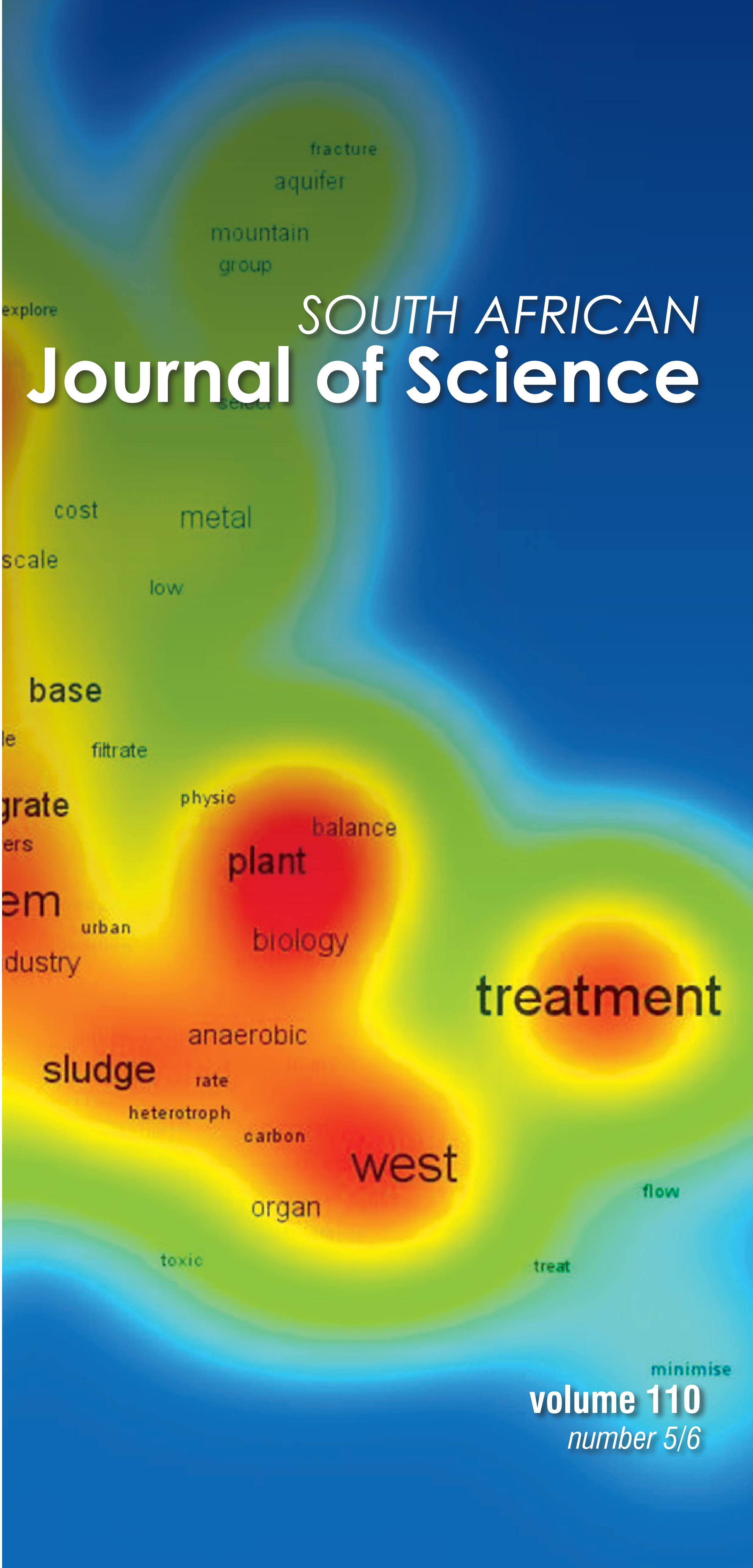


SOUTH AFRICAN Journal of Science

Alternative management strategies to curb rhino poaching

Understanding soil health in South Africa

The future of food and agriculture

Ecological consequences of climate change for freshwater ecosystems

Water research paradigm shifts in South Africa

MAY/JUNE 2014

ISSN: 0038-2353

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

volume 110
number 5/6

MAY/JUNE

ISSN: 0038-2353 (Print) | ISSN: 1996-7489 (Online)

SOUTH AFRICAN Journal of Science

volume 110
number 5/6

Leader

News and Views

Book Review

Commentary

Review Article

Johann Mouton
Centre for Research on Science and Technology, Stellenbosch University

Gansen Pillay
National Research Foundation, South Africa

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

Design and layout
SUN MeDIA Bloemfontein
T: 051 444 2552
E: admin@sunbloem.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 http://mc.manuscriptcentral.com/sajs

Subscriptions
Subscription information can be found at
www.sajs.co.za

Research Article

Dealing with ‘open access’ demons

The possibility of open-access scholarly publications started almost 24 years ago, in response to a growing demand to make research findings free and available to anyone with a computer and an Internet connection, and building on the digital developments of the 1990s. The initiators of the movement came from a wide range of prestigious research-intensive institutions and major research funders including Harvard, the Max Planck Institute, University College London, the University of Montreal and – amongst funders – the Open Society Foundations and the Wellcome Trust. They worked on open-access issues for 10 years before releasing the Budapest Declaration and Guidelines for Open Access Publishing in 2001. Sadly, the Mellon, Carnegie and Ford Foundations seem not to have been supporters at that stage.

In the ensuing 13 years, open access has, for good reason, become an increasingly desirable route to scholarly and scientific publishing. It has also become a complex field in the publishing arena – beset by a number of serious challenges. In particular, open-access journals have come under scrutiny over the past months because of the dubious charging practices and poor, or non-existent, reviewing processes of some. Open access seems to be known as much for inadequate and exploitative publishing practices as for any increase in access.

Readers of *Science* may know that an article published by that journal in October 2013 revealed some startling statistics. Earlier in 2013 a *Science* journalist and molecular biologist, John Bohannon, submitted a seriously flawed manuscript under a range of fictitious names to 304 open-access journals.[1] A number of journals failed to respond, 20% rejected the article and 61%, including some published by Elsevier and SAGE, published the article. Bohannon concluded that a large proportion of open-access journals have lax or no real peer-review systems in place.

Reactions to the *Science* article were swift and clear. *The Guardian’s Higher Education Network*[2] carried a response the very next day from Curt Rice, a professor at the University of Tromsø. Rice argued that the *Science* article demonstrated (almost) the reverse of what it had set out to do: it’s not that there are too many open-access journals that ignore proper reviewing processes, but that there are too many that set out to profit from researchers, and too few that are serious research publishing ventures.

Within a week, the *Economist* had also offered its views on ‘bad science’.[3] Arguing that while it is assumed that the peer-review system ensures that science is ‘self-correcting’, they presented examples of many experiments reported in respectable journals that could not, subsequently, be replicated. Apologists point to many benign reasons: scientists make statistical errors, peer review does take place but reviewers are harried and do not always pick up mistakes or inappropriate conclusions, and universities place more emphasis on publishing than on getting research reporting right. Benign though they may be, these reasons still serve to undermine trust in research and in the current publication system. What is worse, more subtle and clearly obvious limitations and problems with open access are not dealt with because these practices are such an obvious challenge to the industry. Publishers, big and small, have been caught up in pilot trials that have shown that their peer-review practices are inadequate.

But more was to come: in this case, for the respectable publishers Springer and the Institute of Electrical and Electronics Engineers (IEEE) – and for open-access journals in general. In February this year, a French computer scientist, Cyril Labbé, privately informed Springer and the IEEE that he had identified 16 publications by Springer and over a 100 by the IEEE that were computer-generated ‘gobbledygook’.[4] Strangely enough, Labbé’s research had been published online in a Springer journal (*Scientometrics*) in June 2012 but no attention appears to have been paid by Springer to his findings (at the time). The two publishers had subsequently no option but to withdraw over 120 papers from their subscription services after the papers were discovered to be fraudulent. Again, however, the problem was not open access itself, but a fairly cavalier approach to profit-making and a disregard for proper double-blind reviewing – or reviewing of any kind.

Currently, sites on the Internet list the names of 477 ‘predatory’ publishers and 303 ‘predatory’ stand-alone journals.[5] Predatory journals are defined extensively[6] – although the criteria may be summarised as including journals which have dubious practices that are widely considered to be the antithesis of reliable scholarly publishing. The Academy of Science of South Africa does not appear on the list of predatory publishers, nor does the SAJS appear on the list of stand-alone predatory journals.

So what does all this imply for the SAJS, which is already an open–access (and free-to-publish journal)?

Taking account of the problems and dangers exposed by Bohannon and by Labbé, but also the useful insights shared by Rice, it is clear that the SAJS has the potential to contribute substantially to the obvious need for more open-access journals that follow rigorous reviewing processes, offering content that is as reliable as these conditions can assure. This is a critical element in the process of rebuilding trust in open access.

It is true, of course, that nothing is ever free. So if the SAJS does not charge at either end of the publishing process, who does, in fact, pay? The answer is – South African taxpayers. Their annual contribution, however, is a very small portion of the cost of securing just one national key point: science set against royal sports stadiums.

References

1. Bohannan J. Who’s afraid of peer review? Science. 2013;342:60–65.
2. Rice C. Open access publishing hoax: What Science magazine got wrong [homepage on the Internet]. c2013 [cited 2014 Mar 12]. Available from: http://www.theguardian.com/higher-education-network/blog/2013/oct/04/science-hoax-peer-review-open-access
3. Trouble at the lab [homepage on the Internet]. c2013 [cited 2014 May 05]. Available from: http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble
4. Van Noorden R. Publishers withdraw more than 120 gibberish papers. Nature News [serial on the Internet]. 2014 Feb 24 [updated 2014 Feb 25; cited 2014 May 05]. Available from: http://www.nature.com/news/publishers-withdraw-more-than-120-gibberish-papers-1.14763
5. Beall J. Beall’s list [homepage on the Internet]. c2014 [cited 2014 May 05]. Available from: http://scholarlyoa.com/publishers
6. Beall J. Criteria for determining predatory open-access publishers [homepage on the Internet]. c2012 [cited 2014 May 05]. Available from: http://scholarlyoa.com/2012/11/30/criteria-for-determining-predatory-open-access-publishers-2nd-edition

HOW TO CITE: Butler-Adam J. Dealing with ‘open access’ demons. S Afr J Sci. 2014;110(5/6), Art. #a0070, 1 page. http://dx.doi.org/10.1590/sajs.2014/a0070

The unknown underworld: Understanding soil health in South Africa

AUTHORS:
Schalk v.d.M. Louw[1]
John R.U. Wilson[2,3]
Charlene Janion[3]
Ruan Veldtman[2,4]
Sarah J. Davies[3]
Matthew Addison[4,5]

AFFILIATIONS:
[1]Department of Zoology and Entomology, University of the Free State, Bloemfontein, South Africa
[2]South African National Biodiversity Institute, Kirstenbosch National Botanical Gardens, Cape Town, South Africa
[3]Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Stellenbosch, South Africa
[4]Department of Conservation Ecology and Entomology, Stellenbosch University, Stellenbosch, South Africa
[5]Hortgro Science, Stellenbosch, South Africa

CORRESPONDENCE TO:
Schalk Louw

EMAIL:
louws@ufs.ac.za

POSTAL ADDRESS:
Department of Zoology and Entomology, University of the Free Sate, PO Box 339, Bloemfontein 9300, South Africa

KEYWORDS:
soil signature; edaphic factors; sustainability; food security; sustainable agriculture

HOW TO CITE:
Louw SvdM, Wilson JRU, Janion C, Veldtman R, Davies SJ, Addison M. The unknown underworld: Understanding soil health in South Africa. S Afr J Sci. 2014;110(5/6), Art. #a0064, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0064

The need to provide food security to a growing human population in the face of global threats such as climate change, land transformation, invasive species and pollution[1] is placing increasing pressure on South African soils. South Africa is losing an estimated 300–400 million tonnes of soil annually[2], while soil degradation is a major threat to agricultural sustainability[3]. In spite of these problems, treatment of soil health in biodiversity assessment and planning in South Africa has been rudimentary to date.[4,5]

Defining soil health

Soil is a crucial component of the pedosphere, which sustains life, and should therefore be regarded as one of the most important assets held by South Africans. However, in South Africa, soil is a highly neglected research focus in ecosystem service delivery. Studies of ecosystem services often focus on more elegant and tractable systems, such as pollination networks.[6] Currently in South Africa, soils are viewed in certain sectors as resources that can be used to generate short-term gains, rather than assets to be protected and developed. Soils form the basis for food security through agriculture, where processes taking place in the pedosphere result in water retention, nutrient augmentation and soil biodiversity proliferation.

In an effort to facilitate research on soil health, or at least stimulate debate on the topic, we propose that soil health be measured by a combination of abiotic (A) and biotic (B) and socio-economic (S) aspects relative to a benchmark measure, i.e.

$$\text{Soil health} = \frac{\sum A.\ \sum B.\ \sum S}{\text{benchmark}}$$

where A is measured by a subset of soil physicochemical indicators (with subsets determined on the basis of variable thresholds relating to soil type and soil usage); B is determined by standard biodiversity metrics (e.g. species richness, abundance, network or species assemblage connectedness), incorporating biodiversity in the context of applied strategies (e.g. agricultural push–pull systems and mixed cropping) and S is determined by socio-economic values (e.g. monetary value, equity, human well-being). In order to scale each component, benchmark values will also have to be determined that will serve as the denominator for calculation of changing component values over time.[7,8]

Here we use a similar model to that used in above ground environmental analyses (e.g. the IUCN's system analysis in Leverington et al.[9]), but we recognise that the below ground 'closed' medium functions at different tempos and scales. As such, this model is simplistic, yet a degree of ignorance exists about 'understanding' soils,[10] strengthening the notion that soils need to be elevated to mainstream research foci where interactions among the physical, chemical and biological components of soils receive precedence and serve as a point of departure.

For soils to operate in a complex, interacting total system manner, biodiversity in different environments serving different socio-economic requirements can potentially be temporally and spatially separated, e.g. hydroponic farms and conservation areas or an ecological network in which all aspects are incorporated. In some cases this can lead to an overall greater soil health set-up than if all elements were combined in one area at a specific time period (e.g. debate on conservation *versus* agriculture, or conservation *and* agriculture[11] and landscape-scale analysis over seasons[12]).

The three components of soil defined here all contribute to ecosystem services and intersect to provide healthy soils. The model for this soil health index (Figure 1), supported by intersection descriptions and more detailed relevant examples (Figure 2), serves to emphasise that soils are extremely complex and function in multiple roles, and as such have a pivotal role in ecological function. Based on this framework, we formulated several key research questions (Table 1).

The need for foundational work on soil organisms

In the last decade, the diverse roles of soil communities in the ecological function of soils has gained global recognition.[13] Several large multidisciplinary projects in Europe (such as ENVASSO and EcoFINDERS) now focus on soil organisms using holistic approaches incorporating traditional taxonomy[14] and modern molecular techniques[15]. However, in South Africa, like elsewhere in the world, research in the field of soil biology has been neglected compared with research in soil chemistry or soil physics. This scenario has started to change over the past decade or two and South Africa is no exception in this regard. Research on a broad biological basis regarding South African soils has increased since the mid-1990s and these outcomes are published in journals such as the *European Journal of Soil Science*, *Soil Biology and Biochemistry*, *Biogeochemistry*, *Soil Research*, *Geoderma* and the *South African Journal of Plant and Soil*. Sadly, however, this cannot be said of pure foundational research on soil organisms and, despite some notable pioneering experts (e.g. Lawrence[16]), our knowledge of South African soil organisms is largely restricted to taxonomically well-known groups such as ants[17-19] and spiders[20], and even then this knowledge is often fragmented and poorly documented. The need to integrate existing research initiatives was unanimously expressed at a Soil Health Workshop at the XVII Congress of the Entomological Society of Southern Africa in July 2011. This expression led to the formation of SERG (Soil Ecosystem Research Group) – a soil biodiversity research group that provides a platform for linking and promoting research on soil organisms.

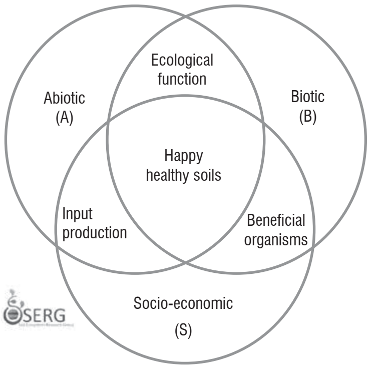
Conceptual model	Area	Definition	Examples
 Ecological function Abiotic (A) Biotic (B) Happy healthy soils Input production Beneficial organisms Socio-economic (S) SERG	A∩B∩S	Happy healthy soils – production landscapes where the inputs are minimal and biodiversity is maintained through sustained soil ecosystem service delivery	- Sustainable flower harvesting from the wild (Figure 2) - Mixed cropping system with conservation tillage
	A∩B	Ecological function – natural ecosystems with significant functional roles. Ecological processes and physicochemical cycles are maintained; soil condition is preserved	- Wetlands - Unfarmed Succulent Karoo ecosystems – dry low primary productivity regions with low or no animal stocking (Figure 2)
	B∩S	Beneficial organisms – soil organisms are used by humans for utilitarian purposes	- Biological control (e.g. entomo-pathogenic nematodes) and vermi-composting - Bio-prospecting
	A∩S	Input production – production systems where physical and chemical properties are manipulated to maintain production	- Mono-cultural intensive agriculture
	B	Biotic – soil biodiversity, including organisms of all different taxonomic and functional groups, which together result in multi-trophic interactions	- Maputo-Pondoland Centre of Endemism
	A	Abiotic – physical and chemical properties of soil	- Iron oxide rich Kalahari sandy soils
	S	Socio-economic – encompasses ownership and land use and subsequent production	- Hydroponics (Figure 2)

Figure 1: Proposed conceptual scheme for defining soil. We consider healthy soils as those that provide abiotic, biotic and socio-economic services.

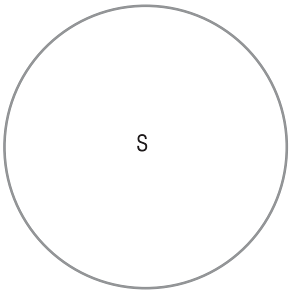
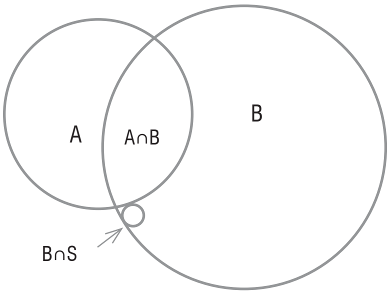
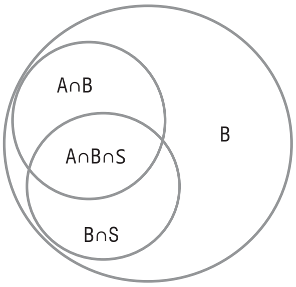
Hydroponics	Nature reserves in the Succulent Karoo	Sustainable harvesting of wild flowers
Approximately 800 ha of hydroponics in RSA in 2002[21]	Knersvlakte Nature Reserve: 24 058 ha[22]	South Africa's Agulhas Plain: 30 597 ha[23]
		
Productivity via hydroponics, means there is no 'soil' at all.[24] However, such systems have no ecosystem functionality. A neighbouring/interlinked soil system is required for soil processes to continue. System will require continual inputs.	Succulent Karoo soils are likely to harbour many endemic species,[25] although they have been poorly studied to date. There are some important functions, e.g. reducing erosion, but very little productive value exists, and in many cases such soils are sensitive to disturbance.[26]	Several initiatives are in place to combine economic value with biodiversity conservation. The selling of flowers (particularly Proteaceae) from nature reserves as 'green' products raises money for environmental restoration and management, as well as local development.

Figure 2: Case studies highlighting how 'soils' differ in abiotic, biotic and socio-economic aspects (based on examples from Figure 1).

Table 1: Key research questions/topics in soil ecology in South Africa

1.	What are the underlying interactions and independent values of sustainable agriculture and intensive agriculture (A∩B∩S & A∩S), taking cognisance that intensive agriculture might not necessarily be unsustainable?
2.	What is the human carrying capacity of a functional, healthy soil, irrespective of whether it is used for crop farming or stock farming?
3.	How much of soil biodiversity is resistant, or resilient or incompatible with disturbance?
4.	How can natural benchmarks for soils in South Africa be determined?
5.	Can the interactions within the total system, e.g. those among production systems, soil organisms and water-based soil nutrient cycling, be analysed? Such analysis could include the following:
•	Interaction and feedback loops between soil organisms and nutrient cycling.
•	Interaction and feedback loops between plants and soil organisms that affect nutrient cycling.
•	Interaction and feedback loops between soil nutrient cycling processes and crop yield.
•	Identification of soil organisms and their nutrient processing qualities.
•	Quantification of nutrient cycles.
•	Identification of species assemblages most beneficial to soil processes and crop yield.

One of the first priorities identified by SERG was the need to collate and mobilise data and collections to consolidate and compare the state of knowledge of each group.

Conclusion

We anticipate that research on soils will be a major initiative linking fundamental and applied research endeavours in the times ahead, especially in the context of climate smart management strategies. Having said this, we do recognise that the establishment of thresholds for biological indicators of soil health is a far greater challenge than the establishment of thresholds for either chemical or physical indicators of soil health, simply because biological indicators are too variable over short periods. Future research endeavours will therefore have to breach this complication.

Acknowledgements

We thank our respective institutions and departments for support in terms of funding and facilities. We also thank the NRF for funding. An anonymous reviewer is gratefully acknowledged for providing many positive and thought-provoking comments which helped improve the manuscript.

References

1. Millennium Ecosystem Assessment. Ecosystems and human well-being: Biodiversity synthesis. Washington, DC: World Resource Institute; 2005.
2. Huntley B, Siegfried R, Sunter C. South African environments into the 21st century. Cape Town: Human and Rousseau and Tafelberg Publishers; 1989.
3. Du Preez CC, Van Huyssteen CW, Mnkeni PNS. Land use and soil organic matter in South Africa 2: A review on the influence of arable crop production. S Afr J Sci. 2011;107(5/6), Art. #358, 8 pages. http://dx.doi.org/10.4102/sajs.v107i5/6.358
4. Department of Environmental Affairs. State of the environment: Land [homepage on the Internet]. No date [cited 2013 Oct 14]. Available from: http://soer.deat.gov.za/22.html
5. Driver A, Sink KJ, Nel JN, Holness S, Van Niekerk L, Daniels F, et al. National biodiversity assessment 2011: An assessment of South Africa's biodiversity and ecosystems. Synthesis report. Pretoria: South African National Biodiversity Institute and Department of Environmental Affairs; 2012.
6. Carvalheiro LG, Veldtman R, Shenkute A, Tesfay GB, Pirk CWW, Donaldson JS, et al. Natural and within-farmland biodiversity enhances crop productivity. Ecol Lett. 2011;14:251–259. http://dx.doi.org/10.1111/j.1461-0248.2010.01579.x
7. Lobry de Bruyn LA, Abbey JA. Characterisation of farmers' soil sense and the implications for on-farm monitoring of soil health. Aust J Exp Agric. 2003;43:285–305. http://dx.doi.org/10.1071/EA00176
8. Gugino BK, Idowu OJ, Schindelbeck RR, Van Es HM, Wolfe DW, Moebius BN, et al. Cornell soil health assessment training manual. New York: Cornell University; 2007.
9. Leverington F, Hockings M, Lemos Costa K. Management effectiveness evaluation in protected areas – a global study. Gatton, Australia: IUCN; 2008.
10. Jones DL, Dennis PG, Owen AG, Van Hees PAW. Organic acid behaviour in soils – misconceptions and knowledge gaps. Plant Soil. 2003;248:31–41. http://dx.doi.org/10.1023/A:1022304332313
11. Balmford A, Green R, Phalan B. What conservationists need to know about farming. Proc R Soc B. 2012;279:2714–2724. http://dx.doi.org/10.1098/rspb.2012.0515
12. Farina A. Principles and methods in landscape ecology. Towards a science of landscape. Dordrecht: Springer; 2006.
13. Decaëns T, Jiminéz JJ, Gioia C, Measey GJ, Lavelle P. The values of soil animals for conservation biology. Eur J Soil Biol. 2006;42:S23–S38.
14. Huber S, Prokop G, Arrouays D, Banko G, Bispo A, Jones RJA, et al., editors. Environmental assessment of soil for monitoring: Volume I. Indicators & criteria. EUR 23490 EN/1. Luxembourg: Office for the official publications of the European communities; 2008.
15. Mulder C, Vonk AJ. Nematode traits and environmental constraints in 200 soil systems: Scaling within the 60–6000 μm body size range. Ecology. 2011;92:10. http://dx.doi.org/10.1890/11-0546.1
16. Lawrence RF. The biology of the cryptic fauna of forests. With special reference to the indigenous forests of South Africa. Cape Town: Balkema; 1953.
17. Robertson HG. Afrotropical ants (Hymenoptera: Formicidae): Taxonomic progress and estimation of species richness. J Hymenopt Res. 2000;9:71–84.
18. Robertson HG. Revision of the ant genus *Streblognathus* (Hymenoptera: Formicidae: Ponerinae). Zootaxa. 2002;97:1–16.
19. Parr CL, Robertson HG, Chown SL. Apomyrminae and Aenictogitoninae: Two new subfamilies of ant (Hymenoptera: Formicidae) for southern Africa. Afr Entomol. 2003;11:128–129.
20. Foord SH, Dippenaar-Schoeman AS, Haddad CR. Chapter 8 – South African spider diversity: African perspectives on the conservation of a mega-diverse group. In: Grillo O, Venora G, editors. Changing diversity in changing environment. Rijeka, Croatia: InTech Publishing; 2011. p. 163–182.
21. Gull C. Study of *Pythium* root diseases of hydroponically grown crops, with emphasis on lettuce [MSc dissertation]. Pretoria: University of Pretoria; 2002.

22. Desmet PG. A systematic plan for a protected area system in the Knersvlakte region of Namaqualand. Stellenbosch: World Wildlife Fund; 1999. Available from: www.pcu.uct.ac.za/resources/reports/ipc9901.pdf

23. Conradie B. Farmers' views of landscape initiatives: The case of the Agulhas plain, CFR. Cape Town: Centre for Social Science Research, UCT; 2010. Available from: www.cssr.uct.ac.za/sites/cssr.uct.ac.za/files/pubs/wp278.pdf

24. Bridgewood L. Hydroponics: Soilless gardening explained. Marlborough, Wiltshire: The Crowood Press; 2003.

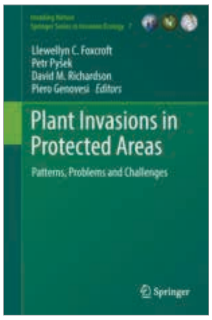
25. Cincotta RP, Wisnewski J, Engelman R. Human population in the biodiversity hotspots. Nature. 2000;404:990–992. http://dx.doi.org/10.1038/35010105

26. Brooks TM, Mittermeier RA, Mittermeier CG, Da Fonseca GAB, Rylands AB, Konstant WR, et al. Habitat loss and extinction in the hotspots of biodiversity. Conserv Biol. 2002;16:909–923. http://dx.doi.org/10.1046/j.1523-1739.2002.00530.x

Protected but vulnerable

BOOK TITLE:
Plant invasions in protected areas: Patterns, problems and challenges

BOOK COVER:



EDITORS:
Llewellyn C. Foxcroft, Petr Pyšek, David M. Richardson and Piero Genovesi

ISBN:
9789400777491

PUBLISHER:
Springer, EUR169.99 (hardcover), EUR142.79 (ebook)

PUBLISHED:
2013

REVIEW TITLE:
Protected but vulnerable

REVIEWER:
Brian W. van Wilgen

EMAIL:
bvanwilgen@sun.ac.za

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

POSTAL ADDRESS:
Department of Botany and Zoology, Stellenbosch University, Matieland 7602, South Africa

HOW TO CITE:
Van Wilgen BW. Protected but vulnerable. S Afr J Sci. 2014;110(5/6), Art. #a0062, 1 page. http://dx.doi.org/10.1590/sajs.2014/a0062

The ongoing transformation of natural ecosystems to provide food, water, fibre, industrial development and space to live for an unprecedented and growing number of people has led to a global environmental crisis. Thousands of species are now threatened with extinction as a direct result of some form of human activity. One of the strategies for preventing a complete transformation of the globe is to set aside protected areas, where species, habitats and ecosystems should be safeguarded from disturbance. However, these protected areas face many challenges, not least of which is invasion by hundreds of newly introduced species that displace the indigenous biota, disrupt ecosystem processes and, in many cases, completely transform the invaded ecosystems. *Plant Invasions in Protected Areas* provides the first global synthesis of invasive alien species in protected areas since a review conducted 25 years ago under the auspices of the Scientific Committee on Problems of the Environment (SCOPE).[1]

The number of protected areas in the world has grown sevenfold over the past half century, from ~20 000 in 1970 to almost 158 000 in 2011, and the network now covers almost 15% of the earth's land surface. Invasive alien species have been present in these areas for a long time, but their impacts are only now really starting to be felt, and are set to grow. Finding ways of dealing with these problems is fast becoming one of the main headaches for managers of many protected areas worldwide. The book sets out to document the understanding of the problem in the context of modern invasion ecology, and to determine how current understanding can support successful management that will limit further negative impacts. Conservation scientists by and large agree that invasive alien plants in protected areas should and can be brought under control, but carrying out such management can also be a wicked problem. As pointed out in the book, matters are complicated by the fact that some people even welcome the alien species,[2] or argue that alien invasions should be accepted as inevitable and perhaps even desirable. In addition, some managers of protected areas try to address these problems by attempting to find beneficial uses for alien species, in most cases leading to further conflict, and failure of control efforts.

The book contains a wealth of information on the origins and status of plant invasions, and their management, in protected areas across the world. The majority of impacts associated with invasive alien species have to date been felt on islands. Prior to human settlement (in some cases very recent), islands were protected from invasions by the long distances from sources of alien species. Consequently, islands proved to be very vulnerable to invasion by newly introduced species. The occupation of ocean islands by people led to the extinction of many unique species, initially directly by hunting, but more recently because islands were swamped by alien species, many invasive. For example, about 70% of the 1000 vascular plant species of the Azores in the North Atlantic are alien. Not surprisingly, therefore, this book has a wealth of information on invasions in protected areas on islands, including those in the Mediterranean, Pacific, Indian, Atlantic, and Southern Oceans. These oceans include iconic World Heritage Sites such as the Galapagos and Aldabra, and the remote islands of the Southern Ocean. There is much to be learnt from islands, and this knowledge is important because severe impacts associated with invasive species, once regarded as island phenomena, are increasingly being felt on continents.

Much of the information is presented in a series of often fascinating case studies, which address the management of both individual species and of areas. African examples include a 'botanical garden' that was established in 1903 in the Amani Nature Reserve, Tanzania (Amani is part of East Usambara, a biodiversity hotspot of global importance). Here over 600 alien woody species were planted in blocks covering 300 ha, leading to the naturalisation of 73 species, 11 of which are now serious invaders. Conservation staff in the Kruger National Park, South Africa, were for a considerable time responsible for the ongoing introduction and cultivation of numerous invasive alien plants in tourist camps and staff villages, despite knowledge of their invasive potential and policies that prevented these practices. Thus, it was often the very people appointed to manage protected areas that set the scene for their degradation through invasion.

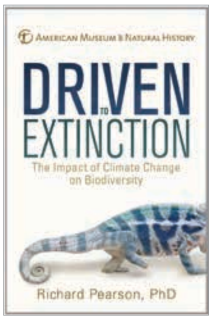
Whether or not this problem can be contained is the focus of the final third of the book. Here the options available to managers are thoroughly reviewed, and the need for a comprehensive approach becomes clear to the reader. Such a comprehensive approach should include preventing new introductions, regular surveillance, early detection and rapid response, eradication where possible, biological control, raising awareness, capacity building, political lobbying, integration of control into protected area management plans, management of invasions outside of the protected area boundary, and ecosystem restoration. The clear message that emerges is that managers need something of a paradigm shift away from 'weed management' (an agricultural concept in which the aim is to control certain established weeds), to management for biodiversity outcomes, in which the goal is to protect biodiversity, and which requires a much broader approach, better planning and regular monitoring. Much experience has been gained from the management, across the globe, of the 135 protected areas that are reviewed in this book. This book is one that should not only be read by students of invasion ecology, but also by conservation policymakers and the managers of protected areas. Whether or not they will is another question, as it is well known that a wide gap exists between research-based texts such as this one, and implementation of the recommendations in practice. This vital aspect is addressed only very briefly in the book under the heading 'Bridging the science–management divide', which refers to high-level coordinating bodies. In my view, much closer partnerships between researchers and on-the-ground managers will be needed – a substantial challenge to the research community. The lessons documented in this book deserve to reach a much wider audience than they probably will, and I would recommend this book to anyone involved with the management of biological invasions anywhere in the world.

References

1. Macdonald IAW, Loope LL, Usher MB, Hamann O. Wildlife conservation and the invasion of nature reserves by introduced species: A global perspective. In: Drake J, Mooney HA, Di Castri F, editors. Biological invasions: A global perspective. Chichester: Wiley; 1989. p. 215–255.
2. Dickie IA, Bennet BM, Burrows LE, Nuñez MA, Peltzer DA, Porté A, et al. Conflicting values: Ecosystem services and invasive tree management. Biol Invasions. 2014;16:705–719. http://dx.doi.org/10.1007/s10530-013-0609-6

BOOK TITLE:
Driven to extinction: The impact of climate change on biodiversity

BOOK COVER:



AUTHOR:
Richard Pearson

ISBN:
9781402772238

PUBLISHER:
Sterling, New York,
R260 (hardcover)

PUBLISHED:
2011

REVIEW TITLE:
Climate change, biodiversity and extinction risk

REVIEWER:
Timothy Kuiper

EMAIL:
timothykuiper@gmail.com

AFFILIATION:
Department of Zoology and Entomology, Rhodes University, Grahamstown, South Africa

POSTAL ADDRESS:
Department of Zoology and
Entomology, Rhodes University,
PO Box 94, Grahamstown 6140,
South Africa

HOW TO CITE:
Kuiper T. Climate change, biodiversity and extinction risk. S Afr J Sci. 2014;110(5/6), Art. #a0063, 1 page. http://dx.doi.org/10.1590/sajs.2014/a0063

Climate change, biodiversity and extinction risk

Warnings about the drastic impacts of climate change on the natural world are nothing new. In an era of sensationalised media and half-truths surrounding the climate change 'debacle', society is in need of hard evidence. *Driven to Extinction* strives to give us just that. Be warned: the evidence tells an uncomfortable story.

Drawing extensively on the scientific literature, the book presents a timely overview of what science can tell us about climate change and its impact on the earth's plants and animals. From poleward and upslope shifts in the distribution of species, to advanced spring phenology, the reader is left convinced of the fingerprint of human-mediated changes in climate on our planet's biodiversity. But are we really facing an impending disaster of species loss? Pearson's level-headed approach to tackling this question is commendable, and in many ways this book presents an unbiased answer.

The foundation and direction of the book is laid in the opening chapter, which argues that the best representation of our current understanding of the matter points to the reality of contemporary climate change and implicates humans in driving the process. The fragility of the relationship between particular species and climate, the significance of contemporary climate change on the broader geological timescale and the role of warming in desynchronising the phenology of closely dependent species are explored in the next three chapters. Chapter 6 looks into the future, using warming scenarios and likely species responses to predict possible rates of extinction in the coming century. Here we are introduced to some alarming numbers – with recent research suggesting that as many as one-quarter of our planet's species may be committed to extinction within the century.

Following this stark picture, Pearson demonstrates equanimity by giving extensive consideration to some alternatives to widespread species loss in the future. It is possible that some species may adapt to warming and change, that complex communities of species may allow for compensation and that drastic ecosystem-level phase shifts in response to climate may not be all bad. Notwithstanding the consideration given to these alternatives, Pearson expresses doubt as to whether these alternatives have the capacity to mitigate the effects of rapid climate change across all species and systems.

Driven to Extinction must be lauded for transforming information from over 150 scientific articles, books and reports into a coherent and captivating tale. Brimming with examples of demonstrated effects on species, communities and ecosystems, the book presents a meticulous exploration of the length and breadth of how climate change is affecting life on our planet and how it might continue doing so. All of this information is condensed into a simply written and easy-to-understand account that just about anyone will be able to appreciate. In short, Pearson has made information from the scientific research and literature freely accessible to the layperson, typifying the purpose of popular science literature. The book is available in softcover, hardcover and ebook formats. The hardcover edition includes full-colour photographs of some of the extraordinary species used as examples in the text.

Despite its substance, the book could have been improved in some ways. The sceptic may notice Pearson's susceptibility to the confirmation trap: he cites and supports only those sources of information that support the overall argument of the book and downplays the alternatives. Furthermore, the oversimplified explanations, broad generalisations and emotive language will perhaps make for a less satisfying read for the expert in the field. Despite these minor criticisms, the book achieves its self-proclaimed purpose: to demonstrate the severe threat that climate change poses for many species.

So what can we do about climate change? The final chapter acknowledges that climate change acts in concert with other important threats to biodiversity such as habitat loss. Conservation strategies must commensurately emphasise the preservation of larger protected areas and corridors that buffer against the effects of climate change. At the end of the day, however, we must address the root of the problem: the notorious rise in greenhouse gases. Although to achieve this will involve globally binding policies and action, the final sentence of the book lays emphasis on the role of the individual: 'It is up to you as a citizen to help chart our course for the future.'

Every responsible citizen should take the time to read this book.

Some thoughts about the future of food and agriculture

AUTHOR:
Louise O. Fresco[1,2]

AFFILIATIONS:
[1]University Professor, University of Amsterdam, Amsterdam, the Netherlands
[2]Honorary Professor, Wageningen University, Wageningen, the Netherlands

CORRESPONDENCE TO:
Louise Fresco

EMAIL:
officefresco-bb@uva.nl

POSTAL ADDRESS:
University of Amsterdam,
PO Box 19268, 1000 GG
Amsterdam, the Netherlands

KEYWORDS:
hunger; food; agriculture; food production; food security

HOW TO CITE:
Fresco LO. Some thoughts about the future of food and agriculture. S Afr J Sci. 2014;110(5/6), Art. #a0066, 2 pages. http://dx.doi.org/10.1590/sajs.2014/a0066

The fact is simple: 842 million people are currently hungry. These people are some of about 2 billion who lack, at least at times, essential nutrients even if they consume enough calories. A century ago, more than half of humankind was severely malnourished in terms of calories and proteins; today this number is only 1 in 8 – much lower than in the past, but still unacceptably high. In the last two decades, more people than ever before – 1 billion – have been lifted out of poverty and hunger. About 75% of this spectacular global reduction occurred in China. Making similar progress in the next decades means tackling Africa and India.

Hunger and poverty mutually reinforce one another. Undernourished women tend to give birth to underweight babies who are prone to stunted growth, cognitive impairment and disease. Poor and hungry people are often marginalised because they lack the initiative or the means to participate in society. Of the people living in extreme poverty, that is, on less than USD1.25 a day, the majority are women and children.

There are three types of situations in which chronic hunger is experienced. In areas of civil unrest and failing governments, people lack the basic means of survival (this situation may also occur temporarily after natural disasters). In rural areas, where most of the poor and hungry live, there are few employment opportunities and partial food self-sufficiency is eroding. In urban areas the cost of living can be prohibitive, and employment can be intermittent, poorly paid and dangerous. The balance of poverty and hunger is shifting towards urban areas, where soon the majority of the poor will live.

Reducing hunger requires several responses. Emergency food aid is needed to alleviate the needs of displaced populations, coupled with peace. Proper child nutrition is needed to prevent damage from undernutrition. Overall, hunger reduction results from increased purchasing power which in turn comes about through economic growth. More equitable growth leads to a more consistent reduction of hunger. In countries in which a large proportion of the population still finds employment in the rural areas, investment in agricultural production and processing is the most efficient way to reduce poverty. Increasing food supply lowers prices while trade helps to overcome seasonal and local shortages.

Overconsumption and inadequate nutrition touch about 1.5 billion people worldwide who are overweight or obese, increasingly so in emerging economies (Mexico has the highest percentage of overweight citizens). There is an inverse relationship between economic class (income) and overconsumption of calories, although this relationship is less linear in the case of hunger. Overconsumption is associated with undernutrition, as both obesity and malnourishment imply nutritional imbalances. Furthermore, children born from undernourished mothers may risk becoming overweight in adolescence.

The world will need 50% more calories in 2030. The world's population is expected to increase by 0.7% annually. Nearly all that growth will take place in developing and emerging markets, most of it in Africa. Increasing urbanisation and income levels are well known to lead to higher demands for animal foods and a greater dietary diversity, including a reliance on fast food. At the same time, 2.7 billion people will then use traditional biomass for cooking: two thirds of sub-Saharan Africa.

Future demand will be influenced by changes in consumer preferences. Middle-class consumers who increasingly consider food as a way to establish identity and health are driving the changes. Africa now numbers 200 million people with middle-class incomes. Combined with rapid ageing of the population, this growing middle-class sector implies new food products and individualised nutrition. Consumer concern has led to increased demands for transparency and regulation for sustainable, animal-friendly and healthy products. Urban populations tend to eat out more and to eat more pre-processed products. These types of consumer demands – and not supply – will be the primary drivers of change in future agricultural production.

Progress in food production in the past half century has resulted from a successful combination of genetic improvements, such as high-yielding varieties, and improved cropping techniques, in particular irrigation and fertilisers. This progress started in Asia with cereals (known as the Green Revolution) and extended to other crops in Asia, such as cassava, potato, sorghum and some livestock species, and to Africa, and evolved to improved soil and water quality and use of organic fertilisers. Today the approach to improving productivity includes optimal resource use efficiency, including energy efficiency, and reducing greenhouse gas emissions along the entire food chain. Other considerations include meeting goals for human health and the environment, including reducing pathogens and improving the uptake and retrieval of nutrients by humans. The next step enhances nutrition through improved food quality and fortification of essential minerals and vitamins, either bred in the plant or animal, or added during processing.

Thanks to decades of agronomic research, we understand that production resources such as water are used more efficiently at higher resource endowment, i.e. when fertilisation or crop protection are optimal. This principle also appears to hold true for animal production. The potential to increase resource efficiency is matched by the availability of underutilised lands. The successful exploitation of Brazil's *cerrado* (savannah) region, hitherto considered marginal land, has led to a reassessment of land resources. Sub-Saharan Africa has ample underutilised land, together with surface water resources, which may allow considerable expansion of irrigated areas (now at 4% as opposed to 40% in Asia).

Food security requires provisions to reduce vulnerability through (regional or local) stocks and economic incentives, e.g. crop and animal insurance. In flood-prone areas, flood risk prevention should be coupled with disaster management including safe areas for people, cattle and stocks. Furthermore, farmers are usually unable to use new technology if they have no secure land title and no access to secure financial and credit services or to marketing and other inputs such as irrigation and fertilisers. Above all, food security requires an enhanced emphasis on food safety in complex and anonymous food chains in which vertical upstream integration leads to increasing concentration of power. Retailers want to ensure supply through contracts and the acquisition of food processors. This dominance, driven by cost-cutting, leads to anonymity of suppliers and has replaced the long-term contracts that existed previously, which often were based on trust. Concentration means that a small group of players determine a large part of the market, resulting in lower prices for producers and potential irresponsible behaviour. The current lack of a level playing field in terms of public health, animal welfare and greenhouse gas emissions may put consumers at risk. The best way to create a level playing field is through international agreements, open borders and trade, and capacity building to bring all countries up to the same standards.

Overall, our food is safer than ever, but structural weaknesses exist, resulting from inadequate regulation, sloppy compliance and even fraud, as well as public health risks posed by high concentrations of animals, in particular pigs and poultry, close to major urban centres. Physical segregation of animals needs to be enforced, as well as protection of workers. Efficiencies in feed supply are to be addressed, in particular the improvement of nutrient uptake through digestion. Antibiotics must be strictly regulated. More can be done to reduce greenhouse gas emissions from livestock, even if trade-offs exist among animal welfare, public health and the environment.

The most pressing food security issue today is the provision of animal proteins. Since 1950, global livestock has grown fivefold and the demand for animal protein is expected to increase by 75% in 2030. No other issue raises more moral, political and technical questions and leads to greater lack of consumer confidence. Alternative protein sources, from plants, algae, insects or even bacteria may reduce the demand for food and feed. It is possible to substitute up to one-third of animal proteins in processed meat. The use of alternative resources, including the retrieval of proteins from waste and insects for feed or food, makes perfect ecological sense. Consumers need to be enticed to make the right – i.e. healthy and sustainable – dietary choices through education and diversification, in particular with respect to proteins, fruits and vegetables.

Labour requirements in developed economies have dropped dramatically in the last century, to 5% of 1900 levels in wheat production, whilst yields grew fivefold. However, in most countries, rural labour availability is declining rapidly because of migration and rising urban wages, while mechanisation remains underdeveloped. As a result, food production becomes increasingly the job of female farmers, under harsh conditions, leading to low labour productivity. Furthermore, the agricultural labour force is aging rapidly. One of the most urgent issues is to entice young and dynamic men and women into agriculture by providing them with the means to become entrepreneurs who earn a decent income and to dispose of labour-enhancing technology. The 500 million or more smallholder farmers of today should not become victims of low capital, a small labour force and low land productivity. Appropriate technology, based on the latest scientific insights, is required. Smallholders do not want to remain small if it means low income and low status. Adolescents and the youth number 1.8 billion today. If we assume that the current smallholders will consolidate and modernise their farms, leading to a reduction by, say 20–25% until 2030, we will need 350–400 million youngsters, nearly one in five, to go into farming and food production.

Climate change may influence plant and animal growth through CO_2 levels, day and night temperatures, precipitation, lengths and dates of onset of growing seasons, rainfall variability, wind speed, pest and disease build up, sea level rise, groundwater tables and salt intrusions. Insect-borne diseases may become an important problem. In regions affected in the relatively short term by reduced precipitation, i.e. Africa, Australia, the Mediterranean and the Middle East, yields are projected to decline as a result of water and heat stress. Many of the challenges of climate change may be met by age-old techniques dealing with the vagaries of weather. However, the speed of change is new, as are the number of people affected. It seems likely that climate change will lead to more fluctuations in production in response to weather variability and possibly increased probabilities of extreme weather events. Global net land-use effects, in terms of area needed for future agricultural needs, seem limited, but local area expansion is expected because of unfavourable weather conditions.

Food is not oil – it is a renewable resource that cannot be exhausted, as food production renews itself with every growing season through photosynthesis. With the possible exception, on a scale of hundreds of years, of phosphate, the inputs for food production (solar radiation, water, nitrogen and potassium) are not globally deficient, even if there may be local shortages. Food is grown by farmers who do not act like a cartel. But food is like oil in the sense that as a set of commodities it is up for speculation. On futures markets, a dry summer in the main producing countries drives up the prices. Prices are likely to rise against the historical low of food prices around 2000 and the long-term downward trend preceding it. Structural and conjectural factors explain the current fluctuations. The extremely rapid growth in demand for food grains, feed and animal and fish products in places like China, India, Brazil and Central Europe has surpassed the growth in production. Production is relatively inelastic and cannot adjust quickly to higher prices, which has led to declining stocks. Increasing energy prices have also affected agriculture through land preparation, transportation and processing costs. Concern about price volatility in recent years has led to suggestions about the creation of global cereal stocks. However, maintaining these stocks is costly and may lead to price distortions. Open trade and transparent markets seem the best guarantee. Overall, rising food prices facilitate investment in agriculture, even though they may be disadvantageous to the urban poor and poor food importing countries.

More than before, public opinion is a driving force in the future of food. The lack of communication by scientists, the private sector and governments has led to severe misconceptions, a lack of trust, false dichotomies and resistance against technology. At best, public opinion alternates between genuine concerns about safety and food nostalgia. The education of the consumer on the realities of food production is important to promote a realistic debate about options.

Lastly, food security is the responsibility of the state. This does not mean that food security necessarily requires state intervention. A strong positive relationship exists between food security and safety and democracy. The state has a role in creating the conditions for sustainable food systems, through a combination of fiscal, legal and policy measures. Markets operate within this framework, and only where markets fail, governments must intervene. Complex and intertwined markets now extend all over the globe and have helped to avoid local food shortage and promote income growth in many hitherto isolated areas. However, such chains require international monitoring.

Never before in human history has the responsibility for the food of so many been borne by such a small minority of farmers, food processors and retailers. A minority that is barely recognised, and often blamed for the ills of environmental damage. The world in which we live is not predictable, but growth is essential to meet the needs of a growing world population with growing aspirations. We know better now than before how to mitigate the negative effects of such growth. The clear trends of the last decades in terms of food production and food security allow us to believe food security and safety for all can be reached in a sustainable manner.

In many ways, South Africa embodies both the challenges and the potential of African food production and consumption. Moreover, through its science and technology and, hopefully, through its enlightened policies, rural programmes and excellent academic institutions, South Africa could play an important role in Africa. The challenge is for South Africa and Africa to live up to the pledges of Maputo and to reverse the declining investment in food and agriculture.

Ubiquitous modelling: In honour of Tony Starfield's 70th anniversary

AUTHORS:
David P. Mason[1]
Michael Sears[2]
Anthony M. Starfield[3]

AFFILIATIONS:
[1]Professor Emeritus, School of Computational and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa

[2]Visiting Professor, School of Computer Science, University of the Witwatersrand, Johannesburg, South Africa

[3]Professor Emeritus, University of Minnesota, St Paul, Minnesota, USA

CORRESPONDENCE TO:
Michael Sears

EMAIL:
michael.sears@wits.ac.za

POSTAL ADDRESS:
School of Computer Science,
University of the Witwatersrand,
Private Bag 3, Wits 2050,
South Africa

KEYWORDS:
ecological modelling; marine modelling; rhino conservation; mathematical modelling; quantitative reasoning

HOW TO CITE:
Mason DP, Sears M, Starfield AM. Ubiquitous modelling: In honour of Tony Starfield's 70th anniversary. S Afr J Sci. 2014;110(5/6), Art. #a0067, 5 pages. http://dx.doi.org/10.1590/sajs.2014/a0067

A 1-day symposium on Computational and Mathematical Modelling was held on 21 April 2012 in the School of Computational and Applied Mathematics at the University of the Witwatersrand in honour of Tony Starfield on the occasion of his 70th anniversary. Starfield was Professor of Applied Mathematics at the University of the Witwatersrand from 1969 to 1979. He spent his career in teaching and doing research in computational and mathematical modelling. The unifying theme of the symposium was computational and mathematical modelling in the broadest sense. Starfield's 60th anniversary was celebrated by a symposium for which the presentations were made by his colleagues of long standing. The proceedings of that symposium were published in the *South African Journal of Science* as the Starfield Festschrift.[1] The symposium on Starfield's 70th anniversary looked to the future of mathematical and computational modelling in the 21st century, with an emphasis on the contributions of a new generation of modellers.

Starfield's keynote paper – 'Ubiquitous modelling' – opened the symposium. He reviewed the development of modelling over the past 50 years; the ideas expressed in his paper were well illustrated by the symposium itself.

Ubiquitous modelling

Seventy years ago one might have defined modelling as the use of mathematics to analyse problems; in other words, applied mathematics. The portfolio of modelling problems would have been dominated by theoretical physics and mechanics. Mathematical techniques determined both the problems selected and the way in which they were posed; there was no point in attempting to model a problem that was mathematically 'intractable'. Advances in techniques (Fourier series, Laplace transforms) opened up new modelling opportunities, but almost all of them within the same neighbourhood of mechanics and theoretical physics.

Modelling problems in what we would now call Operations Research became important during the Second World War and drove the need for more computational power. The development of computers, in turn, revolutionised modelling. At first this revolution was seen as an extension of traditional modelling: numerical techniques and computers could be used to tackle problems that had previously been mathematically intractable. A numerical solution was, however, viewed as an inferior analysis and some journals were slow to publish computational results. I recall J.C. Jaeger, one of the great applied mathematicians of the 20th century, describing the finite element method at a conference in the early 1970s as 'an expensive way to get inaccurate results to irrelevant problems'. Within a few years, finite elements were an indispensable tool in engineering analysis.

The Department of Applied Mathematics at the University of the Witwatersrand was in the forefront of the modelling revolution during the 1970s, beginning with the development of numerical solutions to tackle problems in the traditional portfolio, but recognising very quickly that the real revolution was in exploiting computational power to expand the modelling portfolio into unexplored disciplines. Modelling seminars were started to explore collaborative opportunities in medicine, biology, ecology, wildlife management, building science, socio-economics, and traffic engineering – in other words, ubiquitous modelling!

Computers created the opportunity to explore new disciplines. New disciplines brought new and unusual modelling problems. New problems plus computer power engendered new modelling approaches. New approaches opened up the way for further new problems. The co-evolution of computers, modelling techniques and ubiquitous modelling has been rapid and exciting. Spatial modelling, neural networks, agent-based modelling, visualisation – these are just some of the techniques that a modeller can access using software that is getting easier and easier to use.

But has the philosophy of modelling kept pace with this rapid expansion? If modelling was once constrained by mathematical tractability, that constraint also provided a paradigm for modelling methodology. A modeller knew that it was usually impossible to solve the 'real' problem. The secret was to find a similar problem that was tractable, and then to interpret the results in the light of differences between the real problem and the model. Everybody understood the difference between the 'real world' and the 'model world' and this provided a framework for both appreciating and criticising a modelling paper or report. Moreover, because the number of mathematical techniques was limited, modellers could easily comprehend what other modellers had done. There was a recognised discipline to modelling. Life in the frontier towns of ubiquitous computational modelling is exciting and full of opportunities, but it has produced both good and bad models and the discipline has been slow to follow the pioneers.

What is the discipline of ubiquitous modelling as opposed to mathematical modelling? Jaeger's comment on finite elements was really a call for modelling discipline: when it is possible to model almost anything in a number of different ways, one has to ask whether time and money have been well spent, whether the modelling results fit the problem being addressed, and whether the problem has been carefully framed. The minimal set of questions to be asked about any modelling exercise is:

- Is the 'real world' objective of the exercise clear and well-defined?
- Has the model world been carefully and parsimoniously designed and have all assumptions been noted?
- Is the solution method appropriate, understandable and reproducible?
- Have the results been interpreted back to the real world?

- And, last but certainly not least, how robust are the conclusions? Has a sensitivity analysis explored uncertainties in the data and has an assumption analysis considered the effects of the key assumptions?

The word 'discipline' has been used above in the sense of a modelling methodology, but what about a modelling discipline in the sense of an academic subject? Starfield and Salter[2] argue that the methodology *is* the academic discipline; the above list of questions provides the core of what needs to be taught in any attempt to introduce students to the subject of modelling. Their paper was written in the context of a concerted effort at Oberlin College to introduce modelling across the undergraduate curriculum – ubiquitous modelling – and draws on their experiences in teaching introductory modelling classes to students with backgrounds ranging from literature to neuroscience.

That leads to three important questions:

1. Why should modelling be taught ubiquitously?
2. Can it be taught to students with weak mathematical and numerical skills?
3. How should it be taught?

To provide short answers in reverse order:

Starfield and Salter expand in their 2010 paper[2] on ways to teach the core methodology of modelling. That these can be taught is demonstrated by the success of their classes and is caught by the following (paraphrased) comment of an English major:

> *My friends asked me, over coffee, what classes I was taking. I mentioned Modelling and they burst out laughing and started imagining me showing off designer clothes. It struck me that mannequins who model clothes really are models in the sense that they are designed parsimoniously to achieve a clear objective – to show the clothes to best effect.*

But why should modelling be taught ubiquitously? Firstly, the questions at the core of modelling methodology are in fact at the core of critical thinking in general and experimental science in particular. Moreover, modelling is easy to teach in a way that instantaneously illustrates the power of this methodology. Secondly, we all live in a world that cries out for systematic thinking. One cannot make thoughtful decisions without understanding the interactions and trade-offs between multiple factors. At the very least, to be a responsible citizen, one has to have the ability to ask critical questions about modelling results.

Symposium overview

The application areas addressed by the presentations were wide ranging, involving marine models, land-based ecological models, models in astronomy and astrophysics and models from South African industry. In addition, a panel discussion was held as part of the symposium on rhino conservation in southern Africa as an example of a problem that requires systematic thinking.

Modelling approaches included spatial models, simulation models, and agent-based models as well as mathematical models. The breadth of the papers, both in terms of application area and in terms of methodology, illustrated the idea of ubiquitous modelling in a remarkable way. We present the summary by subject area as that was the order in which the papers were presented. Table 1 summarises the spread of the papers in terms of methodology and approach. (Some of the presentations appear more than once in the table because they covered a variety of models or used multiple approaches.)

Marine process models

The first session was concerned with marine process modelling. Starfield has collaborated for many years with the Marine Research (Ma-Re) Institute at the University of Cape Town.

Modelling plankton production

John Field (former Director of the Ma-Re Institute) presented a paper entitled 'Modelling plankton production by combining sub-surface shipboard measurements with those of the ocean surface from satellites' in which he described the process of developing models that take into account large data sets of daily ocean surface estimates of wind, temperature and chlorophyll from satellites and combining these with sparse ship-based depth profiles of temperature and chlorophyll.[3,4] The models are based on current understanding of the main physical, chemical and biological processes involved in the variability of phytoplankton growth. The temperature and chlorophyll profiles were clustered into a manageable number of optical profiles and dynamical neural networks were then used to predict which profile will occur at a particular place and season, given a few days of surface estimates of wind, temperature and chlorophyll. These profiles are used to give monthly estimates of plankton production over wide areas of the coastal and open ocean.

Starfield ripples in marine social-ecological systems research

Astrid Jarre (SA Research Chair in Marine Ecology & Fisheries, Ma-Re Institute) spoke on behalf of her research team. Their focus is on how to cope with global change in marine social-ecological systems. She described how Starfield had stimulated modelling research in this area, with reference to at least seven projects that were influenced by his modelling philosophy and, in some cases, his modelling methodology for involving both academics and non-academics in conservation management decisions. Examples are: a (system) model of penguin population dynamics in relation to various pressures; a frame-based model looking at different states of the Benguela ecosystem[5,6]; expert systems for ecosystem-based fisheries management; and modelling at the interface between social sciences and fisheries ecology. As an example of ubiquitous modelling, Jarre described how the parsimony and elegance of Starfield's modelling approach had even won the respect of an anthropologist in the group who used Starfield's rapid prototyping philosophy in his MSocSci thesis research.

Modelling in deep sea sonar

Michael Mitchley (Computer Science, University of the Witwatersrand) examined the use of high resolution beam forming methods for sonar bathymetry, including techniques for full-rank covariance matrix estimation and signal enumeration. The modelling of both passive and active sonar was discussed and current challenges were highlighted.

Land-based ecological models

The second session was concerned with ecological modelling of animals on land.

Stability and resilience in seasonal oscillators

Norman Owen-Smith (former Leader of the Centre for African Ecology, School of Animal, Plant and Environmental Sciences, University of the Witwatersrand) presented a paper entitled 'Stability and resilience in seasonal oscillators: Towards mechanistic population models'. Owen-Smith began by observing that models of population dynamics have become fossilised in phenomenological forms that have dubious predictive value, for example, logistic and allied density-dependent equations, Lotka–Volterra models of coupled consumer–resource interactions and Leslie and other matrix models for structured populations. All are unsatisfactory when the reality of global environmental change raises the need for more reliable models to guide difficult decisions for biodiversity conservation.[7] He explored how these alternative model forms might be reconciled and adjusted to represent more faithfully the mechanisms generating changes in population abundance. Populations can be viewed as dissipating systems capturing material and energising resources to maintain their organisation, despite environmental perturbations. The problem for herbivores is that the resource supply varies over the annual

Table 1: Classification of presentations at the workshop

Approach	Model type	Application area	Presenter
Agent based	Economic model	Fisheries system dynamics	A Jarre
Expert system	Computer model	Marine ecosystem dynamics	A Jarre
Frame based	Computer model	Marine ecosystem dynamics	A Jarre
Network model	Representative model	Robotics	D Fanucchi
Neural network	Representative model	Population dynamics	J Field
Simulation	Computer model	Population dynamics	A Jarre
Simulation	Computer model	Ecosystem	D Richardson
Simulation	Computer model	Chemical engineering	A Hutchinson
Simulation	Physical model	Mining engineering	R Kgatle
Simulation	Physical/computer model	Sonar	M Mitchley
Statistical correlation	Physical model	Astrophysics	A Tailor
Statistical extrapolation	Representative model	Population dynamics	J Field
Theoretical	Computational model	Population dynamics	N Owen-Smith
Theoretical	Mathematical model	Population dynamics	J Shaw; all
Theoretical	Physical model	Astrophysics	R Herbst
Theoretical	Physical model	Mining engineering	G Fareo
Theoretical	Economic model	Population dynamics	J Shaw; all

cycle because plant growth is seasonally phased. How then do herbivore populations persist despite seasonal as well as annual variation in the resource supply? This question requires judicially investing the surplus gains of the good times to promote persistence through the lean times. Owen-Smith suggested how this intrinsically disequilibrial system might be simulated by adopting a metaphysiological approach employing a currency of biomass rather than numerical population density. Thereby insights are generated into strategies of resource allocation and adaptive selection among functionally distinct resource types at different stages of the seasonal cycle. Owen-Smith then threw out the challenge: how might such a model system be generalised mathematically and treated analytically?

Modelling the dynamics of complex systems in a highly variable environment

David Richardson (Mathematics and Applied Mathematics, University of Cape Town) began by describing semi-arid rangelands as hierarchical systems with many levels of organisation, ranging from tissues of plants and animals to populations at the ecosystem level. This ecosystem is driven by rainfall which varies widely both between and within years. Rangeland vegetation also varies in both botanical and chemical composition and the spatial distribution of edible material on a plant varies between species. Consequently, mechanisms controlling diet selection and intake differ between veld types. Furthermore, the relative importance of producing meat, milk and fibre varies with the objectives of the pastoralists and with the ecosystem. Variables at one level influence processes at both higher and lower levels. Processes in the different levels operate on widely different timescales and long-term changes depend on the effects of short-term processes. A major difficulty frequently encountered in modelling rangelands is the lack of empirical information between animal productivity, vegetation and rainfall.[8]

Richardson's approach was to simulate the different levels of the system separately and use model output of one level as input for other levels. Model output showed how rainfall variability within years can influence the long-term behaviour of the system and also explained observed unexpected results such as accepted 'good management' leading to a greater probability of rangeland degradation.

How can modelling help save the rhino?

Jo Shaw (Large Mammal Trade, Traffic East/Southern Africa) gave this presentation, which served as a good introduction to the panel discussion and group meetings on rhino. Since 2008, the number of rhinos poached in South Africa has risen dramatically year on year with a total of 448 animals killed during 2011. Rhino horn has been used for thousands of years as a fever-reducing agent as part of traditional Chinese medicine practices. However, research by TRAFFIC, the wildlife trade monitoring network, indicates that more recently there has been increased demand from Vietnam, where rhino horn has a high status cachet. There are approximately 19 000 white rhinos and 1900 black rhinos in South Africa and rhino numbers continue to increase, with births exceeding total mortalities. However, if rhino deaths from poaching continue to rise at the current rate, some scientists estimate that rhino populations will begin to decline by 2015. It is thought that 25% of rhinos in South Africa are privately owned, but accurate data on live rhino numbers and rhino horn stockpiles can be very difficult to obtain. This presentation provided an overview of current understanding of the rhino poaching crisis and raised research questions that could be investigated using computer modelling exercises to inform rhino conservation actions and political decision-making.

Rhino conservation

Trade in rhino horn is prohibited by CITES (Convention on International Trade in Endangered Species). However, illegal trading takes place and the price of rhino horn is believed to be about USD65 000 to USD70 000 per kilogram. A rhino horn weighs about 4–5 kg. The horn can be removed from the rhino down to about 20 mm above the base of the horn without harming the rhino. The horn grows back at a rate of about 40 mm per year.

A panel consisting of Tony Starfield, Norman Owen-Smith, Jo Shaw and David Cumming, formulated three broad modelling questions for small groups of participants to consider:

1. Can a sustainable legal trade capture the market?
2. How can stockpiled rhino horn be used strategically to supply the market?
3. How can a production model for rhino horn, including an economic component, be constructed?

The participants split up into three groups, with each group tackling one question. The groups were not asked to formulate the models mathematically at that stage.

Rhino conservation is essentially a problem in environmental economics. This short exercise framed several plausible modelling projects within the context of environmental economics, which led to the consideration of this problem at the Tenth Mathematics in Industry Study Group which was held at the University of the Witwatersrand from 14 to 18 January 2013.[9]

Models in astronomy and astrophysics

This session contained two presentations on models in astronomy and astrophysics by master's students at the University of the Witwatersrand.

Gravitational torques of nearby barred spiral galaxies

Asha Tailor's research was aimed at using quantitative methods in classifying barred galaxies. Maximum relative gravitational torques were derived and compared for a sample of 40 nearby bright barred disk galaxies. Torques were compared between galaxy pairs and excellent agreement was found between the 3.6 μm and 4.5 μm morphology. The sample used incorporated a wide range of inclination and bar strength values. The tight coupling of 3.6 μm and 4.5 μm morphology provides an excellent opportunity to classify intermediate redshift galaxies out to $z=0.25$. This has important implications for bar-fraction estimates and galaxy accretion and evolution models.

Stellar models

Rhameez Herbst investigated the effect of mass on the radiation flux from a relativistically rotating neutron star. A pulsar was modelled by simulating a relativistically rotating magnetic dipole embedded within a neutron star. The resulting equations retain the mass of the neutron star thereby introducing effects of general relativity on the radiation from the dipole. Exact solutions to the modelling equation were presented as well as plots of energy spectra at different rotational velocities and inclination angles. The results demonstrated that the high speed enhancement of the radiation is always more than compensated for by the frame dragging effect for relativistic neutron star masses leading to a net reduction of radiation from the star.[10]

Models from South African industry

The remaining presentations were based on problems from South African industry and were given by postgraduate students from the School of Computational and Applied Mathematics at the University of the Witwatersrand. The problems had been submitted to one of the Mathematics in Industry Study Group meetings held annually in South Africa since 2004.

Extraction of sugar from shredded cane in a diffuser

The extraction of sugar from shredded cane takes place in a diffuser.[11] A diffuser generally has 12 to 14 components. Ashleigh Hutchinson presented a model in which the sugar cane and the dissolved sugar simultaneously traverse the diffuser in opposite directions. Two coupled first-order difference equations were derived for the sucrose concentration in the juice and the sucrose concentration in the shredded cane. The difference equations were solved analytically and the results were analysed. Spherical voids can occur in the fibre matrix in the diffuser which is undesirable because it inhibits the extraction process. A model was developed which suggests that voids occur as a result of the decrease with depth in the permeability of the fibre matrix.

Rising water table and seismicity

The old, now closed, gold mines in the Witwatersrand are being flooded because pumping of water from the mines has ceased. The mines are intersected by geological faults which are stable because of the frictional resistance to movement and the clamping effect of stresses normal to the fault plane. As the level of the water table rises, water will find its way into the geological faults and thus destabilise them. A simple model was presented by Rahab Kgatle which suggests that the time span for water transport into the faults is likely to be relatively short – months rather than years – and that for deep mines the hydrostatic pressure build up within the filled fractures is likely to increase the risk of fault slip.[12] Slip was predicted to occur along faults not previously prone to slip because of their unfavoured orientation. Under such circumstances, Johannesburg could experience an increase in small seismic events.

An approximate solution to fluid-driven fracture problems

Hydraulic fracturing – in which water is pumped at high pressure in order to open cracks and fissures in rock – has important applications, for example, in releasing gas and oil from shale deposits deep underground. The equations which describe the fracturing process are highly nonlinear and difficult to solve analytically. Because the fluids used in hydraulic fracturing are generally non-Newtonian, a fracturing fluid with power law rheology was considered. The fractures are long and thin and therefore the injection of fluid into the fracture was modelled by Gideon Fareo using lubrication theory. The model also assumed that the excess fluid pressure in the fracture is proportional to the fracture half-width. Fareo found that the fluid velocity averaged across the width of the fracture varied approximately linearly along the fracture. By assuming that the averaged fluid velocity varies exactly linearly along the fracture, an approximate analytical solution for the evolution of the fracture half-width was derived. The approximate analytical solution agreed well with the numerical solution.

Agent tracking and resource allocation on a network

Robots now perform many tasks in industry. In the mining and nuclear industries, they are used to enter locations that are too dangerous for humans. Dario Fanucchi presented a talk on a problem in robotics submitted by the CSIR Mobile Intelligent Autonomous Systems.

Significant work has been done within the last 10 years in the context of robotics on the tracking of autonomous agents in complex domains. Fanucchi introduced mathematical models that capture the motion of simple agents that follow certain dynamical rules and move along a network. The models were then used to develop a strategy for a separate individual, moving on the same or similar network, to maximise the chance of intercepting one or more agents, possibly weighted by importance. Some applications were considered, notably tracking or intersecting insects or birds with known migration patterns and intercepting moving items on a traffic network.

Looking to the future

No modeller can resist attempting to extrapolate, so we end by speculating on the future of the trends and issues we have identified.

Ubiquitous modelling is fuelled by accessible modelling tools. Modellers and students of modelling can learn to use a modelling package in a fraction of the time it would take them to study calculus and differential equations. There are, however, trade-offs: mathematics is capable of generalities and insights that are difficult to obtain via computation – hence the quest for mechanistic population models. It is likely that mathematical models will continue to be useful in this regard, providing a ‘model of a model’: a simplified overview of more detailed computational models (as in the presentation on fluid-driven fractures). However, it is at a more abstract level that mathematics has the most to offer modelling. Computer packages currently allow one to diagram a system model, feed in parameter values, and obtain numerical results for different starting and bounding conditions. What is missing is the ability to understand the meta-properties of the system (stability and resilience, for example). The challenge is to develop modelling packages that routinely provide a meta-analysis.

As powerful modelling tools become easier to use, so they become easier to abuse. There are two responses to this possibility of abuse, both of which are in their infancy. The first is education: there is an urgent need to develop and teach the discipline of modelling. An introductory modelling class should be a requirement for all undergraduate students. The second response is in the development of more intelligent modelling software. Just as word processors routinely check spelling and grammar, so modelling packages could guide and question the user (‘why have you chosen such a large time step?’) One could imagine software that engages modellers in thinking about their model world, suggesting alternative modelling approaches, designing assumption analyses, and providing explanations to the user of the model. Models that are used on a regular basis could monitor their performance, update their parameter values, and suggest changes to the structure of the model. Modelling education and more intelligent software will be symbiotic.

Finally, the diversity of papers presented at this symposium gives a glimpse of the future of modelling in South Africa. Traditional areas of strength such as geomechanics, ecology and resource management will continue to flourish and will likely export ideas and modelling expertise – models that involve stakeholders in resource management, for example, provide a paradigm for all sorts of policy decisions. Modelling successes in the Mathematics in Industry initiative will in turn generate new interests and opportunities. South African involvement in the Square Kilometre Array radio telescope has already led to an upsurge in interest in astronomy and astrophysics; observations from the telescope will create a fertile environment for model development and model testing. Modelling skills developed in one area of application prepare a modeller to work in other areas of application. It follows that providing modelling opportunities and apprenticeships in all of the above will train modellers to deal with the problems of the future, such as those engendered by sustainable energy production and climate change.

Acknowledgements

D.P. Mason and M. Sears thank the Research Committee of the Faculty of Science of the University of the Witwatersrand, Johannesburg, for a grant to run the symposium. We also thank the symposium presenters for their input.

References

1. Mason DP, Wilson DB. Papers to honour A.M. Starfield on his sixtieth anniversary. S Afr J Sci. 2002;98:441–502.
2. Starfield AM, Salter RM. Thoughts on a general undergraduate modelling course and software to support it. Trans R Soc S Afr. 2010;65:116–121. http://dx.doi.org/10.1080/0035919X.2010.513172
3. Demarcq H, Richardson AJ, Field JG. Generalised model of primary production in the southern Benguela upwelling system. Mar Ecol Prog Ser. 2008;354:59–74. http://dx.doi.org/10.3354/meps07136
4. Williamson R, Field JG, Shillington FA, Jarre A, Potgieter A. A Bayesian approach for estimating vertical chlorophyll profiles from satellite remote sensing: Proof-of-concept. ICES J Mar Sci. 2011;68:792–799. http://dx.doi.org/10.1093/icesjms/fsq169
5. Smith MD, Jarre A. Modelling regime shifts in the southern Benguela: A frame-based approach. Afr J Mar Sci. 2011;33(1):17–35. http://dx.doi.org/10.2989/1814232X.2011.572334
6. Starfield AM, Jarre A. Interdisciplinary modelling for an ecosystem approach to management in marine social-ecological systems. In: Ommer RE, Perry RI, Cochrane K, Cury P, editors. World fisheries: A social-ecological analysis. Oxford: Wiley-Blackwell; 2011. p. 105–119. http://dx.doi.org/10.1002/9781444392241.ch6
7. Owen-Smith N. Accommodating environmental variation in population models: Metaphysiological biomass loss accounting. J Anim Ecol. 2011;80:731–741. http://dx.doi.org/10.1111/j.1365-2656.2011.01820.x
8. Richardson FD, Hoffman MT, Gillson L. Modelling the complex dynamics of vegetation, livestock and rainfall in a semi-arid rangeland in South Africa. Afr J Range For Sci. 2010;27:125–142. http://dx.doi.org/10.2989/10220119.2010.520676
9. Mathematics in Industry Study Group 2013 [homepage on the Internet]. c2013 [cited 2014 May 13]. Available from: http://www.wits.ac.za/conferences/misgsa2013.
10. Herbst RS, Qadir A, Momoniat E. The effects of mass on the radiation of a relativistically rotating neutron star. New Astronomy. 2013;25:38–44. http://dx.doi.org/10.1016/j.newast.2013.03.010
11. Breward C, Hocking G, Ockendon H, Please C, Schwendeman D. Modelling the extraction of sugar from sugar cane in a diffuser. Proceedings of the Mathematics in Industry Study Group 2012. Johannesburg: University of the Witwatersrand; 2012. p. 31–57.
12. Fowkes ND, Hocking G, Please CP, Mason DP, Kgatle MR. Rising water table and seismicity. Proceedings of the Mathematics in Industry Study Group 2012. Johannesburg: University of the Witwatersrand; 2012 p. 1–19.

Anything but simple: Inappropriate use of Euclidean distance in Govinder et al. (2013)

AUTHOR:
Victor M.H. Borden[1]

AFFILIATION:
[1]Department of Educational Leadership and Policy Studies, Indiana University, Bloomington, Indiana, USA

CORRESPONDENCE TO:
Victor Borden

EMAIL:
vborden@iu.edu

POSTAL ADDRESS:
Department of Educational Leadership and Policy Studies, Indiana University, 201 N. Rose Avenue, Bloomington, IN 47405-1006, USA

KEYWORDS:
Equity Index; higher education; South Africa; benchmarking

HOW TO CITE:
Borden VMH. Anything but simple: Inappropriate use of Euclidean distance in Govinder et al. (2013). S Afr J Sci. 2014;110(5/6), Art. #a0065, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0065

The ‘Equity Index’ (EI) introduced by Govinder et al.[1] has stimulated critiques addressing a variety of flaws in the use of this allegedly ‘simple and objective’ measure of racial and gender equity among South African higher education institutions. Dunne[2] noted that the use of a mathematical formula and the resultant numerical result provides a false sense of validity and precision. He further described in great technical detail why measures of distance are not as simple as they may seem when portrayed, for explanatory purposes, as the distance between points in a two-dimensional space. Dunne also addresses several issues of substantive validity including the stochastic nature of social measures for which dynamic probabilistic models are required as compared to the mathematical models that serve physical phenomena like measuring distance between objects in space. Moultrie and Dorrington[3] extend this critique, examining other mathematical (double counting) and conceptual (suitability of benchmark) problems.

As a long-time institutional research practitioner within the US context, I was intrigued by the publication of the index and the ensuing critiques as they touch upon the long-standing institutional research practices of peer institution benchmarking.[4-6] Because of the diversity of the US higher education landscape, with over 7000 post-secondary institutions ranging from for-profit, single programme vocational institutions and 2-year community colleges to 4-year regional and comprehensive universities and both public and private research universities, it is not common for us to think of a single measure that can be applied equally to all institutions, or even to those that are internationally competitive for students and staff. Because of this complexity, we are well versed in comparing institutions across a variety of measures and dimensions, including the demographic and academic profile of students, the mix of academic programmes, the types of instructional and non-instructional staff, and revenue sources and expenditure targets. One thing we have learned from this vast experience is that there is no such thing as either a simple or objective measure of institutions in relation to a target (whether that be another institution or a regional or national benchmark).

In the remainder of this critique, I will illustrate the lack of reliability (and therefore questionable validity) of employing a Euclidean distance measure on the concatenated distribution of two sets of proportions (race and gender). Rather than explore the mathematical and technical dimensions of these problems, I will illustrate how the comparison of the 23 South African higher education institutions changes depending on what type of distance measure is used and whether it is used on race and gender separately or combined.

When comparing the ‘position’ of an institution relative to other institutions or to criterion benchmarks like the national representation among racial and gender groups, one must take into account the scale characteristics of the measurement variables (nominal, ordinal, interval, ratio), as well as the statistical relationship (association) among the variables. If one is simply considering race and gender as distinct variables, then it may be suitable to describe these as independent measures (the likelihood of being male or female is not contingent, at least conceptually, on the racial group). However, when the values of a proportional representation variable are portrayed as the values upon which comparisons are based, then, as Moultrie and Dorrington pointed out, there is redundancy. That is, the percentage of males is linearly dependent on the percentage of females (percentage males = 100 – percentage females). Thus, the values of the variable gender have only one degree of freedom. Moreover, as race entails four categories and gender two, if we assume equal probability of each category, the race factor has three times the weight in the characterisation of the position (because race is four groups, there are three degrees of freedom, compared to one for gender). However, race is not uniformly distributed (that is, the general probability for each category is not one divided by the number of categories), so one must take into account the non-linear qualities of proportions across the range values. More prosaically, a 5% point difference has different substantive meaning when an event is rare (e.g. 5%), or more common (e.g. 60%).

There is a wide variety of ways to calculate similarity or difference for use in a positioning (nearest neighbour) analysis. Even if one would like to use a Euclidean-based measure, there are several to choose from. Govinder et al. use the ‘RSSD’ version, that is, the root of the sum of squared differences. If the variables are on notably different scales in terms of variation, it is advisable to first transform the measures to their standardised form (value minus mean, divided by standard deviation). When using percentages, the Chord form of Euclidean distance is recommended, where the values are first subject to a square root transformation. There are several derivatives of the Euclidean form, such as a City Block metric and Minkowski metric that vary the root to which the difference between coordinate points is raised. In addition to Euclidean-based measures, there are correlation-based distance measures (Pearson and Spearman) and the Mahalanobis measure, which takes into account both Euclidean distance and covariance among the variables.

Tables 1 and 2 demonstrate how the calculated distance value and the rank of the 23 South African higher education institutions change depending on which proximity measure is used to calculate the distance from the national benchmark. For these tables, the benchmarks were taken from the Govinder et al. article and the proportions of enrolled students from the Department of Higher Education and Training document, *Statistics on Post-School Education and Training in South Africa: 2011*[7]. The first three proximity measures included in Table 1 are three forms of the Euclidean distance: the RSSD version used by Govinder et al., one based on standardised values for each proportion, and the ‘Chord’ version, which is based on a square root transformation of the original values. In addition, the table shows the results using the Mahalanobis metric, which incorporates the covariance between the variables, and a measure based on the Pearson correlation, which has been reversed (Pearson values range from 1 for the most similar to 0 for the least similar, so the calculated value is subtracted from 1) and multiplied by 1000 to represent the value in integer digits. The rightmost columns of the tables show the rankings among the 23 institutions of the corresponding calculated values.

Table 1 exhibits these various distance measures for the combined race and gender proportions as employed by Govinder et al.[1] The reader is reminded that there are several technical reasons why it is not appropriate to combine these proportions into a single estimation of distance, as noted in the critiques of Dunne[2] and Moultrie and Dorrington[3]. Some of the ramifications for the inappropriateness of doing so are manifest in the variation of calculated distance values and rank in these tables. For example, the Central University of Technology, ranked 2nd using the RSSD calculation, is ranked 11th using the Mahalanobis measure. Durban University of Technology varies considerably by the four measures, as high as 5th using the RSSD and as low as 17th using the Mahalanobis metric.

Table 2 uses the same five measures on the four categories of race. While not suggesting that examining race alone establishes evidence of equity, the benchmarking of distance from the national norms is a cleaner measurement concept than when incorporating race and gender into a single measure. Although the rankings for race alone are not as varied as they are for race and gender combined, they still vary considerably. For example, University of Johannesburg, which is ranked 1st by four measures, is ranked 10th using the Pearson correlation measure. It is also interesting to note that the Chord version of the Euclidean measure, which is generally recommended over RSSD for percentage measures, varies considerably from the RSSD measure.

Establishing equity

Although it is not without controversy, it is instructive to consider how equity is established in other, long-standing methodologies. For example, the US Department of Labor's Office of Federal Contract Compliance, has required since the early 1970s that organisations and businesses that obtain federal contracts establish the equity in both hiring and compensation of their workforce. The compliance requirements revolve around 'labour-market availability' within job groups that are defined

Table 1: Comparison of five distance measures using both race and gender percentages benchmarked against national norms

Institution	Calculated distance value					Rank of distance value				
	Euclidean			Mahal-anobis	Pearson correlation	Euclidean			Mahal-anobis	Pearson correlation
	RSSD	Std	Chord			RSSD	Std	Chord		
University of Johannesburg	9	112	15	25	9	1	2	1	1	3
Central University of Technology, Free State	12	144	17	28	7	2	4	2	11	2
Tshwane University of Technology	17	106	25	26	6	3	1	5	6	1
University of South Africa	17	277	19	27	35	4	16	3	8	12
Durban University of Technology	18	232	33	30	34	5	13	12	17	11
Nelson Mandela Metropolitan University	22	128	22	27	45	6	3	4	9	13
University of Fort Hare	23	192	27	25	11	7	8	7	3	4
Vaal University of Technology	24	207	29	29	14	8	10	8	14	9
University of the Free State	25	248	26	27	73	9	14	6	10	14
University of Limpopo	26	154	37	25	12	10	6	13	2	5
University of Witwatersrand	28	219	32	28	76	11	11	11	12	15
Mangosuthu University of Technology	28	152	45	26	13	12	5	19	7	7
Walter Sisulu University	28	193	42	25	14	13	9	17	4	8
University of Venda	28	159	48	26	12	14	7	22	5	6
University of KwaZulu-Natal	31	426	40	38	113	15	22	16	22	16
North-West University	31	417	30	31	115	16	20	9	19	17
Cape Peninsula University of Technology	33	225	31	32	118	17	12	10	20	18
University of Zululand	33	394	44	29	30	18	19	18	16	10
University of Pretoria	41	283	39	29	209	19	18	15	15	19
Rhodes University	41	281	37	29	210	20	17	14	13	20
University of Western Cape	53	431	47	42	379	21	23	21	23	21
University of Cape Town	54	275	46	31	426	22	15	20	18	22
University of Stellenbosch	84	419	71	38	915	23	21	23	21	23

Source: Republic of South Africa Department of Higher Education and Training[7]

Table 2: Comparison of five distance measures using only race percentages benchmarked against national norms

Institution	Calculated distance value					Rank of distance value				
	Euclidean			Mahal-anobis	Pearson correlation	Euclidean			Mahal-anobis	Pearson correlation
	RSSD	Std	Chord			RSSD	Std	Chord		
University of Johannesburg	8	65	15	25	8.1	1	1	1	1	10
Central University of Technology, Free State	11	73	17	26	1.6	2	2	3	3	1
University of South Africa	11	90	16	25	11.8	3	3	2	2	11
Durban University of Technology	17	222	32	29	39.5	4	18	12	19	12
Tshwane University of Technology	17	105	25	26	2.6	5	4	6	6	2
University of Fort Hare	21	117	27	26	3.0	6	6	7	8	3
Vaal University of Technology	22	121	28	26	3.1	7	7	9	9	4
Nelson Mandela Metropolitan University	22	116	22	26	45.3	8	5	4	4	13
University of the Free State	22	124	24	26	69.1	9	8	5	5	14
North-West University	24	135	27	26	80.7	10	9	8	7	16
University of Limpopo	26	139	37	26	4.8	11	10	14	10	5
University of Zululand	27	145	42	27	5.2	12	11	17	11	9
Walter Sisulu University	27	146	42	27	5.2	13	12	18	12	8
University of Witwatersrand	28	211	32	28	79.2	14	17	11	16	15
Mangosuthu University of Technology	28	150	45	27	5.0	15	13	19	13	7
University of Venda	28	151	48	27	5.0	16	14	22	14	6
University of KwaZulu-Natal	29	382	40	38	123.0	17	22	16	22	17
Cape Peninsula University of Technology	33	223	31	30	136.4	18	19	10	20	18
University of Pretoria	39	211	39	28	250.2	19	16	15	17	19
Rhodes University	39	207	37	27	250.6	20	15	13	15	20
University of Western Cape	52	371	46	40	479.5	21	21	21	23	21
University of Cape Town	54	275	46	28	584.7	22	20	20	18	22
University of Stellenbosch	84	418	71	34	1168.8	23	23	23	21	23

Source: Republic of South Africa Department of Higher Education and Training[7]

according to the wages, job duties and responsibilities, and training requirements. Specifically, the requirements (http://www.dol.gov/ofccp/scaap.htm) note[8]:

> *…federal contractors must conduct availability analyses to determine the percentage of women and minorities who have the skills required to perform the jobs within each job group… Availability involves calculation of minorities and women who are 'available' to work in the job from both external sources (i.e., hired from outside the company) and internal sources (e.g., transfer or promotion of existing employee in the company)…For calculating 'external' availability, you want to consider who is qualified for the job within 'the reasonable recruitment area' for that job. The 'reasonable recruitment area' represents the area from which a contractor usually seeks or reasonably could seek workers for a particular job group.*

Assessing equity in academic programmes can be considered as analogous. To be admitted to an academic programme, students must meet certain basic requirements, such as having completed a secondary education credential and having basic skills suited to a specific programme of study (for example, higher order math skills for engineering and higher order writing skills for communications). Students must also live within commuting distance (except perhaps for UNISA). Comparing proportions of women and racial groups enrolled at a particular university to a generic national benchmark masks all of the availability issues, which are at the root of establishing equity. Throughout my 30 years of experience in using evidence and analysis to address educational access issues, I have found that it is far more constructive to confront directly and as complexly as possible the root causes of inequity, such as those revealed through the many aspects of 'availability'. Conversely, reducing to a single measure such complex phenomena tends to shift attention away from the root causes and can be used by various groups and individuals to absolve the responsibility that we all share in addressing such issues. Establishing equity is anything but simple.

References

1. Govinder KS, Makgoba MW. An Equity Index for South Africa. S Afr J Sci. 2013;109(5/6), Art. #a0020, 2 pages. http://dx.doi.org/10.1590/sajs.2013/a0020

2. Dunne T. Mathematical errors, smoke and mirrors in pursuit of an illusion: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0047, 6 pages. http://dx.doi.org/10.1590/sajs.2014/a0047

3. Moultrie TA, Dorrington RE. Flaws in the approach and application of the Equity Index: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0049, 5 pages. http://dx.doi.org/10.1590/sajs.2014/a0049

4. James GW. Developing institutional comparisons. In: Howard RD, McLaughlin GW, Knight WE, editors. The handbook of instituitonal research. San Francisco, CA: Jossey-Bass; 2012. p. 644–655.

5. Terinzini PT, Hartmark L, L'Orange Jr. WG, Shirley RC. A conceptual and methodological approach to the identification of peer institutions. Res High Educ. 1980;12(4):347–364.

6. McCormick AC, Zhao CM. Rethinking and reframing the Carnegie classification. Change. 2005;37(5):51–57.

7. Department of Higher Education and Training, Republic of South Africa. Statistics on post-school education and training in South Africa. Pretoria: Department of Higher Education and Training; 2011.

8. US Department of Labor. Office of Federal Contract Compliance Programs [homepage on the Internet]. No date [cited 2014 May 09]. Available from: http://www.dol.gov/ofccp/scaap.htm

Further response to Govinder et al. (2014): Flaws in the Equity Index

AUTHORS:
Tom A. Moultrie[1]
Rob E. Dorrington[1]

AFFILIATION:
[1]Centre for Actuarial Research (CARe), University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Tom Moultrie

EMAIL:
tom.moultrie@uct.ac.za

POSTAL ADDRESS:
Centre for Actuarial Research,
University of Cape Town, Private
Bag X3, Rondebosch 7701,
South Africa

KEYWORDS:
transformation; demographics; equity profile; Euclidean distance formula; higher education; South Africa

HOW TO CITE:
Moultrie TA, Dorrington RE. Further response to Govinder et al. (2014): Flaws in the Equity Index. S Afr J Sci. 2014;110(5/6), Art. #a0069, 2 pages. http://dx.doi.org/10.1590/sajs.2014/a0069

We note the response of Govinder et al.[1] to our comments[2] on an earlier article of theirs[3] on the proposed Equity Index (EI) to measure the transformation of higher education in South Africa. Despite their attempts to allay concerns about the intrinsic weaknesses of the EI, many concerns remain. The purpose of this brief response is not to attempt to re-argue each and every point – we leave most to readers to decide – but to highlight some of the more fundamental issues with which the authors failed to deal adequately.

On the mathematics of the Equity Index

Govinder et al. claim that the mathematics of the EI is not and has not been shown to be flawed.[1] This is simply not true. We have pointed out that the formulation preferred by Govinder et al. is not only mathematically incorrect, but also that it leads to double counting. Apart from this, as pointed out by others[4], the underlying logic of using a Euclidean distance as a measure of this sort is both inappropriate and wrong. The authors seem to think that mathematical criticism can be dismissed by simply saying it is not so. And it is this point, above all others, that has not been adequately addressed.

Secondly, the authors' understanding of distance – *vide* the assertion in their response that 'EIs are calculated using the distance formula, mathematically they can indeed be added together or subtracted from each other (as can distances in general)' – is quaint. If one takes Johannesburg as the origin, and one notes that the distance from Johannesburg to Bloemfontein is 400 km, while that from Johannesburg to Durban is 600 km, nothing can be inferred from this information about the distance from Bloemfontein to Durban (650 km). Distances can only be added or subtracted if all relevant points are located on the same straight line passing through the chosen origin. Nonetheless, the authors pursue a red herring by insinuating that we claimed that Table 5 of their paper demonstrated their addition of EIs. On the contrary, we never refer to Table 5, but to Figure 2, which does indeed present the EIs for institutions as being the sum of the respective staffing component EIs. Poor presentation of the data in Figure 2 cannot be attributed to the vagaries of 'greyscale' as the authors suggest. The contradiction between the assertion above that the EIs can be added, and their vigorous (to the point of misrepresenting our argument) defence of their *not* having added EIs in Table 5 is puzzling.

Thirdly, although as we pointed out, and as was repeated by the authors[1], the relative ranking of the higher education institutions (HEIs) does not change when using the correct version of the EI, the authors quietly ignore the point that the range and hence the designation of objective quadrants do change.

Finally, the authors confuse the undoubted utility of being able to decompose the EI by race and sex with mathematical correctness. The fact that on their – mathematically incorrect – version, the EI can be decomposed offers no proof of its mathematical validity, and a 'useful' result is not actually useful if it is premised on flawed logic.

On transformation and legislation

Govinder et al. claim that the EI is 'based on the Constitution' of South Africa. This is trite. The Constitution requires all South Africans to work towards the transformation of our society, with transformation seen as representing the process of change, rather than a defined end-point. We certainly do not contest that transformation is a national imperative or argue that HEIs are exempt from the broader need to transform South African society.

However, labour legislation in South Africa, which also has to be consistent with the Constitution, protects the rights of employed workers. All else being equal, transformation can be expected to be slower in institutions where staff turnover is low. It is not clear whether the authors are proposing raising staff turnover rates (perhaps through making working conditions unpleasant, or perhaps paying people to take on a form of employment that is not counted in the EI) in order to bring about transformation faster. But even if this were the case, the authors seem to have adopted a peculiarly ahistorical stance on the issue of the staffing of South Africa's universities. In our initial comments, we demonstrated (in our Figure 1) that it is likely that much of the poor pace of transformation in South African HEIs might be rooted in South Africa's past. While this is in no way a justification for future lack of transformation, it is naïve to disregard the pernicious legacy of apartheid in terms of who got access to higher education; who got higher degrees; and therefore who is currently qualified to be employed as academic staff at HEIs in South Africa. As we noted, the distribution by race and sex of the South African population aged 25–64 with higher degrees is overwhelmingly skewed towards white South Africans. While this must change, and is changing, it is not helpful during this change to blame the slow pace of demographic change of HEIs on these institutions alone or even on the sector as a whole.

On the data and other aspects of the Equity Index

Govinder et al. suggest that the issue of inclusion of foreigners is a matter that is not yet settled. But as we noted in our Commentary, the germane issue is that the *Employment Equity Act* defines the classes of employee for transformation purposes more specifically than that currently captured by demographic data in a census or on the HEMIS database.

Apart from this, Govinder et al. have misused the census data, treating those with unstated race in the census as foreigners (which most are not) and those born outside South Africa (with stated race) as being South African. However, until the data collection in the census and the HEMIS data is changed to permit the identification of foreigners (as required by EE legislation), the utility of the EI is highly questionable. In addition, in order to compare

EI and ‘quality’, the data on research output (or any other measure of output, such as numbers of graduates) would also have to be equivalently classified, which they are not. Finally, although in society as a whole, foreign-born residents may be comparatively rare, in some HEIs they can constitute a significant minority of staff. Any system that rewards not employing foreign academics implicitly would seem to weight a narrow parochialism and nationalism more highly than the transfer or production of knowledge.

Conclusions

In spite of the length of their rebuttal, readers of this correspondence will note that the authors have largely failed to engage with the major theoretical considerations raised in the commentaries. We re-iterate that the mathematics of their EI is incorrect. We, with other commentators, also repeat that demographic transformation is but one of many aspects of transformation. The danger of an index such as this is that those other aspects are sidelined. Our approach would be to be more inclusive – to build a ‘dashboard of indicators’ of the multiple dimensions of transformation that can be appraised simultaneously, rather than privileging one (supposedly quantifiable) aspect of transformation over those which are less analytically tractable, such as the apparent tension between high research output and being less transformed, or the role of universities in serving the developmental needs of a society at a critical juncture, or the need to balance quality of education against quantity of education, or the need to massify higher education. It is our view that the EI as it currently stands adds little value to the debate on the transformation of higher education in South Africa. Much more careful and rigorous thought is required before such metrics should be applied.

References

1. Govinder KS, Zondo NP, Makgoba MW. Taking the transformation discourse forward: A response to Cloete, Dunne and Moultrie and Dorrington. S Afr J Sci.2014;110(3/4), Art. #a0060, 8 pages. http://dx.doi.org/10.1590/sajs.2014/a0060

2. Moultrie TA, Dorrington RE. Flaws in the approach and application of the Equity Index: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0049, 5 pages. http://dx.doi.org/10.1590/sajs.2014/a0049

3. Govinder KS, Zondo NP, Makgoba MW. A new look at demographic transformation for universities in South Africa. S Afr J Sci. 2013;109(11/12), Art. #2013-0163, 11 pages. http://dx.doi.org/10.1590/sajs.2013/20130163

4. Dunne T. Mathematical errors, smoke and mirrors in pursuit of an illusion: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0047, 6 pages. http://dx.doi.org/10.1590/sajs.2014/a0047

On taking the transformation discourse for a ride: Rejoinder to a response (Govinder et al. 2014)

AUTHOR:
Tim Dunne[1]

AFFILIATION:
[1]Department of Statistical Sciences, University of Cape Town, Cape Town, South Africa

CORRESPONDENCE TO:
Tim Dunne

EMAIL:
tim.dunne@uct.ac.za

POSTAL ADDRESS:
Department of Statistical Sciences, PD Hahn Building, Upper Campus, University of Cape Town, Rhodes Gift 7707, South Africa

KEYWORDS:
demographic divergence; measurement; percentages

HOW TO CITE:
Dunne T. On taking the transformation discourse for a ride: Rejoinder to a response (Govinder et al. 2014). S Afr J Sci. 2014;110(5/6), Art. #a0068, 4 pages. http://dx.doi.org/10.1590/sajs.2014/a0068

This submission explicitly clarifies the fundamental mathematical and logical errors and conceptual inadequacies of three recent contributions by Govinder et al.[1-3], published in the *South African Journal of Science*.

Scientific foundations

Ordination is an act of aggregating pairwise comparisons on a set of elements (objects, persons, events, phenomena) in such a way that the ordering of the elements becomes meaningful. The comparisons are made on the basis of some single composite characteristic of each element. For example, we may specify a notion of extent or size, and then impose a linear ordering on any relevant set of objects which admit comparisons with respect to that characteristic.

The ordering guarantees to the observer that each object (by our chosen criterion) is larger than every object , say to its right, but smaller than every object to its left. We rank elements or objects, but permit tied ranks. This simple device of ranking is profoundly important, and has huge advantages in every sphere of human activity.

The familiar versions of size and extent are counts and measures. These versions are, however, profoundly different in nature and in consequentiality.

For any collection of elements, we may specify a binary property which is either present or absent in each element, and count the number of times the property is present. Effectively, we assign each element a zero for absence and one for presence, and then add all the digits. Count is a form of number that is relevant in handling categorical information.

Measurements are very different from counts, and far more complex, even though we perform measurements routinely in many aspects of everyday life. Measurements address quantity, while counts address multiplicity.

All measurements are fundamentally comparisons, but of a particular kind. We construct measurements by first defining a suitable convenient unit of extent for a particular characteristic of the elements under consideration. For example, we may use grams for mass, metres for lengths and seconds for time. Alternative choices of units are available, such as tonnes, kilometres, hours.

Then we establish or construct a mechanism or device by which a comparison of the extent of an element (e.g. its mass) with the unit of extent of mass (e.g. a kilogram) can be made. This comparison is multiplicative and is represented as a fixed ratio between the two masses. The ratio is fixed in the sense that repeated applications of the mechanism on a particular element will preserve the ratio. The mechanism also has to be reliable, and often we are able to construct an instrument which is able to reproduce the ratio virtually unchanged on repeated applications to the same object or element.

The achievement of science is that such mechanisms have been discovered, and suitable devices have been invented and brought into everyday use. In these discoveries and inventions science has had to read and mimic the architecture of the physical universe and copy the choreography of its intimate laws of cause and effect.

Measuring a characteristic is the result of allowing a mechanism and a unit to operate on an element with that characteristic (inter alia), in order to obtain a value for the fixed ratio, attributable to the element. The number we obtain, coupled with an explicit reference to the unit involved in the mechanism, is the measurement. Thus we might report measurements as 2.59 kg or 4.257 metres.

The ratio is a pure number, expressed to a convenient number of decimal places. For example, a beam of wood, if correctly labelled as having a length of 2 centimetres, should be twice as long as the unit 1 centimetre. Our gauges of the beam width should register 2.0 on any centimetre ruler and 20 on any millimetre ruler, applied to the appropriate part of the beam.

The ratio can only be consistent to a degree determined by the limitations of the instrument we have constructed as the mechanism for obtaining measurements. Thus we allow rounding conventions to limit the number of decimal places to which an instrument is required to preserve the ratio. The desired number of places required is a matter of context and then convenience. The bathroom scale reporting personal body mass in kilograms to one decimal place is adequate for that purpose, but we do not use the same scale for laboratory purposes.

Derived units

We note that besides fundamental units of length, mass, time and arc, we construct derived units for properties such as rate, acceleration, density, momentum and force. Derived units are composite quantities that involve further ratio and product relationships between two or more of the fundamental quantifiable characteristics. The corresponding derived measurements are ratios and products of the observable measurements of distinct fundamental features of the elements or objects in question. A multiplicative arithmetic underlies derived units and measurements of derived variables.

Derived measurements generally involve some degree of decimal fraction. Again, context and convenience give rise to choices of number of decimal places required and of rounding conventions. The decimal place has no independent utility or power. It arises from measurement but it does not in any way constitute measurement.

This remarkable architecture is not an outcome of arithmetic. The spectacular fact is that science has discovered and explored this architecture. Science confirms that the physical universe and its multifold processes are all governed by these elegant regularities. Consequently, our simple arithmetic permits us to harness these regularities into artefacts, machines, processes and constructions that make life more predictable and more comfortable.

The power of measurement lies in the fact that, when combined with insights into the physical world, it permits us to use the laws of cause and effect to reliably predict and engineer desirable outcomes. This benefit is seductive, and it accounts for the pervasive compulsive urge to measure in the social sciences.

Social sciences

The quest for measurement in the social sciences is, however, an illusion. The uncountable variables of social science, whether latent or manifest, whether notional, constructed or real, do not admit units and ratio comparisons. They do not have meaningful zero origins. They permit pairwise orderings of objective and subjective kinds, both of which can be meaningful.

These limitations are not faults in the fabric of the social sciences. They constitute admissions of the complexity of social sciences, where regularity of any kind is necessarily encased and pierced by the confining and transcending effects of human perception and agency, by belief, hope, will, endurance, insight, creativity and ethical imperatives. This list of roots of actions and outcomes in human society is only illustrative, and not intended to be exhaustive.

Nonetheless, in the various social sciences, it is now possible to create subjective instruments that pass peer scrutiny for validity as fit for purpose, but also have some measurement-like characteristics. These instruments involve a designed set of items and associated item scoring systems, for which the total score is a sufficient summary of a proficiency or a degree of attribute.

These test scores have stochastic rather than deterministic interpretations, and are inherently relative orderings rather than absolute measurements. The interpretations have complex relationships with percentages of observations, under specifiable conditions.

In the current debate, the conditions for measurement-like interpretations are not satisfied by the demographic percentages of any population, nor by any mathematical formulae derived from any versions of Govinder et al.'s demographic divergence index (DDI) or its possible generalisations.

Percentage

Percentage is a relational construct characterised by the comparison of the sizes of two objects or elements of interest. One element is declared to be the referent object. Effectively, a ratio comparison is invoked, a number with a decimal fraction is obtained, and then this number is converted to a new type of number called a percentage, through multiplication by 100. To the observer, it will appear that the decimal place in a ratio has been moved two places to the right in the string of digits, and the sign % appended. Note that the ratio is not a unit, but is just a number. Its conversion to percentage and the appending of % does not mean that there is now a unit called % in operation.

It is crucial also to note that the comparison is relative to a particular choice of referent in each context, and not to any specific absolute unit across contexts. The presence of decimal places does not warrant regarding percentage and measurement as equivalent.

Percentage notions are often invoked to describe part-whole relationships. In these cases, the whole is the referent and all the percentages of parts will be reported as between 0% and 100%. This operation may always be applied to part-whole ratio comparisons of measurements, but the result is just a number, not a measurement.

In contrast, in everyday life, percentage is most familiar in the reporting of counts relative to an overall total count. For simplicity of report and communication, a convention of using only distinct or disjoint or mutually exclusive parts of the referent whole, ensures that percentages associated with the parts sum to 100%. This convention avoids double and multiple inclusion induced by any overlapping parts of the whole.

Comparisons of percentages within a common referent context are validly and easily made. It is also admissible to define desired percentages and assess the profile of differences between sets of observed and desired percentages, part by part. The pertinence and utility of any desired percentages is matter for debate. Desired percentages of total counts do not always generate desired integer counts, but often imply desired counts with decimal fractions. The so-called chi-square statistic is recommended for assessing the goodness-of-fit between desired and observed profiles. It takes the decimal fractions and the total count into account. This divergence from perfect fit is obtained by the formula

$$\sum_{i=1}^{i=k} \frac{(O_i - E_i)^2}{E_i}$$

The formula first squares the differences between observed counts (O_i) and expected counts (E_i), and divides each square by the corresponding expected count, for each of k parts. It then sums all k of those relative terms to give an overall divergence value. This divergence index is available in any first-year statistics text.

For perfect agreement between observed and expected count values, the divergence statistic will be zero. When observed and expected count values are relatively close to one another the divergence statistic value is a small positive number. However, the divergence value will be dramatically large when the observed and expected counts differ substantially. Tables exist to inform judgements about the size of the goodness-of-fit statistics.

A large value serves to signal that the observed count differs from the expected count, which may require explanation, understanding and decision-making. The subsequent enquiry should involve both the beliefs which generated the expected counts, and the adequacy with which they can be associated with the context.

It is admissible, but not necessarily relevant or important, to contrast observed counts with any set of expectations. One such set of expected values might be derived from a population profile. We may note that the chi-square statistic is a pure counting number for the objects in the population under observation.

Divergence indices

A whole family of divergence indices for pairs of k-dimensional vectors of numerical values is offered by the Lp-norms:

$$\sqrt[p]{\sum_{i=1}^{i=k} \left|O_i - E_i\right|^p}$$

where p is a positive real number. This family has been widely studied in the literature. The index has clear meanings for particular values, specifically $p = 0, 1, 2$ or ∞.

The special case of $p=2$ includes the divergence index that Govinder et al. advance for percentages and proportions. This application implies the index has to operate on a specific hyperplane on which proportions sum to 1 or, equivalently, percentages sum to 100%.

However, Govinder et al. prefer, advise and fiercely defend one logically flawed special case with $k=6$, for four ethnic or historical race categories and two genders. They purport to be unconcerned that the percentages sum to 200%. The approach of Moultrie and Dorrington[4] is marginalised as an exercise of preference for $k=8$, rather than acknowledged as the only correct approach for race and gender.

Incoherence of approach

To exemplify Govinder et al.'s fallacy, we may approximate a South African population into groups B, C, W and I, with corresponding percentages 76%, 10%, 10% and 4% and gender groups M and F of percentages equal at 50% and 50%. These values are chosen only for simplicity, and more exact census count percentages can be invoked with the same general outcome.

In the intended Govinder et al. policy, any university with a matching profile to the population will be declared to have attained the ideal and constitutional transformation, by virtue of zero differences on these six percentages.

In Table 1, the desired Govinder et al.'s structure is indicated by the two common row and four common column percentages in all six tables. We may demonstrate the fallacy of Govinder et al.'s approach by exhibiting that amongst its immediate outcomes and implications, there are absurdities which defeat the misnamed purpose for which their DDI is intended.

The body of the table is unspecified in Table 1a. The letters *a*, *b* and *c* in Table 1b indicate three areas of arbitrariness in percentages associated with Govinder et al.'s specification. Here we may have three choices satisfying inequalities $0 \leq a \leq 10$, $0 \leq b \leq 10$ and $0 \leq c \leq 4$.

Table 1c to 1f present four possible scenarios, each of which generates a Govinder et al. DDI index value of zero, and hence the Govinder et al. transformation 'jackpot' is attained. Technically there would be an infinite number of distinct race by gender profiles for universities, all of which would be compliant with Govinder et al.'s race criterion and gender criterion for perfect transformation. Only four of these perfect profiles are presented here (Table 1c–1f).

Closer inspection will show that only Table 1f satisfies a condition of independence of race and gender in the pseudo-university population. Race and gender would be mutually uninformative characteristics for a randomly chosen member of the pseudo-university associated with Table 1f. We may note that this possibly attractive inference about internal composition of pseudo-university profile is distinct from notions that the eight percentages in the body of the table mirror underlying South African population percentage realities for race and gender composition.

We now venture to construe the body of any table solely to university selection procedures based on race and gender alone, as per the parody of selection bias offered or imputed by Govinder et al. Then to which of the four universities would a black woman believe her chance of selection would be best, and where would she presume her selection is least likely? Clearly her first inferences would be for 1c as her highest and 1e as her least chance.

Ironically, a white man will interpret and make the same first qualitative inference for his own chances. Neither of these two selection candidates will perceive these four pseudo-universities as equivalent in their equity characteristics. If they have to choose only one university, they might apply to 1c.

In these apparently rational choices there are hidden assumptions: all other things being equal. We may render explicit several pieces of unknown information which have the potential to affect their decisions radically.

Such emerging facts include the size of the universities, the race and gender composition of the applicant cohorts at each university and the rationale offered for the varied forms of glaring inequity apparent despite Govinder et al.'s DDI perfect zero values. Knowing that the university in 1e was 10 times larger than its counterpart in 1c might lead both the applicants to reverse their preferences and prefer 1e over 1c, to maximise the probability of success in selection.

Likewise, information about the underlying composition of the applicant groups from which the pseudo-university admission processes selected to yield these profiles, will affect the interpretation of the two persons exploring their options from the various perfect equity universities of Govinder et al.

The question is simple: will any of these scenarios of perfect Govinder et al. compliance stand up to any rational tests of desirability under the Constitution to which Govinder et al. have made appeal. Universities have many simultaneous obligations and purposes for service to society which require many specificities beyond only race and gender.

Subspaces

The sense in which the Govinder et al. index decomposes into subspace indices of any kind, as claimed in the extract below, is simply mathematically false. Deeming race or gender as spaces or subspaces violates mathematical convention. It is therefore also erroneous to claim that (squared) distances are additive across subspaces.

> *The advantage of our approach is that one can actually determine a race EI (using data points with only four (race) components), EI_r, a gender EI (using data points with only two (gender) components), EI_g, and an overall EI which is a combination of these two previous EIs via $EI = sqrt(EI_r^2 + EI_g^2)$. It was this aggregate combination possibility that led us to use $n=6$. Importantly, the EI_r and EI_g both form subspaces of the EI space using this approach. For $n=8$, we do not have this mathematical structure – one cannot find EI_r and EI_g directly in that space.*[3]

Constitution

The Constitution of the Republic of South Africa, *Act No. 108 of 1996*, as amended up to 2003, has a Bill of Rights at Chapter 2, sections 7 through 37. This set of rights sets out at section 9(3) obligations of the state to ensure no unfair discrimination, directly or indirectly against any person on the basis of any one or more grounds including race, gender, sex, pregnancy, marital status, ethnic or social origin, colour, sexual

Table 1: (a) to (f) Structures of pseudo-university population percentages for distinct scenarios

a	B	C	W	I	
	76	**10**	**10**	**4**	
M					**50**
F					**50**

b	B	C	W	I	
	76	**10**	**10**	**4**	
		a	b	c	**50**
					50

c	B	C	W	I		
	76	**10**	**10**	**4**		
	26	10	10	4	**50**	**M**
	50	0	0	0	**50**	**F**

d	**76**	**10**	**10**	**4**	
M	38	10	0	2	**50**
F	38	0	10	2	**50**
	B	**C**	**W**	**I**	

e	**76**	**10**	**10**	**4**	
	48	0	2	0	**50**
	28	10	8	4	**50**
	B	**C**	**W**	**I**	

f	**76**	**10**	**10**	**4**		
	39	5	5	2	**50**	**M**
	39	5	5	2	**50**	**F**
	B	**C**	**W**	**I**		

orientation, age, disability, religion, conscience, belief, culture, language and birth. A caveat of fair discrimination is provided at section 9(5).

Govinder et al. offer no rationalisation for the suppression of all but two of the outlawed grounds of discrimination in their index.

Their insistence on the entire South African population of all ages takes no account of the narrower age ranges associated with university student attendance and university staff employment.

Moreover Govinder et al. do not allow other relevant criteria for plausibly fair discrimination, such as an appropriate National Senior Certificate outcome, any entry or placement requirements for students, and qualifications or experience for employees.

The adamant position of Govinder et al. is that they have measured inequity, and, on these grounds, have advocated a misnamed DDI as the basis for teaching the universities a lesson or two in policy compliance and budget implications.

Ironically, the word 'equity' appears only twice in the Constitution (section 29.2(a) in respect of basic education and section 29.3(e) in respect of powers of municipalities). The word 'transformation' does not appear, although progressive realisation is addressed. Nowhere does the Constitution invoke the population profile or demographics as criteria or artefacts of its law. This note does not imply that the Constitution abjures use of these terms.

The Constitution and its Bill of Rights are certainly admissible as warrants for imperatives within South African society. It is important to note that the Constitution demands much more than redress of past injustice and deprivation. It advocates a journey to a just society, and progressive realisation as one mode of the journey.

The Constitution does much more than advocate and require transformation. It is as much or even more concerned with the delivery of fundamental rights and efficient delivery of service within the public sector. Included in these imperatives is the preservation of existence and function of institutions for the public good, and their responsibility to keep engaging with transformation.

The Constitution permits error, but it does not require error in the apparent pursuit of its imperatives. The Constitution is fairly understood as requiring our best intentions as citizens in the service of the common agenda that it represents. In requiring good intentions, the Constitution does not advocate good intentions as a sufficient condition for suspending debate on matters of error.

The Constitution demands that we address inequality. It does not require us to presume we can measure inequality. Inequity is explored by making comparisons and answering probing and difficult questions.

It is an ethically fragile and morally dubious strategy to invoke a misguided claim of measurement as the final criterion of debate and decision-making.

Context

Govinder et al. argue that their index has been unfairly misinterpreted by Dunne[5] and debated out of the context into which their DDI was proposed as salient. The second claim is partly true, in a very weak sense.

The context they selected for its justification is such a narrow playing field that their index is virtually context free. Their approach advocates and requires race and perhaps gender as the only determinants of various university outcomes. Simultaneously, in contrast, the context selected by Govinder et al. for the application of their index is an entire gamut of social justice objectives of transformation, universally.

In particular, Govinder et al.'s DDI is proposed for policy monitoring and steering of universities towards claimed public imperatives. For this reason its logical and mathematical structure, its hidden assumptions, its claimed purpose and its fitness for purpose are all the proper focus of scientific enquiry and debate. Effectively, the various critiques have located Govinder et al.'s index in precisely the context to which its inventors seek to have it applied, and exhibited that it is hopelessly flawed.

References

1. Govinder KS, Makgoba MW. An Equity Index for South Africa. S Afr J Sci. 2013;109(5/6), Art. #a0020, 2 pages. http://dx.doi.org/10.1590/sajs.2013/a0020
2. Govinder KS, Zondo NP, Makgoba MW. A new look at demographic transformation for universities in South Africa. S Afr J Sci. 2013;109(11/12), Art. #2013-0163, 11 pages. http://dx.doi.org/10.1590/sajs.2013/20130163
3. Govinder KS, Zondo NP, Makgoba MW. Taking the transformation discourse forward: A response to Cloete, Dunne and Moultrie and Dorrington. S Afr J Sci. 2014;110(3/4), Art. #a0060, 8 pages. http://dx.doi. org/10.1590/sajs.2014/a0060
4. Moultrie TA, Dorrington RE. Flaws in the approach and application of the Equity Index: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0049, 5 pages. http://dx.doi.org/10.1590/sajs.2014/a0049
5. Dunne T. Mathematical errors, smoke and mirrors in pursuit of an illusion: Comments on Govinder et al. (2013). S Afr J Sci. 2014;110(1/2), Art. #a0047, 6 pages. http://dx.doi.org/10.1590/sajs.2014/a0047

Review of carbon dioxide capture and storage with relevance to the South African power sector

AUTHORS:
Khalid Osman[1]
Christophe Coquelet[2]
Deresh Ramjugernath[1]

AFFILIATIONS:
[1]School of Chemical Engineering, University of KwaZulu-Natal, Durban, South Africa
[2]MINES ParisTech, Centre Thermodynamic of Processes (CTP), Fontainebleau, France

CORRESPONDENCE TO:
Deresh Ramjugernath

EMAIL:
Ramjuger@ukzn.ac.za

POSTAL ADDRESS:
School of Chemical Engineering,
University of KwaZulu-Natal,
Howard College Campus, King
George V Avenue, Durban 4041,
South Africa

DATES:
Received: 17 June 2013
Revised: 01 Oct. 2013
Accepted: 22 Nov. 2013

KEYWORDS:
carbon dioxide; capture; storage; emissions; South Africa

HOW TO CITE:
Osman K, Coquelet C, Ramjugernath D. Review of carbon dioxide capture and storage with relevance to the South African power sector. S Afr J Sci. 2014;110(5/6), Art. #2013-0188, 12 pages. http://dx.doi.org/10.1590/sajs.2014/20130188

Carbon dioxide (CO_2) emissions and their association with climate change are currently a major discussion point in government and amongst the public at large in South Africa, especially because of the country's heavy reliance on fossil fuels for electricity production. Here we review the current situation regarding CO_2 emissions in the South African power generation sector, and potential process engineering solutions to reduce these emissions. Estimates of CO_2 emissions are presented, with the main sources of emissions identified and benchmarked to other countries. A promising mid-term solution for mitigation of high CO_2 emissions, known as CO_2 capture and storage, is reviewed. The various aspects of CO_2 capture and storage technology and techniques for CO_2 capture from pulverised coal power plants are discussed; these techniques include processes such as gas absorption, hydrate formation, cryogenic separation, membrane usage, sorbent usage, enzyme-based systems and metal organic frameworks. The latest power plant designs which optimise CO_2 capture are also discussed and include integrated gasification combined cycle, oxy-fuel combustion, integrated gasification steam cycle and chemical looping combustion. Each CO_2 capture technique and plant modification is presented in terms of the conceptual idea, the advantages and disadvantages, and the extent of development and applicability in a South African context. Lastly, CO_2 transportation, storage, and potential uses are also presented. The main conclusions of this review are that gas absorption using solvents is currently most applicable for CO_2 capture and that enhanced coal bed methane recovery could provide the best disposal route for CO_2 emissions mitigation in South Africa.

Introduction

There has been a nearly 100% increase in worldwide CO_2 emissions since 1971. This increase is of great concern to scientists, governments and the public in general as there is general consensus from the greater scientific community that CO_2 – a greenhouse gas – is one of the main contributors to rapid climate change[1] experienced globally, especially in the last few decades.

Globally, 78–83% of CO_2 emissions can be attributed to electricity generation from fossil fuels.[2] In South Africa the situation is no different with almost 93% of the country's electricity needs provided by fossil fuels; 77% of electricity is provided specifically by coal power plants.[3,4] Because of the country's abundant coal reserves, the use of relatively inexpensive coal-derived power is unlikely to cease in the next 200 years.[3] Coal power plant operations have resulted in South Africa's power sector being the 9th highest CO_2 emitting power sector in the world, with an estimated 218 mega tonnes (Mt) of CO_2 emitted each year.[4,5] Although these emissions are low compared to more developed countries, they are higher than the next nine countries in Africa combined.[4] Eskom Ltd., the country's primary electricity utility, is currently the 2nd highest CO_2 emitting company in the world, as a result of its utilisation of pulverised coal (PC) combustion plants.

As can be seen in Figure 1, the most significant CO_2 emission sources in South Africa are situated in the Gauteng, Mpumalanga and Free State Provinces – not surprisingly, as these provinces form the heart of South Africa's coal mining sector and the regions in which most coal power plants are situated. Figure 1 shows not only the CO_2 emissions from coal power plants, but also those from coal-to-liquids industries, gas-to-liquids industries and oil refining processes.

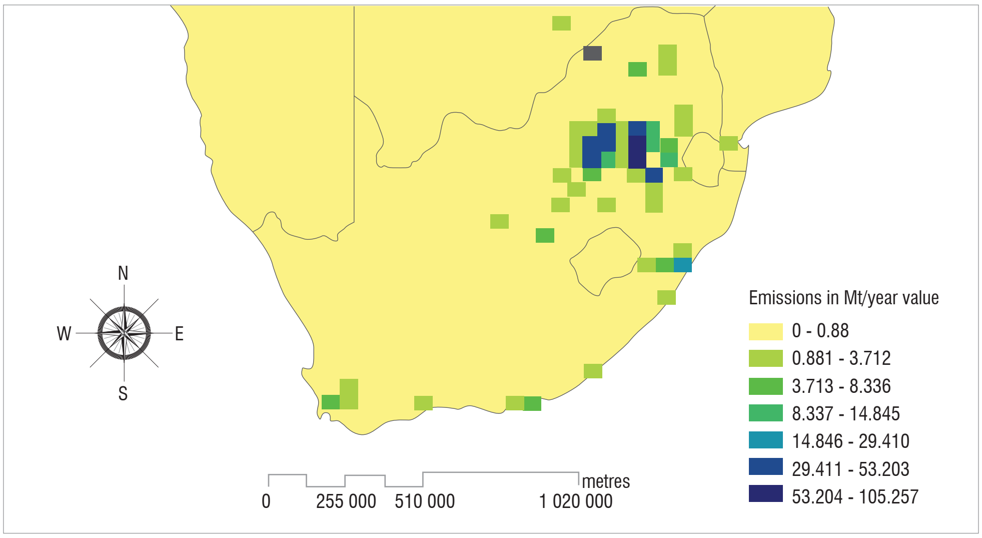


Figure 1: Main CO_2 emission point sources in South Africa.[6]

In an effort to reduce CO_2 emissions and encourage a move towards a cleaner energy strategy, the South African government is considering proposing a CO_2 emissions tax that would be levied on all CO_2 emission sources. Recent debates have suggested a tax rate of R75 to R200 per tonne of CO_2 emitted; with the most recent and currently applicable cost being R120/tonne CO_2 in line with international standards.[7] Considering that South Africa's energy industry emits well over 200 Mt of CO_2 per annum, the proposed levy will result in significant increases in operating costs for companies in this sector. With Eskom having over the last few years almost doubled its electricity tariff, the proposed CO_2 emission tax will further add to the need for Eskom to increase its tariff if it passes on this cost to the consumer. It is therefore imperative that solutions to reduce CO_2 emissions be found.

Carbon capture and storage (CCS) is a promising mid- to long-term solution to reduce CO_2 emissions. This strategy involves capturing CO_2 at power plants and other industries before they are emitted, transporting CO_2 to suitable disposal locations, and either storing CO_2 underground or utilising CO_2 to retrieve high-value products.

In this review, we concentrate on coal power plant operations and their suitability for CCS technology. Techniques that are potentially applicable to CO_2 capture in coal power plants are presented, and CO_2 transportation, storage and potential uses are discussed with specific relevance to South Africa.

Coal power plant operations

Currently, South Africa possesses 14 PC power plants; 7 of them are in the top 30 highest CO_2 emitting power plants in the world.[3,5]

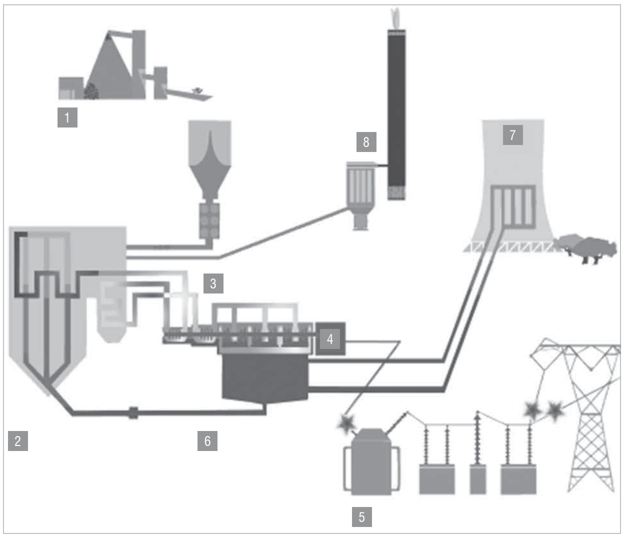
A simplified schematic of a typical PC power plant is shown in Figure 2. In a PC power plant, coal is transported to a pulveriser via conveyor belts and crushed into a powder with a particle diameter of approximately 50 μm. Hot air then blasts the coal into a boiler where it is burnt. The heat generated is used to heat tubes containing water. These tubes can be kilometres long, but are coiled in order to be compact.[8] The water in the heat exchanger tubes is converted into superheated steam at high pressure. The steam is used to drive turbine blades which spin the turbine. The turbine shaft is linked to a generator rotor, which generates electricity using an electromagnet.[8] The electricity flows through transmission lines and transformers to reach consumers at the required voltages. The used steam is then cooled and condensed in cooling towers, and recycled to the boilers for reheating.

The gases that are released during the coal combustion are filtered using bag filters to remove ash. If the gas mixture contains substantial sulphurous and nitrogenous emissions, particularly SO_x and NO_x compounds, then desulphurisation and denitrification processes can be installed to remove them, although these processes have currently not been implemented in South African coal power plants. The remaining gases are emitted through a stack as flue gas. Flue gas composition varies according to coal composition and power plant flue gas treatment processes. The typical composition of flue gases is approximately 12–12.8% CO_2, 6.2% H_2O, 4.4% O_2, 50 ppm CO, 420 ppm NO_x, 420 ppm SO_2 and 76–77% N_2. The flue gas is typically emitted at pressures ranging from 100 kPa to 170 kPa and temperatures of 363.15–412.15 K.[9-11]

PC power plants typically emit CO_2 on a magnitude of 514 kg CO_2/MWh electricity produced.[12]

CO_2 removal from PC power plants entails retrofitting the power plant with a CO_2 capture process to treat the flue gas for selective CO_2 removal before it is emitted through the stack. This mode of CO_2 capture is known as post-combustion capture, because CO_2 capture occurs after coal combustion.

PC combustion is a well-developed and common power plant process that requires a lower investment cost compared to newer technologies. However, CO_2 capture and compression is expensive as the flue gas to be treated is available at unfavourably low pressure and high temperature.



1, coal heap; 2, boilers; 3, superheated steam in turbine; 4, generator rotor; 5, transmission lines; 6, condensed H_2O; 7, cooling tower; 8, stack

Figure 2: Pulverised coal power plant.[8]

Techniques of capturing CO_2 from pulverised coal power plants

Currently, there are many gas separation techniques under investigation for post-combustion CO_2 capture from PC power plants. This section explains the unique properties of CO_2 and presents CO_2 capture techniques which exploit these properties for efficient gas separation, despite the unfavourable conditions of post-combustion flue gas at the stack.

Solubility and pH of CO_2 in H_2O

The solubility of CO_2 in water is 0.9 vol CO_2/vol H_2O or 0.0007 mol CO_2/mol H_2O at 293.15 K.[13,14] CO_2 forms weak carbonic acid when dissolved. This dissolution, however, may reduce the pH of water to as low as 5.5.[12] This finding is important, as it confirms that CO_2 acts as an acid in acid–base reactions, which is vital information in the selection of solvents or sorbents which may be used to absorb or adsorb CO_2.

Gas absorption using solvents

The acidic nature of dissolved CO_2 in water dictates the types of physical and chemical solvents that would potentially be successful for efficient CO_2 absorption. Applicable chemical solvents include amine solvents and solutions, which result in CO_2 absorption by zwitterion formation and easy deprotonation by a weak base.[11] Promising potential physical solvents include chilled ammonia, Amisol and Rectisol solvents,[2] and ionic liquids which consist purely of cations and anions. Huang and Rüther[15] discovered that a Lewis acid type interaction occurs between CO_2 and anions, with CO_2 acting as a Lewis acid and anions acting as a Lewis base.

The selective absorption of CO_2 can be achieved by passing the flue gas through an absorber through which solvent flows counter-currently. CO_2 is selectively absorbed into the solvent and exits through the bottom, while other flue gas components are passed out through the top of the absorber (Figure 3).

The solvent loaded with CO_2 is then heated and sent to a stripping column where desorption occurs. CO_2 is released, while the unloaded solvent is recycled to the absorber.

The advantage of this strategy is that the process is well developed as it is already in use for other gas treatment requirements such as desulphurisation and denitrification processes. There are many possible solvents and solvent mixtures that are under investigation for

CO_2 absorption, including amine and carbonate solvents, as well as ionic liquids.

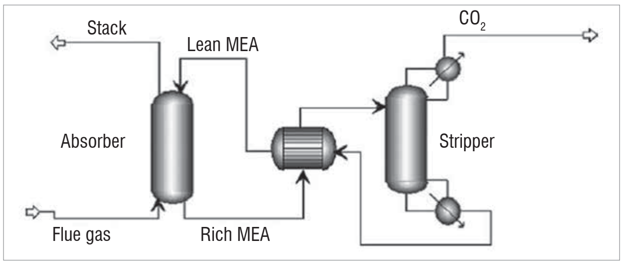


Figure 3: A typical solvent absorption process using mono-ethanolamine (MEA).[2]

The disadvantage is the high energy penalty associated with solvent regeneration in the stripping column. CO_2 absorption increases with decreasing temperature, requiring flue gas to be cooled for CO_2 absorption, as flue gas is available at a relatively high temperature of up to 413 K.[16] However, thereafter, the loaded solvent needs to be heated in the stripping column to release CO_2 and recycle the solvent. There is ongoing research on finding suitable solvents that are easily regenerated with a much lower energy penalty.

Pilot plants for processes of this type have already been set up in Austria and the Netherlands in 2008.[17,18] South Africa's first CO_2 capture plant that would likely include solvent absorption is scheduled to be set up by 2020, pending the success of CO_2 injection projects[19] (Surridge T, 2011, oral and written communication, November 28).

CO_2 capture using dry regenerable sorbents

Figure 4 illustrates a sorbent adsorption process. Flue gas is first cooled and then sent to a carbonation reactor, which is a packed or fluidised sorbent bed reactor. CO_2 is absorbed or adsorbed into the sorbents. This process may be physical or reactive. The sorbent, now loaded with CO_2, is then transferred to a regenerator where it is heated to release the CO_2. The sorbent is then recycled to the carbonation reactor.[16]

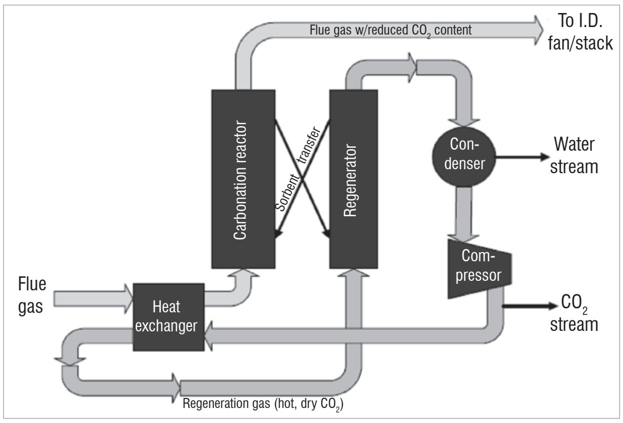


Figure 4: Schematic of the sorbent capture process.[17]

Packed bed reactors are popular for inherently porous activated sorbents while sorbents occurring as pellets, flakes, or fine particulate matter are used in a fluidised bed reactor. The process may operate in continuous or batch mode, depending on the efficiency of solids handling and the CO_2 removal capacity of the process.

Common sorbents under investigation for CO_2 capture include activated coal, sodium carbonate, potassium carbonate and calcium carbonate.[16] CO_2 capture is efficient even at low CO_2 concentrations in the flue gas. Depending on the sorbent and process design, lower regeneration energy requirements can be achieved than those from absorption using amine solvents.[14,17]

The low attrition resistance of many sorbents is a fundamental setback to their implementation as a CO_2 capture technique.[2,18] While single-cycle results seem promising, many sorbents are not robust enough to be used in multi-cycle operation with conventional solids handling techniques. Sorbent pellets may erode or become caked and lose shape. High water content in the flue gas results in further attrition and sorbent caking. Moreover, the expensive nature of solids handling, including conveyor belts and compressed air blast loops which require maintenance, also reduces the feasibility of using sorbents as a CO_2 capture technique.

Research, especially on the introduction of additives and sorbent supports and hybrid processes that combine sorbents with solvents, is being conducted to overcome the current challenges experienced with the use of sorbents.[20] Details of a pilot plant set-up and usage are provided by Manovic et al.[21] who utilised a fixed bed reactor. Fluidised bed pilot projects have also been considered in Canada and Korea.[22,23]

CO_2 molecular size

The CO_2 properties presented by the Asia Industrial Gases Association[13] show that the CO_2 molecule – a carbon atom double bonded to two oxygen atoms – is compact. Also, importantly, the molecule is linear in shape, with a bond length of 116.18 pm, making the molecule approximately 232 pm in length. Figure 5 provides an illustration of the molecule.

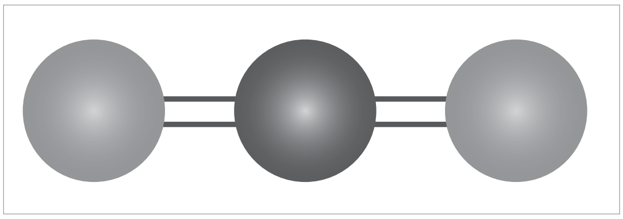


Figure 5: The CO_2 molecule.

By comparison, the diatomic O_2 molecule has a bond length of 120.8 pm, water has a bond length of 102.9 pm and N_2 has a bond length of 109.7 pm.[24] The size and linear shape of the CO_2 molecule in relation to other flue gas components, as well as other properties such as dipole moment and polarisability, facilitates not only the use of sorbents, but also the use of conventional membrane filtration systems, enzymatic membranes and metal organic frameworks to filter out CO_2 from smaller molecules of various components of flue gas.

Membrane separation

Figure 6 illustrates a typical membrane contactor. Flue gas enters into a membrane separation unit. CO_2 selectively permeates through the membrane while other flue gas components do not. Flue gas passes out as stack gas, while CO_2 is recovered and compressed on the other side of the membrane.

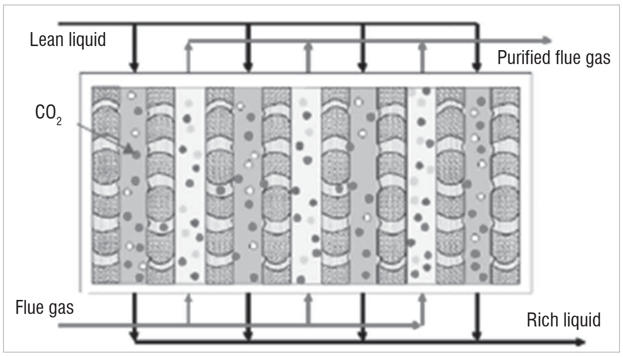


Figure 6: An Illustration of a membrane contactor with solvent.[25]

While membranes can be used on their own, increased efficiency is noted when solvents are used as a sweep fluid to accelerate mass transfer and recover CO_2 on the other side of the membrane. Some solvents, such as ionic liquids, are combined into the membrane pores to increase CO_2 permeability through the membrane.[2]

Common membrane material includes zeolite, ceramic, polymer and silica. More fragile membranes are supported by alumina to increase their robustness. Depending on the type of filtration unit, the process can operate in batch or continuous mode.

The advantage of membranes is that CO_2 can potentially be recovered at high purity. Membrane units are well developed and there is high scope of study regarding membrane types and solvent combinations. If no solvent is used, then solvent regeneration and recycling is not required.

The challenge in implementing membrane separation for CO_2 capture is the high pressure that the process demands. The flue gas needs to be compressed before undergoing filtration in order to achieve a high CO_2 removal rate, which amounts to a high energy penalty. Moreover, many types of membrane material cannot satisfy optimum CO_2 permeability and selectivity constraints and are not robust enough for long-term operation. Satisfying these requirements forms part of ongoing research.

A pilot plant in the Netherlands which accommodates CO_2 capture using membranes combined with solvents was constructed in 2008.[18]

Enzyme-based systems

Instead of using conventional membranes as previously described, enzymes can be used as a liquid membrane suspended between hollow fibre supports for rigidity. As shown in Figure 7, flue gas passes through the liquid membrane. CO_2 is hydrated and permeates as carbonic acid (HCO_3^-) at a faster rate than N_2, O_2 and other flue gas components. CO_2 is recovered under pressure or using a sweep gas on the other side.[2]

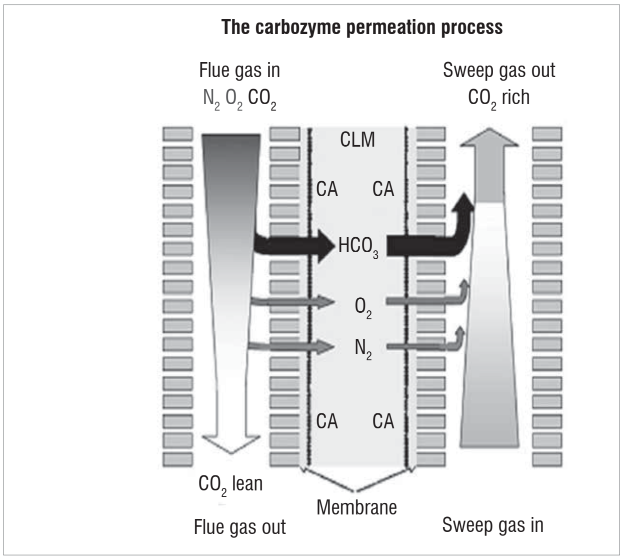


Figure 7: CO_2 separation using the carbonic anhydrase (CA) enzyme.[2]

A popular enzyme for CO_2 capture is carbonic anhydrase. CO_2 recovery with this technique can potentially be as high as 90%.[2] About 600 000 molecules of CO_2 are hydrated by one molecule of carbonic anhydrase.[26] A further advantage is that the heat of absorption of CO_2 into carbonic anhydrase is comparatively low.

Disadvantages include limitations at the membrane boundary layers, long-term uncertainty, and sulphur sensitivity of the enzyme,[26] prompting ongoing research on new enzymes.

Research in this technique has not gone beyond laboratory studies on CO_2 permeability and selectivity.[26,27]

Metal organic frameworks

Metal organic frameworks (MOFs) are hybrid organic/inorganic structures containing metal ions geometrically coordinated and bridged with organic bridging ligands[28] (Figure 8). This arrangement increases the surface area for adsorption, enabling them to be used as sorbents or as nanoporous membranes.

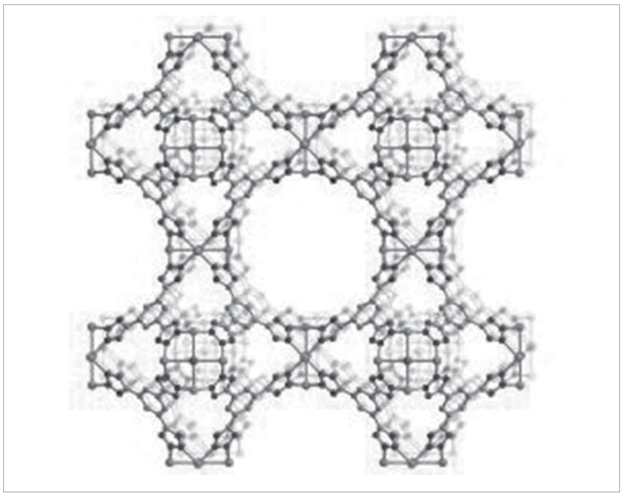


Figure 8: Structure of a typical metal organic framework.[29]

MOFs possess much potential for CO_2 capture because there are hundreds of possible MOFs that can be developed using various combinations of metal ions and organic ligands. They can be tailor-made to suit a particular application such as CO_2 capture.[26] MOFs containing zinc and magnesium ions provide higher CO_2 adsorption and are hence being thoroughly investigated.[30,31] Another advantage is that regeneration energy required is lower than that for conventional sorbents and solvents.[2]

The study of metal organic frameworks is still in its infancy, with investigations being made primarily on a laboratory scale.

CO_2 phase behaviour

The critical point and triple point of CO_2 are 304.25 K and 216.55 K, respectively,[13] and the phase behaviour of CO_2 shown in Figure 9 also allows for CO_2 capture from flue gas by changing conditions of temperature and pressure. Figure 9 shows a wide range of temperature and pressure conditions for the conversion of CO_2 from the gas phase into the liquid phase, as well as into the solid phase for storage. Separation processes that make use of the phase behaviour of CO_2 include cryogenic separation and hydrate formation.

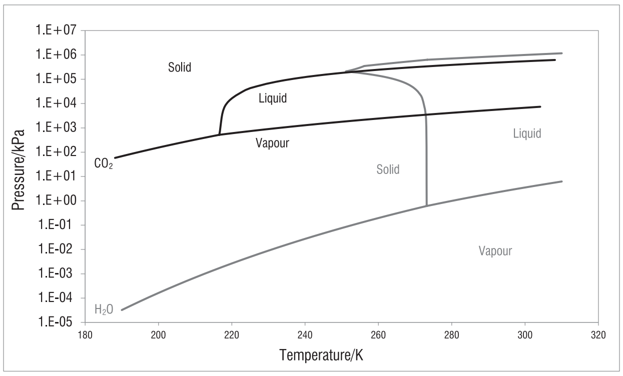


Figure 9: H_2O and CO_2 phase diagram.[11,32]

Cryogenic separation

Cryogenic separation entails the separation of CO_2 from flue gas by a phase change, specifically through cooling flue gas until CO_2 exists in the liquid or solid phase. Figure 9 indicates that vapour–liquid phase change can occur at temperatures between 217 K and 304 K and pressures from 630 kPa to 7396 kPa. In the case of recovering CO_2 in the solid phase at lower temperature, the process is also popularly referred to as CO_2 anti-sublimation.

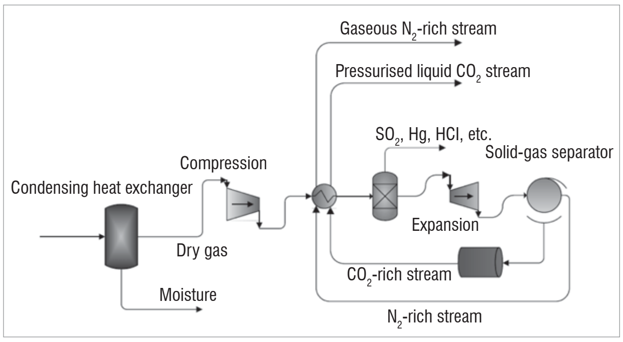


Figure 10: Cryogenic CO_2 capture.[33]

As shown in Figure 10, flue gas is cooled in a heat exchanger and moisture is removed. The resultant dry gas contains CH_4, CO_2, N_2, O_2 and trace components such as Hg, SO_2 and HCl. The dry flue gas is moderately compressed and sent to a heat exchanger where its temperature is lowered to just above the CO_2 solidification point. This temperature varies depending on the operating pressure, which depends on the flue gas conditions from the coal power plant.

SO_2 and other trace compounds from the flue gas are removed using a flash unit which utilises pressure difference to separate components based on volatility. The flue gas then passes through an expander, which causes further cooling and partial precipitation of CO_2. CO_2 is thus separated from the flue gas, which at this point consists primarily of N_2 gas. The CO_2 rich stream is further pressurised and recycled, together with the N_2 rich stream, to the heat exchanger to cool incoming dry flue gas. The CO_2 rich stream undergoes a temperature increase during heat exchange which results in CO_2 being produced in the liquid phase at elevated pressure. N_2 remains in the gaseous phase and is recovered separately.

As an alternative to applying high pressure to compress the flue gas, simulations have proved CO_2 liquefaction to be more energy efficient and cost effective. This technique entails cooling the flue gas instead of compressing it. Energy costs associated with gas compression are reduced, and operating and investment costs for circulation equipment are also reduced.[34]

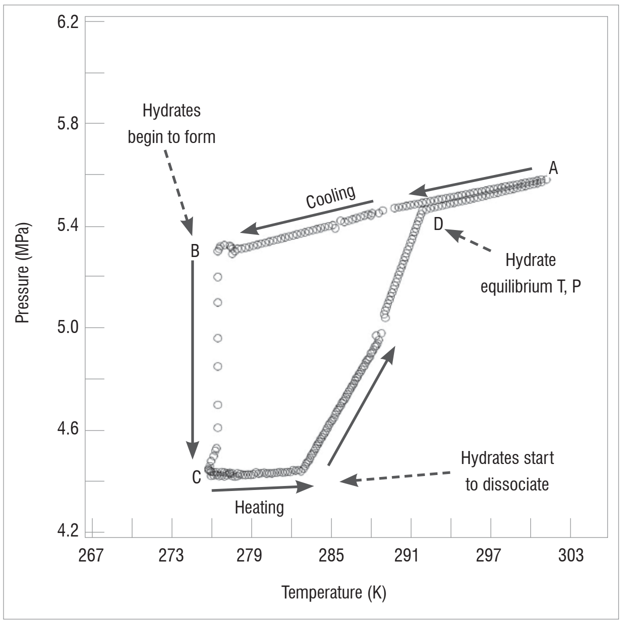
The advantage of cryogenic separation is that CO_2 can potentially be recovered at 99% purity. Refrigeration processes are already well established. Refrigerants such as n-butane, propane, ethane and methane, or a blend of each, can be used.[35]

The disadvantage is the high energy penalty associated with cooling flue gas by refrigeration. Flue gas needs to be cooled to 136–194 K,[35,36] depending on the concentration of CO_2 in the flue gas. Mixed results have been obtained in various studies on the energy penalty and resultant efficiency of cryogenic separation as a CO_2 capture method. Some studies suggest that cryogenic separation possesses an 11–27% energy penalty and is 40% more efficient than conventional absorption.[35,36] However, other studies estimate that the efficiency of cryogenic separation is 3% lower than absorption and membrane processes.[37] Efficiency depends on the CO_2 composition in the flue gas and the degree to which pinch technology can be applied.

The CATO (CO_2 Afvang, Transport en Opslag) programme in the Netherlands, has developed a pilot plant that also accommodates the study of cryogenic separation.[17]

Hydrate formation

A separation process which makes use of CO_2 and H_2O phase behaviour, as well as molecular size and bond lengths, is hydrate formation. This technique entails passing flue gas through a unit containing chilled water at optimum temperature and pressure, causing some components of the flue gas to freeze together with water molecules to form hydrates, which are ice-like crystals in which the gas molecules are trapped inside a cage of water molecules, through hydrogen bonding. Figure 11 shows wide ranges of temperature and pressure that can result in hydrate formation.



A–B: hydrate induction period; B–C: catastrophic growth period; D complete hydrate dissociation.

Figure 11: Pressure–temperature diagram for formation and dissociation of hydrates via the isochoric pressure-search method.[38]

The specific formation of CO_2-water hydrates industrially can require low temperatures of 268.15–298.15 K, and very high pressures of 3000–50 000 kPa.[39] Hydrate formation pressure decreases substantially at temperatures lower than 273 K. Figure 12 shows various hydrate structures. The structures differ depending on the guest molecule. A structure I hydrate is formed with CO_2, as a result of the quadrupole nature of the CO_2 molecule.

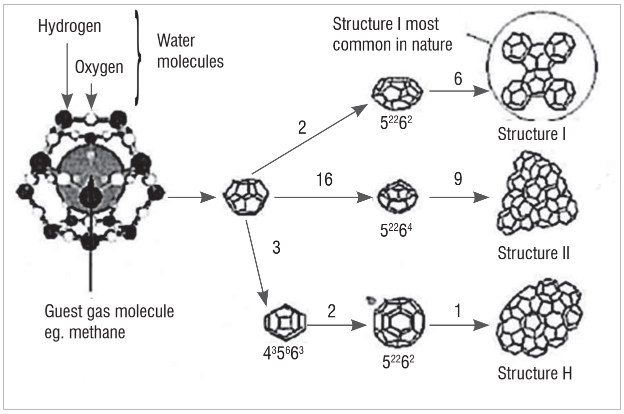


Figure 12: Guest molecule trapped inside water molecule, forming hydrates.[39]

CO_2-water hydrates form and exist as ice crystals in a slurry of water, while other flue gas components remain in the vapour phase and are recovered. CO_2 is thereafter recovered by heating the ice crystals, which releases the CO_2 molecules.

As a result of the size of the CO_2 molecules and the resultant ease of hydrate formation, an advantage of this process is its high selectivity; 99% CO_2 recovery can be achieved.[40] Water is used as an inexpensive recyclable solvent.

The disadvantage is the low temperature and very high pressure required for hydrate formation. Studies are currently being conducted on additives and hydrate formation promoters to reduce the required pressure for hydrate formation, so as to improve the feasibility of the process. Moreover, the handling of slurries results in maintenance problems such as pipeline plugging.

Hydrate formation as a CO_2 capture technique is relatively underdeveloped. There are plans, however, to set up a pilot plant in the USA which caters for hydrate formation.[41]

CO_2 mitigation through the design of new coal power plants

The CO_2 capture techniques described above are investigated primarily for their ability to capture CO_2 from conventional PC power plants. These techniques are intended to be retrofitted in post-combustion mode to existing PC power plants. However, a further option for future coal power plants is to design the coal combustion process in a manner that would result in favourable flue gas composition and conditions, and hence result in more efficient CO_2 capture, from a cost and energy point of view. The main alternative coal combustion processes currently under investigation are integrated gasification combined cycle, oxy-fuel combustion, integrated gasification steam cycle and chemical looping combustion.

Integrated gasification combined cycle

A new alternative power plant process is the integrated gasification combined cycle (IGCC) process. While there are currently no such power plants in South Africa, the process has some advantages over PC power plants and is a more environmentally friendly alternative for new power plant construction.

A simplified schematic of an IGCC power plant is shown in Figure 13. In this process, nearly pure oxygen (O_2) is produced using an air separation unit. The O_2 is sent to a gasifier together with coal. Combustion in the presence of nearly pure O_2 occurs. Coal is partially oxidised to produce a mixture of CO, CO_2 and H_2, collectively known as syngas.[16] The gasifier operates at 3500–7000 kPa and 1255–1644 K. The reactions occurring in the gasifier are[43]:

$C_xH_y + xH_2O \rightarrow xCO + (x + y/2)H_2$ Reaction 1

$C_xH_y + (x/2)O_2 \rightarrow xCO + (y/2)H_2$ Reaction 2

After particulate removal, the syngas is sent to a shift convertor to undergo a water gas shift reaction:

$CO + H_2O \rightarrow CO_2 + H_2$

Reaction 3 (393.15–623.15 K; 15 000 kPa)[16,44]

Steam is utilised in the convertor as a reactant. A gas mixture of CO_2, H_2 and sulphurous and nitrogenous compounds leave the convertor. Unreacted steam is often removed as water. Desulphurisation and denitrification processes are then employed depending on the presence of sulphur- and nitrogen-containing chemicals. The resulting gas mixture contains approximately 50 vol% H_2, 40 vol% CO_2, 2 vol% CO and other trace elements. The gas occurs at 2700 kPa and 310 K.[9]

At this point in the process, CO_2 may be removed using a feasible CO_2 capture technique. CO_2 may then be compressed and stored. After CO_2 capture, H_2 is burned to generate steam at approximately 12 400 kPa[45] which is used to drive the turbines and hence generate electricity.

The electricity generated by the turbine is used to power the gasifier, shift convertor, air separation and compression operations. The remaining electrical energy is then available for commercial use.

IGCC processes are estimated to require higher investment costs than PC processes. However, upon integration of CO_2 capture into the plant, the total investment cost is lower for IGCC processes than for PC processes which are retrofitted for CO_2 capture. IGCC processes also introduce the prospect of pre-combustion capture after shift conversion. CO_2 is captured from flue gas at higher pressure, reducing CO_2 compression costs. It is these advantages that make IGCC an attractive option if the capacity of coal power in South Africa is expanded.

Oxy-fuel combustion

This technique is a modification of the PC power plant. It involves burning coal in nearly pure oxygen.

Oxygen is cryogenically separated from air in an air separation unit. Other air components are emitted into the atmosphere while oxygen is used in the boiler for coal combustion. The resulting heat converts water to superheated steam, for use in steam turbines. The resulting flue gas from combustion is treated for ash and sulphur removal, and thereafter contains CO_2 and water vapour. Water is separated from CO_2 by cooling the flue gas. A schematic of the process is shown in Figure 14.

The main advantage of the process is that the flue gas is available at a high CO_2 concentration of approximately 75.7 mol%,[46] thereby reducing compression costs and facilitating efficient CO_2 removal.[2] Moreover,

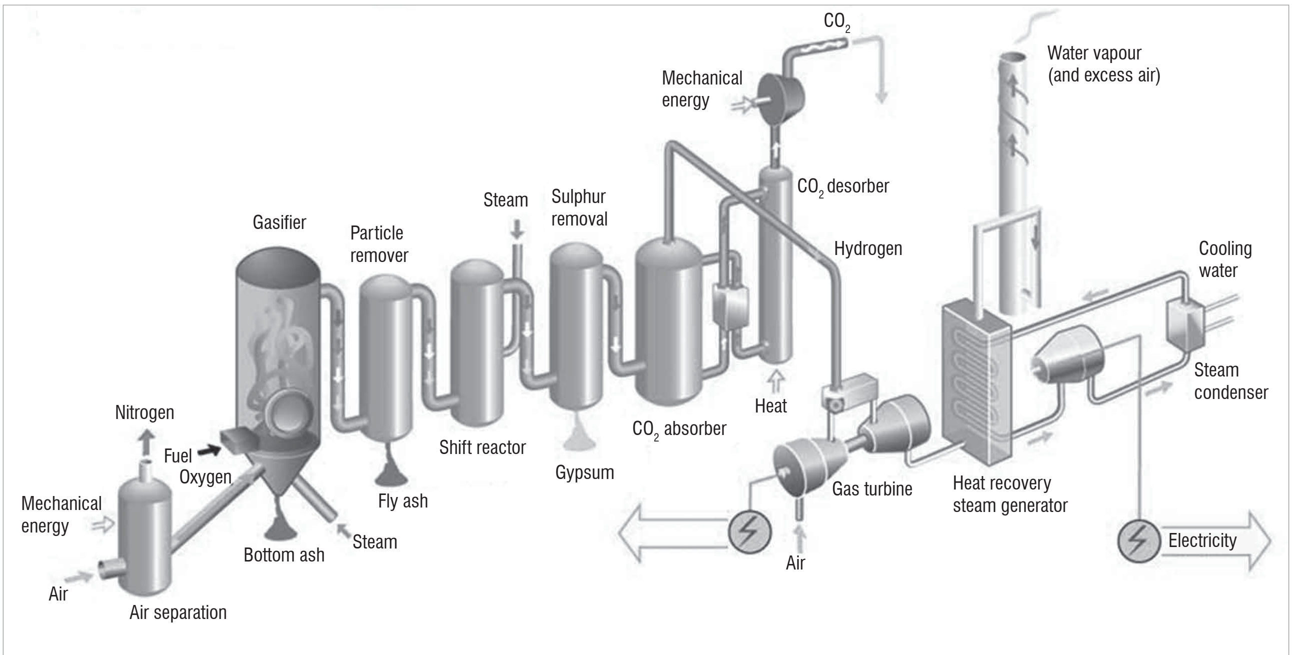


Figure 13: Integrated gasification combined cycle power plant, including CO_2 capture unit.[42]

CO_2 is easily separated from H_2O. The modification of PC to oxy-fuel combustion is also easier than constructing an IGCC process. Oxy-fuel combustion is estimated to inherently result in lower CO_2 emissions than IGCC and conventional PC processes.[46]

The disadvantage is the high flaming temperature at which coal burns in the presence of pure oxygen, which puts much strain on the material of construction.[42] To mitigate this, flue gas is recycled to enable temperature control, as shown in Figure 14. Captured and cooled CO_2 streams may also assist in lowering the temperature of the boiler. Moreover, air separation units require high amounts of energy to obtain pure oxygen from air. Cryogenic methods are also presently accompanied by high energy penalties.

Integrated gasification steam cycle

A US consortium consisting of Siemens Ltd., MAN Ltd., CO_2 Global, Imperial College London, and Jacobs Consultancy has conducted research into a modified IGCC coal combustion process called integrated gasification steam cycle (IGSC), in order to minimise the energy penalty associated with coal power plants possessing CO_2 capture. As seen in Figure 15, waste energy is efficiently utilised through a relatively complex system of recycle streams and turbines of varying pressure.

The process consists of a two-stage combustion system. Coal is gasified in a quench gasifier, which utilises water for temperature control. The resultant syngas contains carbon monoxide, hydrogen gas and oxygen

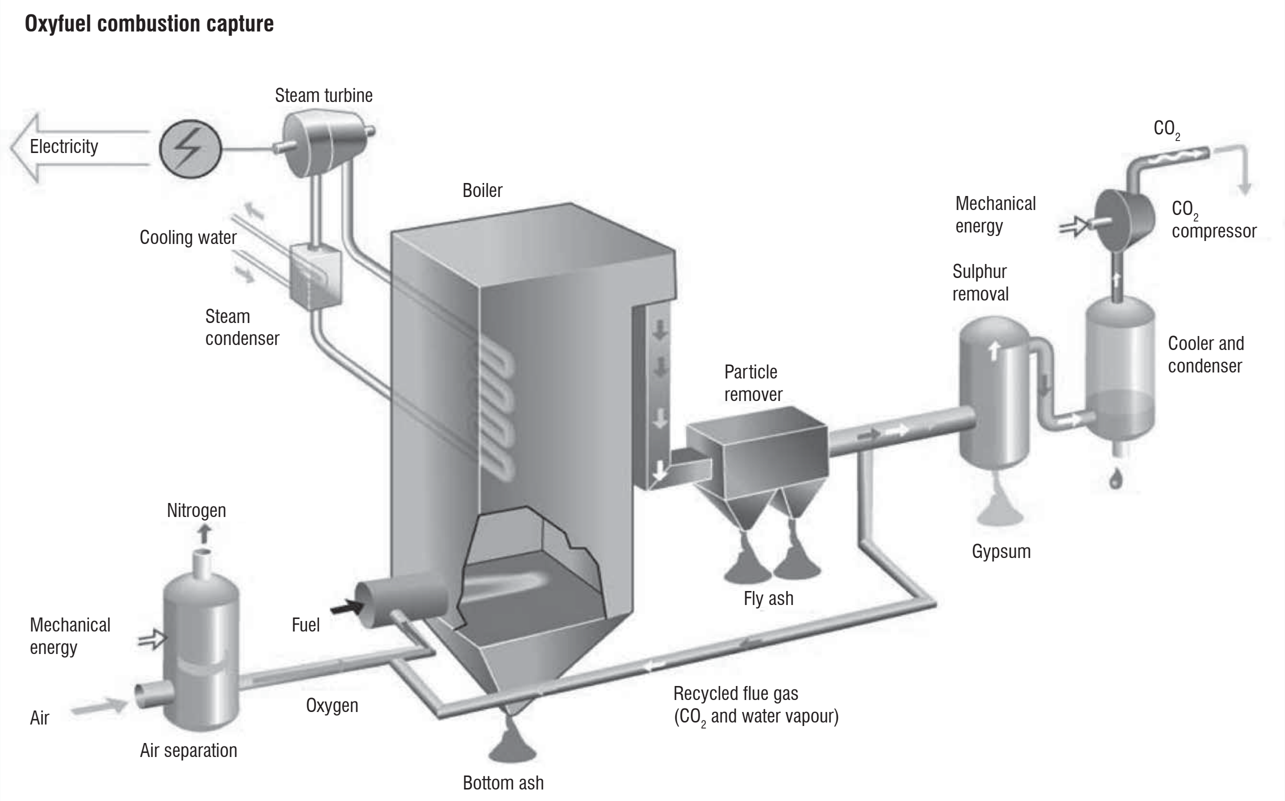


Figure 14: Oxy-fuel combustion capture process.[42]

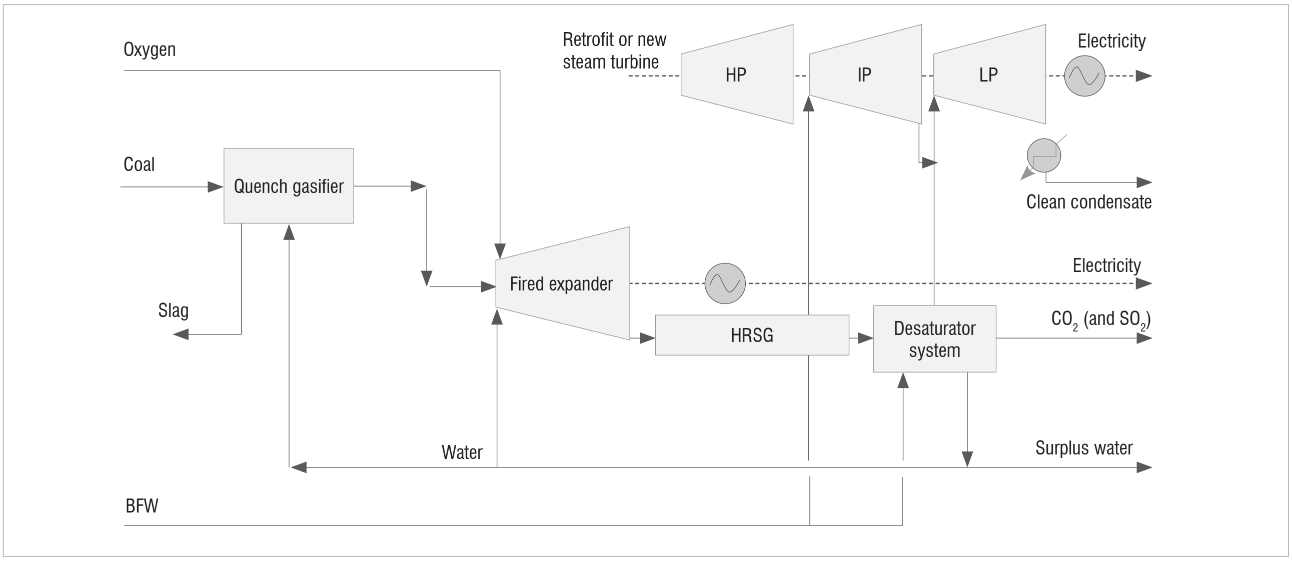


Figure 15: Integrated gasification steam cycle process.[47]

gas, which is passed through a fired expander to generate power. The expander consists of a burner connected to a turbine. Combustion is completed in the expander at temperatures over 1000 K.

The exhaust heat is used to raise high pressure steam in a heat recovery steam generation system, which is then used to power an additional turbine retrofitted to the process. Resultant gases, containing primarily H_2O, CO_2 and trace SO_2, are then cooled in a desaturator to remove H_2O and recover CO_2 in post-combustion mode. The desaturator utilises recycled cooling water and, if optimised, can drive a low-pressure turbine, generating additional power.[47]

The advantage is that the process can potentially obtain 100% CO_2 recovery and increase power plant output by 60% compared to conventional PC power plants.[48] Conventional turbines can be used and CO_2 is available at high pressure.

The capital cost – in the form of the air separation units required to provide oxygen to the gasifier – of IGSC processes is their main drawback. A desulphurisation unit may also be required for coal with a high sulphur content.

Research on IGSC is limited solely to the consortium that invented it, barring all possibility of finding data from other independent sources. There is however, abundant information available from the consortium itself.[47]

Chemical looping combustion

This technique is a further modification of oxy-fuel combustion. Instead of utilising oxygen from an air separation unit for coal combustion, oxygen derived from metal oxides is used.[49] As shown in Figure 16, two fluidised bed reactors are used: the air reactor (1) and fuel reactor (3). Particulate metal or metal oxide is oxidised in the air reactor using air, thereby acting as an oxygen carrier. A cyclone (2) is used to separate the carrier from unreacted components of air, which are emitted as flue gas. The particulate oxygen carrier is transferred into the fuel reactor (3).

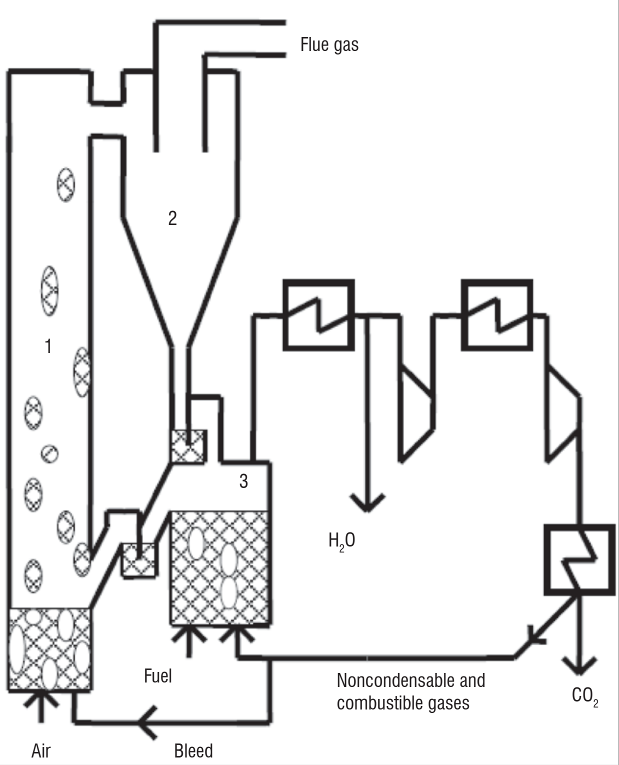


Figure 16: An illustration of chemical looping combustion.[49]

The oxygen carrier is reduced in a combustion reaction with coal and recycled to the air reactor. Flue gas from the fuel reactor contains H_2O and CO_2, and can be used to drive a turbine before being separated by cryogenic means.

The reactions occur typically at 1173.15–1573.15 K.[50] Different metal oxides (such as Fe_2O_3/CuO and $MgAl_2O_4$, nickel, manganese and calcium oxides) can be used as the oxygen carrier.[51,52]

The advantage of chemical looping combustion is that no air separation unit is required, and flue gas contains primarily CO_2 and H_2O, with CO_2 available at 31wt% in the flue gas, which is higher than conventional PC power plant flue gas.[25]

The current disadvantage of chemical looping is the high investment cost required for the technology, which deters research and implementation. Another fundamental challenge to implementation is choosing an ideal oxygen carrier. Current studies show the conversion rate under oxidising conditions, using conventional oxygen carriers to be very fast (nearly 100%/min).[49] However, the occurrence of side reactions with undesirable products is yet to be minimised.[53]

Most research on chemical looping that is currently underway is on the finding of a suitable oxygen carrier.[44] Despite this, a pilot plant has been developed in Sweden to investigate the industrial operation of chemical looping.[49,50]

CO_2 transportation

After the removal of CO_2 from coal power plants and other industrial sources, CO_2 needs to be either transported to locations where it is stored or used in various processes. Currently CO_2 is transported by road. However, the quantity of CO_2 attained from CO_2 capture processes may necessitate that CO_2 be transported via ship or pipeline. In both cases, the required compression pressure can be 10 000–30 000 kPa,[54] depending on the distance and intended disposal or use of CO_2. Compressors may need to be installed every 100 km or so for transport over long distances. Hazards such as the acidic nature of CO_2 gas have to be taken into account when transporting CO_2, particularly if the stream is impure.

CO_2 storage

The high heat of formation of CO_2 of -393.5 kJ/mol[55] provides great difficulty in converting CO_2 to high value products, despite current and recent efforts[56]. While CO_2 has many uses in various commercial enterprises, the sheer amount of CO_2 that would be captured from industrial processes necessitates alternative disposal methods.

The term 'sequestration' describes the removal of CO_2 from the atmosphere and its long-term storage.[57] While there are many options to store CO_2, the most promising strategy involves injecting CO_2 at least 800 m underground, where ambient pressure and temperature are sufficient to result in liquid or supercritical CO_2, and into a formation with an impermeable cap rock so that no substantial leakage may occur.[54]

CO_2 can be stored in geological formations such as former natural gas, oil or brine fields (saline formations), and un-minable coal beds that contain porous rock (Figure 17). There is also the possibility of storing CO_2 in offshore formations of the same nature, but this option is more technically challenging and expensive.

A well-sealed cap rock, containing a layer of shale and clay is essential to prevent upward CO_2 migration and leakage.[58] This form of storage is known as structural trapping and is the most dominant method of trapping gases underground. Residual trapping occurs to a lesser extent, where CO_2 gets trapped in porous rock through displacement and migration. CO_2 may also be trapped through dissolution in underground brine solutions present in porous rock, a mechanism known as solubility trapping. Final trapping of CO_2 occurs upon formation of carbonic acid and, finally, solid carbonate minerals.[59]

A good knowledge of the underground reservoir size is needed to account for horizontal migration of CO_2 and ensure ultimate trapping by geochemical means, such as carbonate formation from reactions with CO_2 and the host rock.[54,58]

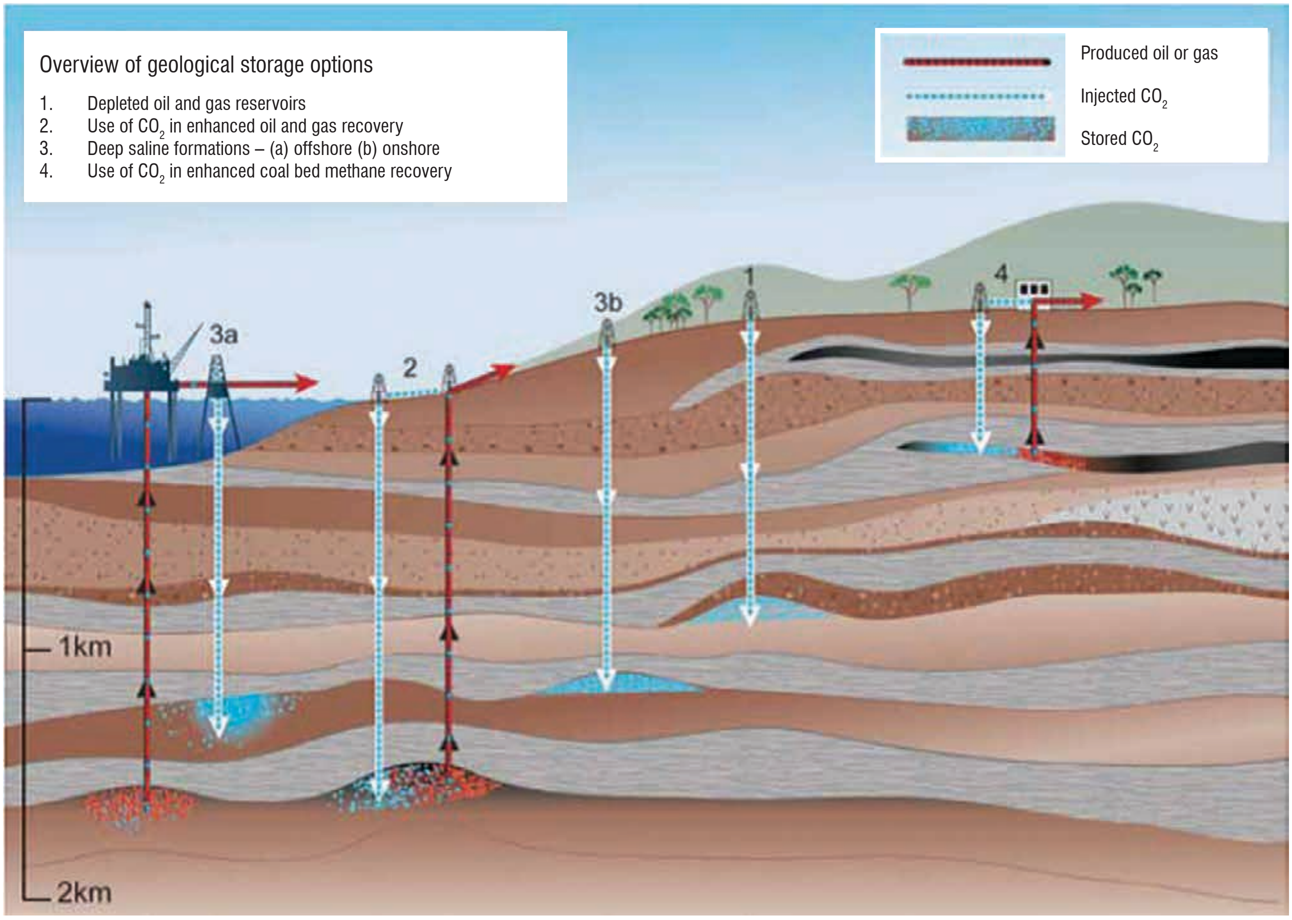

Overview of geological storage options
1. Depleted oil and gas reservoirs
2. Use of CO_2 in enhanced oil and gas recovery
3. Deep saline formations – (a) offshore (b) onshore
4. Use of CO_2 in enhanced coal bed methane recovery
Produced oil or gas
Injected CO_2
Stored CO_2
1
2
3a
3b
4
1km
2km

Figure 17: Geological storage options.[59]

While CO_2 capture is relatively underdeveloped for commercial use, CO_2 sequestration is already prevalent in oil, gas and coal industries in Canada, Algeria, the USA, Norway, the Netherlands, China, Japan, Poland and Australia.[59]

Viljoen[58] and Cloete[60] present areas in South Africa where CO_2 can be stored. These include saline formations (structures of porous rock typically containing mineralised brine solutions in the pores), gas reservoirs and depleted coal mines. The closest area to many of Eskom's operations, and most CO_2-emitting industries in general, are parts of the Northern Karoo, which possesses free saline formations, as well as depleted coal mines. Further formations are shown in Figure 18, which includes large offshore storage opportunities as well.

Figure 18 shows that some possible storage sites are relatively close to the CO_2-emitting sources shown in Figure 1. CO_2 compression and transportation costs via pipelines could be reduced. However, many possible storage sites are far from CO_2 emission sources. CO_2 storage attempts are currently being planned in South Africa, with a test scheduled for 2016. If CO_2 storages tests are successful, CO_2 capture implementation tests should be completed by 2020[19] (Surridge T, 2011, oral and written communication, November 28).

Uses of CO_2 storage

In addition to high pressure CO_2 storage in porous reservoirs, CO_2 may be injected underground to assist in the recovery of high value resources such as oil, natural gas, and methane. These are more feasible solutions that can partially or completely recover the cost of CO_2 capture and storage. Below are a few options for CO_2 storage. (Refer to Figure 17 for illustrations.)

CO_2 enhanced oil recovery

Enhanced oil recovery refers to the strategy of injecting CO_2 into nearly depleted oil wells to pressurise the well and force the remaining oil upwards. CO_2 also reduces the viscosity of the oil for easier extraction. Once the oil is depleted, the CO_2 is sealed off underground. Two boreholes are drilled, one for injecting CO_2 and the other to allow the upward migration of oil. Existing boreholes may also be used. Enhanced oil recovery can increase the recovery rate of oil by 8–15%, and can ultimately increase the recovery of oil by up to 50% of the total oil originally recoverable.[54]

A major enhanced oil recovery project using CO_2 has been conducted in Canada.[59] Enhanced oil recovery projects using other gases such as N_2 and hydrocarbon gas mixtures are also a mature technology.

CO_2 enhanced gas recovery

Enhanced gas recovery refers to the extraction of natural gas from nearly depleted gas reservoirs using CO_2. Natural gas is a mixture composed of methane and various hydrocarbon gases. After conventional extraction, gas reservoirs still contain 10–20% of their initial capacity.[54] Because of the reduced pressure in the reservoir, conventional extraction becomes unfeasible. CO_2 injection increases the pressure in the reservoir. Moreover, CO_2 is denser than natural gas and sinks to lower regions of the reservoir, forcing natural gas upwards. In this way, the reservoir can be completely emptied of all natural gas. An enhanced gas recovery system is in operation in the Netherlands.[59]

CO_2 enhanced coal bed methane recovery

Enhanced coal bed methane (ECBM) recovery involves the extraction of methane gas from coal seams using CO_2. The process of extraction is similar to enhanced gas recovery but is done with coal seams. CO_2 is injected into an un-minable or unfeasibly minable coal bed and methane is forced up through a drilled outlet channel. Conventional extraction techniques recover 50% of methane in coal beds. The use of CO_2 in ECBM has the potential to increase methane recovery to 90% or even 100%.[54]

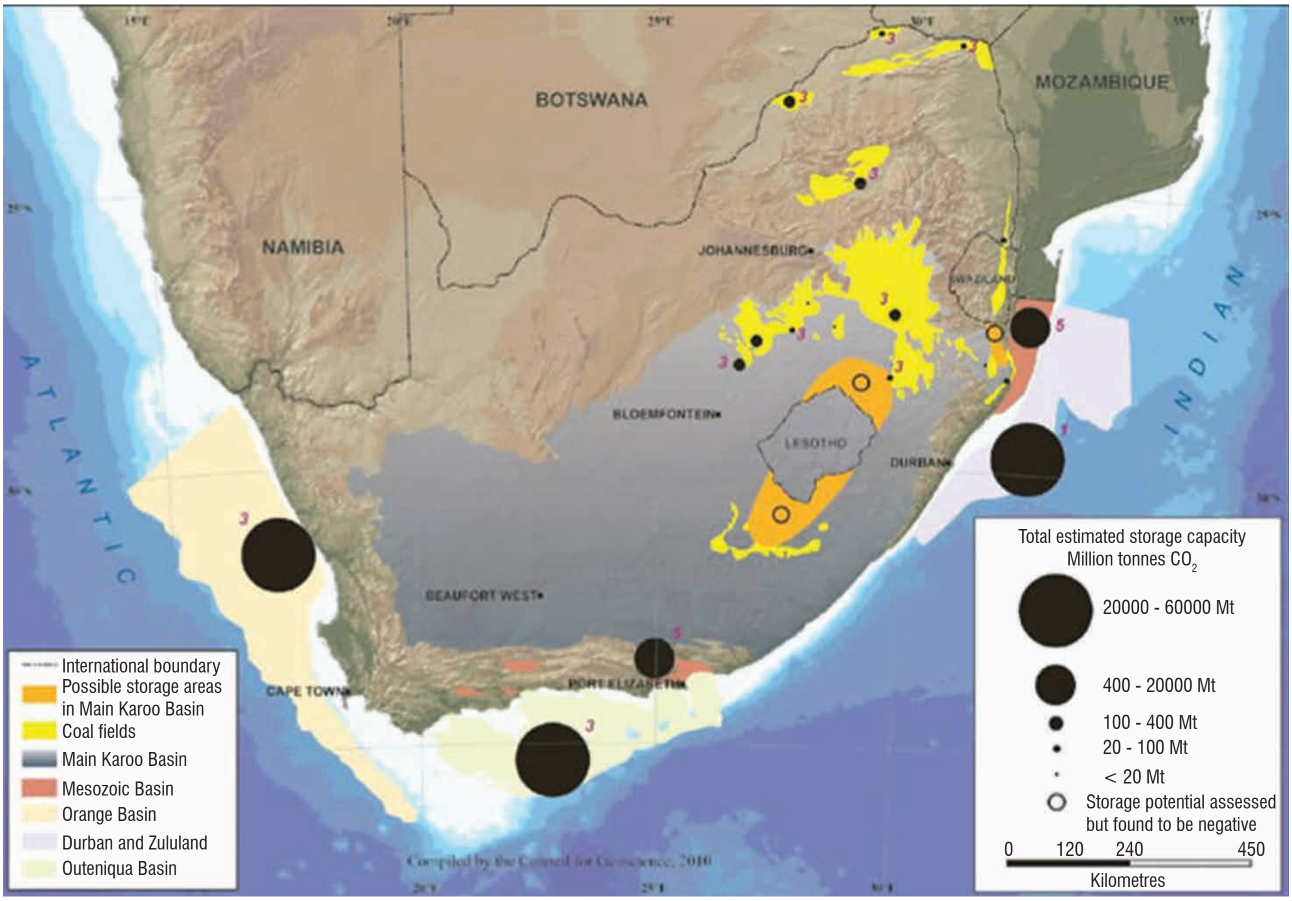


Figure 18: Potential areas for CO_2 storage in South Africa.[61]

While the strategy seems theoretically attractive in terms of methane recovery, the main concern with ECBM recovery is the potential for CO_2 leakage which might occur as a result of the relatively shallow depth of extraction and the permeability of coal seams that are required. Although shallow depth and permeable rock facilitate efficient methane recovery, they also are conducive to permanent, stable CO_2 storage.

Moreover, for leakage to be prevented and for predictable channelling of methane and CO_2 to occur, the coal bed has to be sufficiently thick, which requires an amount of coal to be left un-mined and the mine must be rendered un-minable. A thorough feasibility analysis is required to ensure that ECBM is worth the cost of un-mined coal. Additionally, coal mines that seem uneconomical to mine presently may be economically minable with technological development in the future. However, CO_2 ECBM may render these coal mines permanently un-minable.

ECBM operations are currently underway in Canada, China, Japan and Poland.[59] The strategy has some potential in South Africa because of the vast number of coal beds in the country. However, coal beds offer the least amount of storage potential in South Africa and worldwide because of their shallow depth and capacity compared to other types of potential storage reservoirs, and the permanent loss of currently un-minable coal.

Conclusions

Seven main CO_2 capture techniques have been identified which show great promise as industrial CO_2 emissions mitigation solutions. The technique of CO_2 absorption using solvents was identified as currently the best option for industrial implementation as a capture technique. While other techniques are under consideration, the use of solvents is currently the most likely to be implemented in the South African CCS schedule. Other techniques, while potentially more efficient than the use of solvents, still require substantial research to bring them to the stage of industrial implementation. Four power plant modifications or alternative process designs have been identified and are currently at the pilot plant stage of research. No country has as yet issued a full roll-out of these technologies and it remains unclear as to when these designs would be implemented on an industrial scale. The potential areas of CO_2 storage have been mapped out, with a test injection due to commence in 2016. Three uses of CO_2 storage have been identified to make CO_2 storage more economically attractive, but their applicability to South Africa requires more thorough investigation.

Acknowledgements

This work is based upon research supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation.

Authors' contributions

D.R. was the project leader and K.O.'s PhD supervisor. C.C. was the co-supervisor of K.O.'s PhD. K.O. undertook the bulk of the literature review and writing of the manuscript. D.R. and C.C. conceptualised and edited the manuscript.

References

1. International Energy Agency (IEA). Key world energy statistics – 2010. Paris: International Energy Agency; 2010.

2. Figueroa JD, Fout T, Plasynski S, McIlvried H, Srivastava RD. Advances in CO_2 capture technology – The US Department of Energy's carbon sequestration program. Int J Greenh Gas Con. 2008;2:9–20. http://dx.doi.org/10.1016/S1750-5836(07)00094-1

3. Eskom Power Generation. Eskom – coal power [homepage on the Internet]. c2011 [cited 2011 Feb 28]. Available from: http://www.eskom.co.za/live/content.php?Item_ID=279

4. Carbon Monitoring for Action (CARMA). The 10 largest CO_2 emitting power sectors in the world by country [homepage on the Internet]. c2009 [cited 2011 July 04]. Available from: www.carma.org

5. Carbon Monitoring for Action (CARMA). Top 30 CO_2-emitting power plants in the world [homepage on the Internet]. c2007 [cited 2011 July 04]. Available from: www.carma.org

6. Surridge T. South African activities related to carbon capture and storage [document on the Internet]. c2005 [cited 2009 Jan 28]. Available from: http://www.cslforum.org/documents/pg_RomeMinutespublic.pdf

7. National Treasury. Carbon Tax Policy Paper – Reducing greenhouse gas emissions and facilitating the transition to a green economy. Discussion paper for public comment [document on the Internet]. c2010 [cited 2013 Aug 07]. Available from: http://www.treasury.gov.za/public%20comments/Discussion%20Paper%20Carbon%20Taxes%2081210.pdf

8. Eskom Ltd. Eskom coal power animation [homepage on the Internet]. c2011 [cited 2011 May 25]. Available from: http://www.eskom.co.za/content/Coal.swf

9. National Energy Technology Laboratory (NETL). Doe/Netl advanced carbon dioxide capture r&d program: Technology update. U.S.A.; National Energy Technology Laboratory; c2010 [cited 25 May 2011]. http://www.netl.doe.gov/technologies/coalpower/ewr/pubs/CO2%20Capture%20Tech%20Update%20Final.pdf

10. Brennecke JF, Gurkan BE. Ionic liquids for CO_2 capture and emission reduction. J Phys Chem Lett. 2010;1:3459–3464. http://dx.doi.org/10.1021/jz1014828

11. Asia Industrial Gases Association. Carbon dioxide – Globally Harmonised Document. 7th edn. Singapore: Compressed Gas Association; 2009. Available from: http://www.asiaiga.org/docs/AIGA%20068_10%20Carbon%20Dioxide_reformated%20Jan%2012.pdf

12. US Environmental Protection Agency (EPA). Air emissions [homepage on the Internet]. c2013 [cited 2013 Aug 13]. Available from: http://www.epa.gov/cleanenergy/energy-and-you/affect/air-emissions.html

13. MacColl B. Carbon capture and storage (CCS) – Strategic considerations for Eskom, South African Centre for Carbon Capture and Storage [document on the Internet]. c2011 [cited 2013 Aug 08]. Available from: http://www.sacccs.org.za/wp-content/uploads/2011/11/Day2/CCS%20in%20Utilities%20-%20CCS%20Week.pdf

14. Lentech Water Treatment Solutions. What is carbon dioxide and how is it discovered? [homepage on the Internet]. c2009 [cited 2012 Feb 14]. Available from: http://www.lenntech.com/carbon-dioxide.htm

15. Huang J, Rüther T. Why are ionic liquids attractive for CO_2 absorption? An overview. Aust J Chem. 2009;62:298–308. http://dx.doi.org/10.1071/CH08559

16. Osman K. Carbon dioxide capture methods for industrial sources: A literature review. Energy efficiency and feasibility study [dissertation]. Durban: University of KwaZulu-Natal; 2010.

17. Vierde Nationaal Symposium (VNS). CCS, CATO CO_2 catcher, a CO_2 capture plant treating real flue gas [homepage on the Internet]. c2008 [cited 2011 Feb 12]. Available from: http://www.co2-cato.nl/modules.php?name=CATO&page=79&symposium=true

18. Knudsen JN, Vilhelmsin PJ, Jensen JN, Biede O. Performance review of castor pilot plant at Esbjerg. Vienna: Dong Energy; 2008.

19. Surridge T. South African activities related to carbon capture and storage – September 2005 [document on the Internet]. c2005 [cited 2011 March 08]. Available from: http://www.cslforum.org/documents/pg_RomeMinutespublic.pdf

20. Green DA, Turk BS, Nelson TO, Box P, Gupta RP. Carbon dioxide capture from flue gas using dry regenerable sorbents. Durham, NC: US Department of Energy, National Energy Technology Laboratory; 2004.

21. Manovic V, Anthony EJ, Lu DY. Sulphation and carbonation properties of hydrated sorbents from a fluidized bed CO_2 looping cycle reactor. Fuel. 2008;87:2923–2931. http://dx.doi.org/10.1016/j.fuel.2008.04.023

22. Yi C, Jo S, Seo Y, Lee JB, Ryu CK. Continuous operation of the potassium-based dry sorbent CO_2 capture process with two fluidized-bed reactors. Int J Greenh Gas Con. 2007;1:31–36. http://dx.doi.org/10.1016/S1750-5836(07)00014-X

23. Lu DY, Hughes RW, Anthony EJ. Ca-based sorbent looping combustion for CO_2 capture in pilot-scale dual fluidized beds. Fuel Proc Tech. 2008;89:1386–1395. http://dx.doi.org/10.1016/j.fuproc.2008.06.011

24. Weast RC, Astle MJ, Beyer WH. CRC handbook of chemistry and physics. 64th ed. Boca Raton, FL: CRC Press; 1983.

25. National Energy Technology Laboratory (NETL). Chemical looping process in a coal to liquids configuration [document on the Internet]. c2007 [cited 2009 Jan 29]. Available from: http://www.netl.doe.gov/energy-analyses/pubs/DOE%20Report%20on%20OSU%20 Looping%20final.pdf

26. Trachtenberg MC, Tu CK, Landers RA, Willson RC, McGregor ML, Laipis PJ, et al. Carbon dioxide transport by proteic and facilitated transport membranes. Int J Earth Space.1999;6:293–302.

27. Ge JJ, Cowan RM, Tu CK, McGregor ML, Trachtenburg MC. Enzyme-based CO_2 capture for ALS [document on the Internet]. c2011 [cited 2011 Aug 04]. Available from: http://www.carbozyme.us/publications/P5.pdf

28. Plasynski S, Lang DA, Richard W. Carbon dioxide separation with novel microporous metal organic frameworks [document on the Internet]. c2011 [cited 2011 Aug 04]. Available from: http://www.netl.doe.gov/publications/factsheets/project/Proj315.pdf

29. Long J. New metal organic frameworks in action for capturing carbon dioxide [document on the Internet]. c2010 [cited 2011 Aug 04]. Available from: http://www.greenoptimistic.com/2010/06/02/metal-organic-frameworks-carbon-dioxide-capture/

30. Yazaydın AO, Snurr Q, Park TH, Koh K, Liu J, LeVan MD, et al. Screening of metal-organic frameworks for carbon dioxide capture from flue gas using a combined experimental and modelling approach. J Am Chem Soc. 2009;131:18198–18199. http://dx.doi.org/10.1021/ja9057234

31. Simmons JM, Wu H, Zhou W, Yildirim T Carbon capture in metal-organic frameworks – A comparative study. Energy Env Sci. 2011;4:2177–2185. http://dx.doi.org/10.1039/c0ee00700e

32. Chemicalogic. Steamtab [homepage on the Internet]. c1999 [cited 2012 Feb 14]. Available from: http://www.chemicalogic.com/download/phase_diagram.html

33. Burt S, Baxter A, Baxter L. Cryogenic CO_2 capture to control climate change emissions [document on the Internet]. c2011 [cited 2011 June 21]. Available from: http://www.sustainablees.com/documents/Clearwater.pdf

34. Xu G, Li L, Yang Y, Tian L, Liu T, Zhang K. A novel CO_2 cryogenic liquefaction and separation system. Energy. 2012;42:522–529. http://dx.doi.org/10.1016/j.energy.2012.02.048

35. Clodic D, Hitti R, Younes M, Bill A, Casier F. CO_2 capture by anti-sublimation: Thermo-economic process evaluation. 4th Annual Conference on Carbon Capture and Sequestration; 2005 May 2–5; National Energy Technology Laboratory, Alexandria, VA, USA. p. 1–5.

36. Baltus RE, Culbertson BH, Dai S, Liu H, DePaolo DW. Low-pressure solubility of carbon dioxide in room-temperature ionic liquids measured with a quartz crystal microbalance. J Phys Chem B. 2004;108:721–727. http://dx.doi.org/10.1021/jp036051a

37. Gottlicher G, Pruschek R. Comparison of CO_2 removal systems for fossil-fuelled power plant processes. Eng Conv Manage. 1997;38:173–178. http://dx.doi.org/10.1016/S0196-8904(96)00265-8

38. Sloan ED, Koh C. Clatrate hydrates of natural gases. 3rd ed. New York: CRC Press; 2012.

39. Jadhawar P, Mohammadi AH, Yang JT, Tohidi B. Subsurface carbon dioxide storage through clathrate hydrate formation. Edinburgh: Institute of Petroleum Engineering, Heriot-Watt University; 2006.

40. Chatti I, Delahaye A, Fournaison L, Petitet JP. Benefits and drawbacks of clathrate hydrates: A review of the areas of interest. Eng Con Manage. 2005;46:1333–1343.

41. Tam SS, Stanton ME, Ghose S, Deppe G, Spencer DF, Currier RP, et al. A high pressure carbon dioxide separation process for IGCC plants [document on the Internet]. c2011 [cited 2011 July 09]. Available from: http://www.netl.doe.gov/publications/proceedings/01/carbon_seq/1b4.pdf.

42. Arshad MW. CO_2 capture using ionic liquids [Master's dissertation]. Lyngby: Technical University of Denmark; 2009. Available from: http://docs.google.com/viewer?a=v&q=cache:uF9eKE4Xeg0J:orbit.dtu.dk/getResource%3FrecordId%3D240068%26objectId%3D1%26versionId%3D1+Non+fluorinated+Ionic+Liquids+%2B+CO2+capture&hl=en&pid=bl&srcid=ADGEESg9aXin_GbLKmM6LyI0ZwZlSYo9jdm6WoHXOZShMxVHKwcdqJ9348xr_ET4DibAHbAcF09sbUcIgJSDpEtHdGpt8LdGo4lv02MgmONX0xD9Dj8r9vXvxaAYZI1cbkOF3ovX0axf&sig=AHIEtbRXq7UJoi74_T-CVBR3d5zuDcG5EQ

43. Steeneveldt R, Berger B, Torp TA. CO$_2$ capture and storage: Closing the knowing-doing gap. Chem Eng Res Des. 2006;84:739–763. http://dx.doi.org/10.1205/cherd05049

44. Kanniche M, Bouallou C. CO$_2$ capture study in advanced integrated gasification combined cycle. App Therm Eng. 2007;27:2693–2702.

45. Department of Energy. Cost and performance baseline for fossil energy plants: Bituminous coal and natural gas to electricity. US National Energy Technology Laboratory report, 2007.

46. Davison J. Performance and costs of power plants with capture and storage of CO$_2$. Energy. 2007;32:1163–1176. http://dx.doi.org/10.1016/j.energy.2006.07.039

47. Karmarkar M, Griffiths J, Russell A, Allen R, Austell M, Trusler M. Industrial and utility scale IGSC coal power stations [document on the Internet]. c2011 [cited 2011 Aug 03]. Available from: http://webarchive.nationalarchives.gov.uk/+/http://www.berr.gov.uk/files/file52638.pdf

48. Kent R. New power cycles with carbon capture and sequestration [document on the Internet]. c2011 [cited 2011 Aug 03]. Available from: http://www.wcsawma.org/sitebuildercontent/sitebuilderfiles/34.pdf

49. Mattisson T, Lyngfelt A. Applications of chemical-looping combustion with capture of CO$_2$ [document on the Internet]. c2011 [cited 2011 Aug 08]. Available from: http://www.entek.chalmers.se/∼anly/symp/01mattisson.pdf

50. Mattisson T. Chemical looping combustion using gaseous and solid fuels. International Energy Agency (IEA) Greenhouse Gas R&D Programme [document on the Internet]. c2007 [cited 2011 Aug 08]. Available from: http://www.co2captureandstorage.info/docs/oxyfuel/MTG2Presentations/Session%2006/22%20-%20T.%20Mattisson%20(Chalmers%20University).pdf

51. Wang S, Wang G, Jiang F, Luo M, Li H. Chemical looping combustion of coke oven gas by using Fe$_2$O$_3$/CuO with MgAl$_2$O$_4$ as oxygen carrier. Energy Env Sci. 2010;3:1353–1360. http://dx.doi.org/10.1039/b926193a

52. Fang H, Haibin L, Zengli Z. Advancements in development of chemical-looping combustion: A review. Int J Chem Eng. 2009;2009:1–16. http://dx.doi.org/10.1155/2009/710515

53. Wall T, Liu Y. Chemical looping combustion and CO$_2$ capture: Status and developments [document on the Internet]. c2011 [cited 2011 Aug 08]. http://www.ccsd.biz/publications/files/TN/TN%2032%20Chem%20looping%20updated_web.pdf

54. International Energy Agency (IEA). Prospects for CO$_2$ capture and storage [document on the Internet]. c2004 [cited 2009 Mar 05]. Available from: http://www.iea.org/textbase/nppdf/free/2004/prospects.pdf

55. Perry RH, Green DW. Perry's chemical engineers' handbook. 7th ed. New York: McGraw Hill; 1997.

56. L'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME). Panorama des voies de valorisation du CO$_2$ [Overview of the paths of valorisation of CO$_2$]. Angers: l'Agence de l'Environnement et de la Maîtrise de l'Energie (ADEME); 2010.

57. Teng F, Tondeur D. Efficiency of carbon storage with leakage: Physical and economical approaches. Energy. 2006;32:540–548. http://dx.doi.org/10.1016/j.energy.2006.07.027

58. Viljoen JHA, Stapelberg FDJ, Cloete M. Technical report on the geological storage of carbon dioxide in South Africa [document on the Internet]. c2011 [cited 2011 June 30]. Available from: http://www.ccsconference.co.za/images/presentations/ thinus_cloete.pdf.

59. Intergovernmental Panel on Climate Change (IPCC). Carbon dioxide capture and storage: Summary for policy makers and technical summary [document on the Internet]. c2005 [cited 2011 Jan 26]. Available from: http://www.climnet.org/EUenergy/IPCC_CCS_0905.pdf.

60. Cloete M. Atlas on geological storage of carbon dioxide in South Africa [document on the Internet]. c2010 [cited 2011 June 30]. Available from: http://www.sacccs.org.za/wp-content/uploads/2010/11/Atlas.pdf.

61. CO$_2$ Capture Project. CO$_2$ trapping mechanisms [homepage on the Internet]. c2008 [cited 2013 Aug 12]. Available from: http://www.co2captureproject.org/co2_trapping.html

A review on opportunities for the development of heat pump drying systems in South Africa

AUTHORS:
Thomas Kivevele[1]
Zhongjie Huan[1]

AFFILIATION:
[1]Department of Mechanical Engineering, Tshwane University of Technology, Pretoria, South Africa

CORRESPONDENCE TO:
Thomas Kivevele

EMAIL:
kivevelet@tut.ac.za

POSTAL ADDRESS:
Department of Mechanical Engineering, Tshwane University of Technology, Private Bag X680, Pretoria 0001, South Africa

DATES:
Received: 30 July 2013
Revised: 27 Nov. 2013
Accepted: 28 Nov. 2013

KEYWORDS:
heat pump; drying; specific moisture extraction rate; coefficient of performance; energy efficiency

HOW TO CITE:
Kivevele T, Huan Z. A review on opportunities for the development of heat pump drying systems in South Africa. S Afr J Sci. 2014;110(5/6), Art. #2013-0236, 11 pages. http://dx.doi.org/10.1590/sajs.2014/20130236

Recently, it has been discovered that heat pump drying is an efficient method of drying for drying industries. Heat pumps deliver more heat during the drying process than the work input to the compressor. Heat pump drying is a more advanced method than the traditional South African industrial and agricultural drying methods, such as direct/indirect sunlight, wood burning, fossil fuel burning, electrical heating and diesel engine heating. Heat pump dryers provide high energy efficiency with controllable temperature, air flow and air humidity and have significant energy-saving potential. In the last decade the market for heat pump systems for water heating and space cooling/heating has grown in South Africa, but the development of heat pumps for industrial and agricultural drying is very slow. As a result of high increases in fossil fuel prices and electricity in South Africa, as well as the problem of CO_2 emissions, green energy, energy saving and energy efficiency are imperative. The development of heat pump drying systems in South Africa is an efficient way to solve energy problems in drying applications as this technology is still in its infancy. We review studies on heat pump drying and compare the methods therein with the most common methods of drying in South Africa.

Introduction

Heat pump drying (HPD) is a technology by which materials can be dried at low temperature and in an oxygen-free atmosphere, using less energy than common drying methods. HPD is therefore advantageous for drying biological materials which are thermally sensitive and oxygen sensitive.[1] Drying is a key process in many food industries and in many agricultural countries; large quantities of food products are dried to improve shelf life, reduce packaging cost, lower shipping weights, enhance appearance, encapsulate original flavour and maintain nutritional value.[1] The primary objective of drying is to remove moisture from the food so that bacteria, yeast and mould cannot grow and spoil the food. Economic consideration, environmental concerns and product quality aspects are the three main goals of drying process research in the food industry.[2] In many cases, the drying process is applicable to seasonal biological materials such as fruits and vegetables, so that they can be stored for as long as possible and be available out of season.[3]

Traditional methods of drying biological materials such as fruits, grains, nuts and vegetables have been widely employed in South Africa. The well-known drying methods employ direct/indirect sunlight, wood and fossil fuel burning, electrical heating and diesel engine heating. However, the majority of these methods result in smoke and other emissions which have negative effects on human health and contribute to climate change. In addition, a review of the literature reveals that commercial dryers are highly inefficient. A reason for this inefficiency is that commercial dryers are generally not equipped with heat recovery facilities, whereas heat pumps can provide a very efficient means of recovering both sensible as well as latent heat. A heat pump also delivers more heat than the work input to the compressor. Heat pump assisted dryers are approximately 10 times more efficient than traditional drying systems.[2] A limited number of studies have reported the benefits of HPDs for industrial and agricultural applications in South Africa as depicted in Table 1. Therefore, application of HPD systems in South Africa is of great importance.[4]

Meyer and Greyvenstein[4] performed an economic analysis of drying grain using heat pumps and other methods such as direct electrical heating and diesel engine burning. They discovered that using heat pumps was more economical than using direct electrical heaters, provided that the apparatus was used for 3 months or longer. They also reported that for a working period of less than 3 months, open heat pumps were cheaper than closed heat pumps and vice versa for an active period of 3 months or longer. However, they did not elaborate on the coefficient of performance (COP) of the designed heat pump or the specific moisture extraction rate (SMER) – the main factors which describe the efficiency of the heat pumps when compared to direct electrical dyers.[3]

In South Africa, 80% of farmers use diesel burners to heat the air to dry biological materials such as grains.[4] Most of these diesel burners were installed when diesel was relatively cheap. However, the price of diesel has increased considerably over the past few years and diesel heating is no longer the first choice for drying. South Africa also has experienced an increase in the number and types of drying processes for various industrial, commercial and residential applications, which has resulted in increased energy consumption. It is important to develop the market for HPD systems in South Africa, as the technology is still in its infancy. Most research in this area has been conducted in Europe and Asia (see Table 1). Therefore, the results are based on European and Asian environmental and climatic conditions, and cannot be directly applied to another region with different conditions, like South Africa. With the high consumption of energy in South Africa and pressure from the government and concerned bodies, it is important that such technologies (HPD systems) are investigated for local applications.[4] The objective of this review was to provide an overview of heat pump dryers and a comparison with traditional drying methods in South Africa.

Table 1: Summary of studies on heat pump drying (HPD)

Reference	Location	Application(s)	Conclusion
5,6	Singapore	Green beans	A coefficient of performance (COP) value of above 6 was observed and a specific moisture extraction rate (SMER) above 0.65 was obtained for a material load of 20 kg and a compressor speed of 1200 rpm
7	Turkey	Apples	A system which is composed of the combination of both dryers is considered to be more efficient
8-10	Singapore	Agricultural and marine products	With scheduled drying conditions, the quality of products can be improved
11	Australia	Grain	An open-cycle HPD performed better during the initial stage when the product drying rate was high
4	South Africa	Grains	HPD is more economical than other dryers
12	Brazil	Vegetables (onion)	Better product quality and energy saving of the order of 30% was obtained with HPD
13	Australia	Macadamia nuts	A high quality of dried nuts was observed
14	Norway	Marine products (fish)	The high quality of the dried products was highlighted as the major advantage of HPD
15	Thailand	Agricultural food drying (bananas)	HPD is economically feasible and particularly appropriate for drying materials with a high moisture content
16	New Zealand	Apple	A modified atmosphere heat produces products with a high level of open pore structure, contributing to the unique physical properties
17	Thailand	Garlic and white mulberry	Computer simulation model of the heat pump dehumidified drying was shown to be in good agreement with experimental results
18	Iran	Plums	The optimum temperature of drying for plums is about 70–80 ºC; the SMER of the designed dryer was notably more than conventional types of dryers with respect to saving energy
19	Singapore	Apple, guava and potato	A modified atmosphere heat pump dryer produced better physical properties
20	Brazil	Mango	The energy efficiency was better compared with an electrical resistance dryer
21	Thailand	Fruits (papaya and mango glacé)	Mathematical models of fruits drying using HPD were developed and validated experimentally. The optimum criterion is minimum annual total cost per unit of evaporating water. The effects of initial moisture content, cubic size and effective diffusion coefficient of products on the optimum conditions of HPD were also investigated. Exergy and energy analyses were performed.
22	Turkey	Wool	The SMER was 0.65–1.75 kg/kWh. The COP was 2.47–3.95.
23	New Zealand	Peas	The thin layer drying kinetics model of peas was developed. The model can be used to accurately predict the drying time in a heat pump, thus bringing about energy and cost savings.
24	Turkey	Tropical fruits (kiwi, avocado and banana)	Mathematical models of the drying characteristics of tropical fruits were developed. The results were in good agreement with experimental results.
25	Indonesia	Red chillies	The high quality of the dried chilli was highlighted as the major advantage of HPD compared to sun drying

Historical development of heat pump drying

Traditional methods for drying biological material have been widely employed around the world. As stated before, the most common methods employ direct or indirect sunlight and wood burning. Although these methods are cheap, there are problems associated with them, such as poor-quality dried products, no control over the drying process, possible contamination of the product by dirt, possible interference by rodents and other animals, infestation by insects or mould and exposure of the product to rain and wind, which causes repeated wetting and re-drying. HPD has been found to be more economical than these traditional drying methods.[26]

Heat pumps are heat-generating devices that transfer heat in the opposite direction of spontaneous heat flow by absorbing heat from an area of low temperature and releasing it to a warmer area. Heat pumps are widely used in water and space heating applications. They generally work via vapour-compression cycles or absorption-compression cycles. Although vapour-compression cycles date to 1834, the first commercialised machine was produced in 1850; heat pumps were not originally very popular because of their high installation costs.[27] Heat pumps were first commercially produced in the USA in the 1930s, but only became popular in the 1970s because of reduced operating costs. Approximately one-third of all single-family homes built in the USA were heated by heat pumps in 1984.[27] Recently, progress has been made in alternative industrial applications of heat pumps, especially in the dehumidification and drying of agricultural products.[4,28,29]

Hodgett's[30] and Geeraert's[31] studies reported on the first heat pump dehumidifier in 1973. Hodgett reported that the energy consumption of HPDs was lower than that of conventional steam-heated dryers, and the results concurred with those of Geeraert[31] who studied the application of HPD in timber drying. Tai et al.[32] reported advantages of HPDs such as high energy savings and a wide range of drying conditions with respect to temperature and humidity. Zylla et al.[33] concluded that the SMER increased as the relative humidity of the dryer outlet air increased. Cunney and Williams[34] reported that a well-designed engine-driven heat pump could achieve a reduction of about 30–50% in drying energy cost. Newbert[35] showed that energy consumption could be reduced by 40% when drying malt with a coupled gas engine heat pump dryer. In 1988, about 7% of industrial heat pumps were used for drying. These heat pumps represented an installed capacity of 60 MW.[14] In 1992, Meyer and Greyvenstein[4] analysed the life-cycle cost of HPD applications for grain drying in South Africa.

Principle of HPD

A HPD system consists mainly of two subsystems: a heat pump (refrigeration system) and a drying chamber. Heat pumps can transfer heat from natural heat sources in the surroundings (such as the air, ground or water), or from industrial or domestic waste, or from a chemical reaction or dryer exhaust air. The drying chamber can be a tray, fluid bed, rotary or band conveyor. A basic heat pump dryer consists of a heat pump (including a compressor, a condenser, an expansion valve and an evaporator), a dryer and air cycling circuits, which connect the heat pump and the dryer (Figure 1). The working principle of closed HPDs (as shown in Figure 1) is that the exhausted air from the dryer enters the evaporator of the heat pump, where it is cooled and the moisture in the air is condensed and removed. The cool and dry air from the evaporator then goes into the condenser of the heat pump and is heated. The hot and dry air then enters the dryer and absorbs the moisture in the materials being dried in the dryer and becomes exhausted air at the outlet of the dryer, and the cycle repeats. Because the heat pump retrieves the heat in the exhausted air to heat the air entering the dryer while it removes the moisture in the exhausted air, it achieves a high energy efficiency in the drying of biological materials which are thermally and oxygen sensitive. In an open cycle, exhausted air is not re-circulated and the HPD uses ambient air as the heating source.

The energy efficiency of the HPD is strongly influenced by the relative humidity of the exhausted air from the dryer in the closed-cycle HPDs depicted in Figures 1 and 2. When the relative humidity of the exhausted air is low, a low evaporating temperature in the heat pump evaporator is needed to remove the moisture in the exhausted air, which leads to a large temperature difference between the evaporator and the condenser, resulting in low energy efficiency in the heat pump and the heat pump dryer. The way to remedy this situation is to let some of the exhausted air flow through the evaporator and condenser, and the remaining air bypass the evaporator, and then the two portions mix before the dryer or use a damper as shown in Figure 2 (air mix). This process may raise the relative humidity of the exhausted air and reduces the energy consumption of the heat pump dryer, but it needs a careful capacity control of the heat pump.

Another way to solve the problem of low relative humidity of the exhausted air is by adding a desiccant unit which is parallel with the evaporator to share part of the moisture-removing load, as depicted in Figure 3. There are two parallel air ducts in this system. One is the heat pump air duct which includes air valve 1, the evaporator and condenser. The other is the desiccant duct which includes air valve 2 and the desiccant unit. The desiccant unit is a moisture storage container and is filled with steam adsorbents. The heat pump and the desiccant unit work alternately. The heat pump in this system is different from that in the basic heat pump dryer (Figure 1) in that there are two parallel refrigerant circuits. The first circuit consists of the compressor, valve 1, the heat sink, condenser, expansion valve and evaporator, and the second circuit consists of the compressor, valve 2, heating tubes, expansion valve and the heat sink. The heat pump refrigerant cycles through different circuits in different drying stages. A working period is divided into three stages for the batch drying of biological materials. In the first stage, there is more water in the material and the heat and mass transfer between the hot air and material is extensive, so the relative humidity of the exhausted air is high. In this stage, the heat pump works with valve 1 open and valve 2 closed, and the heat pump refrigerant flows in the first circuit. At the same time, air valve 1 is open and air valve 2 is closed. The exhausted air flows through

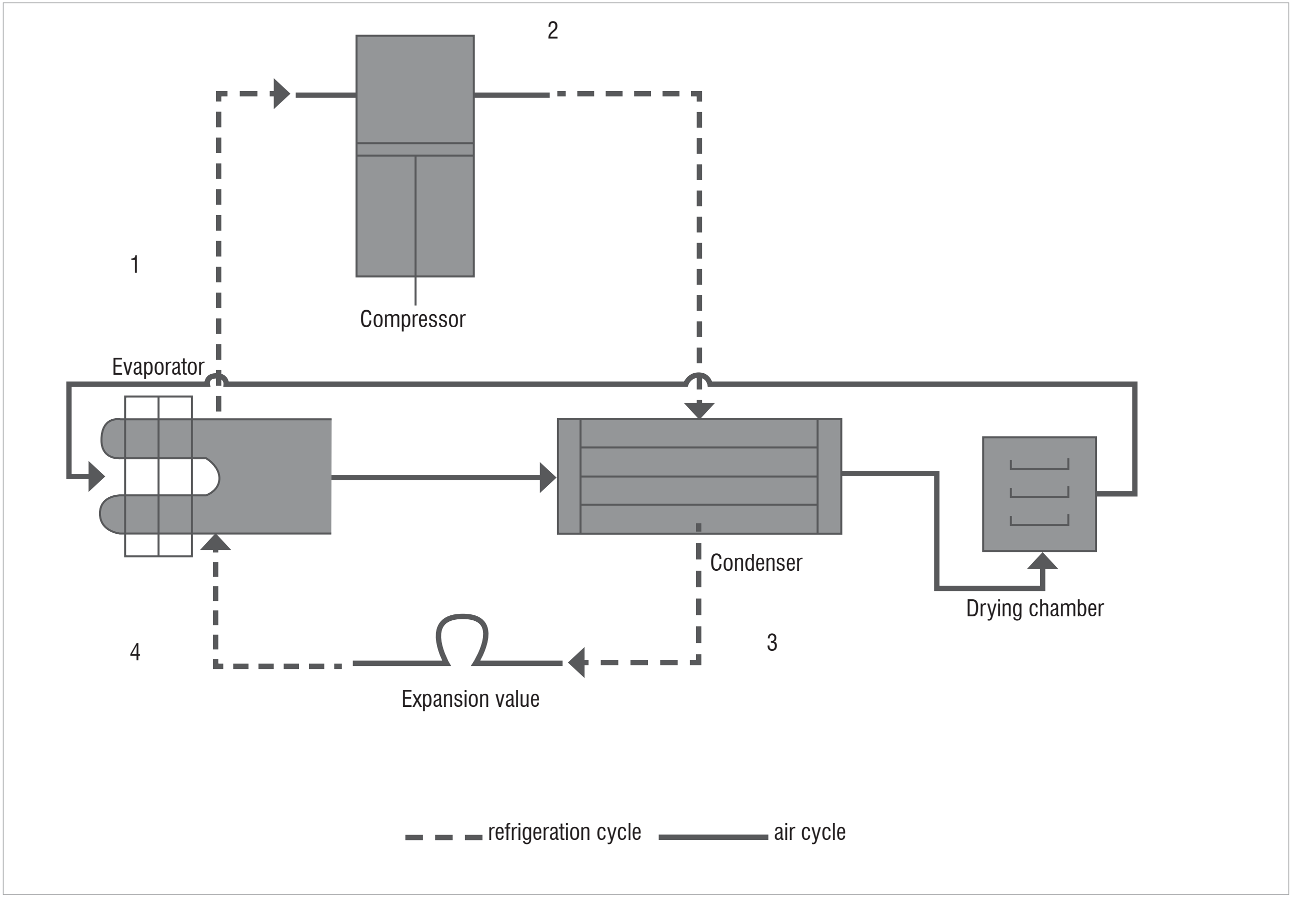


Figure 1: Schematic diagram of a heat pump dryer.

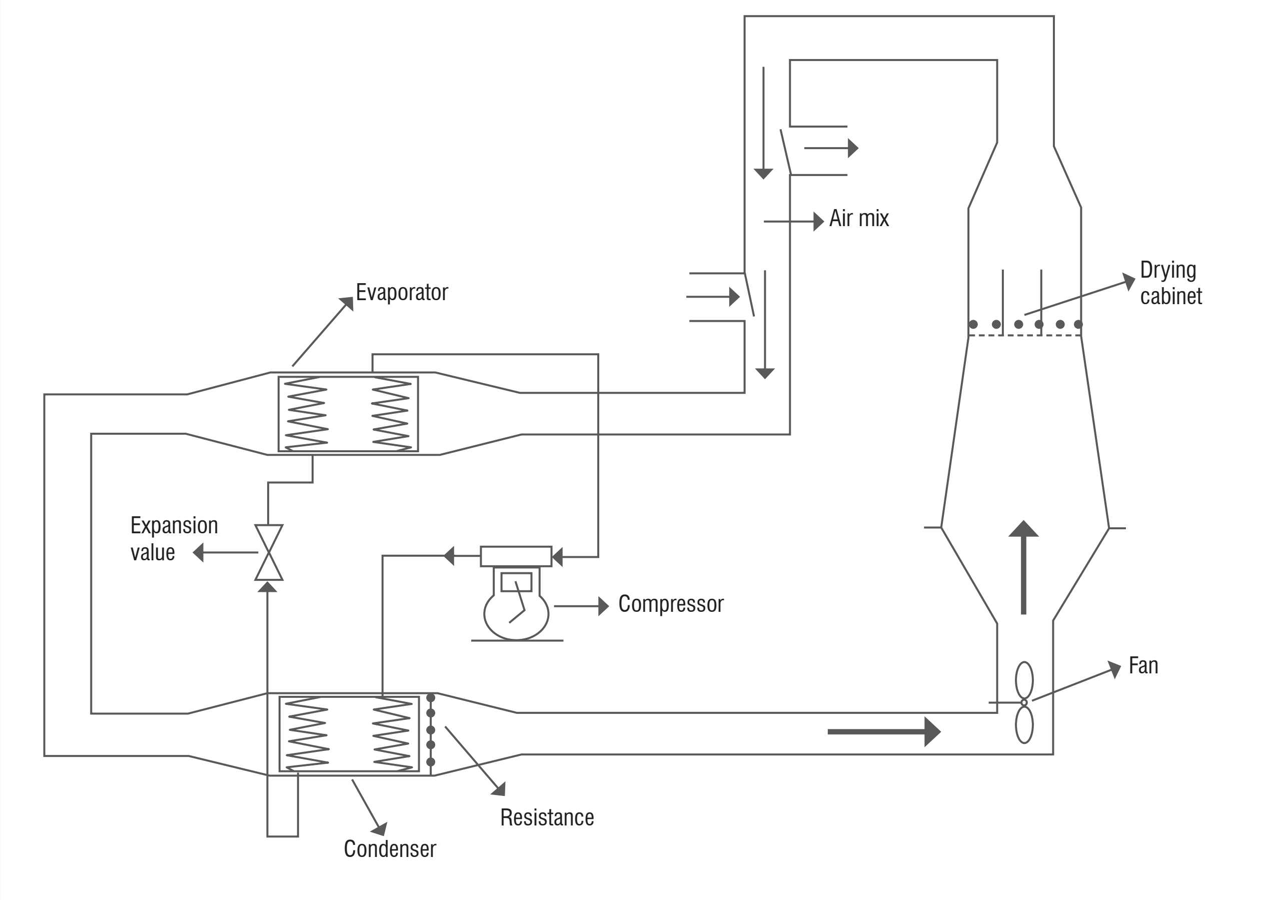


Figure 2: A closed-cycle heat pump dryer.[45]

the heat pump duct, the moisture in the exhausted air is condensed and removed in the evaporator, and then heated in the condenser. After most moisture in the materials has been removed, the hot air cannot get enough water from the materials in the dryer, and so the relative humidity of the exhausted air becomes low, and the drying process enters the second stage.

In the second stage, the heat pump stops, air valve 1 is closed and air valve 2 is open, the exhausted air flows through the desiccant duct, and the desiccant unit begins to work. The moisture in the exhausted air is absorbed by the desiccant, and the dry air is heated simultaneously by the heat of adsorption. When the water content in the materials reduces to the required level, the second stage ends and the drying process of the materials is also over. After the dried materials have been taken out of the dryer, the third stage begins. In this stage, the heat pump works to regenerate the desiccant. As valve 1 is closed and valve 2 is open, the heat pump refrigerant flows in the second circuit, and the heat pump gets heat from the heat sink and heats the desiccant in the desiccant unit. At the same time, air valves 1 and 2 and the dryer door are all open. As fresh air flows in the desiccant unit, it is heated and carries water vapour discharged by the desiccant out of the desiccant unit. When the water content in the desiccant has been reduced to the required level, the third stage ends and the process is complete. The desiccant assisted heat pump dryer has great potential for energy saving for batch drying of thermally sensitive biological materials. The energy consumed by this heat pump can be 30–50% lower than that of a basic heat pump dryer (Figure 1).[36] However, its energy efficiency is mainly affected by steam capacity and the regeneration temperature of the desiccant. The low regeneration temperature can help achieve a high COP of the heat pump (COP_{HP}) and a high SMER.

The performance of HPD is indicated by the COP. The COP is not an efficiency but an energy characteristic. It describes what you get over what you spend. The theoretical and the actual COP can be calculated with the following equation, by either inserting theoretical or real measured values:

$$COP_{HP} = \frac{Q_H}{W_C} \qquad \text{Equation 1}$$

where Q_H is the heat rejected at the condenser and W_c is the work input to the compressor and

$$Q_H = Q_{subcooling} + Q_{condensation} \qquad \text{Equation 2}$$

The integration of a heat pump system into a dryer requires an additional energy consuming unit: a blower. In order to be precise, the energy input for this device must also be included in the calculations. So the overall COP of a heat pump dryer is defined as:

$$COP_{HPD} = \frac{Q_H}{W_c + W_f} \qquad \text{Equation 3}$$

where W_f is the work input to the fan or blower (kJ).

The performance of the dryer is determined by exergy efficiency (η).[37] Recently, the exergy efficiency has been used rather than the energy efficiency in the performance analysis of food processes.[38,39] In drying processes, the driving force behind heat losses is the temperature difference between the drying chamber and the environment.[38] Increasing heat losses and irreversibility decrease the exergy efficiency. The exergetic efficiency of the drying chamber (η) is the ratio of the total

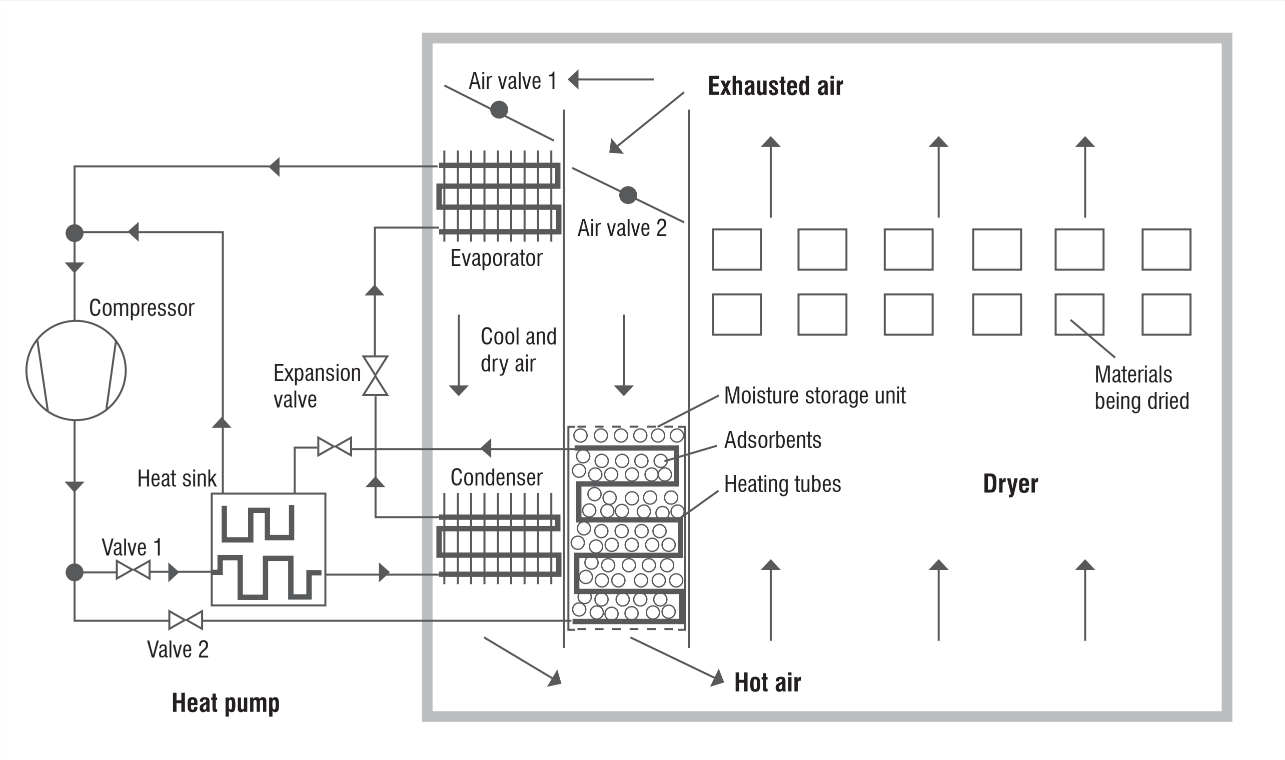


Figure 3: The desiccant-assisted heat pump dryer.

exergy gained by the air stream to the total exergy that enters the system. Thus, the general form of exergy efficiency is written as[38,40]:

$$\eta = \frac{\dot{E}x_{out}}{\dot{E}x_{in}} \times 100 \qquad \text{Equation 4}$$

Therefore, the exergy (*Ex*) values (J/s) at the dryer can be calculated using the general form of the exergy equation applicable for steady-flow systems reported in the literature by Midilli and Kucuk[41] as most applicable for a batch dryer:

$$\dot{E}x = \dot{m}_a C_{pa}\left[(T - T_{ref}) - T_{ref}\ln\frac{T}{T_{ref}}\right] \qquad \text{Equation 5}$$

where $\dot{m}_a$ is the air mass flow rate (kg/s), $C_{pa}$is the thermal capacity of air (J/kg.K), T is the temperature (K) and T_{ref} is the reference (ambient) temperature (K).

The SMER is used to evaluate the performance of a whole heat pump dryer system. The SMER is the only performance measure that considers both the dryer and heat pump system. The SMER reflects directly on how efficient the energy usage is. It is defined as the ratio of the mass of water removed from the product (condensed water in the evaporator in kg) to the required energy for this (kWh).

$$SMER = \frac{\Delta x}{\Delta h} \quad [kg / kWh] \qquad \text{Equation 6}$$

where Δx is the amount of water removed (kg) and Δh is the amount of energy consumed (kJ).

Types of heat pump dryers

There are different types of HPD systems available on the market. Air, chemical, ground source and hybrid systems are discussed in detail here.

Air source heat pump drying systems

Air source heat pumps are the most widely used heat pumps.[42] An air source heat pump (ASHP) is a heating and cooling system that uses outside air as its heat source and heat sink. An ASHP uses a refrigerant system consisting of a compressor and a condenser to absorb heat at one place and release it at another. ASHPs usually are called reverse-cycle air conditioners' when used as space heaters. In domestic use, an ASHP absorbs heat from the outside air and releases it inside during winter, and can often do the converse in summer. When appropriately set up, an ASHP can offer a full central heating solution and domestic hot water with an efficiency of up to 80%.[43,44] However, in the past few years, ASHPs have also become applicable for the drying of biological materials. Xanthopoulos et al.[45] designed a closed-cycle ASHP for drying whole figs. In their study, they presented a mathematical single-layer drying model to predict the drying rate of whole figs. They concluded that, among the models tested, the best model in terms of fit was the logarithmic model. The ASHP is shown schematically in Figure 2. Chua and Chou[46] designed, fabricated and tested a two-stage prototype evaporator heat pump assisted mechanical drying system for enhancing heat recovery, which performed efficiently when compared with other traditional drying methods. More studies on ASHP are outlined in Table 1.

Chemical heat pump drying system

Chemical heat pumps (CHPs) are those systems that utilise a reversible chemical reaction to change the temperature of the thermal energy stored by chemical substances.[47] These chemical substances play an important role in absorbing and releasing heat.[48] The advantages of thermochemical energy storage – high storage capacity, long-term storage of both reactants and products, lower heat loss, etc. – make CHPs a good option for energy upgrading of low temperature heat as well as storage. Sources of low temperature heat are industrial heat waste, solar energy, dryer exhaust, geothermal energy, etc.[49] The working principle of CHP comprises adsorption/synthesis/production and desorption/regeneration/decomposition. The synthesis stage is the cold production stage, which is followed by the regeneration stage, during which decomposition takes place. This regeneration stage can take place in the same or different reactors depending on the system design. Figure 4 depicts a simple CHP with its main components.

In summary, CHP systems are a potentially significant technology for effective energy utilisation in drying. CHPs are designed in a way that they can store energy in the form of chemical energy via endothermic reactions in a suitably designed reactor. Energy is released at various temperatures during the heat-demand period by exo- or endothermic reactions.[42] Several researchers have recommended CHP. Ogura and Mujumdar[50] reported that the calcium oxide hydration/dehydration system was found to be the most feasible for CHP dryer systems from the perspective of temperature level, safety, corrosion and cost. Ogura et al.[51] undertook an experimental study focusing on the heat and mass transfer performance in batch drying using a CHP and found the performance to be highly efficient. Rolf and Corp[52] developed a CHP dryer for the drying of bulky materials such as bark and lumber and recommended this method of drying as it is easy to adapt to any industrial drying process, specifically the drying processes in the pulp and paper industries.

Ground source heat pump drying systems

There is limited information on ground heat pump drying systems, despite the many studies that have been undertaken. Figure 5 is a schematic illustration of a ground source HPD system.[47] This system consists of three subsystems: a ground source heat exchanger, a heat pump system and a drying chamber. The main components of the heat pump system are an evaporator, a condenser, a compressor and an expansion valve. In this system, heat is extracted from the ground by the ground source heat exchanger, which contains a circulating water–antifreeze solution. The heat is transferred to the refrigerant in the evaporator, added to the heat pump cycle, and supplied to the drying chamber. In here, heat is transferred to the drying air and this heated air enters the drying chamber.[42,47]

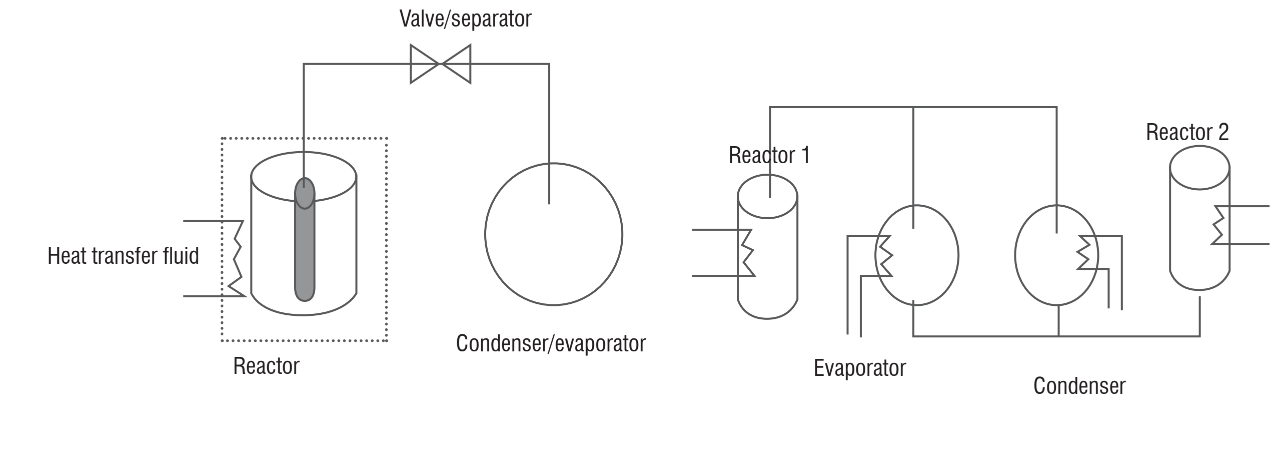


Figure 4: A simple chemical heat pump.[29]

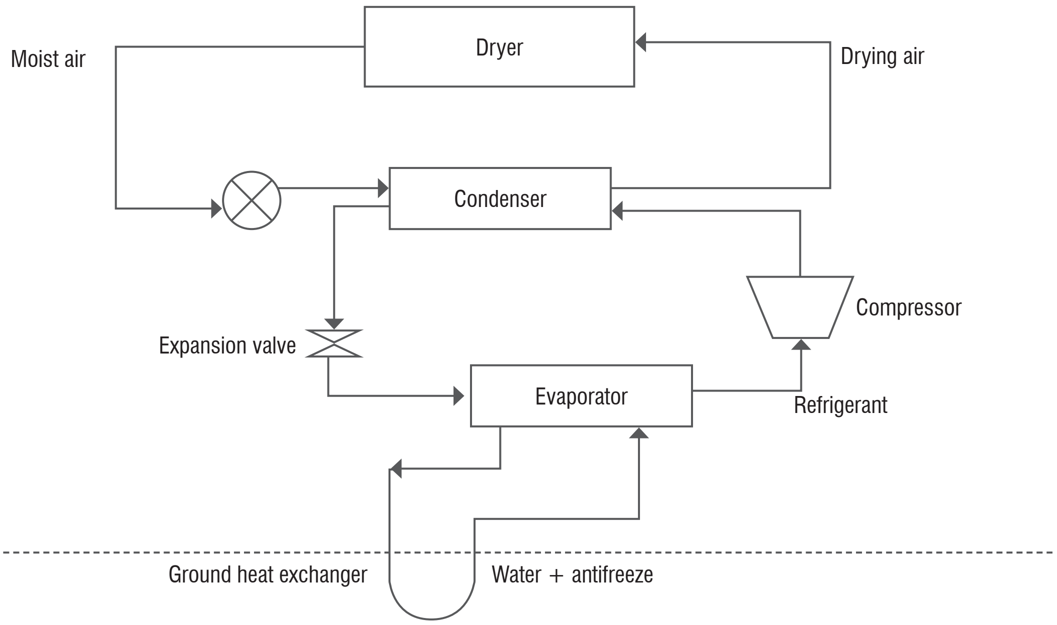


Figure 5: Schematic diagram of a ground source heat pump dryer.[47]

Hybrid heat pump drying systems

The hybrid heat pump drying system includes a solar-assisted heat pump drying system, microwave and infrared or convectional energy.[42] To date, few studies have been conducted on hybrid drying because of its higher initial costs, although this technology has great energy-saving potential compared with other technologies because of its multipurposeful applications. For example, Hawlader and Jahangeer[5] investigated a hybrid solar-assisted heat pump drying system which also functioned as a water heater. Hybrid concepts are likely to make rapid inroads into industrial drying in the coming decades. These concepts consist of intelligent combinations of well-established drying technologies and hence involve less risk. Hybrid HPD systems combine the advantages as well as the limitations of each individual technology. Hence, care must be exercised in their design.[7]

Types of dryers

The two most commonly used dryers in the market are batch dryers and conveyor dryers, which will be discussed further here. However, there are other dryers which are used and reported on in the literature, such as fluidised beds and rotary dryers.[42]

Batch dryers

Batch dryers are more widely reported on than other dryers. Batch drying systems allow total recirculation with a very low air leakage rate, giving rise to high thermal efficiencies.[42] They are also good for low capacity applications, such as laboratory experiments.[28]

Conveyor dryers

Continuous bed drying or /conveyor dryers show promising results compared with batch dryers and are potentially a better option for drying specialty crops. However, very few studies have been done on these types of dryers, possibly because they are suitable for high capacity needs.[42,53]

Advantages and limitations of heat pump dryers

The main advantages of HPD are:

- High energy efficiency (reduced energy consumption) – up to 60% reduction in energy costs compared with traditional drying technologies.
- Controlled temperature profile to meet product requirements – sensors and advanced controllers are used to adjust the temperatures of the condenser and evaporator to obtain suitable drying temperatures, which is not possible with traditional drying methods. It is also possible to control the speed of the fans to achieve optimal air flow.
- More environmentally friendly – 80–100% reduction in emissions of chemicals released as a result of drying some products and up to 60% reduction in CO_2 emissions.
- Consistent output of products – heat pumps can operate 24 hours per day, so the production potential is higher than for traditional drying.
- A wide range of drying conditions (from -20 °C to 80 °C) – because of the moderate climate in South Africa, heat pumps can operate under ambient conditions, which saves energy compared to Asian and European countries.
- Better product quality – HPD conditions such as temperature and air flow rates are controlled to meet specific requirements.
- Business opportunities for both farmers and industry.

The limitations of HPD are:

- Higher initial costs – the initial costs of HPD may be higher than those for traditional drying methods. Most of the initial costs pertain to equipment such as the controllers, compressor, heat exchangers. However, if the period of use is longer than 1 year, the return on investment is received within a short period.
- Refrigerant leakage – refrigerant system pressure can cause cracks in the pipes and valves, resulting in the leakage of refrigerant, which pollutes the environment. Once a leakage occurs, the pressure drops and performance is reduced. In order to reduce gas emissions that harm the environment, the use of green refrigerants such as carbon dioxide and nitrogen oxides has been encouraged.[54,55]
- Maintenance – the compressor, heat exchangers (condensers and evaporator) and refrigerant filters need regular maintenance to keep the dryers operating optimally. Charging of the refrigerant is required when any leak is detected.

Mechanism of drying

The mechanism of drying is a complex phenomenon involving combined heat and mass transfer, which, in most cases, results in products with modified properties. Depending on the drying conditions, food products may undergo various degrees of browning, shrinkage, loss of nutrients, and so on. According to Chou and Chua[1], the degradation of food occurs mainly in three areas: chemical, physical and nutritional (Table 2). It should also be noted that when the product loses moisture during drying, the concentration of nutrients in the remaining mass is increased. Hence, proteins, fats and carbohydrates are present in larger amounts per unit weight in dried food than in fresh food. Foods like fruits and vegetables consist of water, carbohydrates, proteins and fractions of lipids. These compounds are easily modified by high temperatures which results in degraded food quality.[56] Thus, the use of appropriate temperatures is important during drying.

Table 2: Factors that influence food quality during drying

Chemical	Physical	Nutritional
Browning reaction	Rehydration	Vitamin loss
Lipid oxidation	Solubility	Protein loss
Colour loss	Texture	Microbial survival
Gelatinisation	Aroma loss	

The drying of biological materials follows a falling rate profile, and the falling rate period is controlled by the mechanism of liquid and/or vapour diffusion. Thin-layer drying models that describe the drying of these materials fall mainly fall into three categories – theoretical, semi-theoretical and empirical – which are generally based in mass transfer, neglecting the effect of heat transfer. Assuming that the resistance to moisture flow is uniformly distributed throughout the material, the diffusion coefficient is constant and the volume shrinkage is negligible, Fick's second law of diffusion can be stated as[57]:

$$\frac{dM}{dt} = D\,\frac{d^2M}{dr^2} \qquad \text{Equation 7}$$

where M is the local moisture content (kg water/kg dry solids), r is the diffusion path (m), t is the time (s) and D is the moisture dependent diffusivity (m^2/s).

The analytical solution of Equation 7 was given by Crank[57] for various regularly shaped bodies, such as rectangular, cylindrical and spherical. The drying of many foodstuffs such as tomatoes[58], carrots[59], pine nut seeds[60] and pineapple[61] has been predicted using the analytical solution of Equation 7. It should be noted that most cereals, such as rice, corn and wheat, change little in volume during the drying process. Therefore, the analytical solution of Equation 7 applies satisfactorily to the study of these materials. However, for foods with a high moisture content, such as fruits and vegetables, the analytical solution of Equation 7 obtained for constant diffusivity and volume is not always applicable, because shrinkage and diffusivity as functions of moisture content often need to be taken into account. In these cases, simpler models should be a wiser option.[62]

The Lewis model is analogous to Newton's law of cooling and assumes that the internal resistance to moisture movement, and thus the moisture gradients within the material, is negligible.[23] The model considers only the surface resistance and is given by:

$$M = \frac{M - M_E}{M_O - M_E} \exp\,(-k_L t) \qquad \text{Equation 8}$$

where M_O and M_E are the initial and equilibrium moisture contents, respectively, and k_L is the Lewis drying coefficient. This model was used primarily because it is simple. The only drawback of this model is that it tends to overpredict the early stages of the drying curve.[63] The Page model was introduced as a modification of Equation 8 to overcome this shortcoming. This model has produced good fits in predicting drying times of food and agricultural materials.

$$M = \frac{M - M_E}{M_O - M_E} \exp\,(-k_p t^n) \qquad \text{Equation 9}$$

where k_p and n are the Page drying coefficients which determine the precise shape of the drying curve. While neither of them has a direct physical significance, empirical regression equations have been developed which relate both parameters to drying conditions.[64-66] Therefore, the drying rate for the Page equation is given by:

$$\frac{dM}{dt} = (-k_p n t^{n-1})\ (M - M_E) \qquad \text{Equation 10}$$

A modified Page drying coefficient can be defined as:

$$k^* = k_p n t^{n-1} \qquad \text{Equation 11}$$

If $n<1$, k^* decreases during the drying process. Thus higher values of k_p can be used to more closely approximate the diffusion equation in the initial stages of drying without overpredicting drying in the later stages. If $n=1$, Equations 9 and 10 reduce to the Lewis model, approaching the solution of diffusivity equation. Several authors have compared different drying models and have obtained better results for the Page equation than for other existing drying models.[64-67]

Comparison of HPD with South African traditional drying methods

Currently, the market for heat pump systems for water heating and space cooling/heating is well developed in South Africa. However, many developed and other developing countries are interested in applications of heat pumps for energy saving and they have invested much money and time on HPD research. However, the development of heat pumps for industrial and agricultural drying in South Africa is still in its infancy, possibly because:

- The concept of HPDs is not well known in South Africa and is less understood than that of traditional drying methods
- The prices of electricity and fossil fuels have rapidly increased only recently
- There is a lack of techno-economical information regarding HPDs
- Products of HPD have become known in South Africa only recently because the market has increased in the past few years.

A comparison of HPD with the most common South African traditional drying methods for biological materials is given in Table 3. It is clear that the advantages of HPD far outweigh those of traditional South African methods for drying biological materials. South Africa currently is experiencing an increased need for drying processes for various industrial, commercial and residential applications. Therefore, development of HPD systems in South Africa is imperative. The development of HPD systems will not only reduce energy consumption – by 6% if utilised effectively – and be more environmentally friendly, but will also increase business opportunities for both farmers and industrialists.[4]

Table 3: Comparison of heat pump drying with traditional South African drying methods[4,26,28]

Item	Direct/indirect sunlight	Wood burning	Electrical heating	Diesel engine heating	Heat pump dryers
Efficiency	Very low	Low	Low	Low	High
Temperature range (°C)	≤40	10–100	0–100	10–100	10–65
Operation time	During the day	Anytime	Anytime	Anytime	Anytime
Drying air flow	Depends on wind	N/A	Controlled by fan speed	N/A	Controlled by fan speed
Moisture extraction	Depends on weather	Low	Accurate	Low	Accurate
Temperature control	N/A	Fair	Accurate	OK	Accurate
Product quality	Depends on weather	Bad, because of smoke	Good	Bad, because of smoke	Very good
Environmentally friendly	Yes	No	Yes	No, because of CO_2 emissions	Yes
Weather effect	Yes	No, if operation is indoors	No, if operation is indoors	No, if operation is indoors	No, if operation is indoors
Initial capital cost	Low	High	High	High	Very high
Payback period	N/A	N/A	Long	Long	Short
Maintenance costs	Very low	Low	High	High	High
Application range	Limited	Limited	Wide	Limited	Wide
Operational control	N/A	Limited	Available	Available	Available
Noise level	None	Low	Moderate	High	Moderate
Energy-waste level	N/A	High	High	High	Low

N/A, not applicable

Comparison of different drying methods

A comparison of the efficiencies and advantages of heat pump dryers, vacuum and hot air dryers is shown in Table 4. Heat pump dryers have a higher SMER range (1.0–4.0 kg H_20/kWh) than other drying methods. HPD is therefore an efficient and energy-saving alternative for drying industries.

Table 4: Comparison of heat pump drying with vacuum and hot air drying[28]

Item	Hot air drying	Vacuum drying	Heat pump drying
Specific moisture extraction rate (kg H_20/kWh)	0.12–1.28	0.72–1.2	1.0–4.0
Drying efficiency (%)	35–40	≤70	95
Operating temperature range (°C)	40–90	30–60	10–65
Operating % relative humidity range	Variable	Low	10–65
Capital cost	Low	High	Moderate
Running cost	High	Very high	Low

Economic analysis

Limited studies have been conducted comparing the economics of HPD and other convection electrical dryers.[2,68,69] Meyer and Greyvenstein[4] carried out a techno-economic analysis of grain drying using HPDs in South Africa. They found that the life-cycle cost of an electrical heater and diesel engine were three and four times higher, respectively, than that of HPD systems. Teeboonma et al.[21] reported on the optimisation of heat pump fruit dryers and analysed the annual total cost per unit of evaporating water and found that this cost was linearly proportional to both interest rate and electricity price, and decreased with increasing lifetime. In general, heat pumps have enormous potential for saving energy simply because they are the only heat-recovery systems which enable the temperature of the waste heat to be raised to a more useful level.[68]

Conclusion

Different types of heat pump drying systems are appropriate for the drying of many products, especially heat-sensitive products. Many researchers have concluded that HPDs use energy more efficiently than electrical dryers. Also, the quality of heat pump dried products is better than those of conventional drying systems. The widely reported measures for determining efficiency of a HPD system are SMER and COP. Moreover, it has been found that the desiccant-assisted heat pump dryer has a greater energy-saving potential than basic heat pump dryers for batch drying of thermally sensitive biological materials. In comparison with different South African traditional drying methods, the advantages of a HPD are outstanding. Now is the time to develop and expand the applications of HPD systems in South Africa, especially in industrial and agricultural drying. The market for HPDs will benefit South Africa in many sectors and will help to reduce the high energy consumption that South Africa has been experiencing recently.

Acknowledgements

We acknowledge funding from the National Research Foundation (South Africa) and Tshwane University of Technology in support of this study.

Authors' contributions

T.K. was responsible for writing the manuscript; Z.H. was the project leader and provided advice on the data analysis and manuscript structure.

References

1. Chou SK, Chua KJ. New hybrid drying technologies for heat sensitive foodstuffs. Trends Food Sci Tech. 2001;12(10):359–369. http://dx.doi.org/10.1016/S0924-2244(01)00102-9

2. Chua KJ, Chou SK, Yang WM. Advances in heat pump systems: A review. Appl Energ. 2010;87(12):3611–3624. http://dx.doi.org/10.1016/j.apenergy.2010.06.014

3. Patel K, Kar A. Heat pump assisted drying of agricultural produce – An overview.J Food Sci Tech. 2012;49(2):142–160. http://dx.doi.org/10.1007/s13197-011-0334-z

4. Meyer JP, Greyvenstein GP. The drying of grain with heat pumps in South Africa: A techno-economic analysis. Int J Energ Res. 1992;16:13–20. http://dx.doi.org/10.1002/er.4440160103

5. Hawlader MNA, Jahangeer KA. Solar heat pump drying and water heating in the tropics. Sol Energy. 2006;80(5):492–499. http://dx.doi.org/10.1016/j.solener.2005.04.012

6. Hawlader MNA, Chou SK, Jahangeer KA, Rahman SMA, Lau KWE. Solar-assisted heat-pump dryer and water heater. Appl Energ. 2003;74(1–2):185–193. http://dx.doi.org/10.1016/S0306-2619(02)00145-9

7. Aktaş M, Ceylan I, Yilmaz S. Determination of drying characteristics of apples in a heat pump and solar dryer. Desalination. 2009;239(1–3):266–275. http://dx.doi.org/10.1016/j.desal.2008.03.023

8. Chou SK, Chua KJ, Hawlader MNA, Ho JC. A two-stage heat pump dryer for better heat recovery and product quality. Journal of the Institute of Engineers, Singapore. 1998;38:8–14.

9. Chou SK, Chua KJ, Mujumdar AS, Tan M, Tan SL. Study on the osmotic pre-treatment and infrared radiation on drying kinetics and colour changes during drying of agricultural products. ASEAN J Sci Technol Dev. 2001;18(1):11–23.

10. Chua KJ, Mujumdar AS, Chou SK, Hawlader MNA, Ho JC. Convective drying of banana, guava and potato pieces: Effect of cyclical variations of air temperature on convective drying kinetics and colour change. Dry Technol. 2000;18(5):907–936. http://dx.doi.org/10.1080/07373930008917744

11. Theerakulpisut S. Modeling heat pump grain drying system [PhD thesis]. Melbourne: University of Melbourne; 1990.

12. Jangam SV, Thorat BN. Optimization of spray drying of ginger extract. Dry Technol. 2010;28(12):1426–1434. http://dx.doi.org/10.1080/07373937.2010.482699

13. Mason RL, Blarcom AV. Drying macadamia nuts using a heat pump dehumidifier. In: The development and application of heat pump dryers. Seminar papers, 24th March 1993. Brisbane: The Seminar; 1993. p. 1–7.

14. Strommen I, Kramer K. New applications of heat pumps in drying process. Dry Technol. 1994;12(4):889–901. http://dx.doi.org/10.1080/07373939408960000

15. Prasertsan S, Saen-saby P. Heat pump drying of agricultural materials. Dry Technol. 1998;16(1–2):235–250. http://dx.doi.org/10.1080/07373939808917401

16. O'Neill MB, Rahman MS, Perera CO, Smith B, Melton LD. Colour and density of apple cubes dried in air and modified atmosphere. Int J Food Prop. 1998;1(3):197–205. http://dx.doi.org/10.1080/10942919809524577

17. Phoungchandang S. Simulation model for heat pump-assisted dehumidified air drying for some herbs. World J Agric Sci. 2009;5(2):138–142.

18. Chegini G, Khayaei J, Rostami HA, Sanjari AR. Designing of a heat pump dryer for drying of plum. J Res Appl Agric Eng. 2007;52(2):63–65.

19. Hawlader MNA, Perera CO, Tian M. Properties of modified atmosphere heat pump dried foods.Food Eng. 2006;74:392–401. http://dx.doi.org/10.1016/j.jfoodeng.2005.03.028

20. Kohayakawa MN, Silveria-Junior V, Telis-Romero J. Drying of mango slices using heat pump dryer. Proceedings of the 14th International Drying Symposium; 2004 August 22–25; Sao Paulo, Brazil. p. 884–891.

21. Teeboonma U, Tiansuwan J, Soponronnarit S. Optimization of heat pump fruit dryers. J Food Eng. 2003;59(4):369–377. http://dx.doi.org/10.1016/S0260-8774(02)00496-X

22. Oktay Z, Hepbasil A. Performance evaluation of a heat pump assisted mechanical opener dryer. Energ Convers Manage. 2003;44:1193–1207. http://dx.doi.org/10.1016/S0196-8904(02)00140-1

23. Rahman MS, Perera CO, Thebaud C. Desorption isotherm and heat pump drying kinetics of peas. Food Res Int. 1998;30(7):485–491. http://dx.doi.org/10.1016/S0963-9969(98)00009-X

24. Ceylan I, Aktaş M, Doğan H. Mathematical modeling of drying characteristics of tropical fruits. Appl Therm Eng. 2007;27(11–12):1931–1936. http://dx.doi.org/10.1016/j.applthermaleng.2006.12.020

25. Marnoto T, Sulistyowati E, Syahri MM. The characteristic of heat pump dehumidifier drier in the drying of red chili (*Capsicum annum* L). Int J Sci Eng. 2012;3(1):22–25.

26. Chua KJ, Chou SK, Ho JC, Hawlader MNA. Heat pump drying: Recent development and future trends. Dry Technol. 2002;20(8):1579–1610. http://dx.doi.org/10.1081/DRT-120014053

27. Calm JM. Heat pumps in USA. Int J Refrig. 1997;10:190–196. http://dx.doi.org/10.1016/0140-7007(87)90050-8

28. Perera CO, Rahman MS. Heat pump dehumidifier drying of food. Trends Food Sci Tech. 1997;8(3):75–79. http://dx.doi.org/10.1016/S0924-2244(97)01013-3

29. Wongsuwan W, Kumar S, Neveu P, Meunier F. A review of chemical heat pump technology and applications. Appl Therm Eng. 2001;21(15):1489–1519. http://dx.doi.org/10.1016/S1359-4311(01)00022-9

30. Hodgett DL. Efficient drying using heat pump. Chem Eng. 1976;311(July/August):510–512.

31. Geeraert B. Air drying by heat pumps with special reference to timber drying. In: Camatini E, Kester T, editors. Heat pumps and their contribution to energy conservation. NATO Advanced Study Institute Series, Series E, Applied Sciences. Leiden: Noordhoff; 1976. p. 219–246. http://dx.doi.org/10.1007/978-94-011-7571-5_8

32. Tai KW, Devotta S, Watson RA, Holland FA. The potential for heat pumps in drying and dehumidification systems III: An experimental assessment of the heat pump characteristics of a heat pump dehumidification system using R114. Int J Energ Res. 1982;6:333–340. http://dx.doi.org/10.1002/er.4440060404

33. Zylla R, Abbas P, Tai KW, Devotta S, Watson FA, Holland FA. The potential for heat pumps in drying and dehumidification systems I: Theoretical considerations. Energy Res. 1982;6:305–322. http://dx.doi.org/10.1002/er.4440060402

34. Cunney MB, Williams P. An engine-driven heat pump applied to grain drying and chilling. In: Watts GA, Stanbury JEA, editors. Proceedings of the 2nd International Symposium on the Large Scale Applications of Heat Pumps; 1984 Sep 25–27; York, England. Cranfield, Bedford: BHRA; 1984. p. 283–294.

35. Newbert GJ. Energy efficient drying, evaporation and similar processes. J Heat Recov Syst. 1985;5:551–559. http://dx.doi.org/10.1016/0198-7593(85)90223-1

36. Atuonwu JC, Jin X, Van Straten G, Van Deventer Antonius HC, Van Boxtel JB. Reducing energy consumption in food drying: Opportunities in desiccant adsorption and other dehumidification strategies. Procedia Food Sci. 2011;1:1799–1805. http://dx.doi.org/10.1016/j.profoo.2011.09.264

37. Erbay Z, Icier F. Optimization of drying of olive leaves in a pilot-scale heat pump dryer. Dry Technol. 2009;27(3):416–427. http://dx.doi.org/10.1080/07373930802683021

38. Dincer I, Sahin AZ. A new model for thermodynamic analysis of a drying process. Int J Heat Mass Tran. 2004;47(4):645–652. http://dx.doi.org/10.1016/j.ijheatmasstransfer.2003.08.013

39. Corzo O, Bracho N, Vasquez A, Pereira A. Energy and exergy analyses of thin layer drying of coroba slices. J Food Eng. 2008;86:151–161. http://dx.doi.org/10.1016/j.jfoodeng.2007.05.008

40. Akpinar EK. Energy and exergy analyses of drying of red pepper slices in a convective type dryer. Int Commun Heat Mass. 2004;31(8):1165–1176. http://dx.doi.org/10.1016/j.icheatmasstransfer.2004.08.014

41. Midilli A, Kucuk H. Energy and exergy analyses of solar drying process of pistachio. Energy. 2003;28(6):539–556. http://dx.doi.org/10.1016/S0360-5442(02)00158-5

42. Colak N, Hepbasli A. A review of heat-pump drying (HPD): Part 2 – Applications and performance assessments. Energ Convers Manage. 2009;50(9):2187–2199. http://dx.doi.org/10.1016/j.enconman.2009.04.037

43. Daghigh R, Ruslan MH, Zaharim A, Sopian K. Air source heat pump system for drying application. ICOSSSE-63 Conference; 2010 Oct 4–6; Japan. Stevens Point, WI: World Scientific and Engineering Academy and Society (WSEAS); 2010. p. 404–409.

44. Daghigh R, Ruslan MH, Sulaiman MY, Sopian K. Review of solar assisted heat pump drying systems for agricultural and marine products. Renew Sust Energ Rev. 2010;14(9):2564–2579. http://dx.doi.org/10.1016/j.rser.2010.04.004

45. Xanthopoulos G, Oikonomou N, Lambrinos G. Applicability of a single-layer drying model to predict the drying rate of whole figs. J Food Eng. 2007;81(3):553–559. http://dx.doi.org/10.1016/j.jfoodeng.2006.11.033

46. Chua KJ, Chou SK. A modular approach to study the performance of a two-stage heat pump system for drying. Appl Therm Eng. 2005;25(8–9):1363–1379. http://dx.doi.org/10.1016/j.applthermaleng.2004.08.012

47. Colak N; Hepbasli A. Exergy analysis of drying of apple in a heat pump dryer. In: Second International Conference of the Food Industries and Nutrition Division on Future Trends in Food Science and Nutrition; 2005 Nov 27–29; Cairo, Egypt. p. 145–158.

48. Goetz V, Elie F, Spinner B. The structure and performance of single effect solid–gas chemical heat pumps. Heat Recov Syst CHP. 1991;13(1):79–96. http://dx.doi.org/10.1016/0890-4332(93)90027-S

49. Mbaye M, Aidoun Z, Valkov V, Legault A. Analysis of chemical heat pumps (CHPs): Basic concepts and numerical model description. Appl Therm Eng. 1998;18(3–4):131–146. http://dx.doi.org/10.1016/S1359-4311(97)00027-6

50. Ogura H, Mujumdar AS. Proposal for a novel chemical heat pump dryer. Dry Technol. 2000;18(4–5):1033–1053. http://dx.doi.org/10.1080/07373930008917752

51. Ogura H, Yamamoto T, Kage H, Matsuno Y, Mujumdar AS. Effects of heat exchange condition on hot air production by a chemical heat pump dryer using CaO/H2O/Ca(OH)2 reaction. Chem Eng J. 2002;86(1–2):3–10. http://dx.doi.org/10.1016/S1385-8947(01)00265-0

52. Rolf R, Corp R. Chemical heat pump for drying of bark. Annual Meeting of the Canadian Pulp and Paper Association; 1990. Montreal: Technical Section, Canadian Pulp and Paper Association; 1990. p. 307–311.

53. Adapa PK, Schoenau GJ. Re-circulating heat pump assisted continuous bed drying and energy analysis. Int J Energ Res. 2005;29:961–972. http://dx.doi.org/10.1002/er.1103

54. Schmidtt EL, Kliicker K, Flacke N, Steimle F. Applying the transcritical CO_2 process to a drying heat pump. Int J Refrig. 1998;21(3):202–211. http://dx.doi.org/10.1016/S0140-7007(98)00021-8

55. Sarkar J, Bhattacharyya S, Gopal MR. Transcritical CO_2 heat pump dryer: Part 1. Mathematical model and simulation. Dry Technol. 2006;24(12):1583–1591. http://dx.doi.org/10.1080/07373930601030903

56. Mujumdar A. Handbook of industrial drying. 2nd ed. New York: Marcel Dekker; 1987.

57. Crank J. The mathematics of diffusion. Oxford: Pergamon Press; 1975.

58. Hawlader MNA, Uddin MS, Ho AB, Teng ABW. Drying characteristics of tomatoes. J Food Eng. 1991;14:259–268. http://dx.doi.org/10.1016/0260-8774(91)90017-M

59. Reyes A, Alvarez PI, Marquardt FH. Drying of carrots in a fluidized bed: Effects of drying conditions and modeling. Dry Technol. 2002;20(7):1463–1483. http://dx.doi.org/10.1081/DRT-120005862

60. Karatas S, Pinarli I. Determination of moisture diffusivity of pine nut seeds. Dry Technol. 2001;19(3–4):701–708. http://dx.doi.org/10.1081/DRT-100103946

61. Nicoleti JF, Telis-Romero J, Telis VRN. Air-drying of fresh and osmotically pre-treated pineapple slices: Fixed air temperature versus fixed slice temperature drying kinetics. Dry Technol. 2001;19(9):2175–2191. http://dx.doi.org/10.1081/DRT-100107493

62. Madamba PS, Driscoll RH, Buckle KA. The thin-layer dryingncharacteristics of garlic slices. J Food Eng. 1996;29:75–97. http://dx.doi.org/10.1016/0260-8774(95)00062-3

63. Parti M. A theoretical model for thin-layer grain drying. Dry Technol. 1990;8(1):101–122. http://dx.doi.org/10.1080/07373939008959866

64. Wang J. A single-layer model for far-infrared radiation drying of onion slices. Dry Technol. 2002;20(10):1941–1953. http://dx.doi.org/10.1081/DRT-120015577

65. Hossain MA, Bala BK. Thin-layer drying characteristics for green chilli. Dry Technol. 2002;20(2):489–505. http://dx.doi.org/10.1081/DRT-120002553

66. Cronin K, Kearney S. Monte Carlo modeling of a vegetable tray dryer. J Food Eng. 1998;35:233–250. http://dx.doi.org/10.1016/S0260-8774(98)00011-9

67. Panchariya PC, Popovic D, Sharma AL. Thin-layer modeling of black tea drying process. J Food Eng. 2002;52:349–357. http://dx.doi.org/10.1016/S0260-8774(01)00126-1

68. Söylemez MS. Optimum heat pump in drying systems with waste heat recovery. J Food Eng. 2006;74(3):292–298. http://dx.doi.org/10.1016/j.jfoodeng.2005.03.020

69. Hepbasli A, Colak N, Hancioglu E, Icier F, Erbay Z. Exergoeconomic analysis of plum drying in a heat pump conveyor dryer. Dry Technol. 2010;28(12):1385–1395. http://dx.doi.org/10.1080/07373937.2010.482843

Ecological consequences of global climate change for freshwater ecosystems in South Africa

AUTHORS:
Helen F. Dallas[1,2]
Nicholas Rivers-Moore[3]

AFFILIATIONS:
[1]Department of Botany, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
[2]Freshwater Research Centre, Scarborough, South Africa
[3]Centre for Water Resources Research, School of Agricultural, Earth and Environmental Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

CORRESPONDENCE TO:
Helen Dallas

EMAIL:
helen@frcsa.org.za

POSTAL ADDRESS:
Freshwater Research Centre, PO Box 43966, Scarborough 7975, South Africa

DATES:
Received: 01 Sep. 2013
Revised: 21 Nov. 2013
Accepted: 28 Nov. 2013

KEYWORDS:
temperature; flow regimes; habitat; biological assemblages; conservation

HOW TO CITE:
Dallas HF, Rivers-Moore N. Ecological consequences of global climate change for freshwater ecosystems in South Africa. S Afr J Sci. 2014;110(5/6), Art. #2013-0274, 11 pages. http://dx.doi.org/10.1590/sajs.2014/20130274

Freshwater resources in South Africa are under severe pressure from existing anthropogenic impacts and global climate change is likely to exacerbate this stress. This review outlines the abiotic drivers of climate change, focusing on predicted changes in temperature and precipitation. The consequences of global climate change for freshwater ecosystems are reviewed, with effects grouped into those related to water quantity, water quality, habitat and aquatic biological assemblages. Several guiding principles aimed at minimising the potential impact of climate change on freshwater ecosystems are discussed. These guidelines include those focused on water quantity and the maintenance of appropriate environmental flows, integration of global climate change into water quality management, conservation planning for freshwater biodiversity, the promotion of ecosystem resilience, and extending climate change science into policy and public discourse. Proactive assessment and monitoring are seen as key as these will allow for the identification of ecological triggers and thresholds, including thresholds of vulnerability, which may be used to monitor and inform decisions, as well as to improve the ability to forecast based on this knowledge.

Introduction

Freshwater ecosystems are considered to be among the ecosystems most vulnerable to global climate change.[1] Observational records and climate projections provide abundant evidence that freshwater resources have the potential to be strongly impacted by climate change, with wide-ranging consequences for human societies and ecosystems.[1] Observed global trends in precipitation, humidity, drought and run-off indicate that southern Africa is on a negative trajectory with respect to changes associated with climate change.[2] South Africa is a water-stressed country with a mean annual precipitation (MAP) of 500 mm per annum (approximately 60% of the world average[3]), with 65% of the country, especially the arid and semi-arid interior and western regions, receiving on average <500 mm per annum. The eastern half of the country receives between 500 mm and 1000 mm per annum, while a narrow region along the southeastern coastline receives rainfall of 1000 mm to 2000 mm per annum.[3] Southern Africa has been identified as a 'critical region' of water stress, based on an indicator using the ratio of annual withdrawals-to-availability, with more than half of the water management areas in South Africa currently in deficit.[4] The relative dryness of the region, expressed as an aridity index (i.e. the ratio between mean annual potential evaporation and MAP) varies from 1 in the east to >10 in the arid west.[5] Given the most probable scenario of a growing economy and population, climate change has major implications for aquatic ecosystems and for their ability to deliver ecosystem services.

Global and regional climate change models[5-9] predict likely trends in the magnitude and amplitude of event-driven systems, primarily rainfall and air temperature. Changes include shifts in mean condition, variance and frequency of extremes of climatic variables, which result in changes in water quantity, especially in arid and semi-arid regions.[5] Historically, focus on the consequences of global climate change trends has tended to be on terrestrial ecosystems, with less attention given to aquatic ecosystems. In the last decade, focus has shifted to freshwater ecosystems and the number of studies published annually has increased dramatically.[10] This shift follows the recognition that freshwater ecosystems are vulnerable to climate change, that aquatic organisms are highly sensitive to climate change[11,12] and that climate change is expected to worsen freshwater conditions, especially in Mediterranean regions[13]. Several climate model projections warn of widespread biological invasions, extinctions and the redistribution and loss of critical ecosystem functions.[14,15]

This review explores the likely impacts of climate change on freshwater ecosystems, focusing on lotic (rivers) ecosystems, although several issues are also relevant for lentic (lakes, wetlands) systems. It builds on information gleaned during an interactive workshop of climate change and freshwater specialists[16,17] and aims to summarise key issues related to climate change and freshwater ecosystems within the context of South Africa, but with reference to international research. The review discusses the abiotic drivers of climate change and the ecological consequences of climate change to freshwater ecosystems. These consequences are separated into those affecting water quantity, water quality, physical habitat and aquatic biological assemblages. Several guiding principles aimed at minimising the potential impact of climate change on freshwater ecosystems are discussed, including those focused on water quantity and the maintenance of appropriate environmental flows, integration of global climate change into water quality management, conservation planning for freshwater biodiversity, and the promotion of ecosystem resilience. Although specific scientific literature on climate change and freshwater ecosystems in South Africa is limited, relevant studies have been consulted and links have been made to the potential ecological consequences of climate change.

Setting the stage – abiotic drivers of global climate change

General circulation models (GCMs) are a class of computer-driven models for weather forecasting; those that project climate change are commonly called global climate models. GCMs are the core tool for simulating the coupled climate system using physical representations of the atmosphere, land and ocean surface.[6] GCMs simulate the most important features of the climate (i.e. air temperature and rainfall) reliably at a large scale, although, as uncertainties are inherent in CGMs, predictions for rainfall intensity, frequency and spatial distribution have a lower confidence.[5] CGMs are commonly downscaled to enable their outputs to be made relevant to regional- or local-scale climate change scenarios.[8,18]

In South Africa, regional models have developed to a stage where pattern changes at a sub-national scale are made with confidence, while confidence for the magnitude of change is weaker.[18] According to these models, predictions are not uniform within South Africa and climate change is likely to impact most strongly on the western regions, with less of an impact as one moves eastwards. Certain areas are likely to become 'winners' in light of certain projected changes, while other areas are likely to become 'losers' as more water-related stresses are experienced.[19] 'Hotspots' of concern are the southwest of the country, the west coast and, to a lesser extent, the extreme north of South Africa.[19,20] The responses of rainfall and temperature as predicted by global climate change models are summarised in Table 1, with summer and winter rainfall regions given separately where relevant.

Ecological consequences of global climate change

Primary climate change drivers are precipitation, air temperature and evaporative demand. Ecological consequences of global climate change on freshwater ecosystems may be grouped into effects that relate to water quantity, water quality, habitat and biological assemblages (Table 2). Often stressors act in synergistic ways with effects exacerbated through the interaction of two or more effects such as the combined effect of reduced run-off and elevated water temperature. Consequences for biological assemblages are thus often the result of several climate change drivers acting in synergy. In addition, climate change may cause changes in land-use patterns, which in turn may impact on, for example, volumes of fine sediment delivered to river channels. Such feedbacks need to be considered when trying to determine the potential effects of climate change on aquatic ecosystems.

Table 1: Responses of rainfall and air temperature for the summer and winter rainfall regions of South Africa as predicted by global climate change models[5,8,9,119]

Predicted change in climatic factors	
Summer rainfall region (central, north, east)	**Winter rainfall region (southwest)**
Rainfall	
Increase in mean annual precipitation (MAP) of 40 mm to 80 mm per decade in the east, particularly the mountainous areas. Northern and eastern regions likely to become wetter in summer and autumn, especially over regions of steep topography around the escarpment and Drakensberg.	Decrease in MAP of 20 mm to 40 mm per decade. Shorter winter rainfall season, weaker winter pressure gradients, more summer rainfall from January onwards, especially inland and towards the east.
Increase in year-to-year absolute variability of MAP in the east (from 30% up to double).	Decrease in year-to-year absolute variability of annual precipitation.
Wetting trend of varying intensity and distribution, particularly in the east and transitional region. Drying trend in the middle and towards the end of the wet season (i.e. January, April) in northern areas.	Drying trend in the west, mainly in the middle of the rainy season (July) and towards the end of the rainy season (October). Mountainous regions predicted to be relatively stable, while coastal regions likely to become drier.
Greater interannual variability, intensifying in autumn.	Greater interannual variability, more irregular rainfall events.
Increase in intensity of rainfall events.	Increase in the frequency of extreme events, including drought as a result of the predicted poleward retreat of rain-bearing frontal systems.
Air temperature	
Into the IF mean annual temperatures are projected to increase by 1.5–2.5 °C along the coast and by 3.0–3.5 °C in the far interior.	
Into the MDF mean annual temperatures are projected to increase by 3.0–5.0 °C along the coast and by more than 6.0 °C in the interior.	
Interannual variability (standard deviation of the annual mean) of temperature is projected to increase by ~10% over much of South Africa, with increases in excess of 30% in the north. Variability in mountainous areas in the south and west not projected to change (i.e. January, April).	
July (winter) minimum temperatures are projected to increase by a wider range from <2 °C to >6 °C, but with essentially a south to north gradient from the coast to the interior.	
January (summer) maximum temperature is projected to increase by 2–4 °C.	January (summer) maximum temperature is projected to increase by 4–6 °C.
In KwaZulu-Natal, mean daily air temperature is likely to increase by approximately 2.5 °C.	Increase in days with hot, berg winds during December/January/February.

IF, intermediate future (2046–2065); MDF, more distant future (2081–2100).[5]

Note: model predictions are more in agreement for temperature than for rainfall.

Water quantity

Global climate change drivers directly affect the quantity of water in freshwater ecosystems by changing run-off patterns (e.g. mean values, flow variability, duration and timing), increasing the frequency and intensity of extreme events (droughts and floods), and changing groundwater recharge rates (Table 2). A substantial amount of research has been undertaken in South Africa on the likely consequences of climate change on water resources.[5,9] Hydrologically, South Africa has a high-risk climate with a low conversion of rainfall to run-off and very high year-to-year variability (e.g. a 10% change in rainfall can result in up to a 20–30% change in run-off).[19] In addition, run-off response to rainfall is non-linear, with a larger proportion of rainfall being converted to run-off when a catchment is wetter, either because a region is in a high rainfall zone or because the soil water content is high as a result of previous rainfall.[5]

Projected impacts of climate change on hydrological responses have been determined using the Agricultural Catchments Research Unit's (ACRU) agrohydrological modelling system.[5] These impacts were determined using output from one to five GCMs, empirically downscaled to climate station level and adjusted to the 5838 quinary catchments[5]; a quinary is a statistically defined region of uniform topography falling within a quaternary catchment. Quaternary catchments are the principal water management units and have been defined according to a standardised run-off measure per unit area, i.e. drier regions have larger quaternary catchments than areas with higher run-off.[21] Quinary catchments are considered to be physiographically more homogenous than quaternaries and relatively homogeneous hydrologically.[5] Climate values are used as input to the ACRU model based on daily values of rainfall, maximum and minimum temperatures, solar radiation and a reference potential evaporation available for three 20-year climate time slices: the present (1971–1990), the intermediate future (IF: 2046–2065) and the more distant future (MDF: 2081–2100).[5] Predicted hydrological responses include changes in run-off patterns, in the frequency and intensity of extreme events and in groundwater recharge rates.

Table 2: Global climate change drivers and ecological consequences of global climate change in freshwater ecosystems.

Ecological consequence	
Water quantity	Change in run-off patterns (flow variability, duration, timing) Increase in frequency and intensity of extreme events (droughts and floods) Change in groundwater recharge rate
Water quality	Increase in water temperature Increase in organic matter decomposition Decrease in the concentration of dissolved oxygen Changes in nutrient cycles (and carbon cycling) and loads Increase in algal growth and change in eutrophic condition* Increase in the incidence of cyanotoxins* Increase in sedimentation and turbidity Mobilisation of adsorbed pollutants such as metals and phosphorus from the riverbed Increase in transport of dissolved pollutants such as pesticides and pathogens Increased salinisation in semi-arid and arid areas (shallow groundwater and surface water)
Physical habitat	Change in channel geomorphology Decrease in longitudinal and lateral connectivity Change or reduction in aquatic habitat
Biological	Change in aquatic biodiversity Change in phenology and life-history patterns Change in communities Change in species distribution and range Extinction of vulnerable species Increase in the number and spread of invasive and pest species Increase in waterborne and vector-borne diseases

**Consequence is also biological.*

Run-off patterns (flow variability, duration and timing)

Much of South Africa is projected to have increases in annual streamflows by 20% to 30%, regardless of whether it is a year of median flows or a year with the 1:10 year low or high flows.[5] The exception is the southwestern Cape which will have reduced streamflows especially in the wet years. Flow reductions are projected to occur, especially in the 35 years making up the time period between the IF and the MDF (i.e. 2081–2100). Interannual variability is projected to increase (20–30%) in most of the country, with the exception of the southwestern Cape where variability is projected to decrease. Whilst no specific studies exist for South African rivers, and given what is known about the hydroclimatic factors governing run-off, it seems likely that a reduction in streamflow would result in a change in perenniality (rivers) or permanence of inundation (wetlands), with perennial rivers becoming non-perennial and permanent wetlands becoming seasonal or temporary. Further, rivers that are mainly flowing because of surface run-off would be more susceptible to changes in climate compared to rivers with high baseflow indices which would be groundwater fed.

Frequency and intensity of extreme events (droughts and floods)

Researchers have projected that most parts of South Africa are likely to experience reduced frequency, duration and intensity of droughts, with the exception being the west coast and northwest, which will exhibit marked increases in annual droughts.[5] In these regions, an increase in the frequency of extreme events, including drought, is likely as a result of the predicted poleward retreat of rain-bearing frontal systems. Floods and stormflow, i.e. water generated from a specific rainfall event, are projected to increase across South Africa, particularly in the central west where both magnitude and variability of stormflow will increase.[5] An increase in flood frequency is likely to markedly alter many river ecosystems, although the extent to which this happens will depend on deviation from background conditions (e.g. degree of canalisation and catchment hardening) and on how humans respond to the increased flooding, for example through non-structural flood management.[22]

Groundwater recharge rate

Changes in the amplitude, frequency and timing of extreme events may affect groundwater recharge. Projected changes in recharge into groundwater stores are different for median, dry and wet years.[5] Under median conditions into the IF, recharge is projected to increase in a wide band stretching from northeast to southwest (covering over 80% of South Africa), with a small area in the extreme southwest displaying decreases in discharge.[5] In dry year conditions, a northeast to southwest line divides the country, with projected decreases north of the line and increases south of the line. In wet year conditions, a general decrease is projected for the west coast and an increase is projected for 95% of the country.[5] Groundwater is critical for maintenance of 'low flows' and aquatic habitats during the drier periods.

Water quality

Higher water temperatures, increased precipitation intensity, and longer periods of low flows are projected to exacerbate many forms of water pollution, including sediments, nutrients, dissolved organic carbon, pathogens, pesticides, salt and thermal pollution.[1] The quality of water in many South African rivers and wetlands is already widely compromised and climatic drivers therefore act as additional stresses on these ecosystems. Water quality changes, including water temperature, affect the solubility of oxygen and other gases, chemical reaction rates and toxicity, and microbial activity.[23,24] A reduction in the concentration of dissolved oxygen, particularly under the combined effects of high temperature and low flows, is particularly deleterious to aquatic organisms.[23] The effects of water quality variables on aquatic ecosystems have been widely documented[23], with specific studies focusing on particular variables, including water temperature[24]. Recent studies on the link between air and water temperature[25] and the effect of elevated water temperature on aquatic organisms have included experimental laboratory work[26,27] which together with field-based studies[25,28], has allowed for the development of tools for assessing water temperature in river ecosystems[25-30] and scenario prediction for elevated temperatures[31]. Estimating the likely increase in water temperature from predicted changes in air temperature is, however, complex and dependent on insulators and buffers such as solar radiation, groundwater input and shading. One of the key issues is how lapse rates (change in water temperature degrees with every 100 m altitude) will change.

Other water quality variables likely to increase in response to more intense rainfall events (Table 2) include turbidity and nutrients, with sediment washed in from the catchment or, in the case of nutrients, mobilised from the riverbed (e.g. phosphorus). A review paper on the potential impacts of climate change on surface water quality through the lens of UK surface water provides an excellent overview of key issues discussed in this section.[32] Average phosphorus concentration (as orthophosphate) and chemical oxygen demand values indicate that South Africa's freshwater resources are already excessively enriched and are considered to be moderately to highly eutrophic.[32] Further changes in nutrient loads and nutrient cycles (and carbon cycling) may result in increased algal growth, changes in eutrophic condition, as well as increased incidences of cyanotoxins, which affect human health negatively. Most eutrophic rivers and reservoirs in South Africa have as the dominant phytoplankton genera the cyanobacteria *Microcystis* sp. and *Anabaena* sp.[33,34] Other adsorbed pollutants such as metals may also be mobilised, together with increased transport of dissolved pollutants such as pesticides and pathogens. In semi-arid and arid areas salinity may increase as a result of increased evaporation from shallow ground and surface water. Several river systems already have high levels of salinity, for example, the Berg River in the Western Cape.[35] In comparison, salinity levels in the headwaters of the Murray–Darling Basin in Australia are expected to increase by 13–19% by 2050,[36] a situation that may be mimicked in some southern African regions, indicating that under predicted climate change for this region, salinisation would be exacerbated. The synergistic and antagonistic interactions of several water quality variables make it especially difficult to predict the likely consequences of climate change on receiving water bodies, suffice it to say that these consequences are likely to be significant given the levels of stress already imposed on these systems.

Physical habitat

Changes in the amount, seasonal distribution and intensity of rainfall may affect channel geomorphology, longitudinal and lateral connectivity, and aquatic habitat, through changes in run-off. Likely consequences of changes in flow on the geomorphology of river systems depend on the direction of change with increased discharge (e.g. in the eastern region) potentially resulting in channel enlargement and incision, greater channel instability and sinuosity, and increased bank erosion, while decreased discharge (e.g. in the western region) may result in channel shrinkage, greater channel stability, vegetation encroachment, and sedimentation in side channels.[37] Sensitive systems such as fine-grained alluvial streams are likely to be more affected than bedrock channels and armoured stream beds.[37] While local geomorphological studies have not focused specifically on climate change, observations elsewhere are likely to be applicable, with many effects similar to those already observed following the construction of impoundments and abstraction of water.[38]

Loss of longitudinal and lateral connectivity can lead to isolation of populations, failed recruitment and local extinction; the maintenance of natural connectivity patterns is thus essential to the viability of populations of many riverine species and for maintaining instream integrity.[39] Connectivity is typically reduced through flow regulation by dams and is often compounded by other structural modifications such as channelisation.[40] With respect to fish, researchers suggest that functional habitat units (FHU, i.e. 'natural partitions within the river system that contain all the necessary habitat elements to support all life-history stages'[41]) in South African rivers need to be identified, mapped and their connectivities to other FHUs identified and efforts made to protect them in conservation and water allocation strategies.[41]

Flow is a major determinant of physical habitat in streams, which in turn is a major determinant of biotic composition.[39] In South Africa, research has focused on the flow-related dynamics of hydraulic biotopes.[42] A study of four perennial upper rivers of the southwestern Cape, South Africa, revealed that natural and, particularly, manipulated low flows, resulted in consistent, marked declines in physical habitat availability for aquatic invertebrates, with increased habitat fragmentation, hydraulic biotope isolation and dominance by low-velocity shallow biotopes.[43] Invertebrate responses to low-flow disturbances, in contrast, were often river specific, subtle or inconsistent, and required multi-scale lines of evidence for their elucidation.[43] Regions predicted to have decreased flow will therefore likely exhibit increased fragmentation of existing instream and riparian habitats, and resultant loss of habitat and connectivity.

Biological

Thermal and hydrological regimes are master variables driving river ecosystems.[44] Temperature is a primary climate change driver, while flow has been shown to change substantially in response to changes in rainfall patterns.[5] It is therefore likely that climate change will affect aquatic assemblages with biological consequences of climate change acting at several levels, including that of the individual and community. Susceptibility of aquatic organisms to climate change is likely to vary between species and will in part depend on their biological traits. Those species with specialised habitat and/or microhabitat requirements, narrow environmental tolerances or thresholds that are likely to be exceeded at any stage in the life cycle, dependence on specific environmental triggers, dependence on interspecific interactions, and poor ability to disperse to or colonise a new or more suitable area, are likely to be more susceptible.[45] Potential biological consequences of climate change (Table 2) include changes in aquatic biodiversity, changes in individual life-history patterns, changes in communities, changes in species distribution and range, extinction of vulnerable species, increase in the number and spread of invasive and pest species, and an increase in waterborne and vector-borne diseases.

Aquatic biodiversity

Biodiversity in freshwater ecosystems shows substantial impacts from land use, biotic exchange and climate.[46] In the USA, for example, freshwater biodiversity is declining at far greater rates than most affected terrestrial ecosystems.[47] Threats to global freshwater biodiversity can be grouped under five interacting categories: (1) overexploitation, (2) water pollution, (3) flow modification, (4) destruction or degradation of habitat and (5) invasion by exotic species.[48] The inland waters of southern Africa support a high diversity of aquatic species with high levels of endemism, many of which provide direct (e.g. fisheries) and indirect (e.g. water purification) benefits to people.[49] The level of threat to species in South Africa is higher than in other African countries (for which it is about 7%) and 57% of river and 75% of wetland ecosystems are highly threatened,[49-51] while tributaries are in a better condition than main rivers. Researchers have highlighted the substantial threats to riverine biodiversity in South Africa and challenges faced in conserving species and habitats.[52] It has been noted that for every 10% of altered catchment land use, a correlative 6% loss in freshwater biodiversity occurs.[53] The amount of natural vegetation at catchment scale has been found to be a good predictor of river habitat integrity[54], which, in agricultural catchments declines as agriculture exceeds 30–50%[55]. Further, small dams within a catchment impact on water quality and quantity and hence biodiversity.

These existing anthropogenic threats are likely to be further exacerbated by predicted global climate change, leading to greater loss of aquatic biodiversity. For example, dams create discontinuities and downstream water temperature changes that in turn may alter chemical processes that drive energy flows in rivers. This in turn could differentially affect competitive abilities of different species and functional feeding groups, changing the composition of aquatic communities. Potential also exists for reduction or changes to genetic diversity. We see two possible reasons driving this reduction in genetic diversity: reduced flows of individuals within metapopulations because of decreased catchment connectivity, and increased homogenisation of communities because of environmental conditions favouring generalist or opportunist species.[56]

Phenology and life-history patterns

Individual species have life-history parameters that allow them to successfully inhabit aquatic ecosystems; changes in abiotic parameters such as temperature and flow may affect the growth, reproduction and survival of instream species. Data from South Africa, where detailed life-history data has been collected, are limited.[57,58] However, a recent study has provided the first detailed information on the responses of aquatic insect life histories to water temperature and flow, and data on egg development, nymphal growth and oviposition.[59] The study showed that water temperature regimes in rivers of the Western Cape have a measurable impact on aquatic macroinvertebrate life histories. Through a combination of field surveys and laboratory experiments, it was shown that life histories of three target macroinvertebrate species showed differing degrees of flexibility in life-history responses – from subtle changes in the timing of emergence and egg hatching to more extreme differences involving the production of additional generations within a year given differing environmental conditions.[25]

Several studies have examined the reproductive biology of fishes in South Africa and have shown that temperature is an important factor triggering spawning, with temperatures of 18–19 °C triggering spawning of several of South Africa's indigenous fish species.[24] Gonadal development and spawning may also be triggered by water level or flooding. In the Western Cape, the endangered Clanwilliam sawfin, *Barbus serra*, depends on certain key components of the annual flow and temperature regime.[41] Sawfin spawned over a period of about 100 days between November and January, and peak recruitment events were associated with a temperature of ~19 °C and continuously rising temperatures over 7 days or more.[41]

Communities

Aquatic species have had to adapt to variable flows and cope with daily and seasonal ranges of water temperatures.[60] Current community patterns are likely to be in dynamic equilibrium with such abiotic regimes. Consequently, any changes to these regimes in response to climate change would differentially affect different species, which would ultimately be reflected in species patterns. A combination of temperature and the onset and cessation of floods was shown to influence the seasonal pattern of change in an invertebrate community assemblage in a Western Cape river.[58] Under future climate change scenarios, one might therefore anticipate shifts in communities in response to elevated temperatures and changes in frequency, duration and intensity of rainfall events. Certain species are likely to be more or less resilient to changes associated with climate change, with certain species increasing in abundance (winners), while others decrease in abundance (losers). This difference will result in a shift in community structure and possibly lead to a change in trophic status. Thermal tolerances of individual taxa,[26,27] and their ability to withstand changes in water temperatures, may ultimately translate into shifts evident at community levels. Montane stream assemblages are considered to be most vulnerable to climate change because their distribution is most responsive to climatic factors, and elevated sites are isolated from one another, which reduces the scope for altitudinal migration.[61] Changes in flow, particularly flood events, may also lead to changes in riparian vegetation such as reduction in cover and a loss or shift of species, with non-riparian species increasing in abundance as flow decreases. These changes in turn impact on water temperature regimes, particularly in smaller streams in upper reaches.

Species distribution and range

Various studies have reported shifts in distributional ranges of aquatic species[62] with effects being typically species specific, where cold-water organisms are generally negatively affected and warm-water organisms positively affected. Much of this research appears to relate to migratory fish species of economic importance (notably trout[63]) with very few studies on aquatic macroinvertebrates, and virtually no studies applicable to southern Africa. One such example is for range shifts in stoneflies (Plecoptera), in which logistic multiple regression models were developed to predict stonefly responses to temperature change.[64] These models showed mixed success for different species, although predictions broadly agreed with data that the species considered already showed uphill altitudinal shifts of 100–140 m over a 30-year period between 1977 and 2006.[64] Knowledge of species thermophily and rheophily have been used to predict distributional changes associated with climate change.[65] It was hypothesised that thermophobic (i.e. those that favour cool water) taxa and rheophilous (i.e. those that prefer fast water) taxa would have range contractions, being least able to cope with rising temperatures and declines in stream flow, and, conversely, thermophilic taxa (i.e. those that favour warm water) and rheophobous taxa (i.e. those that prefer still water) would expand their ranges, being best placed to take advantage of warmer and more sluggish streams.[65] The study concluded that:

> *Trait analysis has potential for predicting which species will expand their ranges and which will contract, but it needs to be coupled with assessment of how the landscape provides each species with opportunities to track or avoid climate change. Improved quantification of climatically relevant traits and integration of trait analysis with species distribution modelling are likely to be beneficial.*[65]

In a southern African study, researchers demonstrated that a cold-water Western Cape stenotherm could experience a 30% habitat loss in response to a 2 °C increase in mean daily water temperatures.[31] However, predicting changes in species distribution and range is not an exact science, because different taxonomic groups may show different responses to climate change. This is partly a function of dispersal ability and strategy[62], but also because different species show different tolerances to thermal stress, and have different behavioural mechanisms to avoid thermal stress[14].

Extinction of vulnerable species

Freshwater aquatic ecosystems appear to have the highest proportion of species threatened with extinction by global climate change.[66] Recent studies have shown the disproportionate risk of extinctions in mountain ecosystems and, in particular, among endemic species, with rare and stenothermic species likely to become at least locally extinct.[67,68] For example, of the 27 currently recognised indigenous freshwater fish species in the Cape Floristic Region of South Africa, 24 (89%) are endemic to the region and 19 (70%, all endemics) are listed as threatened in the *IUCN Red List of Threatened Species*.[69] Temperature rises will be especially deleterious in high-altitude, fast-flowing streams in which cold stenotherms could lose their thermal refuges.[70] Similar risks exist for the mountainous regions of South Africa that have many of the Gondwanan species.[31] Identifying highly vulnerable species and understanding why they are vulnerable[71] is seen as critical to developing climate change adaptation strategies and reducing biodiversity loss in the coming decades.[72]

Invasive and pest species

Climate change and invasive aquatic species are considered to be two of the most pervasive aspects of global environmental change.[73] As climate change proceeds, aquatic systems may become more vulnerable to invasion[74], and the key climate change drivers – temperature and flow – are likely to determine the invasion and success of exotic and introduced species in rivers.[39,73] Given potential disaggregation in aquatic ecological communities resulting from likely concomitant changes in flow and water temperature regimes, community equilibria are likely to shift. This shift makes communities more vulnerable to invasion by alien species, particularly in species-poor systems where niche space is more open, and as a consequence of increased interbasin transfer schemes. Typically, alien species may out-compete indigenous species; similarly warm-water species may have an advantage over cold-water species as temperature increases. Changes in species patterns could also lead to development of indigenous pest species, as has happened with pest blackfly (Diptera: Simuliidae), which is a threat to livestock, in the Great Fish River (Eastern Cape Province).[75] Invasive aquatic species, including both plants and fish, are already a major threat in many parts of South Africa and one of the greatest threats to biodiversity in the Western Cape is the spread of invasive fish species.[76] Virtually no information is available on the thermal tolerance of indigenous fish species in South Africa, although tolerances are known for several alien species,[77] especially those of aquacultural importance.

Waterborne and vector-borne diseases

Rising temperatures, heavy rainfall and increased flooding are likely to increase the burden of infectious waterborne (e.g. microbial pathogens) and vector-borne (e.g. malaria, bilharzia) diseases, especially for vulnerable populations.[1,78] Heavy rainfall leading to flooding has been associated with increased risk of infection, particularly in developing countries, while temperature affects both the distribution of vectors such as mosquitoes and snails and the effectiveness of pathogen transmission through the vector.[78,79] The incubation period for malaria parasites within the mosquito is strongly temperature sensitive, such that temperature is a major determinant of malaria risk.[79] Malaria transmission in Africa is projected to increase by 16–28% in person-months of exposure by 2100 as a result of projected climate scenarios.[80] This projected increase is related to an increase in distribution (mainly altitudinal) and season length, and, in the Limpopo Province of South Africa, a substantial latitudinal extension.

Discussion

Global climate change is recognised as an additional, amplifying driver of system variability and cannot therefore be viewed in isolation from other stressors.[17,81] Many non-climatic drivers affect freshwater resources at a global scale and the quantity and quality of resources are influenced by, for example, land-use change, construction and management of reservoirs, pollutant emissions, and water and wastewater treatment.[2] South Africa is fortunate in having established widely recognised approaches for determining the 'ecological reserve' – a South African term used to describe what are globally referred to as environmental water requirements. However, there is widespread agreement among scientists that South Africa's aquatic ecosystems have significantly deteriorated since the revision of the *National Water Act (Act 36 Of 1998)*.[82] A key reason given for this deterioration is the widespread failure to operationalise, monitor and enforce ecological reserves – a legislated framework for securing water quality and quantity for the environment.[82] Understanding the likely consequences of climate change for freshwater ecosystems is therefore of critical importance to the future well-being of the resource and society. Several guiding principles, or proactive response options, aimed at minimising the potential impact of climate change on freshwater ecosystems, were formulated during a workshop.[17] These guiding principles include those focused on water quantity and the maintenance of appropriate environmental flows, integration of global climate change into water quality management, conservation planning for freshwater biodiversity, the promotion of ecosystem resilience, and extending climate change science into policy and public discourse. The adoption of a proactive, 'no-regrets' policy with respect to climate change has been widely endorsed in South Africa and elsewhere[17]; such a policy calls for proactive management which includes actions such as restoration, land purchases, and measures that can be taken now to maintain or increase the resilience of rivers.[81] The alternative – reactive management – which involves responding to problems as they arise by repairing damage or mitigating ongoing impacts, has been shown to be inadequate and short-sighted and results in considerable ecological, social and economic consequences and costs in the longer term.[81]

Maintaining appropriate environmental flows

Flow is one of the master variables controlling river ecosystems[44] and it is recognised that a naturally variable flow regime, rather than a static minimum low flow, is required to sustain freshwater ecosystems.[39,83,84] South Africa has been at the forefront of environmental flow research and the DRIFT (Downstream Response to Imposed Flow Transformation) method has been widely applied.[85-87] DRIFT is a scenario-based approach that predicts the bio-physical and socio-economic impacts of proposed water-resource developments on rivers.[82] The maintenance of appropriate environmental flows is considered to be a critical aspect in promoting ecological integrity and reducing ecological consequences of global climate change.[88] In a global analysis of the potential effect of climate change on river basins, it was shown that rivers impacted by dams or with extensive development will require more management interventions to protect ecosystems and people than will catchments with free-flowing rivers.[89] Catchments that are modified have limited ability to absorb disturbances, such as changes in discharge. In contrast, healthy, free-flowing rivers respond to changes in land use and climate through dynamic movements and flow adjustments that buffer against impacts, and are thus more resilient.[89] The rate and magnitude of change relative to historical and recent thermal and flow regimes for each catchment will also determine the impacts of climate change on river ecosystems[81], and changes outside the natural range of flow or temperature variability may have drastic consequences for ecosystem structure and function depending on the rate of change in temperature or discharge relative to the adaptive capacity of species[90].

South Africa has 62 large free-flowing rivers, representing just 4% of South Africa's river length,[50] with the remaining rivers dammed. The approximately 600 large dams and more than 500 000 small dams[91] that regulate South African rivers significantly affect their discharge and hydrological regimes.[75,92-97] Dams, interbasin water transfers[98] and abstraction of water threaten the maintenance of appropriate environmental flows in our rivers. Worldwide management actions have been recommended for dammed river systems.[89] Consideration should be given to undertaking a review of the justification and viability of existing water infrastructure, including opportunities to re-engineer such infrastructure to incorporate better environmental and social performance measures. Use should be made of infrastructure management as an opportunity to retrofit dams to build ecological resilience, for example, retrofit outlet valves to allow environmental flows to be released and install fishways[99] to facilitate passage of fish – although in some instances this may be inadvisable (e.g. the Western Cape) as it would facilitate invasion of alien fish. In addition, an evaluation should be undertaken of the appropriateness of interbasin water transfers and vulnerability of donor and recipient riverine biota to climate change.

Integrating global climate change into water quality management

In recognising that climate change would be a further contributing factor to existing water quality problems, management options for reducing these effects need to be examined. Water quality issues in South Africa are a major cause for concern, with contributing sources including mines, sewage effluent discharges, industrial effluents, non-point source pollution from agricultural sources, acid atmospheric depositions, over-abstraction of groundwater, and excessive soil losses and sedimentation.[100,101] Water quality impacts could be ameliorated or decreased through application and implementation of the ecological reserve and Water Resources Classification processes, and identification of a dynamic baseline that incorporates climate change. The actual mechanism for achieving these objectives is through the determination of resource quality objectives (RQOs) per Water Management Area, which relate to the quality of water resources in terms of their quantity, quality, habitat and biota, as provided for in South Africa's *Water Act of 1998*. RQOs may be established for priority 'resource units' (i.e. river reaches identified as having value to society; for example, presence of a unique community of aquatic macroinvertebrates). Each priority resource unit will have associated indicators linked to numerical limits (which should also incorporate thresholds of potential concern that inform river managers of a problematic trend before the system state changes). These are defined based on a narrative vision (for example, that the water quality be suitable to maintain a Gondwanan relict macroinvertebrate community of conservation importance), and provide the yardsticks from which to develop a monitoring programme. RQOs are determined based on a seven-step process,[102] and are gazetted together with their associated classification and ecological reserve. In theory, RQOs should be updated every five years. While there is no explicit provision for integrating global climate change into water quality management, the five-yearly revisions make this a possibility. Thus, in practice, an RQO could be set for water temperatures, with an indicator chosen that reflects changes in water temperatures, such as a thermally sensitive aquatic macroinvertebrate species. This species could be monitored over the succeeding five years, until the time to review the RQOs applies. Should it be found that the numerical limits were repeatedly exceeded, i.e. that a downward trend was observed, then questions should be raised in connection with how to enhance system resilience. Potentially, increasing river connectivity increases the adaptive capacity of the river in question by better enabling river biota to undergo range shifts to avoid habitat stress.

Conservation planning for freshwater biodiversity

The principles of conservation planning[103] can be applied to identify biodiversity threats. In South Africa, recent approaches to systematic conservation planning for freshwater biodiversity have shifted from 'representation' to 'representation and persistence'.[104,105] These approaches present four key principles to consider when planning for the persistence of freshwater biodiversity: (1) selecting ecosystems of high ecological integrity, (2) incorporating connectivity, (3) incorporating areas important to population persistence and (4) identifying additional natural processes that can be mapped.[104,105] A critical assumption in conservation planning is that a conservation target (such as percentage species area) is the minimum area needed to ensure representivity and persistence. In South African freshwater conservation planning, 20% targets are typically used for both species occurrences and freshwater types.[106] However, conserving river systems continues to present challenges because a river reach cannot be assumed to be protected by virtue of its being within a protected area, as it is affected by cumulative upstream influences as well as downstream connections.

Recently, biodiversity conservation has also shifted from a species to a landscape approach,[107] which has resulted in a number of implications for conservation planning. The first is a move away from individual species targets (although these still play a role in conservation plans) and a move towards river types as surrogates for freshwater biotic communities. The second is the need to recognise and incorporate differential rates of turnover of species and communities with downstream distance, with implications for setting higher targets in upstream areas where turnover rates are higher.[108]

Planning for climate change is, however, in its infancy. Global climate change causes changes in the spatial configuration of habitats, and static conservation planning approaches are not adequate to deal with such changing environmental gradients.[70] The question arises as to whether protected areas are an effective conservation strategy in the face of climate change. Ideally, reserves should be planned around biomes which are expected to have reduced representation. In the absence of data, this planning requires either individual species models that may be simplistic in complex community landscapes, or species models that inevitably include uncertainties because of difficulties in accounting for stochastic events, synergies and interactions between multiple species. For reserve design to be resilient, corridors and connectivity are required.[109] Connectivity must be planned in spatial and temporal dimensions, to counter disrupted hydrological and thermal time-series signatures (changes in frequency, duration, magnitude and timing of flow or thermal events) resulting from, for example, dam construction, water abstractions and land-use changes.[110,111] Connectivity in stream channels and riparian corridors becomes critical as species distributions change relative to protected area boundaries. Restoring freshwater ecosystem connectivity (e.g. with fish passages) could be a key mechanism to enable freshwater biota to move to new areas. However, connectivity is not always a panacea; it sometimes is a double-edged sword because of the spread of aliens. Careful assessment of risks and advantages in establishing corridors is necessary.[109]

Connectivity includes maintaining flow signatures (the magnitude, timing, frequency and duration of flow events), which are linked to life-history cues, and may be lost because of increased construction of dams and disrupted hydrological and thermal regimes.[110,111] It has been recommended that the integration of environmental flow assessment and systematic conservation planning would be mutually beneficial and provide an ideal platform for integrated water resource management.[112] Changes to the flow regime under various climate change scenarios[5] may be used in environmental flow assessment to examine the likely impacts of development and conservation within a changing climate. There will be increased difficulty in meeting sufficient flows to maintain the character of all systems, leading to difficult dilemmas in choosing one system over another.[109]

There is also growing consensus amongst freshwater scientists that water temperatures should be incorporated into environmental flow guidelines.[113] Increases in water temperatures have more severe consequences for the higher-altitude, stenothermic cold-water species, whose distributions are expected to shrink as water temperatures increase. These consequences should be of particular concern to conservation planners as the upper zones of river systems are typically where species turnover, ecotones and taxonomic diversity are highest. It therefore becomes imperative to maintain river connectivity in these upper reaches to facilitate system adaptation and species range shifts.

Promoting ecosystem resilience

Resilience is the capacity of reduced or impacted populations or communities to recover after a disturbance.[114] The resilience approach is founded on the understanding that the natural state of a system is one of change rather than one of equilibrium, even though the magnitude and type of change is not always predictable.[115] It is the ecological concept that reflects the capacity of natural systems to recover from environmental change and thus persist into the future. Systems that are more resilient are better able to adapt to changes in climate[116] and ecosystem resilience is seen as key to reducing the consequences of global climate change on aquatic ecosystems. Stressed ecosystems have lower resilience. To enhance the resilience of freshwater ecosystems and minimise impacts, specific, proactive restoration, rehabilitation, and management actions are recommended.[89] Ways to promote ecosystem resilience include, for example, maintaining environmental flows,[117] restoring habitat and connectivity, and recognising the link between catchment condition and freshwater ecosystem health. As mentioned, free-flowing rivers in largely undeveloped catchments are expected to be resilient in the face of climate change, while the need for restoration/rehabilitation and proactive management may be quite high in dammed and developed river systems.[89] The resilience of a catchment may also, for example, be influenced by the intactness of its tributaries, which often act as refuges for aquatic biota.[118]

Enhancing resilience of freshwater ecosystems in South Africa requires the application and implementation of the Resource Directed Measures, which consist of three main elements: classification and the reserve and resource quality objectives.[17,82] It is therefore very important that the water resources status and the ecological flow requirements of these resources be determined for an effective national scale allocation process (National Water Resources Strategy) and resource protection. Rehabilitation of riparian zones and landscapes are considered important.[17] Global climate change increases the urgency to institute the freshwater conservation measures that are already desirable, to increase resilience.

Extending climate change science into policy and public discourse

Ultimately, the vulnerability of freshwater systems to climate change depends on national water management and the desire and ability to instigate management options that have the potential to lessen its consequences.[1] Governance and integration of plans and policies in a holistic manner that incorporates all levels of governance, especially to avoid conflicts between climate, energy and water policies, is recommended.[17] Engaging top level leadership is important, with potential economic savings linked to adaptation more likely to receive reaction and support from government, whose decisions are often not driven by issues such as biodiversity and conservation.[17] Engagement at a local scale, which in reality is the scale at which climate change is going to be felt, is as important as institutional support.

Conclusions

Proactive assessment and monitoring are key as these would allow for the identification of ecological triggers and thresholds, including thresholds of vulnerability, which may be used to monitor and inform decisions, as well to improve the ability to forecast based on this knowledge. Identification of ecological reference sites for long-term monitoring and routine monitoring of these and other impacted sites within a framework of established biomonitoring programmes are critical. This monitoring would facilitate detection of change, both in response to non-climate as well as climate change induced effects, although the ability of the various monitoring tools to facilitate this may still need to be validated. One of the key challenges facing freshwater ecologists is to develop a suite of tools for detecting the impacts of climate change in complex natural systems that can be applied across multiple spatio-temporal scales and levels of organisation.[12] Integration of long-term, empirical survey data with models and manipulative experiments will facilitate the development of mechanistic, and hence predictive, understanding of responses to future change.[12]

Acknowledgements

We gratefully acknowledge the Water Research Commission (WRC) of South Africa for funding our research and in particular Mr Bonani Madikizela and Mrs Una Wium for their support and enthusiasm over the years. We also thank the WRC and World Wide Fund for Wildlife (WWF) for funding the Climate Change and Freshwater Resources workshop that formed the basis for this review, and the workshop participants for their valuable contributions. The comments of two anonymous reviewers are also much appreciated.

Authors' contributions

H.D. is the project leader and researcher and wrote the manuscript; and N.R-M. is a researcher on the project and contributed to the manuscript.

References

1. Bates BC, Kundzewicz ZW, Wu S, Palutikof JP. Climate change and water. Technical paper of the Intergovernmental Panel on Climate Change. Geneva: IPCC Secretariat; 2008.

2. Kundzewicz ZW, Mata LJ, Arnell NW, Döll P, Kabat P, Jiménez B, et al. Freshwater resources and their management. Climate change impacts, adaptation and vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2007.

3. Zucchini W, Nenadi O. A web-based rainfall atlas for southern Africa. Environmetrics. 2006;17:269–283. http://dx.doi.org/10.1002/env.748

4. Alcamo J, Henrichs T. Critical regions: A model-based estimation of world water resources sensitive to global changes. Aquat Sci. 2002;64:352–362. http://dx.doi.org/10.1007/PL00012591

5. Schulze RE. A perspective on climate change and the South African water sector. Water Research Commission report 1843/2/11. Pretoria: Water Research Commission; 2011.

6. Hewitson B, Reason C, Tennant W, Tadross M, Jack C, MacKellar N, et al. Dynamical modelling of present and future climate systems. Water Research Commission report 1154/1/04. Pretoria: Water Research Commission; 2004. http://dx.doi.org/10.1002/joc.1314

7. Schulze RE. Climate change and water resources in Southern Africa. Water Research Commission report 1430/1/05. Pretoria: Water Research Commission; 2005.

8. Hewitson BC, Crane RG. Consensus between GCM climate change projections with empirical downscaling: Precipitation downscaling over South Africa. Int J Climatol. 2006;26:1315–1337.

9. Lumsden TG, Schulze RE, Hewitson BC. Evaluation of potential changes in hydrologically relevant statistics of rainfall in southern Africa under conditions of climate change. Water SA. 2009;35:649–656. http://dx.doi.org/10.4314/wsa.v35i5.49190

10. Filipe A, Lawrence JE, Bonada N. Vulnerability of stream biota to climate change in mediterranean climate regions: A synthesis of ecological responses and conservation challenges. Hydrobiologia. 2013;719:331–352.

11. Durance I, Ormerod SJ. Climate change effects on upland stream macroinvertebrates over a 25-year period. Glob Change Biol. 2007;13:942–957. http://dx.doi.org/10.1111/j.1365-2486.2007.01340.x

12. Woodward G, Perkins DM, Brown LE. Climate change and freshwater ecosystems: impacts across multiple levels of organization. Philos Trans R Soc Lond B Biol Sci. 2010;365:2093–2106. http://dx.doi.org/10.1098/rstb.2010.0055

13. Dallas HF. Ecological status assessment in Mediterranean rivers: Complexities and challenges in developing tools for assessing ecological status and defining reference conditions. Hydrobiologia. 2013;719:483–508. http://dx.doi.org/10.1007/s10750-012-1305-8

14. Sunday JM, Bates AE, Dulvy NK. Thermal tolerance and global redistribution of animals. Nat Clim Change. 2012;2:686–690.

15. Pereira HM, Leadley PW, Proença V, Alkemade R, Scharlemann JPW, Fernandez-Manjarrés JF, et al. Scenarios for global biodiversity in the 21st century. Science. 2010;330:1496–1501. http://dx.doi.org/10.1126/science.1196624

16. Dallas HF, Rivers-Moore NA. Adaptation to the consequences of climate change for freshwater resources. Starter document produced for the Water Research Commission and the World Wide Fund for Wildlife. Cape Town: The Freshwater Consulting Group and Ezemvelo KZN Wildlife; 2009.

17. Dallas HF, Rivers-Moore NA. Future uncertain – Climate change and freshwater resources in South Africa. Technical report produced for the Water Research Commission and the World Wide Fund for Wildlife. Cape Town: The Freshwater Consulting Group and Ezemvelo KZN Wildlife; 2009.

18. Hewitson B, Tadross M, Jack C. Scenarios from the University of Cape Town. In: Schulze RE, editor. Climate change and water resources in southern Africa: Studies on scenarios, impacts, vulnerabilities and adaptation. Water Research Commission report 1430/1/05. Pretoria: Water Research Commission; 2005.

19. Stuart-Hill S, Schulze R, Colvin J. Handbook on adaptive management strategies and options for the water sector in South Africa under climate change. Water Research Commission report 1843/3/11. Pretoria: Water Research Commission; 2011.

20. Mukheibir P, Ziervogel G. Framework for adaptation to climate change in the city of Cape Town (FAC4T). Report by the Energy Research Centre and the Climate Systems Analysis Group, University of Cape Town. Submitted to City of Cape Town: Environment Resource Management; 2006.

21. Midgley DC, Pitman WV, Middleton BJ. Surface water resources of South Africa 1990, Vol I-VI. Water Research Commission reports 298/1.1/94 to 298/6.1/94. Pretoria: Water Research Commission; 1994.

22. Poff NL. Ecological response to and management of increased flooding caused by climate change. Philos T Roy Soc Lond B Biol Sci. 2002;360:1497–1510. http://dx.doi.org/10.1098/rsta.2002.1012

23. Dallas HF, Day JA. The effect of water quality variables on aquatic ecosystems: A review. Water Research Commission technical report 224/04. Pretoria: Water Research Commission; 2004.

24. Dallas HF. Water temperature and riverine ecosystems: An overview of knowledge and approaches for assessing biotic response, with special reference to South Africa. Water SA. 2008;34:393–404.

25. Dallas HF, Rivers-Moore NA, Ross-Gillespie V, Eady B, Mantel S. Water temperatures and the ecological reserve. Water Research Commission report 1799/1/12. Pretoria: Water Research Commission; 2012.

26. Dallas HF, Ketley ZA. Upper thermal limits of aquatic macroinvertebrates: Comparing critical thermal maxima with 96-LT50 values. J Therm Biol. 2011;36:322–327. http://dx.doi.org/10.1016/j.jtherbio.2011.06.001

27. Dallas HF, Rivers-Moore NA. Critical thermal maxima of aquatic macroinvertebrates: Towards identifying bioindicators of thermal alteration. Hydrobiologia. 2012;679:61–76. http://dx.doi.org/10.1007/s10750-011-0856-4

28. Dallas HF, Rivers-Moore NA. Micro-scale heterogeneity in water temperature. Water SA. 2011;37:505–512.

29. Rivers-Moore NA, Mantel S, Dallas HF. Prediction of water temperature metrics using spatial modelling in the Eastern and Western Cape, South Africa. Water SA. 2012;38:167–176. http://dx.doi.org/10.4314/wsa.v38i2.2

30. Rivers-Moore NA, Dallas HF, Morris C. Towards setting environmental water temperature guidelines: A South African example. J Env Manag. 2013;128:380–392. http://dx.doi.org/10.1016/j.jenvman.2013.04.059

31. Rivers-Moore NA, Dallas HF, Ross-Gillespie V. Life history does matter in assessing potential ecological impacts of thermal changes on aquatic macroinvertebrates. Riv Res Appl. 2013;29:1100–1109. http://dx.doi.org/10.1002/rra.2600

32. Oberholster PJ, Ashton PJ. State of the nation report: An overview of the current status of water quality and eutrophication in South African rivers and reservoirs [document on the Internet]. c2008 [cited 2013 Nov 21]. Available from: http://npconline.co.za/MediaLib/Downloads/Home/Tabs/Diagnostic/MaterialConditions2/An%20overview%20of%20the%20current%20status%20of%20water%20quality%20in%20South%20Africa.pdf

33. Harding WR, Paxton BR. Cyanobacteria in South Africa: A review. Water Research Commission report TT 153/01.165. Pretoria: Water Research Commission; 2001

34. Van Ginkel CE. A national survey of the incidence of cyanobacterial blooms and toxin production in major impoundments. Internal report no. N/0000/00/DEQ/0503. Pretoria: Resource Quality Services, Department of Water Affairs and Forestry; 2004.

35. Van Rensburg LD, De Clercq WP, Barnard JH, Du Preez CC. Case studies from Water Research Commission projects along the Lower Vaal, Riet, Berg and Breede Rivers. Water SA. 2011;37:739–749. http://dx.doi.org/10.4314/wsa.v37i5.11

36. Pittock B. Climate change: An Australian guide to the science and potential impacts. Canberra: Australian Greenhouse Office; 2003.

37. Goudie AS. Global warming and fluvial geomorphology. Geomorphology. 2006;79:384–394. http://dx.doi.org/10.1016/j.geomorph.2006.06.023

38. Arturo Elosegi, Sergi Sabater. Effects of hydromorphological impacts on river ecosystem functioning: A review and suggestions for assessing ecological impacts. Hydrobiologia. 2012;712(1):129–143. http://dx.doi.org/10.1007/s10750-012-1226-6

39. Bunn SE, Arthington AH. Basic principles and ecological consequences of altered flow regimes for aquatic biodiversity. Env Manag. 2002;30:492–507. http://dx.doi.org/10.1007/s00267-002-2737-0

40. Ward JV, Stanford JA. Ecological connectivity in alluvial river ecosystems and its disruption by flow regulation. Regul River. 1995;11:105–119. http://dx.doi.org/10.1002/rrr.3450110109

41. Paxton BR. The influence of hydraulics, hydrology and temperature on the distribution, habitat use and recruitment of threatened cyprinids in a Western Cape river, South Africa [PhD thesis]. Cape Town: University of Cape Town; 2008.

42. Wadeson RA, Rowntree KM. Application of the hydraulic biotope concept to the classification of instream habitats. Aquat Ecosys Health. 1998;1:143–157. http://dx.doi.org/10.1080/14634989808656911

43. Tharme RE. Ecologically relevant low flows for riverine benthic macro-invertebrates: Characterization and application [PhD thesis]. Cape Town: University of Cape Town; 2010.

44. Poff NL, Zimmerman JKH. Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows. Freshwater Biol. 2010;55:194–205. http://dx.doi.org/10.1111/j.1365-2427.2009.02272.x

45. Desta H, Lemma B, Fetene A. Aspects of climate change and its associated impacts on wetland ecosystem functions: A review. J Am Sci. 2012;8:582–596.

46. Sala OE, Chapin IFS, Armesto JJ. Global biodiversity scenarios for the year 2100. Science. 2000;287:1770–1774. http://dx.doi.org/10.1126/science.287.5459.1770

47. Ricciardi A, Rasmussen JB. Extinction rates of North American freshwater fauna. Conserv Biol .1999;13:1220–1222. http://dx.doi.org/10.1046/j.1523-1739.1999.98380.x

48. Dudgeon D, Arthington AH, Gessner MO, Kawabata Z, Knowler DJ, Lêvêque C, et al. Freshwater biodiversity: Importance, threats, status and conservation challenges. Biol Rev. 2006;81:163–182. http://dx.doi.org/10.1017/S1464793105006950

49. Darwall WRT, Smith KG, Tweddle D, Skelton P. The status and distribution of freshwater biodiversity in southern Africa. The IUCN Red List of Threatened Species – Regional Assessment. Gland: IUCN; 2009.

50. Nel JL, Driver A, Strydom W, Maherry A, Petersen C, Hill L, et al. Atlas of freshwater ecosystem priority areas in South Africa: Maps to support sustainable development of water resources. Pretoria: Water Research Commission; 2011.

51. Driver A, Sink K, Nel JL, Holness S, Van Niekerk L, Daniels F, et al. National Biodiversity Assessment 2011: An assessment of South Africa's biodiversity and ecosystems – Synthesis Report. Pretoria: South African National Biodiversity Institute and Department of Environmental Affairs; 2012.

52. Ashton P. Riverine biodiversity conservation in South Africa: Current situation and future prospects. Aquat Conserv. 2007;17:441–445.

53. Weitjers MJ, Janse JH, Alkemade R, Verhoeven JTA. Quantifying the effect of catchment and use and water nutrient concentrations on freshwater river and stream biodiversity. Aquat Conserv. 2009;19:104–112.

54. Amis MA, Rouget M, Balmford A, Thuiller W, Kleynhans CJ, Day J, et al. Predicting freshwater habitat integrity using land-use surrogates. Water SA. 2006;33:215–222.

55. Allan JD. Landscapes and riverscapes: The influence of land use on stream ecosystems. Annu Rev Ecol Sys. 2004;35:257–284. http://dx.doi.org/10.2989/16085914.2012.763110

56. Eady BR, Rivers-Moore NA, Hill TR. Relationship between water temperature predictability and aquatic macroinvertebrate assemblages in two South African streams. Afr J Aquat Sci. 2013;38:163–174.

57. King JM. The distribution of invertebrate communities in a small South African river. Hydrobiologia. 1981;83:43–65. http://dx.doi.org/10.1007/BF02187150

58. Ractliffe SG. Disturbance and temporal variability in invertebrate assemblages in two South African rivers [PhD thesis]. Cape Town: University of Cape Town; 2009.

59. Ross-Gillespie V, Dallas HF. Water temperatures and the reserve (WRC Project: K5/1799): Key life-history traits and thermal cues of selected aquatic macroinvertebrates. Report number 1799/20 produced for the Water Research Commission. Freshwater Research Unit, University of Cape Town and the Freshwater Consulting Group; 2011.

60. Minshall GW, Petersen RC, Nimz CF. Species richness of streams of different size from the same drainage basin. Am Nat .1985;125:16–38. http://dx.doi.org/10.1086/284326

61. Bush A, Nipperess D, Turak E, Hughes L. Determining vulnerability of stream communities to climate change at the landscape scale. Freshwater Biol. 2012;57:1689–1701. http://dx.doi.org/10.1111/j.1365-2427.2012.02835.x

62. Heino J, Virkkala R, Toivonen H. Climate change and freshwater biodiversity: Detected patterns, future trends and adaptations in northern regions. Biol Rev. 2009;84:39–54. http://dx.doi.org/10.1111/j.1469-185X.2008.00060.x

63. Rieman BE, Isaak D, Adams S. Anticipated climate warming effects on bull trout habitats and populations across the interior Columbia Basin. T Am Fish Soc. 2007;136:1552–1565. http://dx.doi.org/10.1577/T07-028.1

64. Sheldon AL. Possible climate-induced shift of stoneflies in a southern Appalachian catchment. Freshwater Sci. 2012;31:765–777. http://dx.doi.org/10.1899/11-135.1

65. Chessman BC. Biological traits predict shifts in geographical ranges of freshwater invertebrates during climatic warming and drying. J Biogeog. 2012;39:957–969. http://dx.doi.org/10.1111/j.1365-2699.2011.02647.x

66. Millennium Ecosystem Assessment. Ecosystems and human well-being: Synthesis. Washington, DC: Island Press; 2005.

67. Williams SE, Bolitho EE, Fox S. Climate change in Australian tropical rainforests: An impending environmental catastrophe. Philos Trans R Soc Lond B Biol Sci. 2003;270:1887–1892. http://dx.doi.org/10.1098/rspb.2003.2464

68. Andreone F, Cadle JE, Cox N, Glaw F, Nussbaum RA, Raxworthy CJ, et al. Species review of amphibian extinction risks in Madagascar: Conclusions from the global amphibian assessment. Conserv Biol. 2005;19:1790–1802. http://dx.doi.org/10.1111/j.1523-1739.2005.00249.x

69. Tweddle D, Bills R, Swartz E, Coetzer W, Da Costa L, Engelbrecht J, et al. The status and distribution of freshwater biodiversity in southern Africa. In: Darwall WRT, Smith KG, Tweddle D, Skelton P, editors. The status and distribution of freshwater biodiversity in southern Africa. Gland & Grahamstown: IUCN & South African Institute for Aquatic Biodiversity; 2009. p. 21–37.

70. Turak E, Marchant R, Barmuta LA, Davis J, Choy S, Metzeling L. River conservation in a changing world: Invertebrate diversity and spatial prioritization in south-eastern coastal Australia. Mar Freshwater Res. 2011;62:300–311. http://dx.doi.org/10.1071/MF09297

71. Lauzeral C, Leprieur F, Beauchard QD, Oberdorff T, Brosse S. Identifying climatic niche shifts using coarse-grained occurrence data: A test with non-native freshwater fish. Glob Ecol Biogeog. 2010;20:407–414. http://dx.doi.org/10.1111/j.1466-8238.2010.00611.x

72. Staudinger MD, Grimm NB, Staudt A, Carter SL, Chapin FS, Kareiva P, et al. Impacts of climate change on biodiversity, ecosystems, and ecosystem services: Technical input to the 2013 National Climate Assessment. Cooperative report to the 2013 National Climate Assessment; 2012.

73. Rahel FJ, Olden JD. Assessing the effects of climate change on aquatic invasive species. Conserv Biol. 2008;22:521–533. http://dx.doi.org/10.1111/j.1523-1739.2008.00950.x

74. Sorte CJB, Ibáñez I, Blumenthal DM, Molinari NA, Miller LP, Grosholz ED, et al. Poised to prosper? A cross-system comparison of climate change effects on native and non-native species performance. Ecol Lett. 2013;16(2):261–270. http://dx.doi.org/10.1111/ele.12017

75. Rivers-Moore NA, De Moor FC, Morris C, O'Keeffe J. Effect of flow variability modification and hydraulics on invertebrate communities in the Great Fish River (Eastern Cape Province, South Africa), with particular reference to critical hydraulic thresholds limiting larval densities of *Simulium chutteri* Lewis (Diptera, Simuliidae). River Res Appl. 2007;23:201–222. http://dx.doi.org/10.1002/rra.976

76. Lowe SR, Woodford DJ, Impson DN, Day JA. The impact of invasive fish and invasive riparian plants on the invertebrate fauna of the Rondegat River, Cape Floristic Region, South Africa. Afr J Aquat Sci. 2008;331:51–62. http://dx.doi.org/10.2989/AJAS.2007.33.1.6.390

77. Rural Fisheries Programme. A manual for rural freshwater aquaculture. Water Research Commission report TT 463/P/10. Pretoria: Water Research Commission; 2010.

78. Hunter PR. Climate change and waterborne and vector-borne disease. J Appl Microbiol Symp Suppl. 2003;94:37S–46S.

79. Paaijmans KP, Read AF, Thomas AB. Understanding the link between malaria risk and climate. Proc Natl Acad Sci USA. 2009;106:13844–13849. http://dx.doi.org/10.1073/pnas.0903423106

80. Tanser FC, Sharp B, Le Sueur D. Potential effect of climate change on malaria transmission in Africa. Lancet. 2003;362:1792–1798. http://dx.doi.org/10.1016/S0140-6736(03)14898-2

81. Palmer MA, Lettenmaier DP, Poff NL, Postel SL, Richter B, Warner R. Climate change and river ecosystems: Protection and adaptation options. Environ Manage. 2009;44:1053–1068. http://dx.doi.org/10.1007/s00267-009-9329-1

82. King J, Pienaar H. Sustainable use of South Africa's inland waters. Water Research Commission report TT 491/11. Pretoria: Water Research Commission; 2011.

83. Poff NL. Managing for variation to sustain freshwater ecosystems. J Water Res Pl-Asce. 2009;135:1–4. http://dx.doi.org/10.1061/(ASCE)0733-9496(2009)135:1(1)

84. Poff NL, Richter BD, Arthington AH. The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards. Freshwater Biol. 2009;55:147–170. http://dx.doi.org/10.1111/j.1365-2427.2009.02204.x

85. Brown CA, Joubert A. Using multicriteria analysis to develop environmental flow scenarios for rivers targeted for water resource development. Water SA. 2003;29:365–374.

86. King J, Brown C, Sabet H. A scenario-based holistic approach to environmental flow assessments for rivers. River Res Appl. 2003;19:619–639.

87. King J, Brown C. Environmental flows: Striking the balance between development and resource protection. Ecol Soc. 2006;11:26. http://dx.doi.org/10.1002/rra.709

88. Pittock J, Hansen L, Abell R. Running dry: Freshwater biodiversity, climate change and protected areas. Biodiversity. 2008;9:30–37. http://dx.doi.org/10.1080/14888386.2008.9712905

89. Palmer MA, Reidy Liermann CA, Nilsson C, Flörke M, Joseph Alcamo J, Lake PS, et al. Climate change and the world's river basins: Anticipating management options. Front Ecol Environ. 2008;6:81–89. http://dx.doi.org/10.1890/060148

90. Poff L, Brinson M, Day Jr J. Aquatic ecosystems and global climate change: Potential impacts on inland freshwater and coastal wetland ecosystems in the United States. Philos Trans R Soc Lond B Biol Sci.2002;360:1497–1510. http://dx.doi.org/10.1098/rsta.2002.1012

91. Department of Water Affairs and Forestry, South Africa (DWAF). Management of the water resources of the Republic of South Africa. Pretoria: DWAF; 1986.

92. O'Keeffe JH, De Moor FC. Changes in the physico-chemistry and benthic invertebrates of the Great Fish River, South Africa, following an interbasin transfer of water. Regul River. 1988;2:39–55. http://dx.doi.org/10.1002/rrr.3450020105

93. O'Keeffe JH, Palmer RW, Byren BA, Davies BR. The effects of impoundment on the physiochemistry of two contrasting southern African river systems. Regul River. 1990;5:97–110. http://dx.doi.org/10.1002/rrr.3450050202

94. Palmer RW, O'Keeffe JH. Transported material in a small river with multiple impoundments. Freshwater Biol. 1990;24:563–575. http://dx.doi.org/10.1111/j.1365-2427.1990.tb00733.x

95. Palmer RW, O'Keeffe JH. Downstream effects of impoundments on the water chemistry of the Buffalo River (Eastern Cape), South Africa. Arch Hydrob. 1990;202:71–83. http://dx.doi.org/10.1007/BF02208128

96. Palmer RW, O'Keeffe JH. Downstream effects of a small impoundment on a turbid river. Arch Hydrob. 1990;119:457–473.

97. Palmer RW, O'Keeffe JH. Distribution and abundance of blackflies (Diptera: Simuliidae) in relation to impoundments in the Buffalo River, Eastern Cape, South Africa. Freshwater Biol. 1995;33:109–118. http://dx.doi.org/10.1111/j.1365-2427.1995.tb00391.x

98. Snaddon C, Davies B, Wishart M. A global overview of inter-basin water transfer schemes: Ecological socio-economic and socio-political implications and recommendations for their management. Water Research Commission report TT 120/00. Pretoria: Water Research Commission; 2000.

99. Bok A, Rossouw J, Rooseboom A. Guidelines for the planning, design and operation of fishways in South Africa. Water Research Commission report 1270/2/04. Pretoria: Water Research Commission; 2004.

100. Ashton P. An overview of the current status of water quality in South Africa and possible trends of change. Pretoria: CSIR Water Ecosystems and Human Health Research Group; 2009.

101. Schulze R, Meigh J, Horan M. Present and potential future vulnerability of eastern and southern Africa's hydrology and water resources. S Afr J Sci. 2001;97:150–160.

102. Department of Water Affairs (DWA). Procedures to develop and implement resource quality objectives. Pretoria: Department of Water Affairs; 2011.

103. Margules CR, Pressey RJ. Systematic conservation planning. Nature. 2000;405:243–253. http://dx.doi.org/10.1038/35012251

104. Roux DJ, Nel JL. Freshwater conservation planning in South Africa: Milestones to date and catalysts for implementation. Water SA. 2013;39:151–163.

105. Nel JL, Reyers B, Roux DJ, Impson ND, Cowling RM. Designing a conservation area network that supports the representation and persistence of freshwater biodiversity. Freshwater Biol. 2011;56:106–124. http://dx.doi.org/10.1111/j.1365-2427.2010.02437.x

106. Rivers-Moore NA, Goodman PS, Nel JL. Scale-based freshwater conservation planning: Towards protecting freshwater biodiversity in KwaZulu-Natal, South Africa. Freshwater Biol. 2011;56:125–141. http://dx.doi.org/10.1111/j.1365-2427.2010.02387.x

107. World Bank. Flowing forward: Freshwater ecosystem adaptation to climate change in water resources management and biodiversity conservation. Working Note 28; 2010.

108. Rivers-Moore NA. Turnover patterns in fish versus macroinvertebrates – Implications for conservation planning. Afr J Aquat Sci. 2012;37:301–309. http://dx.doi.org/10.2989/16085914.2012.708857

109. Dunlop M, Brown PR. Implications of climate change for Australia's National Reserve system: A preliminary assessment. Report to the Department of Climate Change, Australia; 2008.

110. Richter BD, Baumgartner JV, Powell J, Braun DP. A method for assessing hydrologic alteration within ecosystems. Conserv Biol. 1996;10:1163–1174. http://dx.doi.org/10.1046/j.1523-1739.1996.10041163.x

111. Richter BD, Baumgartner JV, Wigington R, Braun DP. How much water does a river need? Freshwater Biol. 1997;37:231–249. http://dx.doi.org/10.1046/j.1365-2427.1997.00153.x

112. Nel JL, Turak E, Linke S, Brown C. A new era in catchment management: Integration of environmental flow assessment and freshwater conservation planning. Mar Freshwater Res. 2010;62:1–10.

113. Olden JD, Naiman RJ. Incorporating thermal regimes into environmental flows assessments: Modifying dam operations to restore freshwater ecosystem integrity. Freshwater Biol. 2010;55:86–107. http://dx.doi.org/10.1111/j.1365-2427.2009.02179.x

114. Hildrew AG, Giller PS. Patchiness, species interactions and disturbance in the stream benthos. In: Hildrew AG, Giller PS, Raffaeli D, editors. Aquatic ecology: Scale, pattern and process. London: Blackwell Science; 1994.

115. Nelson DR, Adger WN, Brown K. Adaptation to environmental change: Contributions of a resilience framework. Annu Rev Env Resour. 2007;32:395–419. http://dx.doi.org/10.1146/annurev.energy.32.051807.090348

116. Lawler JL. Climate change adaptation strategies for resource management and conservation planning. Ann N Y Acad Sci. 2009;1162:79–98. http://dx.doi.org/10.1111/j.1749-6632.2009.04147.x

117. Arthington AH, Naiman RJ, McClain ME. Preserving the biodiversity and ecological services of rivers: New challenges and research opportunities. Freshwater Biol. 2010;55:1–17. http://dx.doi.org/10.1111/j.1365-2427.2009.02340.x

118. Nel JL, Roux DJ, Maree G, Kleynhans CJ, Moolman J, Reyers B, et al. Rivers in peril inside and outside protected areas: A systematic approach to conservation assessment of river ecosystems. Divers Distrib. 2007;13:341–352. http://dx.doi.org/10.1111/j.1472-4642.2007.00308.x

119. Midgley GF, Chapman RA, Hewitson B, Johnston P, De Wit M, Ziervogel G, et al. A status quo, vulnerability and adaptation assessment of the physical and socio-economic effects of climate change in the Western Cape. Report to the Western Cape Government, Cape Town, South Africa. Report no. ENV-S-C 2005-073. Stellenbosch: CSIR; 2005.

Modelling new particle formation events in the South African savannah

AUTHORS:
Rosa T. Gierens[1]
Lauri Laakso[2,3]
Ditte Mogensen[1]
Ville Vakkari[1]
Johan P. Beukes[3]
Pieter G. van Zyl[3]
Hannele Hakola[2]
Alex Guenther[4,5]
Jacobus J. Pienaar[3]
Michael Boy[1]

AFFILIATIONS:
[1]Department of Physics, Helsinki University, Helsinki, Finland
[2]Research and Development, Finnish Meteorological Institute, Helsinki, Finland
[3]Unit for Environmental Sciences and Management, North-West University, Potchefstroom, South Africa
[4]Pacific Northwest National Laboratory, Richland, Washington, USA
[5]Washington State University, Pullman, Washington, USA

CORRESPONDENCE TO:
Rosa Gierens

EMAIL:
rosa.gierens@helsinki.fi

POSTAL ADDRESS:
Department of Physics, Helsinki University, PO Box 48, Helsinki 00014, Finland

DATES:
Received: 15 Apr. 2013
Revised: 14 Oct. 2013
Accepted: 30 Oct. 2013

KEYWORDS:
atmospheric modelling; aerosols; nucleation; boundary layer; savannah

HOW TO CITE:
Gierens RT, Laakso L, Mogensen D, Vakkari V, Beukes JP, Van Zyl PG, et al. Modelling new particle formation events in the South African savannah. S Afr J Sci. 2014;110(5/6), Art. #2013-0108, 12 pages. http://dx.doi.org/10.1590/sajs.2014/20130108

Africa is one of the less studied continents with respect to atmospheric aerosols. Savannahs are complex dynamic systems sensitive to climate and land-use changes, but the interaction of these systems with the atmosphere is not well understood. Atmospheric particles, called aerosols, affect the climate on regional and global scales, and are an important factor in air quality. In this study, measurements from a relatively clean savannah environment in South Africa were used to model new particle formation and growth. There already are some combined long-term measurements of trace gas concentrations together with aerosol and meteorological variables available, but to our knowledge this is the first detailed simulation that includes all the main processes relevant to particle formation. The results show that both of the particle formation mechanisms investigated overestimated the dependency of the formation rates on sulphuric acid. From the two particle formation mechanisms tested in this work, the approach that included low volatile organic compounds to the particle formation process was more accurate in describing the nucleation events than the approach that did not. To obtain a reliable estimate of aerosol concentration in simulations for larger scales, nucleation mechanisms would need to include organic compounds, at least in southern Africa. This work is the first step in developing a more comprehensive new particle formation model applicable to the unique environment in southern Africa. Such a model will assist in better understanding and predicting new particle formation – knowledge which could ultimately be used to mitigate impacts of climate change and air quality.

Introduction

Savannahs are highly dynamic systems covering an area of approximately 16 x 10^6 km^2, or 11.5% of the global land surface. They are sensitive to climate and land-use alterations, which can lead to fast changes in biomass and soil properties.[1] These areas lie mostly in the developing world where vast land-use changes take place.[2] Savannahs affect the regional and global climate, but the interactions between the biosphere and the atmosphere have not undergone intense investigation and are therefore still not well understood.[3] In order to understand the global climate, it therefore is important to study the savannah environment.

Aerosols modify the climate directly by affecting the Earth's radiative budget by scattering and absorbing solar radiation, as well as indirectly by acting as cloud condensation nuclei and thereby changing the properties of the clouds, which again alter the radiation budget. According to recent studies based on global modelling and observations, new particle formation contributes to these aerosol effects.[4-8] In addition to the climate impacts, aerosols are an important factor for air quality and thus influence human health. In developed countries, strong measures have been taken to improve air quality, but in the developing world aerosols harm the health of hundreds of millions of people.[9]

Africa is one of the less studied continents with regard to atmospheric aerosols both from a measurement and a modelling point of view. Biomass burning is a significant source of aerosols in Africa, but previous studies have shown that there are also several other sources.[10] A few modelling studies using regional and global models have investigated South Africa,[11,12] but none have used detailed chemistry and aerosol dynamics. In this case study, we used measurements conducted at a semi-clean savannah site in South Africa to simulate new particle formation and growth. The measurements were carried out with a transportable measurement trailer.[13] The observations were used as input to the model to constrain the result. The model used in this work was MALTE (Model to predict new Aerosol formation in the Lower TroposphereE).[14,15] Our aim was to investigate the processes creating secondary organic aerosols and explore what factors are important for the growth of these aerosols, especially in southern Africa. Furthermore, we wanted to test the model in a savannah environment, which is very different from the boreal forest ecosystem that the model was developed for. In addition to being a method to enhance our understanding of the processes studied, modelling also allows one to extrapolate to places without measurements. The reliability of such extrapolated results will obviously depend on knowledge of the local conditions. Therefore modelling is a useful tool for compensating when measurements are sparse, such as for areas like southern Africa.

Methods

Site characteristics

The measurements were conducted at a station in the Botsalano Game Reserve (located at 25.54°S and 25.75°E, 1400 m above sea level) in the North West Province of South Africa. Botsalano is located 50 km north of the nearest city, Mafikeng, which has a population of 260 000, and approximately 100 km south of Gaborone (population 200 000), the capital of Botswana. The mining and metallurgical region in the vicinity of the cities of Rustenburg and Brits in the Western Bushveld Indigenous Complex,[16] which is one of the largest regional pollution sources, is located approximately 150 km east of Botsalano. This area is a platinum group mineral, chromium and base metal mining and metallurgical extraction region, which results in enhanced SO_2 and associated sulphate emissions. Otherwise, the nearby region east of Botsalano is quite sparsely populated. The Johannesburg–Pretoria megacity, with more than 10 million inhabitants and heavy industry, is located 300 km east to southeast of the site.[17] There are some small cities and industry in the region, but the area west and south of Botsalano has few anthropogenic activities. There are some small local pollutant sources such as traffic and biomass burning. Considering all the aforementioned, Botsalano can be considered to be representative of a semi-clean background site, with occasional higher pollutant concentrations associated with easterly winds.[18,19] The location of Botsalano and surroundings is shown in Figure 1 and a detailed description of the site is given by Laakso et al.[10]

The Botsalano Game Reserve is located in dry bushveld, with a fairly homogeneous vegetation of grass and thin tree stands.[10] The dominant trees and shrubs are mostly *Acacia*, *Rhus*, *Ziziphus*, *Vitex* and *Grewia*, which are typical for a savannah biome. There is also different grass and herbaceous species. The height of the canopy is approximately 8 m. Animals roam freely through the reserve. The large-scale meteorology is characterised by a high degree of stability and anticyclonic circulation taking place more than 50% of the time year-round.[20] As a result of the high stability, the vertical mixing is limited and thus creates situations where air masses are contaminated either by industrial sources or by biomass burning.[21] Re-circulation occurs on regional and sub-continental scales and often in the order of several days.[20] New particle formation events occurred on 69% of the days when measurements were taken, which is a relatively high frequency for continental boundary layer conditions, and only 6% of the days showed clearly no signs of new particle formation.[18]

Measurements

The measurements were made with a mobile measurement trailer. The trailer is aimed to be a self-sufficient and mobile monitoring station requiring only three-phase power and periodic maintenance of the instruments. In addition to measuring instruments, the trailer is equipped with a GPRS modem, so that data can be automatically, and wirelessly, copied to a server, enabling remote monitoring of data quality. A detailed description of the trailer is given by Petäjä et al.[13] During the measurement period described here (July 2007 – January 2008), the trailer was not moved. All data were quality controlled for unreliable measurements, which occurred mostly as a result of frequent electricity breaks.[10]

For measuring the particle number distribution, a differential mobility particle sizer was used with a size range from 10 nm to 840 nm and a time resolution of 7.5 min. The meteorological instruments were mounted on a mast located on the roof of the trailer at a height of 3.7 m above the ground. The meteorological parameters together with trace gas concentrations were logged at a time resolution of 1 min. The measured parameters, together with the instruments used, are shown in Table 1. The gas concentration data were corrected based on on-site calibrations. Volatile organic compound (VOC) measurements were conducted using adsorbent tubes filled with Tenax-TA and Carbopack-B with a sampling time of 2 h. The samples were analysed using a thermodesorption instrument (Perkin-Elmer TurboMatrix 650) attached to a gas chromatograph (Perkin-Elmer Clarus 600) and a mass selective detector (Perkin-Elmer Clarus 600T). The column used was a DB-5MS (60 m, 0.25 mm, 1 μm) column. The sample tubes were desorbed at 300 °C for 5 min and cryofocused in a Tenax cold trap (-30 °C) prior to injecting the sample into the column by rapidly heating the cold trap (40 °C/min) to 300 °C. A five-point calibration was conducted using liquid standards in methanol solutions. Standard solutions were injected onto adsorbent tubes and flushed with nitrogen flow for 5 min in order to remove methanol. The VOC sampling system did not

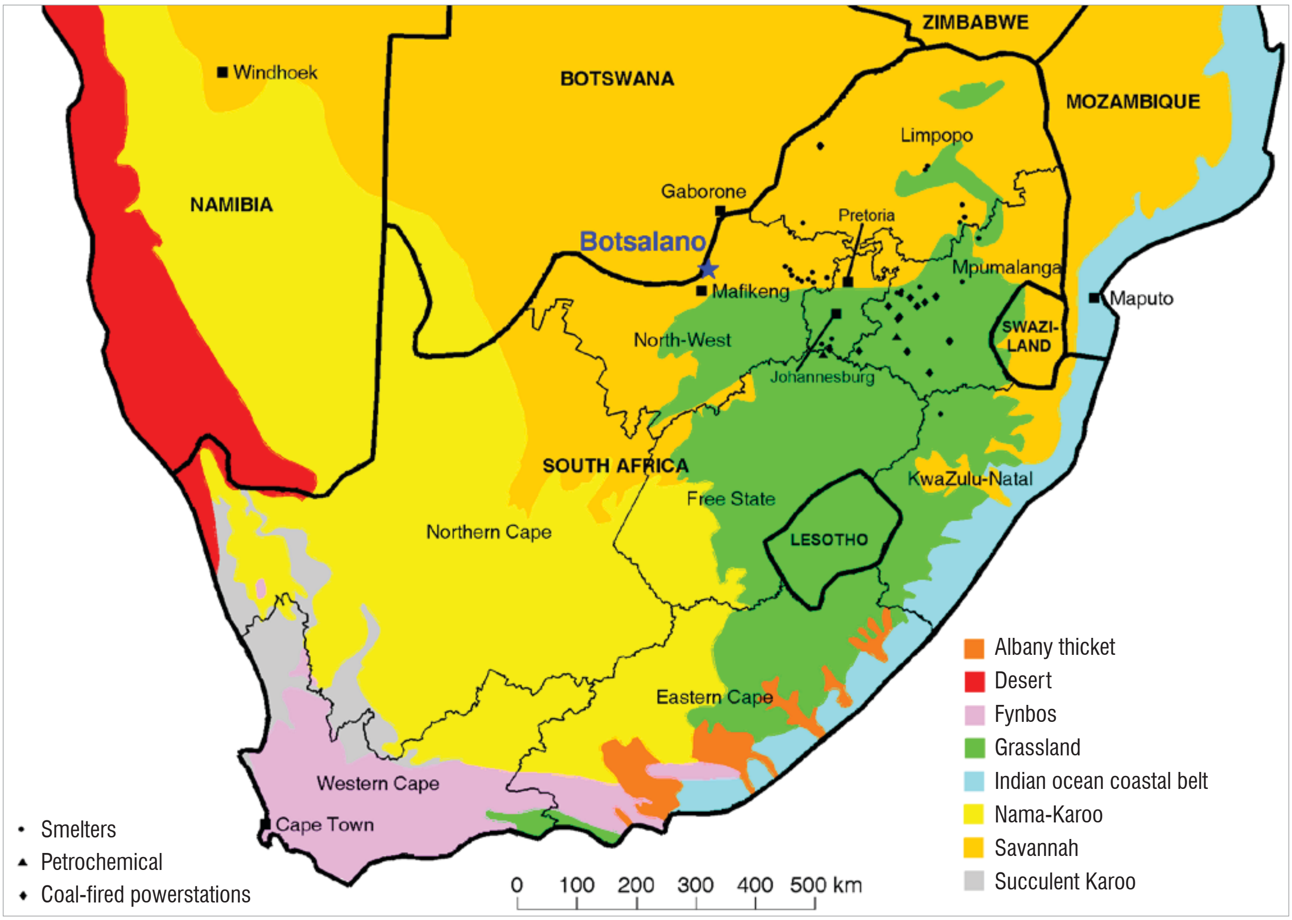


Figure 1: Location of the Botsalano measurement site in South Africa within a regional context. The nearest city, Mafikeng, as well as Johannesburg, Pretoria and Cape Town are shown. All major atmospheric point sources within an approximately 500-km radius around Botsalano and the biomes in southern Africa are also indicated.

include ozone removal and therefore the VOC mixing ratios may be considered as the lower limits only for the ozone reactive compounds such as monoterpenes and isoprene. VOC measurements were made only intermittently, which had to be taken into account when selecting the days to be studied.

Days selected

We chose 6 days (7–10 and 14–15 October 2007) during the local spring; each day had different characteristics with respect to trace gas and particle concentrations. This period coincided with the beginning of the growing season. On these days, the temperature varied between a night-time low of 5 °C and a daytime maximum of 32 °C, and the relative humidity from 9% to 99%. The predominant wind direction was from the north, but various directions were observed (Table 1). There was a nucleation event during all of the studied days, which included days with clean background air as well as higher pollutant concentrations. Most of the days were sunny, and rain occurred only once during the evaluated period.

Model description

MALTE is a one-dimensional model that simulates new particle formation and growth in the lower troposphere, from the surface up to 3 km above the ground, covering the entire boundary layer. It includes modules for the boundary layer meteorology, chemical and aerosol dynamic processes, and for emissions from the canopy.[14] In this study we used a further developed version, in which the one-dimensional version of the model SCADIS is used to improve the meteorology scheme.[15] This modification was made to improve the estimates of vertical turbulent fluxes of heat, moisture and scalars.

MALTE has a flexible number of vertical layers. The layer height increases logarithmically from the ground upwards, such that the resolution is highest near the ground. A total of 51 layers from the ground to a height of 3000 m were chosen for this study, with 14 of the layers located inside the canopy (lowest 8 m). Based on meteorological conditions at each level, the detailed chemical and aerosol dynamic processes were solved. Meteorology, emissions, chemistry and aerosols were combined with a split-operator approach: the meteorological conditions, as well as fluxes of particles and gases, were calculated with a 10-s time step for 60 s. Afterwards, emissions from the canopy, chemical reactions and aerosol dynamic processes at every atmospheric layer were solved with a time step of 60 s. Then the meteorological module started again and the process continued. From the meteorological model point of view, changes in the gas phase and in particle concentrations appeared instantly after each minute.[15]

Boundary layer meteorology

For the boundary layer processes, the one-dimensional version of the model SCADIS (scalar distribution) was applied. A detailed description of the model is given by Sogachev et al.[22-24] This part of the model solves the meteorological conditions in and above the atmospheric boundary layer, the vertical transport by turbulence, as well as the interactions with and within the canopy. SCADIS includes a set of equations for momentum, continuity, heat, humidity and transport of a passive tracer. For solving the turbulent fluxes, the one-and-a-half order closure is applied. The canopy is described with multiple layers. SCADIS includes parameterisations for drag forces of leaves and for radiation transfer, distinguishing between sunlit and shaded leaf area, and thus aims to properly describe interactions between vegetative canopy and the atmosphere. In the model, a horizontally homogenous vegetation canopy is assumed. Penetration of solar radiation inside the canopy was calculated based on the leaf area index, and gave the average radiation flux at each level. In a bushy area like Botsalano, this was a challenging approach. Because of the one-dimensional nature of the model, there is no separation between the area between the trees and 'inside' the trees. From the model point of view, the canopy was treated as if it were horizontally homogenous with very sparse foliage. Prognostic equations for soil moisture were also included.

Emissions from the canopy

Isoprene and monoterpene emission rates have been reported for many important savannah woody (tree and shrub) genera in southern Africa including the dominant genera at Botsalano Game Reserve.[25-28] In contrast, there are few measurements of biogenic emissions from herbaceous savannah vegetation. These species are generally assumed to make a negligible contribution to the total biogenic VOC flux, but this assumption could result in underestimated emissions. The five dominant woody genera at Botsalano include one with high isoprene and monoterpene emissions (*Rhus*), one low monoterpene emitter (*Grewia*) and two very low emitters (*Vitex* and *Ziziphus*). The fifth genera, *Acacia*, is more diverse and includes high isoprene emitters (e.g. *A. nigrescens* and *A. polycantha*), high monoterpene emitters (e.g. *A. tortilis*) and species with low emissions of both (e.g. *A. nilotica*).[25,28] In this context, the leaf level emission rate categories are based on the definitions of Guenther et al.[29] The isoprene emission classification comprises high ($\geq$14 μg/(g h)), low (0.1 to 14 μg/(g h)) and very low ($<$0.1 μg/(g h)). The classes for monoterpenes are high ($\geq$1.6 μg/(g h)), low (0.1 to 1.6 μg/(g h)) and very low ($<$0.1 μg/(g h)).[29] The high variability of the woody covered fraction and the diversity of high and low emitters results in very different emission rates from the various savannah landscapes. Otter et al.[26] combined a regional database of plant species composition and isoprene and monoterpene emission rates to estimate biogenic emission rates from land-cover types in southern Africa. They found

Table 1: Instruments used for measuring meteorological parameters and trace gas concentrations at the Botsalano Game Reserve. The range of values measured during October 2007 are presented to describe the characteristics of the savannah environment.

Parameter	Instrumentation (manufacturer, country)	Measured range
Temperature, relative humidity	Rotronic MP 101A (Rotronic, Switzerland)	5.3 – 32 °C, 9.4 – 99%
Wind speed	Vector W200P (Vector Instruments, United Kingdom)	0.1 – 20 m/s
Precipitation	Thies 5.4103.20.041 (Thies Clima, Germany)	0 – 36 mm/h
Wind direction	Vector A101ML (Vector Instruments, United Kingdom)	0 – 360°
Photosynthetically active radiation	LiCor LI-190SB (Li-Cor, USA)	0.0015 – 2.1 μmol/s
SO_2	Thermo-Electron 43S (Thermo Scientific, USA)	0.1 – 74 ppb
NO_x	Teledyne 200AU (Teledyne API, USA)	0.1 – 11 ppb
CO	Horiba APMA-360 (Horiba, Japan)	76 – 280 ppb
O_3	Environnement s.a 41M (Environnement S.A, France)	5.6 – 73 ppb

a large range in isoprene (<1 to >13 g/m^2 per year) and monoterpene (<0.5 to >5 g/m^2 per year) emission rates for different southern African savannahs.

Emissions of isoprene, monoterpenes and other VOCs from the canopy were calculated with the model MEGAN (Model of Emissions of Gases and Aerosols from Nature).[30,31] The compounds included were: isoprene, 2-methyl-3-buten-2-ol, β-pinene, α-pinene, methanol, acetone, acetaldehyde, formaldehyde, formic acid, acetic acid, methane, sabinene, Δ^3-carene, myrcene, limonene, trans-β-ocimene, β-caryophyllene, α-farnescene, 'other monoterpenes' (for those not mentioned here) and 'other sesquiterpenes' (for those not mentioned here). The isoprene and monoterpene emission factors used for this simulation are based on the estimates by Otter et al.[26] for the savannah type located at the Botsalano Game Reserve. The emission rates for all other compounds are based on the emission factors recommended by Guenther et al.[31] The emission rates depend also on leaf temperature and available solar radiation for leaves in the sun and shade, which are calculated separately for every model level inside the canopy. No anthropogenic VOC sources were included in the model.

Gas phase chemistry

The gas phase concentrations of all compounds were calculated for every model time step by the Kinetic PreProcessor.[32] In total, 770 chemical species and 2148 chemical and photochemical reactions chosen from the Master Chemical Mechanism (MCM) version 3.2 [33,34] via http://mcm.leeds.ac.uk./MCM/ were included for the model runs. These numbers include atmospherically relevant inorganic compounds and reactions together with the full oxidation chemistry paths for isoprene, 2-methyl-3-buten-2-ol, β-pinene, α-pinene, methanol, acetone, acetaldehyde, formaldehyde, formic acid, acetic acid and methane. We also included first-order oxidation reactions among $^{\bullet}OH$, O_3, $NO_3^{\bullet}$ and the following organic compounds: sabinene, Δ^3-carene, myrcene, limonene, ocimene (to simulate the chemistry of trans-β-ocimene), 'other monoterpenes' (those not mentioned here), β-caryophyllene, farnesene (to simulate the chemistry of α-farnescene), and 'other sesquiterpenes' (other than the two mentioned here). The full chemistry paths for these compounds (we are referring to the following compounds: sabinene, Δ^3-carene, myrcene, ocimene, 'other monoterpenes', farnesene, and 'other sesquiterpenes') are not included, as they are unknown and not provided by MCM. In the case of limonene and β-caryophyllene, for which the full near-explicit MCM chemistries are available, we chose not to include them because of computational costs; limonene and β-caryophyllene comprise 539 and 591 compounds, respectively, and MALTE is not written in parallel. Furthermore, as the predicted emissions of limonene and β-caryophyllene are relatively low compared with, especially, α-pinene and β-pinene, it is reasonable to exclude the chemistry of these minor components.

Aerosol dynamics

We used the size-segregated aerosol dynamics model UHMA (University of Helsinki Multicomponent Aerosol model)[35] to simulate the aerosol dynamic processes occurring in our study. UHMA describes in detail new particle formation and growth, including both particle coagulation and multicomponent condensation. The fixed-sectional method with 38 size bins was applied to represent particle size distributions. The model has an option to include several different nucleation mechanisms. We first used kinetic nucleation,[36] in which critical clusters are formed by collisions of sulphuric acid molecules (H_2SO_4) or other molecules containing sulphuric acid. The nucleation rate J_{kin} is calculated by

$$J_{kin} = K \cdot [H_2SO_4]^2 \qquad \text{Equation 1}$$

where K is the kinetic coefficient that contains details of the nucleation process, such as the probability of a collision of two molecules containing sulphuric acid resulting in the formation of a stable critical cluster.

It has been suggested that gases other than sulphuric acid participate in the nucleation process; for example ammonia or low volatile organic vapours might play a role.[37,38] The second nucleation mechanism used in this work assumes the formation of stable critical clusters by collision of sulphuric acid or a molecule containing sulphuric acid and a low volatile organic compound. The nucleation rate J_{org} is calculated in the same way as by Lauros et al.[15]:

$$J_{org} = P \cdot \upsilon \cdot [H_2SO_4] \cdot [C_{org}] \qquad \text{Equation 2}$$

where $[C_{org}]$ represents the concentration of low volatile organic gases (in the model these are the first stable reaction products of α-pinene, β-pinene and isoprene with OH), υ is the collision rate of the molecules and the coefficient P contains details about the nucleation process, similarly as for K in Equation 1.

The newly formed nanosize clusters grow by condensation of sulphuric acid and low volatile organic compounds (C_{org} alike as in nucleation)[14] following the nano-Köhler theory.[39] For particles reaching a size of a few nanometres, sulphuric acid and the first stable reaction products of α-pinene, β-pinene and isoprene generated by reactions with $^{\bullet}OH$, O_3 and $NO_3^{\bullet}$ participate in the growth of the particles. Oxidation products from α-pinene, β-pinene and isoprene reactions were chosen to represent the oxidation products of the VOCs, as there is no currently available method in MALTE to estimate saturation vapour pressure. Other monoterpenes (myrcene, sabinene, limonene, Δ^3-carene and ocimene) and sesquiterpenes (farnescene and β-caryophylllene) were not chosen, because full chemistry for these compounds was not included. For the remaining organic compounds, we predicted concentrations that are too low and expected saturation vapour pressures that are too high for these compounds to make a significant condensation contribution. We scaled the concentration of condensing vapours by taking a fraction of 0.05 of the abovementioned compounds gas phase concentrations based on measurement fit.

The model calculates dry deposition for newly formed particles as well as deposition to the canopy. Because MALTE does not simulate the formation of clouds or precipitation, wet deposition was not included.

Model set-up

All measured meteorological parameters were used in the model. Temperature, absolute humidity and wind speed were used for nudging with a factor of 0.1. In this way we were able to simulate the meteorological conditions in the model more closely to the observations, which allowed some of the effects of synoptic scale weather phenomena to be included. Upper boundary values for temperature, humidity and wind speed are required in order to simulate the vertical profiles. Because of the non-existence of soundings in the region of Botsalano, the values from the ERA-Interim reanalysis[40] provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) were used.

The simulation of emissions from the canopy depends on the standard emission potentials, which in this study were based on the measurements reported in the literature.[25,26,28] In the beginning of the simulation the concentrations of all emitted gases and their reaction products were set to zero. For this reason the model run was started one day prior to the studied days, so that the concentrations could accumulate and the chemistry related to these species could stabilise.

From the measured trace gas (O_3, SO_2, NO, NO_x, and CO) concentrations, 15-min averages were calculated and used as input to the model. Data points with concentrations below zero (caused by measurement error) were set to the detection limit of the instruments. For periods of missing data, gap filling was applied (that is, concentrations for missing data points were calculated based on the gradients at the same time on the previous and following days). Every midnight, the particle number concentration in each size bin in the model was set to the measured concentration in order to initialise each study day and to correct the background aerosol concentrations. For the initialisation, the boundary layer was assumed to have a constant concentration and a height of 300 m. Above the boundary layer the concentration was set to 20% of that inside the boundary layer, reflecting the dilution of background aerosol loading in the residual layer.

Results and discussion

Meteorological conditions

The meteorological conditions were evaluated for a 2-week period, including the days selected. The model reproduced the measured temperature well most of the time, but during the warmest hours of midday the model underestimated the measured temperature by 3–5 °C on two out of three days simulated (Figure 2). Almost as common, but less extreme, was the overestimation during the night, which did not exceed 2.6 °C. The model provided an average temperature of the area – both in and out of shade – but the measurements were taken at one location that was not shaded by the canopy. If this approach explained the underestimated daytime temperatures, the difference would not be present above the canopy. The temperatures produced by the model at 10 m were the same as those at 4 m, except for some nights with strong vertical gradients. Even if there were no measurements made above the canopy, it is unlikely that any known physical process caused by the sparse vegetation would have had such a strong effect that the model could not recognise. Therefore the temperatures underestimated by the model are not merely a consequence of inclusion of the model values representing shaded areas. Furthermore, as seen from the middle panel in Figure 2, the simulated absolute humidity was underestimated for most of the studied period. These two issues might be connected, as an underestimation of temperature leads to an underestimation of evaporation, which results in a too low concentration of water vapour in the air. One significant factor affecting the temperature near the surface is the boundary values given for the upper boundary at 3000 m. We investigated the effects of errors in the upper boundary values by altering the upper boundary temperature and found that errors in this temperature can only partly explain the underestimated surface temperature. Another reason for the underestimated midday temperature might be errors in the albedo, the ground heat capacity, or the heat flux from the ground, none of which have been measured in the area and can thus not be validated. These effects could also explain the overestimated temperatures modelled for some nights. All of the explanations mentioned, except the upper boundary values, play a role in the energy balance, which, according to Grote et al.[3], is poorly understood in savannah ecosystems.

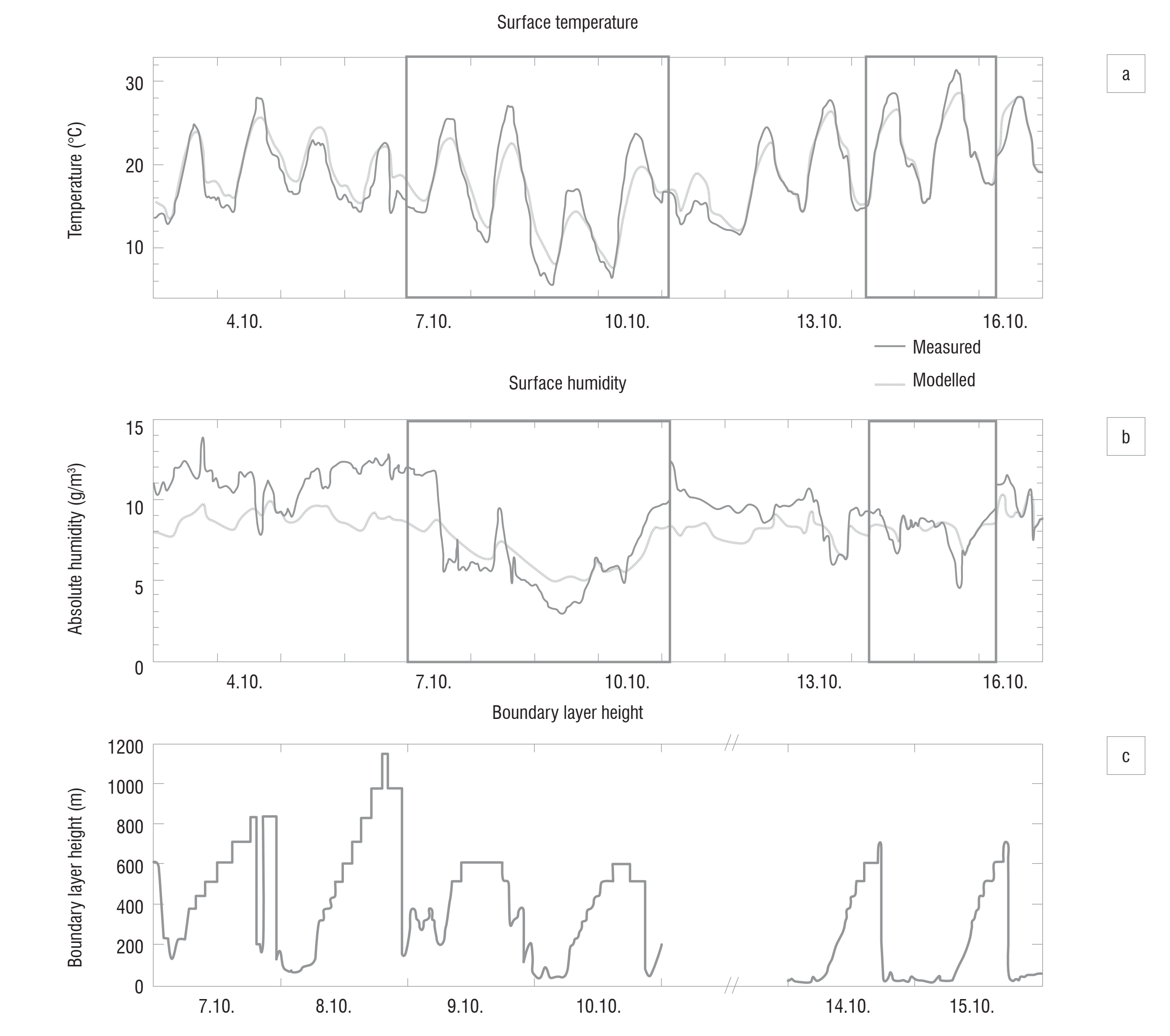


Figure 2: Modelled and measured (a) temperature and (b) absolute humidity during 3–16 October 2007. Modelled values shown are at the model layer with height 4.0 m, the layer closest to the measurement height 3.7 m. Boxes show the days chosen for the study of particles. The modelled temperature does not show as a strong diurnal cycle as measured temperature, underestimating at midday and overestimating at night-time. (c) Simulated boundary layer height for the selected days.

For the vertical mixing of particles, the amount of turbulence and the height of the boundary layer are important. The atmospheric boundary layer is well mixed, but transportation from the boundary layer to the free atmosphere is weak. Figure 2 shows the daily evolution of the atmospheric boundary layer. The night-time boundary layer, in particular, was shallow, which was caused by strong stable stratification typical of the region. The sudden downward peak during the late hours on 7 October was caused by numerical instability in the model[41]. Lauros et al.[15] compared the vertical profiles of the model to measurements made by radiosonde soundings, and stated that the model suggested a too high level of stability and too weak mixing. As the sub-continental conditions in Botsalano are very stable, we cannot make the same conclusion without comparison with measurements. The night-time boundary layer on the last 2 days was very shallow (below 100 m). Turbulent kinetic energy indicates the effectiveness of turbulent mixing. During the last 3 nights, turbulent kinetic energy was approximately zero, indicating that, based on the model, no vertical mixing of particles or gases took place. This scenario leads to high concentrations near ground level for compounds with sources inside the canopy. We therefore need to consider these high levels of stability during the last 2 nights when evaluating the gas and particle concentrations.

Gas phase concentration

For particle formation and growth, a good estimate of the sulphuric acid concentration is essential. Sulphuric acid is traditionally thought to be produced by $^{\bullet}$OH oxidation of SO_2. Therefore it is important to consider these compounds when evaluating the model's ability to predict the sulphuric acid concentration. Modelled $^{\bullet}$OH concentration (a), measured SO_2 concentration (b) and modelled sulphuric acid concentration together with an estimate of the sulphuric acid concentration based on the sulphuric acid proxy (c) are presented in Figure 3. The sulphuric acid proxy was calculated as

$$[H_2SO_4]_{proxy} = k_3 \; \frac{SO_2 \text{ x } Glob}{CS} \qquad \text{Equation 3}$$

where $k_3 = 1.4 \text{ x } 10^{-7} \text{ x Glob}^{-0.70}$, Glob is the global radiation, and CS is the condensation sink as given by Petäjä et al.[42] This method was used by Vakkari and co-workers[18] to estimate the measured sulphuric acid concentrations in Botsalano, and is shown here to provide comparison with their work. As expected, $^{\bullet}$OH showed a clear diurnal cycle, which also reflected the sulphuric acid concentration pattern. Large fluctuations

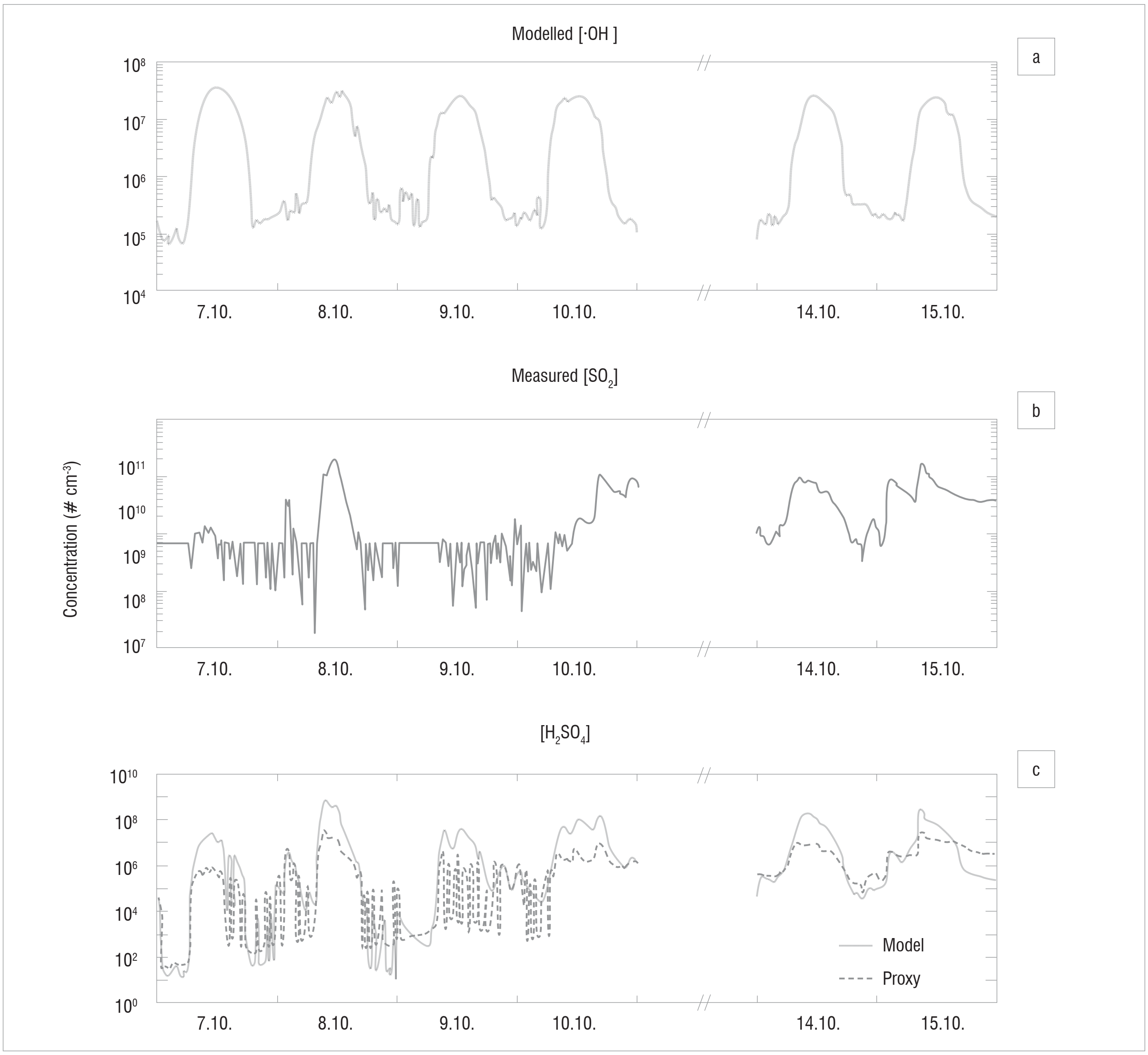


Figure 3: (a) Modelled ·OH concentration, (b) measured SO_2 concentration (gap filling applied) and (c) sulphuric acid concentration according to simulations and calculated by a proxy based on the work by Petäjä et al.[40]

of the sulphuric acid proxy were especially pronounced on 9 October as a result of relatively large variations in low SO_2 concentrations. The simulated sulphuric acid concentration was higher (in the order of 10^8 cm^{-3}) than measurements from other sites[42]; compared with the sulphuric acid proxy, we generally seem to overestimate the concentration. Occasional higher concentrations of sulphuric acid seem to be related to the high concentrations of SO_2 caused by advection. However, no sulphuric acid concentration measurements exist at this station or anywhere close by and our model simulations show good agreement between measured and modelled sulphuric acid concentration for another station.[14] We made sensitivity runs with doubled emission rates for the organic compounds (similar to Boy et al.[43]), which decreased the •OH concentration and thereby decreased the sulphuric acid concentration. However, we did not observe a significant decrease in the sulphuric acid concentration and because we do not have any •OH concentration measurements to compare with, we chose to use the original modelled sulphuric acid concentrations for simulations of aerosol formation and growth in the next section.

For completion, we also present the measured concentrations of CO (a), NO and NO_2 (b) in Figure 4. Increased concentrations of CO (and SO_2) are usually an indication for pollution events. In Botsalano, this pollution is observed especially when the wind blows from a southeasterly direction. Such peaks are seen on 8 October and 10 October, with a twofold higher CO concentration and 10–20-fold higher SO_2 concentrations compared with the times in-between. During 14 and 15 October, there seem to be different air masses present because the trace gas concentrations, as well as the particle number distribution, clearly vary (see 'Particles'). On these days, the wind blew from all directions, not only a southeasterly direction.

Other important vapours participating in the growth of the particles are low volatile organic compounds. Only limited VOC measurements were available during four of the selected days, and the concentrations were mostly below the detection limit. Measured and modelled concentrations of isoprene and the sum of monoterpenes (α-pinene, camphene, sabinene, β-pinene, Δ³-carene and limonene) are visualised in Figure 5. The measured values (dots and crosses) are from a 5.7-m sampling height and in the figure are shown at the middle of the 2-h sampling period. The emission rate of isoprene was highest during daytime because the emission of isoprene is driven by light- and temperature-dependent enzymatic synthesis. Shortly after the emission starts to increase, the boundary layer begins to develop as a result of radiative heating of the surface. Emitted gases are then mixed in a larger volume of air, which lowers the gas concentration. The maximum isoprene concentrations are therefore found in the morning and in the evening. In the case of the monoterpenes, the mixing conditions dominate over the emission rate, leading to higher concentrations during the night than during the day; this pattern has been observed at other sites[44]. The stable nights during 14–15 October resulted in especially high concentrations for both isoprene and the monoterpenes. The measured concentrations of isoprene are one order of magnitude lower and the measured monoterpene concentrations are up to two orders of magnitude lower than what the model predicted. In part, this may be a result of differences in the emission factors or the plant species composition data between the regional model and the Botsalano site. The sum of the measured monoterpenes is calculated based on samples of individual monoterpenes. However, the low observed values seem to be because of the lack of ozone removal in the sampled air. Very recent data by Jaars[45] for the Welgegund measurement site in South Africa that lies close to the ecotone of the grassland and savannah biomes, indicated that monthly median isoprene and monoterpene concentrations vary between 2.4 x 10^8 to 1.25 x 10^9 cm^{-3} and 2.4 x 10^8 to 3.1 x 10^9 cm^{-3}, respectively. These values were obtained by utilising ozone removal prior to sampling, as suggested by Héllen et al.[46], and are much closer to

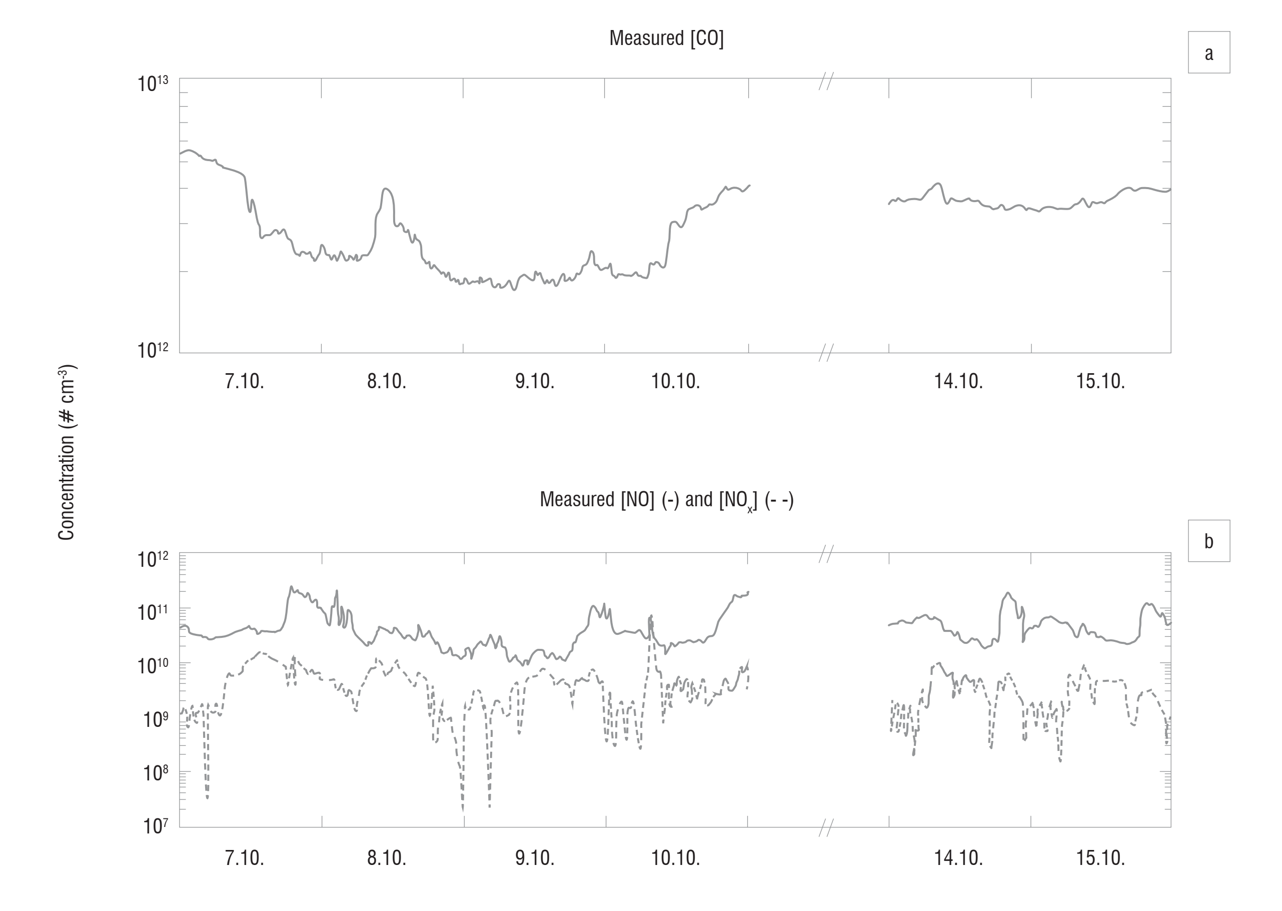


Figure 4: Measured (a) CO, (b) NO (–) and NO_x (--) concentrations in October 2007 (gap filling applied when necessary).

the modelled values than the measured values reported in this paper. Because of the level of uncertainty in the measurements, we therefore cannot make any solid conclusions on the reliability of the model based on this comparison.

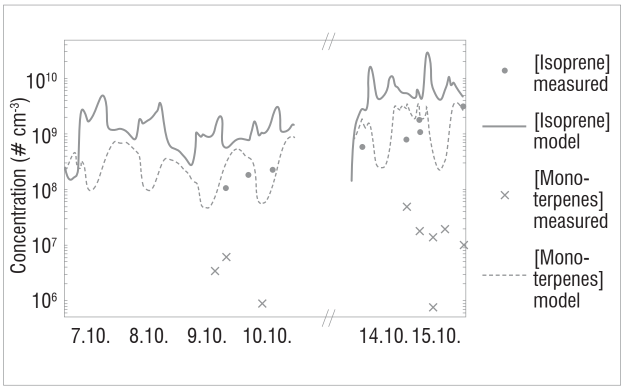


Figure 5: Modelled (–) and measured (●) isoprene concentration, and modelled (--) and measured (x) monoterpenes concentration during 7–10 and 14–15 October 2007. The monoterpenes concentration is the sum of α-pinene, camphene, sabinene, β-pinene, carene, limonene, myrcene and ocimene concentrations, both for modelled and measured values. Measured values are shown at the middle of the 2-h sampling period and are from a height of 5.7 m and the simulated values are from a model height of 5.8 m.

The modelled concentrations of the first stable reaction products of α-pinene, β-pinene and isoprene with $^{\bullet}$OH, O_3 and $NO_3^{\bullet}$, which in the simulations are forming and growing the particles, are shown in Figure 6. The low level of mixing, together with the higher concentrations of monoterpenes (Figure 5), is reflected in the high concentrations of the reaction products during the last 2 days. The concentrations of the reaction products generated from monoterpene and isoprene reactions with $^{\bullet}$OH (dotted line) and O_3 (dashed line), showed similar behaviour to those previously described for isoprene – a high formation rate during the day but, with effective mixing, the concentration decreased or did not continue to increase. The reaction products generated from monoterpene and isoprene reactions with $NO_3^{\bullet}$ (solid line) showed diurnal cycles driven by the vertical mixing. This process caused the concentration to be higher if the surface layer was very stable, which was the case during the last 2 days (14–15 October), especially during the night. The stable stratification also leads to strong vertical gradients, as the VOCs are emitted in the canopy.

Particles

The ability of the two selected nucleation mechanisms – kinetic and organic-induced nucleation – to reproduce the observed number concentration of particles was evaluated by holding the nucleation coefficients K (in Equation 1) and P (in Equation 2) constant for all simulated days. The modelled events, together with the observations, are shown in Figure 7. To allow more detailed comparison, the formation rate of 10-nm particles (J_{10}) and the growth rate for 10–20-nm particles were calculated from observations and both simulations for each day (Table 2).

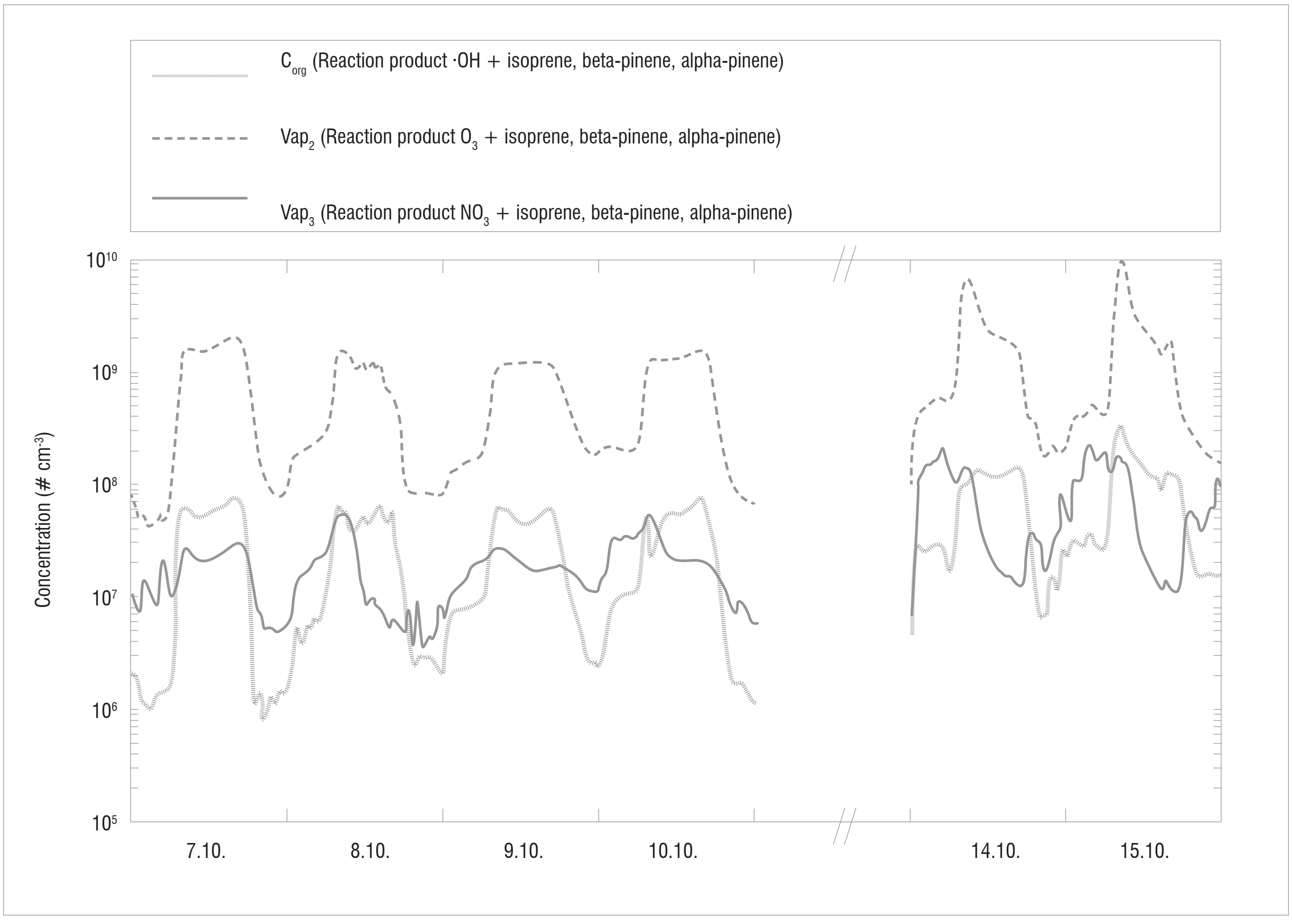


Figure 6: Modelled reaction products of volatile organic compounds (isoprene, α-pinene and β-pinene) with ·OH, O_3 and NO_3 during 7–10 and 14–15 October 2007. The first mentioned participates in new particle formation, and all in particle growth in the simulation.

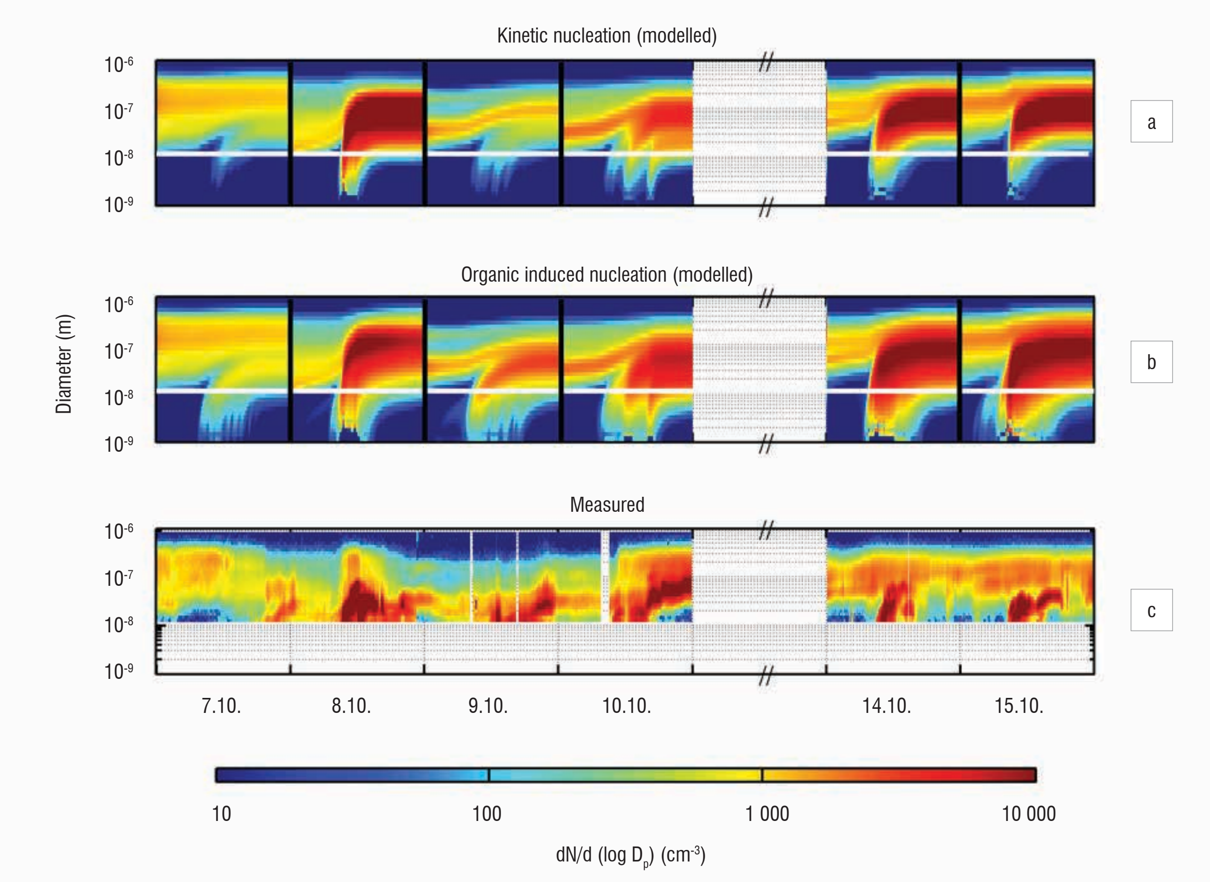


Figure 7: Modelled (a) kinetic nucleation and (b) organic-induced nucleation and (c) measured particle number size distribution for 7–10 and 14–15 October 2007. In (a) and (b) the white line shows the detection limit of the instrument at 10 nm, and the black lines model initialisation in the beginning of each day. The simulated concentrations are from a model height of 4.0 m.

Table 2: Growth rates (GR), particle formation rates (J_{10}) and sulphuric acid concentrations calculated from measurements, the two different simulations and for the sulphuric acid proxy. J_{10}-values shown are the means over 4 h after the start of the nucleation event. Sulphuric acid concentrations from both the simulation and proxy are given for the time defined by simulation (kinetic nucleation). For completeness, the estimated starting times of the events, defined from J_{10} time series, are shown.

Date		7 Oct	8 Oct	9 Oct	10 Oct	14 Oct	15 Oct
Event starting time (HH:MM)	Observed	17:52	08:13	10:43	08:15	08:09	07:29
	Model (kinetic)	09:30	08:48	08:45	08:14	07:21	08:09
	Model (organic)	08:24	08:18	07:45	08:12	07:12	08:00
Sulphuric acid concentration (# cm^{-3})	Proxy	6.0E+5	1.8E+7	2.2E+6	3.0E+6	9.4E+6	2.1E+7
	Model	2.0E+7	3.1E+8	3.7E+7	5.0E+7	1.7E+8	1.4E+8
GR 10-20nm (nm/h)	Observed	5.0	14	2.1	7.6	19	9.2
	Model (kinetic)	4.6	28	3.6	9.9	12	17
	Model (organic)	3.5	33	5.3	9.4	16	14
J10 (cm^3/s)	Observed	0.074	2.5	0.091	0.18	0.93	0.72
	Model (kinetic)	0.0071	1.2	0.0085	0.027	0.20	0.24
	Model (organic)	0.024	0.84	0.079	0.096	0.84	1.2

On the first day (7 October 2007) the measurements showed an increasing concentration of nucleation mode particles at approximately 18:00 (Figure 7). A sudden decrease in the concentration of Aitken mode particles in the middle of the day can also be seen, probably as a result of a different type of air mass being advected to the site. The formation rate of the particles was the lowest observed in the studied period (Table 2), which, together with the high concentrations of Aitken mode particles for the first half of the day, could have lead to the late beginning of the observed event: in the morning the few newly formed particles were coagulating on bigger particles (which had a high concentration), and when the background aerosol loading decreased, it allowed for the newly formed particles to start growing and reach the size at which they could be observed (10 nm), later than on the other days. This change of air mass cannot be accounted for in the model, and therefore this effect is not seen in the results of the simulations. Both nucleation approaches also led to particle formation in the simulations, only much earlier on the day (Figure 7, Table 2). The values for growth rate agree well with those observed, but kinetic nucleation underestimated J_{10} by an order of magnitude.

On 8 October, strong advection of SO_2 was observed within the measurements, leading to a high sulphuric acid concentration (Figure 3), and thereby causing a high rate of particle formation (Table 2). The simulations also show relatively high formation rates at 10 nm, although lower than observed. The growth rate was overestimated by the model. The results on 9 October are similar to those on 7 October – low sulphuric acid and a good agreement with both J_{10} and growth rate, except for the kinetic nucleation approach that underestimated the formation rate. The daytime sulphuric acid concentration was the lowest during these 2 days, and as kinetic nucleation assumes particle formation from sulphuric acid, it is to be expected that this approach is sensitive to changes in sulphuric acid concentration.

Of all the simulated days, 10 October showed the best agreement with the observations when considering J_{10}, growth rate and the timing of the start of the event. During the last 2 days (14–15 October), both particle formation mechanisms gave similar results. The sulphuric acid concentration during these days was slightly elevated, and the modelled monoterpene and isoprene concentrations were the highest of the simulated period. In the evening of 14 October there was some rain, but at this point the observed concentrations were already lower. Throughout the last 2 days, the SO_2 and CO concentrations (Figure 3 and Figure 4) were generally higher than during the first days, indicating that the air was more polluted at this time than during the first 4 days.

The difference in sulphuric acid concentration among the days leads to a greater difference in the simulated particle concentrations (Figure 7), indicating that sulphuric acid alone cannot be responsible for new particle formation, and the dependency on its concentration was weaker than assumed in the model. The kinetic nucleation approach particularly depends too strongly on sulphuric acid concentration (Table 2). This finding is in good agreement with observations at other sites.[15] The organic nucleation approach gave J_{10} with a better agreement with the observed for all days except 8 October. When changing the approach for particle formation, the mechanism to grow particles in the model was the same. The difference in the growth rates between the simulations can be explained by the changes in coagulation and condensation when the number of particles in lower size bins changed. The high growth rate on 8 October was probably caused by the very high sulphuric acid concentration, and the above mean growth rate on 14 and 15 October by high concentrations of both sulphuric acid and the condensing organic vapours (Figure 6).

In our simulations we assigned the coefficient K in Equation 1 a value of 8×10^{-18} cm^3/s, which is lower than the values reported for other sites.[15,37] For the organic-induced nucleation, P in Equation 2 was set to 1×10^{-5} cm^3/s. To compare with other results, we calculated the coefficient

$$K_{org} = P \cdot \upsilon \qquad \text{Equation 4}$$

which varied from 3.8×10^{-13} cm^3/s to 9.4×10^{-13} cm^3/s with a mean value of 5.4×10^{-13} cm^3/s. These values are at the low end of the distribution reported by Paasonen and co-workers[38], but not out of range. Compared with other sites, the sulphuric acid concentration in Botsalano is higher, and the K and K_{org} have lower values, resulting in a nucleation rate that is similar to that reported from other sites.[38] The air mass history analysis by Vakkari et al.[18] for the same site indicated that the highest formation and growth rates were related to the highest VOC emissions, but not to the highest estimated sulphuric acid concentrations, which suggests that the dependency on sulphuric acid is not as strong as for some other sites.

The use of a column model is based on the assumption that the area is homogeneous enough to exclude transport of particles by horizontal advection. Figure 7 shows that this requirement was not always fulfilled at Botsalano. Many of the differences between the observed and simulated particle concentrations can be explained by an air mass with different aerosol properties being advected at the site. This scenario is evident on 8 October and on the last 2 days studied (Figure 7), when the number concentration changed so rapidly that it cannot be explained by aerosol dynamic processes. The concentrations of Aitken mode particles are affected the most, because it takes time for the particles to grow. If the newly formed particles are not present at the site after some time, the modelled concentration is highly overestimated. Figure 8 shows modelled 1-h averages of number concentrations against measured concentrations for 10–30-nm and 30–100-nm particles. Results from simulations with organic-induced nucleation are shown with crosses (x) and kinetic nucleation with circles (○). The modelled number concentration of 30–100 nm was highly overestimated repeatedly by both nucleation mechanisms (Figure 8a). The modelled 10–30 nm number concentration by kinetic nucleation approach, on the other hand, shows a tendency to underestimate the concentration (Figure 8b). The organic-induced nucleation approach did not result in a similar tendency. The mechanism to grow particles was the same in all simulations, only the path to particle formation was changed. Figure 8 shows that the two nucleation approaches resulted in different 10–30-nm number concentrations, whereas the 30–100-nm concentrations were more alike. The similarity in the pattern for the 30–100-nm particle concentration, despite the difference in the smaller size range, indicates that advection indeed plays a key role and that the difference between the model and the measurements cannot be a consequence of error in the particle growth scheme alone. Figure 7 suggests that the heterogeneity of the region has the greatest effect during the afternoons and evenings. While advection was the main cause of the difference between observed and modelled particle concentrations, it is also likely that some variability was caused by the inaccuracy of the parameterisation for growth of the particles.

Conclusions

MALTE simulations of the dynamics of temperature and humidity in the savannah environment need improvement; the diurnal cycle of temperature was not strong enough. However, considering that exchange of water and energy between the atmosphere and the savannah ecosystem is not well understood,[3] and that uncertainties originate from the boundary values, the model performance was satisfactory. The simulations showed a high level of stability and a shallow boundary layer during night-time, with significant effects on the concentrations of gases with sources within the canopy.

The sulphuric acid concentrations in the model were relatively high, but can be explained by the high level of $^{\bullet}$OH and SO_2 concentrations. The SO_2 concentration depended strongly on the origin of the air mass present at the site, and sudden changes took place repeatedly. The measured monoterpene and isoprene concentrations differed from the measurements by one to two orders of magnitude. The lack of ozone removal in the sampling was probably the main cause of this difference. Although the problems with the values can be discussed, the simulated diurnal variation was reasonable. The reaction products of isoprene, α-pinene and β-pinene with $^{\bullet}$OH, O_3 and $NO_3^{\bullet}$ provided possibly the highest uncertainties for our simulations on the formation and growth of particles.

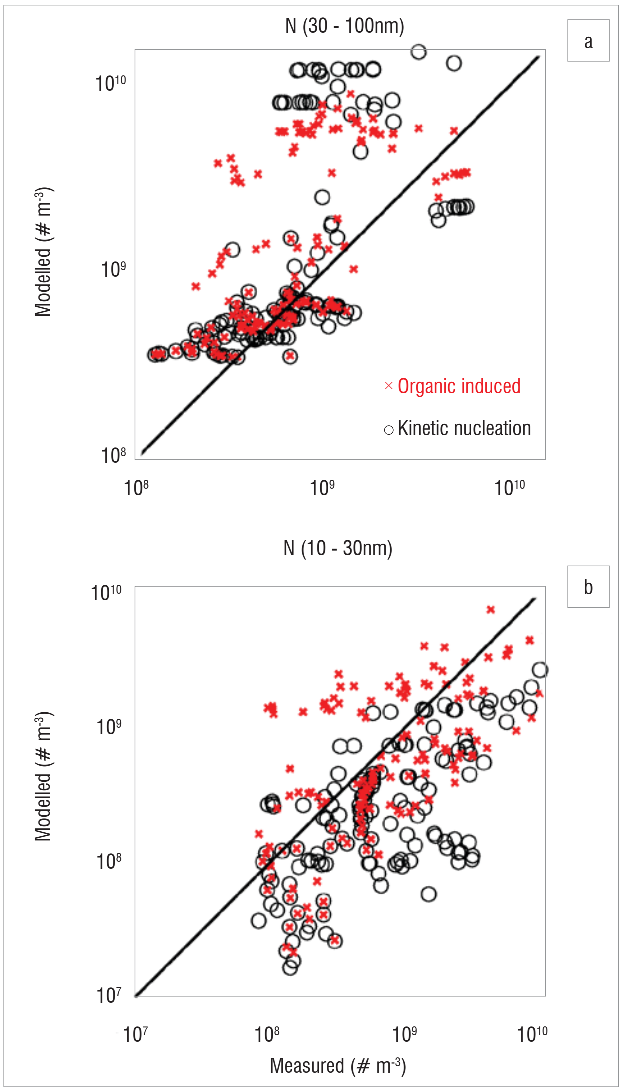


Figure 8: Modelled 1-h averages of number concentrations against measured for (a) 30–100-nm and (b) 10–30-nm particles, including all studied days. Results from simulations with organic-induced nucleation are shown with crosses (x) and kinetic nucleation with circles (○), and are from a model height of 4.0 m.

One of the most important causes of differences between the modelled and measured particle number concentrations was advection. The site was quite heterogeneous with a high frequency of horizontal advection of either clean or polluted air masses, which make the use of a column model challenging. Organic-induced nucleation was better in reproducing observed particle formation and growth rates on all days except one (Table 2), as well as number concentration for 10–30-nm particles (Figure 8). The coefficients describing the nucleation processes, K and P, were low compared with studies of other sites. Choosing a constant value for these coefficients proved not to be a valid approximation, and other variables for describing the nucleation process should be introduced.

The next step in developing this work further is to use the Model to Simulate the concentrations of Organic vapours and Sulphuric Acid (SOSA)[47] to model particle formation at Welgegund, a station approximately 200 km southeast of Botsalano. The station has the same measurements as Botsalano, but also some additional observations that can be used for driving and validating the model conditions, such as sensible heat flux and soil humidity and temperature. A much more comprehensive VOC measurement campaign was also conducted at this site, during which time ozone removal was applied. SOSA is very similar to MALTE – both include the same modules (or versions of them) – but it is written in parallel so the computational time will not limit the work in the same way as it did in this study. Running longer time periods will give a better understanding on different conditions and the possibility to study, for example, seasonal variability.

Possible climate change and air quality related impacts in southern Africa could be significant. Therefore, continued modelling developments as presented in this paper to better understand and predict new particle formation for this unique environment are critical in mitigating possible negative impacts.

Acknowledgements

This work was supported by the Finnish Center of Excellence in Physics, Chemistry, Biology and Meteorology of Atmospheric Composition and Climate Change, Helsinki University Centre for Environment (HENVI). The CSC–IT Center for Science Ltd is gratefully acknowledged for computational resources.

Authors' contributions

R.G. performed the simulations, analysed the data and wrote most of the manuscript. D.M. provided the chemistry module and wrote the parts concerning chemistry. L.L. was in charge of the project, did the actual VOC sampling in Botsalano, cleaned all the data together with V.V. and contributed to the writing of the manuscript. V.V. also contributed to the writing of the manuscript. J.P.B. and P.G.v.Z. aided in contextualising the results within a South African context and made some conceptual contributions. H.H. arranged the VOC sampling equipment and VOC analysis and contributed to the writing of the manuscript. A.G. developed the model for emissions from the canopy and contributed to the writing of the manuscript. J.J.P. made conceptual contributions. M.B. supervised the modelling work and supported the writing process.

References

1. Scholes RJ, Hall DO. The carbon budget of tropical savannas, woodlands and grasslands. In: Breymeyer AI, Hall DO, Melillo JM, Agren GI, editors. Global change: Effects on coniferous forests and grasslands. Chichester: John Wiley and Sons; 1996. p. 69–100.

2. Land use, land-use change, and forestry. Intergovernmental Panel on Climate Change special report. Cambridge: Cambridge University Press; 2000.

3. Grote R, Lehmann E, Brümmer C, Brüggemann N, Szarzynski J, Kunstmann H. Modelling and observation of biosphere-atmosphere interactions in natural savannah in Burkina Faso, West Africa. Phys Chem Earth. 2009;34:251–260. http://dx.doi.org/10.1016/j.pce.2008.05.003

4. Spracklen DV, Carslaw KS, Kulmala M, Kerminen V-M, Sihto S-L, Riipinen I, et al. Contribution of particle formation to global cloud condensation nuclei concentrations. Geophys Res Lett. 2008;35,L06808.

5. Merikanto J, Spracklen DV, Mann GW, Pickering SJ, Carslaw KS. Impact of nucleation on global CCN. Atmos Chem Phys. 2009;9:8601–8616. http://dx.doi.org/10.5194/acp-9-8601-2009

6. Kerminen V-M, Paramonov M, Anttila T, Riipinen I, Fountoukis C, Korhonen H, et al. Cloud condensation nuclei production associated with atmospheric nucleation: A synthesis based on existing literature and new results. Atmos Chem Phys. 2012;12:12037–12059. http://dx.doi.org/10.5194/acp-12-12037-2012

7. Yu F, Luo G. Simulation of particle size distribution with a global aerosol model: Contribution of nucleation to aerosol and CCN number concentrations. Atmos Chem Phys. 2009;9:7691–7710. http://dx.doi.org/10.5194/acp-11-1083-2011

8. Yu F. A secondary organic aerosol formation model considering successive oxidation aging and kinetic condensation of organic compounds: global scale implications. Atmos Chem Phys. 2011;11:1083–1099. http://dx.doi.org/10.5194/acp-11-1083-2011

9. UNEP and C4. Asian Brown Cloud: Climate and other environmental impacts. Nairobi: UNEP; 2002.

10. Laakso L, Laakso H, Aalto PP, Keronen P, Petäjä T, Nieminen T, et al. Basic characteristics of atmospheric particles, trace gases and meteorology in a relatively clean southern African savannah environment. Atmos Chem Phys. 2008;8:4823–4839. http://dx.doi.org/10.5194/acp-12-4449-2012

11. Mann GW, Carslaw KS, Ridley DA, Spracklen DV, Pringle KJ, Merikanto J, et al. Intercomparison of modal and sectional aerosol microphysics representations within the same 3-D global chemical transport model. Atmos Chem Phys. 2012;12:4449–4476. http://dx.doi.org/10.5194/acp-12-4449-2012

12. Laakso L, Merikanto J, Vakkari V, Laakso H, Kulmala M, Molefe M, et al. Boundary layer nucleation as a source of new CCN in savannah environment. Atmos Chem Phys Discuss. 2012;12:8503–8531. http://dx.doi.org/10.5194/acpd-12-8503-2012

13. Petäjä T, Vakkari V, Pohja T, Nieminen T, Laakso H, Aalto PP, et al. Transportable aerosol characterization trailer with trace gas chemistry: Design, instruments and verification. Aerosol Air Qual Res. 2013;13:421–435.

14. Boy M, Hellmuth O, Korhonen H, Nilsson ED, ReVelle D, Turnipseed A, et al. MALTE – Model to predict new aerosol formation in the lower troposphere. Atmos Chem Phys Discuss. 2006;6:3465–3512. http://dx.doi.org/10.5194/acpd-6-3465-2006

15. Lauros J, Sogachev A, Smolander S, Vuollekoski H, Sihto S-L, Laakso L, et al. Particle concentration and flux dynamics in the atmospheric boundary layer as the indicator of formation mechanism. Atmos Chem Phys. 2011;11:5591–5601. http://dx.doi.org/10.5194/acp-11-5591-2011

16. Venter AD, Vakkari V, Beukes JP, Van Zyl PG, Laakso H, Mabaso D, et al. An air quality assessment in the industrialised western Bushveld Igneous Complex, South Africa. S Afr J Sci. 2012;108(9/10), Art. #1059, 10 pages. http://dx.doi.org/10.4102/sajs.v108i9/10.1059

17. Lourens A, Butler T, Beukes JP, Van Zyl PG, Beirle S, Wagner TK, et al. Re-evaluating the NO_2 hotspot over the South African Highveld. S Afr J Sci. 2012;108(11/12), Art. #1146, 6 pages. http://dx.doi.org/10.4102/sajs.v108i11/12.1146

18. Vakkari V, Laakso H, Kulmala M, Laaksonen A, Mabaso D, Molefe M, et al. New particle formation events in semi-clean South African savannah. Atmos Chem Phys. 2011;11(7):3333–3346. http://dx.doi.org/10.5194/acp-11-3333-2011

19. Vakkari V, Beukes JP, Laakso H, Mabaso D, Pienaar JJ, Kulmala M, et al. Long-term observations of aerosol size distributions in semi-clean and polluted savannah in South Africa. Atmos Chem Phys. 2013;13:1751–1770. http://dx.doi.org/10.5194/acp-13-1751-2013

20. Tyson PD, Garstang M, Swap R. Large-scale recirculation of air over southern Africa. J Appl Meteorol. 1996;35:2218–2236. http://dx.doi.org/10.1175/1520-0450(1996)035<2218:LSROAO>2.0.CO;2

21. Swap RJ, Annegarn HJ, Suttles JT, King MD, Platnick S, Privette JL, et al. Africa burning: A thematic analysis of the Southern African Regional Science Initiative (SAFARI 2000). J Geophys Res. 2003;108:8465. http://dx.doi.org/10.1029/2003JD003747

22. Sogachev A, Menzhulin G, Heimann M, Lloyd J. A simple three-dimensional canopy – planetary boundary layer simulation model for scalar concentrations and fluxes. Tellus. 2002;54B:784–819.

23. Sogachev A, Panferov O. Modification of two-equation models to account for plant drag. Bound-Lay Meteorol. 2006;121:229–266. http://dx.doi.org/10.1007/s10546-006-9073-5

24. Sogachev A. A note on two-equation closure modelling of canopy flow. Bound-Lay Meteorol. 2009;130:423–435. http://dx.doi.org/10.1007/s10546-008-9346-2

25. Guenther A, Otter L, Zimmermann P, Greenberg J, Scholes R, Scholes M. Biogenic hydrocarbon emissions from southern African savannas. J Geophys Res. 1996;101:25859–25865. http://dx.doi.org/10.1029/96JD02597

26. Otter L, Guenther A, Wiedinmyer C, Fleming G, Harley P, Greenberg J. Spatial and temporal variations in biogenic volatile organic compound emissions for Africa south of the equator. J Geophys Res. 2003;108(D13):8505.

27. Greenberg J, Guenther A, Harley P. Eddy flux and leaf-level measurements of biogenic VOC emissions from Mopane woodland of Botswana. J Geophys Res. 2003;108(D13):8466.

28. Harley P, Otter L, Guenther A, Greenberg J. Micrometeorological and leaf-level measurements of isoprene emissions from a southern African savanna. J Geophys Res. 2003;108(D13):8468.

29. Guenther A, Zimmerman P, Wildermuth M. Natural volatile organic compound emission rate estimates for United States woodland landscapes. Atmos Environ. 1994;28(6):1197–1210. http://dx.doi.org/10.1016/1352-2310(94)90297-6

30. Guenther A, Karl T, Harley P, Wiedinmyer C, Palmer PI, Geron C. Estimates of global terrestrial isoprene emissions using MEGAN (Model of Emissions of Gases and Aerosols from Nature). Atmos Chem Phys. 2006;6:3181–3210. http://dx.doi.org/10.5194/acp-6-3181-2006

31. Guenther AB, Jiang X, Heald CL, Sakulyanontvittaya T, Duhl T, Emmons LK, et al. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): An extended and updated framework for modeling biogenic emissions. Geosci Model Dev. 2012;5(6):1471–1492. http://dx.doi.org/10.5194/gmd-5-1471-2012

32. Damian V, Sandu A, Damian M, Potra F, Carmichael GR. The Kinetic PreProcessor KPP – A software environment for solving chemical kinetics. Comput Chem Eng. 2002;26:1567–1579. http://dx.doi.org/10.1016/S0098-1354(02)00128-X

33. Jenkin ME, Saunders SM, Pilling MJ. The tropospheric degradation of volatile organic compounds: A protocol for mechanism development. Atmos Environ. 1997;31:81–104. http://dx.doi.org/10.1016/S1352-2310(96)00105-7

34. Saunders SM, Jenkin ME, Derwent RG, Pilling MJ. Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part A): Tropospheric degradation of non-aromatic volatile organic compounds. Atmos Chem Phys. 2003;3:161–180. http://dx.doi.org/10.5194/acp-3-161-2003

35. Korhonen H, Lehtinen KEJ, Kulmala M. Multicomponent aerosol dynamics model UHMA: Model development and validation. Atmos Chem Phys. 2004;4:757–771. http://dx.doi.org/10.5194/acp-4-757-2004

36. McMurry P, Friedlander S. New particle formation in the presence of an aerosol. Atmos Environ. 1979;13:1635–1651. http://dx.doi.org/10.1016/0004-6981(79)90322-6

37. Kulmala M, Kerminen V-M. On the growth of atmospheric nanoparticles. Atmos Res. 2008;90:132–150. http://dx.doi.org/10.1016/j.atmosres.2008.01.005

38. Paasonen P, Nieminen T, Asmi E, Manninen HE, Petäjä T, Plass-Dülmer P, et al. On the roles of sulphuric acid and low-volatility organic vapours in the initial steps of atmospheric new particle formation. Atmos Chem Phys. 2010;10:11223–11242. http://dx.doi.org/10.5194/acp-10-11223-2010

39. Kulmala M, Kerminen V-M, Anttila T, Laaksonen A, O'Dowd CD. Organic aerosol formation via sulphate cluster activation. J Geophys Res. 2004;4, Art. 04205.

40. Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Q J Roy Meteor Soc. 2011;137:553–597. http://dx.doi.org/10.1002/qj.828

41. Troen IB, Mahrt L. A simple model of the atmospheric boundary layer; sensitivity to surface evaporation. Bound-Lay Meteorol. 1986;37:129–148.

42. Petäjä T, Mauldin III RL, Kosciuch E, McGrath J, Nieminen T, Paasonen P, et al. Sulfuric acid and OH concentrations in a boreal forest site. Atmos Chem Phys. 2009;9:7435–7448. http://dx.doi.org/10.5194/acp-9-7435-2009

43. Boy M, Kulmala M, Ruuskanen TM, Pihlatie M, Reissell A, Aalto PP, et al. Sulphuric acid closure and contribution to nucleation mode particle growth. Atmos Chem Phys. 2005;5:863–878. http://dx.doi.org/10.5194/acp-5-863-2005

44. Mogensen D, Smolander S, Sogachev A, Zhou L, Sinha V, Guenther A, et al. Modelling atmospheric OH-reactivity in a boreal forest ecosystem. Atmos Chem Phys. 2011;11:9709–9719. http://dx.doi.org/10.5194/acp-11-9709-2011

45. Jaars K. Temporal assessment of volatile organic compounds at a site with high atmospheric variability in the North-West Province [MSc thesis]. Potchefstroom: North-West University; 2013.

46. Hellén H, Kuronen P, Hakola H. Heated stainless steel tube for ozone removal in the ambient air measurements of mono- and sesquiterpenes. Atmos Environ. 2012;57:35–40. http://dx.doi.org/10.1016/j.atmosenv.2012.04.019

47. Boy M, Sogachev A, Lauros J, Zhou L, Guenther A, Smolander S. SOSA – A new model to simulate the concentrations of organic vapours and sulphuric acid inside the ABL – Part I: Model description and initial evaluation. Atmos Chem Phys. 2011;11:43–51. http://dx.doi.org/10.5194/acp-11-43-2011

Equilibrium ion exchange studies of Ni^{2+} on homoionic forms of clinoptilolite

AUTHORS:
Joshua Gorimbo[1]
Blessing Taenzana[1]
Kutemba Kapanji[1]
Linda L. Jewell[2]

AFFILIATIONS:
[1]School of Chemical and Metallurgical Engineering, University of the Witwatersrand, Johannesburg, South Africa
[2]Department of Chemical Engineering, University of South Africa, Johannesburg, South Africa

CORRESPONDENCE TO:
Linda Jewell

EMAIL:
jewelll@unisa.ac.za

POSTAL ADDRESS:
Department of Chemical Engineering, University of South Africa, Private Bag X6, Florida 1710, South Africa

DATES:
Received: 26 Sep. 2012
Revised: 22 Oct. 2013
Accepted: 08 Nov. 2013

KEYWORDS:
clinoptilolite; ion exchange; nickel removal; heavy metals; zeolites

HOW TO CITE:
Gorimbo J, Taenzana B, Kapanji K, Jewell LL. Equilibrium ion exchange studies of Ni^{2+} on homoionic forms of clinoptilolite. S Afr J Sci. 2014;110(5/6), Art. #2012-0027, 7 pages. http://dx.doi.org/10.1590/sajs.2014/20120027

A natural zeolite (clinoptilolite) that is mined in KwaZulu-Natal, South Africa, was evaluated for the removal of Ni^{2+} from wastewater. In particular, the effect of zeolite modification on Ni^{2+} removal from synthetic wastewater was investigated. The natural clinoptilolite was pretreated with 2 M metal chlorides for 24 h to yield near homoionic Na^+, K^+ and Ca^{2+} forms. A comparison of the isotherms for the Na^+-Ni^{2+}, K^+-Ni^{2+}, Ca^{2+}-Ni^{2+} and natural-Ni^{2+} systems gave insight into how the displaced ion affects the selectivity of the clinoptilolite for Ni^{2+}. The Na^+, K^+ and natural forms show highly selective convex isotherms whereas the Ca^{2+} form has a concave graph suggesting that the selectivity series is $Ca^{2+} > Ni^{2+} >$ (Na^+, K^+, natural). Thermodynamic properties revealed that the Ni^{2+} sorption capacity increases as the values of the equilibrium constant and Gibbs free energy increase with increasing temperature from 298.15 K to 348.15 K. The enthalpy change was positive for all forms of clinoptilolite; values of 26.00 kJ/mol, 18.72 kJ/mol and 42.05 kJ/mol were obtained for exchange of Ni^{2+} into Na^+, K^+ and Ca^{2+} forms, respectively. The positive changes in enthalpy provide an indication that the sorption reaction is endothermic for Ni(II). The Gibbs free energy values were all negative except for Ca^{2+}-exchanged clinoptilolite at 298.15 K and 308.15 K, for which the values were positive 3.10 kJ/mol and 0.53 kJ/mol, respectively. The entropy values for Ni^{2+} sorption were also positive; values of 0.12 kJ/mol.K, 0.08 kJ/mol.K and 0.14 kJ/mol.K were obtained for the Na^+, K^+ and Ca^{2+} forms, respectively. As expected, the enthalpy obtained from the Van't Hoff plot is dependent not only on the metal ion being adsorbed, but also on the ion being displaced. Pretreatment of the zeolite enhances the removal efficiency provided that monovalent ions are used for the pretreatment. Thus clinoptilolite is an effective low-cost absorbent for the removal of Ni^{2+} from aqueous solutions.

Introduction

Heavy metal pollution is an environmental problem of concern worldwide. The increasing levels of heavy metals in the environment represent a serious threat to human health and ecological systems. Soluble and mobile heavy metal species are non-biodegradable and tend to bioaccumulate in living organisms causing various diseases and disorders. The Ni^{2+} ion is one such heavy metal which is frequently encountered in raw wastewater streams from industries, such as those for the manufacture of magnetic tape, electroplating, jewellery, welding rods, dental procedures and pigments; Ni^{2+} is also used as a catalyst in oil hydrogenation.[1] Several methods – such as ion exchange, solvent extraction, reverse osmosis, precipitation and adsorption – have been proposed for the treatment of wastewater contaminated with heavy metals.[2-4] Among several chemical and physical methods, the adsorption of heavy metals onto zeolites has been found to be superior to other techniques because of the capability of the zeolite to remove several cations simultaneously from an aqueous solution through ion exchange.[3-7] The history of ion exchange on zeolites has been reported by several authors.[3,8-10] One of the important properties of zeolites is that they show selectivity in adsorption, i.e. they possess different affinities for different ions.[11,12]

The effect of converting the zeolite initially to a homoionic form for the removal of heavy metal ions from wastewater has been studied by many authors.[13-19] These authors pointed out that zeolite in homoionic forms exhibits a significantly increased ability to remove heavy metals from wastewater. NaCl is most often used as the pretreatment agent. Prior to any ion-exchange application, most exchangeable ions from the structure of the material are removed by pretreatment and replaced by more easily removable ones. Pretreatment of natural zeolites with, for example, acids, bases and surfactants, is also used to improve their ion-exchange capacity. Most pretreatment operations increase the content of a single cation, called a homoionic form.

The evaluation of the ion-exchange properties of zeolites is based on equilibrium data for a particular exchange reaction. On the basis of these data, the main thermodynamic properties, such as the equilibrium constant (K_{eq}) and Gibbs free energy ΔG°, can be computed using a suitable model. The use of a reliable model for the exchange process is particularly important when one needs to predict the ion-exchange behaviour of the zeolites for varying compositions of the aqueous phase based on experimental data. Much research has been done with regard to modelling the equilibria of ion exchange in zeolites.[20-24]

Fitting of adsorption isotherm equations to experimental data is often an important aspect of data analysis. In most previous studies, Ni^{2+} adsorption with clinoptilolite was generally examined with the Langmuir, Dubinin–Radushkevitch and Freundlich isotherms.[25-27]

The Langmuir equation assumes that the adsorbed species forms a monolayer. But monolayer formation is possible only for a dilute solution. Under high concentration conditions the assumption is no longer valid as adsorbates accumulate to form multiple layers. The Langmuir equation also assumes that adsorbed molecules do not interact with each other laterally. This is impossible as weak forces of attraction exist even between molecules of the same type. Another assumption is that all the sites on the solid surface are equivalent in size and shape and have equivalent affinity for adsorbate molecules, i.e. the surface of the solid is homogeneous. But real solid surfaces are heterogeneous. Because clinoptilolite minerals have high surface irregularities, the adsorption models (Langmuir and Freundlich) should not be used to explain the adsorption equilibrium phenomenon.[28]

Zeolites have been found to be highly efficient in removing heavy metals from wastewater and the costs involved are still significantly below competing technologies. Although major breakthroughs have been made towards the use of zeolites in environmental remediation, most researchers have focused on heavy metal ions such as Pb^{2+}, Cu^{2+} and Co^{2+} rather than on Ni^{2+} removal. Two interesting observations with respect to Ni^{2+} adsorption on clinoptilolite have been made by Sprynskyy et al.[29] After its initial fast uptake from mixed metal feed concentrations, the second phase of adsorption is characterised by desorption and is referred to as an 'inversion phenomenon'. It is thought that this phenomenon is caused by the counter-diffusion of displaced extra-lattice cations from the deeper layers of the zeolite. As high concentrations of these counter-ions diffuse out of the zeolite pores, the Ni^{2+} is displaced and readsorbed. This displacement indicates that Ni^{2+} adsorbs in the mesopores, rather than in the zeolite channels. It was observed by Sprynskyy et al.[29] that when metal ions like Pb^{2+} are adsorbed on clinoptilolite, there is usually only a slight difference between the loading capacities of these metal ions when single component feeds are compared to mixed feeds. This finding implies that metal ions are sorbed on specific sites. However, with Ni^{2+}, adsorption significantly decreases in mixed feeds as a consequence of competition with other metals. Sprynskyy et al.[29] concluded that the adsorption of Ni^{2+} ions is not site specific. Moreover, the zeolite selectivity for Ni^{2+} is generally low, hence improvements in Ni^{2+} selectivity are important for industrial applications, especially when implementation of more stringent standards for discharge of heavy metals into receiving environments are taken into consideration. It has been proven that converting zeolite to homoionic forms improves selectivity, but we are not aware of any study which compares different homoionic forms for the removal of Ni^{2+} from aqueous solutions. In this study, physical (i.e. pore diameter and volume) and thermodynamic differences between these forms have been found; such data have not been published before.

Theory: Ion-exchange isotherms

Ion-exchange isotherms are plots of the equilibrium fraction of an exchanging ion in solution against the equilibrium fraction of the same ion in the zeolite at the same temperature. The isotherms are plotted in terms of equivalent cation fraction of the ion in the solution (X_{sol}) against that in the solid (X_{zeo}) in accordance with the analysis of Inglezakis et al.[23] The equivalent cation fraction in solution and on the clinoptilolite is calculated by using Equations 3 and 4 given below.

Construction of ion-exchange isotherms

In general, the ion-exchange reaction between a solution containing the cation A^{zA+} (where A is a cation of valence z_A) and the B form of clinoptilolite (B being a cation of valence z_B) may be written as[30]:

$$z_B A^{z_A+}{}_{(aq)} + z_A(BL_{z_B}) \rightleftharpoons z_A B^{z_B+}{}_{(aq)} + z_B(AL_{z_A}) \qquad \text{Reaction 1}$$

in which L is a portion of the clinoptilolite framework holding a unit negative charge and the subscript *aq* denotes the solution phase. The equivalent fraction of the exchanging cation in the solution (X_{sol}) is therefore given by:

$$X_{sol} = \frac{Z_A m_s^A}{Z_A m_s^A + Z_B m_s^B} \qquad \text{Equation 2}$$

where m_s^A and m_s^B are the molarities of the ions A and B in solution, respectively. The equivalent cation fraction in the clinoptilolite is given by:

$$X_{zeo} = \frac{z_A (M_{A,i} - M_{A,f})V}{W.CEC} \qquad \text{Equation 3}$$

where *W* is the zeolite mass in grams, *V* is the solution volume used in litres, $M_{A,i}$ and $M_{A,f}$ are the initial and final concentrations of the exchanging ion (in moles per litre) and CEC is the cation-exchange capacity of the zeolite (in eq/g).[23]

Experimental methods

The raw zeolite sample used in this study was obtained from Pratley (Pty) Limited (Kenmare, South Africa) which mines the zeolite in KwaZulu-Natal (South Africa). All chemicals and reagents used for experiments and analysis were analytical grade supplied by Merck Ltd. (Johannesburg, South Africa).

Zeolite preparation

Small grains of zeolite were obtained by first using a hammer to break the as-received samples into small pieces. These small pieces were then ground using a pestle and mortar and sieved to yield fractions differing in diameter from 0.60 mm to 0.85 mm. Sieving was repeated several times to minimise the retention of smaller grains in a sample with a larger size range. Prior to the batch adsorption experiments, the crushed zeolite was washed with distilled water three times to remove the surface dust, and then dried in an oven at 343.15 K for 24 h until a constant weight was attained.

Preparation of homoionic forms

Near homoionic forms of Na^+, K^+ and Ca^{2+} clinoptilolite were generated by treating 30-g batches of the purified clinoptilolite with 300 mL of 2 M chloride salts (NaCl, KCl and $CaCl_2$ for Na^+, K^+ and Ca^{2+}, respectively). The mixtures were then placed in a Labex shaking incubator (Edenvale, South Africa) at 298.15 K for 24 h at a speed of 200 rpm. The solutions of the Cl^- salts were replaced with fresh ones for a further 24 h. The treated clinoptilolite grains were washed several times with distilled water to eliminate excess metal chlorides, and dried in an oven at 343.15 K for 24 h. Treated zeolite fractions were used as adsorbents for Ni^{2+} removal. Metal chloride treatment was conducted based on the findings of previous studies that alkali and alkaline earth metals are cheap, commonly available and are the most effective exchangeable ions for heavy metal removal.[2]

Equilibrium studies

Equilibrium studies were done as follows. A stock solution (1000 mg/L) of Ni^{2+} was prepared by dissolving 2.04 g of $NiCl_2{\cdot}6H_2O$ in 1 L of distilled water. The composition of the synthetic aqueous solution used in this study was based on those used previously and the concentrations were within the range of typical industrial wastewaters, that is, 0.1 mg/L to 100 mg/L.[28]

Synthetic samples were prepared to give Ni^{2+} concentrations of 20 mg/L, 40 mg/L, 60 mg/L, 80 mg/L and 100 mg/L by adding an appropriate amount of $NiCl_2{\cdot}6H_2O$ stock solution to deionised water. The masses of clinoptilolite used in the experiments ranged from 0.2 g to 1.5 g and the solution volume ranged from 10 mL to 100 mL. Before adding the adsorbents, the pH of the solution was adjusted to 7, i.e. the pH at which H^+ and Ni^{2+} competition is minimal, using either 0.1 M NaOH or 0.1 M HNO_3 solution, following the method of Gaus and Lutze[26]. The zeolite mass to solution volume ratios and the aqueous mixture compositions used in the experiments were designed to yield a relatively evenly spaced distribution of points along the ion-exchange isotherm as well as significant differences in the initial and final concentrations of the cation in solution. No background electrolyte was added during the ion-exchange experiments.

The clinoptilolite-zeolite samples were put in 250-mL Erlenmeyer flasks and agitated in a Labex shaking incubator at 298.15 K, 308.15 K, 323.15 K, 333.15 K or 348.15 K for 24 h.[2] After equilibrium was established, the clinoptilolite was separated from the solution by centrifugation and the pH of the supernatant was adjusted to 7. The equilibrium concentration of the exchangeable ions (Na^+, K^+, Ca^{2+}) and Ni^{2+} in the samples was determined by atomic absorption spectrometry (AAS, Varian 55B). In order to calculate the experimental error, all adsorption experiments were performed in triplicate, which enabled us to account for the errors caused by instability of the isothermal bath temperature and the calibration of the AAS.

The metal concentration in the liquid phase was determined at the beginning (C_o) and at the end (C_f) of the adsorption. The following equation was used to compute the percentage uptake of the metal by the clinoptilolite:

$$\text{Sorption }\% = \frac{(C_o - C_f)100}{C_o} \qquad \text{Equation 4}$$

Results

The chemical analysis of the natural and homoionic clinoptilolite samples is given in Table 1.

Table 1: Chemical composition (wt %) of treated clinoptilolite samples determined by X-ray flourescence

Component	Weight (%)			
	Natural	Na form	K form	Ca form
SiO_2	77.24	77.37	76.13	76.50
Al_2O_3	13.30	13.18	13 09	12.93
Fe_2O_3	0.14	0.13	0.14	0.12
FeO	1.09	1.01	1.14	0.12
MnO	0.04	0.01	0.02	0.02
MgO	0.86	0.71	0.67	0.80
CaO	1.40	0.77	0.87	3.35
Na_2O	1.94	3.84	0.23	1.43
K_2O	3.78	2.94	7.78	3.77
TiO_2	0.14	0.14	0.14	0.14
Total	100.44	100.01	100.00	99.97

The elemental analysis of the natural and pretreated clinoptilolite revealed that it is mainly composed of SiO_2, Al_2O_3 and Fe_2O_3 with very low amounts of MnO and TiO_2 in the framework. The extra-framework ions Na^+, K^+, Mg^{2+} and Ca^{2+} showed considerable variation depending on the pretreatment agent used and this change was mainly at the expense of Na^+ and Ca^{2+} content. SiO_2 ranged from 76.13% to 77.37% and Al_2O_3 ranged from 12.93% to 13.30%.

In the natural sample about 50% of the exchangeable ions were K^+. Complete exchange of cations was impossible to achieve, especially for the Na^+- and Ca^{2+}-treated clinoptilolite as it is assumed that K^+ present in the clinoptilolite did not exchange significantly with other cations. This behaviour is attributed to the location of K^+. Based on the chemical composition of the natural clinoptilolite used in this study, the molecular formula for the natural zeolite was:

$(Na_{1.25}K_{1.6})(Ca_{0.49}Mg_{0.43})(Al_{5.14}Fe_{0.31})(Si_{25.67}Ti_{0.04})O_{62}$

The pretreatment process led to the production of different forms of zeolite depending on the treatment agent used. The molecular formulae of the pretreated samples were as follows:

Na^+-clinoptilolite:

$(Na_{2.45}K_{1.24})(Ca_{0.27}Mg_{0.35})(Al_{5.13}Fe_{0.31})(Si_{25.5}Ti_{0.04})O_{62}$

K^+-clinoptilolite:

$(Na_{0.15}K_{3.36})(Ca_{0.31}Mg_{0.34})(Al_{5.21}Fe_{0.35})(Si_{25.7}Ti_{0.04})O_{62}$

Ca^{2+}-clinoptilolite:

$(Na_{0.88}K_{1.53})(Ca_{1.14}Mg_{0.38})(Al_{4.84}Fe_{0.29})(Si_{24.35}Ti_{0.033})O_{59}$

The cation-exchange capacity (CEC, milliequivalents (meq) per gram) of the zeolite samples was derived from the analysis given in Table 1. The calculation of CEC was based on the assumptions that (1) Al^{3+} and Fe^{3+} substitute for Si^{4+} in the tetrahedral sites and result in a negatively charged structure, (2) this negative charge is balanced by the alkali and alkaline earth ions in the intracrystalline cation-exchange sites and (3) any other exchangeable ions present in a homoionic form of clinoptilolite are assumed to occupy inaccessible sites; for instance, the K^+, Ca^{2+} and Mg^{2+} present in the Na^+-clinoptilolite are in inaccessible exchange sites and do not participate in the ion-exchange process. An additional assumption is that negligible mineral impurities are present in the sample.[22]

Experimental CECs of clinoptilolite were determined and were found to range from 1.38 meq/g to 2.29 meq/g for all forms of clinoptilolite (for natural and pretreated clinoptilolite).

Previous studies showed that the South African zeolite comprises mainly clinoptilolite (80–85%) and coexists with impurities of opaline, cristobalite, K-feldspar and traces of sanidine.[2] From the X-ray flourescence results, the Si:Al ratio was calculated to be 5.14. The chemical composition, the theoretical exchange capacity and the Si:Al ratio (generally ranging from 4 to 5.5) are typical for clinoptilolite.[31] Low SiO_2 members are enriched with Ca^{2+}, whereas high SiO_2 clinoptilolite is enriched with K^+, Na^+ and Mg^{2+}. It was found that the natural zeolite was predominantly clinoptilolite.

Effects of ion exchange

The surface area and pore volume data for Na^+, K^+ and Ca^{2+} form clinoptilolite are presented in Table 2. The ionic radii taken from Cotton and Wilkinson[32] have also been included in Table 2. The data for the Na^+ and Ca^{2+} forms are within experimental error, while the K^+ form has a higher surface area and pore volume.

$$2Na^+Zeo^-_{(s)} + Ni^{2+}(Cl^-)_{2(aq)} \rightleftharpoons Ni^{2+}(Zeo^-)_{2(s)} + 2Na^+Cl^-_{(aq)} \qquad \text{Reaction 5}$$

$$2K^+Zeo^-_{(s)} + Ni^{2+}(Cl^-)_{2(aq)} \rightleftharpoons Ni^{2+}(Zeo^-)_{2(s)} + 2K^+Cl^-_{(aq)} \qquad \text{Reaction 6}$$

$$Ca^{2+}(Zeo^-)_{2(s)} + Ni^{2+}(Cl^-)_{2(aq)} \rightleftharpoons Ni^{2+}(Zeo^-)_{2(s)} + Ca^{2+}(Cl^-)_{2(aq)} \qquad \text{Reaction 7}$$

Table 2: Structural characteristics of different homoionic forms of the zeolite

Adsorbent	Surface area	Pore volume	Ionic radii (Å)
	S_{BET} (m²/g)	V_t (cm³/g)	
Na form	13.54	3.11	Na^+ (1.16)[32]
K form	16.49	3.79	K^+ (1.52)[32]
Ca form	14.31	3.28	Ca^{2+} (1.14)[32]

Note: *The surface area and pore volume data were obtained from Brunauer Emmett Teller (BET) analysis. The ionic radii are values from Cotton and Wilkinson[32].*

Figures 1 to 4 present the experimental ion-exchange results. The initial and final concentration of Ni^{2+} was measured as well as the concentration of the displaced ions, Na^+ for the Na^+ form, K^+ for the K^+ form and Ca^{2+} for the Ca^{2+} form. These measured concentrations, zeolite masses and the volumes of solutions were used in the construction of ion-exchange isotherms.

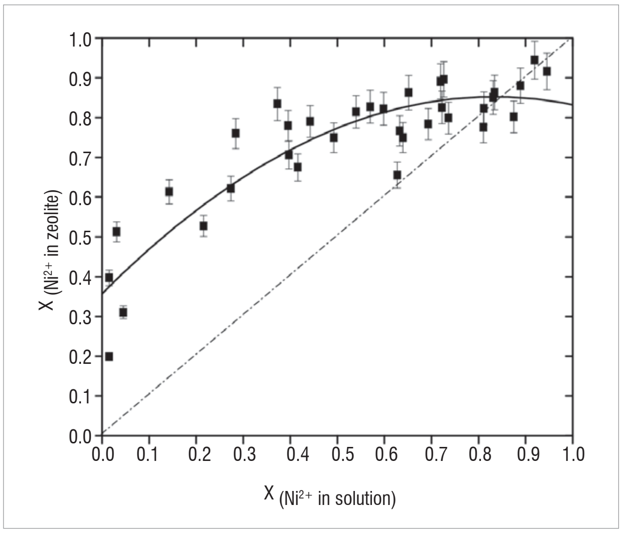


Figure 1: The Na^+-Ni^{2+} isotherm at 298.15 K. $X_{(Ni^{2+}\ in\ soln)}$: equivalent fraction of ingoing nickel in the liquid phase. $X_{(Ni^{2+}\ in\ Zeo)}$: equivalent fraction of ingoing cation in the solid phase. Percentage error bars for all points are indicated. The solid curve is a polynomial fit of order 2.

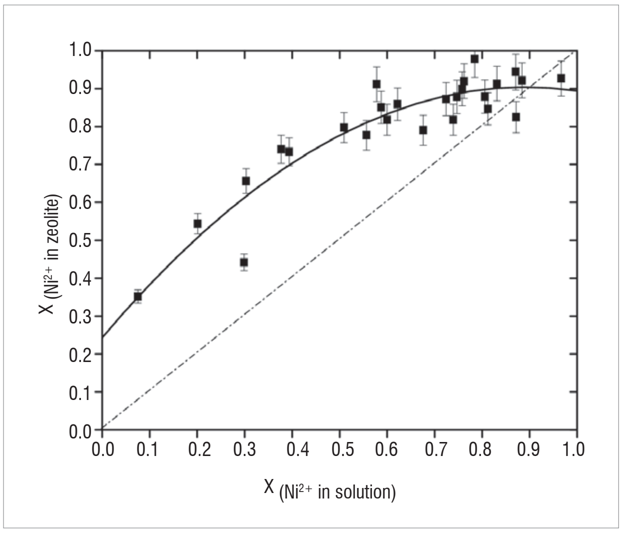


Figure 2: The K^+-Ni^{2+} isotherm at 298.15 K. $X_{(Ni^{2+}\ in\ soln)}$: equivalent fraction of ingoing nickel in the liquid phase. $X_{(Ni^{2+}\ in\ Zeo)}$: equivalent fraction of ingoing cation in the solid phase. Percentage error bars for all points are indicated. The curve was fitted to the isotherm data using a polynomial of order 2.

A comparison of the isotherms (Na^+-Ni^{2+}, K^+-Ni^{2+}, natural-Ni^{2+} and Ca^{2+}-Ni^{2+}) gives insight into how the displaced ion affects the selectivity of the clinoptilolite for Ni^{2+}. The Na^+-Ni^{2+} isotherm indicates a strong selectivity for Ni^{2+} through the entire range of zeolite composition. It is observed that the isotherms shown in Figures 1 to 3 do not proceed to completion (that is, do not attain $X_{solution} = 1$ for $X_{zeolite} = 1$). This result is attributed to the fact that in these systems only a fraction of the total CEC is available to the incoming cations as a result of crystallite occlusion; nonetheless the CEC is close to 1 (0.9 ± 0.1). The same observation was reported by Breck[33]. When the isotherm exhibits a convex profile, e.g. for Na^+-Ni^{2+}, K^+-Ni^{2+} and natural-Ni^{2+} systems, the uptake of Ni^{2+} from the solution is known to follow a Langmuir-type isotherm.[34] The Na^+, K^+ and natural forms show highly selective convex graphs, whereas the Ca^{2+} form has a concave graph, suggesting the selectivity series to be $Ca^{2+} > Ni^{2+} >$ (Na^+, K^+, natural). This selectivity series is in agreement with the investigation done by Colella[34] on Italian clinoptilolite. Figure 4 demonstrates that the clinoptilolite has a greater affinity for Ca^{2+} than for Ni^{2+}. This observation is consistent with the expectation that the Ca^{2+} ion has a relatively lower desorption ratio than Na^+ and K^+ because of the strong interaction (ionic in nature) between clinoptilolite and Ca^{2+} (it is difficult to destroy the Ca^{2+}–O–Al bonds). A similar observation was made by Ćurković et al.[35] for Serbian clinoptilolite. A remarkably high selectivity of natural-, K^+- and Na^+-exchanged clinoptilolite for Ni^{2+} is observed.

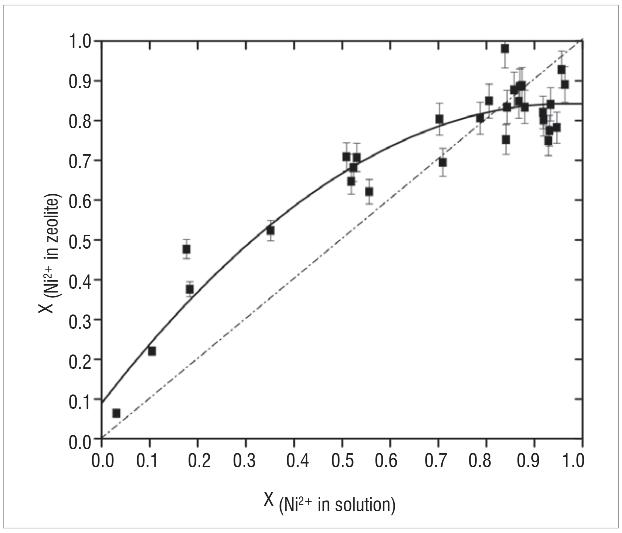


Figure 3: The natural-Ni^{2+} isotherm at 298.15 K. $X_{(Ni^{2+}\ in\ soln)}$: equivalent fraction of ingoing nickel in the liquid phase. $X_{(Ni^{2+}\ in\ Zeo)}$: equivalent fraction of ingoing cation in the solid phase. Percentage error bars for all points are indicated. The curve was fitted to the isotherm data using a polynomial of order 2.

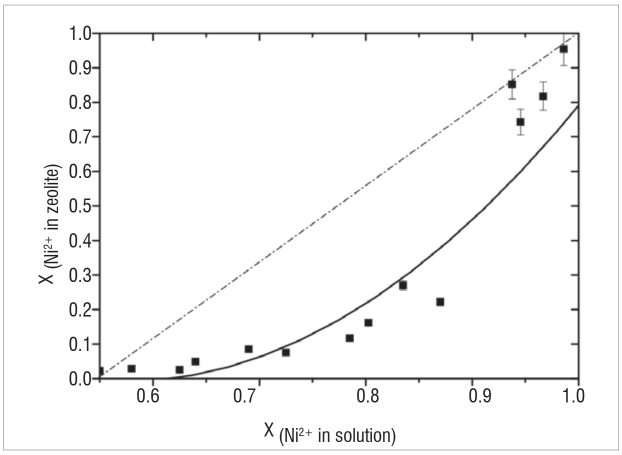


Figure 4: The Ca^{2+}-Ni^{2+} isotherm at 298.15 K. $X_{(Ni^{2+}\ in\ soln)}$: equivalent fraction of ingoing nickel in the liquid phase. $X_{(Ni^{2+}\ in\ Zeo)}$: equivalent fraction of ingoing cation in the solid phase. Percentage error bars for some points are indicated. The curve was fitted to the isotherm data using a polynomial of order 2.

In order to obtain the parameter capable of describing the selectivity/non-selectivity of a given form of zeolite, the thermodynamic equilibrium constant (K_{eq}) was evaluated using the following procedure outlined by Khan and Singh[36]:

$$K_{eq} = a_z/a_s = f_z C_z / f_s C_s \qquad \text{Equation 8}$$

where a_z denotes the activity of adsorbed Ni^{2+}, a_s is the activity of Ni^{2+} in solution at equilibrium, C_z is the milligrams of Ni^{2+} adsorbed per litre of solution in contact with the clinoptilolite surface, C_s denotes the milligrams per litre of Ni^{2+} in solution at equilibrium, f_z is the activity coefficient of the adsorbed Ni^{2+} and f_s is the activity coefficient of the

Ni^{2+} in solution. The method used to calculate ion activities is the one proposed by Debye and Huckel[37] in 1923 in which the ionic strength of the solution is calculated first and then the activity coefficient. Because the activity coefficient approaches unity at very low concentrations, Equation 8 can be re-written as

$$\lim_{C_z \to 0} \frac{C_z}{C_s} = \frac{a_z}{a_s} = K_{eq} \quad \text{which reduces to} \qquad \text{Equation 9}$$

$$K_{eq} = \frac{C_z}{C_s} \qquad \text{Equation 10}$$

The values of K_{eq} were calculated as the ratio of the equilibrium concentration of Ni^{2+} on the zeolite and in solution attained after 24 h of adsorption. The standard free energy change on adsorption (ΔG°) was calculated using the following equations:

$$\Delta G° = \text{-RT} \ln K_{eq} \qquad \text{Equation 11}$$

Several researchers have reported ΔG° values for adsorption in zeolites.[23,25] In their calculations they only reported $\Delta G°_{ads}$, and did not mention the $\Delta G°_{des}$ of the ion being displaced. By considering these two parameters, a true Gibbs free energy of exchange ($\Delta G°_{exc}$) can be computed which is representative of the whole system. Therefore Equation 12 is used to calculate the standard free energy of reaction:

$$\Delta G°_{exc} = \Delta G°_{ads} - \Delta G°_{des} \qquad \text{Equation 12}$$

where $\Delta G°_{exc}$ denotes the Gibbs free energy of ion-exchange reaction, $\Delta G°_{ads}$ is the Gibbs free energy of adsorption of Ni^{2+} onto clinoptilolite and $\Delta G°_{des}$ is the Gibbs free energy of desorption of exchangeable ions in the clinoptilolite (Na^+, K^+ and Ca^{2+}). R is the universal gas constant (8.314 J/mol.K) and T is the temperature in Kelvin.

The ΔS° and ΔH° values were obtained from the slope and intercept of the Van't Hoff plots (plots of the natural logarithm of equilibrium constant of the reaction versus the reciprocal of temperature in Kelvin).

$$\ln K_{eq} = \Delta S°/R - (\Delta H°/R)1/T \qquad \text{Equation 13}$$

The enthalpy of the reaction can be found from the gradient of the plot, which equals -ΔH°/R, and the entropy is the intercept, ΔS°/R. Table 3 reports the thermodynamic values at different temperatures.

The equilibrium constants K_{eq} derived from this study were quite high compared to those reported by Argun[27] despite using the same type of zeolite. Argun reported very low K_{eq} values of 3.28, 2.97 and 2.65 at 293.15 K, 313.15 K and 333.15 K, respectively, using Turkish natural clinoptilolite from the Langmuir isotherm constants to approximate the equilibrium constant, whereas, in this study, we found K_{eq} to be 23.69 at 298.15 K, which increased as temperature was increased (Table 3). A re-evaluation of the isotherm data in Argun's[27] study (after 3-h contact time instead of 60 min as in the original paper) at 293.15 K and at an initial concentration of Ni^{2+} of 25 mg/L for 1 g of clinoptilolite yielded a K_{eq} of 18 for the natural-Ni^{2+} system, which is very close to the values obtained in this study. The recalculated ΔG° value at 293.15 K is -7.04 kJ/mol (from -2.89 kJ/mol). Our results demonstrate that the reliance on linearised Langmuir equations potentially limits the ability to model sorption data accurately.

The thermodynamic quantities reported in Table 3 are in accordance with the selectivity series depicted from the isotherm plots. The Gibbs free energy of natural and pretreated clinoptilolite were evaluated, and the spontaneity of adsorption is seen to follow the series Na^+-form > natural-form > K^+-form > Ca^{2+}-form. In all cases, the free energy (ΔG°) of the Ni^{2+} sorption was negative, suggesting that the spontaneity of the process increased with increasing temperature. The values of the ΔG° also confirm that the maximum adsorption is obtained with the Na^+-exchanged clinoptilolite followed by the natural, K^+ and Ca^{2+} forms.

The ΔH° was positive in all forms of clinoptilolite and ranged from 18.72 kJ/mol to 42.05 kJ/mol (Table 1), which indicates that the sorption reaction is endothermic for Ni^{2+}. These values were calculated from plots of ln K_{eq} versus 1/T. The linear nature of the plot indicates that the mechanism of adsorption is not changed as temperature is changed. But the amount of adsorption is changed because the supply of thermal energy is different. The endothermic nature of the adsorption processes shows that these processes are not energetically stable.[30] If the values of ΔH° for Ni^{2+} adsorption had been within the range of 8.4–12.6 kJ/mol, then one could propose that the adsorption process was ionic in nature.[15] However, the values obtained in this study were greater than 12.6 kJ/mol, which indicates that the mechanism for the adsorption of these ions in zeolites is not ion exchange.

The Ni^{2+}-Na^+ system

The ion-exchange isotherms shown in Figure 1 demonstrate the extreme selectivity of Na^+-clinoptilolite for the incoming Ni^{2+} cations, which is confirmed by the high values of the equilibrium constant ($K_{eq}>1$) in Table 3 and the corresponding negative values of the free energies of exchange. The data also indicate that the value of K_{eq} increases with increasing temperature from 298.15 K to 348.15 K in all forms of the zeolite.

The enthalpy of exchange is positive for the Ni^{2+}-Na^+ clinoptilolite. This positive value is explained on the one hand by the substitution of Na^+ cations by Ni^{2+} cations, which have a greater heat of hydration (-406 kJ/mol for Na^+ and -2105 kJ/mol for Ni^{2+})[32] in the aqueous phase, and on the other hand by the greater interaction energy of the Ni^{2+} cations with the exchange centres of clinoptilolite as a result of their similarity in size.

The K^+-Ni^{2+} and natural-Ni^{2+} systems

The K^+-Ni^{2+} and natural-Ni^{2+} systems presented highly selective convex isotherms when Ni^{2+} was adsorbed; this finding is confirmed by the high values of the exchange constant ($\sim K_{eq}>16$ at 323.15 K). The enthalpies of exchange for both of these systems were positive and were almost of the same magnitude (18.75 kJ/mol for K^+-Ni^{2+} and 22.38 kJ/mol for natural-Ni^{2+}). This result suggests that K^+ is common to both systems as an exchangeable ion.

Table 3: Values of various thermodynamic parameters for the adsorption of Ni (II) in solution onto natural and pretreated clinoptilolite

Form of zeolite	**Thermodynamic parameters**											
	K_{eq}					$\Delta G°_{reaction}$ (kJ/mol)					ΔH° (kJ/mol)	ΔS° (kJ/mol.K)
	298.15 K	308.15 K	323.15 K	333.15 K	348.15 K	298.15 K	308.15 K	323.15 K	333.15 K	348.15 K		
Na form	29.60	44.21	53.85	83.75	144.48	-9.12	-10.54	-11.63	-13.49	-15.90	26.00	0.12
K form	8.73	11.26	16.45	21.73	24.47	-5.58	-6.46	-7.82	-8.90	-9.82	18.72	0.08
Ca form	0.35	0.989	1.34	2.2	9.01	3.10	0.53	-0.34	-3.94	-6.11	42.05	0.14
Natural form	23.69	24.53	39.87	60.45	76.73	-8.50	-9.70	-10.69	-12.20	-13.50	22.38	0.10

The Ca^{2+}- Ni^{2+}system

Distinct from the other systems, the Ni^{2+}-Ca^{2+} system yielded a concave isotherm, clear non-selectivity and the equilibrium constants (K_{eq}) were less than 1 at 298.15 K and 305.15 K (Table 3). The Gibbs free energy of exchange was positive at these temperatures, but a further increase in temperature resulted in the spontaneity of the reaction. The ΔG° values were positive at 298.15 K and 308.15 K (3.098 kJ/mol and 0.527 kJ/mol, respectively) and positive at temperatures greater than 323.15 K. Pabalan[22] also reported a positive ΔG° value of 4.19 kJ/mol at 298.15 K for a system involving Ca^{2+}-Na^+ using American clinoptilolite.

The substitution of the doubly charged Ca^{2+} cations with doubly charged Ni^{2+} ions proved to be an endothermic exchange process (positive enthalpy). The entropy for the Ca^{2+}-Ni^{2+}exchange system was positive because in the aqueous phase the strongly hydrated doubly charged cations (Ca^{2+}) were substituted by the similarly hydrated Ni^{2+} cations.

Conclusion

Evaluation of the thermodynamic parameters K_{eq}, ΔG°, ΔH° and ΔS° provided insight into the mechanism of Ni^{2+} sorption by the zeolite. The results of this research showed that clinoptilolite is an effective low cost adsorbent for the removal of Ni^{2+} from aqueous solution. Pretreatment of the zeolite enhances the removal efficiency if monovalent ions are used, and selectivity depends on the type of the exchangeable ions on the zeolite. Treating the zeolite with $CaCl_2$ decreased the zeolite's ability to remove Ni^{2+} from aqueous solution. Ca^{2+} bind more strongly to the zeolite than sodium and potassium ions. The thermodynamic parameters revealed that Ni^{2+} sorption in clinoptilolite is spontaneous and endothermic. The surface area and pore volume data presented for the Na^+ and Ca^{2+} forms were within experimental error, while the K^+ form had a higher surface area and pore volume.

Acknowledgements

This work was supported financially by the National Research Foundation of South Africa in the form of a Thuthuka grant and grantholder bursaries. The clinoptilolite was generously donated by Pratley (Pty) Limited (Kenmare, South Africa).

Authors' contributions

J.G., B.T. and K.K. collected the data; L.L.J. supervised the work; J.G. wrote the first draft of the manuscript; and all authors contributed to revisions of the manuscript.

References

1. Lalhruaitluanga H, Prasad MNV, Radha K. Potential of chemically activated and raw charcoals of *Melocanna baccifera* for removal of Ni(II) and Zn(II) from aqueous solutions. Desalination. 2011;271(1–3):301–308. http://dx.doi.org/10.1016/j.desal.2010.12.055

2. Kapanji K, Jewell LL. Ion exchange of Cu^{2+}, Ni^{2+} and Co^{2+} with South African clinoptilolite. In: WISA 2008 Biennial Conference Proceedings; 2008 May 18–22; Sun City, South Africa [proceedings on the Internet]. c2008 [cited 2014 March 22]. Available from: http://www.ewisa.co.za/literature/files/2008_110.pdf

3. Erdem E, Karapinar N Donat R. The removal of heavy metal cations by natural zeolites. J Coll Int Sci. 2004;280(2):309–314. http://dx.doi.org/10.1016/j.jcis.2004.08.028

4. Inglezakis VJ, Stylianou M, Loizidou M. Ion exchange and adsorption equilibrium studies on clinoptilolite, bentonite and vermiculite. J Phys Chem Solids. 2010;71(3):279–284. http://dx.doi.org/10.1016/j.jpcs.2009.12.077

5. Al-Haj Ali A, El-Bishtawi R. Removal of lead and nickel ions using zeolite tuff. J Chem Tech Biotech. 1997;69(1):27–34. http://dx.doi.org/10.1002/(SICI)1097-4660(199705)69:1<27::AID-JCTB682>3.0.CO;2-J

6. Baker HM, Massadeh AM, Younes HA. Natural Jordanian zeolite: Removal of heavy metal ions from water samples using column and batch methods. Environ Monit Assess. 2009;157(1–4):319–330. http://dx.doi.org/10.1007/s10661-008-0537-6

7. Blanchard G, Maunaye M, Martin G. Removal of heavy metals from waters by means of natural zeolites. Water Res. 1984;18(12):1501–1507. http://dx.doi.org/10.1016/0043-1354(84)90124-6

8. Buasri A, Chaiyut N, Phattarasirichot K, Yongbut P, Nammueng L. Use of natural clinoptilolite for the removal of lead (II) from wastewater in batch experiment. Chiang Mai J Sci. 2008;35(3):447–456.

9. Çoruh S, Ergun ON. Ni^{2+} removal from aqueous solutions using conditioned clinoptilolites: Kinetic and isotherm studies. Environ Prog Sus Energy. 2009;28(1):162–172. http://dx.doi.org/10.1002/ep.10316

10. Dubinin MM, Isirikyan AA, Regent NI. Isotherms of the equilibrium exchange of Mg^{2+}, Ca^{2+}, Zn^{2+}, and Cd^{2+} ions on NaA zeolite. Bull Acad Sci USSR Div Chem Sci. 1974;23(6):1172–1177. http://dx.doi.org/10.1007/BF00923072

11. Dyer A, Enamy H. The sodium-calcium exchange in zeolite A. Zeolites. 1985;5(2):66–67. http://dx.doi.org/10.1016/0144-2449(85)90073-9

12. Dyer A, Zubair M. Ion-exchange in chabazite. Micropor Mesopor Mater. 1998;22(1–3):135–150. http://dx.doi.org/10.1016/S1387-1811(98)00069-9

13. Oter O, Akcay H. Use of natural clinoptilolite to improve water quality: Sorption and selectivity studies of lead(II), copper(II), zinc(II), and nickel(II). Water Environ Res. 2007;79(3):329–335. http://dx.doi.org/10.2175/106143006X111880

14. Yuan J, Yang Y. Ion exchange equilibria between clinoptilolite and aqueous solutions of K^+-Na^+-Ca^{2+} and K^+-Na^+-NH^4. Ion Exch Adsorp. 2008;24(6):496–503.

15. Günay A, Arslankaya E, Tosun I. Lead removal from aqueous solution by natural and pretreated clinoptilolite: Adsorption equilibrium and kinetics. J Hazard Mater. 2007;146(1–2):362–371. http://dx.doi.org/10.1016/j.jhazmat.2006.12.034

16. Jänchen J, Ackermann D, Stach H, Brösicke W. Studies of the water adsorption on zeolites and modified mesoporous materials for seasonal storage of solar heat. Sol Energ. 2004;76(1–3):339–344.

17. Nour El-Dien FA, Ali MM, Zayed MA. Thermodynamic study for the (NH^{4+}-K^+) exchange on K-saturated clinoptilolite. Thermochim Acta. 1997;307(1):65–75. http://dx.doi.org/10.1016/S0040-6031(97)00316-X

18. Sun X, Xi C, Hou Z. Study on modification and fluoride-adsorption capacity of zeolite. In: Xiong F, Luo Q, editors. Proceedings 2010 International Conference on Challenges in Environmental Science and Computer Engineering (CESCE), 2010 March 6–7. Wuhan, China. Los Alamitos, CA: IEEE Computer Society; 2010. p. 354–357.

19. Taffarel SR, Rubio J. On the removal of Mn^{2+} ions by adsorption onto natural and activated Chilean zeolites. Miner Eng. 2009;22(4):336–343. http://dx.doi.org/10.1016/j.mineng.2008.09.007

20. Wang X, Hu H, Sun C. Removal of copper (II) ions from aqueous solutions using Na-mordenite. Sep Sci Tech. 2007;42(6):1215–1230. http://dx.doi.org/10.1080/01496390701241956

21. White DA, Franklin G, Bratt G, Byrne M. Removal of manganese from drinking water using natural and modified clinoptilolite. Proc Saf Envir Prot Trans AIChE B. 1995;73(3):239–242.

22. Pabalan RT. Thermodynamics of ion exchange between clinoptilolite and aqueous solutions of Na^+, K^+ and Ca^{2+}. Geo Cosmo Acta. 1994;58(21):4573–4590. http://dx.doi.org/10.1016/0016-7037(94)90192-9

23. Inglezakis VJ, Loizidou MD, Grigoropoulou HP. Equilibrium and kinetic ion exchange studies of Pb^{2+}, Cr^{3+}, Fe^{3+} and Cu^{2+} on natural clinoptilolite. Water Res. 2002;36(11):2784–2792. http://dx.doi.org/10.1016/S0043-1354(01)00504-8

24. Jama MA, Yucel H. Equilibrium studies of sodium-ammonium, potassium-ammonium, and calcium-ammonium exchanges on clinoptiolite zeolite. Sep Sci Tech.1989;24(15):1393–1416. http://dx.doi.org/10.1080/01496398908050659

25. Fridriksson T, Neuhoff PS, Viani BE, Bird DK. Experimental determination of thermodynamic properties of ion-exchange in Heulandite: Binary ion-exchange experiments at 55 and 85 °C involving Ca^{2+}, Sr^{2+}, Na^+, and K^+. Am J Sci. 2004;304(4):287–332. http://dx.doi.org/10.2475/ajs.304.4.287

26. Gaus H, Lutze W. Equilibrium studies with Ca/Sr zeolite A. J Phys Chem. 1976;80(27):2948–2950. http://dx.doi.org/10.1021/j100908a007

27. Argun ME. Use of clinoptilolite for the removal of nickel ions from water: Kinetics and thermodynamics. J Hazard Mater. 2008;150(3):587–595. http://dx.doi.org/10.1016/j.jhazmat.2007.05.008

28. Rajic N, Stojakovic D, Jovanovic M, Logar NZ, Mazaj M, Kaucic V. Removal of nickel(II) ions from aqueous solutions using the natural clinoptilolite and preparation of nano-NiO on the exhausted clinoptilolite. Appl Surf Sci. 2010;257(5):1524–1532. http://dx.doi.org/10.1016/j.apsusc.2010.08.090

29. Sprynskyy M, Buszewski B, Terzyk A, Namieśnik J. Study of the selection mechanism of heavy metal (Pb2+, Cu2+, Ni2+, and Cd2+) adsorption on clinoptilolite. J Colloid Interface Sci. 2006;304:21–28. http://dx.doi.org/10.1016/j.jcis.2006.07.068

30. Caputo D, Pepe F. Experiments and data processing of ion exchange equilibria involving Italian natural zeolites: A review. Micropor Mesopor Mater. 2007;105(3):222–231. http://dx.doi.org/10.1016/j.micromeso.2007.04.024

31. Altin O, Ozbelge O, Dogu T. Use of general purpose adsorption isotherms for heavy metal-clay mineral interactions. J Colloid Interface Sci. 1998;198:130–140. http://dx.doi.org/10.1006/jcis.1997.5246

32. Cotton FA, Wilkinson G. Advanced inorganic chemistry. New York: John Wiley and Sons; 1988.

33. Breck DW. Zeolite molecular sieves. New York: John Wiley and Sons; 1974.

34. Colella C. Ion exchange equilibria in zeolite minerals. Miner Dep. 1996;31(6):554–562. http://dx.doi.org/10.1007/BF00196136

35. Ćurković L, Cerjan-Stefanović S, Filipan T. Metal ion exchange by natural and modified zeolites. Water Res. 1997;31(6):1379–1382. http://dx.doi.org/10.1016/S0043-1354(96)00411-3

36. Khan AA, Singh RP. Adsorption thermodynamics of carbofuran on Sn (IV) arsenosilicate in H^+, Na^+ and Ca^{2+} forms. Coll Surf. 1987;24(1):33–42. http://dx.doi.org/10.1016/0166-6622(87)80259-7

37. Debye P, Hückel E. Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen [On the theory of electrolytes I: Lowering of freezing point and related phenomena]. Phys Zeits. 1923;24:185–206. German.

Rethinking the science–policy interface in South Africa: Experiments in knowledge co-production

AUTHOR:
Mark Swilling[1]

AFFILIATION:
[1]School of Public Leadership, Stellenbosch University, Stellenbosch, South Africa

CORRESPONDENCE TO:
Mark Swilling

EMAIL:
swilling@sun.ac.za

POSTAL ADDRESS:
School of Public Leadership, Stellenbosch University, Private Bag X1, Matieland 7602, South Africa

DATES:
Received: 22 Aug. 2013
Revised: 11 Nov. 2013
Accepted: 13 Nov. 2013

KEYWORDS:
policy formulation; reflexive approach; transdisciplinary approach; sustainability; stakeholder engagement

HOW TO CITE:
Swilling M. Rethinking the science–policy interface in South Africa: Experiments in knowledge co-production. S Afr J Sci. 2014;110(5/6), Art. #2013-0265, 7 pages. http://dx.doi.org/10.1590/sajs.2014/20130265

This article contributes to the increasingly significant discussion about the science–policy interface. The challenge therein is that such a discussion tends to revolve around two seemingly mutually exclusive approaches: the reflexive approach inspired by Maarten Hajer's work that deconstructs the discourses of participatory policymaking, and the more normative transdisciplinary approaches that legitimise researchers as active change agents. With reference to a discussion of three South African case studies characterised by practical involvement of researchers in change processes, it is concluded that both approaches have merit and can improve the other: the reflexive approach could benefit from a better understanding of appropriate research methods for facilitating authentic engagement and participation, and the transdisciplinary approach could benefit from some reflexive caution about the change agent roles of researchers. The dynamics of the case studies and conclusions are significant in light of the fact that the South African research community is being influenced by re-alignments in the global scientific research community, resulting in an increasing emphasis on the need to do transdisciplinary research. For example, the adoption by some of the most significant global scientific associations in the natural and social sciences of the Future Earth platform at the Rio+20 conference in 2012 reflects most clearly this re-alignment. Researchers would be well advised to critically engage this agenda rather than presume it means little more than a rewording of traditional interdisciplinary approaches.

Introduction

The role of academic researchers in formulating policies about sustainability has drawn increasing interest from a wide range of different perspectives.[1] In this article, two particularly influential perspectives, which have in common a focus on the role of the researcher at the science–policy interface, are addressed. Inspired mainly by the work of the Dutch social scientist Maarten Hajer, one group is interested in a reflexive approach that reveals how researchers transform their craft into advocacy but rarely admit to their discursive role in complex power relations.[2-5] The second group works with theories of transdisciplinary research to develop a normative approach that intentionally promotes researchers as active 'co-producers' of problem-solving knowledge.[6,7] Perspectives that deal with the institutional dynamics of the science–policy interface[8] are not addressed.

While both traditions addressed here favour the active engagement of researchers in policy processes, they are concerned with very different dimensions. However, it will be argued that the reflexive approach could benefit from a greater appreciation of the practical methodologies and methods of co-production, while the transdisciplinary approach could be more reflexive about the consequences (and potential dangers) of combining researcher and advocacy roles. Considering the central place given to 'transdisciplinary research' in the new Future Earth programme adopted by the global science community at the so-called Rio+20 conference in 2012, it is an appropriate time to deliberate on these matters.

A synthesis of the reflexive and transdisciplinary approaches is used to reflect on three case studies from the South African context. These cases reveal the intricate dynamics of the science–policy interface where the search is on for ways of formulating sustainable solutions to South Africa's challenges.[8,9]

Reflexive approach

Since 1994 it has become common practice for South African scientists, academics and professional researchers to be drawn into the policy formation process as drafters of policy documents and background 'research papers'. Quite often the resulting policy is justified on the grounds that the policy formulation process was legitimate, because the inputs were not simply the 'subjective' perspectives of stakeholders and politicians but also 'objective' analyses of scientists and researchers. Fortunately, recent research[8] has started to raise questions about the validity of this 'required by science' discourse, and often draws on Maarten Hajer's influential work.[2-5]

Instead of accepting that participation of stakeholders and experts by definition improves policy content, Hajer deploys a constructivist approach to question what he calls the 'staged performances'[4] put on by policy managers who are obliged by their political and managerial masters to produce consensual outcomes. Usually this 'performance' means setting up processes in ways that reinforce consensus and suppress conflict.

Hajer's approach recognises the 'performative dimension of policy deliberation'. Like staging a dramatic production, the policy manager is effectively the orchestrator of a process that not only has a 'script', but also a physical 'setting' (e.g. venues organised in a certain way) within which a production is 'staged' (by a specific set of actors mandated to participate in the process), and a particular pattern of 'performances' (e.g. chaired/facilitated in a certain way, in a certain language) that are all equally important in shaping the final outcome as well as who can and cannot participate or whose voice carries more weight. To focus only on the script (i.e. the formal outcome of the process which is the text) is to miss the full significance of what is going on.

This application of theatre theory is what Hajer calls the 'dramaturgy' of the policy deliberation process.[3] It helps us to understand why formalistic participatory processes are often meaningless: stakeholders often just play out predetermined roles as defined by the script, setting and stage manager. It also helps us to understand how 'even with the same cast, policy deliberation can change face through experiments with new settings and stagings'[3].

Policy deliberation will not change just because there are skilled facilitators in place, or because there is improved capacity to be reflexive. It may have a better chance of happening if informed by particular methodologies and methods for co-producing knowledge. To achieve this enriched understanding of the policy deliberation process, it may help to apply the transdisciplinary research approach to give greater emphasis to a more reflexive practice at the science–policy interface.

Transdisciplinary approach

Transdisciplinary research has emerged as a mode of knowledge co-production that goes well beyond the traditional understanding of interdisciplinary research. The best way to define it is that it entails conducting interdisciplinary research with – rather than for – society in order to co-produce socially robust solutions to complex societal problems that can no longer be approached and solved by mono- or even interdisciplinary approaches.[6,7,10-14] Following Latour, this shift from 'for' to 'with' opens up a Pandora's box of old and new debates about the profoundly relational character of knowledge that can no longer be reduced to the quantitative enumerations regarded as sacrosanct by the natural sciences.[15]

Global warming, natural resource depletion and increasing poverty are just a few complex societal problems warranting a transdisciplinary response. They are complex because they are truly planetary-level problems, and because they are being produced by both nature and society and have long-term consequences for both. These 'hybrid' problems can no longer be approached by treating the 'natural' and the 'social' as two fundamentally different and unconnected realities which must, in turn, be worked on separately by the natural and social sciences in isolation of society. This divided approach can only result in producing partial knowledge of these problems, whereas the need today is clearly for integrated solutions based on integrated knowledge sets.[16]

To justify its claim as a new mode of knowledge co-production, it has been critical for researchers working from a transdisciplinary approach to establish the approach within the scientific community as credible and scientific. They have attempted to demonstrate that it is necessary to start with shared real-world problem statements which can then be translated into scientific problem statements and research questions. This outline then provides the basis for research that is co-produced with societal actors to produce knowledge that is relevant to societal actors and valid 'scientific knowledge'.

The transdisciplinary approach has very much been focused on the discovery, design and production of appropriate transdisciplinary methods that are replicable in different contexts. These methods are intended to successfully integrate quantitative and qualitative theoretical knowledge with socially generated transformative knowledge, to produce 'scientifically valid' and 'socially useful' knowledge. But in so doing, the transdisciplinary approach has – using Hajer's terms – not only produced a new script, but created the justification for a much bolder and comprehensive dramaturgy that the average researcher is now expected to manage. In practice, this creates for the researcher a more complex mode of double participation – as both 'participating insider' and as 'observing stranger'.

A reflexive note

Three cases from the South African context have been selected to highlight both the need for co-production of knowledge to address real-world problems and the need for critical reflexivity about the practices of engagement by researchers at the science–policy interface. They have been written up in ways that reveal how the use of words like 'objective' and 'subjective' are misleading because they imply that only experts can know the objective world 'out there' and that the subjective world 'in here' is not therefore relevant. Both 'worlds' interact as researchers position themselves within an inescapable paradox: embedded within the power relations of the contexture while at the same time entrusted as producer of analyses in accordance with the formal rules of science.

The first case is about formulating a 10-year policy framework for the national Department of Science and Technology which is a department mandated to invest in science and technology research. The second case refers to an ambitious urban regeneration strategy for Cape Town by the Western Cape Provincial Government (WCPG). Finally, the third case concerns the iShack project in the university town of Stellenbosch, aimed at addressing the challenge of incremental upgrading of informal settlements.

In all three cases I was involved as a researcher with specific knowledge expertise, but I was also an advocate: I strongly supported the need for government to invest in global change research over the long term; I also supported the policy intent of the WCPG with respect to urban regeneration in Cape Town, and I have for many years been concerned about the negative political consequences of the 'wait-for-the-grid' approach to in-situ upgrading of informal settlements. This was the 'contexture' of my role as a researcher: I was acting in various capacities, namely from active participant and facilitator to expert researcher who deployed the transdisciplinary research methodology. Denial to escape this paradox is not an option; it can only be recognised and incorporated into the analysis offered here.

Case 1: Global Change Grand Challenge

The Department of Science and Technology (DST) is a national government department responsible for the formulation of long-term policies and strategies aimed at supporting the transition from a resource-intensive to a knowledge-based economy.

In its long-term policy framework for accelerating innovation entitled *Innovation Towards a Knowledge-Based Economy 2008-2018*[17] the DST defined five 'grand challenges' that would then guide future investments in science and technology. One of these challenges was that of global change science, with a focus on climate change and a broader interest in transition to a more sustainable mode of economic production and consumption.

Significantly, the document entitled *Global Change Grand Challenge National Research Plan*[18] that was completed in June 2009 contains the following key sentence:

> *An inclusive process involving a wide cross-section of the science and policy communities in South Africa was followed to develop a detailed implementation plan for the first of these, i.e. enhancing our scientific understanding. This process has culminated in the development of this 10-year national research plan for the Global Change Grand Challenge (the Global Change Research Plan).*

Unfortunately, this 'inclusive process' is not discussed any further. The assumption created by this report is that the Global Change Grand Challenge (GCGC) unproblematically reflects a consensus reached by 'the science and policy communities' that participated in the various levels of engagement, i.e. the 'lead editors' who wrote the document, the 'editorial panel' comprising eminent academics who played an oversight role, and the 'contributing authors'. The key actors were drawn from the Council for Scientific and Industrial Research (CSIR) (which is a government-controlled and funded 'science council'); the National Research Foundation (NRF) which is a government agency that manages government funding for scientific research undertaken mainly by universities (specifically the South African Environmental Observation Network); officials from the DST; and academics from a select group of universities (with the Universities of Cape Town, Stellenbosch and Witwatersrand playing leading roles).

The 10-month process over 2008/2009 involved the following:

- intensive initial meetings between DST and CSIR to finalise a Terms of Reference;
- a series of facilitated consultative discussions and workshops involving about 50 scientists, most of whom were drawn from the CSIR, but with significant involvement by academics from various universities;
- submissions of written proposals from most of the participants, mostly referring to general conceptual issues but also practical implementation challenges related to expenditure and integration;
- intensive cycles of drafting and redrafting following comments submitted by participants;
- final approval by the DST and the Minister of Science and Technology.

The end result was a research framework divided into four themes as depicted in Table 1.

Significantly, the 'Understanding a changing planet' and 'Reducing the human footprint' themes reflected the research foci of the natural scientists, and there were already substantial funding pipelines in place for this research. The other two themes were unfunded and reflected the perspectives of those mainly interested in the economic and social dimensions of sustainability transitions.[19]

The process described above that led up to the formulation of the GCGC plan cannot be described as a research-based policy formulation process. Nor did it come remotely close to the notion of co-produced knowledge to address a real-world problem as envisaged by the transdisciplinary approach. It was, instead, more like a carefully staged policy negotiation process to craft a script that would protect and enlarge existing funding flows for Themes 1 and 2 and create new funding flows for Themes 3 and 4 (Table 1).

Themes 1 and 2 summarise the essence of the earth system science portfolio managed largely but not exclusively by the Applied Centre for Climate and Earth Systems Science (ACCESS) which is, in turn, managed by the CSIR. As a major consortium of universities and state-controlled research agencies/councils, ACCESS is South Africa's pre-eminent global change initiative within the natural sciences. It aims to secure an annual budget of EUR10 million. However, it does not address Themes 3 and 4. Hence those interested in opening up new funding flows to address the issues raised under Themes 3 and 4 needed to mount convincing arguments about the need to extend the scope of global change research beyond the traditional boundaries of earth system science.[20-22]

Besides the link between funding flows and policy arguments, there were four other aspects of the policy-formulation process that are worth noting, which shed light on the role researchers play as the stage managers of the science–policy interface:

- Competing conceptions of global change. As the majority of researchers involved were from the natural sciences and associated with ACCESS, they shared the earth system perspective on global change[20] which emphasises the importance of understanding extremely rapid changes in the global earth system. However, there was a minority of mainly social scientists whose conception of global change drew on material flow analysis[23], the Multi-Level Perspective[24,25] and the economics of socio-technical transitions[26], which emphasises the complex dynamics of transition and the importance of sustainability-oriented innovation systems[23]. In the end, the earth system perspective was reflected in Themes 1 and 2, and the transition perspective was reflected in Themes 3 and 4. Unsurprisingly, the bulk of the funding was allocated to Themes 1 and 2.
- Tension between research for deepening the understanding of systems and transdisciplinary research for co-producing knowledge. The failure to allocate funds later on for Research Chairs dealing with global and national sustainability transitions using a transdisciplinary approach (Theme 4) reflects the lower priority enjoyed by the transition/sustainability perspectives despite government policy commitments to a green economy and transition to sustainable development. Furthermore, in the first call for proposals for the first national research conference on global change (26–28 November 2012) issued by the NRF and DST, none of the transition themes (in Themes 3 and 4) were listed as topics for paper submissions. After objections, this was later changed to include the full span of topics.
- Limited involvement of the policy community. Besides the ongoing involvement of Imraan Patel, a senior DST official, there is little evidence that key policymakers within the DST and other government departments affected by the GCGC plan (such as the Department of Environmental Affairs) were involved in the policy formulation process.
- Weak connections to the private sector and civil society. The private sector and civil society stakeholders were effectively excluded from the process, despite considerable experience and expertise in connecting innovations to implementation.

In conclusion, the process of formulating the GCGC was stage managed by a tightly networked group of researchers with a vested interest in reproducing a policy framework that favoured funding flows into research programmes that they managed. This situation was certainly true of the earth systems research community, whereas the transitions research community – which included myself – aspired to secure these funding flows. Although nominally a partnership with the 'policy community', the DST set the stage for the performance but did little to guide it. The script and performance was carefully orchestrated primarily by the CSIR to

Table 1: Research framework outlined in the *Global Change Grand Challenge National Research Plan*

Understanding a changing planet	Reducing the human footprint	Adapting the way we live	Innovation for sustainability
1. Observation, monitoring and adaptive management 2. Dynamics of the ocean around southern Africa 3. Dynamics of the complex internal earth systems 4. Linking the land, air and sea 5. Improving model predictions at different scales	1. Waste minimisation methods and technologies 2. Conserving biodiversity and ecosystem services, e.g. clean drinking water 3. Institutional integration to manage ecosystems and ecosystem services	1. Preparing for rapid change and extreme events 2. Planning for sustainable urban development in a South African context 3. Water security for South Africa 4. Food and fibre security for South Africa	1. Dynamics of transition at different scales – mechanisms of innovation and learning 2. Resilience and capability 3. Options for greening the developmental state

Source: Department of Science and Technology[18] direct research funding along distinct pathways. Unsurprisingly, there is little evidence that this policy process was influenced by transdisciplinary methods of co-production of knowledge for innovation.

Case 2: Cape Town Central City Regeneration Initiative

In contrast to the researcher-driven GCGC process, the Cape Town Central City Provincial Government Regeneration Initiative (henceforth CTRI for short) was initiated by the Provincial Minister of Transport and Public Works. The initial policy formation phase, led by key officials in his department, occurred between December 2009 and May 2010.[27]

After his appointment as the WCPG's Minister of Transport and Public Works, Robin Carlisle initiated an informal networking process with a few individuals within and outside government to formulate a terms of reference for what eventually became the CTRI. He decided to build on two existing partnership agreements: one with the Cape Town Partnership led by Andrew Boraine and the other with the Cape Higher Education Consortium (CHEC) led by Nasima Badsha. The Cape Town Partnership is a non-profit that was established in 1999 as a partnership between the City of Cape Town (CCT) and the South African Property Owners Association and has spearheaded the regeneration and marketing of Cape Town's central business district (CBD). CHEC is also a non-profit established by the four Western Cape universities to coordinate joint activities, in particular collaborations with the WCPG and CCT.

CHEC was contracted by the Department of Transport and Public Works (DTPW) to constitute a Steering Committee that would assist the department's Regeneration Team, led by Francois Joubert, to formulate the overarching policy framework. Academics mainly from the Universities of Stellenbosch, Western Cape and Cape Town were drawn in, plus Andrew Boraine and Nasima Badsha, as well as well-known architects and planners from the consulting world, namely Mokena Makeka and Barbara Southworth (who was previously head of planning in the CCT).

The initial strategic intent of the CTRI was to catalyse an impactful urban regeneration initiative that would simultaneously double the floor space of the CBD with major economic consequences *and* resolve a key financial problem facing the DTPW. This case arose because the department was responsible for a large stock of government buildings that were not only dysfunctional as office buildings, but also more expensive to operate than what it would have cost to lease the buildings from the private sector. The Minister's vision was that strategically located public assets could be re-invented and then used as leverage (via sale or renovation) to catalyse large-scale private sector investments in urban regeneration within the precincts where the public sector buildings are located.

The primary task of the Steering Committee was to write up this proposal in the form of a policy document. Although the time period was too short for this task to be a genuine transdisciplinary research process, real expertise from different sources was mobilised and integrated via a set of discovery-oriented engagements that did result in significant debate, exploration and synthesis. The process involved the following engagements:

- regular meetings of a core coordination group;
- less regular, broader and more formal meetings of the Steering Committee to brainstorm key ideas and strategies (these meetings were the most crucial turning points in the process);
- increasingly frequent meetings later in the process which involved consultants working on the drafting of the policy document;
- research work undertaken mainly by Masters student Katherine Hyman[24] to collect and read key planning documents, especially those regarding infrastructure;
- ongoing informal interactions with key stakeholder groups from the private sector, consulting industry and CCT;
- a crucial stakeholder workshop on 9 April convened by CHEC that brought in key players from the property development industry, consulting firms, CCT, WCPG and the universities to discuss what was by then a draft policy framework;
- intense interactive engagements during the drafting phase which was concentrated into the months of April and May 2010.

Although the DTPW did not initially assume that sustainability was going to be an integral part of the argument and vision of the final policy document, it gradually became clear that there would be one overriding obstacle to the achievement of the vision, and that was the lack of an adequate urban infrastructure (specifically with respect to energy, solid waste, transport, water and sanitation). In contrast to what the consultants and some officials were saying, further research by the academics showed that there were real infrastructure constraints and that solutions using 'business-as-usual' technologies would be prohibitively expensive. This conclusion opened up the space for the introduction of a sustainability perspective, referring specifically to new technologies for treating sewage, using water more efficiently, designing and operating energy-efficient buildings, generating renewable energy, recycling solid waste and introducing non-motorised mobility and public transport.

However, it would be incorrect to ignore the fact that the WCPG and the CCT had over the previous several years evolved a range of policy and strategy documents that expressed commitments to a more sustainable use of resources and less negative impacts on the natural environment.[28-30] These documents created a legitimating language that key politicians, such as the Premier of the WCPG, tended to draw on to express future visions and plans. The dense network of NGOs and university-based researchers that deal with sustainability issues in Cape Town also have strong working relationships with the WCPG and the CCT. These relationships, together with the existence of influential business-linked groups interested in sustainability (e.g. Cambridge Programme for Industry, Accelerate Cape Town, Sustain Our Africa), have built up an accepted body of expertise and general awareness that infuses public and policy discourse. Without this discursive environment, it would not have been possible to introduce specific ideas into the CTRI policy document about sustainability-oriented urban infrastructure alternatives.

The final document was handed over to the DTPW on 17 May 2010, paving the way for phases 2 (precinct planning) and 3 (implementation) of the project. The final version captured a vision for the central city as a space that needs to be productive, connected, innovative, cohesive, sustainable and safe. In October 2013, the first major Brownfields Redevelopment Initiative was announced to realise the CTRI vision, namely the so-called Two Rivers Urban Park project envisaged for a 300-ha area largely owned by state agencies. This particular project emerged from a group of officials at provincial and city level working with key individuals from the universities and the private sector.

In conclusion, this case is about a knowledge partnership actively solicited and led by a government department (with full political support) that involved universities, the property development sector and consultants in a process that was not just about negotiating language to express a consensus to satisfy a political perspective. What mattered was not merely the content of the final report (the script), but also the setting (meetings at CHEC offices and offices of the private sector) and the process (input from the university and property development sectors) which validated and legitimised the final product. A rapid process of interactive discovery and debate informed by intensive information gathering and stakeholder engagement made it possible to co-generate a policy framework that has continued to inspire subsequent work and retain political support. Researchers were given space to investigate, raise questions, criticise the findings of consultants and facilitate learning processes that formed part of the joint planning process. Given the complexities, in this case our advocacy was informed by our research rather than the other way round.

Key criticisms would be the absence of involvement of non-governmental organisations or broader civil society sectors, and the fact that the CCT was only brought into the process towards the end which resulted in implementation delays.

Case 3: Incremental upgrading: The iShack initiative

Soon after 1994, the South African government introduced an ambitious housing programme to address the legacy of apartheid. The result was the construction of 2.9 million houses by 2010 – one of the highest rates of housing delivery to the poor in the world. Nevertheless, shrinking household sizes and population growth meant that by 2004, the housing backlog had grown from 1.5 million to 2.1 million housing units. To make matters worse, houses were built on cheap land to reduce costs, resulting in the bulk of housing being located on the urban peripheries far from places of employment and access to services. This peripheral location of low-income settlements resulted in the ballooning of bus transport subsidies to offset the rising costs of getting them to work and exacerbated household poverty in these settlements.

To remedy this problem, a new housing policy was introduced by the Department of Human Settlements in 2004 called *Breaking New Ground: A Comprehensive Plan for the Development of Sustainable Human Settlements* (commonly referred to as BNG). A key component of this new policy was acceptance of the need for in-situ upgrading of informal settlements rather than relocating them. This policy resulted in what is now called the Upgrading of Informal Settlements Programme (UISP). The Minister of Human Settlements signed his performance agreement with the President in 2010 which committed him to the upgrading of 400 000 shacks by 2014.

In early 2011, a group of postgraduate students decided to focus their research on an illegal informal settlement of 6000 people called Nkanini (which means 'take by force'), located within walking distance of Stellenbosch.[18] The initial research question was: what does in-situ upgrading (as specified by the UISP) mean in practice from the perspective of the average shack dweller living in Nkanini? A transdisciplinary research methodology was adopted, but it was recognised from the start that the relevant formal stakeholders could not be identified as required by mainstream transdisciplinary approaches. Instead, direct relationships needed to be established with the community, which included students moving into the community to experience living in a shack, building relationships with individuals and mounting visible campaigns such as the painting of shacks using bright colours and designs. Contact was made with the Informal Settlement Network (ISN), which is a social movement active in the Stellenbosch area, supported in turn by Shackdwellers International (SDI) (www.sdinet.org). A working relationship of sorts was also established with relevant officials in the Stellenbosch Municipality who were, in turn, working formally with ISN/SDI.

It became apparent early on that in practice, the UISP means delivery by the municipality (subject to funding from higher levels of government) of electricity (street lights only), water, sanitation, roads, and stormwater and solid waste services. However, this service provision can happen only if the settlement has been legally recognised as permanent and the land has been rezoned as residential. Neither of these conditions were met in the case of Nkanini: it is one of the few informal settlements in South Africa which has been formally declared illegal and therefore needs to be removed – a threat that has never been carried out. Even if Nkanini were legally entitled to be there, then in-situ upgrading would in practice mean waiting for the electricity and water grids to arrive, with minimal services for solid waste collection in the meantime. The WCPG has calculated that on average it takes 8 years after legalisation or rezoning for communities to be connected to the electricity and water grids after formal commencement of the upgrading process. Even then, all the community is likely to receive is street lighting, not electrical connections to each unit.

In short, the problem statement became: upgrading means 'wait for the grids to arrive'. The research question became: what could be done between now and the arrival of the grids to improve quality of life? The fact that development has come to mean 'trust and wait' effectively demobilises civil society because there is nothing to organise communities around that can result in tangible, immediate improvements to daily life. Instead, activists discover where the state intends delivering next and stay one step ahead by organising people around what is going to get delivered anyway in a uniform top-down manner. This situation is not only a recipe for a weak civil society, but also effectively undermines democracy. This scenario contextualises the significance of the transformation-oriented research question. Note how different this approach is to the most common research questions asked about informal settlements, namely 'why do they exist?', 'what are the living conditions?' (which are both about systems knowledge) or, occasionally, the target knowledge question, 'what is the solution?'. 'What can be done now by members of the community?' is a transformation knowledge question.

After many months of informal interactions with the community, officials and ISN, and informed by initial research on UISP and BNG, it was decided that ecological design methods may open up an alternative way of thinking about a genuine incremental approach to upgrading, one that avoids all the negative consequences of the 'trust and wait' approach. Working with engineers and an ecological architect, a design was generated for an 'improved shack' – the 'iShack'[31,32]. This design amounted to a 14.2-m^2 shack that included:

- insulation in the walls and roof, covered with cardboard painted with fire retardant paint;
- a thermal mass for passive heating and cooling by using a 1-m high adobe wall along the back of the shack together with a floor made from fired clay bricks reclaimed from the landfill;
- a north–south orientation plus a roof overhang on the north side for shade in summer and solar penetration in winter;
- correctly sized and located windows for lighting and ventilation;
- a 25-W solar panel to power three LED lights and a cell phone charger;
- a gutter to capture rainwater.

Working with an informal group of local leaders and with permission from the local authority, a very poorly built wooden shack inhabited by a single mother with three young children was identified for replacement. After building a new iShack for her, the old shack was demolished as required by the local authority. A neighbouring shack was retrofitted with insulation and a solar unit. Environmental monitoring equipment was installed in both shacks, plus a neighbouring non-retrofitted shack in order to generate comparisons. The results showed conclusively the benefits of the intervention, which included 4–6 h of extra thermal comfort each day, reduced fire risk and improved lighting.

Four rather dramatic consequences flowed from the erection of the iShack, the retrofitting of two neighbouring shacks and related research on sanitation and solid waste. Firstly, a process of social mobilisation within the community started to take place around demands for incremental upgrading. The core group of community members who worked with the students accumulated skills and knowledge, including attending training modules paid for by the project. In other words, what started off largely as a rather limited technical intervention spiralled out into a wider community mobilisation process. Secondly, the Bill and Melinda Gates Foundation (which also funds the international work of SDI) requested a funding proposal to take the work forward, resulting in a grant of USD 250 000 in June 2012. Thirdly, the government's Green Fund decided to allocate another R1.7 million to help the project reach scale. Fourthly, in 2013, the Stellenbosch Municipality changed its indigent policy to provide for the transfer of the free basic electricity subsidy to non-grid connected shack dwellers – an unprecedented innovation. In addition, the iShack project has attracted extensive media attention in the mainstream and local press and resulted in four television appearances.

Driven by problem-solving research, the envisaged end result is a viable social enterprise that makes it possible to organise informal settlements around tangible material improvements. Once a community realises the benefits of cooperative action, they will have in place social and institutional structures that will make it possible to continue to struggle for further improvements, such as secure land rights and access to subsidies for housing.

Some officials and SDI staff have openly criticised the researchers for crossing the boundary between being researchers and becoming activists. Others have argued that it is possible to be a researcher-activist: that is to use research to articulate alternatives and win ground as an activist. Because this transdisciplinary process is taking place within such a volatile context of highly unequal power relations, the dramaturgy of the process has become a key focus of attention. Compared to the contestations over who stage manages the process and related performances (researchers versus officials versus SDI staff versus community leaders), the script itself is almost irrelevant. Yet it was the technical breakthroughs about alternative infrastructure solutions, derived from interactions with particular groups of shack dwellers, that produced the social effects, including recognition that Nkanini was there to stay. In short, a limited well-managed process has triggered a secondary and much wider drama involving a set of players that have political agendas that may be incompatible with the original vision of community empowerment.

In conclusion, the iShack case demonstrates how researchers can actively engage with communities to co-produce solutions to real-world problems. Yet they are part of wider processes that they can ill afford to be naïve about. By adapting the transdisciplinary research methodology to this specific context, researchers actively perturbed the social fabric of everyday life to demonstrate a viable alternative to the state's top-down approach to in-situ upgrading. Instead of a 'trust and wait' approach, an authentic incrementalist approach emerged. It was only after the model was demonstrated and elaborated that it was possible to formalise a working relationship with the other stakeholders to upscale the model. A transdisciplinary research approach has continued to be implemented with issues such as sanitation and solid waste removal included in the research agenda with equally dramatic social effects. The challenge, however, has been to ensure that researchers and the community-based co-researchers remain reflexive about their roles at all times. Success can breed an arrogance that undermines the humility needed to effectively engage the complex power dynamics that saturate communities like Nkanini. Self-conscious recognition of the power of the script to shape the settings and performances in this particular contexture will determine whether the researchers will be able to continue to work in such an embedded manner in future.

Conclusion

It was argued at the outset that the reflexive approach is interested in contextualising the dramaturgy of deliberative policy processes to reveal the limits of rote practices that result in meaningless formalistic outcomes. Meanwhile, the transdisciplinary approach mounts a normative argument in favour of researchers as co-producers of problem-solving knowledge. While the former focuses on roles and discourse, the latter focuses on methodologies and methods for practically realising co-production. It was suggested that both are needed to understand how researchers engage in real-world policy processes about sustainability-oriented innovations. Those who advocate the transdisciplinary approach need to be more reflexive, and those who argue for reflexivity may need to take more seriously the importance of particular methodologies and methods of actual co-production.

The three cases reveal the degree to which researchers become advocates and how they operationalise knowledge partnerships. The case narratives help to highlight the paradox faced by researchers who are both active performers within particular settings and the designated script writers with a responsibility to analyse and facilitate understanding. Whereas research was merely supportive of pre-determined positions in Case 1, in Cases 2 and 3, in which the outcomes were less clear at the outset, research was able to inform and shape the end result. Case 1 demonstrated a more traditional approach with very limited co-production of knowledge, but with researchers advocating specific policy frameworks that simultaneously served their own institutional interests and put in place a 10-year government commitment to fund what is South Africa's first coherently structured sustainability-oriented research agenda. It is significant that this plan combined earth system science and sustainability transitions perspectives, with the bulk of funding going towards the former.

In Case 2 a significant degree of co-production of knowledge involving public, private and university-based stakeholders took place, with researchers playing less of an advocacy role as a result of strong leadership by government officials. Although private sector stakeholders were engaged, civil society stakeholders were excluded. Sustainability was not initially emphasised, but over time researchers played the key role in revealing the need to broaden the script to include sustainability-oriented innovations with respect to future urban infrastructure investments. The settings and processes of engagement were conducive for learning in this regard.

Case 3 was explicitly motivated by a transdisciplinary co-production approach involving a particular community in which researchers acted as both knowledge producers and as advocates for a particular sustainability-oriented solution. Unlike Cases 1 and 2, government was not initially a participant in the process. Nor was it possible to assume the existence of a formalised setting for engagements between organised stakeholders, because none of these conditions existed prior to the process. However, the impact of the original research results triggered a much wider secondary drama that transformed what was a limited technical intervention into a much wider social mobilisation and institution-building process.

To conclude, further research is needed on the micro-dynamics of the actual science–policy interface. This work should entail reflexive research that analyses the interactions of the actors themselves, paying particular attention to the dynamics of problem identification, knowledge production and problem solving, and the roles played by particular actors as performance changes require changes in the script. This analysis, in turn, will help implementers of the transdisciplinary approach to become more critically aware of their actual roles, impacts and unavoidable biases. These three cases reveal how important it was for researchers to actively engage in policy processes to achieve particular outcomes that may not have been achieved without the learning that research makes possible. But it would be naïve to ignore that these processes are saturated by the dynamics of power, institutional interests and agenda setting by researchers themselves and by key players that researchers can rarely control or counter.

Engagement will always come at a price. The key to balancing the cost is how reflexive researchers will be in analysing their own practices and mistakes as they navigate ever-changing scripts, stagings and performances as they learn to use transdisciplinary methodologies and methods.

Acknowledgements

Research funding from the National Research Funding made this article possible. The assistance of Natalie Meyer is gratefully acknowledged.

References

1. Proceedings of the Berlin Conference on the Human Dimensions of Global Environmental Change; 2012 Oct 5–6; Berlin, Germany. Available from: www.berlinconference.org/2012

2. Hajer M. The politics of environmental discourse: Ecological modernization and the policy process. New York: Oxford University Press; 1995.

3. Hajer M. A frame in the fields: Policymaking and the reinvention of politics. In: Hajer M, Wagenaar M, editors. Deliberative policy analysis: Understanding governance in the network society. Cambridge: Cambridge University Press; 2003. http://dx.doi.org/10.1017/CBO9780511490934.005

4. Hajer M. Setting the stage: A dramaturgy of policy deliberation. Admin Soc. 2005;36(6):624–647. http://dx.doi.org/10.1177/0095399704270586

5. Naicker I. The role of science in issues advocacy: Invasive alien plants in the fynbos vegetation of South Africa [PhD thesis]. Cambridge: University of Cambridge; 2012.

6. Lang D, Wiek A, Bergmann M, Stauffacher M, Martens P, Moll P, et al. Transdisciplinary research in sustainability science: Practice, principles, and challenges. Sust Sci. 2012;7(1 suppl):25–43. http://dx.doi.org/10.1007/s11625-011-0149-x

7. Scholz RW. Environmental literacy in science and society: From knowledge to decisions. Cambridge: Cambridge University Press; 2011.

8. Watson R. Turning science into policy: Challenges and experiences from the science–policy interface. Phil Trans R Soc B. 2005;360(1454):471–477. http://dx.doi.org/10.1098/rstb.2004.1601

9. Burns M, Weaver A. Exploring sustainability science: A southern African perspective. Stellenbosch: Sun Press; 2008.

10. Hadorn GH, Pohl C. Handbook of transdisciplinary research. Dordrecht: Springer; 2008.

11. Hirsch-Hadorn G, Bradley D, Pohl C, Rist S, Wiesmann U. Implications of transdisciplinarity for sustainability research. Ecol Econ. 2006;60:119–128. http://dx.doi.org/10.1016/j.ecolecon.2005.12.002

12. Regeer BJ, Bunders JFG. Knowledge co-creation: Interaction between science and society: A transdisciplinary approach to complex societal issues. Amsterdam: RMNO; 2009.

13. Scholz RW, Tietje O. Embedded case study methods: Integrating quantitative and qualitative knowledge. London: SAGE; 2002.

14. Thompson-Klein J. Prospects for transdisciplinarity. Futures. 2004;36(4):515–526. http://dx.doi.org/10.1016/j.futures.2003.10.007

15. Latour B. Reassembling the social: An introduction to actor-network theory. Oxford: Oxford University Press; 2005.

16. Morin E. Homeland earth. Cresskill, NJ: Hampton Press; 1999.

17. Department of Science and Technology, Republic of South Africa. Innovation towards a knowledge-based economy 2008-2018. Government Printer: Pretoria; 2008.

18. Department of Science and Technology, Republic of South Africa. Global Change Grand Challenge National Research Plan. Government Printer: Pretoria; 2009.

19. Swilling M, Annecke E. Just transitions: Explorations of sustainability in an unfair world. Cape Town and Tokyo: UCT Press & United Nations University Press; 2012.

20. Clark WC, Crutzen PJ, Schellnhuber HJ. Science for global sustainability: Toward a new paradigm. Harvard University John. F. Kennedy School of Government Working Paper Series. Cambridge, MA: Harvard University John. F. Kennedy School of Government; 2005.

21. Rockstrom J, Steffen W, Noone K, Persson A, Chapin FS, Lambin EF, et al. Planetary boundaries: Exploring the safe operating space for humanity. Ecol Soc. 2009;14(2), Art. #32.

22. Steffen WA, Sanderson A, Tyson PD, Jäger J, Matson PA, Moore III B, et al. Global change and the earth system: A planet under pressure. Berlin: Springer; 2004.

23. Fischer-Kowalski M, Haberl H. Socioecological transitions and global change: Trajectories of social metabolism and land use. Cheltenham, UK: Edward Elgar; 2007.

24. Smith A, Voss JP, Grin J. Innovation studies and sustainability transitions: The allure of the multi-level perspective and its challenges. Res Pol. 2010;39(4):435–448. http://dx.doi.org/10.1016/j.respol.2010.01.023

25. Van den Bergh JCM, Truffer B, Kallis G. Environmental innovation and societal transitions: Introduction and overview. Env Inn Soc Tran. 2011;1:1–23. http://dx.doi.org/10.1016/j.eist.2011.04.010

26. Perez C. Technological revolutions and financial capital: The dynamics of bubbles and golden ages. Cheltenham, UK: Elgar; 2002.

27. Hyman K. Economic development, decoupling and urban infrastructure: The role of innovation for an urban transition in Cape Town [Master's thesis]. Stellenbosch: Stellenbosch University; 2010.

28. Swilling M, editor. Sustaining Cape Town. Stellenbosch: SunMedia; 2010.

29. Pieterse E, editor. Counter-currents. Johannesburg: Jacana; 2012.

30. Cartwright A, Parnell S, Oelofse G, Ward S. Climate change at the city scale: Impacts, mitigation and adaptation in Cape Town. London: Earthscan; 2012.

31. Keller A. Conceptualising a sustainable energy solution for in situ informal settlement upgrading [Master's thesis]. Stellenbosch: Stellenbosch University; 2012.

32. Swilling M, Tavener-Smith L, Keller A, Von der Heyde V, Wessels B. Rethinking incremental urbanism: Co-production of incremental informal settlement upgrading strategies. In: Van Donk M, Gorgens T, Cirolia L, editors. Pursuing partnership-based approaches to incremental upgrading in South Africa. Johannesburg: Jacana; forthcoming.

A technique to determine the electromagnetic properties of soil using moisture content

AUTHOR:
Petrus J. Coetzee[1]

AFFILIATION:
[1]Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Petrus Coetzee

EMAIL:
hcoetzee@gew.co.za

POSTAL ADDRESS:
Department of Electrical,
Electronic and Computer
Engineering, University of
Pretoria, Private Bag X20,
Hatfield 0028, South Africa

DATES:
Received: 22 July 2013
Revised: 21 Nov. 2013
Accepted: 24 Nov. 2013

KEYWORDS:
electromagnetic properties; soil moisture; time domain reflectometry; ground constants

HOW TO CITE:
Coetzee PJ. A technique to determine the electromagnetic properties of soil using moisture content. S Afr J Sci. 2014;110(5/6), Art. #2013-0225, 4 pages. http://dx.doi.org/10.1590/sajs.2014/20130225

Accurate electromagnetic ground constants are required for applications such as modelling of ground wave propagation of radio signals and antennas above a real, imperfect earth and for use in geological surveys and agricultural applications. A simple method to determine the ground parameters (conductivity and relative dielectric constant) for any radio frequency is outlined here. The method has been verified over the 2–30-MHz frequency range but should be applicable up to several GHz. First, a low cost, commercial soil moisture meter using time domain reflectometry techniques is used to determine the soil moisture percentage. Then previously published universal soil models implemented on a programmable calculator or a PC are used to calculate the required constants at the frequency of interest according to the measured moisture percentage. The results obtained by this method compare favourably with those obtained by the input impedance of a low horizontal dipole technique. The received signal strength of a ground wave, HF transmission also compares favourably with that predicted by GRWave using ground constants calculated by the soil moisture technique. This method offers significant advantages in terms of simplicity, speed and cost when compared with current techniques.

Introduction

Accurate electromagnetic ground constants are required for many applications including modelling of ground wave propagation of radio signals, calculating clutter and reflection in radar applications, calculating the soil penetration of an electromagnetic wave and modelling antennas above a real, imperfect earth. Electromagnetic ground constants also are widely used in geological surveys and agricultural applications.

Various sources and techniques are available to determine the required constants but these approaches are generally cumbersome, may require specialised instrumentation and may not be applicable to the frequency of interest. The International Telecommunications Union (ITU) published various graphs for relative permittivity (ε_r) and conductivity (σ) in ITU-R Recommendation 527–3.[1] However, the validity and usefulness of these ITU publications are questionable. The Institute of Electrical and Electronics Engineers (IEEE) made the following statement regarding the published data[2]:

> *The International Telecommunications Union (ITU) has published world surface conductivity maps for a number of frequency bands, although these are no longer being updated. The curves of conductivity and relative permittivity in ITU-R Recommendation 527–3 exhibit no dispersion in the band 3–30 MHz, whereas measured values show significant dispersion in the band for which surface soils typically can show characteristics from lossy conductors to lossy dielectrics. The real and imaginary parts of the complex relative permittivity form a Hilbert transform pair. As a result, the conductivity and relative permittivity are not independent variables. Their mutual coupling is described by the Kramers–Kronig relations. Therefore, the ITU values for the HF band are inconsistent with the results of complex variable theory and are in error.*

This statement relating to a professional ITU standard is extraordinary and illustrates the problem with conventional values for soils at radio frequencies. The source of this misconception seems to be a report by Pearce et al.[3] in which they incorrectly concluded that the relative dielectric constant is basically constant from 50 MHz to over 500 MHz. The message is clear: published graphs and tables, even by international bodies like the ITU, are not reliable or accurate enough and a better solution is required. In this paper, work done in obtaining a universal soil model is reviewed and measured data using this model are presented.

Universal soil model

During the 1970s, Conrad L. Longmire and H. Jerry Longley, working for the Mission Research Corporation, developed a universal soil impedance model[4] for the Defence Nuclear Agency. During October 1975, Longmire and Ken S. Smith expanded this model, making it valid from 1 Hz to 10 GHz.[5] This research was conducted to quantify the effect of an electromagnetic pulse generated by a high altitude nuclear explosion coupling into structures and underground cables through the soil.

Electromagnetic pulse is a very fast nanosecond time domain pulse with a frequency spectrum beyond 100 MHz. Longmire modelled soil as a resistor-capacitor (RC)–transmission line with the variation of conductivity and the dielectric constant with frequency as a function of water content. Thus, if one knows the water content of the soil, one can predict with good accuracy what the value of σ and ε_r will be at a specific frequency. Longmire and Smith based their research on work done by Scott[6] in 1971.

Scott[6] has presented results of measurements of the electrical conductivity and dielectric constant, over the frequency range of 100 Hz to 1 MHz, for many samples of soil and rock. He noted that the results for the many samples could be correlated quite well in terms of just one parameter: the water content. By averaging his data, he produced a set of curves, $\varepsilon_r(f)$ and $\sigma(f)$, as functions of frequency (f) for various values of water content. Thus, if one knows the conductivity or the dielectric constant at one frequency, one can estimate both as functions of frequency using Scott's 'universal' curves. If the water content of a soil is known, it is possible to predict what its dielectric constant and conductivity will be at a specific frequency with generally useful accuracy.

Longmire observed that all of Scott's curves for $\varepsilon_r(f)$ would very nearly coincide with each other if displaced to the right or left, that is, that there is just one curve for $\varepsilon_r(f/f_0)$, where f_0 scales with water content. Longmire's contribution was thus to show how to use the frequency-dependent parameters to formulate a time-domain treatment of electromagnetic problems. The time-domain method solved Maxwell's equations in dispersive soils, based on the assumption that each volume element of the soil could be represented by an RC network. The real and imaginary parts of this model are related to the conductivity and the dielectric constant, respectively. A consequence of the RC network model is that the variation of dielectric constant and conductivity with frequency are not independent. It is also clear that the dielectric constant increases with frequency and conductivity decreases with frequency. In terms of the RC network, this means that as the water content is varied, only the R values change, while the C values remain fixed.

Wilkenfeld measured the conductivity and dielectric constant of several samples of grout and concrete over the frequency range 1–200 MHz. From the data published by Scott[6] and Wilkenfeld, Longmire and Smith developed a universal soil impedance (actually, admittance) model that operates from 1 Hz to 10 GHz and includes both Scott's and Wilkenfeld's data. The 10% moisture curves are taken as references and are scaled to the left or right for different moisture values.

The surface wave component of an electromagnetic wave propagates along and is guided by the earth's surface, similar to the way in which an electromagnetic wave is guided along a transmission line. Charges are induced in the ground by the surface wave. These charges travel with the surface wave and create a current in the ground. The ground carrying this current can be represented by a leaky capacitor (a resistance R shunted by a capacitive reactance C). The characteristics of the ground as a conductor can therefore be represented by an equivalent parallel RC circuit, where the ground's conductivity can be simulated with a resistor and the ground's dielectric constant by a capacitor. Figure 3 is a generalised version of the Debye model.[7] At medium-wave frequencies (300 kHz–3 MHz) the soil characteristics are dominated by the resistance, but at higher frequencies the soil is both resistive and capacitive.

In Figure 3, R_0 is the resistance at zero frequency (direct current) and C_∞ is the capacitance at infinite frequency. The other branches provide transient responses with various time constants.

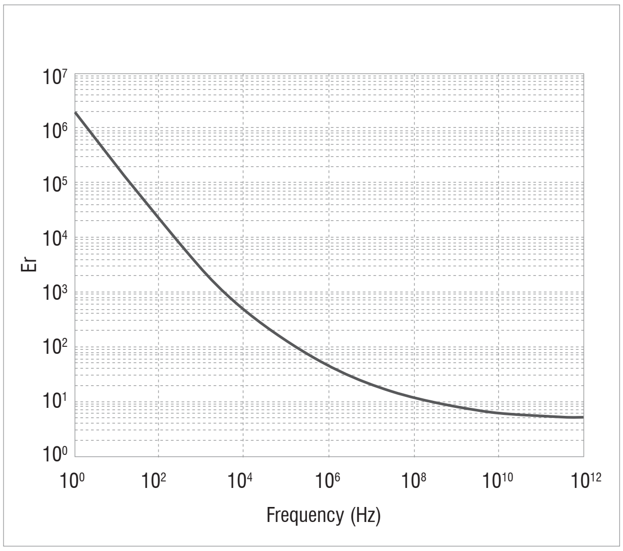


Figure 1: Longmire's[5] universal curve for dielectric constant (ε_r) as a function of frequency for a 10% moisture content.

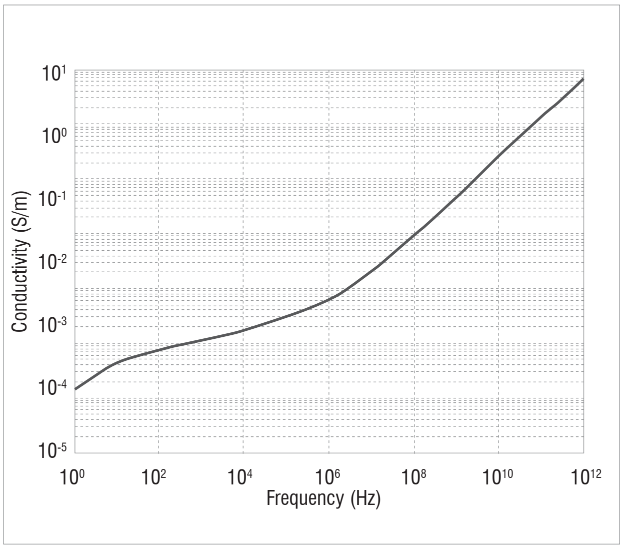


Figure 2: Longmire's[5] universal curve for conductivity (σ) as a function of frequency for a 10% moisture content.

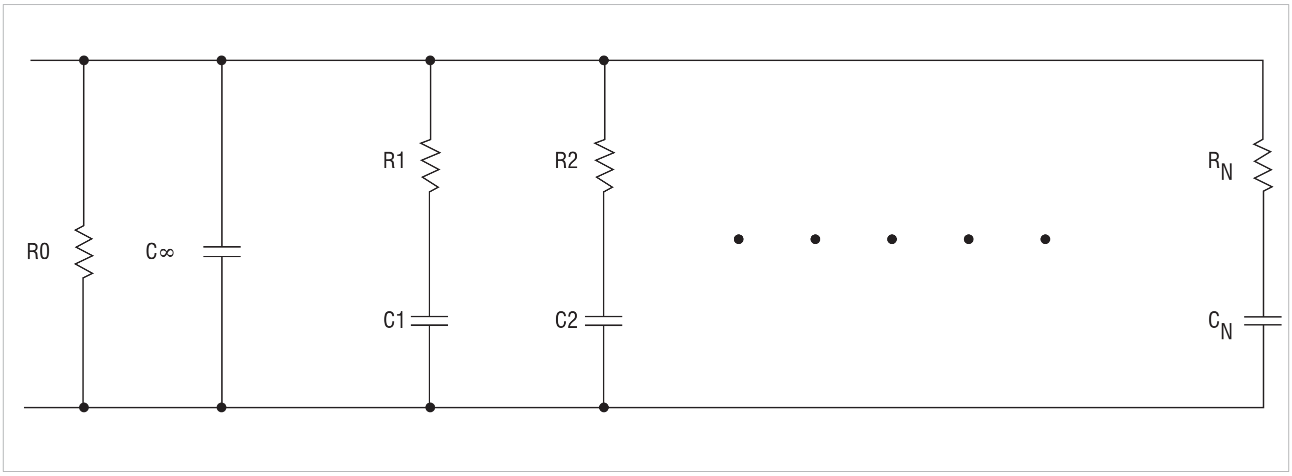


Figure 3: Universal resistor–capacitor (RC) network of the characteristics of ground as a conductor.[5]

Determining the soil moisture content

Various sources and techniques are available to determine the required parameters – relative dielectric constant and ground conductivity – but they all tend to be cumbersome in some way or another. According to the IEEE the main techniques are:

- direct current resistivity methods
- surface impedance methods using very low frequency (up to 20 kHz) signals
- propagation studies in which the receiver is sometimes located underground (also limited to very low frequency)
- wave tilt method
- self-impedance methods
- mutual impedance methods
- time domain reflectometry

The drawback of all these techniques, with the possible exception of time domain reflectometry, is that they cannot easily provide the required soil parameters at a specific frequency.

Time domain reflectometry

Time domain reflectometry (TDR) uses a narrow pulse of electromagnetic energy at one end of a parallel transmission line located in a lossy material. The characteristics of the reflected waveform are influenced by the dielectric properties of the medium. In the 1980s, Clarke Topp, a soil physicist working for Agriculture Canada, was approaching the problem from the time domain direction and, without knowledge of Longmire's work, used TDR to measure soil conductivity and dielectric constant and from there determined the moisture content.[8] Soil physicists have since extensively researched and documented TDR methods to obtain soil moisture content.

Originally used primarily for testing high-speed communication cables, TDR is a complex electronic technology. The early development of TDR for the unusual application of measuring water content in soils began somewhat by chance and continued almost in spite of the goals set by the supporting organisations. TDR was originally used on coaxial transmission lines filled with soils in the laboratory. However, coaxial transmission lines are not practical for measurements in the field and techniques were developed to use parallel transmission lines consisting of two parallel rods placed in the soil. Originally, TDR cable testers were used for measurements in the field and the TDR waveform from the oscilloscope screen was later manually measured using a ruler. Clearly there was a need for a TDR instrument that could measure soil water content directly instead of recording travel time as cable lengths.

With the advent of powerful microprocessors it became possible to perform the entire full waveform signal processing on a single integrated circuit and directly display the soil moisture content on a liquid crystal display. These modern microprocessors are very compact and power efficient, making battery-powered portable equipment possible.

Commercial equipment employing the TDR technique to measure soil moisture content is now freely available and is used to measure the soil water content on golf courses and in the agricultural sector.

It is thus now possible to easily and accurately determine soil moisture content using RF techniques (TDR) and to use the moisture percentage in conjunction with Longmire's soil impedance model to determine the conductivity and dielectric constant for the applicable frequency of interest. This technique is a major breakthrough and promises enhanced accuracy for the modelling of the effect of a real, imperfect ground on electromagnetic signals.

The low horizontal dipole

In free space, a half-wave ($\lambda/2$) dipole antenna has an impedance of very nearly $72 + j0\ \Omega$ at resonance. This input impedance changes as the antenna is brought closer to the surface of an imperfect earth. An antenna can thus be used as a 'geological probe'. Nicol[9] published graphs to determine the electromagnetic parameters of soil (ε_r and σ) from measuring the input impedance of a thin, half-wave dipole at heights of 0.05 λ and 0.02 λ. Note the interdependency of all the variables, including operating frequency, in Nicol's work.

An antenna modelling program (EZNEC)[10] based on numerical electromagnetic code (NEC–2) can be used to calculate the input impedance of a thin dipole close to the ground, according to the defined ground constants. Using the graphs of Longmire and Smith[5] to determine the soil constants for 10% soil moisture at a frequency of 5 MHz yields $\varepsilon_r = 24.61$ and $\sigma = 0.0046$. The wavelength at 5 MHz is nearly 60 m; a half-wave dipole antenna is thus nearly 30 m in length. If the height is taken as 0.02 λ (1.2 m) EZNEC calculates the input impedance at 5 MHz as: $Z = 79.72 + j69.71\ \Omega$. According to the graph published by Nicol[9] the input impedance is approximately $80 + j70\ \Omega$. The values correlate rather well. The values also correlate well when the process is repeated for a dipole at a height of 0.05 λ. It is, however, a bit difficult to read the values accurately from Nicol's graphs. Longmire and Smith's technique is considerably more exact.

Practical application of time domain reflectometry

The received power level of a ground wave, high-frequency (3–30 MHz) signal between Pretoria and the National Antenna Test Range at Paardefontein,[11] north of Pretoria, was measured and compared to the value calculated by the GRWave computer model. GRWave is based on the theory of Rotheram[12,13]. The program was modified for execution on a PC by Dr John Cavanagh of the Naval Surface Warfare Centre in July 1988.[14] Later, the CCIR adopted the program to compute ground wave transmission loss.[15] GRWave is published by the ITU.[16] The program can be used to determine transmission loss and field strength transmission loss from the designated transmitter to the designated receiver. The GRWave model considers a smooth (no terrain obstacles), homogeneous (a single set of ground constants), spherical earth bounded by a troposphere with exponential height variation. GRWave uses three different methods to calculate field strength depending on wavelength (λ), path length (d) and antenna height (h) relative to the earth's radius (a). At longer distances ($d > \lambda^{1/3}a^{2/3}$ and $h < \lambda^{2/3}a^{1/3}$), the residue series is used; at shorter distances ($h > \lambda^{1/3}a^{2/3}$ and $h < \lambda^{2/3}a^{1/3}$), the model employs the extended form of the Sommerfeld[17] flat-earth theory, and geometric optics are used to calculate field strength at distances not covered by either residue series or the Sommerfeld theory ($h > \lambda^{2/3}a^{1/3}$ and d within the radio horizon). GRWave requires frequency, polarisation, power, ground relative dielectric constant and conductivity, lower and higher antenna heights, and distance as inputs.

A 100-W HF transmitter in conjunction with a wideband, monopole antenna with known characteristics operating against a ground screen was used at the Pretoria site. A continuous wave signal was transmitted on the selected frequency under command of the receive site.

A calibrated receiving antenna (Rohde & Schwarz HFH2-Z1, Munich, Germany) and a test receiver (Rohde & Schwarz ESH 3) were used to measure the received field strength (in dBμV/m) at Paardefontein. The measured field strength was converted to received power (dBm) and compared to the value calculated by GRWave for the specified operating conditions.

A Spectrum Technologies Field Scout TDR 300 (Aurora, IL, USA) soil moisture meter was used to determine the moisture percentage at various points between Pretoria and Paardefontein. The average moisture content was 17.3%, from which the conductivity and relative dielectric constant were calculated for the applicable test frequency.

The tests were conducted over a distance of 23.56 km. The measured and the GRWave calculated received power levels are compared in Figure 4. As seen in Figure 4, the measured and calculated results correlate very well. The fact that Paardefontein is located in open countryside with no major mountains between Paardefontein and the test transmitters' location in Pretoria contributed positively to the result. Hilly terrain would probably have a negative impact on the results.

Table 1: Calculated conductivity and relative dielectric constant over the 2–30-MHz range for a 17.3% ground moisture content

Frequency (MHz)	Calculated conductivity (S/m)	Calculated relative dielectric constant (ε_r)
2	0.004	41
3	0.005	34.8
4	0.006	31.4
5	0.007	29.3
6	0.008	27.9
7	0.008	26.9
8	0.009	26
9	0.009	25.3
10	0.010	24.7
11	0.010	24
12	0.011	23.5
13	0.011	23
14	0.011	22.5
15	0.012	22.1
16	0.012	21.6
17	0.013	21.2
18	0.013	20.8
19	0.013	20.4
20	0.014	20.1
21	0.014	19.7
22	0.014	19.4
23	0.015	19.11
24	0.015	18.82
25	0.016	18.56
26	0.016	18.30
27	0.016	18.07
28	0.017	17.84
29	0.017	17.63
30	0.018	17.43

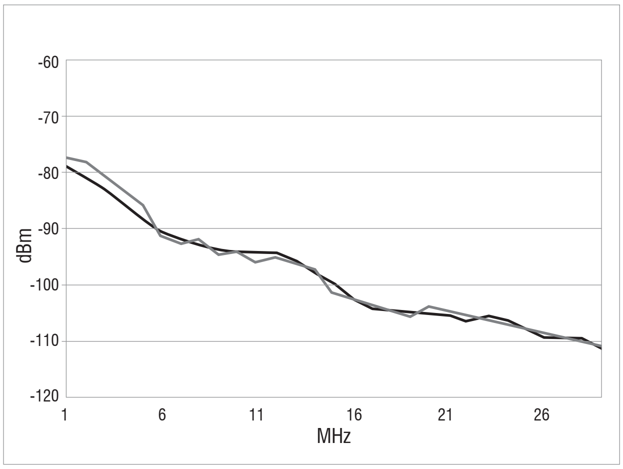


Figure 4: Measured (grey line) and calculated (black line) received power for a high-frequency ground wave signal between Pretoria and Paardefontein.

Conclusion

The universal soil impedance model and moisture percentage technique offers significant advantages in terms of simplicity, speed and cost in determining the electromagnetic properties of soil (ε_r and σ) at any frequency of interest when compared with current techniques.

With the correct electromagnetic ground constants for the applicable frequency, it is now possible to more accurately model ground wave propagation, transmitter area coverage, ground penetration of a RF signal, radiation patterns and input impedances of HF antennas as well as ground reflections at reflective antenna test ranges such as the National Antenna Test Range at Paardefontein.

Acknowledgement

I gratefully acknowledge the support of my colleague Dion F. Janse van Rensburg in the development of the above technique.

References

1. ITU–R. Recommendation P. 527–3: Electrical characteristics of the surface of the earth. ITU–R Recommendations P Series (Radiowave Propagation). Geneva: ITU; 2000.
2. IEEE. Guide for measurements of electromagnetic properties of earth media. IEEE Std 356–2001. New York: Institute of Electrical and Electronics Engineers; 2002.
3. Pearce DC, Hulse WH, Walker JW. The application of the theory of heterogeneous dielectrics to low surface area soil systems. IEEE Trans Geosci Electron. 1973;11:167. http://dx.doi.org/10.1109/TGE.1973.294311
4. Longmire GL, Longley HJ. Time domain treatment of media with frequency-dependent electrical parameters. Theoretical Notes, note 113. Santa Barbara, CA: Mission Research Corporation; 1973.
5. Longmire GL, Smith KS. A universal impedance for soils. Topical report for period 1 July 1975 to 30 Sept. 1975. Report no. MRG-N-214. Santa Barbara, CA: Mission Research Corporation; 1975.
6. Scott JH. Electrical and magnetic properties of rock and soils. Electromagnetic Pulse Theoretical Notes, note 18. AFWL EMP 2–1. April 1971.
7. Debye P. Polar molecules. Mineola, NY: Dover; 1929.
8. Topp GC, Davis JL, Annan AP. Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. Water Resour Res. 1980;16:574–582. http://dx.doi.org/10.1029/WR016i003p00574
9. Nicol JL. The input impedance of horizontal antennas above an imperfect earth. Radio Sci. 1980;15:471–477. http://dx.doi.org/10.1029/RS015i003p00471
10. Lewallen RW. EZNEC v.4.0.39. Beaverton, OR: EZNEC; 2007. Available from: www.eznec.com
11. Armscor Defence Institutes. The South African National Antenna Test Range [homepage on the Internet]. No date [cited 2013 Nov 21]. Available from: www.paardefontein.co.za
12. Rotheram S. Ground wave propagation I: Theory for short distances. Proc IEE Part F. 1981;128:275–284.
13. Rotheram S. Ground wave propagation II: Theory for medium and long distances and reference propagation curves. Proc IEE Part F. 1981;128:285–295.
14. Cavanagh JF. GRWAVE release 2. Washington DC: Naval Surface Warfare Centre; 1985.
15. CCIR Report 714-2. Ground wave propagation in an exponential atmosphere. Geneva: ITU; 1990.
16. ITU-R. rsg3-grwave. Geneva: ITU; 2009. Available from: http://www.itu.int/oth/R0A0400000F/en
17. Sommerfeld AN. The propagation of waves in wireless telegraphy. Ann Physik. 1909;28:665–736. http://dx.doi.org/10.1002/andp.19093330402

Management strategies to curb rhino poaching: Alternative options using a cost–benefit approach

AUTHORS:
Sam M. Ferreira[1,2]
Michèle Pfab[3]
Mike Knight[4,5]

AFFILIATIONS:
[1]Scientific Services, SANParks, Skukuza, South Africa
[2]School of Biological and Conservation Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa
[3]Applied Biodiversity Research, South African National Biodiversity Institute, Pretoria, South Africa
[4]Park Planning & Development, SANParks, Port Elizabeth, South Africa
[5]Centre for African Conservation Ecology, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

CORRESPONDENCE TO:
Sam Ferreira

EMAIL:
sam.ferreira@sanparks.org

POSTAL ADDRESS:
SANParks Scientific Services,
PO Box 401, Skukuza 1350,
South Africa

DATES:
Received: 02 Nov. 2012
Revised: 31 Oct. 2013
Accepted: 25 Nov. 2013

KEYWORDS:
anti-poaching; awareness; CITES; demand–supply; rhino horn trade

HOW TO CITE:
Ferreira SM, Pfab M, Knight M. Management strategies to curb rhino poaching: Alternative options using a cost–benefit approach. S Afr J Sci. 2014;110(5/6), Art. #2012-0055, 8 pages. http://dx.doi.org/10.1590/sajs.2014/20120055

The combination of increasing demand and high black market prices for rhino horn in Asian markets has fueled an escalation in rhino poaching since 2007, particularly in South Africa. This situation has in turn resulted in greatly increased rhino protection costs, loss in confidence by the private sector in rhinos, loss of revenue to conservation authorities and reduced rhino population growth rates. Within current CITES processes, management responses to threats posed by poaching to rhino persistence fall within a mixture of reactive responses of increased protection and law enforcement and some pro-active responses such as demand reduction tactics, along with a parallel call for opening a legal trade in horn. These rhino management strategies carry different risks and benefits in meeting several conservation objectives. An expert-based risk–benefit analysis of five different rhino management strategies was undertaken to assess their potential for delivering upon agreed rhino conservation objectives. The outcomes indicated that benefits may exceed risks for those strategies that in some or other format legally provided horn for meeting demand. Expert risk–benefit approaches are suggested to offer a rational, inclusive and consensus generating means of addressing complex issues such as rhino poaching and augmenting the information used within the CITES decision-making processes.

Introduction

Poaching of Africa's rhinos has escalated exponentially from an average loss of 0.17 rhinos per day (a total of 62 rhinos) in 2007, to 2.04 rhinos per day (a total of 745 rhinos) in 2012.[1] This escalation has raised concerns amongst conservationists about the long-term survival of the species.[2] South Africa, with 82% (or 20 954 rhinos) of the continent's rhino population, has been most affected by poaching, losing 1805 (or 75%) of the 2387 rhino poached since 2006.[3] Of particular concern is the 1.75-fold ($\underline{+}$0.64 s.d.; $n=5$) increase in the annual rate of poaching, which accounts for 3.2% of the South African rhino population size in 2012. Although this loss is currently sustainable, it is predicted that South Africa's rhino population will start to decline by 2015–2016[1,4] if the increasing rate of poaching is not quickly addressed. Kenya and Zimbabwe lost 2.2% and 4.1%, respectively, of their rhino populations in 2012; Namibia was much lower amongst the four major African range states, with a loss of 0.04%.[2] The eight remaining minor rhino range states, which collectively conserve about 500 rhinos, had losses in 2012 ranging from 0% in Swaziland to 100% in Mozambique,[3] indicative that poaching is affecting the entire African continent.

The rapid rise in rhino poaching has been driven by an exponential increase in the illegal demand and black market price for rhino horn in south east Asia, especially Vietnam and China.[5,6] This increased demand for horn has not only come from the traditional Chinese medicine users, but has also been brought about by anecdotes of the unproven cancer reducing properties of rhino horn together with its newly found status symbol and general entrepreneurial uses, all supported by thriving regional economies with a higher disposable income than previously.[2,7] The inelastic relationship between the increasing demand and restricted supply influences the high black market prices for rhino horn,[8] making the product attractive to criminals and organised crime syndicates.[7,9] Increasing involvement by syndicated organised crime can have greater degrading effects on society at large.[10] Furthermore, it has been argued that trade bans, such as that over the sale of rhino horn,[11] exasperate the situation, driving up the black market prices for rhino horn even further and increasing pressure on wildlife populations.[12,13]

With the increasing value of rhinos, especially their horns, protection costs have soared, making rhinos a liability to state conservation authorities, private and communal landowners alike.[14] The private sector, which owns 24% of the South African rhino population on a further 2 million hectares of land, plays an integral role in conservation of the species[15] and wildlife habitat. No longer are the benefit streams from tourism, limited trophy hunting and live sales of rhinos sufficient to offset increased security costs for rhinos (especially in South Africa), and some private rhino owners are opting out of rhino conservation.[1,12] This situation is of major concern to rhino conservationists as it will lead to a lower carrying capacity for surplus rhinos, a reduction in the population growth, reduced essential revenue for the conservation authorities and a general devaluing of the important wildlife industry.[1,16]

Responses to escalating rhino poaching range from traditional increased law enforcement and protection (including conservation buffer zones) and demand reduction approaches[17,18] (such as targeted Asian awareness campaigns) to those advocating a regulated legal trade in horn.[13,14] If anything, these seemingly opposing strategies have tended to polarise the rhino debate,[19] with the pros and cons of alternative strategies in a logical, consensus building framework remaining unexplored. Approaches such as participatory risk–benefit analyses may facilitate consensus decisions and have been advocated as a way to evaluate various management strategies directed at curbing rhino poaching.[17]

The challenge is that these evaluations cannot be made using only biological information; there is a need to move beyond traditional debates and decisions[11,20] that have to date only considered two strategies: (1) no trade in raw rhino horn with an associated intense law enforcement campaign and (2) unrestricted trade in raw rhino horn. We report on the outcome of a workshop involving a variety of experts to collectively identify and evaluate alternative strategies, focusing on rhino horn as a commodity, and making use of a basic conceptual model of drivers influencing demand and supply.[17] The aims of the exercise were to (1) achieve a consensus understanding regarding potential drivers of rhino poaching by considering common economic theories and opinions, (2) identify a suite of alternative management strategies, irrespective of present legal constraints, by collating existing proposals and adapting or proposing new ones, (3) evaluate the risks and benefits of each management strategy for rhino persistence within South Africa, Africa and Asia, as well as for other conservation values, economic values and societal expectations and (4) recommend consensus[21] best-practice management strategies.

Material and methods

Workshop participation

A total of 45 experts were invited, of whom 30 participated. The participants had expert interest and experience across a broad spectrum of fields including traditional (n=1) and resource economics (3), law (2), enforcement and compliance (9), conservation science (11) and ethics (4). In addition, the attendees had a common interest in rhinos and were representative of various value systems associated with conservation (10), animal welfare (2), animal rights (2), national (7) and provincial government (4) and private rhino ownership (5). We provided a brief overview on rhino conservation status[1] as well as an introduction to the requirements for innovative thinking,[17] given that present approaches have apparently had limited success in curbing the incentives for rhino poaching.[4] To ensure that we had a wide spectrum of viewpoints represented, we asked participants to express their expectations from the workshop and categorised these into 19 categories. The authors remained independent in facilitating the process of the assessment.

Developing understanding of poaching drivers

Following a participatory objective setting process,[22] participants collectively agreed on a list of rhino conservation objectives, along with their expected challenges and costs. Differential effects can only be evaluated within a common understanding of how various factors may interact and influence rhino poaching.

Participants were introduced to risk–benefit approaches for evaluating various management strategies[23,24] (Box 1). Traditional risk methodology[23] focuses on describing all events or outcomes associated with a strategy as risks, whether these have positive or negative consequences. This approach is challenging in participatory workshop processes and discussions tend to provide clarity when participants have a reference framework of risks and benefits. Our approach thus accommodated this aspect. The group agreed to use the integrated framework approach[17] in understanding the complexities associated with the relationship between the supply and demand for rhino horn, and aimed to develop a common understanding of how this relationship potentially influenced the price of horn and the incentive to poach rhinos.

In order to potentially meet the agreed rhino conservation objectives, participants proposed a number of alternative management strategies. These strategies were grouped into those that primarily attempted to reduce the demand for rhino horn, affect the supply of rhino horn or both. This approach allowed strategically aligned strategies to be identified and grouped and then evaluated through a risk–benefit analysis.

Risk–benefit analysis

Potential outcomes or developments (such as an increase in poaching or an increase in the number of populations) associated with each objective were identified by participants in a participatory manner as being either a risk or benefit in delivering on the objectives. Each outcome or development was assessed in a spreadsheet model in terms of its possible impact, likelihood of occurrence and certainty of happening, following the scoring shown in Table 1. The 30 participants collectively listed various outcomes or developments for a specific scenario and then, following a discussion, a consensus was reached and scores assigned. Impact relates to the extent of an outcome's perceived effect on the objective. Likelihood provides a scoring for the possibility that the outcome or development will occur, while the certainty provides an indication of the confidence of it actually occurring. Risks were assessed in relation to the negative outcomes, while benefits focused on the positive outcomes. Participants defined the relative importance of each objective, used in weighting components, at the end of the risk–benefit analyses of the various management strategies following the same collective discussion approach as for scoring risks and benefits. As part of the analyses, each strategy was assessed for its logistical costs (challenges) and benefits (opportunities). In addition, the relative financial resources required for or generated in delivering on a specific strategy were also estimated by the group (see Box 1 for details).

Box 1: Risk–benefit analyses

Different rhino conservation objectives (i) may carry different importance (w_i) for different stakeholders and experts. An event or outcome that occurs in association with the implementation of a specific strategy may carry risks (r), benefits (b) or both. Risks for event j are defined as the product of the impact or effect ($e_{r,j,i}$) it will have on objective i and the likelihood ($p_{r,j,i}$) that event j will actually realise.[23] Similarly, benefits associated with event j are defined as the product of the impact or effect ($e_{b,j,i}$) it will have on objective i and the likelihood ($p_{b,j,i}$) that event j will actually realise. The total consequence for objective i of an event $_j$ is scaled by the importance of objective i. For risks and benefits, that is, $w_i e_{r,j,i}$ and $w_i e_{b,j,i}$, respectively.

The overall consequences of the risks and benefits on several events is the average consequence of events j on objective i defined as

- $Risk = w_i \sum_{n=1}^{n=j} \frac{e_{r,i,j}\, Pr,i,j}{nj}$
- $Benefit = w_i \sum_{n=1}^{n=j} \frac{e_{b,i,j}\, Pb,i,j}{nj}$

The complete risk–benefit profile associated with events j influencing objective i then reduces to an estimate k_i where

- $Kj = w_i \sum_{n=1}^{n=j} \frac{e_{b,i,j}\, Pb,i,j}{nj} - w_i \sum_{n=1}^{n=j} \frac{e_{r,i,j}\, Pr,i,j}{nj}$

Operational elements (z_a), usually comprising costs and logistics, use a similar structure

- Costs: $Z_a = w_a \sum_{n=1}^{n=j} \frac{e_{g,a,j}\, Pg,a,j}{nj} - w_a \sum_{n=1}^{n=j} \frac{e_{l,a,j}\, Pl,a,j}{nj}$ with g referring to events that lead to gains and/or to those that lead to losses.
- Logistics: $Z_a = w_a \sum_{n=1}^{n=j} \frac{e_{o,a,j}\, Po,a,j}{nj} - w_a \sum_{n=1}^{n=j} \frac{e_{c,a,j}\, Pc,a,j}{nj}$ with o referring to events that provide opportunities and c to those that lead to challenges.

The complete risk–benefit-logistic–cost profile of a management strategy collapses to

- $Q_m = \frac{\sum_{n=1}^{n=k} k_j}{n_k} + \frac{\sum_{n=1}^{n=a} z_a}{n_a}$

Strategies are prioritised so that $Q_1 > Q_2 > Q_3 > Q_4 > \ldots\ldots Q_m$.

Table 1. Scores used by participants in the risk–benefit analyses of various management strategies

Score assigned	Impact	Likelihood	Certainty	Value	Availability
1	Very low	Negligible	Uncertain	Very low	Very low
2	Low	Unlikely	Some uncertainty	Low	Low
3	High	Likely	Some certainty	High	High
4	Very high	Definite	Certain	Very high	Very high

Results

Workshop participation and expectations

Participants ($n=30$) had a broad diversity of expectations in relation to addressing current rhino management issues. Importantly, 37% of participants appreciated the need to explore alternative rhino conservation strategies, with 24% advocating the need for an integrated approach (Figure 1).

Drivers of rhino poaching incentives

The integrated framework[17] and supply–demand relationships that influence the price of horn provided an understanding of the main drivers affecting the relationship between the demand and supply of rhino horn. These relationships indicated the complex nature that different management strategies may potentially have in affecting the price of horn and the incentive to poach rhinos. The central preposition of this framework was that the demand for horn is driven primarily by traditional and new uses, while the supply of horn is affected by attempts to restrict (or eliminate) or provide horn via legal means. It was assumed that if supply is increased, a reduction in price may result and the incentive to poach should decrease. However, the nature of that relationship may differ substantially depending on the specific use, the consumer country being considered and the elasticity of the demand for horn. Participants expected that, at least in some consumer countries, an increase in the supply of rhino horn should lead to a substantial decline in the price of that commodity. Incentives to poach rhinos were also expected to increase in a non-linear positive relationship between the price of horn and the risk of detection. This risk which acts as a disinvestment to poach consists of two elements: the fear of being detected and arrested, and the magnitude of the punishment.[25] It was suggested that critical thresholds would exist in this relationship, which apply to increased protection, increased judicial sentences, shoot-to-kill protection, increased demand reduction strategies and provision of horn (Figure 2). When the incentives created by the increasing price of horn outweigh disincentives to poach, an escalation in poaching results and the need to change approaches arises.[26]

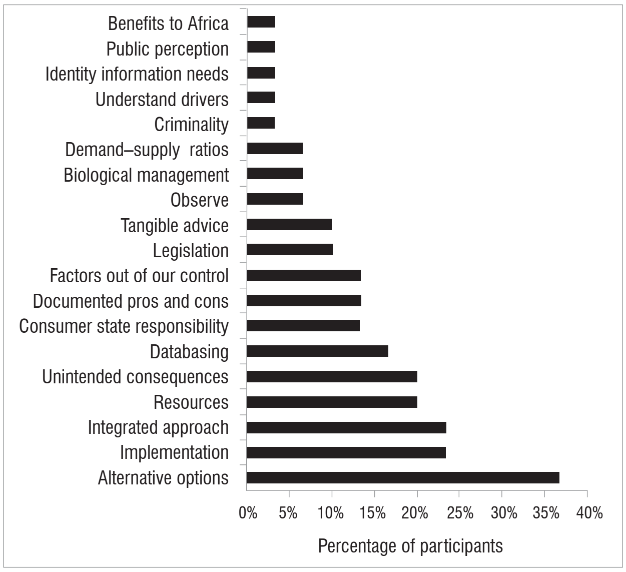


Figure 1. A summary of the expectations and issues that the 30 participants expressed at the workshop, some of which fell into more than one category.

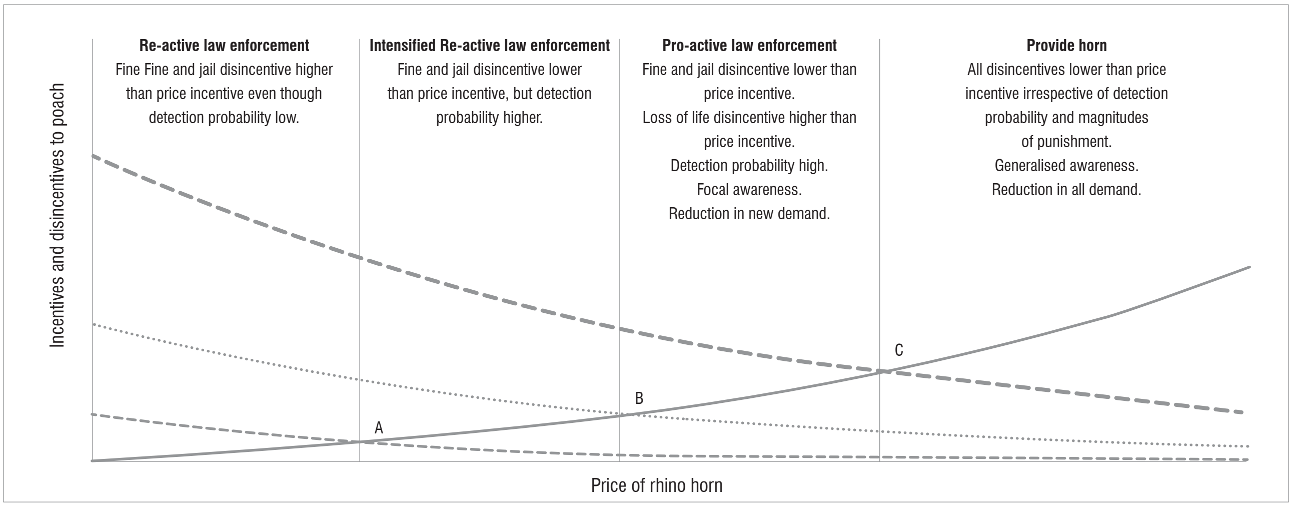


Figure 2. Expected relationships between incentives as well as disincentives to poach and the price of rhino horn. Participants expected non-linear increases in poaching incentives (solid line) with an increase in the price of rhino horn. Disincentives to poach varied – re-active law enforcement (lower dashed line) as was the case until recently, relied on jail sentences and fines as sufficient deterrents to poach. When incentives to poach exceed re-active law enforcement disincentives (A), then intensified re-active law enforcement (dotted line) increases detection probabilities and serves as a disincentive. When price increases to such an extent that incentives to poach exceed intensified re-active law enforcement disincentives (B), authorities may impose pro-active law enforcement (upper dashed line) with high probabilities of detection and additional risks such as loss of life as sufficient deterrents to poach. Such activities can be complimented by focal awareness programmes directed at reducing demand. A significant critical threshold is reached when price is so high that the incentive to poach exceeds all disincentives (C). At this threshold, authorities are best placed to change tack completely and provide rhino horn in parallel with aggressive awareness campaigns.

Objectives and alternative management strategies

Participants identified six objectives that should be achieved through management strategies (Table 2). Four of these objectives have strong conservation outcomes; one gives recognition to economic values associated with rhinos from the direct and indirect values attributed to both the horn and live rhino markets; and the last considers the values of stakeholders both in range as well as in consumer states. The group weighted the objectives with regard to their perceived relative importance.

Table 2. Objectives and their importance weighting of different management strategies

Importance	Objective
Essential	Conservation of rhinos in South Africa, including their population size and range. Conservation of other biodiversity components associated with the protected areas/properties rhinos occur on.
Most important	Conservation of rhinos elsewhere in Africa, including their population size and range. Sustained direct and indirect economic values of both African rhino species.
Important	Conservation of rhino in Asia, including their population size and range.
Some importance	Expectations of and benefits to stakeholders within range as well as consumer states.

A total of 17 possible management strategies for curbing rhino poaching in South Africa were identified (Table 3). These possibilities were grouped into three strategic themes that focused on reducing supply, reducing demand and increasing supply. These strategies were then further grouped into eight strategic management categories. The consumer response category focused on increased diplomatic pressure on the consumer states as well as increased domestic trade restrictions, while the international awareness category addressed demand reduction.

A total of six suggested detailed strategies were more focused on restricting the supply of horn to the illegal market. Half of these were more directed at providing for disincentives to criminal involvement in the trade in horn, such as through enhanced law enforcement, while the remaining three – such as the creation of rhino horn alternatives or dehorning animals – were considered indirect approaches aimed at restricting the supply of horn.

Eight detailed strategies were focused on increasing the supply of rhino horn to the market, through direct donations, provision of live rhinos, and restricted or unrestricted trade in rhino horn.

A total of five different detailed strategies from across the spectrum of management categories were selected for the analysis. Each strategy, along with the mechanism through which it was thought to work, is described below.

1. **Direct disincentives to poaching through increased local law enforcement (the status quo)**: This strategy is maintaining the current status quo in which the ban on the trade in horn nationally and internationally is retained and for which there are slightly improved law enforcement and anti-poaching efforts and associated intelligence gathering in the range states. The primary focus of this strategy is to curb poaching through disrupting criminal syndicates and providing direct disincentives for poaching. This approach is the basis for numerous current fundraising initiatives. The strategy is suggested to reduce the availability of rhino horn through local law enforcement and intense anti-poaching programmes. It aims to discourage poaching by increasing the risk of being arrested and prosecuted. With a restriction on supply of horn, but no concomitant reduction in demand in the consumer states, the price of rhino horn is expected to rise, increasing further poaching pressure.
2. **Increased international awareness (demand reduction)**: This strategy is the same as the above status quo situation, but includes an intensification of awareness and government law enforcement interventions in consumer states to reduce the use of rhino horn. This approach also targets conduit states through diplomatic awareness. Substitution with alternative products and increased enforcement in consumer states may form part of this approach,

Table 3. Categorisation of the 17 proposed alternative detailed rhino management strategies

Strategic theme	Strategic management category	Detailed management strategy
Reduce demand	Consumer state responses	Diplomatic pressure and legal actions Strict domestic measures in consumer states
	International awareness	**Consumer state public awareness programmes**
Reduce supply	Indirect disincentives	Creating and providing rhino horn substitutes Breeding or surgical creation of hornless rhinos Dehorning rhinos
	Direct disincentives	**Law enforcement and compliance – status quo** Destroy all stockpiles No hunting and no national trade in live rhino and rhino horn
Increase supply	Horn stock donation	Buy all horn stock and donate to consumer states
	Provision of live rhinos	Trade in live animals to establish out of range populations for horn harvesting **Lease of live animals to establish out of range populations for horn harvesting**
	Restricted trade	**Trade in horn nationally** Medicinal horns traded nationally in powdered or whole form Medicinal horns traded internationally in powdered or whole form
	Unrestricted trade of horn	**Trade in horn internationally from stockpiles and natural mortality** **Trade in horn internationally from harvested horn**

Strategies evaluated in the cost–benefit analysis are shown in bold.

but were considered to be relatively minor components. This strategy was suggested to reduce the demand for rhino horn, and along with a shrinkage in supply, there would be a reduction in the price of horn and a subsequent reduction in the incentive to poach. There was some uncertainty as to whether differential rates in the shrinkage of supply and reduction in demand may affect the price of horn and the incentive to poach. This strategy is currently being widely advocated.

3. **Provision of live rhinos to consumer states to breed for horn**: This strategy involves providing live rhinos to consumer states through international trade. Harvesting of horns from these ex-situ rhino populations would be permitted. It also includes the status quo scenario plus a ban on both domestic and international trade in rhino horn in range states. It envisages the trade being compliant with current CITES (the Convention on International Trade in Endangered Species of Wild Fauna and Flora) restrictions, but with consumer states allowing domestic trade in horn products derived from local harvesting. The strategy proposed here assumes a leasing agreement for the rhinos (i.e. involves a financial return to the lessor) and benefit sharing (50:50) between range and consumer states. To avoid genetic contamination of South African rhino populations, no progeny would be allowed to return to South Africa. Effectively this strategy provides for parts of the rhino horn market. With a greater reliable supply of horn to the market, the price is expected to decline, thus reducing the incentive to poach. The associated harvesting of horn may not be able to provide for certain markets because of cultural preference of horn originating from free-ranging rhino, thereby limiting impact on demand–supply dynamics. This strategy also carries considerable uncertainty regarding the possible effects of achieving the conservation objectives.

4. **Restricted trade in rhino horn**: In this scenario the national moratorium on the trade in horn in South Africa would be lifted, allowing only for domestic trade. This strategy is expected to lead to the stockpiling of rhino horn. Participants essentially considered this strategy as a stepping stone to international trade. The strategy envisages a well-regulated and controlled internal trade with appropriate database management and reporting being in place. Effectively this strategy provides for parts of the rhino horn market, albeit only locally in South Africa. Sales and expansion of the rhino range in South Africa were expected. This strategy is anticipated to have limited positive impact on the global demand–supply dynamics, and thus limited influence on the price of rhino horn and hence incentives to reduce poaching.

5. **Unrestricted trade in rhino horn**: This strategy allows for the international trade in rhino horn. The approach includes the situation described in the status quo strategy above, plus a well-regulated domestic trade in rhino horn within South Africa along with the required law enforcement and compliance mechanisms. Harvested horn from dehorned rhino plus stockpiled horn from natural deaths in both private and state populations would be allowed to be sold. The approach requires a legitimate trading partner, as well as compliance with CITES and international systems for tracking and monitoring of the rhino horn to reduce laundering of illegally obtained horn. A central selling organisation was advocated as the trading mechanism based on free market principles with certified buyers. Effectively this strategy provides for all components of the rhino horn market. It is envisaged that the demand–supply ratio should lead to a drop in the price of horn and a reduction in the incentives for poaching. There was uncertainty around the reduction in the price of horn possibly stimulating further demand from a growing, wealthier Asian middle class, thus maintaining demand and poaching incentives.

Risk–benefit analyses

The potential mechanisms by which the six conservation objectives are met and, in turn, affect the demand and supply of rhino horn to the market for each of the five management strategies are described in Table 4. The descriptions include each strategy's logistical challenges and opportunities, and relative financial costs and revenue generation

Table 4. Summary of the risks and benefits associated with five management strategies to curb rhino poaching in South Africa

Strategy	Detailed risk–benefit analysis
Direct disincentives (status quo)	Overall, risks were dominated by an expected general degradation of all conservation values with some measure of certainty. Negative consequences for economic values were highlighted, but participants were relatively uncertain about their impacts and the likelihood of them materialising. The strategy, however, carries some certainty regarding risks to the public perception and conservation reputation, including some costly logistical challenges. Overall, risks substantially exceeded benefits (Figure 3).
International awareness (demand reduction)	It was acknowledged with some uncertainty that declines in poaching rates and enhanced conservation effects may result from this strategy. Degradation of conservation values were thought to be more likely, but also with considerable uncertainty. There was also some uncertainty about what the consequences might be for the economic value of rhinos. This strategy carried some logistical challenges with considerable costs, but benefits to South Africa's reputation were anticipated. Overall, the strategy carried more risks than benefits (Figure 3).
Provision of live animals to consumer states to breed for horn	Direct poaching effects on rhinos may diminish, but indirect effects were identified as having potentially high negative impacts on the conservation values of rhinos in South Africa. There was general uncertainty about other conservation consequences, although some were perceived to be beneficial. The effect on the value of live rhinos was thought to be high in the short term, and to diminish over time. It was anticipated, although with some uncertainty, that there may be risks associated with animal welfare, ethics and South Africa's conservation reputation, but that benefits would more than likely accrue from consumer states being appreciative of the recognition of their traditional values. A large number of logistical challenges were envisaged. It was anticipated that revenue could be generated, creating further opportunities. This option carried more perceived benefits than risks (Figure 3).
Restricted trade in horn	Several consequences associated with horn stockpiling were identified. It was imagined that rhino owners may want to have the option to sell their rhino horn stockpiles, because at present holding rhino horn is a security risk. It was envisaged that this management option would open the possibility to locally perfect the trading mechanism with its checks and balances for later roll-out to the international market. Generally the risk to conservation objectives remained, with much uncertainty and additional expectations of high logistical challenges associated with innovative criminal activities and stockpile management. Some opportunities associated with a number of anticipated conservation incentives were noted, in addition to the increased economic value of live rhinos through hunting opportunities. Overall this approach carried nearly equal risks and benefits (Figure 3).
Unrestricted trade in rhino horn	Because of an anticipated reduction in poaching incentives, several benefits were identified for conservation and economic value objectives. These assessments, however, only carried some certainty. Enhancement of South Africa's conservation reputation was anticipated, even though there were some risks associated with South Africa's support of medicinal uses that may have limited value. The strategy was perceived to carry considerable challenges associated with establishing a legitimate trading partner, regulated trade procedures, as well as high costs associated with lobbying internationally to achieve CITES compliance, although these costs were anticipated to be offset by increased financial gains. Generally this strategy had substantially more benefits than risks (Figure 3).

potentials. A comparison of the five alternative management strategies (Figure 3) suggests that rhino poaching may be best addressed by management strategies that generate benefits at least equal to or higher than the associated risks involved in the supply of horn to the market. Overall, unrestricted international trade in rhino horn produced the best risk–benefit score, while the worst case scenario, in which risks substantially exceeded benefits, was provided by the status quo strategy.

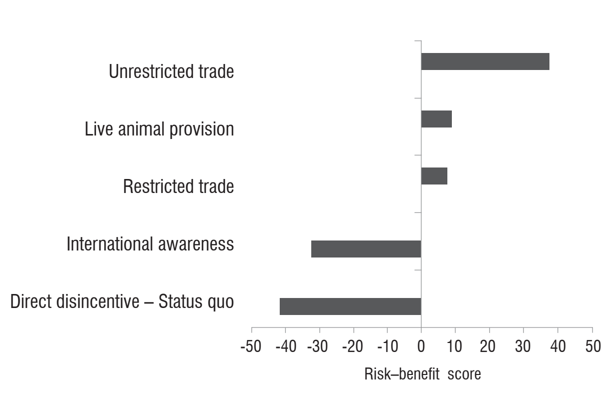


Figure 3. Relative comparison of risks and benefits for five evaluated management strategies to curb rhino poaching in South Africa. Negative scores are indicative of risks outweighing benefits, while positive scores indicate strategies in which benefits outweigh risks.

Discussion

Despite substantial efforts to curb rhino poaching, rhinos continue to be poached for their highly priced horns.[1] Alternative strategies thus need to be seriously considered to ensure the persistence of Africa's rhinos. We made use of an expert-based risk–benefit analysis in an attempt to tease out the cost–benefit relationships influenced by the main drivers of rhino poaching for horn in this complicated and complex international resource use industry. Moral complexity and factual uncertainty often derail such debates.[27]

A large part of factual uncertainty stems from opinions, expectations and assumptions not robustly tested by appropriate information. Curbing rhino poaching epitomises such challenges. In the absence of factual certainty, several approaches are used, most notably that of adaptive management.[28] In such instances, adaptive management proceeds with some opinion about how a system may be working and implements responses followed by processes allowing learning by doing. The opinions that workshop participants had about responses to curb rhino poaching were translated into predictions through discussions on what the outcomes might be. These discussions reflected many management strategies or options to meet desired outcomes,[29] as noted in the wide variety of expectations and options discussed by the workshop participants. The risk–benefit approach thus benefitted from this diversity of opinions towards the agreed central outcome of seeking solutions to curb the escalation in rhino poaching and the need for innovative thinking.

Although a potentially large number of management strategies were proposed, the five consolidated strategies that were assessed provided a diversity of approaches to achieving the agreed objectives. In addition, the risk–benefit approach had the value of enticing expert participants to collectively debate and decide upon various issues associated with each strategy,[30] thus drawing away from the limited polarised trade/no trade paradigm that has tended to dominate most previous deliberations.[14,19] Our results suggest that there were more benefits than risks associated with strategies that increase the supply of horn, but this outcome may be constrained by the limited expertise within the workshop on Asian culture and markets.

CITES processes are cumbersome at best[7] and often ignore ecological realities.[31] Although the provisions of CITES[7] are invaluable for achieving sustainable international trade, they are vulnerable to lobbying tactics that distract focus from the conservation and sustainable use agenda. Given the international scope of the illegal trade in rhino horn,[5,6] and the dichotomy of conservation philosophies, it would be of value to repeat the risk–benefit approach used in the present study with other range states, and also with the Asian consumer states. The latter would be important in elucidating the understanding of local Asian trade in rhino horn and the international implications of the illegal trade. The importance of the risk–benefit approach would be its non-confrontational and consensus building nature that allows for a common understanding of the details and socio-ecological-economic linkages.[32] We envisage that such an approach could draw consumer states in as potential partners for finding a solution to the rhino poaching problem. Risk–benefit approaches may thus provide a useful basis for a participatory consultation process to inform CITES processes and decisions.

As many of the decisions dealing with the cause of rhino poaching rely upon an understanding of the economic drivers and processes, any risk–benefit analysis involving all range and consumer states must have a well-informed resource/economics presence in addition to other knowledge areas such as conservation, private rhino ownership, animal ethics, animal welfare, animal rights, law enforcement, Chinese traditional medicinal experts, market analysts, jurisprudence and diplomatic representation. It is clear from the South African workshop that the broader the base of expertise, the more informed the outcome.

Disregarding these shortcomings, the conceptual framework[17] adopted by participants highlighted three important realisations. Firstly, curbing rhino poaching requires a variety of responses that may have different levels of effectiveness depending on the incentive/disincentive interaction at a specific time. Rhino poaching is an illegal activity driven by the complex interaction between risks and rewards of committing a crime.[25] Incentives driven by the international price of rhino horn are relatively simple and tend to increase exponentially as the price of the commodity increases. However, the impacts of disincentives remain more elusive given their broad scope and the variability in their effectiveness at different prices of the commodity. A key challenge arises when the price of rhino horn reaches such high values that incentives outweigh all disincentives. Reactive responses such as enhanced law enforcement and dehorning focus on the symptoms of the problem and will be ineffective, as evidenced by the rhino poaching surge observed during recent years in South Africa.[1] In essence, more of the same law enforcement will be in vain. Conservation authorities must be adaptive and switch to more wide reaching solutions that focus on the cause of the problem, such as the demand for horn. Doing so does not mean that basic law enforcement and protection should be discarded.

Secondly, not all tools and strategies are available to conservationists to address the rhino poaching problem, which provides conservation authorities with particular challenges. The provision of rhino horn through international trade is not feasible in the foreseeable future given lengthy CITES processes.[7] An alternative response focusing on the cause of the problem is to consider strategies of providing horn that are not trade related,[4] and thus not constrained by CITES processes. The possible positive effects of this response would be negated given the ban on domestic trade in rhino horn in consumer states.[33]

The third realisation is that if rhino range states are to be effective in conserving their rhinos, they need to be decisive and flexible in changing their management approaches. In the short term, authorities are forced to be reactive and focus on the symptoms of rhino poaching in their management responses, which are progressively more demanding, expensive, technically advanced and complex, in an attempt to increase the risk to would-be poachers, a known significant deterrent for crime[25] below certain threshold prices for the commodity. The opportunity costs of redirecting such resources away from other conservation activities can be detrimental to other conservation outcomes. Collapsing (or undermining) organised crime links is part of this complexity, while focal awareness campaigns, particularly in Vietnam where new demands have reportedly surfaced,[5] can disrupt the exponential cascade of continual increases in rhino horn retail prices. Such disruption of the horn supply chains could provide important respite and allow time to fully explore other broad-based systemic strategies and ways to provide rhino horn when price incentives outweigh all disincentives.

Being mindful that our results are dependent upon subjective assessments and understanding, as well as the persuasive powers of participants, the assessment is only indicative; yet it provides important insights to inform an adaptive management approach to addressing this subject. An important product of this process was the identification of potential information gaps, especially those with potentially major impacts for which there were a high degree of uncertainty. Some of these included the need to: (1) identify the potential opportunity costs to conservation associated with an increase in the supply and/or demand reduction strategies; (2) understand the potential threats and the magnitude of these threats associated with implementing different rhino management strategies on South Africa's reputation; (3) understand the animal rights and welfare issues with regard to live animal trade strategies; (4) understand the international and national legal implications of the different strategies; (5) assess the potential impact that increased supply and demand reduction may have on the potential to increase demand in consumer states; (6) assess the potential impact of local rhino management policies and actions on regional and continental rhino populations; and (7) understand detailed market linkages and drivers of rhino horn consumption in consumer states.

Conclusions

Curbing rhino poaching requires integrated and flexible approaches. A restriction in the availability of strategies constrains authorities to respond effectively to the underlying causes of rhino poaching. Conservation authorities are thus forced to engage in progressively more aggressive and costly law enforcement activities to the detriment of other conservation values. Risks and costs of the present status quo management strategy substantially outweighed any benefits and any additional measures to enhance this strategy should be approached with caution. Even the simultaneous application of demand reduction strategies may not reduce the incentive enough to reduce the threat of poaching. South African conservation authorities will remain compromised if short-term pro-active law enforcement activities fail to disrupt organised crime syndicates to reduce poaching and if alternative management options to supply rhino horn within CITES processes do not become readily available.

Acknowledgements

We thank Antoinette Botha (South African National Biodiversity Institute) who very competently made all the logistical arrangements for the expert workshop at short notice. We also thank all the workshop participants.

Authors' contributions

S.M.F. was the project leader; S.M.F., M.H.K and M.P. were responsible for workshop design and facilitation; S.M.F. wrote the manuscript and M.H.K and M.P. provided conceptual contributions and manuscript revisions.

References

1. Knight M. African Rhino Specialist Group report. 10th Meeting of the IUCN/SCC African Rhino Specialist Group. Pachyderm. 2012;52:7–19.

2. Thomas R. Surge in rhinoceros poaching in South Africa. TRAFFIC Bulletin. 2010;23:3.

3. Emslie R. African rhinoceroses – Latest trends in rhino numbers and poaching: An update to Doc 54-2-Annexe 2 from the IUCN Species Survival Commission's (IUCN/SSC) African Rhino Specialist Group to the CITES Secretariat pursuant to Resolution Conf. 9.14 (Rev. CoP15). Gland: IUCN; 2013.

4. Ferreira SM, Botha JM, Emmett, M. Anthropogenic influences on conservation values of white rhinos. PLOS ONE. 2012;7(9), e45989. http://dx.doi.org/10.1371/journal.pone.0045989

5. Milliken T, Shaw J. The South Africa – Viet Nam rhino horn trade nexus: A deadly combination of institutional lapses, corrupt wildlife industry professionals and Asian crime syndicates. Johannesburg: TRAFFIC; 2012.

6. Turton C. Review of the trade in rhinoceros horn in Viet Nam. Viet Nam: TRAFFIC Southeast Asia; 2009.

7. CITES. Conservation of and trade in African and Asian rhinoceroses. [document on the Internet]. c2012 [cited 2012 Mar 24]. Available from: http://www.cites.org/eng/notif/2012/E014.pdf

8. 't Sas Rolfes M. The economics of rhino extinction. Endangered Wildlife. 1990;2:4–9.

9. Blecher E, Thomas K, Muradzikwa S, Smith L, De Villiers P. Economics. 2nd ed. Cape Town: Oxford University Press; 2010.

10. Hollway W, Jefferson T. The risk society in an age of anxiety: situating fear of crime. Brit J Sociol. 1997;48:255–266. http://dx.doi.org/10.2307/591751

11. CITES. Consideration of proposals for the amendment of appendices I and II. 1995: Prop.10.28 [document on the Internet]. c1995 [cited 2012 Jan 25]. Available from: http://www.cites.org/sites/default/files/eng/cop/10/prop/E-CoP10-P-28.pdf

12. Rivalan P, Delmas V, Angulo E, Bull LS, Hall RJ, Courchamp F, et al. Can bans stimulate wildlife trade? Nature. 2007;447:529–530. http://dx.doi.org/10.1038/447529a

13. Biggs D, Courchamp F, Martin R, Possingham H. Legal trade of Africa's rhino horns. Science. 2012;339:1038–1039. http://dx.doi.org/10.1126/science.1229998

14. Child B. The sustainable use approach could save South Africa's rhinos. S Afr J Sci. 2012;108(7/8), Art. #1338, 4 pages. http://dx.doi.org/10.4102/sajs.v108i7/8.1388

15. Knight M. South Africa's rhino success story gets sad new poaching chapter. SWARA. 2011;Oct–Dec:28–32.

16. Fourie R. The rhino moratorium curse. Pretoria: Wildlife Ranching South Africa; 2011.

17. Ferreira SM, Okita-Ouma B. A proposed framework for short-, medium- and long-term responses by range states to curb poaching for African rhino horns. Pachyderm. 2012;51:52–59.

18. TRAFFIC. Creative experts' meeting on messaging to reduce consumer demand for tigers and other endangered wildlife species in Vietnam and China: Meeting report. Gland: TRAFFIC International; 2012.

19. Sohrabian S. Beyond biodiversity: Sustainable development implications of South Africa's "rhino wars" [homepage on the Internet]. c2012 [cited 2014 April 22]. Available from: http://ictsd.org/i/news/bioresreview/142048/

20. 't Sas Rolfes M. Rhinos: Conservation, economics and trade-offs. London: Institute of Economic Affairs; 1995.

21. Morgan TJH, Rendell LE, Ehn M, Hoppitt W, Laland KN. The evolutionary basis of human social learning. P Roy Soc B. 2012;279:653–662. http://dx.doi.org/10.1098/rspb.2011.1172

22. Biggs H, Rogers KR. An adaptive system to link science, monitoring and management in practice. In: Du Toit JT, Rogers KH, Biggs HC, editors. The Kruger experience: Ecology and management of savanna heterogeneity. Washington DC: Island Press; 2003. p. 59–80.

23. Mentis M. Environmental risk management in South Africa. Pretoria: Mike Mentis; 2010.

24. Schoemaker PJH. Multiple scenarios developing: Its conceptual and behavioral basis. Strategic Manage J. 1998;14:193–213. http://dx.doi.org/10.1002/smj.4250140304

25. Becker GS. Crime and punishment: An economic approach. J Political Econ. 1968;76:169–217. http://dx.doi.org/10.1086/259394

26. Messer KD. Protecting endangered species: When are shoot-on-sight policies the only viable option to stop poaching? Ecol Econ. 2010;69:2334–2340. http://dx.doi.org/10.1016/j.ecolecon.2010.06.017

27. Dickson P, Adams WM. Science and uncertainty in South Africa's elephant culling debate. Environ Plann C. 2009;27:110–123. http://dx.doi.org/10.1068/c0792j

28. Roux DJ, Foxcroft LC. The development and application of strategic adaptive management within South African National Parks. Koedoe. 2011;53(2), Art. #1049, 5 pages. http://dx.doi.org/10.4102/koedoe.v53i2.1049

29. Schoemaker PJH. Multiple scenario developing: Its conceptual and behavioral basis. Strategic Manage J. 1998;14:193–213. http://dx.doi.org/10.1002/smj.4250140304

30. Armstrong JS. The value of formal planning for strategic decisions: Review of empirical research. Strategic Manage J. 1982;3:197–211. http://dx.doi.org/10.1002/smj.4250030303

31. Van Aarde RJ, Ferreira SM. Elephant demography challenges CITES trade resolutions. Environ Conserv. 2009;36:8–10. http://dx.doi.org/10.1017/S0376892909005438

32. Gaylard A, Ferreira S. Advances and challenges in the implementation of strategic adaptive management beyond the Kruger National Park – Making linkages between science and biodiversity management. Koedoe. 2011;53(2), Art. #1005, 8 pages. http://dx.doi.org/10.4102/koedoe.v53i2.1005

33. CITES. Conservation of and trade in African and Asian rhinoceroses. Resolution Conf. 2010; 9.15 (Rev. CoP15) [document on the Internet]. No date [cited 2012 Jan 24]. Available from: www.cites.org/eng/res/all/09/E09-14R15.pdf

Water research paradigm shifts in South Africa

AUTHORS:
Raymond Siebrits[1]
Kevin Winter[1]
Inga Jacobs[2]

AFFILIATIONS:
[1]Environmental and Geographical Science, University of Cape Town, Cape Town, South Africa
[2]Water Research Commission, Pretoria, South Africa

CORRESPONDENCE TO:
Kevin Winter

EMAIL:
kevin.winter@uct.ac.za

POSTAL ADDRESS:
Environmental and Geographical Science, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa

DATES:
Received: 18 Sep. 2013
Accepted: 25 Nov. 2013

KEYWORDS:
research; paradigms; South Africa; horizon scanning; scientometrics

HOW TO CITE:
Siebrits R, Winter K, Jacobs I. Water research paradigm shifts in South Africa. S Afr J Sci. 2014;110(5/6), Art. #2013-0296, 9 pages. http://dx.doi.org/10.1590/sajs.2014/20130296

We performed a scientometric analysis of water research publications extracted from four decades of South African related papers to identify paradigms and paradigm shifts within water research in South Africa. Between 1977 and 1991, research publications are dominated by research into technical and engineering solutions, as well as designs and plans to secure water supply. From 1992 to 2001, publications on water pollution, water quality, water resource management and planning are prominent. The second major paradigm is observed from 2001 to 2011 in which the emphasis is on planning, modelling, catchment-scale studies and a multidisciplinary approach to research. Another transition period, towards the end of 2011, is characterised by uncertainty, although it also shows the prominence of key concepts such as participation, governance and politics in water management. The second aim of this study was to identify and prioritise current and future water research questions through the participation of a wide range of researchers from across the country, and to relate these questions to research paradigms, issues and concerns in water in South Africa. Over 1600 questions were collected, reduced in number and then prioritised by specialists in the water sector. The majority (78%) of questions offered by respondents in the South African case study dealt with relatively short- to medium-term research requirements with 47% of questions focused on medium-term issues such as supplying water, service delivery and technical solutions.

Introduction

Limited historical data are available to describe water research in South Africa over the first half of the 20th century. Many authors recognise that this period was dominated by technological developments, breakthrough research and projects in water storage and transfer, as well as the positivist approach to nature and development.[1-3] For example, the development of irrigation in agriculture played a major role in shaping early 20th century water policies, infrastructure and socio-economic development in South Africa.[2]

A new era in water research in South Africa began with the promulgation of the *Water Research Act No. 34 of 1971*. The Act led to the formation of the Water Research Commission (WRC) and the Water Research Fund with the purpose of initiating, managing and financing water research. The objectives of the WRC, as stated in the Act, were to co-ordinate, promote and encourage research in respect of a wide range of purposes and activities.[4]

Furthermore, the *National Water Act No. 36 of 1998* contributed to both forging radical changes in water resource management and new directions in research.[5] This legislation replaced many previous inconsistent acts that focused on water security, supply-side interventions and water as an economically exploitable resource. A shift in the political landscape, marked by the first democratic elections in South Africa in 1994, contributed to a major shift in the existing water resource management paradigm. Legislative reform was timeous amidst growing concerns about the state of the country's waterways and the rising capital expenses in supply schemes, coupled with growing environmental concerns globally.[6-8] South Africa is lauded as being the first country in the world to have promulgated national water legislation which uses water to achieve societal transformation and focuses attention on environmental and social justice.[7] However, although the *National Water Act* introduces a paradigm shift towards a socially equitable and just resource management society,[9] South Africa continues to experience a water deficit that requires urgent management, mitigation and interventions.

We have identified the prevailing paradigms that have influenced the history of water research in South Africa by analysing the publication output over the last four decades. In identifying research questions proposed by a range of researchers active in the water sector in South Africa, we also aim to gain insight into future water research questions and approaches.

One of the challenges lies in finding acceptable methods of identifying paradigms and interpreting these through an historical analysis as well as in contemporary times. The bulk of the theoretical and methodological arguments for question prioritisation, as well as further detailed results and discussion, can be found in Siebrits et al.[10] A brief overview of the processes is provided along with the final results: the priority water research questions for South Africa formulated in the study completed in 2012.

Emerging paradigms in water resource management

A paradigm can identify a conceptual framework that is composed of a class of common elements, theories, laws and generalisations that is widely acknowledged within a scientific school of thought or discipline. Paradigms also shift for a variety of reasons and under various influences. According to Kuhn[11], when enough significant anomalies have accrued against a current paradigm, then the scientific discipline is thrown into a state of crisis. During this crisis, new ideas, and even those previously discarded, are tested further. A paradigm shift occurs only after a given discipline or significant thinking in a field of knowledge changes, and only when this change is widely recognised. Sometimes paradigms only gain ground because of some dramatic and unforeseen verification, such as a shift in legislation or policy, or for personal or aesthetic reasons in which they may appear neater, simpler or more elegant than their older counterparts do.

When new paradigms appear, however, they are rarely complete. More often, they are the products of relatively sudden and unstructured events that arise from an enlightened moment in which previously hidden components of knowledge suddenly come into view.[11] New knowledge sometimes has the power to cause an anomaly leading to an unexpected change in the world view of those holding one or another paradigm. Thus, a change of world view begins when a significant anomaly is recognised within an existing paradigm. The challenge is often to identify the anomaly and then to recognise the significance or importance of the phenomenon. The signals and changes in paradigms, with respect to paradigm changes in water resource management, provide the context in which to explore corresponding changes in the water research enterprise in South Africa.

One of the earliest paradigms in water resource management began at the start of the 20th century and is most often acclaimed as the hydraulic mission because it is characterised by major engineering activities involving the construction of water infrastructure to capture, store and distribute water. This period is also described as the heroic engineering phase[12] which is noted for its immense scale of projects and plans. This phase also finds support in modernist or positivist philosophical beliefs which consider it possible to control and manage nature. The majority of water projects in this period were concerned with supplying more water, more efficiently, to more areas.[1,13]

The demand side of water resource management, which is also a management paradigm, focuses attention on how to manage water demand and use. This shift is influenced to an extent by various social advocacy movements, but is also driven by increasing recognition of resource scarcity, heightened interest in sustainable development considerations, post-modern philosophies and increased prominence of environmental justice, equity and democratisation of resources.[1,14]

On a larger, global scale, Allan[15] makes further use of paradigms in explanations of global changes in water resource management. His work focuses on the development of an analytical method to address the problem of water resource allocation. Allan's contribution lies in identifying paradigms that are reliant on economic, legal and political factors that influence the water sector in semi-arid countries. This shift in paradigms is represented in a transition of five water management paradigms, each with its own distinct focus and function.

The first of the five paradigms is referred to as the pre-modern paradigm, which spanned from 1850 to the beginning of the 19th century, and which was dominated by a general increase in water supply and usage. It occurred in an era of the hydraulic mission in which ingenuity and engineering efforts abounded.[12] The second paradigm from the early to late 20th century was characterised by industrial modernity and again featured an increase in activity in the hydraulic mission. In this phase, water demand increased because agricultural activity shifted from subsistence to commercial-based economies, followed by further demands on water resources as a result of the rapid increase in industrial activity.

The third paradigm in Allan's[15] framework, which occurred in industrialised nations from the 1960s, shows a shift towards sustainable resource management and a concerted effort to redress the damage done by previous paradigms. The fourth is characterised by a period of economic expansion (particularly in the North) and by smart economic decisions that offer several environmental advantages, but is also characterised by a general decline in the hydraulic mission. Finally, the fifth paradigm is dominated by political and institutional change which becomes increasingly aligned with global shifts towards sustainability and also a rapid decline in the hydraulic mission.

Overall, there are elements within Allan's management paradigms that are in alignment with similar developments in water resources in South Africa. The assumption too is that the water research enterprise in South Africa corresponds reasonably well with these changing paradigms in water resource management. The occurrence of different paradigms, as suggested by Allan, especially the third, fourth and fifth, are of particular interest in this study because these paradigms represent a period in which the research effort should be detectable in the research publication evidence and in the formulation of research questions that was undertaken as a part of this study.

Paradigms in research publications: A scientometric analysis

Given the evolution of water management paradigms in South Africa and also globally, the question then arises as to whether scientific research has responded to these paradigmatic changes, and if so, in what ways. Has research driven these management paradigms or has it lagged behind?

The field of bibliometrics provides useful insights into the health of a country's national innovation system and is also a necessary and integral part of science policy. For the purposes of this paper, bibliometrics is synonymous with scientometrics, that is, the analysis of science and scientific output. Scientometrics is used interchangeably with bibliometrics and informetrics.[16] The three fields all refer to the study of science, knowledge, and knowledge management and production. In this study, scientometrics is considered as the 'study of the quantitative aspects of science as a discipline or economic activity'[17]. It is used to analyse the evolution and academic output of water research within a socio-political and historical context.

Scientometric methods begin with the collection of a series of appropriate publications or reference material followed by a network analysis. Sets of keywords and/or noun phrases amongst the journal articles are analysed with respect to their frequency to each other within an article and between articles. This approach results in a topic/word/concept co-occurrence network. In this analysis, statistical algorithms, such as cluster analysis and multidimensional scaling, are used as the foundation of scientometrics.

Studies have tested the strength of accepted scientometric methods and showed that scientometrics is robust and reliable even on a coarse level.[18,19] Arguably, the main societal impact of scientometrics has been the creation of the impact factor and analysis of research, researchers, publications and journals.

Scientometric maps have a number of advantages in the output of research analysis. Researchers argue that scientometric maps are an important means of conveying the results of the method.[20,21] The maps allow for the representation of diverse and large sets of data, but they remain heuristic tools that can be used to explore and consider plural perspectives to inform decisions and evaluations.[21] In this study, maps are the main means of representing results of patterns and trends in water research in South Africa.

Identifying water research questions

Scientometric analysis of published works provides an interpretative account that helps to identify patterns of change and to understand the relationships that influence these trends. However, scientometrics is not an appropriate method for determining future water research questions. For this purpose, we used a form of horizon scanning to identify future research questions and strategies using methodological elements similar to studies undertaken by Sutherland and Woodroof[22] which are to: (1) scope the issue, (2) gather information, (3) spot signals, (4) watch trends, (5) make sense of the future and (6) agree on the response. In this study, we used a similar approach which is supported by a collaborative, multi-stakeholder process to identify and examine threats or trends in society, the environment or a sector, and to identify needs that will enable appropriate questions.[22,23]

Research methods

Scientometric analysis

A conceptual narrative on water research in South Africa is central for the discussion on water research paradigms, knowledge and appropriate adaptive capacity. Many authors have discussed how these approaches provide an objective and evidence-based means of assessing the state of a research or scientific field[16,24,25]. Scientometric methods are based on two assumptions. Firstly, that 'scientific knowledge can be represented as a network of concepts or ideas, and that these elementary entities can be aggregated to form macro-structures'[26]. Secondly, if mapped or represented in a structured manner then 'it is assumed that each map is a snapshot at a distinct point in time of what is actually a changing and evolving structure of knowledge'[26]. The key data for this method are research outputs, in the form of publications, collaborations, intellectual property, policy influence and applications.

Locating relevant water-related publications objectively and comprehensively is a challenge within itself. This challenge stems from the definition of water research used herewith. In this study, the journal search set comprised a twofold approach. Firstly, journals that published five or more articles with the search terms 'water' and 'South Africa' (or derivatives thereof) were included. Secondly, snap polls and pilot surveys undertaken towards the end of 2011 that included questions asking practitioners where they published and read South African water-related research were included. The results from the significant publication count criteria and stakeholder input amounted to 171 publications forming part of the journal search set. These journal titles were then added to the query and searched further. The final search query was for journal articles that contained 'water' and 'South Africa' in their topic within the journal search set. Searches were performed on Thomson Reuters Web of Science.

Scientometric queries require that the content or topic of research outputs must be analysed. The most widely used method for this analysis is the co-word analysis of research publications, particularly within their title and abstract.[25,27-30] The methodological foundation of co-word analysis is the idea that the co-occurrence of words describes the contents of documents. By measuring the relative intensity of these co-occurrences, it is possible to establish a simplified representation of a field's concept networks.[27] Co-word analysis examines the frequency of individual words or word phrases within a data point (publication) and concurrently across the data set. The more frequently a word or phrase appears in a data point, the more relevant that topic becomes within the data points.

The final output from this process is a network visualisation file (.net) for the most frequent keywords within the data set and their relative frequency towards other keywords. This network file is the fundamental output of mapping and visualising networks within scientometrics. The most common method of visualising or representing these outputs is through network maps. These maps are simplistic representations of the networks themselves and represent the strength of topics and their interrelationship (associated strength) with other topics. The majority of network maps use size and distance as indicators of certain attribute properties or relationships.

There are limitations in the use and interpretation of scientometric maps because the output provides only a representation of relationships among terms found in published content. The results should be interpreted with caution even though the evolution of scientometric methods represents the most effective known method of simply representing scientific relationships, output or developments on a particular scale.

The search for water research questions

A form of horizon scanning was used in this study to identify and evaluate research questions that are currently being asked by researchers. There are three main methodological steps that are typically used: (1) identify and create a collaborative stakeholder network, (2) collect data from this network regarding research expertise, opinions on research considerations and research questions and (3) analyse this data by allowing the network to deliberate the results and produce a final set of results of research opinion and questions.

A substantial taxonomy of horizon scanning methods used in identifying and prioritising future research questions, scenarios and needs is provided by Sutherland and Woodroof[23]. They follow a combination of open fora, trend analysis, questionnaire and expert consultation. There are multiple reasons for attempting to engage a wide variety of stakeholders. Arguably a strength, and at the same time a weakness of the current study, was the intention to involve a wide variety of stakeholders with an interest in water and water research, and to engage these participants through the *voice* of a research initiative, rather than through that of the researchers. The intention was to make communication professional, allow for branding to be created, and to enable a common identity if other researchers began working on, or in association with, the project. The detailed methodological steps and substantial outputs from this process can be found in Siebrits et al.[10] while relevant results are presented here.

Results and analysis

Research output

The number of journal articles and research reports published per year from 1977 to 2011 as identified through the search is shown in Figure 1. The stacked column graphic shows the proportion of WRC research reports, *Water SA* articles and other journal articles. In summary, there is an increase in annual publication counts, a rise in *Water SA* articles and a marked increase in WRC research reports. The increase in the proportion of other journal articles from the early 1990s until the present is notable. South Africa's water-related research output has steadily increased and the research is found in more diverse, international journals.

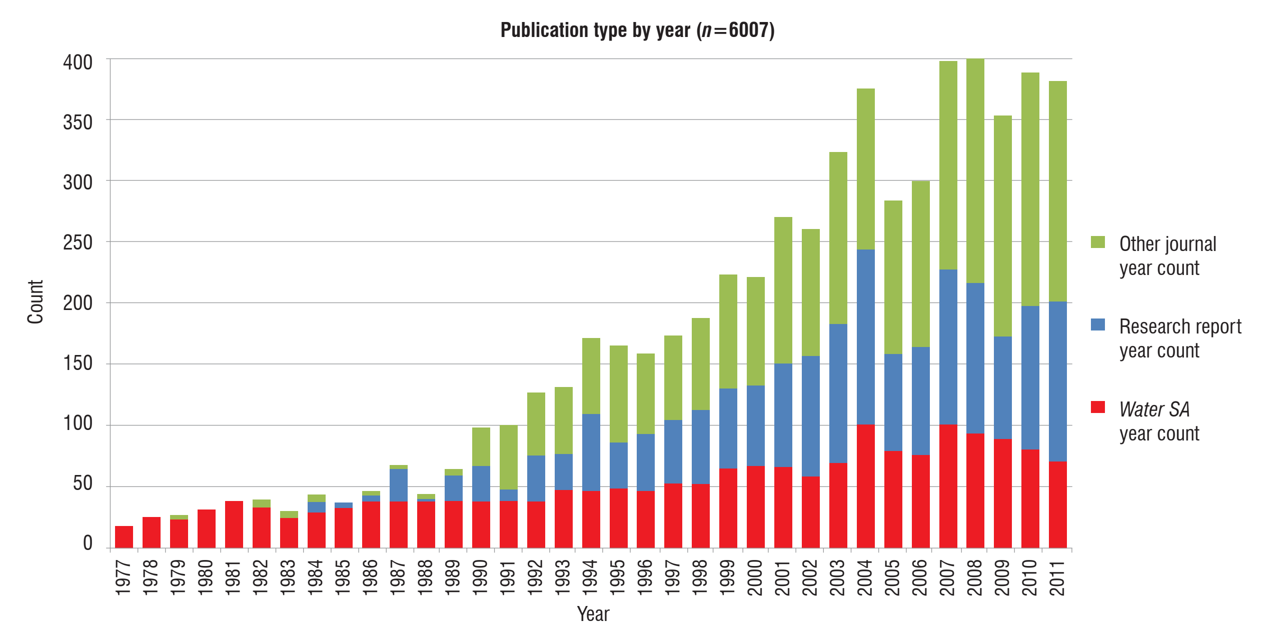


Figure 1: Publication type by year for all data points and all data sources.

A scientometric map was created using Sci2 and VOS viewer and is displayed for illustration purposes in Figure 2. The figure comprises results for a specific time slice, shown in label and density format. Label format presents more prominent words in the network as larger spheres. The closer spheres or words are to each other the more interrelated they are in the network. Colours represent clusters or sets of related words as they emerged from the network (i.e. more general relationships). Lines represent significant connections between words of the same or different clusters. Density format presents the identical map but uses warmer colours to display words and clusters of greater prominence with the colour contours conforming to how strongly related clusters are. Word size represents general prominence as per label view.

The map shows the time slice from 2002 to 2006 (Figure 2). It illustrates a range of emerging research fields and a general increase in overall publications. The word 'management' becomes pronounced along with terms such as 'community', 'impact' and 'application'. It is also during this phase that the word 'integrate' becomes increasingly prominent.

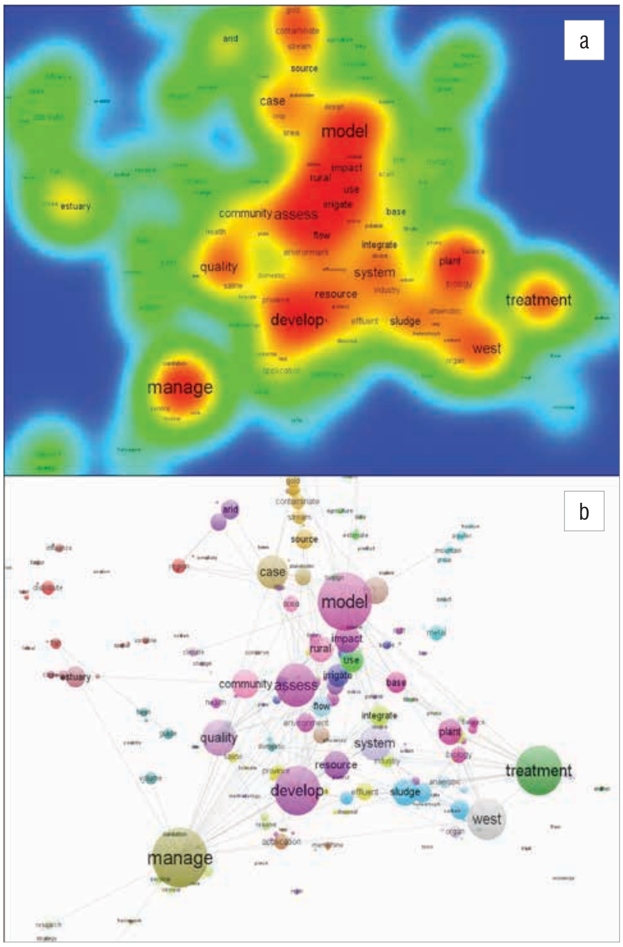


Figure 2: (a) Density and (b) label format bibliometric maps based on keywords in publications from 2002 to 2006 (*n*=1545).

The stakeholders captured by the research signed up and engaged in the process for numerous reasons. Some simply wanted to remain informed of the process and results. Some saw an opportunity to participate in the surveys and discussions while others used the portal to seek further information about water research. When the study was completed in December 2012, there were 2260 unique stakeholder contacts on the database.

The stakeholders contained within the database were diverse in their involvement in the South African water sector but appeared well connected within the water sector networks. Figure 3 displays the organisations represented in the database. Overall, stakeholders in the database were affiliated to 572 organisations or institutions.

Question gathering

Overall, 1075 stakeholders were contacted directly via individual telephone calls during May and June 2012 to be alerted about the survey. By the time the main survey closed in December 2012, 641 completed responses had been received. Of the 1674 questions submitted, 4629 keywords/categories were provided of which 844 of these were unique. These keywords/categories can be seen as the descriptive data of the submitted questions and guided the identification of themes for the workshop. Table 1 shows the top 40 keywords in the data set along with the number of counts per keyword in brackets.

The most striking result is the 245 occurrences of the keyword 'management'. A large proportion of the submitted questions had a management-oriented line of inquiry. The questions were further categorised into six themes:

1. Change – building socially resilient and adaptive responses to social, climate and general environmental change
2. Data – capturing of quality data through strategic monitoring, and with reliable analysis, modelling and scientific reporting
3. Ecosystems – protection, conservation, restoration and productive use of healthy ecosystem services
4. Governance – integrated, strategic adaptive management
5. Innovation – investment in infrastructure and research for innovation
6. Resources – protection, conservation, treatment and management of water resources for equitable growth and development

Following further refinements to the questions, which included filtering for duplicates and suitability as well as quality control through reasonable testing, a total of 401 questions were presented as the input data to the Water Research Horizon Scanning Workshop in October 2012 in Cape Town. Delegates were tasked to reduce the 401 questions to approximately an eighth in total number. The final data set amounted to 59 priority water research questions across the six themes. The workshop question prioritisation which constituted the central output result from the workshop is presented in the publication by Siebrits et al.[10]

Research output and links to paradigms

South Africa has undergone significant changes in the output and structure of water research over the past four decades. There has been substantial growth in output with a total relevant sample publication record of 6007 articles and research reports and a current annual output of over 350 articles and reports per year. The number and sources of journal articles over this period have increased and diversified while WRC research report output has also increased, albeit at a slower rate.

The emergence of two main areas of research or fields of specialisation in the democratic transition (1992–1996) period is supported by greater diversity of publications than in previous years. The engineering or technical research outputs cluster together and again focus on treatment systems, processes and evaluation. This time the clustering is associated with management-based and planning-oriented research which is pronounced in the words 'catchment', 'develop' and 'urban'. Although somewhat dispersed, water quality and algae also emerge as topics of research concern.

A transition period in water research occurred over a period that became increasingly focused on quality constraints, fields of management and planning. Words such as 'review', 'model', 'community' and 'geography' begin to appear in the research publications. The emergence of these words supports the beginning of paradigm changes as a result of water deficits towards end-use efficiency as outlined by Ohlsson and Turton[14]. The beginning of paradigm changes also indicates that the second transition of Turton and Meissner[31] occurred with a new social contract around water that came not only from a new political regime and democratic transition that focused on redistribution, but also one that was spurred on by a movement of South African environmentalism, the beginning of the global sustainability debate and the rise of civil

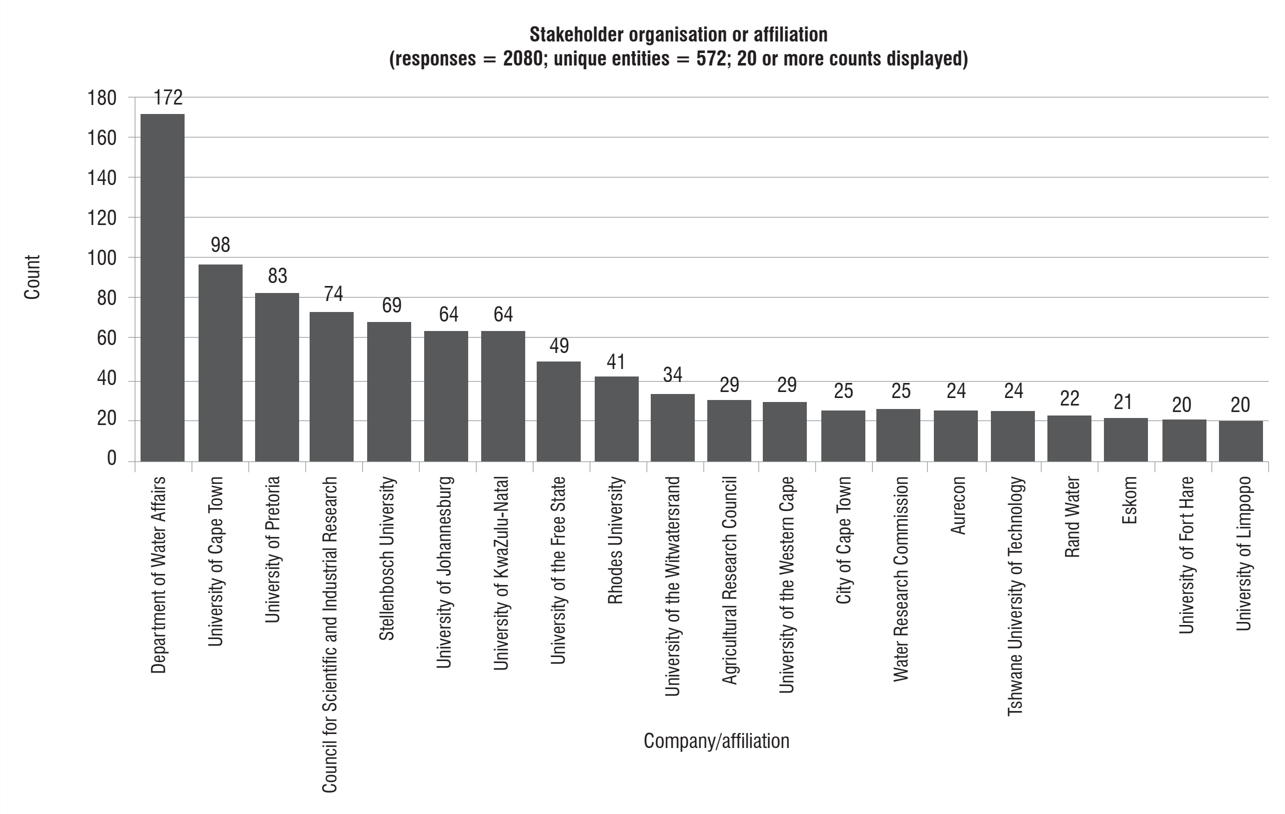


Figure 3: Distribution of stakeholder organisation or affiliation.

Table 1: Top 40 research questions by keyword provided by stakeholders

management (245)	groundwater (79)	sanitation (54)	technology (44)	conservation (27)
treatment (136)	hydrology (73)	services (53)	policy (43)	capacity (26)
quality (118)	mining (73)	education (51)	rural (43)	energy (26)
supply (103)	health (72)	research (50)	use (37)	human (26)
wastewater (99)	economics (65)	monitoring (49)	wetlands (34)	planning (26)
agriculture (94)	catchment (55)	resources (49)	environmental (30)	urban (26)
pollution (83)	change (54)	ecology (47)	industry (30)	waste (25)
governance (80)	climate (54)	river (45)	demand (29)	alternatives (25)

The number of counts per keyword is indicated in brackets.

society activism. Here marginal uncertainty begins to creep into the understanding of water affairs as described by Allan[15] and the need to model, plan around catchments and include other disciplines (especially from the humanities) begin to become considerations in the research environment.

The period 1997–2001, around the major transformation of South Africa's water laws and post-establishment of the national Constitution, shows a strong polarisation between the main technical and management-orientated disciplines. Words such as 'develop', 'manage' and 'assess' become more prominent while the technical focus diminishes. Researchers began to focus further on understanding the broader water context, to use systems approaches and to begin to plan for more than just engineering solutions. These results support the view that a transition was still underway with regard to the dominant paradigms, but the word system had shifted noticeably towards the management- and development-related research disciplines and away from the technical.

The most recent decade of water research represents the greatest change in water research paradigms. It represents over half (3456 of 6007) of the collected and analysed publications, and constitutes the most representative sample of current recent water research. In this period, words become clustered and centralised, with images being most clustered in their centres with few stand-alone concentration areas. This pattern indicates how research has become more diverse yet interconnected and a shift towards other disciplines. This is most prominent in the first series of the millennium analysed (2002–2006) with an emphasis on concepts such as management, modelling and development. These observations point to research that is directed towards dealing with current issues and societal benefits and needs.

Between 2007 and 2011 there appears to be a significant interconnectedness of specific keywords with many others. Here management has become a key research theme, which is connected to almost every other keyword or area of interest. All major areas of water research received fair attention and prominence in the results, from treatment systems to catchments, modelling, communities, development and biological concerns. The word 'integrated' is increasingly prominent and linked to management, suggesting a dominant thrust in water research activity over this time. The growing prominence of climate-related research also highlights growing global interest in environmental change. Another interesting emerging field is groundwater research. While this topic has been present alongside general hydrological keywords and concepts, during this period the development and impact of the groundwater theme appears to become more independent than before.

The research effort in South Africa appears to have evolved into a new set of paradigms, albeit it tentative and uncertain, in which some emphasis is given to the social sciences disciplines and to concepts of governance and management. There is also evidence of research that focuses more attention on demand-side applications and interests, and integrated management. However, a third or reflexive transition phase does not appear just yet.[15] Keywords that relate to the green economy or risk awareness are not yet prominent. What is obvious is an increase in the prominence of collaboration across multiple disciplines over the last decade.

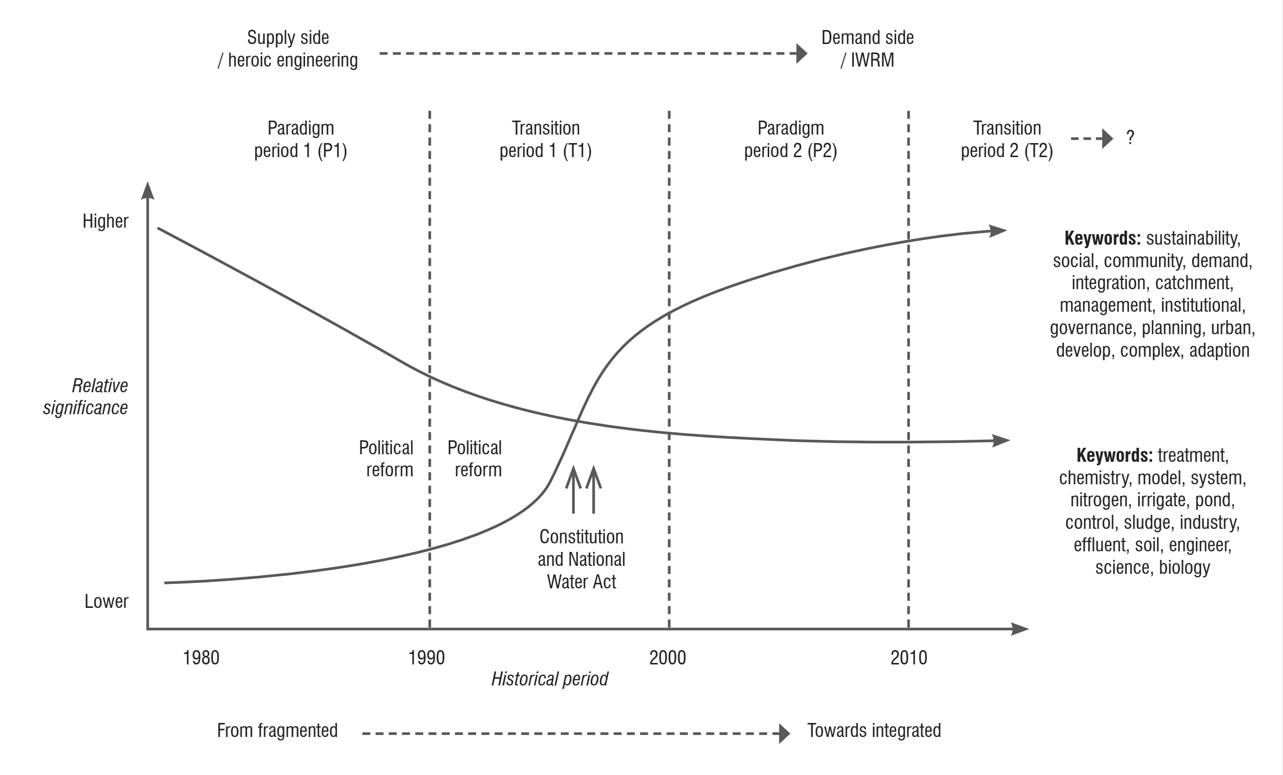
In brief, the scientometric analysis of South African published works on water research over the past four decades shows two reasonably distinct paradigms (Figure 4). The first paradigm occurs in a period dominated by the quest to supply water, which is interrupted dramatically by changes in the political landscape. The Constitution, the *National Water Act* among others, and the shift in the national balance of power, introduce the next paradigm shift and an emphasis on integrated water resource management. This new paradigm is characterised by a research effort that is centred on new themes and concepts such as sustainability, community, governance and adaptation. The shift from the 1980s, once dominated by research efforts that focused on treatment, technical interventions, chemistry and so forth, now features research interests, themes and approaches such as integrated water resource management and multidisciplinary studies in water research.

Finally, it is interesting to observe what is not prominent in the scientometric results. Topics and themes such as data quality and integrity, law, rights, access, licensing and culture are noticeably absent from most of the scientometric outputs. This absence does not necessarily mean that they are being ignored, but rather that they are receiving less attention than other research disciplines and specialisations. The absence of these terms does not necessarily alter the observed paradigmatic shift, but may suggest that the South African water research field is not yet ready to move on to another water paradigm – at least not in the immediate future.

Identifying and prioritising questions and the link to paradigms

The launch and strategies undertaken through the Aqua d'UCT initiative far surpassed expectations with regard to participation, uptake and response. The robust and yet diverse nature of the results and community interaction during the study was shown by the steady growth of interest from approximately 600 to over 2000 stakeholders on the research contact database by the time the study was completed in 2012.

The most salient findings of the survey indicated that many respondents wanted longer and more substantial research projects to be funded and established, yet the majority of research questions that were offered were categorised as short- to medium-term projects taking only 1 to 3 years to complete. Nevertheless, these questions reflect the diverse research disciplines and specialisations as suggested by the keywords such as 'management', 'governance', 'planning', 'education', 'policy', 'planning' and 'alternatives' being most prominent. However, those questions of a more technical nature relating to treatment, quality and pollution, hydrology, climate, supply and ecology dominate the input data set.


Supply side
/ heroic engineering
Demand side
/ IWRM
Paradigm
period 1 (P1)
Transition
period 1 (T1)
Paradigm
period 2 (P2)
Transition
period 2 (T2)
?
Higher
Relative
significance
Lower
Political
reform
Political
reform
Constitution
and National
Water Act
1980
1990
2000
2010
Historical period
From fragmented
Towards integrated
Keywords: sustainability, social, community, demand, integration, catchment, management, institutional, governance, planning, urban, develop, complex, adaption
Keywords: treatment, chemistry, model, system, nitrogen, irrigate, pond, control, sludge, industry, effluent, soil, engineer, science, biology

IWRM, integrated water research management

Figure 4: Paradigms and transitions emerging from scientometric analyses.

The survey results form a substantial collection of research questions from water research stakeholders. The process of reducing the survey data set into something manageable for prioritisation at the final workshop was also a rigorous one. The reduction from 1603 initial to 59 priority research questions followed similar methods to those used by Sutherland et al.[32] The only significant change was to gather the initial data set of questions from a broad and larger community rather than from key specialists only.

As mentioned earlier, the final set of research questions is presented in Siebrits et al.[10] Here we have analysed the content of the questions further and organised the results into two categories. The first category is a combination of words and concepts from each question placed in three columns – short-, medium- and long-term research challenges. The method is a subjective one. Some researchers will interpret the questions differently according to factors such as context and experience. In addition, no question resides exclusively within a particular level of challenge. For example, a short-term research question that seeks to understand the skills gap in the water sector might identify the extent of the problem and the causal factors, but skills development is more likely to be a long-term issue that requires careful monitoring, evaluation and interventions that are informed by research that falls into a long-term research challenge.

The second category organises the questions into indicators of knowledge and knowledge management. For example, the solutions to many research questions are already known from previous studies conducted elsewhere in South Africa or in the world, whereas questions dealing with long-term issues and grand challenges require a much greater commitment to knowledge construction.

Table 1 maps all 59 questions against this matrix. The results were not tested with the stakeholders, but are presented as a contribution to thinking about how best to incorporate multiple criteria into the development and organisation of the research question bank.

In general, the presentation (Table 2) confirms three important observations. Firstly, over 78% of the questions that were offered and refined at the workshop seek to address short- to medium-term research questions, typically questions dealing with service delivery, sanitation, access to water, pricing and water quality. Secondly, the majority of the questions confirms the existence of a transition paradigm, similar to what was identified earlier in the scientometric analysis. These questions deal with issues of intermediate concern but are also dominated by issues of integration, data and information systems, social change, planning decisions and further development of regulations. The majority of the questions is organised in the medium-term category. The uncertainty is created partly by the long list of questions that seek to address multiple issues that beset the South African water research landscape. Finally, there is a small set of questions that is arguably more closely aligned with issues and concerns that features some elements of Allan's[15] third, fourth and fifth paradigm. Here the questions deal with medium- to long-term critical concerns of sustainability, establishing green economies and implementing new forms of integrated, adaptive governance. These kinds of questions pose extraordinary challenges necessitating considerable financial and institutional support. Two examples of these kinds of questions are:

- How can innovative process technologies, including nanotechnology, be applied to benefit water and wastewater treatment process?
- What are the life cycle and systematic impacts of acid mine water and how can these be managed, mitigated, remediated and beneficiated?

Delegates acknowledged that the workshop was an energising and interesting collaborative exercise. While there were some obvious

Table 2: Categorisation and organisation of the research question data set

Research questions: **Extent of challenge**	**Short term** **(immediate: 2 to 3 years to complete these projects)**	**Medium term** **(10 to 20 years to complete these projects)**	**Long term** **(grand challenges: more than 20 years to complete)**
High level of knowledge required together with financial investment, expertise, capacity, leadership, etc.	Integrated planning; challenge of rapid urbanisation; threat of mining on water quality; poor regulation; freshwater pollution; access to water; managing fit of purpose water; rural water management; impacts of eutrophication	Early warning systems; impacts on health; economic and social value of ecosystems and services; ecological reserve determination; sustainable abstraction; sustainable catchments; applying business model to catchments; improving treatment of emerging contaminants; understanding water and energy nexus; water use in agro-industry; improving footprint assessment; risks and impacts of acid mine drainage; building water-sensitive settlements	Food and water security; ensuring functional, safe, socially just service delivery; understanding and addressing global scale change impacts on water resources; addressing ecological risks that also impact on society; analysing emerging micro-pollutants and treatments; treating pathogens; using nanotechnology in water treatment
Reasonably high level of knowledge required to meet these challenges	Challenge of meeting supply and demand; waterborne sanitation challenges; access to efficient water-based services; societal contribution to total water services; open-access data sources; impact of invasive species on water resources; skills gaps; water losses; socially just water pricing; efficient water use; regulating illegal water use	Enabling relevance of water research; determining societal and environmental value of water; improving real-time data capture; using data for decision support systems; accountability in governance of water; efficacy in bio-remediation projects; protection of ecological systems; improving public education and response; developing innovation in information systems; ring-fencing water costs; enforcement of water quality; establishing co-operative governance; improving communication networks and tools; managing sediment accumulation; use of remote sensing in monitoring; addressing agro-hydrological drainage; climate change and water scarcity and threats; water resource management in informal settlements; importance of water in urban design and planning; innovation and development in water re-use; impacts of shale gas fracturing; responding to growth and development; ecosystem protection; policy and legislation response to scarcity and poor water quality	Analysis, assessment, monitoring, reporting and regulation of groundwater at national scale; determining risks and impacts of using different forms of treated water for irrigation
Relatively easy to build the required level of knowledge (much is already known)		Improving quality and access to hydrological data; understanding change drivers in water resources management; transitioning to sustainable development	

gaps in the representation of participants, delegates were pleased to interact with diverse leaders in the field. Most delegates appreciated the quality of exchange and interaction during the formal and informal activities. Positive comments were also received about the organisation, facilitation, the venue and structure of the workshop. Many said that the structured approach to the workshop made the best use of time in order to achieve the intended product.

The strongest criticism from delegates was that the approach and methods used at the workshop were not designed to identify horizon scanning research questions per se. Rather, delegates said that they felt coerced into responding to questions that were put before them. Moreover, delegates felt that it was difficult to develop new questions that were of an horizon scanning, long-term nature for a number of reasons: the groups were too diverse, there was insufficient time to consider and develop meaningful questions and the process was too demanding for the facilitators which resulted in tasks being carried out in a mechanistic manner against a tight time frame.

Delegates were also critical of the fact that they had to work with a large number of questions that were poorly formulated. Problematic questions came in a number of forms: they were often about immediate issues; they were not valid research questions; they were too broad to be categorised in a chosen theme; they were often limited to disciplines and fields within the natural sciences; and many did not show any insight into what might lie on the horizon.

The final list of questions may well be indicative of the current state of thinking amongst researchers generally. The majority of the questions focused on end-use efficiency, demand-side management and technical solutions, with only a few dealing with research to further innovation, progressive forms of governance, and the integration of other sectors and their respective alignment with water resource management. This apparent lack of creativity and innovative thinking also explains the frustration that some delegates felt in that they were expecting to be able to provide foresight into drafting new types of questions that would create a new set of paradigms and lead future thinking.

Conclusions

Scientometric results show that the publication record for water-related research in South Africa contained 6007 publications from 1977 to 2011. WRC research reports amounted to 1760 (29.30%) of this total. The remainder were peer-reviewed journal articles. Of these journal articles, 1829 (30.45%) were published in *Water SA*. The publication record also increased in number dramatically since 1990, with more articles being published annually than each previous year throughout the data set.

Paradigms were identified through the scientometric mapping methods using the publication record to show a history of water research from 1977 to 2011. Overall, the research output focused predominantly on management, development, models, quality and system treatment. Technical matters are dominant in the historical record but other paradigms such as allocative efficiency, uncertainty and risk are also present to lesser degrees. The change in paradigms is observed when these results are examined over successive time periods.

Two major paradigm approaches were observed in the analysis of water research publications along with one significant transition period. The first set of paradigms, from 1977 to 1991, emphasises the hydraulic mission in which research and implementation aimed to secure supply and understand basic natural systems. This period is dominated by engineering and laboratory-related disciplines. The 'getting more' and 'supply management' paradigms are characterised by efforts to ensure water supply, drainage and the development of the sewered city – mainly engineering and biological-related research efforts. In the following 10 years (1992–2001), there is a transition in which quality constraints and fields of management and planning become prominent. This paradigm is in response to changes in water deficits and a focus on end-use efficiency. A second transition occurs with a new social contract around water at a time when the new political regime enters government in a period of democratic transition, growing environmentalism and a rise of civil society activism. The need to plan, model catchments and include other disciplines becomes evident in the research environment.

The question prioritisation activities using horizon scanning methods provided an opportunity to engage with a wide and diverse population of water research stakeholders and practitioners. The survey resulted in a substantial collection of research questions from water stakeholders and researchers. Many questions deal with immediate to medium-term concerns while only a few aim to tackle long-term or systemic problems. Others are coupled or integrated questions that cover a number of disciplines.

There are recognised limitations to this study. The simplification of scientometrics causes a potential loss in detail and context. The interpretations of output maps remain subjective but the method does provide powerful, macro perspectives of a research area. However, albeit simplistic, the methods used in the field of scientometrics are repeatable and are not dependent on the choice of experts and their opinions which may vary as the choice of the participants changes in peer reviews.[33]

It is recommended that further detailed mapping and analysis be done on publications to explore the reasons that might cause paradigm shifts as well as to understand what is missing in the existing body of knowledge. Horizon scanning has many inappropriate elements for the South African context as it is limited to a degree by its reach and participation. It is recommended that further prioritisation activities are undertaken to guide research but that these are expert lead and informed at the earliest stage before taking the results to a wider audience for consultation. The current state of questioning does, however, provide an overall perspective of what a large and diverse group of research stakeholders and practitioners believes is important, even if these may not deal with long-term challenges but are rather more situated in addressing current and pressing research needs.

Acknowledgements

This publication was based on the Water Research Commission funded project K5/2170. The project leaders wish to thank the reference group participants, the workshop delegates who gave of their time to prioritise research questions and debate research in South Africa and the project staff who contributed in their specific tasks and fields.

Authors' contributions

R.S. performed most of the experiments and wrote the manuscript; K.W. was the project leader; and I.J. was the funding manager and expert advisor.

References

1. Tempelhoff J, Hoag H, Ertsen M. Water history and the modern. Water History. 2009;1:81–82. http://dx.doi.org/10.1007/s12685-009-0014-3

2. Tewari D. A detailed analysis of evolution of water rights in South Africa: An account of three and a half centuries from 1652 AD to present. Water SA. 2009;35:693–710. http://dx.doi.org/10.4314/wsa.v35i5.49196

3. Allan T. Water in international systems: A risk society analysis of regional problemsheds and global hydrologies. Presented at: The 1999 Oxford University Conference on Water Resources and Risk; 1999 March 22; Oxford, UK. Available from: http://awiru.co.za/pdf/allantony2.pdf

4. Republic of South Africa (1971) Water Research Act No. 34 of 1971. Pretoria: Government Printers.

5. Republic of South Africa (1998) National Water Act No. 36 of 1998. Pretoria: Government Printers.

6. Herold C. The water crisis in South Africa. 2009 Des Midgley Memorial Lecture presented at the 14th SANCIAHS Symposium; 2009; University of KwaZulu-Natal, Pietermaritzburg, South Africa.

7. Funke N, Nortje K, Findlater K, Burns M, Turton A, Weaver A, et al. Redressing inequality: South Africa's new water policy. Environment. 2007;49(3):10–23. http://dx.doi.org/10.3200/ENVT.49.3.10-25

8. Schreiner B. Water services: Yesterday, today and tomorrow – a strategic perspective. In: Proceedings of the 2006 Water Institute of South Africa Biennial Conference; 2006 April 21–25; Durban, South Africa. Johannesburg: WISA; 2006. Available from: http://www.ewisa.co.za/literature/files/331%20Schreiner.pdf

9. Fallenmark M, Rockstrom J. The new blue and green water paradigm: Breaking new ground for water resources planning and management. J Water Res Pl-ASEC. 2006;132(3):129–132. http://dx.doi.org/10.1061/(ASCE)0733-9496(2006)132:3(129)

10. Siebrits R, Winter K, Barnes J, Dent M, Ekama G, Ginster M, et al. Priority water research questions for South Africa developed through participatory processes. Water SA. 2014;40(2):199–209. http://dx.doi.org/10.4314/wsa.v40i2.2

11. Kuhn T. The structure of scientific revolutions. Chicago, IL: University of Chicago Press; 1962.

12. Turton A. The role of science in deepening democracy: The case for water in post-apartheid South Africa. J Transdiscipl Res S Afr. 2009;5(1):9–28.

13. Van Vuuren L. What's in a name: Looking back at the start of public water governance. The Water Wheel. 2009;July/August:38–41.

14. Ohlsson L, Turton A. The turning of a screw: Social resource scarcity as a bottle-neck in adaptation to water scarcity. Stockh Water Front. 2000;1:10–11.

15. Allan J. Water in the environment/socio-economic development discourse: Sustainability, changing management paradigms and policy responses in a global system. Gov Oppos. 2005;40(2):181–199. http://dx.doi.org/10.1111/j.1477-7053.2005.00149.x

16. Hood W, Wilson C. The literature of bibliometrics, scientometrics and informetrics. Scientometrics. 2001;52(2):291–314. http://dx.doi.org/10.1023/A:1017919924342

17. Tague-Sutcliffe J. An introduction to informetrics. J Inf Process Manage. 1992;28:1–3. http://dx.doi.org/10.1016/0306-4573(92)90087-G

18. Klavans R, Boyack K. Quantitative evaluation of large maps of science. Scientometrics. 2006;6(3):475–499. http://dx.doi.org/10.1007/s11192-006-0125-x

19. Rafols I, Leydesdorff L. Content-based and algorithmic classifications of journals: Perspectives on the dynamics of scientific communication and indexer effects. J Assn Inf Sci Technol. 2009;60(9):1823–1835. http://dx.doi.org/10.1002/asi.21086

20. Roessner D. Quantitative and qualitative methods and measures in the evaluation of research. Res Evaluation. 2000;9(2):125–132. http://dx.doi.org/10.3152/147154400781777296

21. Rafols I, Porter A, Leydesdorff L. Science overlay maps: A new tool for research policy and library management. J Assn Inf Sci Technol. 2010;61(9):1871–1887. http://dx.doi.org/10.1002/asi.21368

22. Sutherland W, Woodroof H. The need for environmental horizon scanning. Trends Ecol Evol. 2009;24(10):523–527. http://dx.doi.org/10.1016/j.tree.2009.04.008

23. Shackleton C, Scholes B, Vogel C, Wynberg R, Abrahamse T, Shackleton S, et al. The next decade of environmental science in South Africa: A horizon scan. S Afr Geogr J. 2011;93(1):1–14. http://dx.doi.org/10.1080/03736245.2011.563064

24. LaRowe G, Ambre S, Burgoon J, Ke W, Borner K. The scholarly database and its utility for scientometrics research. Scientometrics. 2009;79(2):219–234. http://dx.doi.org/10.1007/s11192-009-0414-2

25. Todrov R. Representing a scientific field: A bibliometric approach. Scientometrics. 1989;15(5):593–605. http://dx.doi.org/10.1007/BF02017072

26. Small H. Macro-level changes in the structure of co-citation clusters. Scientometrics. 1993;26(1):5–20. http://dx.doi.org/10.1007/BF02016789

27. Janssens F, Leta J, Glanzel W, De Moor B. Towards mapping library and information science. Inf Process Manage. 2006;42:1614–1642. http://dx.doi.org/10.1016/j.ipm.2006.03.025

28. Wallin J. Bibliometric methods: Pitfalls and possibilities. Basic Clin Pharmacol Toxicol. 2005;97:261–275. http://dx.doi.org/10.1111/j.1742-7843.2005.pto_139.x

29. Borner K, Mane K. Mapping topics and topic bursts in PNAS. Proc Natl Acad Sci. 2004;101(1):5287–5290. http://dx.doi.org/10.1073/pnas.0307626100

30. Noyons E, Van Raan E. Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. J Assn Inf Sci Technol. 1998;49(1):68–81.

31. Turton A, Meissner R. The hydrosocial contract and its manifestation in society: A South African case study. In: Turton A, Henwood R, editors. Hydropolitics in the developing world: A southern African perspective. Pretoria: African Water Issues Research Unit; 2002.

32. Sutherland W, Fleishman E, Masica M, Pretty J, Rudd M. Methods for collaboratively identifying research priorities and emerging issues in science and policy. Methods Ecol Evol. 2011;2(3):238–247. http://dx.doi.org/10.1111/j.2041-210X.2010.00083.x

33. Pouris A. Peer review in scientifically small countries. R&D Manage. 1998;18(4):333–340. http://dx.doi.org/10.1111/j.1467-9310.1988.tb00608.x

Climate and the mfecane

AUTHORS:
Michael Garstang[1,2]
Anthony D. Coleman[3]
Matthew Therrell[4]

AFFILIATIONS:
[1]Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia, USA
[2]Simpson Weather Associates Inc, Charlottesville, Virginia, USA
[3]Battlescenes Tours, Dundee, South Africa
[4]Department of Geography, University of Alabama, Tuscaloosa, Alabama, USA

CORRESPONDENCE TO:
Michael Garstang

EMAIL:
mxg@swa.com

POSTAL ADDRESS:
Department of Environmental Sciences, University of Virginia, PO Box 400123, Charlottesville, VA 22904-4123, USA

DATES:
Received: 31 July 2013
Revised: 11 Nov. 2013
Accepted: 26 Nov. 2013

KEYWORDS:
southern African drought; Natal history; coupled ocean–atmosphere model; tree-ring analysis; volcanism

HOW TO CITE:
Garstang M, Coleman AD, Therrell M. Climate and the mfecane. S Afr J Sci. 2014;110(5/6), Art. #2013-0239, 7 pages. http://dx.doi.org/10.1590/sajs.2014/20130239

The mfecane is thought to be a massive upheaval and devastation of Nguni tribal chiefdoms in the second decade of the 19th century in what is now KwaZulu-Natal and the Eastern Cape of South Africa. Other historians have challenged this extreme interpretation suggesting that the use of the term mfecane be discontinued. We show that pervasive cycles of drought and cold periods in southern Africa are significantly amplified and extended by volcanic eruptions and that, in particular, the eruption of Tambora in 1815 triggered a prolonged and extreme climatic event which bears all of the characteristics ascribed to the mfecane. These findings are supported by a coupled ocean–atmosphere numerical model and by tree-ring rainfall and sea surface temperature analyses, suggesting that the term mfecane is an appropriate description of a singular climatic event.

Introduction

The term 'mfecane' has been used to describe a period of extreme privation, widespread famine, depopulation and displacement of people over a large area of southeast South Africa (Figure 1). Thought to have begun in the second decade of the 19th century, the phenomenon may have had antecedent contributing conditions and has been seen to have had extensive subsequent sociological and political repercussions potentially extending well into the 20th century.

The range of arguments on the mfecane advanced by southern African historians, cover the full spectrum from the ascription of late 20th century political decisions[1] to the mfecane to abandoning the use of the term mfecane entirely.[2]

The purpose of this paper is not to enter into the complex broader issues implied by or ascribed to the mfecane but to explore, firstly, the inherent large-scale atmospheric controls on the climate of Africa, and, secondly, whether a specific geophysical event combined to precipitate such a human catastrophe. The resulting climatic anomaly led to conditions akin to what may have been described by the original meaning of the term in Xhosa (*ukufaca*) which is 'to be weak, emaciated by hunger', where the stress is upon 'famine'.[2]

Later use of the word mfecane, placing emphasis upon 'extreme famine', would then be entirely consistent with the environmental conditions which we will describe, irrespective of any subsequent consequences or interpretations.

Previous work has not addressed either of the above potential influences or employed such findings to interpret the full impact of environmental factors. Such evidence has not been called upon to identify among multiple causes, a set of circumstances which mark the event as distinct from others and provide a basis for the use of the term 'mfecane'.

We will first draw attention to the unique geographical position of the continent of Africa relative to global-scale circulations of the atmosphere. Two large-scale global circulations dominate the climate of the subtropical latitudes of Africa, inducing pervasive dryness and periodic, if not entirely predictable, decadal cycles of wet and dry conditions.

Secondly, we will examine the superimposition of the effects of the world's largest known volcanic eruption upon the larger-scale climatic periodicities as the critical contributing factor to the occurrence of extreme climatic conditions in the region of interest. Precipitation anomalies derived from a large-scale coupled ocean–atmosphere model will be interpreted for the time period and location.

Sea-surface temperatures (SST) in the adjacent Indian Ocean that coincide with dry and wet conditions derived from a tree-ring analysis confirm the model-predicted SST fields and precipitation anomalies.

Large-scale atmospheric circulations over Africa

The continent of Africa is centred upon the geographic equator, extending approximately 35 degrees of latitude polewards in each hemisphere (Figure 1). This apparently trivial geographical detail has in fact profound implications for the continent of Africa. Africa is indeed unique among the continents in its equator-centric location.

Two major atmospheric circulations – Hadley and Walker – are centred on the equator. The impact upon Africa of these two circulations, underemphasised in the literature, results first in curtailing and second in creating high variability in the rainfall over much of this continent outside of the equatorial region. About 75% of the continent receives less than 500 mm of rainfall per year: a direct product of the Hadley circulation. This meager rainfall is subject to high variability amplified by the periodic dry and wet decadal cycles induced by the Walker circulation.

The Hadley circulation

Surface wind fields driven by solar heating over the meteorological equator and the deflecting forces (Coriolis) of the rotating earth, result in the converging flow of the northeast and southeast trade winds. Figure 2a shows the Hadley circulation in its simplest but, for our purposes, most important form: mass convergence at the surface along the meteorological equator results in upward motion, cooling and condensation of the warm moist equatorial air, the formation of deep convective clouds and the production of the high rainfalls of the equatorial region approximately 15°N and S of the geographic equator. The descending poleward limbs of the Hadley cell, initially caused by high altitude (15 km) radiative cooling and increased density of the air, result in sinking, compression,

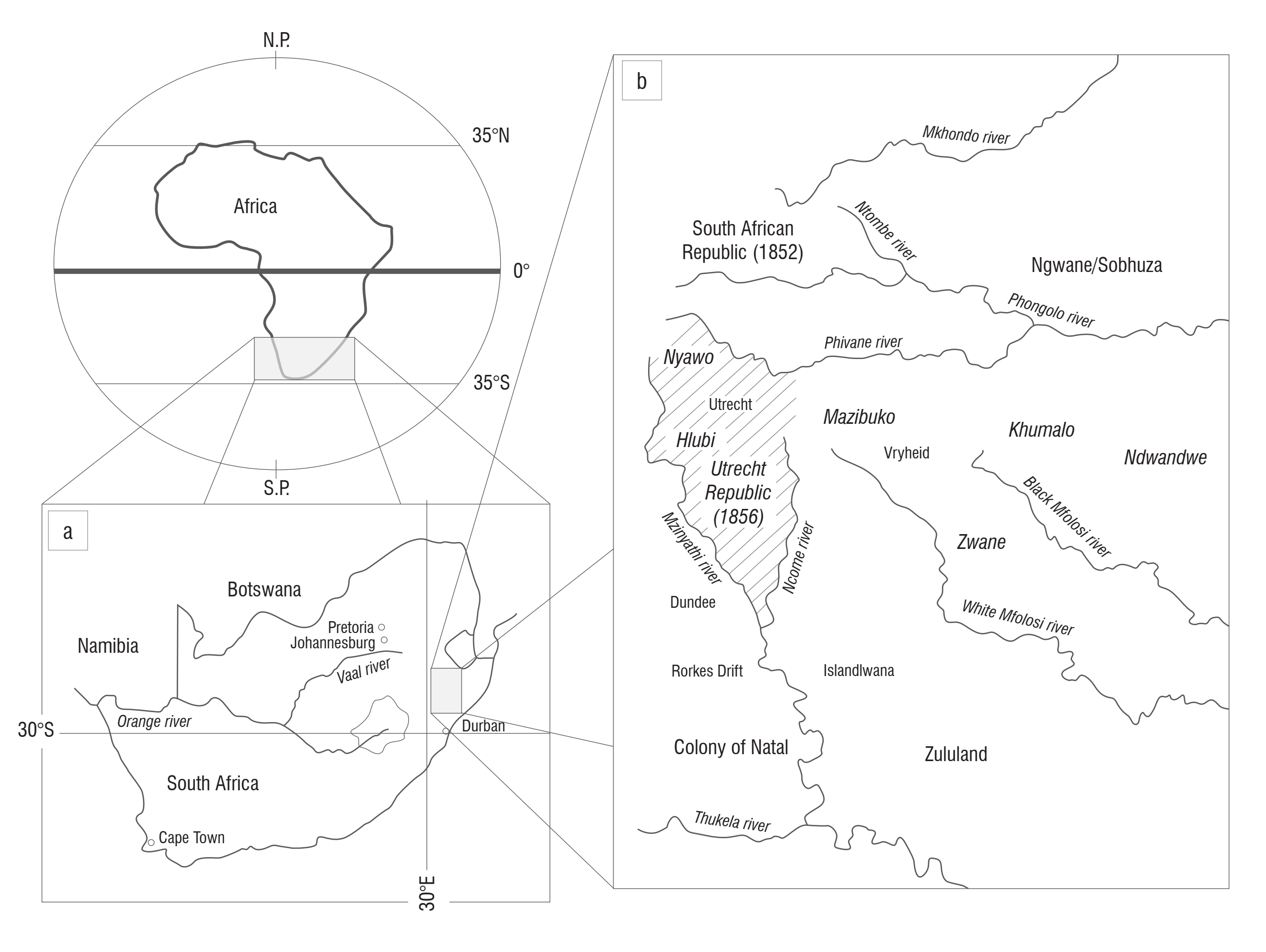


Figure 1: (a) The continent of Africa together with contemporary political boundaries and key cities, rivers and mountains of southern Africa and (b) the location of the Colony of Natal and Zululand together with tribal groups in the early to mid-19th century.

drying and warming of the atmosphere and the formation of the semi-permanent subtropical anticyclones and the large hot deserts of Africa.

Figure 2b shows the more complex manifestation of the Hadley circulation in the form of the Intertropical Convergence Zone (ITCZ) and the Zaire Air Boundary (ZAB) and their seasonal migrations.[3] With the exception of the rain-producing regions of the convergence zones, the dominant effect of the Hadley circulation over Africa is a suppression of rainfall. The result is that much of Africa is dry, with an annual rainfall unable to support agriculture without supplemental (irrigation) water.

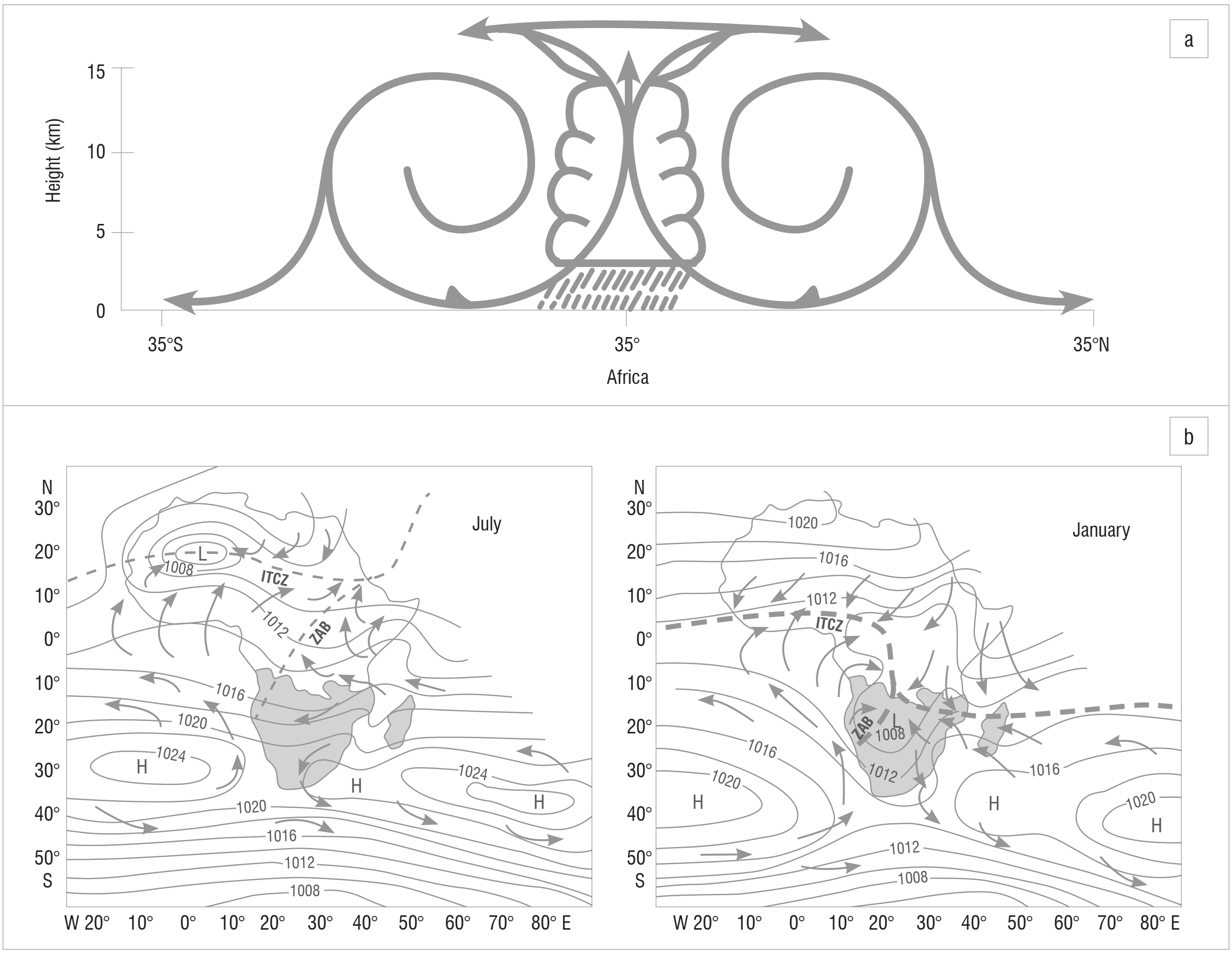
For rain to have been produced in the region of 19th-century Natal (Figure 1), the pervasive downward motions of the Hadley cell must have been countered by generation of upward motion. On the large scale, poleward excursions of the ITCZ can bring cyclonically circulating air in the form of easterly waves or equatorial vortices or even remnants of tropical cyclones (hurricanes) into the region. These phenomena are irregular and highly seasonal, occurring most often at the height of summer. Failing organised weather systems, daytime heating of the surface in summer, in the presence of moist air inflow from the Indian Ocean (warm Agulhas current), can develop afternoon thunderstorms and rain and could have made the coast of Natal wetter than the interior. Only rarely in the winter will frontal systems or low-pressure areas originating over the Southern Oceans, and reaching the Cape, penetrate into and beyond what was 19th-century Natal (Figure 1).

The Walker circulation

Extending around the world, centres of upward and downward motion form a chain of wet and dry regions located along the meteorological equator (Figure 3).[4] Warming and cooling of the sea surface along this equatorial band, particularly in the far eastern Pacific Ocean, together with land masses create a longitudinal distribution in the centres of wet and dry conditions. When the waters of the far eastern Pacific warm, the cold upwelling off the coast of central America is replaced with warm surface water, nutrients from the depths are cut off and fishing ends. The fishermen of the region call this the El Niño. This warming of the surface water is the low phase of the Walker circulation shown in Figure 3a.

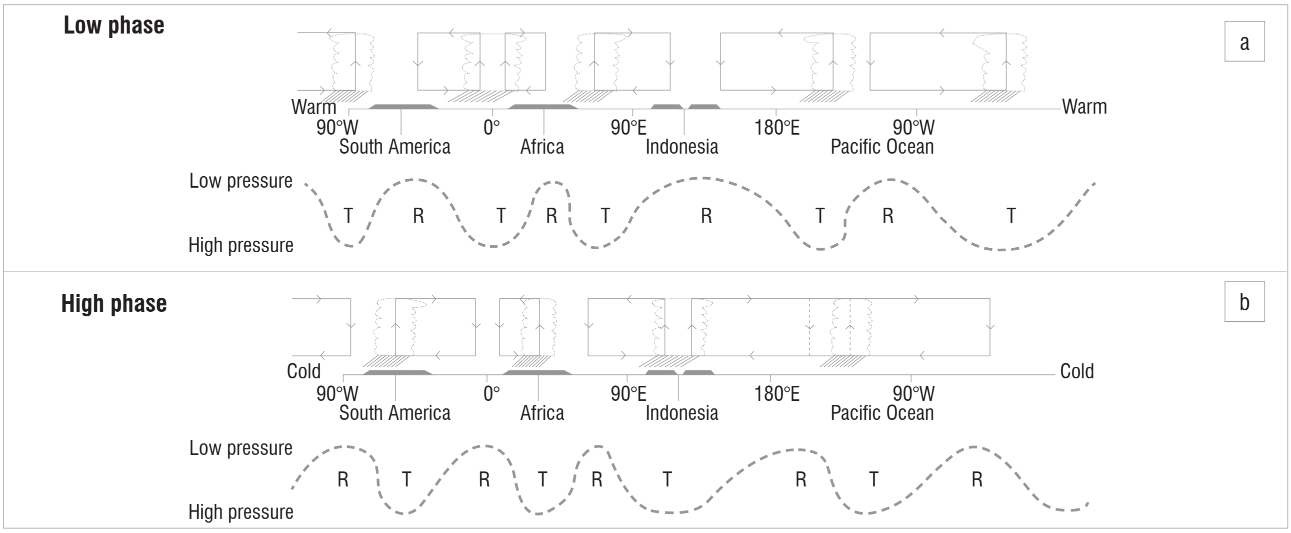
When the warm waters of the eastern Pacific retreat to be replaced by cold upwelling water, the La Niña has returned with the high phase of the Walker circulation (Figure 3b). The effects of the Pacific Ocean oscillations are transmitted around the globe through changes in the surface pressure fields as shown in Figure 3. The surface pressure patterns result in alternating regions of wet (trough) and dry (ridge) conditions within the subtropics of the globe. Figure 3 shows that southern Africa would experience wet conditions under the high phase (La Niña) and dry conditions under the low phase (El Niño).

The east/west Walker circulation and the north/south Hadley circulation, are thus central to Africa. The periodic oscillations of trough (wet) and ridge (dry) pressure patterns extending into both hemispheres of Africa result in near decadal wet and dry periods that have been isolated for the summer rainfall regions of southern Africa.[5] Other climatic forcing functions embedded within the adjacent Atlantic and Indian Oceans reinforce or interfere with the above climatic controls, increasing variability and reducing predictability.



Source: Hobbs et al.[3], p. 10

Figure 2: (a) Schematic of the Hadley circulation over Africa in the form of a vertical cross-section from the surface to 15 km extending along the north/south length of Africa. Ascending air, cloud and rain dominate the equatorial region; descending air, drying and warming dominate the greater part of the continent. (b) The mean surface pressure (mb) and wind (heavy arrows) fields for each season: July, northern hemisphere summer/southern hemisphere winter, and January, northern hemisphere winter/southern hemisphere summer. The Intertropical Convergence Zone (ITCZ) and the Zaire Air Boundary (ZAB) are shown by heavy dashed lines.



Source: Modified after Garstang and Fitzjarrald[4]

Figure 3: East to west globally encircling Walker circulation centred on the meteorological equator resulting in a trough (T)–ridge (R) oscillation in the surface pressure fields with accompanying wet (rain) and dry (drought) regions extending globally around the equator. The low phase (a) occurs when the far eastern Pacific warms (El Niño). The high phase (b) occurs when the far eastern Pacific cools (La Niña).

Sea-surface temperatures

The El Niño–Southern Oscillation (ENSO) initiates changes in SST in the far eastern equatorial eastern Pacific, triggering the reversals in the Walker Circulation.[6,7] Atlantic SSTs strongly influence the position of the ITCZ over Africa.[8,9] Both observations and model studies suggest that the ITCZ moves away from cooler SSTs, potentially reducing the rainfall.

Figure 4a, based on the 10 driest years according to tree-ring analysis in Zimbabwe from 1900 to 2000, shows colder ocean temperatures off the east coast of southern Africa compared to the 10 wettest years (Figure 4b),[10] supporting the model results shown in Figure 5 of a 20% decrease in annual precipitation.

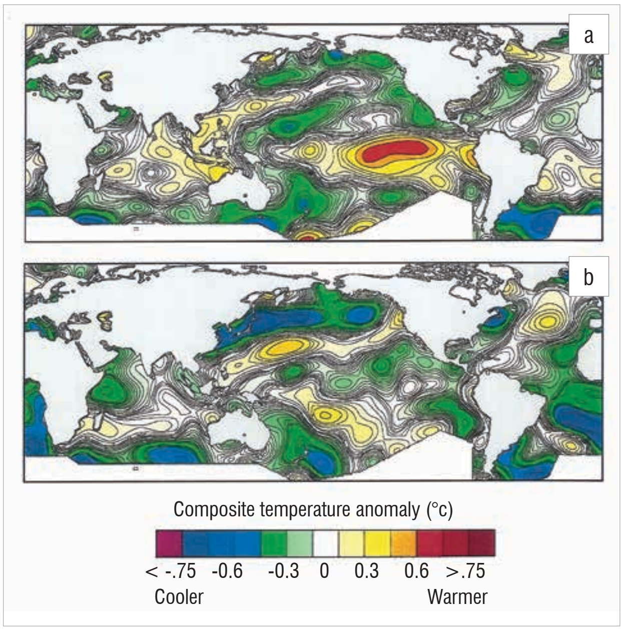


Figure 4: Composite sea-surface temperature anomalies for the (a) 10 driest and (b) 10 wettest years in the reconstructed rainfall record shown in Figure 6.

Role of volcanism

Numerous researchers, notably Lamb[11] and Bryson and Goodman[12], have pointed to volcanism as a major factor influencing climate, including SST and precipitation. More recently, Haywood et al.[13] and others[14,15] have modelled the climatic effects of large explosive volcanic eruptions. They concluded that large asymmetrical stratospheric aerosol loadings produced by the volcanoes and concentrated in the northern hemisphere result in droughts in the Sahel. Haywood et al.[13], using the United Kingdom's Meteorological Office's coupled global atmosphere–ocean Hadley model (HadGEM), were able to show that the four driest Sahelian summers during the period 1900–2010 were preceded by substantial northern hemisphere volcanic eruptions.

We drew upon simulations conducted at the Hadley Centre to assess the impact of Tambora upon the climate of southeastern southern Africa. To do this, we used Hadley model simulations of volcanic eruptions whose optical depths most closely reflect conditions believed to have been present in the Tambora eruption. These optical depths are primarily controlled by the amount and height to which sulphur particles are injected into the atmosphere. These results are then used to predict the percentage change in the precipitation over southern Africa. The result (Figure 5) is conditioned by the fact that the volcanic explosivity index (VEI) of Tambora is two orders of magnitude greater than the VEI used for the simulation.

On 15 August 1815, Tambora – a volcano on the island of Simbawa in Indonesia – erupted, expelling some 140 gt of magma.[16] At position seven on the nine-point logarithmic VEI, Tambora is the largest known historical eruption. Located at 8.2°S latitude, 118°E longitude and lofting some 60 mt of sulphur to an altitude of 45 km, well into the stratosphere, this plume entered both the southern and northern hemispheres. Forming a global aerosol sulphate veil, reflecting short-wave radiation, there were pronounced climate repercussions in both hemispheres. Extending in the northern hemisphere as far north as Canada and Europe, 1816 was known as the 'year without a summer'.[16] With the volume and character of the aerosols (high sulphur yield) above the stable barrier of the tropopause, it is estimated that the effect of this veil may have persisted for up to 5 years. The shortest growing season (less than 50% of normal) in southern Maine, New Hampshire and eastern Massachusetts between 1790 and 1840 was recorded for the year of 1816.[16]

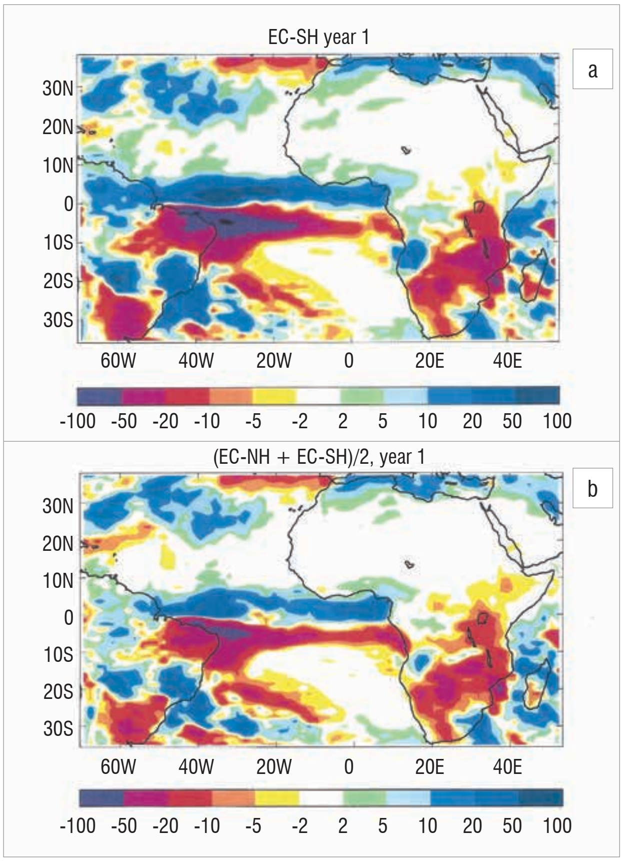


Figure 5: Model results from HadGEM for 1 year following an El Chichon-scale eruption for (a) injection of the aerosol load into the southern hemisphere and (b) injection of the aerosol load into both hemispheres, showing the per cent of induced change in annual precipitation along the lower colour-coded bar.

Evidence of the impact of Tambora on the climate of the southern hemisphere following the 1815 eruption is much less detailed than that for the northern hemisphere. Grab and Nash[17,18] used documentary evidence to study climatic variability during cold seasons in Lesotho between 1833 and 1900. They found, however, that the coldest period during the 19th century was the years following the 1815 eruption of Tambora. An unknown eruption preceded Tambora in 1809.[19] While Grab and Nash[17] were unable to document in detail the consequences of these two eruptions in southeastern southern Africa, they were able to describe the consequences of the eruptions of Amagura, also known as Toku, in the Fijian Island group (18°S, 174°W), in June 1846, and Krakatau (6.1°S, 105.4°E) in August 1883. Both of these eruptions were of lesser magnitude measured on the VEI than Tambora (Tambora 7, Amagura estimated at 4 and Krakatau 6). For the lowlands of Lesotho, Grab and Nash[17] found three consecutive years of very severe winters and one additional year of severe winter from 1847 to 1850 following the eruption of Amagura. These winters resulted in extensive loss of human life and livestock. Despite the low sulphur yield of Krakatau, four consecutive years (1884–1887) of severe winters and early frosts followed that eruption.

The Hadley Centre's coupled atmosphere–ocean model results for a volcano the size of El Chichon, which injected 7–12 $MtSO_2$ into the stratosphere, were then used to estimate the impact upon the precipitation of southern Africa. This volcano has been estimated to be the second most climatically important volcano of the last 50 years.[20] Simulations were performed to determine the potential impact on southern Africa for aerosol injections into (1) the northern hemisphere, (2) the southern hemisphere and (3) both hemispheres (see Haywood et al.[13] for further details). Figures 5a and 5b show the results for December–April rainfall. For this relatively modest eruption, there was a robust drying of 10–50 mm/month, equivalent to about a 20% decrease in annual mean rainfall, regardless of whether the injection was in both hemispheres or in one hemisphere only. For an eruption as massive as Tambora (60 $MtSO_2$), these model results suggest that the impact across much of the summer rainfall region of southern Africa could have been far more severe, both in terms of the reduction in rainfall and the duration of the impact, which could have been 4 to 5 years following the eruption owing to the longer residence time of aerosol injected to higher regions of the stratosphere.

Of interest is the model result which suggests that the coastal margin of southeastern southern Africa might, even under the extreme conditions created by Tambora, not have been as severely impacted as the interior of the subcontinent. This possibility, in turn, suggests that the more coastal dwelling population may have escaped the worst consequences of the drought, but further multi-model assessments would be needed to assess the robustness of this result.

Nevertheless, the conclusions drawn from the Hadley model results, supported by the tree-ring based SSTs which indicate cooling and drying, suggest that the impact of Tambora on the climate of the region could have significantly exceeded that suggested by the above evidence. It may be added that the methodology employed represents one of the few ways in which climatic aberrations may be illuminated in a location and at a time which has little other supporting material available.[17,18]

Hall[21], using oral history and the analysis of tree rings, suggested that there were three major changes in climate in southeastern southern Africa during the 18th and 19th centuries: 1700–1750 (dry), late 1700s (wet) and 1800–1820 (dry). The drought in the early part of the 19th century was referred to by the Nguni people as the 'madlathule' (let one eat what he can and say naught)[22] and is coincident with a serious breakdown of social, political and economic institutions in the region.[23]

The tree-ring work in Zimbabwe[10] confirms the dry period identified by Hall[21] in the first two decades of the 19th century (Figure 6) as well as the extended dry period coincident with the eruption of Krakatau (1883) described by Grab and Nash[17]. The rainfall reconstructed from our tree-ring data shown in Figure 6 suggests that the effects of the unknown eruption in 1809, followed by that of Tambora, combined to produce an extended period of 15 years of drought in the southeastern region.

Of the six prolonged drought periods in the reconstructed rainfall record shown in Figure 6, three are associated with volcanism (1809–1823: unknown eruption in 1809 and Tambora in 1815; 1838–1870: Amagura in 1846; 1880–1896: Krakatau in 1883). The drought of the 1920s, which lasted into 1933, is well known for its devastation of farming in Natal but was not associated with any volcanism.[24]

The accumulated evidence cited above suggests that the combined eruptions of the 1809 volcano and Tambora in 1815 created devastating drought and cold conditions in southeastern southern Africa for much of the second decade of the 19th century. These conditions initiated by the 1809 event, became extreme after 1815 and likely extended over much of the summertime rainfall region of southeastern and central southern Africa, reducing rainfall and the growing seasons to less than half of the mean annual values (Figures 5a and 5b).

Nicholson et al.[25] show a ubiquitous period of aridity in the earliest years of the 19th century (see Figure 10 in Nicholson et al.[25]) based upon a composite precipitation data set compiled for Africa, including the Kalahari, over the last two centuries. Holmgren et al.[26], using isotopes of oxygen in stalactites from the Limpopo Province of South Africa, show extreme dry conditions (see Figure 4 in Holmgren et al.[26]) prevailing at the end of the 18th century.

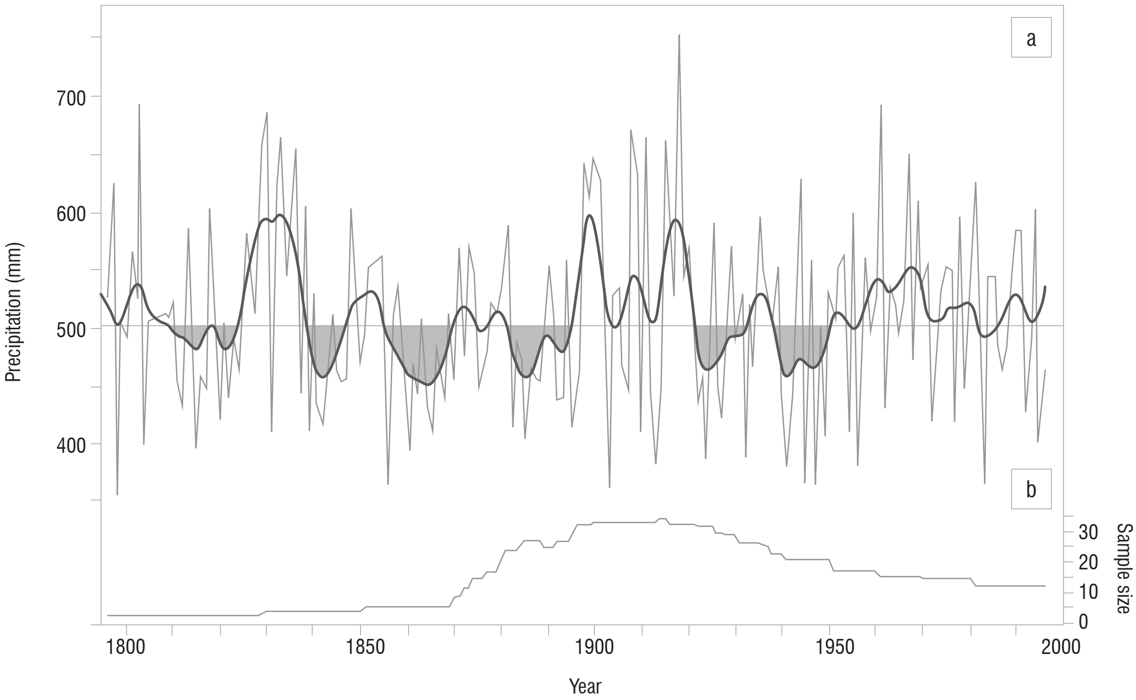


Figure 6: (a) Annual (thin line) and 10-year smoothing spline (heavy line) of reconstructed November–February regional rainfall for Zimbabwe according to tree-ring data from 1796 to 1996. (b) Sample size over time from 21 trees. The shading indicates six major decadal-length dry periods in Zimbabwe during the 19th and 20th centuries.

Other environmental factors

Ballard[23], Guy[27], Mcl Daniel[28], Eldredge[29] and Gump[30] have all drawn attention to the role played by agricultural practices, soil characteristics and fertility and the carrying capacity of the pastures of the region east and southeast of the Drakensberg towards the Indian Ocean coast. Acocks'[31] reconstruction of vegetation for the above region from the 15th to the 20th centuries suggests that, over time, good grazing is replaced by poorer species in the wetter regions and not at all in the drier regions. Both Guy[27] and Gump[30] argue that significant environmental degeneration had occurred by the end of the 18th century in the Zululand region.

Reduction in carrying capacity progressively limited the time of year and the area available for grazing while the increasing numbers of livestock and people heightened the pressure on these natural resources. By the early 19th century, Wilson[32] estimated that the number of cattle in the Zululand region likely exceeded the number of people.

Maize may have largely replaced millet and sorghum by the beginning of the 19th century in the Zululand region. Millet and sorghum are drought resistant, grow well in relatively poor sandy soils, require little rain, mature quickly and are suited to short growing seasons. Both crops store well and provide sustenance over multiple seasons. Maize, while possessing some advantages over millet and sorghum such as higher crop yields, has serious relative disadvantages. Maize requires higher rainfall than either millet or sorghum, and requires well-drained soils high in nitrogen, phosphorus and potash. Without supplemental nutrients, maize quickly exhausts soils. Maize is also subject to insect damage while stored.

While cumulatively over time, all of the above factors eventually lead to ecological disequilibrium, such a change is gradual rather than sudden and is spatially inhomogeneous. The exception to such a response could be the result of a sudden occurrence of extreme conditions such as a drought, which could precipitously trigger the simultaneous collapse over a large region.

Summary and conclusions

The rainfall regime of southern Africa, encompassing most of present day KwaZulu-Natal, the far Eastern Cape, the lowlands of Lesotho and much of the interior plateau, is subject to the large-scale equatorial circulations of the earth's atmosphere. The primary effect of the large-scale circulations upon the rainfall of the region is to impose pervasive dry conditions with a high level of variability. Such variability can take the form of decadal fluctuations in rainfall resulting, alternately, in droughts and floods.

Irregular volcanic activity, occurring particularly in the global equatorial and tropical regions, introduces an anomalous response in rainfall, potentially interfering with the larger-scale rainfall production. The causal relationship between rainfall and the eruption is a cooling effect induced by the injection of particles from the eruption into the stratosphere and a regional cooling of the sea surface.

Dry conditions as a result of the large-scale circulations can be reinforced by a volcanic eruption and can lead to extreme and protracted drought. Reduction in solar insolation reduces surface temperatures. The suppression of rainfall, the absence of clouds and the presence of low humidity result in cold nights with severe frost, reinforcing the already cold conditions. Damage to crops and pasture and the curtailment of the growing season follow.

Such extreme changes in climate brought about by the coincidence of large-scale atmospheric scales of motion and a major global tectonic event, occur within a period of months or less, and over large spatial areas much greater in extent than the above described region.

Observational and model calculations of the impact of a major volcanic eruption support the above time and space characteristics as well as the magnitude of the resulting drought. These characteristics, while amplified by ecological disequilibrium, occur with a suddenness and uniformity over a wide area not associated with a gradual and inhomogeneous environmental degradation. While such an event follows upon and is connected to both antecedent and subsequent conditions, it differs from these 'before' and 'after' states by the uniqueness of its time and space characteristics.

The magnitude and temporal and spatial identity of the change in climate over Zululand and the adjacent regions as a result of the eruption of Tambora constitute a clearly recognisable and unusual state. Historical evidence of such a state, mainly in the form of extremely dry and cold conditions with associated famine and mortality, would support the recognition of these conditions as those described for the 'mfecane'.

Climatic disruptions in the dry tropics of Africa continue into modern times. Loss of life to famine caused by drought occurs from Somalia through the Sahel to the west coast of North Africa. Volcanic eruptions have preceded three of the four driest Sahelian summers during the past 110 years. Hundreds of thousands of people have lost their lives and tens of millions have been displaced by droughts in the subtropics of north Africa.[13] The lessons of the mfecane remain as important today as they were two centuries ago. The place and role of the mfecane should remain as marking a singular geophysical and human event in the history of southeastern southern Africa.

Acknowledgements

We thank Prof. Jim Haywood and Dr Andrew Jones of the Hadley Centre of the United Kingdom's Met Office for their assistance in running simulations of volcanic eruptions on the Centre's coupled atmosphere–ocean general circulation model (HadGEM), work supported by the Joint DECC and Defra Integrated Climate Programme, DECC/DEFRA (GA01101). We also thank Prof. Joe Miller for scholarly guidance on the manuscript and Dr David Hondula for preparation of the figures (both at the University of Virginia) and Mary Morris for bibliographic research, and preparation and submission of the manuscript.

Authors' contributions

M.G. carried out the meteorological and climatological analysis and compilation including initiating the model simulation of the impact of the volcanic eruption on the climate of southern Africa using the United Kingdom's Meteorological Office's coupled ocean–atmosphere model. M.G. also assembled, submitted and served as corresponding author on the paper. A.D.C. initiated the research into the impact on the climate of southeastern southern Africa of the eruption of Tambora in 1815 and contributed to sections on the changes in crops and agriculture and the role of drought on the people and crops of the region. M.T. carried out the tree-ring analyses of rainfall in Zimbabwe and produced the climatological record of the sea surface temperatures for the 10 driest and 10 wettest years over the period 1857–1996.

References

1. Omer-Cooper JD. The Zulu aftermath: A nineteenth century revolution in Bantu Africa. London and Ibadau: Longman; 1966.
2. Cobbing J. The case against the mfecane. Paper no 144:1984:27. University of the Witwatersrand African Studies Institute Seminar Series. Seminar paper presented in Johannesburg; March 1984.
3. Hobbs JE, Lindesay JA, Bridgman HA, editors. Climates and the southern continents: Present, past and future. West Sussex: John Wiley and Sons; 1998.
4. Garstang M, Fitzjarrald DR. Observations of surface to atmosphere interactions in the tropics. New York: Oxford University Press; 1999.
5. Tyson PD. Climatic changes and variability in southern Africa. Cape Town: Oxford University Press; 1986.
6. Kane RP. Periodicities, ENSO effects and trends of some South African rainfall series: An update. S Afr J Sci. 2009;105:199–207.
7. Stuecker MF, Timmerman A, Jin FF, McGregor S, Ren HL. A combination mode of the annual cycle and the El Niño/Southern Oscillation. Nat Geosci. 2013;6:540–544. http://dx.doi.org/10.1038/ngeo1826
8. Chang P, Zhang R, Hazeleger W, Wen C, Wan X, Ji L, et al. Oceanic link between abrupt changes in the North Atlantic Ocean and the African monsoon. Nat Geosci. 2008;1:444–448. http://dx.doi.org/10.1038/ngeo218

9. Shanahan TM, Overpeck JT, Anchukaitis KJ, Beck JW, Cole JE, Dettman DL, et al. Atlantic forcing of persistent drought in West Africa. Science. 2009;324:377–380. http://dx.doi.org/10.1126/science.1166352

10. Therrell MD, Stable DW, Ries LP, Shugart HH. Tree-ring reconstructed rainfall variability in Zimbabwe. Clim Dynam. 2006;26:677–685. http://dx.doi.org/10.1007/s00382-005-0108-2

11. Lamb HH. Volcanic dust in the atmosphere: With a chronology and assessment of its meteorological significance. Philos Trans R Soc Lond A. 1970;266:425–533. http://dx.doi.org/10.1098/rsta.1970.0010

12. Bryson R, Goodman BM. Volcanic activity and climatic changes. Science. 1980;207:1041–1044. http://dx.doi.org/10.1126/science.207.4435.1041

13. Haywood JM, Jones A, Bellouin N, Stephenson D. Asymmetric forcing from stratospheric aerosols impacts Sahelian rainfall. Nat Clim Change. 2013;3:660–665. http://dx.doi.org/10.1038/nclimate1857

14. Joseph R, Zeng N. Seasonally modulated tropical drought induced by volcanic aerosol. J Clim. 2011;24:2045–2060. http://dx.doi.org/10.1175/2009JCLI3170.1

15. Timmreck C. Modeling the climatic effects of large explosive volcanic eruptions. WIREs Clim Change. 2012;3:545–564. http://dx.doi.org/10.1002/wcc.192

16. Oppenheimer C. Climatic, environmental and human consequences of the largest known historic eruption: Tambora volcano (Indonesia) 1815. Prog Phys Geogr. 2003;27:230–259. http://dx.doi.org/10.1191/0309133303pp379ra

17. Grab SW, Nash DJ. Documentary evidence of climate variability during cold seasons in Lesotho, southern Africa 1833–1900. Clim Dynam. 2010;34:473–499. http://dx.doi.org/10.1007/s00382-009-0598-4

18. Nash DJ, Grab SW. A sky of brass and burning winds: Documentary evidence of rainfall variability in the Kingdom of Lesotho, Southern Africa, 1824–1900. Clim Change. 2010;101:617–653. http://dx.doi.org/10.1007/s10584-009-9707-y

19. Betraud C, Van Ypersele JP, Berger A. Volcanic and solar impacts on climate since 1700. Clim Dynam. 1999;15:355–367. http://dx.doi.org/10.1007/s003820050287

20. Haywood JM, Jones A, Clarisse L, Bourassa A, Barnes J, Telford P, et al. Observations of the eruption of the Sarychev volcano and simulations using the HadGEM2 climate model. J Geophys Res. 2010;115(D21). http://dx.doi.org/10.1029/2010JD014447

21. Hall M. Dendroclimatology, rainfall and human adaptation in the later iron age of Natal and Zululand. Ann Natal Mus. 1976;22:693–703.

22. Coleman AD, Garstang M. The disputed territory: A cause of the Anglo-Zulu war – Re-examined. Adv Hist Studies. Forthcoming 2014.

23. Ballard C. Drought and economic distress: South Africa in the 1800s. J Interdis Hist. 1986;XVII(2):359–378. http://dx.doi.org/10.2307/204770

24. Arnold P. Tom and Ethel: The story of a soldier settlement. Pietermaritzburg: Pam Arnold; 1990.

25. Nicholson SE, Dezfuli AK, Klotter D. A two-century precipitation dataset for the continent of Africa. Bull Amer Meteor Soc. 2012;93:1219–1231. http://dx.doi.org/10.1175/BAMS-D-11-00212.1

26. Holmgren K, ŐBerg H. Climate change in southern and eastern Africa during the past millennium and its implications for societal development. Environ Dev Sustain. 2006;8:185–195. http://dx.doi.org/10.1007/s10668-005-5752-5

27. Guy J. Ecological factors in the rise of Shaka and the Zulu Kingdom. In: Marks S, Atmore A, editors. Economy and society in pre-industrial South Africa. London: Longman; 1980. p. 102–119.

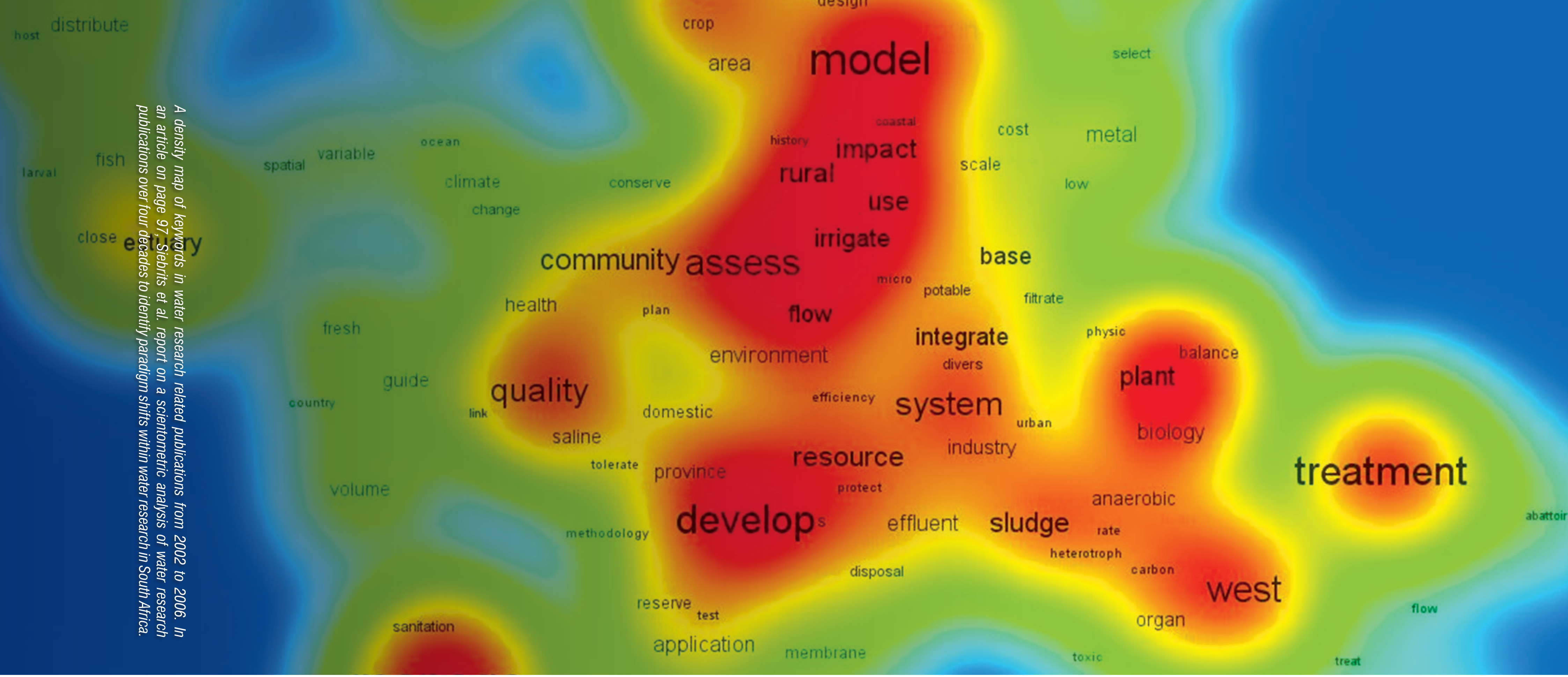
28. McI Daniel B. A geographical study of pre-Shakan Zululand. S Afr Archeol Bull. 1973;55:23–31.

29. Eldredge EA. Drought, famine and disease in nineteenth-century Lesotho. Afr Econ Hist. 1987;16:51–93.

30. Gump J. Origins of the Zulu Kingdom. The Historian. 1988;50:525–527. http://dx.doi.org/10.1111/j.1540-6563.1988.tb00757.x

31. Acocks PH. Veld types of South Africa. Memoirs of the Botanical Survey of South Africa. 1953;28:13.

32. Wilson M. The Nguni people. In: Oxford history of South Africa. London: Oxford University Press; 1968. p. 107–108.



A density map of keywords in water research related publications from 2002 to 2006. In an article on page 97, Siebrits et al. report on a scientometric analysis of water research publications over four decades to identify paradigm shifts within water research in South Africa.

APPLYING SCIENTIFIC THINKING IN THE SERVICE OF SOCIETY

Our vision is to be the apex organisation for science and scholarship in South Africa, internationally respected and connected, its membership simultaneously the aspiration of the country's most active scholars in all fields of scientific enquiry, and the collective resource for the professionally managed generation of evidence-based solutions to national problems.

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

T +27 12 349 6600/21/22 | **F** +27 86 576 9514

WWW.ASSAF.ORG.ZA