direction and extent of temporal trends in the long-term soil moisture data set.

These statistical methods have been successfully used in detecting trends and changes in long-term soil moisture time series.[14,23] The main advantages of the non-parametric statistical methods are that missing values are allowed and these tests do not make any assumptions about the distribution of the data.[37,38] Furthermore, these methods have low sensitivity to outliers and heterogeneous time series.[39] These statistical tests were run in XLSTAT software (https://www.xlstat.com/en/).

Mann–Kendall test

The Mann–Kendall test statistic S of Kendall[40] was used in this study to assess the monotonic trends in the soil moisture over time. The test statistic *S* is calculated based on Mann[41], Kendall[40] and Yue et al.[37] as:

$$S = \Sigma_{i=1}^{n-1} \Sigma_{j=i+1}^{n} sgn(x_j - x_i), \qquad \text{Equation 3}$$

where *n* is the number of data points, x_j and x_i are data values in time series at time *j* and *i* (*j* > *i*), respectively. Furthermore, $sgn(x_j - x_i)$ is the sign function given by:

$$sgn(x_j - x_i) = \begin{cases} -1 \; if \; (x_j - x_i) < 0 \\ 0 \; if \; (x_j - x_i) = 0 \\ +1 \; if \; (x_j - x_i) > 0 \end{cases} \qquad \text{Equation 4}$$

For a sample size *n*>10, a normal approximation to the Mann–Kendall test may be used.[40] The variance statistic is then computed as:

$$Var(S) = \frac{n(n-1)(2n+5)-\Sigma_{t=1}^{m} t_i(t_i-1)(2t_i+5)}{n} \qquad \text{Equation 5}$$

where *n* is the number of observations and t_i are the ties of the sample time series. The standard normal variable (Z_s) was used to identify the direction of the trend and its significance:

$$Z_s = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}, \text{ if } S>0 \\ 0, \; if \; S=0 \\ \frac{S+1}{\sqrt{Var(S)}}, \text{ if } S<0 \end{cases} \qquad \text{Equation 6}$$

where positive Z_s values indicate an increasing trend while negative values indicate a decreasing trend. The significance of the trends was tested at the significance levels of 95% and 99%.

The Theil–Sen slope estimator

The Theil–Sen slope estimator of Sen was used to give an indication of the magnitude of the linear trends in the soil moisture over time. According to Da Silva et al.[38], a linear model *f (t)* can be described as:

$$f(t) = Q_i + B, \qquad \text{Equation 7}$$

where Q_i is Sen's slope and *B* is the constant. To derive an estimate of Q_i, the slopes of all data pairs are calculated:

$$Q_i = \frac{x_j - x_k}{j - k}, \; i=1, 2...N, \qquad \text{Equation 8}$$

where X_j and X_k are data values at time *j* and *k* (*j*>*k*), respectively. The median of Sen's slope is calculated as:

$$Z_s = \begin{cases} \frac{S-1}{\sqrt{Var(S)}}, \text{ if } S>0 \\ 0, \; if \; S=0 \\ \frac{S+1}{\sqrt{Var(S)}}, \text{ if } S<0 \end{cases} \qquad \text{Equation 9}$$

The sign of Q_{med} reflects the data trend direction, whereas its value gives the magnitude of the slope of the trend. A positive Q_{med} value indicates an increasing trend while a negative value indicates a decreasing trend over time.

Results and discussion

Variability of the long-term soil moisture data set

The results of the statistical tests on the monthly averages of soil moisture for 39 years at all stations are presented in Table 3. The monthly mean soil moisture values ranged between 68.51 mm and 92.64 mm at Bainsvlei station in September and February, respectively. The monthly mean soil moisture values ranged between 100.26 mm and 114.63 mm at Bronkhorstspruit station in August and January, respectively. The monthly mean soil moisture values ranged between 33.68 mm and 37.10 mm at Mandeni station in January and October, respectively. Despite the highest annual rainfall received at Mandeni station, Bronkhorstspruit station had the highest soil moisture (108.55 mm) while Mandeni station had the lowest (35.13 mm). The highest soil moisture content at Bronkhorstspruit station could be attributed to higher water-holding capacity as a result of relatively high clay and organic carbon contents as also reported by Myeni[29]. The lowest soil moisture content at Mandeni station could be attributed to the low water-holding capacity of sandy soils, which dominated this site.[29] The results further showed the seasonal soil moisture pattern, with wet conditions in summer and dry conditions in winter months. The findings of our study agree with the findings of Pan et al.[18], who reported that soil moisture peaked in February and was minimal in July in the summer regions of South Africa.

Table 3: Basic statistics and Mann–Kendall trend analysis of soil moisture for 39 years (1980–2018) at all stations

Station	Month	Minimum (mm)	Maximum (mm)	Mean (mm)	Standard deviation	Mann–Kendall test	Sen's slope
Bainsvlei	January	59.95	124.61	85.27	17.39	0.136	0.291
	February	61.61	130.56	92.64	19.07	0.028	0.053
	March	62.45	131.56	91.32	14.62	-0.109	-0.223
	April	67.25	141.29	90.82	18.38	0.028	0.071
	May	61.64	116.31	79.64	14.49	0.128	0.200
	June	59.49	120.60	74.14	13.80	0.142	0.187
	July	59.07	88.70	69.15	8.77	0.042	0.043
	August	58.82	93.40	68.68	8.91	-0.023	-0.021
	September	58.72	122.03	68.51	12.30	-0.220	-0.157
	October	58.75	126.67	74.57	14.69	-0.117	-0.145
	November	61.18	119.12	82.21	14.73	-0.090	-0.217
	December	61.71	115.90	82.13	14.82	0.069	0.125
	Annual	66.34	94.25	79.84	6.21	-0.009	-0.004
Bronkhorstspruit	January	103.21	131.19	114.63	6.43	0.001	0.000
	February	102.94	143.06	113.88	9.34	-0.042	-0.049
	March	100.09	134.51	114.16	8.77	-0.163	-0.176
	April	98.69	140.43	109.52	8.61	0.055	0.053
	May	96.95	125.55	105.32	6.71	0.152	0.106
	June	96.76	125.07	102.79	5.44	0.009	0.002
	July	96.74	106.99	100.46	2.32	-0.112	-0.021
	August	96.74	108.52	100.26	2.22	-0.171	-0.021
	September	98.00	119.43	102.19	4.18	-0.260	-0.067
	October	100.09	124.33	109.44	5.71	0.015	0.013
	November	102.06	131.95	114.71	7.57	0.001	0.004
	December	106.61	133.29	115.52	6.37	0.015	0.016
	Annual	104.21	113.26	108.55	2.24	-0.015	-0.005
Mandeni	January	19.50	49.30	33.68	10.04	0.409*	0.553
	February	19.75	47.94	34.53	8.84	0.371*	0.444
	March	19.38	52.66	34.43	8.15	0.328*	0.362
	April	21.99	47.83	36.72	6.57	0.182	0.146
	May	20.61	49.21	34.99	7.56	-0.155	-0.146
	June	20.41	49.83	34.42	8.95	-0.444	-0.480
	July	20.26	50.93	35.41	10.05	-0.409	-0.547
	August	19.66	47.21	34.31	7.85	-0.341	-0.363
	September	24.06	45.84	35.15	5.72	-0.136	-0.106
	October	21.82	49.31	37.10	7.83	0.385*	0.420
	November	20.53	49.19	35.80	8.73	0.466**	0.526
	December	19.65	51.14	35.07	9.91	0.393*	0.585
	Annual	30.33	43.09	35.13	3.17	0.317*	0.129

*p<0.05, **p<0.001

The Mann–Kendall test and Sen's slope statistical tests were applied to the time series of soil moisture estimates from 1980 to 2018 at the three stations, and the trend analysis for all months and the whole year are also presented in Table 3. The results of the Mann–Kendall test at the Bainsvlei station show a marginal increasing trend of soil moisture in January, February, April, May, June, July and December, while the remaining months show a non-significant decreasing trend. For the Bronkhorstspruit station, the results show a marginal decreasing trend of soil moisture in February, March, July, August and September, while the remaining months show a marginal increasing trend. The results further show that soil moisture increased significantly from October to March, while the remaining months show a marginal decreasing trend at the Mandeni station. These findings suggest that wet seasons have become wetter while dry seasons have become drier at the eastern coastal regions in recent years.

In South Africa, an increase in air temperature and the variability of rainfall is expected as a result of predicted climate change.[10] Therefore, understanding the effects of air temperature and rainfall on soil moisture is critical in the determination of the impacts of climate variability on soil moisture status in this region. The regression graph of the mean annual air temperature and mean annual soil moisture indicate that air temperature explains about 2% of the variation in soil moisture at Bainsvlei and Bronkhorstspruit stations, but only 1% at Mandeni station (Figure 3). Results also indicate the negative linear relationship between air temperature and soil moisture as expected. The regression graph of the mean annual rainfall and mean annual soil moisture indicate that more than 70% of the variation in soil moisture can be explained by air temperatures across all stations (Figure 4). The results also indicated a positive and significant effect of rainfall on soil moisture status as expected.

To investigate the potential impacts of climate variability on soil moisture changes, long-term trends in soil moisture were compared with rainfall and air temperature trends (Figure 5). The mean annual soil moisture results indicate a marginal decrease in soil moisture from 1980 to 2018 at the Bainsvlei and Bronkhorstspruit stations, at a rate of -0.02 and -0.001 mm per annum, respectively. Furthermore, the trends indicate that Bainsvlei and Bronkhorstspruit stations are becoming warmer, with increases of 0.04 and 0.02 °C per annum, while mean annual rainfall shows decreasing trends at a rate of -0.97 and -1.05 mm per annum, respectively. An increase in temperatures at the Bainsvlei and Bronkhorstspruit stations could have enhanced the rate of ET(t) which removes moisture from the soil and decreases soil moisture content. However, T_{air} is not the only climatic factor controlling the rate of ET(t), because U and RH also play a critical role. Furthermore, the rate of ET(t) is also limited by the amount of soil moisture available in the soil, such that ET(t) will be limited if the soil moisture content is below the wilting point, even though T_{air} could be increasing.[42] Therefore, the relationship between T_{air} and soil moisture is not explicit, as also noted by Cheng et al.[14] These findings are in agreement with Wang et al.[43] who noted that the effect of temperature on soil moisture is relatively low as the result of low soil moisture available for evapotranspiration in semi-arid regions. The findings of this study suggest that Bainsvlei and Bronkhorstspruit stations are gradually becoming drier as a result of decreasing trends of rainfall with possibly a small influence of increasing air temperature.

In contrast, there was a significant increase in mean annual soil moisture at the Mandeni station, at a rate of 0.11 mm per annum. The increase in soil moisture at the Mandeni station could be attributed to the observed significant increasing trend of rainfall at a rate of 15.89 mm per annum. The findings of this study suggest a strong correlation between rainfall and soil moisture and agree with previous studies that have reported that soil moisture closely follows trends of rainfall, whether drying or wetting.[14,17,43] Furthermore, the findings suggest that Mandeni station is gradually becoming wetter as a result of the increasing trend of rainfall, even though air temperatures are also increasing.

Overall discussion

Long-term temporal variation in soil moisture revealed that 1983, 1992, 1998 and 2015 were the driest years while 1987 and 2000 were the wettest years. These findings confirm the extreme droughts and floods that were experienced in this region in these years.[11] The occurrence of floods in South Africa is often associated with tropical cyclones while the occurrence of droughts is often associated with the El Niño-Southern Oscillation phenomenon.[11] The findings of this study confirm the suitability of the model estimates to capture variation in soil moisture very well. Therefore, the model estimates could be reliably used to provide long-term soil moisture data sets for climatic research.

The findings of this study indicate that Bainsvlei and Bronkhorstspruit stations located inland are experiencing drier conditions while the Mandeni station located in the coastal region is experiencing wetter conditions, especially in the summer months. The findings are consistent with those of previous studies which predicted that eastern coastal parts of South Africa are expected to become wetter while the inland parts are expected to be drier as a result of predicted climate change.[11,12,44,45] The expected drying of inland parts is likely to pose water scarcity challenges while the wetting of eastern coastal parts is likely to induce erosion and flood risks. Furthermore, changes in soil moisture attributed to climate variability are likely to affect various sectors – such as agriculture and water supply – that are primarily dependent on soil moisture availability.

The findings of this study suggest that air temperatures have been increasing across South Africa, at an average of 0.36 °C per decade over the past recent 39 years. These findings are consistent with the observed increase in air temperatures at a rate of 0.4 °C per decade over the past 54 years (1961–2014) in the southern African region, as reported by Davis and Vincent[11]. Furthermore, these findings are also consistent with the observed increase in global average temperature at a rate of 0.6 °C per decade estimated by IPCC[10].

The findings of this study confirm that climate variability and change are likely to change soil moisture content in South Africa, as also noted by Cheng et al.[14] However, the findings also suggest that the influences of climate change on soil moisture will vary with region and climatic conditions. Therefore, understanding the factors that affect soil moisture variability at the local scale is critical for the development of informed adaptation strategies to support efficient management and sustainable use of natural resources.

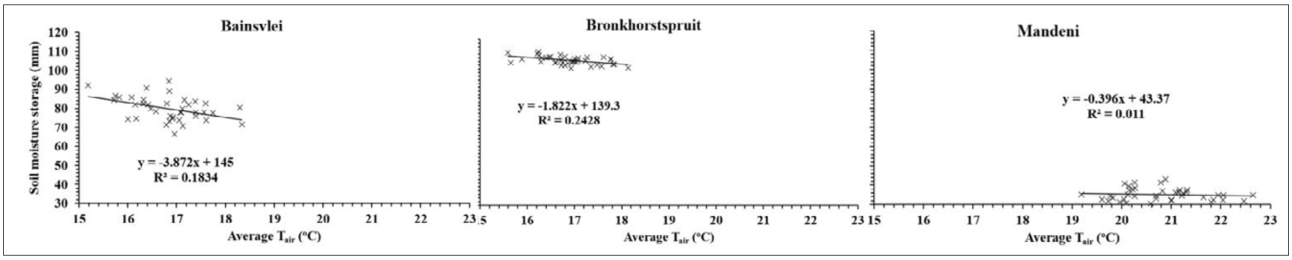


Figure 3: Regression plots of average annual air temperature (T_{air}) and average annual soil moisture over 39 years at three different locations.

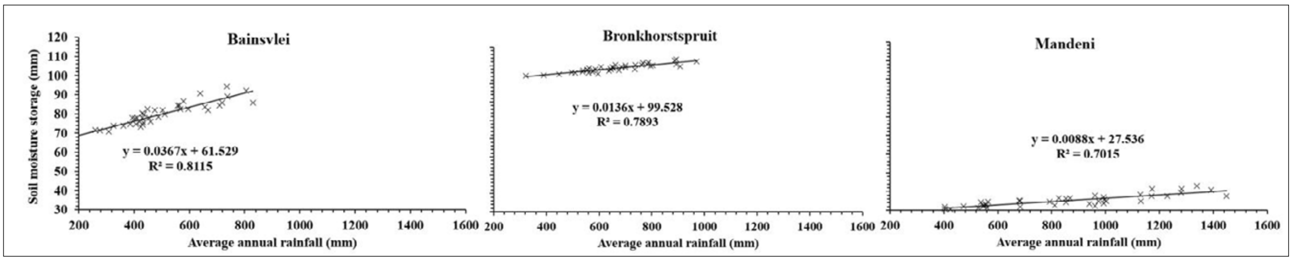


Figure 4: Regression plots of average annual rainfall and average annual soil moisture over 39 years at three different locations.

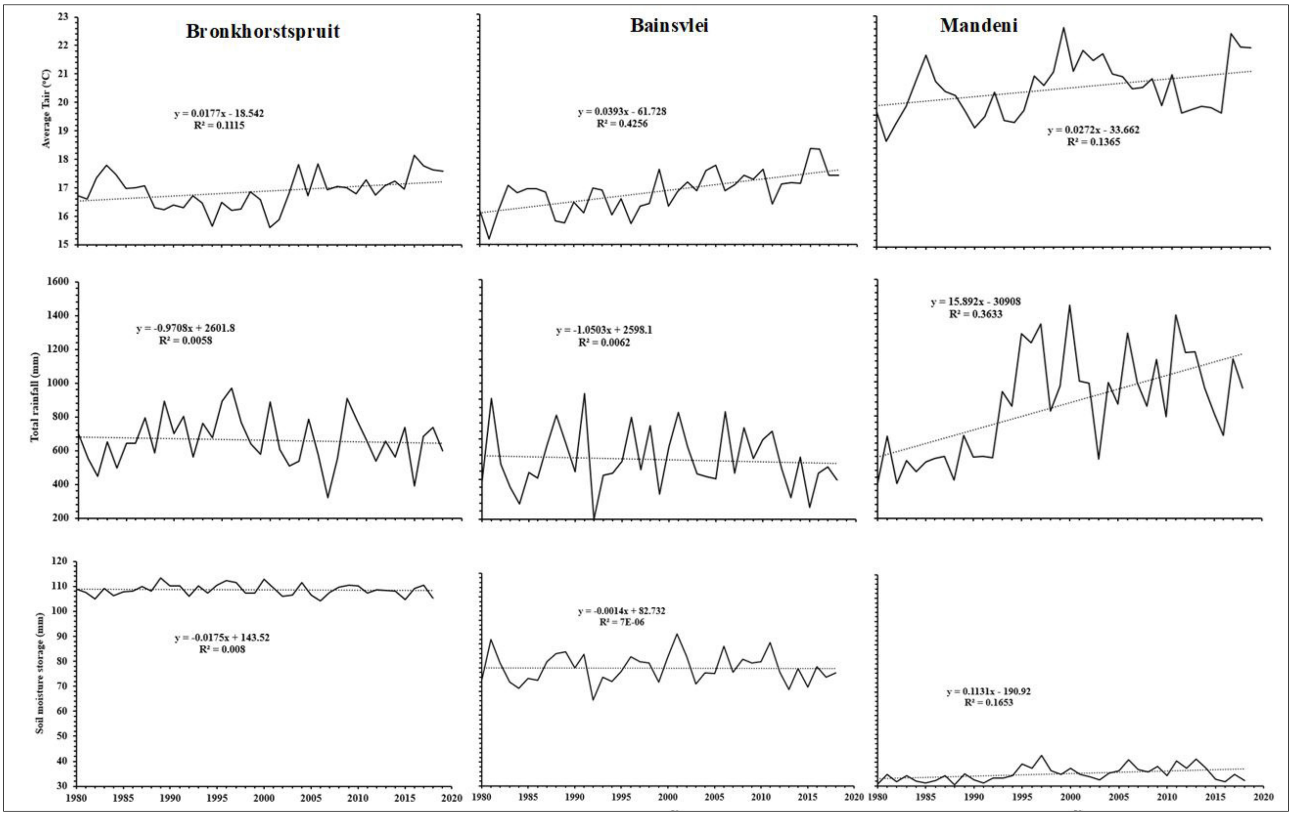


Figure 5: Temporal variations and linear trends of average air temperature (T_{air}), annual rainfall and soil moisture over the last 39 recent years at different stations.

Conclusions

Soil moisture is a critical parameter in the forecasting and assessment of weather-induced extreme events, which are likely to increase as a consequence of the expected climate change in this region. In this study, a water balance model was used to reconstruct long-term soil moisture data sets from 1980 to 2018 (39 years) in three stations that represent the different agroclimatic conditions of South Africa. Additionally, long-term changes and variability of moisture were examined to investigate the potential impacts of climate variability on soil moisture.

The results of the study show a marginal decreasing trend of annual soil moisture at the Bainsvlei and Bronkhorstspruit stations located inland. In contrast, the Mandeni station located in the coastal region is gradually becoming wetter as a result of the increasing trend of rainfall, despite the increase in air temperatures. These findings suggest that inland regions are becoming drier while coastal regions are becoming wetter, especially in the summer months in this country.

Our study confirms that increasing climate variability and climate change are likely to alter the soil moisture content status in this country, although their effects will vary with agroclimatic conditions. Therefore, there is a vital need for the understanding of factors that affect soil moisture variability at the local scale for the development of informed adaptation and mitigation strategies. Our study also demonstrates the suitability of the model estimates to provide comprehensive soil moisture data sets for weather and climate research studies, given that long-term and representative in-situ soil moisture measurements are often lacking in many countries, especially in developing countries.

Acknowledgements

Dr Thandile Mdlambuzi (South African Sugarcane Research Institute) is gratefully acknowledged for his technical support during fieldwork. We also thank Drs Garry Paterson and Thomas Fyfield (Agricultural Research Council) for proofreading and editing the manuscript.

Competing interests

We declare that there are no competing interests.

Authors' contributions

L.M.: Conceptualisation, methodology, data collection and analysis, writing – original. M.E.M. and A.D.C.: Methodology, review and editing, supervision.

References

1. Du C, Wu W, Liu X, Gao W. Simulation of soil moisture and its variability in East Asia. Remote Sens Model Ecosyst Sustain III. 2006;6298(1986):62982F. https://doi.org/10.1117/12.690643

2. Brocca L, Ciabatta L, Massari C, Camici S, Tarpanelli A. Soil moisture for hydrological applications: Open questions and new opportunities. Water (Switzerland). 2017;9(2):140–160. https://doi.org/10.3390/w9020140

3. El Masri B. Examining the spatial and temporal variability of soil moisture in Kentucky using remote sensing data. Biomed J Sci Tech Res. 2017;1(7):1–4. https://doi.org/10.26717/BJSTR.2017.01.000604

4. Fischer EM, Seneviratne SI, Vidale PL, Lüthi D, Schär C. Soil moisture-atmosphere interactions during the 2003 European summer heat wave. J Clim. 2007;20(20):5081–5099. https://doi.org/10.1175/JCLI4288.1

5. Seneviratne SI, Wilhelm M, Stanelle T, Van Den Hurk B, Hagemann S, Berg A, et al. Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. Geophys Res Lett. 2013;40(19):5212–5217. https://doi.org/10.1002/grl.50956

6. Huang J, Van Den Dool HM, Georgakakos KP. Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts. J Clim. 1996;9(6):1350–1362. https://doi.org/10.1175/1520-0442(1996)009<1350:AOMCSM>2.0.CO;2

7. Meng L, Quiring SM. A comparison of soil moisture models using soil climate analysis network observations. J Hydrometeorol. 2008;9(4):641–659. https://doi.org/10.1175/2008JHM916.1

8. Teuling AJ, Seneviratne SI, Stöckli R, Reichstein M, Moors E, Ciais P, et al. Contrasting response of European forest and grassland energy exchange to heatwaves. Nat Geosci. 2010;3(10):722–727. https://doi.org/10.1038/ngeo950

9. GCOS. The Global Observing System for Climate: Implementation needs. Geneva: World Meteorology Organisation; 2016. Available from: https://library.wmo.int/doc_num.php?explnum_id=3417

10. IPCC. Summary for policymakers. In: Field CB, Barros VR, Dokken DJ, Mach KJ, Mastrandrea MD, Bilir TE, et al., editors. Climate change 2014: Impacts, adaptation, and vulnerability. Part A: Global and sectoral aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK / New York: Cambridge University Press; 2014. p. 1–32. Available from: https://www.ipcc.ch/site/assets/uploads/2018/02/ar5_wgII_spm_en.pdf

11. Davis CL, Vincent K. Climate risk and vulnerability: A handbook for southern Africa. 2nd ed. Pretoria: CSIR; 2017.

12. Jury MR. South Africa's future climate: Trends and projections. In: Knight J, Rogerson CM, editors. The geography of South Africa. Cham: Springer; 2019. p. 305–312. https://doi.org/10.1007/978-3-319-94974-1_33

13. Brocca L, Zucco G, Moramarco T, Morbidelli R. Developing and testing a long-term soil moisture dataset at the catchment scale. J Hydrol. 2013;490:144–151. https://doi.org/10.1016/j.jhydrol.2013.03.029

14. Cheng S, Guan X, Huang J, Ji F, Guo R. Long-term trend and variability of soil moisture over East Asia. J Geophys Res Atmos. 2015;120(17):8658–8670. https://doi.org/10.1002/2015JD023206

15. 15. Coopersmith EJ, Bell JE, Cosh MH. Extending the soil moisture data record of the U.S. Climate Reference Network (USCRN) and Soil Climate Analysis Network (SCAN). Adv Water Resour. 2015;79:80–90. https://doi.org/10.1016/j.advwatres.2015.02.006

16. 16. Hao L, Sun G, Liu Y, Zhou G, Wan J, Zhang L, et al. Evapotranspiration and soil moisture dynamics in a temperate grassland ecosystem in Inner Mongolia, China. Trans ASABE. 2016;59(2):577–590. https://doi.org/10.13031/trans.59.11099

17. Stillman S, Ninneman J, Zeng X, Franz T, Scott RL, Shuttleworth WJ, et al. Summer soil moisture spatiotemporal variability in southeastern Arizona. J Hydrometeorol. 2014;15(4):1473–1485. https://doi.org/10.1175/JHM-D-13-0173.1

18. Pan N, Wang S, Liu Y, Zhao W, Fu B. Global surface soil moisture dynamics in 1979-2016 observed from ESA CCI SM dataset. Water. 2019;11(5):883. https://doi.org/10.3390/w11050883

19. Narasimhan B, Srinivasan R, Arnold JG, Di Luzio M. Estimation of long-term soil moisture using a distributed parameter hydrologic model and verification using remotely sensed data. Trans Am Soc Agric Eng. 2005;48(3):1101–1113. https://doi.org/10.13031/2013.18520

20. Yin Z, Ottlé C, Ciais P, Guimberteau M, Wang X, Zhu D, et al. Evaluation of ORCHIDEE-MICT-simulated soil moisture over China and impacts of different atmospheric forcing data. Hydrol Earth Syst Sci. 2018;22(10):5463–5484. https://doi.org/10.5194/hess-22-5463-2018

21. Nandintsetseg B, Shinoda M. Seasonal change of soil moisture in Mongolia: Its climatology and modelling. Int J Climatol. 2011;31(8):1143–1152. https://doi.org/10.1002/joc.2134

22. Meng X, Mao K, Meng F, Shen X, Xu T, Cao M. Long-term spatiotemporal variations in soil moisture in North East China based on 1-km resolution downscaled passive microwave soil moisture products. Sensors. 2019;19(16):3527. https://doi.org/10.3390/s19163527

23. Dorigo WA, Xaver A, Vreugdenhil M, Gruber A, Hegyiová A, Sanchis-Dufau AD, et al. Global automated quality control of in situ soil moisture data from the International Soil Moisture Network. Vadose Zo J. 2013;12(3):1–21. https://doi.org/10.2136/vzj2012.0097

24. Mittelbach H, Casini F, Lehner I, Teuling AJ, Seneviratne SI. Soil moisture monitoring for climate research: Evaluation of a low-cost sensor in the framework of the Swiss soil moisture experiment (SwissSMEX) campaign. J Geophys Res Atmos. 2011;116(5):1–11. https://doi.org/10.1029/2010JD014907

25. RoTimi Ojo E, Bullock PR, Fitzmaurice J. Field performance of five soil moisture instruments in heavy clay soils. Soil Sci Soc Am J. 2015;79(1):20. https://doi.org/10.2136/sssaj2014.06.0250

26. Dostálová A, Doubková M, Sabel D, Bauer-Marschallinger B, Wagner W. Seven years of advanced synthetic aperture radar (ASAR) global monitoring (GM) of surface soil moisture over Africa. Remote Sens. 2014;6(8):7683–7707. https://doi.org/10.3390/rs6087683

27. Oroza CA, Bales RC, Stacy EM, Zheng Z, Glaser SD. Long-term variability of soil moisture in the southern Sierra: Measurement and prediction. Vadose Zo J. 2018;17(1):1–9. https://doi.org/10.2136/vzj2017.10.0178

28. Malekian R, Gordon R, Madani A, Robertson S. Evaluation of the versatile soil moisture budget model for a humid region in Atlantic Canada. Can Water Resour J. 2014;39(1):73–82. https://doi.org/10.1080/07011784.2014.888891

29. Myeni L. Optimizing monitoring networks for accurate and continuous in situ soil moisture dataset across South Africa [PhD thesis]. Pietermaritzburg: University of KwaZulu-Natal; 2020.

30. Moeletsi ME, Shabalala ZP, De Nysschen G, Walker S. Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over the Free State Province, South Africa. Water SA. 2016;42(3):466–474. https://doi.org/10.4314/wsa.v42i3.12

31. Agricultural Research Council – Institute for Soil Climate and Water (ARC-ISCW). Agro-climatology database [database on the Internet]. c2019 [cited 2019 Nov 14]. Available from: http://www.arc.agric.za/arc-iscw/Pages/Climate-Monitoring-Services.aspx

32. Conradie DCU. South Africa's climatic zones: Today, tomorrow. Paper presented at: International Green Building Conference and Exhibition: Future Trends and Issues Impacting on the Built Environment; 2012 July 25–26; Johannesburg, South Africa. Available from: http://researchspace.csir.co.za/dspace/handle/10204/6064

33. Burn DH, Elnur MAH. Detection of hydrologic trends and variability. J Hydrol. 2002;255(1–4):107–122. https://doi.org/10.1016/S0022-1694(01)00514-5

34. Shabalala ZP, Moeletsi ME, Tongwane MI, Mazibuko SM. Evaluation of infilling methods for time series of daily temperature data: Case study of Limpopo Province, South Africa. Climate. 2019;7(7):86. https://doi.org/10.3390/cli7070086

35. Abraha MG, Savage MJ. Comparison of estimates of daily solar radiation from air temperature range for application in crop simulations. Agric For Meteorol. 2008;148(3):401–416. https://doi.org/10.1016/j.agrformet.2007.10.001

36. DeLiberty TL, Legates DR. Interannual and seasonal variability of modelled soil moisture in Oklahoma. Int J Climatol. 2003;23(9):1057–1086. https://doi.org/10.1002/joc.904

37. Yue S, Pilon P, Cavadias G. Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. J Hydrol. 2002;259(1–4):254–271. https://doi.org/10.1016/S0022-1694(01)00594-7

38. Da Silva RM, Santos CAG, Moreira M, Corte-Real J, Silva VCL, Medeiros IC. Rainfall and river flow trends using Mann–Kendall and Sen's slope estimator statistical tests in the Cobres River basin. Nat Hazards. 2015;77(2):1205–1221. https://doi.org/10.1007/s11069-015-1644-7

39. Asfaw A, Simane B, Hassen A, Bantider A. Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin. Weather Clim Extrem. 2018;19:29–41. https://doi.org/10.1016/j.wace.2017.12.002

40. Kendall MG. Rank correlation measures. 4th ed. London: Charles Griffin; 1975. p. 15–22.

41. Mann HB. Nonparametric tests against trend. Econom J Econom Soc. 1945:245–259. https://doi.org/10.2307/1907187

42. Allen RG, Pereira LS, Smith M, Raes D, Wright JL. FAO-56 dual crop coefficient method for estimating evaporation from soil and application extensions. J Irrig Drain Eng. 2005;131(1):2–13. https://doi.org/10.1061/(ASCE)0733-9437(2005)131:1(2)

43. Wang Y, Yang J, Chen Y, Fang G, Duan W, Li Y, et al. Quantifying the effects of climate and vegetation on soil moisture in an Arid Area, China. Water (Switzerland). 2019;11(4):1–16. https://doi.org/10.3390/w11040767

44. Lumsden TG. Evaluation of potential changes in hydrologically relevant statistics of rainfall in southern Africa under conditions of climate change. Water SA. 2009;35(5):649–656. https://doi.org/10.4314/wsa.v35i5.49190

45. MacKellar N, New M, Jack C. Observed and modelled trends in rainfall and temperature for South Africa: 1960–2010. S Afr J Sci. 2014;110(7/8), Art. #2013-0353. https://doi.org/10.1590/sajs.2014/20130353

AUTHORS:
Stefanie Schütte[1]
Roland E. Schulze[1]
Mary Scholes[2]

AFFILIATIONS:
[1]Centre for Water Resources Research, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[2]School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Stefanie Schütte

EMAIL:
Schuttes@ukzn.ac.za

DATES:
Received: 02 Apr. 2020
Revised: 30 Sep. 2020
Accepted: 04 Nov. 2020
Published: 28 May 2021

HOW TO CITE:
Schütte S, Schulze RE, Scholes M. Impacts of soil carbon on hydrological responses – a sensitivity study of scenarios across diverse climatic zones in South Africa. S Afr J Sci. 2021;117(5/6), Art. #8118. https://doi.org/10.17159/sajs.2021/8118

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☒ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Yali Woyessa

KEYWORDS:
organic matter, impact on hydrology, soil water holding capacity, hydrological modelling

FUNDING:
National Research Foundation (South Africa)

Impacts of soil carbon on hydrological responses – a sensitivity study of scenarios across diverse climatic zones in South Africa

Soil organic carbon (SOC) content and the water holding capacity of soils are two properties which link the carbon and hydrological cycles. Hydrological model inputs seldom include soil carbon as a parameter even though soil carbon content is known to influence soil water retention capacities. This study is a sensitivity analysis of changes in hydrological responses when the model inputs include different soil carbon percentages for the topsoil horizon. Sensitivities of hydrological responses such as transpiration, runoff volumes, the stormflow component of runoff and extreme runoff events to SOC content were quantified under various climatic conditions in South Africa. The soil water holding capacities at the drained upper limit (i.e. field capacity), permanent wilting point and saturation were calculated for the topsoil horizon, using SOC dependent pedo (soil)-transfer functions for different soil carbon scenarios and locations in South Africa. These variables, together with other pre-determined soil- and location-related inputs, as well as 50 years of daily climate, were then used as inputs in a process-based hydrological model. Overall, it was found that increased SOC content in the topsoil horizon leads to an increase in transpiration, a reduction in runoff, especially in its stormflow component, and a reduction of extreme runoff events. However, these changes are relatively small compared to the influence of climate, particularly of rainfall amount and distribution.

Significance:

- Organic carbon content of the soil and the water holding capacity of soils link the carbon and hydrological cycles.
- Management interventions that increase SOC lead to win-win situations because, in addition to climate change mitigation, plant water availability improves, and overall surface runoff ‘flashiness’ becomes more regulated.
- While rainfall amount and distribution over space and time remain the most critical determinants of hydrological responses, increased SOC in the topsoil horizon leads to increases in transpiration and thus plant growth, and to a reduction in runoff, especially in its stormflow component, and hence to a small reduction of severe flooding events.

Introduction

The amount of soil organic carbon (SOC), and more broadly soil organic matter, is directly linked to the chemical, physical and biological properties of soil. Soil texture, rather than SOC, is the main determinant of soil water holding capacity, i.e. the volume of water that can be held by the soil. However, SOC also plays a role, thereby linking the carbon and hydrological cycles.[1-3] Mechanisms of general soil water absorption and retention are explained in soil science textbooks.[4] SOC itself has a water retention effect through affecting soil structure and adsorption properties[5], as well as through soil aggregation and associated pore space distribution[6].

The SOC impact on soil water retention depends on the type of soil, on soil carbon content and on the amount of water in the soil at a given point in time.[5-11] A positive correlation of SOC with water retention and/or selected water potentials is extensively reported in the literature[8-13], with limited exceptions[13-15]. The importance of SOC in estimating water retention is affected by textural composition[5,16], with this effect being of higher importance in coarse-textured soils than in fine-textured soils[5,8,16]. A US database analysis linked an increase of 1% (of soil weight) in SOC content to a 2% to >5% increase in plant available water content.[17] Soil water is a key controller of metabolic processes in the soil and of plant growth and productivity.[18] Changes in soil water retention with SOC additions affect the timing and duration of plant water availability, and are especially valuable in low carbon soils.[8]

The quantification of the relationship between SOC content and soil water retention has been reported as part of selected pedo (soil)-transfer functions which are empirical relationships between parameters of soil characteristics and more readily obtainable data on soil properties.[19,20] Soil water retention is commonly measured at suctions of 33 kPa and 1500 kPa. It is assumed that these water holding volumes are indicators of that particular soil’s hydrological variables of drained upper limit (DUL) and permanent wilting point (WP), respectively, while plant available water is the water held between DUL and WP.[21] Rawls et al.[5] and Saxton and Rawls[16] developed equations for water retention at suctions of 33 kPa and 1500 kPa which included clay, sand and silt content, as well as SOC, based on US soil databases, with the equations of Rawls et al.[5] being the more robust.[22] Soil porosity equations were also developed by Rawls et al.[7]

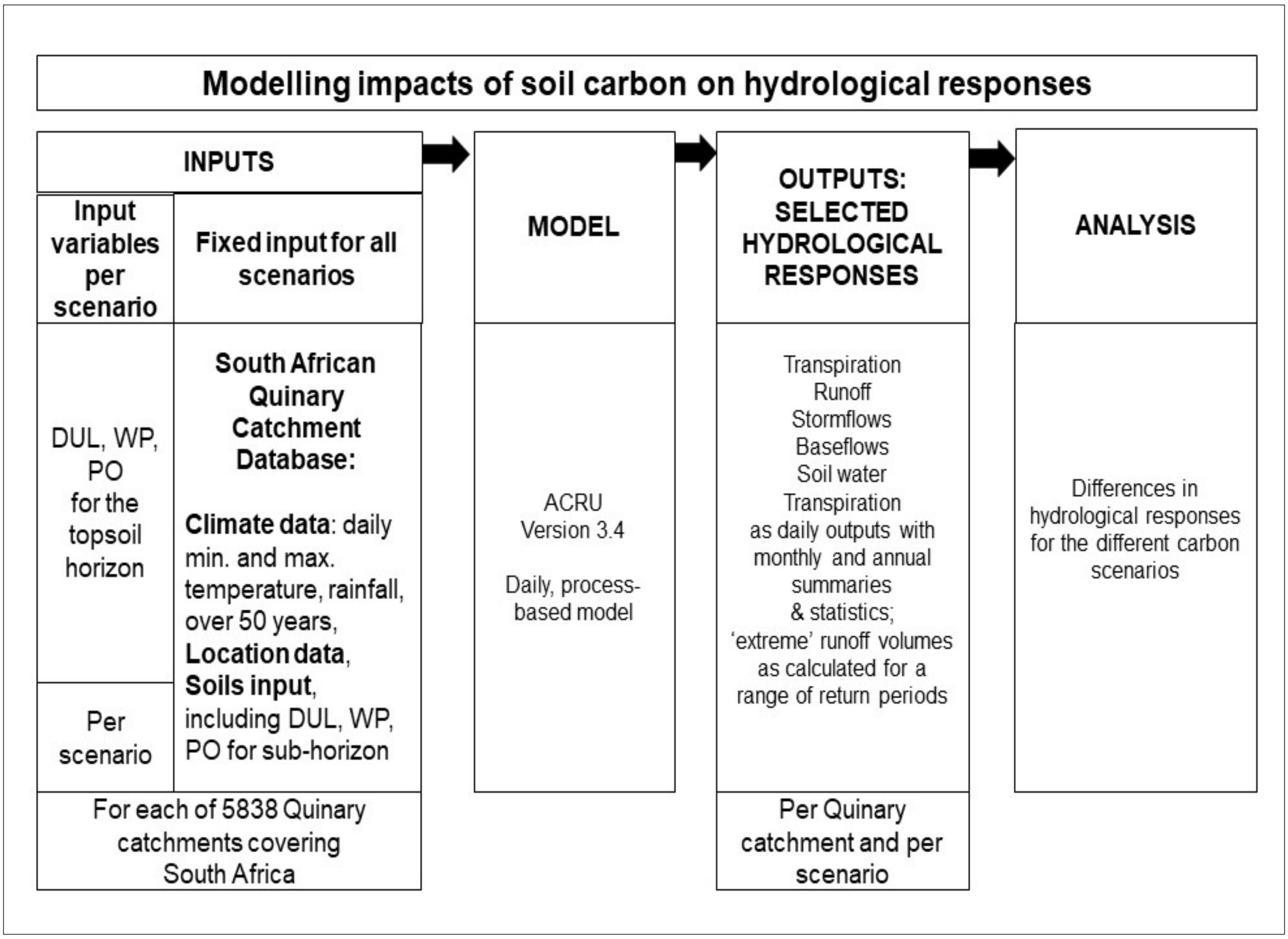
Soil organic carbon in South Africa is generally low and spatially highly variable.[23] With hydrological responses expected to change in South Africa in the next 50 years due to climate and land use change, there has to date been no baseline study to illustrate how soil carbon content impacts hydrological responses. This study focuses on the sensitivity of hydrological processes – e.g. stormflow and plant physiological responses, specifically transpiration rate – to varying amounts of soil carbon. The aim of this study was to use a process-based daily time-step hydrological model to explicitly include soil carbon contents across seven hydroclimatic zones using SOC-dependent pedo-transfer functions for different soil carbon scenarios and locations in South Africa. The results of this study would then indicate the SOC threshold at which hydrological processes are impacted upon and where these soils are located in South Africa.

Methodology

Six scenarios with varying carbon content were defined. One scenario was based on the actual SOC contents in the topsoil horizon, as derived from the soil carbon database[24,25], with an average SOC of 1.2% and a range between 0 and 12%. To be able to calculate sensitivities to changes in SOC content, hypothetical doubling and halving of the actual SOC amount were undertaken, with the three carbon scenarios being termed 'C_{actual}', 'C_{half}' and 'C_{double}'. The half scenario was included because land-use change mostly results in a reduction of SOC; however, conservation agriculture, irrigation and afforestation could result in an increase – hence the double scenario. In addition to the above scenarios, unrelated to actual SOC contents, assumptions of hypothetical SOC contents of 1%, 2% and 4% were made, with these three carbon scenarios being termed 'C_1' 'C_2' and 'C_4'. This approach was used to exclude the impact of the spatial variability of actual SOC in order to determine the sensitivity of the modelled hydrological responses. While a change of SOC from 1% to 4% is perhaps unrealistic, this was chosen to show more extreme changes. While hydrological modelling includes the soil profile properties, and therefore properties of the top- and subsoil horizons, we focused only on changes of SOC in the topsoil horizon, where substantial changes are more likely.

The soil-dependent hydrological soil water variables of DUL, WP and porosity (PO) for the topsoil horizon were calculated for each carbon scenario using the pedo-transfer functions by Rawls et al.[5], but corrected as per Nemes et al.[22] to read SOC rather than soil organic matter. First, soil textural contents of clay and sand, as well as SOC, were obtained from the Soil Profile Database.[26] The conversion from point values to area values has been explained in detail in Schütte et al.[24]

A schematic on the more detailed methodology of modelling impacts of soil carbon on hydrological responses is shown in Figure 1. To model hydrological responses in southern Africa, the Quinary Catchments Database[27] is frequently used. In the Quinary Catchments Database, South Africa, Lesotho and Eswatini (formerly known as Swaziland) were delineated into 5838 hydrologically relatively homogeneous response units, the so-called Quinary catchments, which are hydrologically interlinked with each linked to a 50-year data set of daily climate as well as location (e.g. altitude and slope), natural vegetation and soil properties. This existing database was used in this study to model hydrological responses, but with the DUL, WP and PO values of the topsoil horizon in the Quinary Catchments Database replaced with the newly calculated values. By using this approach to model the various scenarios, per-scenario results of hydrological responses can be obtained on a Quinary catchment spatial resolution across southern Africa, with the responses including transpiration, runoff, and its components of stormflow and baseflow, all for a statistically median year, for the 1:10 dry year, the 1:10 wet year, as well as for design 1-day, 2-day and 3-day runoff events calculated for a range of return periods by volumes.


Modelling impacts of soil carbon on hydrological responses
INPUTS
Input variables per scenario
Fixed input for all scenarios
DUL, WP, PO for the topsoil horizon
Per scenario
South African Quinary Catchment Database:
Climate data: daily min. and max. temperature, rainfall, over 50 years,
Location data,
Soils input, including DUL, WP, PO for sub-horizon
For each of 5838 Quinary catchments covering South Africa
MODEL
ACRU Version 3.4
Daily, process-based model
OUTPUTS: SELECTED HYDROLOGICAL RESPONSES
Transpiration
Runoff
Stormflows
Baseflows
Soil water
Transpiration
as daily outputs with monthly and annual summaries
& statistics;
'extreme' runoff volumes as calculated for a range of return periods
Per Quinary catchment and per scenario
ANALYSIS
Differences in hydrological responses for the different carbon scenarios

DUL, drained upper limit; WP, wilting point; PO, porosity

Figure 1: Schematic describing modelling impacts of soil carbon on hydrological responses.

Table 1: Selected stations, their locations, their representative Quinary catchment and characteristics, monthly means of daily maximum temperature (°C), monthly rainfall (mm) and of A-Pan equivalent evaporation totals (mm) for the period 1950–1999 for the seven hydroclimatic zones, after Hughes[29]

Station Quinary name Quinary number	Latitude Longitude Elevation (masl)	Acocks[32] Vegetation type & Dominant soil texture	Monthly mean of climatic variable (°C or mm)	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Ann
Mount Edgecombe U20M3 Quinary 4707	29˚42'S; 31˚02'E 82.9 m	Coastal Forest and Thornveld (#01) Loam	Daily maximum temperature	24	25	26	27	27	27	25	24	23	22	23	23	25
			Rainfall	96	96	118	124	117	102	61	38	17	22	38	61	888
			A-Pan evaporation	92	99	127	111	97	100	81	73	63	66	76	82	1068
Mara A71D3 Quinary 327	23˚09'S; 29˚33'E 918.8 m	Arid Sweet Bushveld (#14) Loamy Sand	Daily maximum temperature	28	28	29	29	28	27	26	24	22	22	24	26	26
			Rainfall	25	57	78	76	55	34	24	8	4	1	3	9	375
			A-Pan evaporation	138	142	151	145	127	130	109	97	84	85	103	122	1433
Upington D73E3 Quinary 2025	28˚27'S; 21˚25'E 851.6 m	Orange River Brokenveld (#32) Loamy Sand	Daily maximum temperature	29	32	35	35	35	32	28	24	21	21	23	27	29
			Rainfall	12	18	21	28	34	41	26	12	4	3	4	3	204
			A-Pan evaporation	165	189	216	213	177	162	116	91	73	79	101	135	1716
Elsenburg G22G3 Quinary 2700	33˚51'S; 18˚50'E 181.4 m	Coastal Rhenoster-bosveld (#46) Loam	Daily maximum temperature	22	25	27	28	29	27	24	20	18	17	17	19	23
			Rainfall	49	39	25	17	21	29	81	113	133	116	105	66	796
			A-Pan evaporation	117	145	168	169	142	124	85	60	46	47	61	82	1246
Outeniqua K30B1 Quinary 3307	33˚55'S; 22˚28'E 965.5 m	Knysna Forest (#04) Loam	Daily maximum temperature	18	19	20	21	21	21	20	18	17	16	16	16	19
			Rainfall	109	94	86	91	91	101	80	66	51	50	84	82	985
			A-Pan evaporation	82	93	103	97	79	80	62	50	43	45	53	64	850
Cedara U20E1 Quinary 4686	29˚31'S; 30˚17'E 1101.5 m	Natal Mist Belt Ngongoni Veld (#45) Loam	Daily maximum temperature	22	23	25	25	25	25	23	21	19	19	20	22	22
			Rainfall	84	107	131	136	109	101	48	25	12	14	30	45	842
			A-Pan evaporation	109	117	148	130	112	111	87	72	61	66	82	97	1189
Roodeplaat A21A3 Quinary 12	25˚55'S; 28˚21'E 1541.7 m	Bankenveld (#61) Loam	Daily maximum temperature	26	26	26	27	27	25	23	21	18	18	21	24	24
			Rainfall	74	111	111	126	82	86	45	15	5	4	5	26	689
			A-Pan evaporation	135	140	147	150	126	123	95	80	65	71	91	116	1338

The analysis of hydrological responses shown here is focused on seven strategic locations within South Africa, each represented by its respective Quinary catchment. These seven selected locations are considered to be representative of different climatic regimes and natural vegetation zones in South Africa and have been used as sample locations in previous studies.[28,29] The selected locations, together with the natural vegetation types in these zones, are shown in Figure 2. Table 1 shows the selected locations' identifiers, elevations, Quinary catchment names and numbers, their dominant natural vegetation and soil types, as well as mean monthly rainfall and potential evaporation, and monthly means of daily maximum temperature for the 50 years of observed and/or infilled data[30] (1950–1999), with the different climatic zones, according to the frequently used international Köppen classification[31] provided in the text.

Roodeplaat is in Köppen Climate Zone Cwb (winters long, dry and cool), with a mean annual precipitation (MAP) of 689 mm, mainly in the summer months (October to March). Mara is in Köppen Climate Zone BSh (semi-arid, hot and dry), with a low MAP of 375 mm. Upington has a very low MAP of 204 mm (Köppen Climate Zone BWh, arid, hot and dry). Elsenburg is in the winter rainfall region, with a MAP of 796 mm, and is in Köppen Climate Zone Csb (summers long, dry and cool). Outeniqua is in Köppen Climate Zone Cfb (wet all seasons, summers long and cool) and experiences rainfall throughout the year, with slightly lower rainfall in the cool winter months, with a MAP of 985 mm. Cedara is in Köppen Climate Zone Cwb (winters long, dry and cool), with a MAP of 842 mm, mainly in the summer months. Mount Edgecombe has a MAP of 1068 mm and is in Köppen Climate Zone Cfa (wet all seasons, summers long and hot, but wetter in summer than in winter). The locations' MAPs show a wide range from 204 mm to 1068 mm, while the annual mean temperature ranges from 19 °C to 29 °C. There is also a large elevation range, from 83 m to 1542 m.

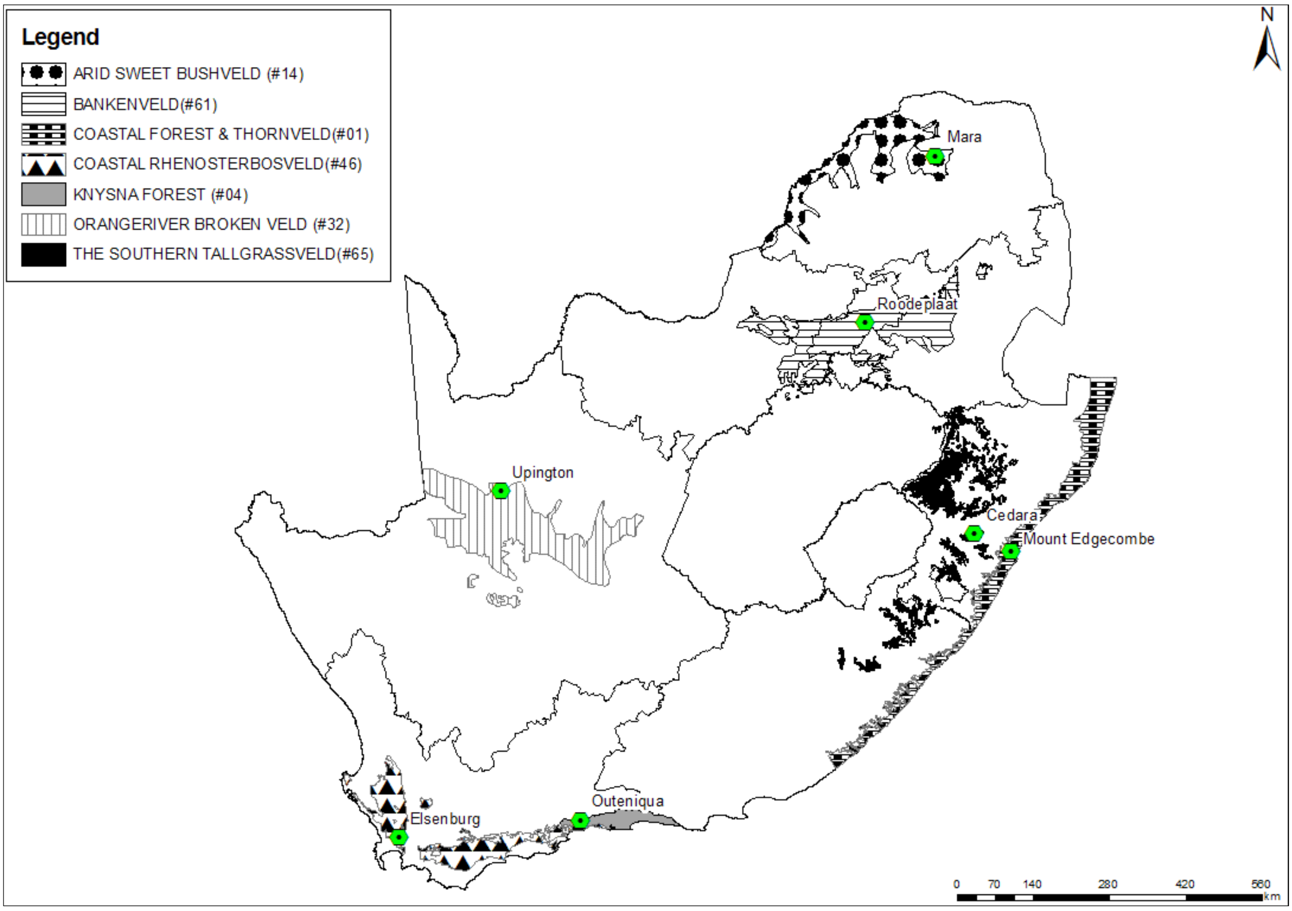

Legend
ARID SWEET BUSHVELD (#14)
BANKENVELD(#61)
COASTAL FOREST & THORNVELD(#01)
COASTAL RHENOSTERBOSVELD(#46)
KNYSNA FOREST (#04)
ORANGERIVER BROKEN VELD (#32)
THE SOUTHERN TALLGRASSVELD(#65)
N
Mara
Roodeplaat
Upington
Cedara
Mount Edgecombe
Outeniqua
Elsenburg
0 70 140 280 420 560
km

Figure 2: Locations of the seven hydroclimatic zones selected after Schulze[28]. The vegetation types represented by these zones are labelled according to Acocks[32]. See also Table 1.

The widely verified process-based daily time-step ACRU Model[28] was used first to simulate and explore the baseline hydrological characteristics of the seven hydroclimatic zones assuming naturally occurring vegetation types according to Acocks[32]. These simulations included volumes and monthly distributions of baseflow and stormflow. The model was then used to simulate the impacts of the various SOC scenarios. The model takes into account the atmosphere–soil profile–plant–water continuum of the landscape. Daily precipitation that reaches the soil surface after interception by vegetation either infiltrates and moves from topsoil horizon to subsoil horizon and possibly groundwater, or the water runs off as stormflow or (slow, delayed) baseflow to discharge into rivers.[28]

Results

The calculated ACRU input variables changed as a result of SOC changes; for example for Quinary catchment No. 4686, which represents Cedara for the C_1 and C_4 scenarios: the topsoil horizon DUL increased from 0.301 m/m (C_1) to 0.335 m/m (C_4), WP increased from 0.179 m/m (C_1) to 0.181 m/m (C_4) and PO increased from 0.454 m/m to 0.496 m/m.

Figure 3 shows the runoff results for a daily time slice of 5 months for one selected Quinary catchment (No. 4686) representing Cedara for the C_1 and C_4 scenarios. The runoff events are highly dependent on the magnitude and timing of the rainfall events. The C_4 scenario provided evidence that the higher SOC percentages reduced daily peaks compared to the C_1 scenario.

Impacts of the 1%, 2% and 4% SOC scenarios for the same Cedara catchment, for a period of 1 year (Figure 4), show accumulated annual transpiration of 345 mm (C_1 and C_2) and 352 mm (C_3), thus showing an increase in transpiration of 6 mm (2%) from the C_1 to the C_4 scenario. Runoff decreased by 16 mm (equivalent to 13%) from 125 mm (C_1) to 120 mm (C_2) to 109 mm (C_4). The stormflow component of runoff was reduced by 11 mm (11%, for C_1 to C_4) from 106 mm (C_1) to105 mm (C_2) to 95 mm (C_4), and the baseflow was reduced by 5 mm (or 26%) from 19 mm (C_1) to 15 mm (C_2) to 14 mm (C_4).

Changes in median annual transpiration for the 50 years of modelled daily values at the various locations are shown in Figure 5, representing plant water usage, with large differences, as expected, among the locations, being the lowest in arid Upington (Köppen Zone BWh) and the highest in moist Mount Edgecombe (Köppen Zone Cfa). With an increase in SOC, transpiration hardly changed for Roodeplaat, Mara, Cedara and Upington. However, at Elsenburg, in the winter rainfall zone and with a more temperate climate (Köppen Zone Csb), transpiration increased by 12 mm, equivalent to 9%, for a change in SOC from 1% to 4%, and increased by 6 mm, equivalent to 4%, for a change in SOC from 1% to 2%. Mount Edgecombe (in the Cfa climate zone, wet all seasons, summers long and hot) shows a transpiration increase of 14 mm, equivalent to 3%, for a change in SOC from 1% to 4%, but hardly any change (6 mm or 1%) when changing from 1% to 2% SOC. Generally, however, these locations show an increase in transpiration with increased SOC.

The runoff figures (not shown) in a 1:10 dry year vary from no runoff for all carbon scenarios for arid Upington, to a runoff of 56 mm in the C_1 scenario, 53 mm in the C_2 scenario, 43 mm in the C_4 scenario, and 60 mm, 58 mm and 54 mm, respectively, for the C_{half}, C_{actual} and C_{double} scenarios for Elsenburg. In a year of median responses, the runoff ranges between 1 mm for Upington in the C_4 scenario to 196 mm, 193 mm and 187 mm for Elsenburg (winter rainfall zone, Csb) for, respectively, the C_{half}, C_{actual} and C_{double} scenarios. For a 1:10 wet year, runoff ranges between 21 mm for Upington (dry, BWh) for the C_4 scenario, to 468 mm, 463 mm and 452 mm at Mount Edgecombe (wet, Cfa) for the C_{half}, C_{actual} and C_{double} scenarios (not shown).

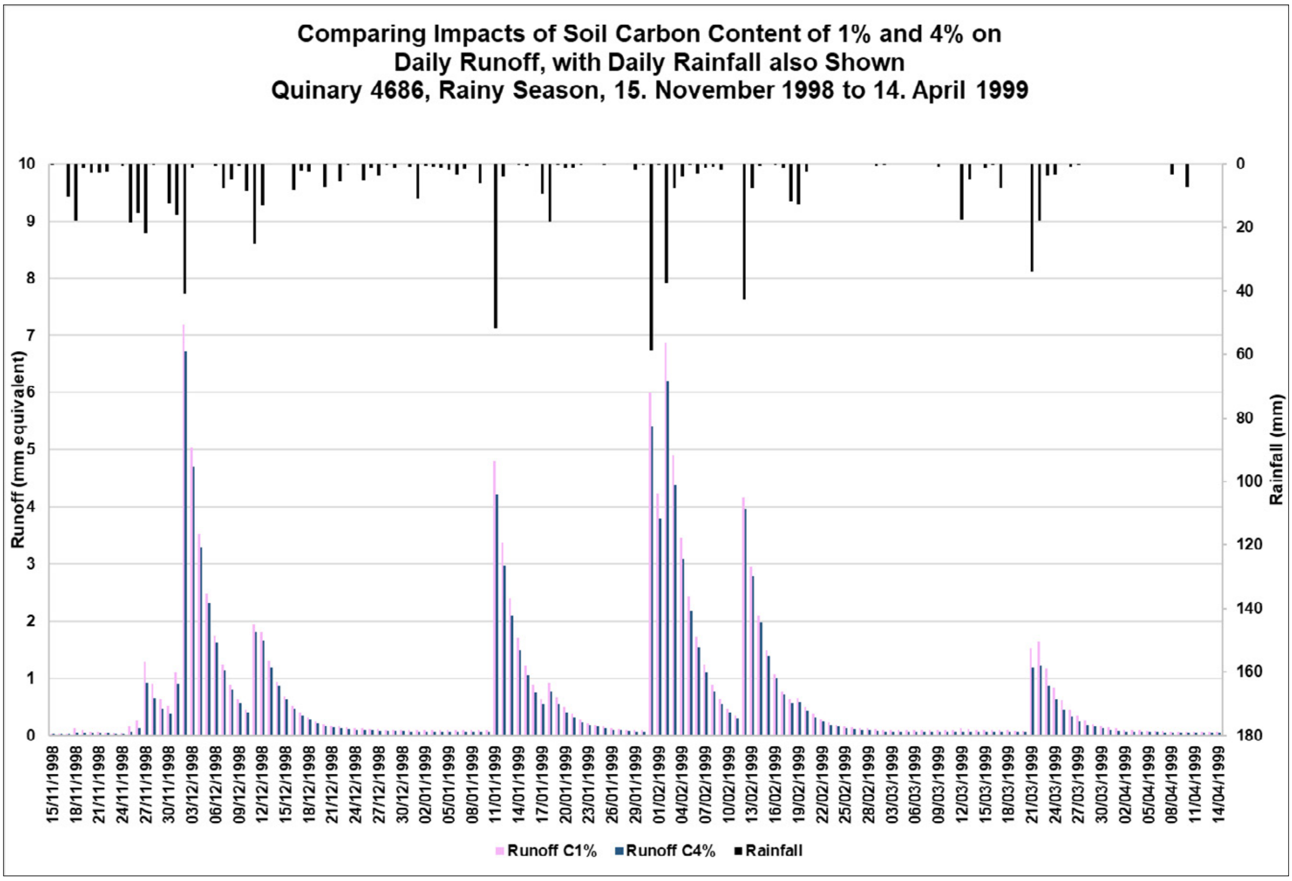


Figure 3: Simulated daily runoff for soil organic carbon of 1% (light pink) and 4% (dark blue), for the Quinary catchment representing Cedara during a 3.5-month period during the rainy season, with daily rainfall shown on the secondary axis.

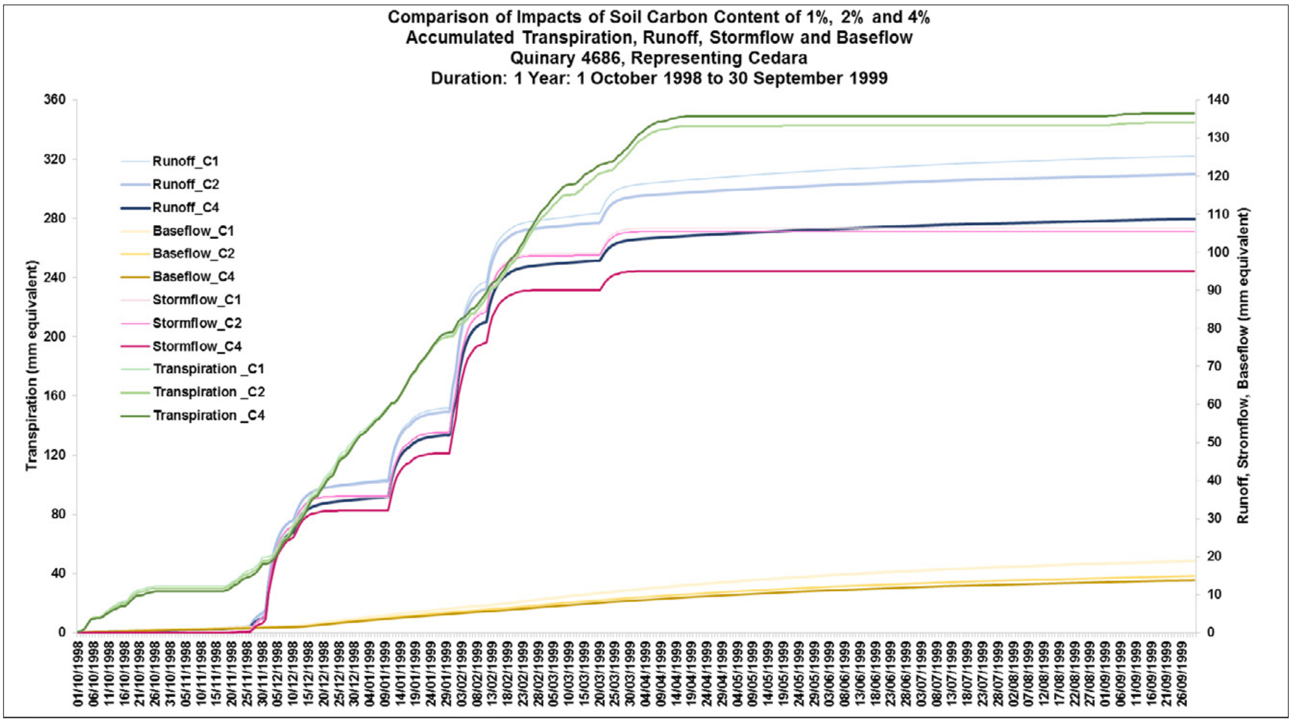


Figure 4: Daily accumulated transpiration (left *y*-axis), as well as accumulated runoff, stormflow and baseflow (right *y*-axis, all in mm) for a one-year period for the C_1, C_2 and C_4 scenarios, using Quinary Catchment 4686 representing Köppen Climate Zone Cwb at Cedara.

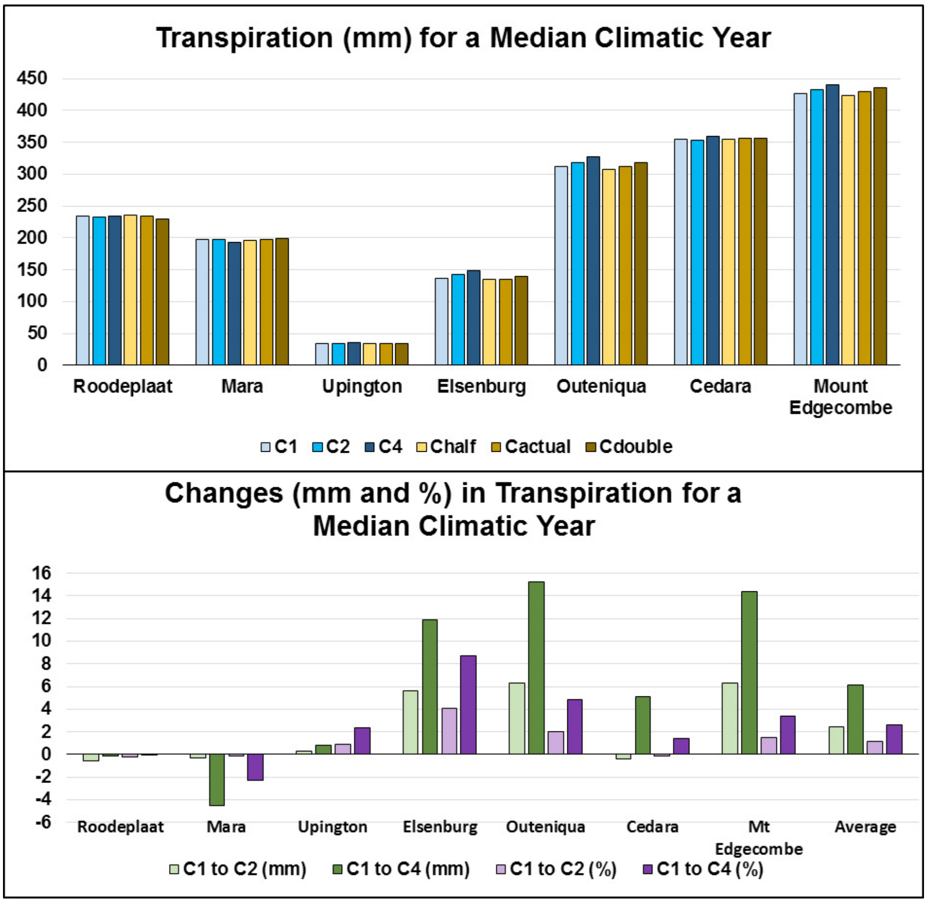


Figure 5: Accumulated transpiration (mm, top) for various soil carbon scenarios at selected locations in South Africa, as well as changes in transpiration from the C_1 to C_2 and from the C_1 to C_4 scenario (% and mm, bottom graph), for a median climatic year.

Selected changes in runoff results (Figure 6) show the impact of SOC to vary. While no impact is seen in Upington in a 1:10 dry year, because there is no runoff anyway, substantial sensitivities to SOC are seen for the other wetter areas. The largest absolute reduction of 24 mm is at Cedara in a 1:10 wet year when changing SOC from 1% to 4% SOC, with the largest relative reduction (but only a small absolute reduction) in runoff for Upington at 44% in a median year with a change in SOC from 1% to 4%.

Stormflows are rapid surface or near surface flows and are generally the major component of total runoff in most parts of South Africa. The highest results are from Mount Edgecombe (wet, Cfa) where stormflows are modelled at 330 mm, 328 mm and 310 mm for the C_1, C_2 and C_4 scenarios (not shown). Changes (mm and %) in stormflows in a 1:10 dry year, a median year and a 1:10 wet year for changing scenarios from the C_1 to the C_4 scenario (Figure 6) show the biggest absolute change, for a 1:10 wet year at Cedara (wet, Cwb) with a 24 mm reduction for a change from 1% to 4% SOC, while the biggest relative change is for Upington (dry, BWh) with a 27% reduction. In summary, an increase in SOC can lead to significant reductions in stormflows, but this depends on the inherent climate of an area and whether it is a dry, median or wet year.

Baseflows are the slow-release component of runoff and are the only water source in rivers in the non-rainy season while being a major component of runoff in the winter rainfall region. Most important in this sensitivity study are baseflows in the 1:10 dry year, with no baseflows for any of the SOC scenarios generated at Roodeplaat, Mara and Upington and very little at Cedara. The highest annual baseflows are found at Elsenburg (winter rainfall, temperate climate) with respectively 58, 56 and 52 mm for the C_{half}, C_{actual} and C_{double} SOC scenarios (not shown). In the cases where baseflow occurred, generally, a small reduction in baseflows was evident, although in relative terms this could be high, with up to 99% for Upington for a change of SOC from 1% to 4% in a 1:10 wet year (not shown).

Changes in more extreme runoff design events for 1-day and 3-day accumulated magnitudes for design return periods of 2-, 5-, 10- and 50-year return periods are shown in Figure 7. While the Quinary catchments at Elsenburg (winter rainfall) and Outeniqua (all year rainfall) show no significant changes, the highest absolute reduction was at Mount Edgecombe (wet, Cfa), from 2.9 mm equivalent runoff for a 3-day event for the 2-year return period to 4.4 mm for a 3-day event for the 50-year return period. Relative reductions were highest for Upington (dry, BWh), up to 20% for a 2-day runoff event for a 50-year return period (not shown), with a reduction of 2.2% for 3-day and 2-day runoff events for a 50-year return period.

Overall, it was found that increased SOC content in the topsoil horizon leads to an increase in transpiration, a reduction in runoff, especially in the stormflow component, and to a reduction of extreme runoff events.

Discussion and conclusions

While most soil profile and location specific properties cannot be changed, land use management can influence the amount of carbon in the soil, especially in the topsoil horizon. The sensitivities of hydrological responses such as transpiration, total runoff, the stormflow component, and extreme events to changes in SOC content at a number of diverse locations within South Africa were quantified using a hydrological process modelling approach. Relevant hydrological soil variables of soil water content at DUL (or field capacity), WP and PO, i.e. at saturation, were calculated for the topsoil horizon, using pedo-transfer functions which include various amounts of carbon representing different soil carbon scenarios. These soil carbon scenarios were then used as inputs to a process-based hydrological model at Quinary catchment resolution, with other soil profile, location and natural vegetation properties remaining as per the standard South African Quinary Catchments Database. Differences in hydrological responses between the scenarios were assessed for a number of climatically diverse areas within South Africa ranging from desert to sub-tropical climates.

Soil water holding capacities impacted by SOC were found to be a link between the carbon and the hydrological cycles as reported in the literature.[1-3] For the location studies, SOC was shown to impact hydrological responses, but the magnitude of these changes is strongly influenced by rainfall regimes and varies between the different climatic zones, location and soil properties. In assessing runoff on a daily basis for Cedara, for example, an increase in SOC led to a reduction in the conversion of rainfall to runoff, with the peak runoff magnitudes generally being reduced. Changes in runoff range between insignificant in very dry areas, to up to 24 mm of absolute reduction for Cedara in a 1:10 wet year, when modelling a change from 1% to 4% SOC, with the largest relative reduction (but only a small absolute reduction) in runoff for Upington at 44% for a median year when SOC is changed from 1% to 4%. The significant reductions in runoff results are mainly from stormflows, but with also more muted reductions in baseflows. An increase in SOC leads to transpiration increases, as was expected and found by others.[8] With an increase in SOC, shifts from runoff, and especially from the stormflow component, towards transpiration are seen. With increased SOC, the soil holds water more *in situ* in the landscape, with this water being available for plant transpiration and growth, which in turn leads to a reduction in runoff. On the other hand, when there is very little rain, as is in the case of Upington in a 1:10 dry year, then there is no runoff for any of the soil carbon scenarios.

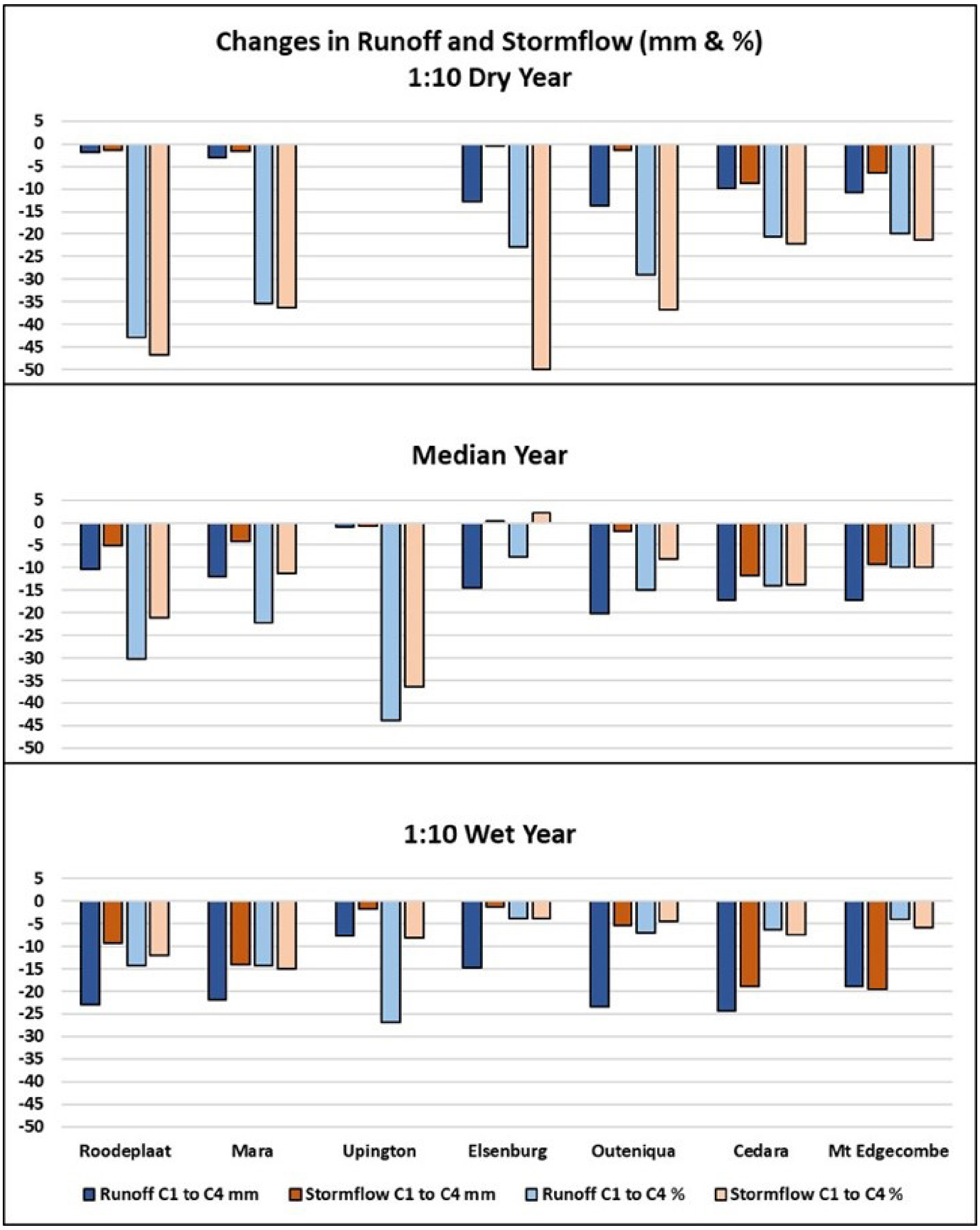


Figure 6: Changes in runoff and stormflow (mm and %) for carbon scenarios of 1% to 4% in a 1:10 dry year, a median year and a 1:10 wet year, for selected locations in South Africa.

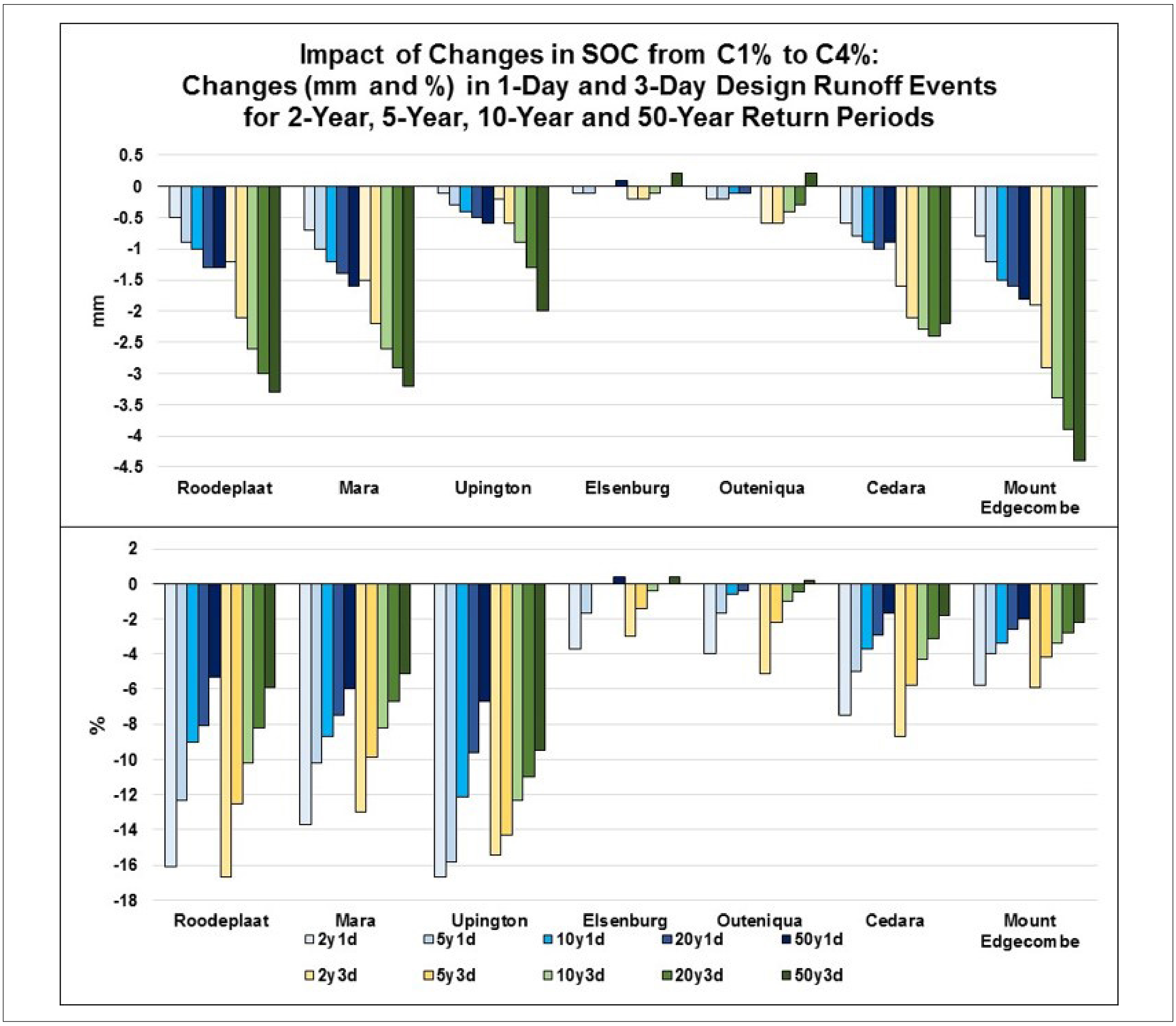


Figure 7: Changes in 1-day and 3-day design runoff events for return periods of 2, 5, 10, 20 and 50 years, with absolute changes (mm, top graph) and percentage changes (bottom graph) shown for selected areas in South Africa.

Not all areas show a change in extreme runoff with SOC changes, but most show a slight reduction in extreme runoff events with an increase in SOC content. When expressed as relative changes, this reduction is higher in smaller floods with shorter return periods compared to changes in larger floods with higher return periods. However, when expressed as absolute changes, the reductions are higher for larger floods with higher return periods compared to smaller floods with shorter return periods. Overall, an increase in soil carbon is shown to reduce extreme runoff events in most areas, but with different magnitudes. Increases in soil carbon should thus help to reduce some flood damage, thereby providing an important ecosystem service.

For the first time in South Africa, sensitivities of hydrological responses to SOC content changes have been calculated for selected locations with widely differing climatic regimes, with the results of this study confirming those in the literature.[12] In this study, a quantification of the overall reduction in runoff, and especially in stormflows, has been presented. Land management practices that increase carbon content would retain more water in the soil profile which would be available for plant use, and would thus usually lead to reduced runoff and flood events, but the impacts are limited and, again, depend on climatic, soil and location factors. Increased SOC, with increased plant water availability, is an additional benefit to climate change mitigation and thus presents a win-win situation.

More research is recommended to update the South African hydrological soil property databases, incorporating the new DUL, WP and plant available water values. While we examined only changes in SOC content in the topsoil, this study could be expanded to the subsoil horizon as well. The methodology developed in this study could also be used for sensitivity studies elsewhere in South Africa. Bearing in mind uncertainties regarding input values of carbon content, climate and soil variables, as well as pedo-transfer functions established elsewhere in the world, further improvements to impact modelling can be made if locally derived equations of WP, DUL and PO, which include a soil carbon factor, and improved model inputs, become available. Further research is also recommended to study the impact of actual changes in SOC on hydrological responses in South Africa over a historical period, as well as on SOC impacts on plant growth in the form of changes to soil water and plant stress-free days, for agricultural crop yield and primary production assessments.

Acknowledgements

We acknowledge the National Research Foundation (South Africa) for financial support of S.S.'s PhD studies through the Southern African Systems Analysis Programme. And the Agricultural Research Council for allowing the use of their databases. We also thank the two anonymous reviewers whose comments helped to improve this paper.

Competing interests

We declare that there are no competing interests.

Authors' contributions

S.S. devised the methodology, did the modelling and the analyses and wrote the paper. R.E.S. and M.S. provided technical input and provided editing support in their roles as PhD supervisor and co-supervisor, respectively.

References

1. Falkowski PG, Scholes RJ, Boyle E, Canadell J, Canfield D, Elser J, et al. The global carbon cycle: A test of our knowledge of earth as a system. Science. 2000;290128176(13):291–296. https://doi.org/10.1126/science.290.5490.291

2. Mu Q, Zhao M, Running SW. Evolution of hydrological and carbon cycles under a changing climate. Part III: Global change impacts on landscape scale evapotranspiration. Hydrol Process. 2011;25(26):4093–4102. https://doi.org/10.1002/hyp.8367

3. Lal R. Agricultural activities and the global carbon cycle. Nutr Cycl Agroecosystems. 2004;70(2):103–116. https://doi.org/10.1023/B:FRES.0000048480.24274.0f

4. Sumner ME, editor. Handbook of soil science. Boca Raton, FL: CRC Press; 2000.

5. Rawls WJ, Pachepsky YA, Ritchie JC, Sobecki TM, Bloodworth H. Effect of soil organic carbon on soil water retention. Geoderma. 2003;116(1–2):61–76. https://doi.org/10.1016/s0016-7061(03)00094-6

6. Hudson BD. Soil organic matter and available water capacity. J Soil Water Conserv. 1994;49(2):189–194.

7. Rawls WJ, Nemes A, Pachepsky Y. Effect of soil organic carbon on soil hydraulic properties. Dev Soil Sci. 2004;30:95–114. https://doi.org/10.1016/S0166-2481(04)30006-1

8. Ankenbauer KJ, Loheide SP. The effects of soil organic matter on soil water retention and plant water use in a meadow of the Sierra Nevada, CA. Hydrol Process. 2017;31(4):891–901. https://doi.org/10.1002/hyp.11070

9. Kay BD, Da Silva AP, Baldock JA. Sensitivity of soil structure to changes in organic carbon content: Predictions using pedotransfer functions. Can J Soil Sci. 1999;77(4):655–667. https://doi.org/10.4141/S96-094

10. Resurreccion AC, Moldrup P, Tuller M, Ferré TPA, Kawamoto K, Komatsu T, et al. Relationship between specific surface area and the dry end of the water retention curve for soils with varying clay and organic carbon contents. Water Resour Res. 2011;47(6):1–12. https://doi.org/10.1029/2010wr010229

11. Da Costa A, Albuquerque JA, Da Costa A, Pértile P, Da Silva FR. Water retention and availability in soils of the State of Santa Catarina-Brazil: Effect of textural classes, soil classes and lithology. Rev Bras Ciência do Solo. 2013;37(6):1535–1548. https://doi.org/10.1590/S0100-06832013000600010

12. Franzluebbers A. Water infiltration and soil structure related to organic matter and its stratification with depth. Soil Tillage Res. 2002;66(2):197–205. https://doi.org/10.1016/S0167-1987(02)00027-2

13. Shaxson TF. Re-thinking the conservation of carbon, water and soil: A different perspective. Agron Sustain Dev. 2006;26:9–19. https://doi.org/10.1051/agro:2005054

14. Lal R. Physical properties and moisture retention characteristics of some Nigerian soils. Geoderma. 1978;21(3):209–223. https://doi.org/10.1016/0016-7061(78)90028-9

15. Danalatos NG, Kosmas CS, Driessen PM, Yassoglou N. Estimation of the draining soil moisture characteristic from standard data as recorded in routine soil surveys. Geoderma. 1994;64(1–2):155–165. https://doi.org/10.1016/0016-7061(94)90095-7

16. Saxton KE, Rawls WJ. Soil water characteristic estimates by texture and organic matter for hydrologic solutions. Soil Sci Soc Am J. 2006;70(5):1569. https://doi.org/10.2136/sssaj2005.0117

17. Olness A, Archer D. Effect of organic carbon on available water in soil. Soil Sci. 2005;170(2):90–101. https://doi.org/10.1097/00010694-200502000-00002

18. Shaxson TF. Re-thinking the conservation of carbon, water and soil: A different perspective. Agron Sustain Dev. 2006;26:9–19. https://doi.org/10.1051/agro:2005054

19. Pachepsky YA, Rajkai K. Tóth B. Pedotransfer in soil physics: Trends and outlook – A review. Agrokémia és Talajt. 2015;64(2):339–360. https://doi.org/10.1556/0088.2015.64.2.3

20. Van Looy K, Bouma J, Herbst M, Koestel J, Minasny B, Mishra U, et al. Pedotransfer functions in Earth system science: Challenges and perspectives. Rev Geophys. 2017;55(4):1199–1256. https://doi.org/10.1002/2017RG000581

21. Bauer A. Influence of soil organic matter on bulk density and available water capacity of soils. Farm Res. 1974:44–52. https://library.ndsu.edu/repository/handle/10365/24299

22. Nemes A, Timlin DJ, Pachepsky YA, Rawls WJ. Evaluation of the pedotransfer functions for their applicability at the U.S. National Scale. Soil Sci Soc Am J. 2009;73(5):1638. https://doi.org/10.2136/sssaj2008.0298

23. Du Preez CC, Van Huyssteen CW, Mnkeni PNS. Land use and soil organic matter in South Africa 1: A review on spatial variability and the influence of rangeland stock production. S Afr J Sci. 2011;107:27–34. https://doi.org/10.4102/sajs.v107i7/8.354

24. Schütte S, Schulze RE, Paterson G. Identification and mapping of soils rich in organic carbon in South Africa as a climate change mitigation option. Pretoria: Department of Environmental Affairs; 2019. Available from: https://www.environment.gov.za/sites/default/files/reports/identificationandmapping_soilsrich_organiccarboninsouthafrica.pdf

25. Schulze RE, Schütte S. Mapping soil organic carbon at a terrain unit resolution across South Africa. Geoderma. 2020;373:114447. https://doi.org/10.1016/j.geoderma.2020.114447

26. Soil Survey Staff, ARC-ISCW. 1:250 000 scale land type survey of South Africa. Pretoria: Agricultural Research Council –Institute for Soil, Climate and Water; no date.

27. Schulze RE, Horan MJC, Kunz RP, Lumsden TG, Knoesen D. Methods 2: Development of the Southern African Quinary Catchments Database. In: Schulze RE, Hewitson BC, Barichievy KR, Tadross MA, Kunz RP, Horan MJC, et al. Methodological approaches to assessing eco-hydrological responses. Pretoria: Water Research Commission; 2010; p. 63–74.

28. Schulze RE. Hydrology and agrohydrology – A text to accompany the ACRU 3.00 Agrohydrological Modelling System. Pretoria: Water Research Commission; 1995.

29. Hughes CJ. Degradation of ecological infrastructure and its rehabilitation for improved water security [unpublished doctoral thesis]. Pietermaritzburg: University of KwaZulu-Natal; 2018.

30. Lynch SD. Development of a raster database of annual, monthly and daily rainfall for southern Africa. Pretoria: Water Research Commission; 2004.

31. Schulze RE, Schütte S. Climate zones and climate change. In: Schulze RE, editor. Handbook for farmers, officials and other stakeholders on adaptation to climate change in the agriculture sector within South Africa Section B: Agriculture's natural capital in South Africa: A climate change perspective. Pietermaritzburg: Department Agriculture, Forestry and Fisheries; 2016. p. 173–179. Available from: https://www.daff.gov.za/daffweb3/Branches/Forestry-Natural-Resources-Management/Climate-Change-and-Disaster-Management/Climate-Change-Unit

32. Acocks JPH. Veld types of southern Africa. 3rd ed. *Memoirs* of the *Botanical Survey* of *South Africa 57*. Pretoria: Botanical Research Institute; 1988. p.1–146.

The urban water metabolism of Cape Town: Towards becoming a water sensitive city

AUTHORS:
J. Ffion Atkins[1]
Tyrel Flügel[2,3]
Rui Hugman[2]

AFFILIATIONS:
[1]Department of Environmental and Geographical Sciences, University of Cape Town, Cape Town, South Africa
[2]Umvoto Africa (Pty) Ltd, Cape Town, South Africa
[3]Department of Geography and Environmental Studies, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Ffion Atkins

EMAIL:
ffion.atkins@uct.ac.za

DATES:
Received: 14 July 2020
Revised: 02 Oct. 2020
Accepted: 22 Nov. 2020
Published: 28 May 2021

HOW TO CITE:
Atkins JF, Flügel T, Hugman R. The urban water metabolism of Cape Town: Towards becoming a water sensitive city. S Afr J Sci. 2021;117(5/6), Art. #8630. https://doi.org/10.17159/sajs.2021/8630

ARTICLE INCLUDES:
☒ Peer review
☒ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☒ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Yali Woyessa

KEYWORDS:
water sensitive cities, performance indicators, urban water management, managed aquifer recharge, water equity

FUNDING:
South African National Research Foundation (grant no. 114696)

To improve its resilience to increasing climatic uncertainty, the City of Cape Town (the City) aims to become a water sensitive city by 2040. To undertake this challenge, a means to measure progress is needed that quantifies the urban water systems at a scale that enables a whole-of-system approach to water management. Using an urban water metabolism framework, we (1) provide a first city-scale quantification of the urban water cycle integrating its natural and anthropogenic flows, and (2) assess alternative water sources (indicated in the New Water Programme) and whether they support the City towards becoming water sensitive. We employ a spatially explicit method with particular consideration to apply this analysis to other African or Global South cities. At the time of study, centralised potable water demand by the City amounted to 325 gigalitres per annum, 99% of which was supplied externally from surface storage, and the remaining ~1% internally from groundwater storage (Atlantis aquifer). Within the City's boundary, runoff, wastewater effluent and groundwater represent significant internal resources which could, in theory, improve supply efficiency and internalisation as well as hydrological performance. For the practical use of alternative resources throughout the urban landscape, spatially explicit insight is required regarding the seasonality of runoff, local groundwater storage capacity and the quality of water as it is conveyed through the complex urban landscape. We suggest further research to develop metrics of urban water resilience and equity, both of which are important in a Global South context.

Significance:

- This research provides the initial groundwork of quantifying the magnitude of the urban water cycle of the City of Cape Town at an annual timescale, in relation to becoming a water sensitive city. The urban water metabolism framework used in this study provides important insight to assess whole-of-system urban water dynamics and to benchmark progress towards becoming water sensitive. By quantifying the magnitude of flows into and out of the urban system, this research sheds light on the opportunities to improve circularity in the urban water cycle. The spatial approach adopted here provides a platform to interrogate the urban landscape and its role in the urban water cycle. By using data products that are available via national data sets or remote sensing, this approach can be applied to other African or Global South where data is characteristically scarce. Further work is required to establish metrics that can adequately describe urban water resilience and equity.

Introduction

Current urban water management practices are challenged by climate change, increasing per-capita water demand and growing populations. Already, many cities around the world are vulnerable to water shortages[1-3] and it has been suggested that one in six large cities is likely at risk of a significant water deficit by 2050[4]. Recent water crises experienced by Bengaluru (India)[5], Los Angeles (USA)[3] and Cape Town (South Africa)[6-8] highlight potential consequences of failing to account for the future water demands of a city within the context of climate change related stresses.

The City of Cape Town (hereafter the City) is largely dependent upon the surface storage of rainfall from the surrounding catchment areas to supply water to its residential, industrial and commercial users. A severe multi-year drought between 2015 and 2018 highlighted the vulnerability of a growing city being reliant solely on the surface storage of rainfall. Although the event was considered an extremely rare (1 in 300 year) drought[9], there is strong evidence for drying and warming of the regional climate systems[8] with observed, and modelled, long-term increases in aridity for most of southern Africa[9-11]. Additionally, reduced precipitation patterns have been observed in other Mediterranean-like climates of the southern hemisphere[12,13], strengthening the prediction that severe droughts, like that experienced in 2015–2018, are likely to occur more frequently in the Cape Town region[8].

While demand management strategies were highly effective in reducing water consumption by 50% during the 2015–2018 drought[14], it is clear that demand management alone will be an insufficient adaptation measure for future climate scenarios. The Water Strategy for Cape Town[15], which was published in response to the 2018 water crisis, recognises the need to take a whole-of-system approach and represents a welcome shift in urban water management paradigm and practice. The New Water Programme[15] outlines the various planned contributions of alternative sources (groundwater, desalination and water reuse) and interventions such as reducing water demand, clearing alien vegetation and augmenting surface storage. The Water Strategy also highlights the City's commitment to transition to a water sensitive city and the 'optimal use stormwater and urban waterways for the purpose of flood control, aquifer recharge, water reuse and recreation'[15]. The water crisis faced by the City highlighted the need to be able to consider management interventions in relation to the dynamics of the urban system as a whole, which comprises social-ecological-technical, hydrological and economic components and processes. This research lays the groundwork of quantifying the magnitude of the urban water cycle at an annual timescale in order to contribute to an assessment of the water sensitivity of the City.

Water sensitive cities

A water sensitive city approaches urban water management as a holistic system, gives water due prominence in the design of urban areas and is underpinned by three key pillars[16]:

1. Cities as supply catchments: access to diverse water sources, both centralised and decentralised.
2. Cities providing ecosystem services: the urban landscapes actively support and supplement the natural environment.
3. Cities comprising water sensitive communities: emphasising the importance of socio-political capital for water sensitive behaviours and decision making.

Quantifying water sensitivity

The initial step for determining a city's degree of water sensitivity is quantifying the rate and direction of flows. This quantification of flows into, within, and out of urban areas is often referred to as urban metabolism[17] and is a powerful empirical analysis of the society–nature interaction[18]. Urban metabolism focuses on quantifying the fluxes of energy, materials and nutrient flows into and out of urban areas[19,20] and is readily applied into fields of urban planning and design[19]. For the most part, the field quantifies flows that are anthropogenically driven (e.g. water and energy consumption) and is only recently gaining traction as a framework to assess urban water cycles[21] and performance[22-24].

Performance indicators are important in the implementation, assessment and communication of progress towards sustainability goals.[22] There are several methods to benchmark the performance of water management, using indicators such as Sustainable Cities Index[25], Green Cities Index[26], City Blueprint[27] and more recently the Water Sensitive Cities Index[22]. A City Blueprint assessment was done for Cape Town[28], evaluating the water governance processes and capacity required to implement water sensitive urban design (WSUD). Their assessment highlighted the shift in governance processes related to water scarcity during the 2018 water crisis and the potential to adopt WSUD through successful implementation of policies such as Management of Urban Stormwater Impacts Policy and the Flood Plain and River Corridor Management Policy. A review of methods used to evaluate urban water performance by Renouf and Kenway[24] found that, with the exception of the Water Sensitive Cities Index, evaluation criteria are often misaligned from the visions and objectives of urban water management. At the macroscale (the city as a whole), there is little quantitative assessment to monitor progress towards (and gauge the performance of a city against) its urban water management visions and objectives. Earlier research by Kenway and colleagues[21] addressed this void, applying a mass balance analysis of several cities in Australia. They proposed an urban water metabolism framework (UWMF) to quantify and assess the performance of a city, which was also the first attempt to quantify the term 'water sensitive city'. They integrated both the anthropogenic (bulk water supply, wastewater effluent, reuse) and natural hydrological flows (precipitation, evapotranspiration, runoff, groundwater infiltration) into the water mass balance. From this metabolic framework, they derived performance indicators that allowed each city's water management effectiveness to be compared. They argue that quantifying these volumes is the first step towards designing a water sensitive city[21], and this also provides a benchmark for measuring progress and comparison between different cities. A method of accurate comparison is important to allow for cities to learn best practices from one another. Together, the metabolic framework and performance indicators provide powerful metrics to guide urban water management and policy along a trajectory towards being water sensitive.

Urban metabolism in Cape Town

Several studies have quantified the fluxes of materials, energy and water into Cape Town (at a city scale) using varying methods such as an ecological footprint assessment[29] an economy-wide material flow analysis[30]; systems analysis[31] and urban metabolism[32]. In terms of mass flux, water comprises the largest component – about 96% of the total material flows into and out of Cape Town[29] – a fairly consistent proportion observed in all urban metabolism studies[17,21,33]. Currie et al.[32] conducted a broad assessment of the urban metabolism of Cape Town, quantifying energy, water, people, transport, food and solid waste. For water, they primarily focused on municipal water services, i.e. the anthropogenic flows, and illustrated the average volumes of water per year that flow into and out of the City, compartmentalising these volumes into various sources (the different reservoirs) and pathways (i.e. domestic/commercial/industrial demand) within the City. While their analysis was comprehensive and provides novel insight into the spatial patterns of water consumption, from an urban water cycle perspective they did not include hydrological flows or alternative water supply (i.e. reuse, desalination, groundwater or runoff). An earlier study by Ahjum and Stewart[31] aimed to assess the energy costs associated with the various demand and alternative (municipal) supply scenarios. They quantified parts of the urban water cycle from a systems perspective, did include groundwater supply (and recharge), and omitted all other components of the hydrological cycle (e.g. rainfall, runoff). Thus there exists a gap in the quantification of how the various flows into/within and out of the City are integrated and in some cases interdependent.

This research builds on such existing work but integrates the anthropogenic and hydrological components into one framework. It aims to contribute to the growing inventory of urban water metabolism within the context of holistic urban water management and the commitment to becoming a water sensitive city. The choice of spatially appropriate methods employed in this study is considered in terms of their applicability to other African and Global South cities, where data of the urban environment (e.g. stormwater runoff rates) are often scarce[34,35]. Using the UWMF proposed by Renouf et al.[24], we assess whether the interventions proposed in the New Water Programme[14] support the City towards becoming water sensitive. The particular objectives of this research are to (1) conduct a water mass balance analysis for the City of Cape Town, using available spatial data products (e.g. mean annual precipitation, landcover, evapotranspiration), keeping in mind scalability and comparability across cities, (2) assess performance of the urban water cycle in relation to the principles of water sensitivity, and (3) assess performance of the urban water cycle under the New Water Programme.

Methods

The urban water cycle of Cape Town was quantified as a mean annual average and was assessed in the context of a water sensitive city under a 'normal' non-drought scenario and under the New Water Programme. In brief, the steps taken were to:

1. Define the system boundary (Figure 1) and quantify all parameters of the urban water cycle (Figure 2), both anthropogenic and hydrological flows.
2. Conduct mass balance analyses (Equation 1 and Table 1) of the urban water cycle in a non-drought 'normal' scenario (Figure 2a) and under the New Water Programme (Figure 2b).
3. Assess the water sensitive performance of the urban water cycle (Table 2) using indicators stipulated in Renouf et al.[36]
4. Test and compare the performance of a hypothetical water cycle under the proposed New Water Programme.

System boundary

The system is defined as the City of Cape Town Metropole (Figure 1). Within the metropolitan boundary there are two primary aquifers (Cape Flats and Atlantis) and secondary aquifers (Table Mountain Group, TMG). The primary sand aquifers are found directly below the urban (Cape Flats Aquifer) and peri-urban (Atlantis) environments, and are a direct part of the urban hydrological cycle. Furthermore, they are a source for centralised and decentralised water supply, as well as subject to managed aquifer recharge (MAR) from treated wastewater and stormwater runoff. Although there are TMG aquifers within the urban boundary (i.e. Table Mountain itself), in this paper, we did not consider TMG as a source of groundwater supply within the City. References to TMG are in regard to aquifers beyond the metropolitan boundary (surrounding Steenbras) and are thus considered inputs into the system rather than internal flows.

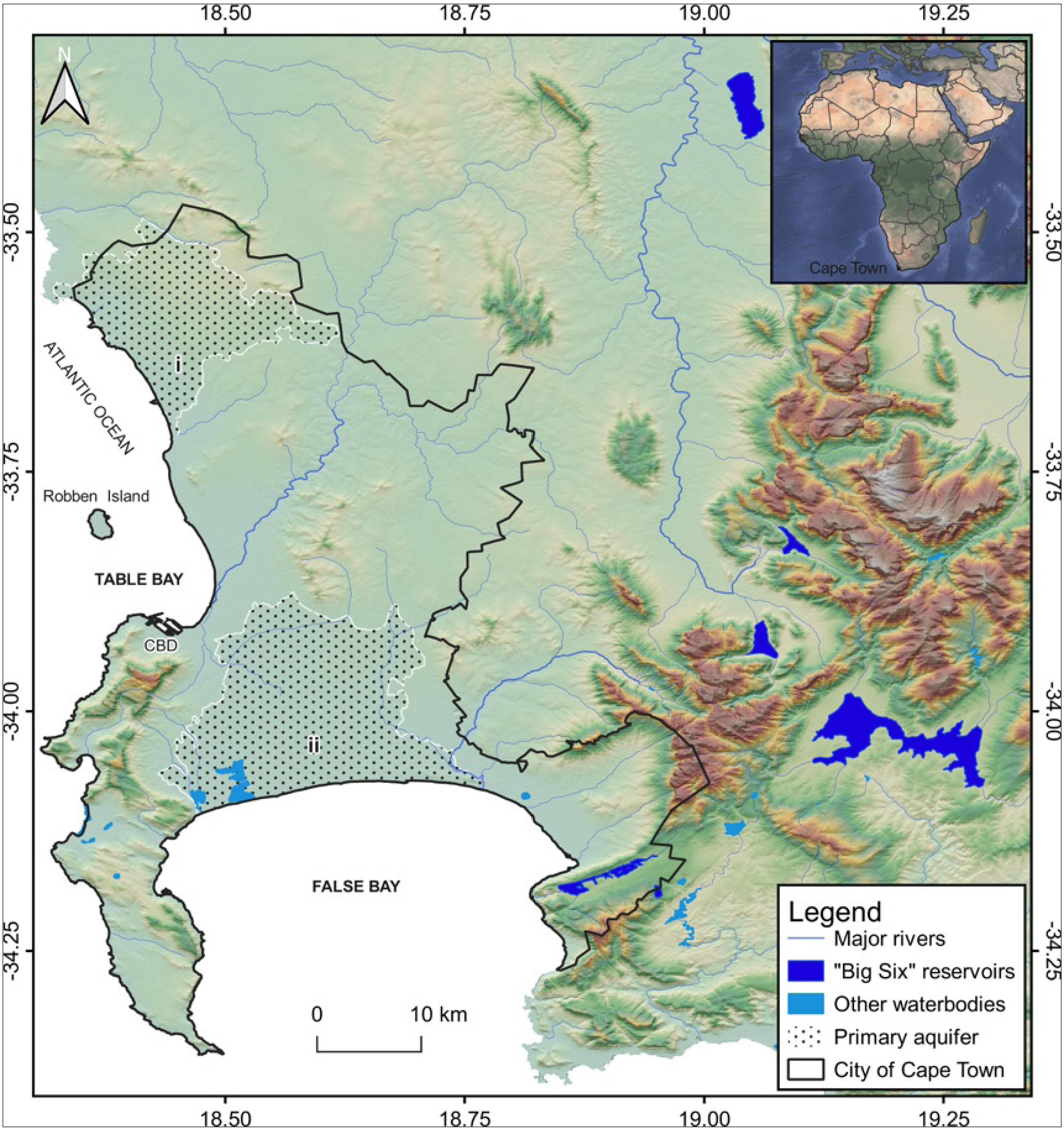


Figure 1: The City of Cape Town metropolitan boundary (heavy black line) and its surrounding area with rivers and the surface storage ('Big Six') reservoirs (referred to in text as WCWSS) highlighted in deep blue. Atlantis (i) and Cape Flats Aquifer (ii) shown with stippling.

Mass balance

The urban water cycle comprises anthropogenic and hydrological flows of water into, within and out of the city (Figure 2). Anthropogenic flows represent the volumes of water consumed and discharged by the urban areas serviced by the City with the point of entry as the water treatment plants and point of exit as the wastewater treatment works (WWTW). Hydrological flows represent precipitation, runoff, and groundwater recharge that occur within the defined system boundary, as well as surface and groundwater discharge out of the system into the ocean. Water that flows into the urban system via rivers and aquifers is not accounted for, as it is considered environmental flow, as per Renouf et al.[36] Decentralised groundwater abstraction and non-potable reuse are assumed here to be their own separate outputs for convenience sake, but as these rates may increase in the future, they could equally be considered to leave the system as evapotranspiration or groundwater discharge.

The water mass balance assumes a steady state and follows Equation 1:

$$Qi = Qo,$$

$$(P+C+Si) + Rw + MAR = (W+Rs+Et+Dg+Cufw+Gd+Rwn) - Rw - MAR \qquad \text{Equation 1}$$

where Q_i is the sum of all inputs and Q_o is the sum of all outputs (including losses). Inputs consist of centralised bulk water supply (C) which comprises surface supply (Cs), desalination (Cd), TMG aquifer (Cg), precipitation (P) and surface water inflow (Si). Outputs consist of wastewater effluent (W), runoff (Rs), groundwater abstraction (Dg), groundwater discharge (Gd), non-potable reuse (Rwn) and losses (Cufw).

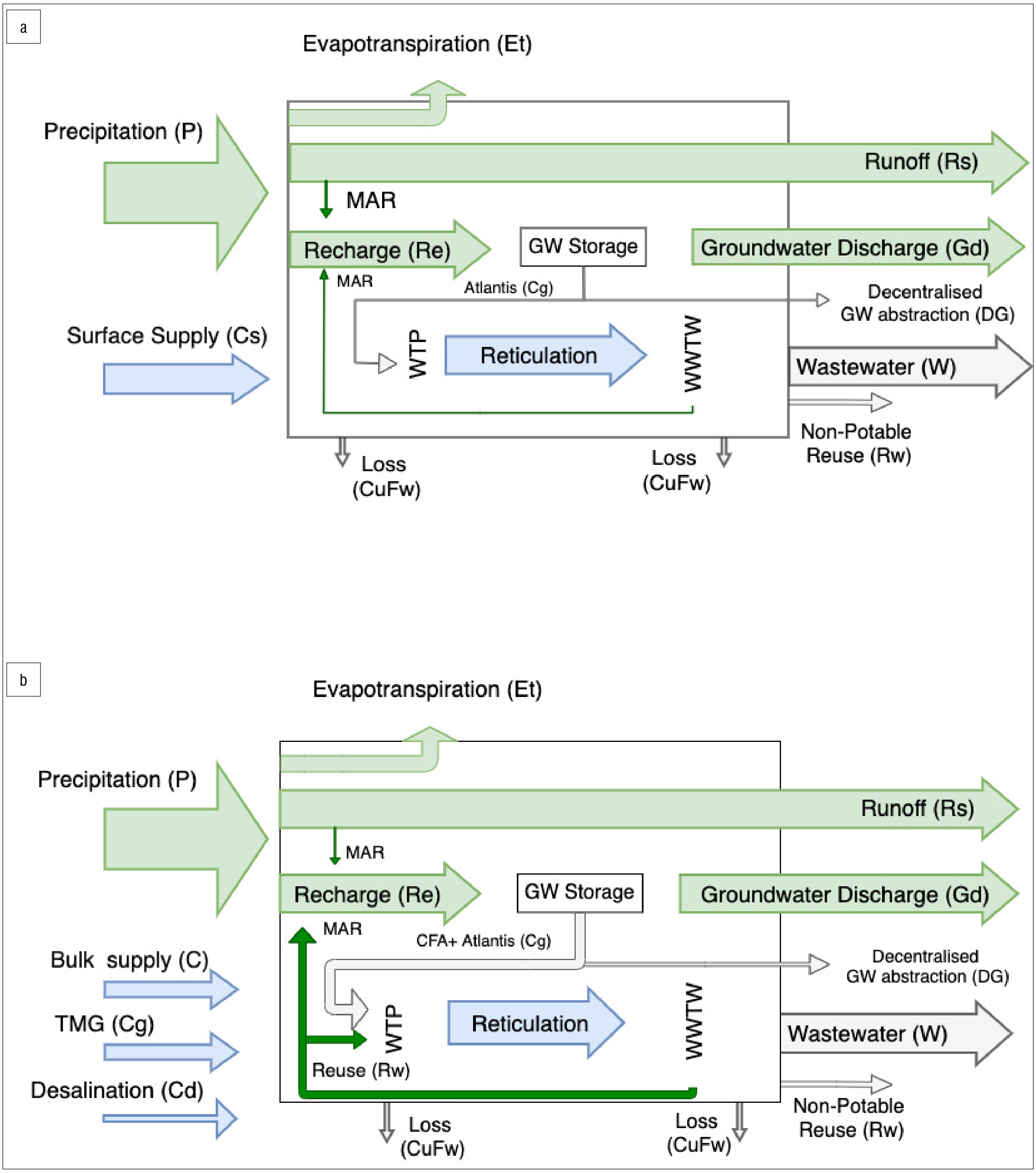


Figure 2: Conceptual model of (a) pre-drought, baseline urban water cycle of Cape Town (Scenario 1) and (b) the cycle under the proposed New Water Programmes (Scenario 2). Blue arrows indicate anthropogenic inputs, light green arrows are hydrological flows, grey arrows are flows of non-potable quality, and dark green arrows are recycled water. The size of the arrows conceptually indicates the magnitude of the labelled parameter.

Water recycling terms (Rw and MAR) refer to potable reuse (as per the New Water Programme) and managed aquifer recharge (MAR) and are included as both Q_i and Q_o, but are subtracted from outputs.

Estimating flows

Anthropogenic flows

Monthly flow data for bulk water supply and wastewater effluent were obtained from the City. Annual averages of both were calculated, using the data available for the time period between 2008 and 2012. Recycled water volumes were sourced from the City of Cape Town.[14] The Western Cape Water Supply System (WCWSS) gets its water from six major rainfed dams that supply the City of Cape Town (~64%), agriculture (~29%) and other municipalities in the region (~7%).[14] In this study we focus solely on the City of Cape Town and its use of WCWSS water. We omit water use of WCWSS water by other municipalities and agriculture owing to paucity of reliable data, but acknowledge this use merits further research. We do include the peri-urban agricultural area as it constitutes significant decentralised groundwater abstraction within the City's boundary. Virtual water and other forms of imported and exported water (e.g. imported in food and bottles) have not been included in this analysis but also merit future investigation.

Groundwater abstraction rates, both centralised and decentralised, were assumed to be the maximum permitted allowance, as stipulated in the water use licence. For decentralised abstractions, the allowed abstraction rate (7.6 GL/year) refers to the Philippi Horticultural Area – an urban farming locale within the city boundaries. Although considered agricultural use, these abstraction rates have been included owing to the direct use of the Cape Flats Aquifer which is an important decentralised supply relevant to this research. Private/residential abstraction rates are omitted from this study owing to paucity of data, although are expected to have increased during and since the drought (2015–2017). Centralised abstraction and MAR rates were assumed to be the maximum allowable rate as stipulated in the water use licence under the New Water Programme for the City.

Table 1: Type and sources of data and their time period (pre-drought)

Parameter		Data source	Data period
	Land use data	South African land-cover data sets	2014
Anthropogenic flows			
Cs	Surface supply	CoCT	2008–2012
Cd	Desalination	Water Strategy[15]	
Cg	Groundwater (TMG)	Water Strategy[15]	
W	Wastewater	City of Cape Town[37]	2008–2012
Rwn	Non-potable reuse	City of Cape Town[37]	2008–2012
Cufw	Loss	City of Cape Town[14]	
Rw	Water reuse	City of Cape Town[37]	
MAR	Managed aquifer recharge	Water Use Licence, Water Outlook[14]	
Hydrological flows			
P	Precipitation	South African Water Resources Book of Maps[38]	Annual average
Et	Evapotranspiration	Schulze et al.[39]	Annual average
Gd	Groundwater abstraction (CFA, Atlantis)	City of Cape Town[37]	Annual average
Rs	Stormwater runoff	SANRAL[40] runoff coefficients (see supplementary material)	
Re	Groundwater recharge	Water Balance Method (Re = P-Et-Rs)	Annual average
	Urban population	Stats SA	

Hydrological flows

We aimed to use spatial products that were readily available and locally relevant. A mean annual precipitation map was obtained from South African Water Resources Geographic Information System book of maps.[38] A mean annual evapotranspiration map was generated by Schulze et al.[39], using the FAO Penman–Monteith method[40], and was based on daily maximum and minimum temperatures, on a 1.7x1.7 km grid for 50 years, and empirically determined month-by-month gridded values of actual vapour pressure and daily gridded values of solar radiation. Runoff was calculated as per the rational method using runoff coefficients according to SANRAL[41], taking into consideration soil type, slope and land use (Supplementary tables 1–3). Several other methods to calculate runoff were explored and compared; these are detailed in the supplementary material. The values of hydrological parameters presented in the analysis represent the sum of all pixels within the metropolitan boundary. We acknowledge the potential for uncertainties associated with these spatial products to propagate error through to the final assessment of the City's performance and thus tested the sensitivity of performance indicators to these hydrological parameters, which is detailed in Supplementary table 7.

All flows were then visualised using a Sankey diagram which was constructed with the help of Python code published by Lupton and Allwood[42].

Performance indicators

Performance indicators for urban water efficiency, water supply internalisation and hydrological performance were calculated following Renouf et al.[24,36] (Table 2). Renouf et al.[24,36] also include more indicators such as water-related energy efficiency and nutrient-related water efficiency, but owing to the paucity of necessary data for a robust assessment of such indicators, we have focused on the three aforementioned. Calculation of hydrological performance required estimating annual stormwater runoff and groundwater recharge in a pre-urbanised, or historical, environment. Historical runoff estimates were calculated following the same method for present day runoff except the urbanised areas were replaced by the South African National Vegetation map (City of Cape Town Open Data Portal). Details of runoff coefficient classifications are in Supplementary table 4.

Table 2: Urban water performance indicators adopted and adapted from Renouf et al.[36] Cext refers to all centralised, external sources for the Western Cape Water Supply System (WCWSS) and includes surface dams, desalination and Table Mountain Group (TMG) aquifer. Cint refers to internal sources of centralised (groundwater from Atlantis or Cape Flats aquifers and recycled water) and decentralised (e.g. rainwater harvesting) water.

Performance indicator	Description	Equation
Urban water efficiency	Total external water use per capita (kL/capita/year)	C_{ext}/Population
Water supply internalisation	Proportion of total urban water demand met by internally harvested/recycled water	$(C_{int} + D)/$ $(C_{int}+C_{ext}+D)$
Hydrological performance	Ratio of post- (i) to pre-urbanised (o) annual stormwater runoff (Rs) and groundwater recharge (Re)	Rs(i)/Rs(o), Re(i)/Re(o)

Scenario assessment

Two separate urban water cycles were quantified: the cycle which is considered 'normal' or non-drought (2008–2012) (Figure 2a) and is hereafter referred to as Scenario 1; and the cycle under the scenario of the New Water Programme (Figure 2b and Table 1), hereafter referred to as Scenario 2. For the New Water Programme, effective yields of the alternative water sources were derived from the Water Strategy[15] and are detailed in Table 3. In this water cycle, supply/demand remain as they are for Scenario 1 in order to assess the impact of the alternative supply interventions on the water cycle as a whole.

Table 3: Effective yields of the alternative water sources for the New Water Programme as outlined in the Water Strategy[15]. Units stipulated in the original report are given in megalitres per day and million kilolitres per annum, the latter of which equates to gigalitres per year (GL/year), as reported in this study.

Source	Pre-drought yield (GL/year)	Effective yield (GL/year)
Groundwater (CFA and Atlantis)	3.3	20.1
Groundwater (Table Mountain Group)	0	18.3
Reuse	0	25.6
Desalination	0	18.3
Berg River Dam augmentation[a]		14.6
Alien vegetation clearing[a]	-	20.1
Demand management[a]	-	25.6
Managed aquifer recharge (CFA and Atlantis)	4.65[b]	25.5[c]

[a] *These are not tested in this framework explicitly as they are considered to contribute to existing surface water capacities rather than as alternative resources.*

[b] *Atlantis (rough estimate)*

[c] *Value comes from Water Outlook[14]*

Results and discussion

Mass balance

All data were collated into a mass balance analysis using Equation 1 and are summarised in Table 4 and Figure 3. In Scenario 1, total inputs equate to 1796 GL/year comprising precipitation and bulk water supply; total outputs equate to 1471 GL/year comprising wastewater effluent, runoff, evaporation, groundwater discharge, decentralised groundwater abstraction, recycled water and accounted for losses; and internal flows amount to 745 GL/year as groundwater recharge and abstraction. In Scenario 2, total inputs are reduced to 1754 GL/year where bulk water supply is divided into surface water, centralised groundwater abstraction and centralised desalination; internal flows are increased to 803 GL/year to include centralised groundwater abstraction from the internal primary aquifers, water recycling for potable use and MAR; outputs marginally decreased to 1709 GL/year, mostly reflected in reduced wastewater effluent discharge and reduced groundwater discharge.

Assuming long-term averages are in equilibrium, inputs should equate to outputs, and thus we find a 2% error in the water cycle as a whole for both scenarios. As divided into anthropogenic and natural flows, 7% and 3% errors are found respectively. With regard to the 7% error in anthropogenic flows, we propose that it is a result of unaccounted for losses from the system. Bulk water supply (325 GL/year) far exceeded effluent released from WWTW (235 GL/year), representing a 27% loss from the system, as measured from points of entry via water treatment plants to the points of exit (WWTW). On average, the City reports ~15% loss between the water treatment plant and consumers (water that is sold); the remainder is loss between the consumer and WWTW discharge (including irrigation of gardens, pools, leaking sewers, and loss from the WWTW itself). Known losses (up to ~15%) are described in Figure 3 as ‘loss’, and what is unknown is described as ‘Unaccounted for’, in order to achieve conservation of mass within the analysis. The simplification of household behaviour and losses between consumers and the WWTW (mentioned in the methods section) has a considerable impact on macroscale dynamics and merits further, more detailed investigation. The 3% error in hydrological flows likely represents inaccuracies that exist in the spatial data and methods used to estimate natural flows, and are evaluated in the following section.

Table 4: Water mass balance of the water cycle pre-drought scenario (Scenario 1), and the hypothetical water cycle under the New Water Programme (Scenario 2)

		Scenario 1 (GL/year)	Scenario 2 (GL/year)
	Input		
P	Precipitation	1471.4	1471.4
Csw	Bulk water supply (surface water dams)	324.9	246.0
Cg	Centralised groundwater abstraction (TMG)	0	18.3
Cd	Centralised desalination	0	18.3
Sub-total		**1796.2**	**1753.9**
	Internal flow		
Cg	Centralised groundwater abstraction (CFA+Atl)	3.3	20.1
Rw	Recycled water (potable use)	0.0	25.6
Rw(MAR)	Managed aquifer recharge	0.0	16.1
Re	Groundwater recharge	741.7	741.7
Sub-total		**745.0**	**803.4**
	Output		
Dr	Decentralised rainwater harvest	0.0	0.0
Dg	Decentralised groundwater abstraction	26.9	26.9
W	Wastewater effluent	234.8	157.5
Cufw	Known losses	48.6	84.7
Rw	Recycled water (non-potable use)	18.8	18.8
Rs	Surface runoff	492.3	466.8
Gd	Groundwater discharge	0.0	710.8
ET	Evapotranspiration	711.6	218.6
Sub-total		1751.7	1709.6
	Water balance (total) (error)	**2%**	**2%**
	Water balance (anthropogenic flows) (error)	7%	7%
	Water balance (hydrological flows) (error)	1%	3%

Estimating hydrological flows

The total volume of rainfall within the City boundary is 1471.4 GL/year, equating to an average of 605 mm/year. Of this, an average of 219 GL/year is lost to evapotranspiration, 492 GL/year to runoff and 742 GL/year to recharge (Table 4 and Figure 4). These respectively equate to spatial averages of 15%, 33% and 50% of total rainfall. The spatial distribution of each hydrological flow is inherently heterogeneous (Figure 4), where spatial estimates of evapotranspiration vary spatially between 0 and 30% of precipitation, runoff varies between 17% and 96%, and recharge between 0 and 83% (Supplementary figure 2).

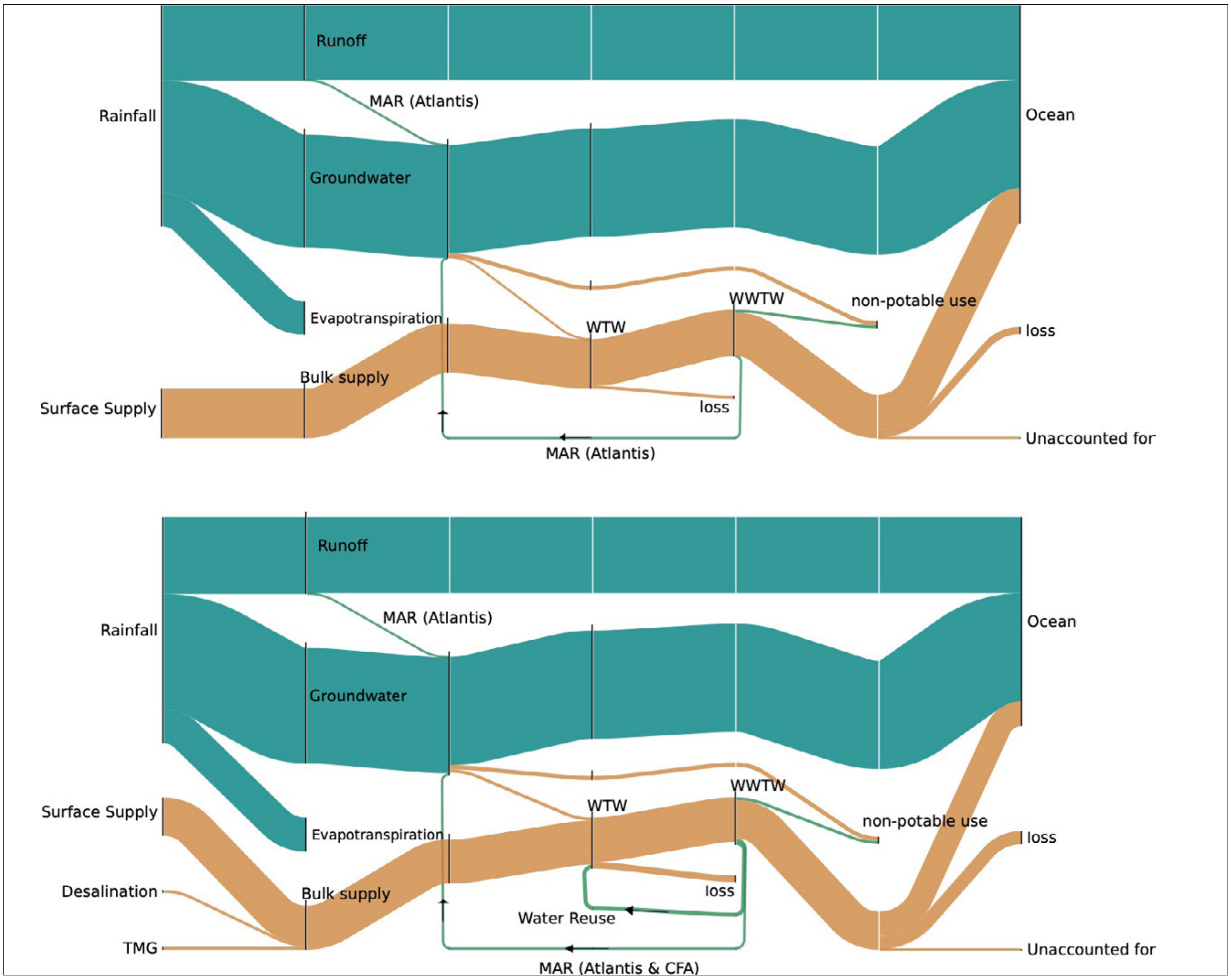


Figure 3: Sankey diagrams illustrating the balance of inputs and outputs of anthropogenic (brown) and natural (teal) water in the City of Cape Town, representing average gigalitres per annum for the years 2008–2012. (a) Scenario 1 represents non-drought conditions where external water from WCWSS represents external supplies from the surface 'Big Six' reservoirs. What is extracted is considered 'surface supply' and passed through a water-treatment plant into the reticulation system and through the wastewater treatment works (WWTW) and eventually discharged into rivers and/or the ocean. Known losses occur at several stages of the reticulation network and the term 'unaccounted for' water has been included to allow conservation of mass, representing potential error. Groundwater is recharged via managed aquifer recharge (MAR) using runoff and treated effluent. (b) Scenario 2 represents the same data set but includes the partial replacement of 'surface supply' via diversification of external sources from desalination and Table Mountain Group (TMG) aquifers. In addition, resource internalisation (green lines) takes the form of *direct reuse* (recycled water from WWTW back to WTW) as well as *indirect reuse* (from WWTW to groundwater) in the form of MAR.

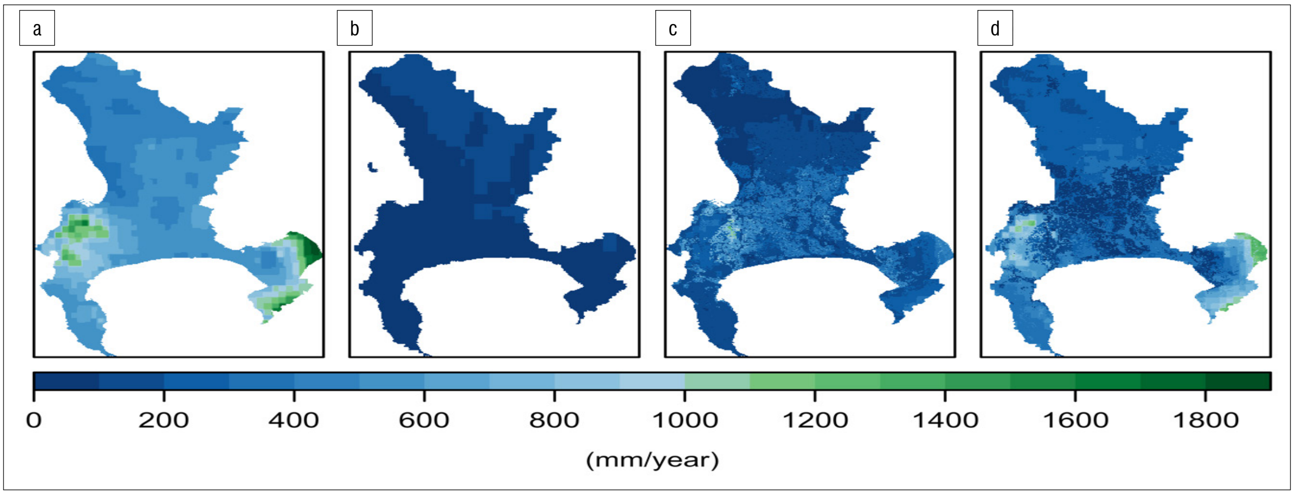


Figure 4: Annual averages of (a) precipitation (from Bailey and Pitman[38]), (b) evapotranspiration (from Schulze et al.[39]), (c) runoff (from SANRAL[40]), and (d) recharge (using the water balance method: Re= MAP-Et-Rs where Re is groundwater recharge, MAP is mean annual precipitation, Et is evapotranspiration and Rs is stormwater runoff).

Several methods were used to calculate runoff rates, resulting in wide variation in estimates of runoff (see Supplementary figure 1). A comparison was made between all runoff estimates of this study and a stormwater runoff model of the Zeekoevlei catchment using the software PCSWMM.[43] An average runoff estimate using PCSWMM was 187.6 mm/year[43] and our estimates ranged between 191 mm/year and 277 mm/year for the same catchment area. All runoff estimates of this study were marginally higher than estimates by Okedi[43], likely representing storm conditions, but all estimates were within a reasonable range. Using the coefficients from SANRAL[40] did result in the highest runoff value (277 mm/year) for the Zeekoevlei catchment, but using these coefficients also gave recharge values that were most consistent with those reported in the literature for the Atlantis aquifer[44]. Recharge rates in Atlantis were reported to be between 10% and 30.3% of rainfall, with an average of 16%.[41(and references therein)] The recharge rates calculated in this study are particularly high when compared to the average reported by Jovanovic et al.[44] and it is likely that they represent potential rather than effective recharge. If this is the case, evapotranspiration rates[41] may be an underestimate. Earlier studies into urban evapotranspiration have observed evapotranspiration to constitute 40–80% of a city's annual water balance losses.[45] However, Hobbie et al.[46] (see their supplementary material) conducted a water mass balance using recommended coefficients from the Minnesota Stormwater Manual[47] and estimated average evapotranspiration in their urban watershed to be 14.5% of precipitation. Their watershed comprised urban (impervious), peri-urban and rural areas, similar to the landcover types found in the urban boundary of this study. Evapotranspiration is inherently patchy within an urban environment and the data product used[41] has a resolution of 74 m which may not have captured the full variability. In light of this, we did assess the applicability of using MODIS16A3 product for the years of this study (2008–2012), and we found, for all years, the core urban area to be flagged as 'no data' and thus deemed unusable for spatial calculations of recharge. We assessed the potential for this unavoidable uncertainty in evapotranspiration products to propagate further error into the performance indicators (Supplementary table 7). This highlights both the challenges of working with spatial data in the urban environment and the scope to validate remote sensing products in highly heterogenous contexts.

Estimates of stormwater runoff rates using PCSWMM from more catchments within the urban environment would enable a more statistically robust comparison between methods. However, the generation of such estimates is a time-consuming, data-intensive process and if such estimates do already exist they have not been made available through the City of Cape Town Open Data Portal or formal channels of communication. The benefit of the methods employed in this study are that they are easily applied to landcover data that can be obtained from national (including satellite-derived) data sets that can be applied to many data scarce (namely in situ data) urban areas at multiple scales.

Limitations

We acknowledge several simplifications in the conceptualisation of the urban water cycle, most notably with regard to aggregates of household fluxes. For example, not all water that enters a household enters the sewer system, as some is lost due to gardening/recycling of grey water/filling of pools and we acknowledge that this may be significant when combined at the city scale and particularly so during drought. We have assumed these pathways to be aggregate losses out of the system between the point of entry (water treatment plant) and exit (WWTW), rather than contributing to, for example, non-potable water recycling or decentralised storage. Future work would investigate these fluxes as aggregate contributions of household behaviour on the macroscale water budget.

Performance scenario planning

To assess the metabolic performance of the current urban water cycle, performance indicators were calculated following Renouf et al.[24,36,48], as presented in Table 5. Pre-drought (2018), 99% of centralised water inputs into the City were supplied from external sources (C_{ext}) via WCWSS, with other decentralised sources such as groundwater and water recycling for potable and non-potable uses. Rainwater harvesting was assumed to be negligible; however, this is likely to have changed due to behavioural changes induced by the 2015–2018 water crisis.

Table 5: Performance indicators as stipulated using Renouf et al.[36] for the current water cycle and the cycle under the proposed New Water Programme which assumes alternative sources replace surface storage supply (WCWSS) relative to their full capacity, as stipulated in Table 2, and that demand remains as per 2008–2012 data

Performance indicator	'Pre-drought' cycle	New water programme
Urban water efficiency (kL/capita/year)	77	66
Water supply internalisation	13%	25%
Hydrological performance *Runoff*	1.5	1.5
Hydrological performance *Recharge*	0.83	0.84

Efficiency

Urban water efficiency is an indicator of the environmental water use of the urban system, expressed as a rate of environmental water withdrawal per inhabitant per year (kL/capita/year).[36] The pre-drought per capita water withdrawal from the environment amounts to 77 kL/capita/year (210 L/capita/day). The environmental water demand calculated by Renouf et al.[36] for Australian cities was relatively higher (between 92 and 182 kL/capita/year). With the proposed New Water Programme, per capita environmental water withdrawal decreases to 66 kL/capita/year (183 L/capita/day), reflecting the internalisation of water sources via direct and indirect reuse. When compared to the results of Australian cities, the City is seemingly water efficient. However, this performance indicator is misleading in this context as it assumes that per capita water usage is equal across the population. We posit that this 'efficiency' reflects rather the vast disparities in access to and use of water in Cape Town.[29,32,49] The 2011 Census estimates reported that 96.6% of households in Cape Town have piped access to water (via a public tap) within 200 m from home, and 87% of households have piped water within their dwelling or yard.[32,50] However, this access varies considerably across the City[51] – for example, only 50% of households in Site C Khayelitsha have piped water inside their dwelling or yard[50]. Per capita usage in areas where access to water is solely via a public tap will be considerably lower than in more affluent areas.

The water reuse scheme, which is intended to increase from 18 GL/year to 43 GL/year, is one approach to improving the efficiency of the urban system but still remains a small contribution (~12%) to water supply. Representing urban water efficiency spatially would highlight the disparities in efficiency (i.e. high-end and low-end users) that a bulk (average) quantification cannot. By disaggregating water efficiencies spatially, urban planners and decision-makers would be better informed as to where to focus policy and funding.

Supply internalisation

The New Water Programme is designed to augment supply capacity into the City. Most of the designated augmented supplies come from outside the urban boundary as desalination, surface supply (Voëlvlei Dam) or TMG aquifer. Currently, supply internalisation is 13%, where internal sources comprise groundwater (8%) and recycled water (5%). With the New Water Programme, supply internalisation could increase up to 25%, assuming that water reuse schemes as well as MAR (including groundwater abstraction) contribute their full capacity (see Table 2).

Further, improving supply internalisation would entail adopting a fit-for-purpose approach[52], by diversifying supply and reducing reliance on external sources of water for potable use. The contributions of greywater systems, rainwater harvesting, larger-scale stormwater harvesting and the use of groundwater for hospitals/irrigation of school fields and parks have not been accounted for here due to a paucity of data. As they become more significant and more established components of the system, they will merit inclusion in further studies.

Hydrological performance

The hydrological performance is an indicator of the degree to which natural hydrological flows have increased or decreased relative to pre-urbanised flows[36], and is a ratio of post- to pre-urbanised annual flows. A ratio of greater than 1 means that the magnitude of the annual flow is larger than that of the pre-developed landscape, and a ratio of less than 1 means it is smaller. Currently, runoff is at a ratio of 1.5 and recharge at 0.83; effectively, runoff is greater than for pre-developed landscapes and recharge is smaller due to the increase in impervious services. With the New Water Programme, runoff does not change, but recharge increases marginally to 0.84, reflecting the volumes of water redirected from WWTW effluent to MAR (Figure 3). The sensitivity of hydrological performance of recharge to evapotranspiration was assessed by varying evapotranspiration by an arbitrary ±10% (Supplementary table 7). Results show a mean hydrological performance recharge of 0.83 (±0.02), indicating that this indicator is not sensitive to uncertainty in the evapotranspiration data.

Practicality of using alternative sources

Assessing the major outputs of water from the system, wastewater could theoretically replace centralised supply to the City by 63%, and stormwater runoff by 189%, which would not only improve hydrological performance of the City, but supply internalisation and diversification as well. These figures represent a theoretical upper limit of a closed-loop system and assume no wastewater is discharged into rivers and all stormwater is captured, which is not feasible, or even desirable (e.g. environmental flow requirements[53]). Utilising such resources is challenging; the storage and purification of alternative water sources would require a significant shift in management practices and a commitment to a holistic approach to urban water management.[16] However, there are many examples where MAR has been used as a method for storing and treating urban stormwater[54-56] or recycled wastewater[57].

In Cape Town, MAR has been in successful operation in the peri-urban area of Atlantis since the 1980s.[44,54] This indirect recycling of treated domestic wastewater and stormwater has proven an effective water conservation measure[54] at a small scale (equating to ~1% of total bulk water supply). Modelling studies by Mauck[58] have shown that MAR on the Cape Flats, as part of a larger water sensitive urban design strategy, can also mitigate the migration of pollution plumes from urban sources. Although there are several risks of implementing MAR in an urban landscape (specifically localised flooding and poor quality water), these can be significantly mitigated via the integration of (artificial) wetlands and detention ponds that act to purify water[59] and, in some cases, recharge the aquifer. As wetlands in the area are hydrologically diverse in character[60], further research is required to assess how feasible, from a hydro-ecological perspective, integrating wetlands into stormwater treatment could be at such a spatial scale. An estimated capacity of ~13 GL can be stored within the city-wide network of stormwater detention ponds[43] and the potential to address the need for storage by using real-time control techniques is currently under investigation[61]. Successfully making use of stormwater resources within the urban environment provides possibilities to restore wetland and vlei (shallow lake) ecosystems throughout the City to create blue and green spaces[62], in addition to mitigating against drought and flood in the form of storage. In order to fully explore the potential to store stormwater runoff in particular, the seasonal fluctuations in groundwater storage need to be included in this mass balance, as does better parameterisation of groundwater/surface water interaction in both wetland and river systems. Further still, the quality of stormwater is often a key challenge to its usability; applying the water-related nutrient efficiency indicator suggested by Renouf et al.[48] would be a useful starting point in quantifying city nutrient budgets.

Towards water sensitivity and resilience

The performance indicators used in this study do well to assess the urban water metabolism in terms of resource efficiency, reduced reliance on external sources and improved hydrological function. However, these indicators, derived in a developed country context, do not account for socio-economic disparities in access to water. Addressing development and inequity issues, characteristic in developing countries, is noted as one of the fundamental tenets of a water sensitive design[16] for the South African context[63]. Cape Town has a marked degree of inequality in terms of access to resources[30], and in particular water[29,49], where only 83% of the population has access to running water within their residential abode or backyard[32]. The implementation of water sensitive projects requires simultaneously addressing equitable access to basic services.[64] Tackling equitable service delivery to informal settlements, in particular, is an unrelentingly complex challenge for all cities in South Africa and many others of the Global South.[65] Developing an indicator of water equity that accounts for access to water services, both centralised and decentralised, is a crucial consideration in moving towards water sensitivity.

It is worth noting that while urban metabolism provides a useful framework to benchmark *sustainability* of the urban water cycle, there remains a lack of adequate metrics to quantify its *resilience*. While there are distinct differences between these terms, they are often used interchangeably in the literature[66,67], which adds challenges for the creation of adequate metrics. As recently defined by Elmqvist et al.[68], sustainability is an increase in efficiency and optimisation of resource use (including equitable access) and resilience is the capacity of a system to recover from disaster events and return to desired functions. Quantifying the urban water metabolism highlights where sustainability of the water cycle could be improved and achieved. However, its resilience – an emergent property of the complex urban system – is much harder to define and thus quantify. More challenging yet in its definition is the uncertainty associated with predictions of future climate and socio-economic systems and the many potential scenarios of what a 'New Normal'[49] may look like. Including the socio-ecological functions of the urban environment (e.g. water purification of wetlands, storage capacity or behavioural changes in water use) into the urban metabolism framework may be a start in developing resilience metrics for the urban water cycle.

Conclusion

This research offers a first systemic quantification of the various components of the urban water cycle of Cape Town, comprising both anthropogenic and hydrological variables at a steady state. Despite scarce, and low confidence, data used to derive several variables at the macroscale, the urban water metabolism framework has proven a useful tool in assessing the performance of the City in the context of becoming water sensitive. Broadly, the performance indicators highlight that water efficiency and supply internalisation improve under the proposed New Water Programme. Hydrological performance improves only marginally; increases in groundwater recharge reflect the volumes of water reclaimed from treated effluent for MAR. No improvement to hydrological performance regarding runoff is achieved and highlights the need to reduce runoff and increase recharge further within the urban landscape. Water reuse and MAR are proposed interventions that strongly drive improved performance, and yet their contributions remain small by comparison with externally sourced water. This reinforces the benefits of 'closing the loop' in the City's commitment to becoming water sensitive, notwithstanding the likely future of static or decreasing water inputs.

The UWMF performance indicator framework is shown to be a valuable tool in assessing the impact of management decisions and interventions on the urban water cycle of the City and its commitment to becoming water sensitive. However, including an indicator to assess water equity is critical if the framework is to be used in a developing country context and will enhance the applicability of this approach in more cities of the Global South. Finally, while the framework does well to assess the sustainability of the urban water cycle, measuring its resilience will require integrating metrics that can adequately represent more facets of the society–nature interaction of a complex adaptive urban system.

Acknowledgements

We acknowledge the City of Cape Town for the water consumption data provided and are grateful to several anonymous reviewers from the Department of Environmental and Geographical Sciences at UCT and Umvoto for helpful comments on early drafts of the manuscript. All opinions and findings are that of the authors and do not necessarily reflect the opinions of University of Cape Town, Stellenbosch University, Umvoto Africa and/or City of Cape Town.

Competing interests

J.F.A. has no conflict of interest. T.F. and R.H. are employees of Umvoto Africa, a Cape Town based geohydrology consultancy that was contracted to undertake groundwater studies for the City of Cape.

Authors' contributions

J.F.A. conceptualised the study and methodology, collated the different data requirements, conducted the spatial data analyses and budget analysis, created the figures and wrote the manuscript. T.F. conceptualised the study, conducted the spatial analyses, created the boundary map and revised the manuscript. R.H. conceptualised the study, conducted the budget analyses and revised the manuscript.

References

1. Padowski JC, Gorelick SM. Global analysis of urban surface water supply vulnerability. Environ Res Lett. 2014;9(10), Art. #104004. https://doi.org/10.1088/1748-9326/9/10/104004

2. Richter BD, Abell D, Bacha E, Brauman K, Calos S, Cohn A, et al. Tapped out: How can cities secure their water future? Water Policy. 2013;15(3):335–363. https://doi.org/10.2166/wp.2013.105

3. Pincetl S, Porse E, Mika KB, Litvak E, Manago KF, Hogue TS, et al. Adapting urban water systems to manage scarcity in the 21st century: The case of Los Angeles. Environ Manag. 2019;63(3):293–308. https://doi.org/10.1007/s00267-018-1118-2

4. Flörke M, Schneider C, McDonald RI. Water competition between cities and agriculture driven by climate change and urban growth. Nat Sustain. 2018;1(1):51–58. https://doi.org/10.1038/s41893-017-0006-8

5. Goldman M, Narayan D. Water crisis through the analytic of urban transformation: An analysis of Bangalore's hydrosocial regimes. Water Int. 2019;44(2):95–114. https://doi.org/10.1080/02508060.2019.1578078

6. Shepherd TG, Boyd E, Calel RA, Chapman SC, Dessai S, Dima-West IM, et al. Storylines: An alternative approach to representing uncertainty in physical aspects of climate change. Clim Change. 2018;151(3–4):555–571. https://doi.org/10.1007/s10584-018-2317-9

7. Wolski P. How severe is Cape Town's 'Day Zero' drought? Significance. 2018;15(2):24–27. https://doi.org/10.1111/j.1740-9713.2018.01127.x

8. Sousa PM, Blamey RC, Reason CJC, Ramos AM, Trigo RM. The 'Day Zero' Cape Town drought and the poleward migration of moisture corridors. Environ Res Lett. 2018;13(12):124025. https://doi.org/10.1088/1748-9326/aaebc7

9. Otto FEL, Wolski P, Lehner F, Tebaldi C, Van Oldenborgh GJ, Hogesteeger S, et al. Anthropogenic influence on the drivers of the Western Cape drought 2015–2017. Environ Res Lett. 2018;13(12), Art. #124010. https://doi.org/10.1088/1748-9326/aae9f9

10. Feng S, Fu Q. Expansion of global drylands under a warming climate. Atmos Chem Phys. 2013;13(19):10081–10094. https://doi.org/10.5194/acp-13-10081-2013

11. Lehner F, Coats S, Stocker TF, Pendergrass AG, Sanderson BM, Raible CC, et al. Projected drought risk in 1.5°C and 2°C warmer climates: Drought in 1.5°C and 2°C warmer climates. Geophys Res Lett. 2017;44(14):7419–7428. https://doi.org/10.1002/2017GL074117

12. Garreaud RD, Alvarez-Garreton C, Barichivich J, Boisier JP, Christie D, Galleguillos M, et al. The 2010-2015 megadrought in central Chile: Impacts on regional hydroclimate and vegetation. Hydrol Earth Syst Sci. 2017;21(12):6307–6327. https://doi.org/10.5194/hess-21-6307-2017

13. Smith IN, McIntosh P, Ansell TJ, Reason CJC, McInnes K. Southwest Western Australian winter rainfall and its association with Indian Ocean climate variability. Int J Climatol A J R Meteorol Soc. 2000;20(15):1913–1930. https://doi.org/10.1002/1097-0088(200012)20:15<1913::AID-JOC594>3.0.CO;2-J

14. City of Cape Town. City of Cape Town Water Outlook 2018 [document on the Internet]. c2018 [cited 2020 Oct 02]. Available from: https://resource.capetown.gov.za/documentcentre/Documents/City%20research%20reports%20and%20review/Water%20Outlook%202018%20-%20Summary.pdf

15. City of Cape Town. Our shared water future: Cape Town's water strategy [document on the Internet]. c2019 [cited 2020 Oct 02]. Available from: https://resource.capetown.gov.za/documentcentre/Documents/City%20strategies,%20plans%20and%20frameworks/Cape%20Town%20Water%20Strategy.pdf

16. Wong THF, Brown RR. The water sensitive city: Principles for practice. Water Sci Technol. 2009;60(3):673–682. https://doi.org/10.2166/wst.2009.436

17. Wolman A. The metabolism of cities. Sci Am. 1965;213:179–190. https://doi.org/10.1038/scientificamerican0965-178

18. Fischer-Kowalski M. Society's metabolism: The intellectual history of materials flow analysis: Part I: 1860–1970. J Ind Ecol. 1998;2(1):61–78. https://doi.org/10.1162/jiec.1998.2.1.61

19. Kennedy C, Pincetl S, Bunje P. The study of urban metabolism and its applications to urban planning and design. Environ Pollut. 2011;159(8–9):1965–1973. https://doi.org/10.1016/j.envpol.2010.10.022

20. Niza S, Rosado L, Ferrão P. Urban metabolism: Methodological advances in urban material flow accounting based on the Lisbon case study. J Ind Ecol. 2009;13(3):384–405. https://doi.org/10.1111/j.1530-9290.2009.00130.x

21. Kenway S, Gregory A, McMahon J. Urban water mass balance analysis. J Ind Ecol. 2011;15(5):693–706. https://doi.org/10.1111/j.1530-9290.2011.00357.x

22. Beck L, Brown RR, Chesterfield C, Dunn G, De Haan F, Lloyd S, et al. Beyond benchmarking: A water sensitive city. Paper presented at: OzWater'16; 2016 May 10–12; Melbourne, Australia.

23. Paul R, Kenway S, McIntosh B, Mukheibir P. Urban metabolism of Bangalore City: A water mass balance analysis. J Ind Ecol. 2018;22(6):1413–1424. https://doi.org/10.1111/jiec.12705

24. Renouf M, Kenway SJ. Evaluation approaches for advancing urban water goals. J Ind Ecol. 2017;21(4):995–1009. https://doi.org/10.1111/jiec.12456

25. Arcadis. Sustainable Cities Water Index [document on the Internet]. c2016 [cited 2020 Oct 02]. Available from: https://images.arcadis.com/media/0/6/6/%7B06687980-3179-47AD-89FD-F6AFA76EBB73%7DSustainable%20Cities%20Index%202016%20Global%20Web.pdf

26. EIU. Asian Green City Index: Assessing the environmental performance of Asia's major cities. Munich: Economist Intelligence Unit; 2011.

27. Van Leeuwen CJ, Koop SHA, Sjerps RMA. City blueprints: Baseline assessments of water management and climate change in 45 cities. Environ Dev Sustain. 2016;18(4):1113–1128. https://doi.org/10.1007/s10668-015-9691-5

28. Madonsela B, Koop S, Van Leeuwen K, Carden K. Evaluation of water governance processes required to transition towards water sensitive urban design-an indicator assessment approach for the City of Cape Town. Water. 2019;11, Art. #292. https://doi.org/10.3390/w11020292

29. Gasson B. The ecological footprint of Cape Town: Unsustainable resource use and planning implications. Paper presented at: The National Conference of South African Planning Institution; 2002 September 19–20; Durban, South Africa..

30. Hoekman P, Von Blottnitz H. Cape Town's metabolism: Insights from a material flow analysis. J Ind Ecol. 2017;21(5):pp1237-1249. https://doi.org/10.1111/jiec.12508

31. Ahjum F, Stewart TJ. A systems approach to urban water services in the context of integrated energy and water planning: A City of Cape Town case study. J Energy South Afr. 2014;25(4):59–70. https://doi.org/10.17159/2413-3051/2014/v25i4a2239

32. Currie PK, Musango JK, May ND. Urban metabolism: A review with reference to Cape Town. Cities. 2017;70:91–110. https://doi.org/10.1016/j.cities.2017.06.005

33. Decker EH, Elliott S, Smith FA, Blake DR, Rowland FS. Energy and material flow through the marine environment. Annu Rev Energy Environ. 2000;25(1):685–740. https://doi.org/10.1146/annurev.energy.25.1.685

34. Currie P, Lay-Sleeper E, Fernandez JE, Kim J, Musango JK. Towards urban resource flow estimates in data scarce environments: The case of African cities. J Environ Prot (Irvine, Calif). 2015;6(9):1066–1083. https://doi.org/10.4236/jep.2015.69094

35. Vrebos D, Staes J, Vandenbroucke T, D'Haeyer T, Johnston R, Muhumuza M, Kasabeke C, Meire P. Mapping ecosystem service flows with land cover scoring maps for data-scarce regions. Ecosyst Serv. 2015;13:28–40. https://doi.org/10.1016/j.ecoser.2014.11.005

36. Renouf MA, Kenway SJ, Lam KL, Weber T, Roux E, Serrao-Neumann S, et al. Understanding urban water performance at the city-region scale using an urban water metabolism evaluation framework. Water Res. 2018;137:395–406. https://doi.org/10.1016/j.watres.2018.01.070

37. City of Cape Town. City of Cape Town draft water strategy [document on the Internet]. c2019 [cited 2020 Oct 02]. Available from: https://resource.capetown.gov.za/documentcentre/Documents/City%20strategies,%20plans%20and%20frameworks/Cape%20Town%20Water%20Strategy.pdf

38. Bailey AK, Pitman WV. Water resources of South Africa (WR2012): Book of maps. Version 1. WRC report no. TT 684/16. Pretoria: Water Research Commisson; 2016.

39. Schulze RE, Maharaj M, Moult N. South African atlas of climatology and agrohydrology. WRC report 1489/1/06 Section 13.3. Pretoria: Water Research Commission; 2007.

40. Allen RG, Pereira LS, Raes D, Smith M. Crop evapotranspiration: Guidelines for computing crop water requirements. FAO irrigation and drainage paper 56. Rome: FAO; 1998.

41. Kruger E, editor. Drainage manual: Application guide. 6th ed. Pretoria: The South African National Roads Agency, SOC Ltd.; 2013.

42. Lupton RC, Allwood JM. Hybrid Sankey diagrams: Visual analysis of multidimensional data for understanding resource use. Resour Conserv Recycl. 2017;124:141–151. https://doi.org/10.1016/j.resconrec.2017.05.002

43. Okedi J. The prospects for stormwater harvesting in Cape Town, South Africa using the Zeekoe Catchment as a case study [PhD thesis]. Cape Town: University of Cape Town; 2019.

44. Jovanovic N, Bugan RDH, Tredoux G, Israel S, Bishop R, Marinus V. Hydrogeological modelling of the Atlantis aquifer for management support to the Atlantis Water Supply Scheme. Water SA. 2017;43(1):122–138. https://doi.org/10.4314/wsa.v43i1.15

45. Grimmond CSB, Oke TR. An evapotranspiration-interception model for urban areas. Water Resour Res. 1991;27(7):1739–1755. https://doi.org/10.1029/91WR00557

46. Hobbie SE, Finlay JC, Janke BD, Nidzgorski DA, Millet DB, Baker LA. Contrasting nitrogen and phosphorus budgets in urban watersheds and implications for managing urban water pollution. Proc Natl Acad Sci USA. 2017;114(16):4177–4182. https://doi.org/10.1073/pnas.1618536114

47. Minnesota Pollution Control Agency. The simple method for estimating phosphorus export. In: Minnesota Stormwater Manual [webpage on the Internet]. c2017 [updated 2020 Apr 08; cited 2020 Oct 02]. Available from: https://stormwater.pca.state.mn.us/index.php?title=The_Simple_Method_for_estimating_phosphorus_export

48. Renouf M, Serrao-Neumann S, Kenway SJ, Morgan E, Low Choy D. Urban water metabolism indicators derived from a water mass balance – Bridging the gap between visions and performance assessment of urban water resource management. Water Res. 2017;122:669–677. https://doi.org/10.1016/j.watres.2017.05.060

49. Enqvist JP, Ziervogel G. Water governance and justice in Cape Town: An overview. Wiley Interdiscip Rev Water. 2019; e1354. https://doi.org/10.1002/wat2.1354

50. City of Cape Town. Trends and change – 10 years: Census 2001 – Census 2011. Cape Town: City of Cape Town; 2012.

51. Currie PK, Musango JK, May ND. Urban metabolism: A review with reference to Cape Town. Cities. 2017;70(June):91–110. https://doi.org/10.1016/j.cities.2017.06.005

52. Kenway SJ, Lam KL, Sochacka B, Renouf MA. Integrated urban water systems. In: Newton P, Prasad D, Sproul A, White S, editors. Decarbonising the built environment: Charting the transition. Singapore: Springer Singapore; 2019. p. 287–304. https://doi.org/10.1007/978-981-13-7940-6_15

53. Gerten D, Hoff H, Rockström J, Jägermeyr J, Kummu M, Pastor AV. Towards a revised planetary boundary for consumptive freshwater use: Role of environmental flow requirements. Curr Opin Environ Sustain. 2013;5(6):551–558. https://doi.org/10.1016/j.cosust.2013.11.001

54. Bugan RDH, Jovanovic N, Israel S, Tredoux G, Genthe B, Steyn M, et al. Four decades of water recycling in Atlantis (Western Cape, South Africa): Past, present and future. Water SA. 2016;42(4):577–594. https://doi.org/10.4314/wsa.v42i4.08

55. Clark R, Gonzalez D, Dillon P, Charles S, Cresswell D, Naumann B. Reliability of water supply from stormwater harvesting and managed aquifer recharge with a brackish aquifer in an urbanising catchment and changing climate. Environ Model Softw. 2015;72:117–125. https://doi.org/10.1016/j.envsoft.2015.07.009

56. Radcliffe JC, Page D, Naumann B, Dillon P. Fifty years of water sensitive urban design, Salisbury, South Australia. Front Environ Sci Eng China. 2017;11(4):7. https://doi.org/10.1007/s11783-017-0937-3

57. Bekele E, Toze S, Patterson B, Higginson S. Managed aquifer recharge of treated wastewater: Water quality changes resulting from infiltration through the vadose zone. Water Res. 2011;45(17):5764–5772. https://doi.org/10.1016/j.watres.2011.08.058

58. Mauck BA. The capacity of the Cape Flats Aquifer and its role in water sensitive urban design in Cape Town [PhD thesis]. Cape Town: University of Cape Town; 2017.

59. Verhoeven JTA, Meuleman AFM. Wetlands for wastewater treatment: Opportunities and limitations. Ecol Eng. 1999;12:5–12. https://doi.org/10.1016/S0925-8574(98)00050-0

60. Brown C, Magoba R. Rivers and wetlands of Cape Town: Caring for our rich aquatic heritage. WRC report no. TT 376/08. Pretoria: Water Research Commission; 2009.

61. Okedi J, Armitage NP. Benefits of real time control for catchment scale stormwater harvesting in Cape Town, South Africa. In: Mannina G, editor. New trends in urban drainage modelling UDM 2018 green energy and technology. Cham: Springer; 2018. p. 587–591. https://doi.org/10.1007/978-3-319-99867-1_101

62. Brodnik C, Holden J, Marino R, Wright A, Copa V, Rogers B, et al. Jumping to the top: Catalysts for leapfrogging to a water sensitive city. IOP Conf Ser Earth Environ Sci. 2018;179(1):12034. https://doi.org/10.1088/1755-1315/179/1/012034

63. Armitage N, Fisher-Jeffes L, Carden K, Winter K, Naidoo V, Spiegel A, et al. Water Sensitive Urban Design (WSUD) for South Africa: Framework and guidelines. WRC report no. TT 588/14. Pretoria: Water Research Commision; 2014.

64. Fisher-Jeffes L, Carden K, Armitage N, Borwa A. A water sensitive urban design framework for South Africa. T Reg Plan. 2017;71(1):1–10.

65. Seeliger L, Turok I. Averting a downward spiral: Building resilience in informal urban settlements through adaptive governance. Environ Urban. 2014;26(1):184–199. https://doi.org/10.1177/0956247813516240

66. Elmqvist T. Development: Sustainability and resilience differ. Nature. 2017;546:352. https://doi.org/10.1038/546352d

67. Johannessen Å, Wamsler C. What does resilience mean for urban water services? Ecol Soc. 2017;22(1):1. https://doi.org/10.5751/ES-08870-220101

68. Elmqvist T, Andersson E, Frantzeskaki N, McPhearson T, Olsson P, Gaffney O, et al. Sustainability and resilience for transformation in the urban century. Nat Sustain. 2019;2(4):267–273. https://doi.org/10.1038/s41893-019-0250-1

Monitoring changes in South Africa’s surface water extent for reporting Sustainable Development Goal sub-indicator 6.6.1.a

AUTHOR:
Heidi van Deventer[1,2]

AFFILIATIONS:
[1]Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Department of Geography, GeoInformatics and Meteorology, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Heidi van Deventer

EMAIL:
HvDeventer@csir.co.za

DATES:
Received: 19 Aug. 2020
Revised: 02 Feb. 2021
Accepted: 03 Feb. 2021
Published: 28 May 2021

HOW TO CITE:
Van Deventer H. Monitoring changes in South Africa’s surface water extent for reporting Sustainable Development Goal sub-indicator 6.6.1.a. S Afr J Sci. 2021;117(5/6), Art. #8806. https://doi.org/10.17159/sajs.2021/8806

ARTICLE INCLUDES:
☒ Peer review
☐ Supplementary material

DATA AVAILABILITY:
☐ Open data set
☐ All data included
☒ On request from author(s)
☐ Not available
☐ Not applicable

EDITOR:
Yali Woyessa

KEYWORDS:
SDG 6, wetlands, remote sensing, lacustrine wetlands, hydroperiod

FUNDING:
Council for Scientific and Industrial Research (South Africa)

For the first progress reporting on the Sustainable Development Goal sub-indicator 6.6.1a in 2020, the South African and global statistics related to wetlands were compared. Firstly, in terms of the total wetland extent, the South African National Wetland Map version 5 (NWM5) represented 87% more inland, surface aquatic ecosystems than the Global Surface Water (GSW) product. More than half of the lacustrine systems and none of the palustrine and arid systems in NWM5 are represented in the GSW layer. Secondly, in terms of changes in the extent of wetlands, both the global and South African statistics showed a decreasing trend in the spatial extent of surface aquatic ecosystems in South Africa. These trends should be further investigated against systematic assessments of decadal drought periods. The hydroperiod information (permanent, seasonal and ephemeral inundation periods) of the GSW products show that South African lacustrine wetlands do not have a single dominant class (≥70% of the extent of a polygon) of inundation, but consist of a mosaic of these classes.

Significance:

- The South African National Wetlands Map version 5 represents 87% more of the extent of lacustrine, palustrine and arid wetlands than the Global Surface Water products that are used for progress reporting on the Sustainable Development Goal sub-indicator 6.6.1a.
- South African and global statistics suggest a decline in the extent of lacustrine wetlands, although a systematic comparison with decadal drought periods is required to confirm these trends.
- South African lacustrine wetlands consist of a mosaic of hydroperiod classes (permanent, seasonal and ephemeral inundation periods) with no individual class dominating (≥70% of the extent of) wetlands polygons.

Aquatic ecosystems play a pivotal role in water provision, but globally these systems are impacted and at the brink of collapse.[1] Monitoring ecosystem changes in surface aquatic systems using remote sensing is critical to facilitate urgent intervention strategies to safeguard these vulnerable and threatened ecosystems. Space-borne satellite images now enable monitoring at scales ranging from landscape to global.

Seventeen Sustainable Development Goals (SDGs) were adopted in 2015 by member states of the United Nations and 2020 was the first year of reporting progress on the targets identified. In terms of aquatic ecosystems, SDG target 6.6 states: ‘By 2020, protect and restore water-related ecosystems, including mountains, forests, wetlands, rivers, aquifers and lakes’. Global statistics derived from the Global Surface Water (GSW) products[2] were released to assist each country in reporting sub-indicator SDG 6.6.1a, which pertains to the spatial extent of surface aquatic ecosystems. In 2019, South Africa released the latest National Biodiversity Assessment of 2018, which included updates to the National Wetland Map version 5 (NWM5) and artificial wetlands spatial data layers.[3] These data sets provide a new opportunity for comparing the spatial extent of surface aquatic ecosystems between the different data sets. In addition, other GSW products are also considered for characterising the hydroperiod of South African wetlands.

SDG target 6.6. has only one indicator (6.6.1): ‘Change in the extent of water-related ecosystems over time’.[4] Aquatic ecosystems included for this target are ‘wetlands, lakes, estuaries, artificial ecosystems (dams), rivers and aquifers’. Indicator 6.6.1 aggregates three of the sub-indicators, namely change in (a) spatial extent, (b) water quality and (c) water quantity, while a fourth sub-indicator, related to the ecological condition of ecosystems, is reported separately. Two tiers of reporting are facilitated. Tier 1 uses globally available data from remote sensing derivatives and countries can validate these quantities against their own methods and/or data sets. Tier 2 allows for countries to report their own statistics for the sub-indicators of SDG 6.6.1.

To facilitate Tier 1 reporting for the spatial extent of aquatic ecosystems (sub-indicator 6.6.1.a), the United Nations extracted data from recent GSW products which quantify the total extent and changes in the spatial extent of surface aquatic ecosystems. The GSW products were derived from Landsat images since 16 March 1984, and more recently Sentinel-2 images after 10 October 2015.[2] The Freshwater Ecosystems Explorer application (www.sdg661.app) enables countries to view or download data for reporting.

The average extent of inundation between the years 2000 and 2004 was taken as the reference against which changes in the extent of surface water is measured for 6.6.1.a. These years were mainly drought free, i.e. they did not include a 60-month or 24-month decadal drought, with the exception of 2004, in which an intense summer drought affected 26% of the summer-rainfall region that covers about 90% of the areal extent of the country.[5]

The Freshwater Ecosystem Explorer statistics report 6144 km^2 of wetlands for South Africa (Table 1). The statistics show that, since the reference period, permanent water reduced by nearly 20% while seasonal water increased by 32%. In addition, the minimum water extent of reservoirs (artificial) decreased by 25%, whereas the maximum water extent of reservoirs decreased by only 11%. These results suggest that South Africa may not yet have recovered from the 2015/2016 drought. However, no systematic assessment of the severity and geographic extent of this drought is yet available for the whole of South Africa, which would be necessary for comparing these to other decadal droughts described in Malherbe et al.[5]

The total spatial extent of surface aquatic ecosystems as reported for the SDG 6.6.1.a sub-indicator was compared to three data sets of South Africa (Table 1). Three research groups have mapped surface aquatic systems using various data sources, for different purposes, categories and intended uses.[3,6,7] Using NWM5 and the artificial wetlands layer yielded the largest spatial extent of wetlands, with a total of nearly 44 000 km^2.[3] These inland wetlands, and the extent of some of the rivers, include permanently and seasonally inundated (lacustrine), vegetated (palustrine), arid (with no dominant inundation or vegetation cover) and artificial systems/ water bodies. Van Deventer et al.[3] report extensive omission errors in aquatic ecosystems, with 76% of the national extent mapped at a low confidence. The statistics used for reporting on the SDG 6.6.1.a sub-indicator therefore underreport the spatial extent of inland wetlands by at least 87% of those mapped in NWM5 and the artificial wetland layers. Note that the GSW products do not include palustrine and arid systems at all, while these are included NWM5, in addition to the lacustrine systems.

The South African 2013/2014 land cover data[6] have been used to track changes against a historical reference point of 1990 (Table 2)[8]. The year 1990 is a suitable reference, because it was just before the 1991–1995 decadal drought.[5] In the comparative study, three categories of water are reported to have reduced between 1990 and 2014, including a loss of nearly 1% of artificial water bodies, 0.5% mining water and 24% natural water (Table 2). The report notes, however, that the data derived from remote sensing products may include some shadows erroneously classified as wetlands. Despite the errors noted for both the GSW and land cover products, both studies observed a decrease in the spatial extent of surface aquatic ecosystems over the past 20–30 years. These general trends should be further investigated through a systematic assessment of the decadal drought cycles[5], and for specific regions to evaluate whether they relate to the natural variation in rainfall or to climate change trends associated with increasing temperatures and evapotranspiration.

Table 1: Descriptive statistics related to the spatial extent of surface wetland types available on the Freshwater Ecosystems Explorer for reporting on Sustainable Development Goal sub-indicator 6.6.1.a, in comparison to three South African data sets

Wetland type (down) / Source (across)	SDG 6.6.1.a statistics of 2020 (km^2)[2]	Land cover (2013/2014)[6]	DRDLR:NGI (2016)[7]	SAIIAE and NWM5[3]
Total extent of wetlands	6 144.3 km^2	14 738.5 km^2	28 099.3 km^2	43 804.3 km^2
Permanently inundated (lakes and rivers)	-209 km^2 (-19.7%)	Water permanent 3 919.6 km^2 (includes natural and artificial systems)	2 309.3 km^2 including perennial water bodies and the extent of some rivers	Lacustrine wetlands: 2 787.2 km^2 (11% of inland wetlands in South Africa)
Seasonally inundated (lakes and rivers)	+190 km^2 (+31.8%)	Water seasonal 631.5 km^2 (includes natural and artificial systems)	3 765.4 km^2 (including non-perennial water bodies and the extent of some rivers)	Palustrine wetlands: 14 479.3 km^2 (55% of inland wetlands in South Africa)
Permanently or seasonally inundated systems	n/a	Wetlands 10 187.5 km^2 (includes natural and artificial systems)	n/a	Rivers: 11 462.3 km^2
Arid (not predominantly inundated or vegetated)	n/a	Not distinguished from other classes	21 917.3 km^2	Arid: 9 091.6 km^2 (34% of inland wetlands in South Africa)
Artificial	Minimum water extent: -379 km^2 (-24.9%) Maximum water extent: -185 km^2 (-11.5%)	Not distinguished from other classes.	107.3 km^2 of dams, reservoirs and other structures mapped	5 983.9 km^2

DRDLR:NGI, Department of Rural Development and Land Reform: National GeoInformation; NWM5, National Wetland Map version 5; SAIIAE, South African Inventory of Inland Aquatic Ecosystems; SDG, Sustainable Development Goals

Table 2: Changes in the spatial extent of surface aquatic ecosystems between 1990 and 2014[8]

Category	Year 1990 (km^2)	Year 2014 (km^2)	Change (km^2)	Percentage change (%)
Artificial water bodies	3 062.3	3 006.2	-56.0	-0.9
Mining water	121.6	100.9	-20.8	-0.4
Natural water (including shadows)	2 837.9	1 390.6	-1 447.3	-24.0
Total	**6 021.8**	**4 497.7**	**-1 524.1**	**-25.3**

Table 3: Global Surface Water products[2] for two selected study areas in South Africa. The outlines of the National Wetland Map (NWM5[3]) polygons are shown in grey. Legend files provided by the Global Surface Water Explorer were used as is for colours and categories displayed in the legends and maps.

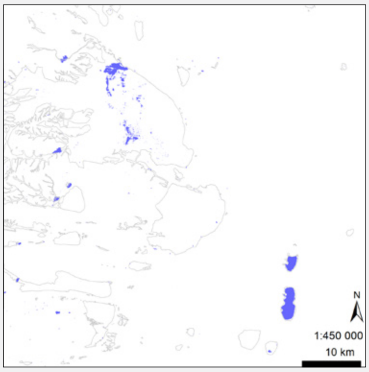
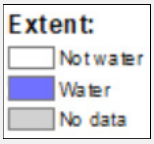
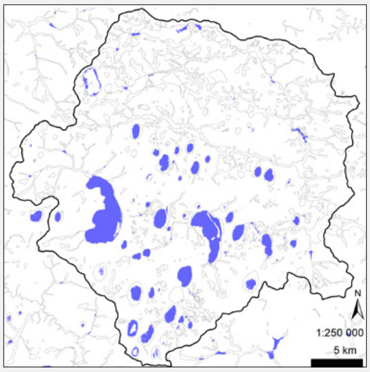
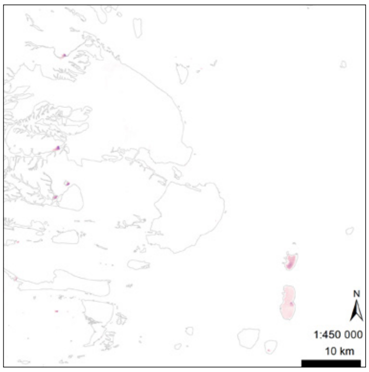
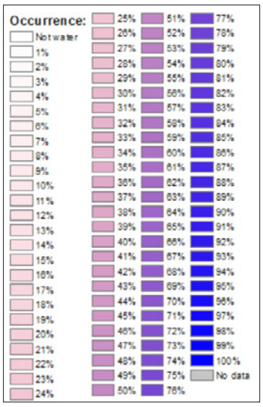
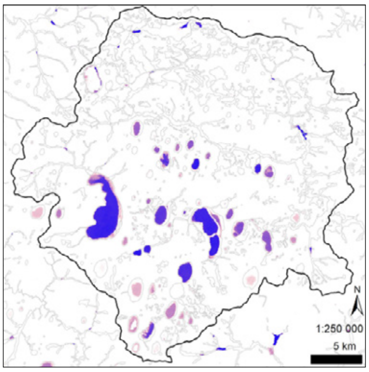
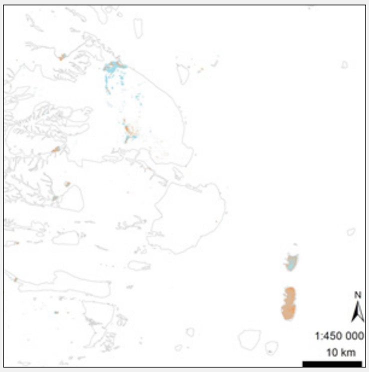
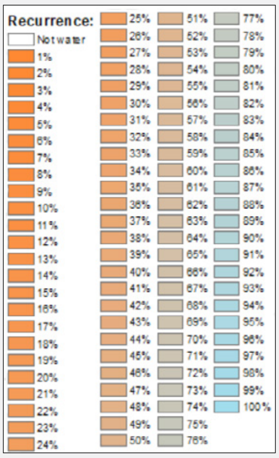
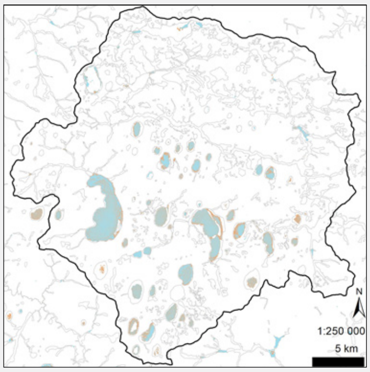
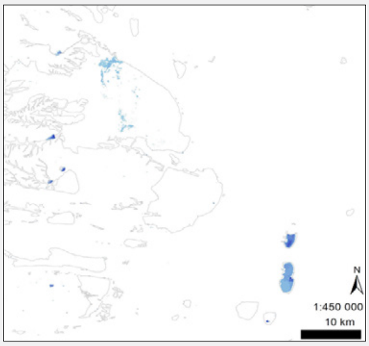
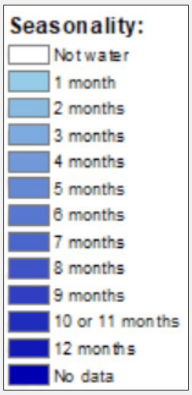
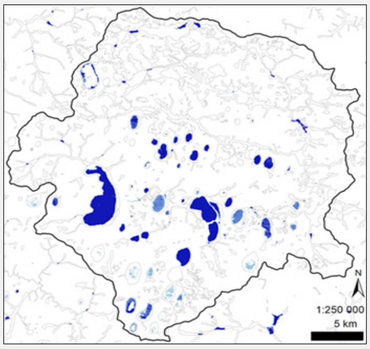
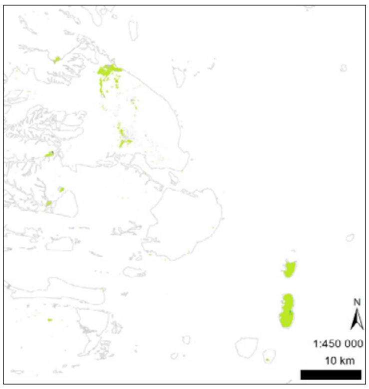
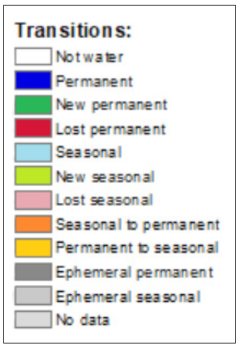
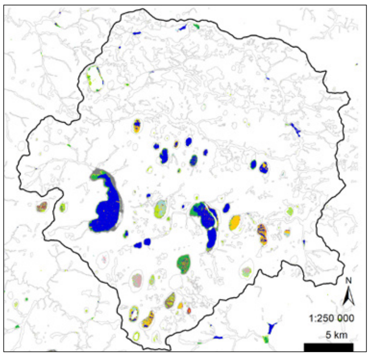
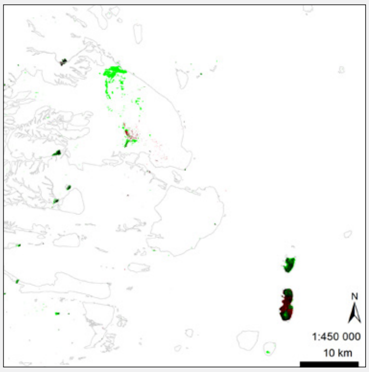
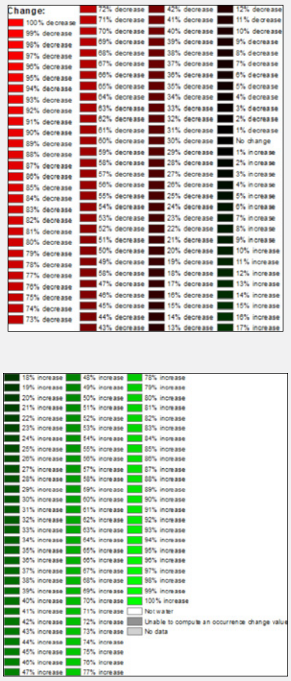
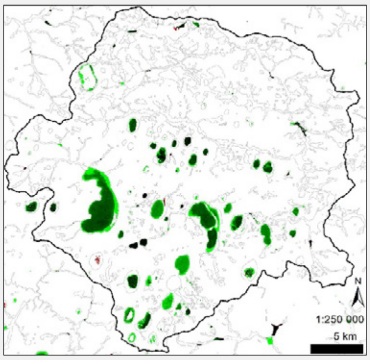
Product (down) / site (across)	Hakskeenpan, Northern Cape Province	Legend	The Mpumalanga Lakes District (Quaternary catchment W55A), Mpumalanga Province
Extent: maximum water extent (all locations ever mapped as water)			
Occurrence: frequency with which water was present on the surface since March 1984		 Occurrence: Not water 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12% 13% 14% 15% 16% 17% 18% 19% 20% 21% 22% 23% 24% 25% 26% 27% 28% 29% 30% 31% 32% 33% 34% 35% 36% 37% 38% 39% 40% 41% 42% 43% 44% 45% 46% 47% 48% 49% 50% 51% 52% 53% 54% 55% 56% 57% 58% 59% 60% 61% 62% 63% 64% 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77% 78% 79% 80% 81% 82% 83% 84% 85% 86% 87% 88% 89% 90% 91% 92% 93% 94% 95% 96% 97% 98% 99% 100% No data	
Recurrence: how frequently water returned from one year to another (expressed as a percentage), or a measurement of the degree of interannual variability in the presence of water		 Recurrence: Not water 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12% 13% 14% 15% 16% 17% 18% 19% 20% 21% 22% 23% 24% 25% 26% 27% 28% 29% 30% 31% 32% 33% 34% 35% 36% 37% 38% 39% 40% 41% 42% 43% 44% 45% 46% 47% 48% 49% 50% 51% 52% 53% 54% 55% 56% 57% 58% 59% 60% 61% 62% 63% 64% 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77% 78% 79% 80% 81% 82% 83% 84% 85% 86% 87% 88% 89% 90% 91% 92% 93% 94% 95% 96% 97% 98% 99% 100%	

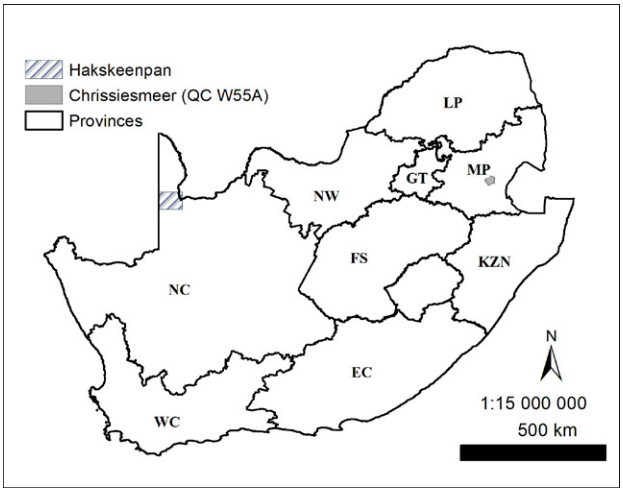
Table 3 continued

<table>
<tr><th>Product (down) / site (across)</th><th>Hakskeenpan, Northern Cape Province</th><th>Legend</th><th>The Mpumalanga Lakes District (Quaternary catchment W55A), Mpumalanga Province</th></tr>
<tr><td><b>Seasonality</b> describes the intra-annual distribution of water</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td><b>Transition</b> between permanent water, seasonal water and land classes can be determined between any two years of observation; transitions between the first and last year of observation*</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td><b>Change</b>
in water occurrence intensity between two epochs (16 March 1984 to 31 December 1999, and 1 January 2000 to 10 October 2015)*</td><td>
</td><td>
</td><td>
</td></tr>
</table>

*1984 was part of a 60-month decadal drought cycle[5].

Table 4: Number of polygons and percentage per type inundated, as derived from the Global Surface Water transition data layer[2] for inland wetlands and rivers mapped in National Wetland Map version 5 of South Africa[3]

Hydrogeomorphic unit (down)/ Inundation type (across)	Inundated		Not inundated		Total	
	No.	%	No.	%	No.	%
Channelled valley bottom	2 299	18.8	9 962	81.2	12 261	7.8
Depression	9 825	11.7	74 154	88.3	83 979	53.2
Flat	34	6.1	519	93.9	553	0.4
Floodplain	867	25.2	2 580	74.8	3 447	2.2
River	2 692	25.2	8 007	74.8	10 699	6.8
Seep	3 061	7.5	37 546	92.5	40 607	25.7
Unchannelled valley bottom	773	12.5	5 409	87.5	6 182	3.9
Total	**19 551**	**12.4**	**138 177**	**87.6**	**157 728**	**100.0**



EC, Eastern Cape; FS, Free State; GT, Gauteng; KZN, KwaZulu-Natal; LP, Limpopo; MP, Mpumalanga; NC, Northern Cape; NW, North West; WC, Western Cape

Figure 1: Location of the two study areas, Hakskeenpan and Chrissiesmeer, in the Northern Cape and Mpumalanga Provinces, respectively. See Table 3 for more information.

The Global Surface Water Explorer (GSWE, https://global-surface-water.appspot.com/) enables the viewing and download of several data layers (at 30 m spatial resolution) related to the spatial extent of inundation (Table 3, Figure 1). The transition data layer was used to investigate the three different hydroperiod classifications of permanently, seasonally and ephemerally inundated systems. The tiles were merged into a single raster using ArcGIS 10.6[9], and statistics were calculated (Tabulate Area) for each polygon of the inland wetlands and rivers of NWM5. In total, 19 551 polygons (12%) of the 157 728 inland wetlands and rivers have signatures of inundation according to the GSW transition data layer (Table 4). This totals 1359.1 km^2 of inland wetlands, which is about half of the 2787.2 km^2 of the lacustrine systems identified using the 2013/2014 National Land Cover data[7] for NWM5.[3] On average, the majority of inland wetlands and rivers showed that no more than 22% of their spatial extents are inundated, with the exception of depressions, which had an average of 29% of their extents inundated.

Fewer than 1% of the number of inland wetlands and river polygons in NWM5 had an extent of ≥70% of polygon in any one of the three hydroperiod classes of the GSW transition product. Consequently, a new approach for assigning hydroperiod to a single wetland unit needs to be investigated.

The author provided information to the South African Department of Water and Sanitation concerning the reference period and the underrepresentation of wetlands in the global data sets relative to NWM5, as described in this article. The Department's SDG team incorporated this information in their Response to the United Nations Environment Programme 2020 data drive.[10]

Acknowledgements

This work was funded by the Council for Scientific and Industrial Research (CSIR). I am grateful to my colleague, Steven Weerts, who provided proofreading and editing. Thank you to the external reviewers whose suggestions offered improvement to the manuscript.

Competing interests

There are no competing interests to declare.

References

1. Díaz S, Settele J, Brondízio ES, Ngo HT, Guèze M, Agard J, et al., editors. Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) summary for policymakers of the Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Bonn: IPBES; 2019. https://doi.org/10.5281/zenodo.3553579

2. Pekel J-F, Cottam A, Gorelick N, Belward AS. High-resolution mapping of global surface water and its long-term changes. Nature. 2016;540:418–422. https://doi.org/10.1038/nature20584

3. Van Deventer H, Van Niekerk L, Adams J, Dinala MK, Gangat R, Lamberth SJ, et al. National Wetland Map 5 – An improved spatial extent and representation of inland aquatic and estuarine ecosystems in South Africa. Water SA. 2020;46(1):66–79. https://doi.org/10.17159/wsa/2020.v46.i1.7887

4. United Nations (UN). Integrated Monitoring Guide for SDG 6: Step-by-step monitoring methodology for indicator 6.6.1 on water-related ecosystems Version 20 [document on the Internet]. c2017 [cited 2021 Feb 02]. Available from: http://www.unwater.org/app/uploads/2017/05/Step-by-step-methodology-6-6-1_Revision-2017-01-20_Final-1.pdf

5. Malherbe J, Dieppois B, Maluleke P, Van Staden M, Pillay DL. South African droughts and decadal variability. Nat Hazards. 2016;80(1):657–681. https://doi.org/10.1007/s11069-015-1989-y

6. GeoTerraImage Pty Ltd (GTI) 2013-2014 South African National Land-Cover Dataset. Pretoria, South Africa: GTI; 2015.

7. Department of Rural Development and Land Reform: National Geo-Information (DRDLR:NGI). Provincial geodatabases of hydrological databases exported from GeoMedia in March 2016. Cape Town: DRDLR:NGI; 2016.

8. GeoTerraImage Pty Ltd (GTI). South Africa land cover water feature splits (1990-2013/14). Data users report and meta data (Version 2). Pretoria: GTI; 2016.

9. Environmental Systems Research Institute (ESRI). ArcGIS desktop 10.6. Redlands, CA: ESRI; 1999–2017.

10. South African Department of Water and Sanitation (DWS). Sustainable Development Goal 6: 2020 data drive. South African response to the United Nations Environment Programme. Pretoria: DWS; 2020. Available from: https://www.dws.gov.za/projects/sdg/Water%20related%20Indicator%20Reports.aspx