SOUTH AFRICAN Journal of Science

volume 111
number 7/8

New satellite-based weather forecasting tool for data sparse regions in Africa

Pancreatic islet regeneration – a potential cure for diabetes?

Directed genetic modification of African horse sickness virus by reverse genetics

Fertiliser potential of fly ash – a waste product of coal-fired electricity generation

Polarisation-encoded quantum key distribution: the future of data security

JULY/AUGUST 2015

ISSN: 0038-2353

ASSAf
ACADEMY OF SCIENCE OF SOUTH AFRICA

JULY/AUGUST

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

SOUTH AFRICAN
Journal of Science

volume 111

number 7/8

Leader

News & Views

Scientific Correspondence

Book Review

Review Article

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

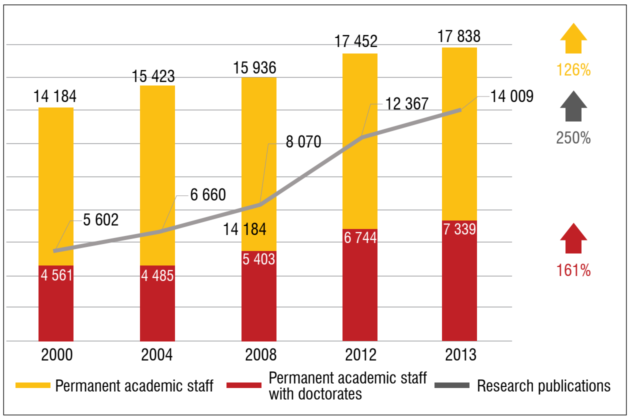
Research Article

An investment in knowledge....producing new interest

An investment in knowledge always pays the best interest.

Benjamin Franklin

When the Department of Education – now the Department of Higher Education and Training (DHET) – replaced the 'SAPSE-110' higher education funding system with the 'New Funding Formula' (NFF) in 2004, it is unlikely that they imagined it would give rise to an entire new industry within university administrations: the management of journal and book publication outputs by academic committees and by research support departments. Yet this is exactly what happened. What also happened was a dramatic rise in research publications and the graduation of students with master's and PhD degrees. Charles Sheppard[1] of the Nelson Mandela Metropolitan University gathered data from the DHET's HEMIS (Higher Education Management Information System) and provided the overview of research outputs shown in Figure 1.



Source: Sheppard[1]

Figure 1: Research output of South African universities.

In brief, publication outputs increased by 18.9% between 2000 and 2004, by 30.7% between 2004 and 2008 after the NFF had been introduced, and then by a further 53.2% in the years between 2008 and 2012 – an increase of 250% over the entire period covered by Sheppard's data. And while the number of academic staff increased by 126% over the review period, the number of academic staff members with doctorates increased by 161%.

Comparative data for the periods prior to 2000 are not easily found, but Bawa and Mouton[2] published the publication outputs for the years 1990 to 1998 (Figure 2).

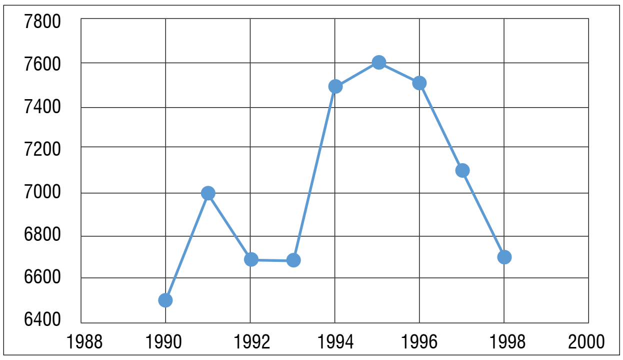
Bawa and Mouton[2] advance several hypotheses, drawn from different sources, to explain the fall of productivity between 1991 and 1992, and again between 1995 and 1998 – but the implication to be drawn from the data provided by Sheppard[1] is that the decline continued through to 2000, then initially rose slowly but gathered considerable momentum after the introduction of the NFF.

Apparent correlations between calendar dates, education policies and research productivity do not prove inevitable causality – but it would be hasty to ignore possible coincidence in the absence of other possible explanations. At the very least, the data would seem to imply that investment in knowledge production can prove to be a positive incentive.

Over the years since 2004, the rules, requirements and processes on which the award of research output funds has been based have changed, and the role of agencies has changed, with the Academy of Science of South Africa (ASSAf) taking on the major assessment and recommendation activities of 2014 and 2015.

In March this year, more substantial changes were announced, which will take effect from the 2016 round of assessments and grants. Details of the changes have been summarised by most universities, but the broad picture that emerges from the March policy is a thought-provoking reflection on current conditions in the world of academic cooperation and publishing.

The section that deals with journals, for instance, reminds universities to be aware of predatory journals, with, it would seem, the intention of alerting researchers to these scams whilst also ensuring that papers in these journals are not listed for awards. More stringent conditions are also specified for the recognition of accreditation of journals: local journals are encouraged, through their Editors-in-Chief, to apply for local accreditation even if they are not Web of Science or Scopus listed – and to ensure that they meet the DHET accreditation requirements.



Source: Bawa and Mouton[2]

Figure 2: Scientific output of South African universities from 1990 to 1998.

Conference proceedings, keynote addresses, works in progress, short papers, brief communications and technical notes may now also be considered for awards, provided that the conference is approved by the DHET. In addition, more than 60% of contributions published in conference proceedings must emanate from multiple institutions.

The new conditions for awards for books are, perhaps, the most onerous and, at the same time, the most generous. The maximum subsidy for a book is now 10 units, although the award will depend on the length of the book: at the lower end, a book length of 60–90 pages will generate 2 units, while at the upper end (300 or more pages), a book will earn the maximum 10 units. A chapter in a book will qualify for two units. Books (and their chapters) will, however, require a written justification (of no more than 500 words) signed by the author of the book, or the general editor, explaining the contribution that the book makes to scholarship. This justification may not be an abstract of the contents or preface of the book, but must, rather, describe the methodology used as well as the unique contribution made to knowledge production. The justification will also need to clarify that the book or chapter for which subsidy is claimed disseminates original research and new developments within the specific discipline.

HOW TO CITE: Butler-Adam J. An investment in knowledge....producing new interest. S Afr J Sci. 2015;111(7/8), Art. #a0117, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0117

The justification must also include an unambiguous declaration that no part of the work was plagiarised or published elsewhere, and specify the target audience. A statement from the institution's evaluation committee is also required, indicating that it has thoroughly checked both the previous and current editions and affirming that at least 50% of the work has not been published previously, and that the book has been peer reviewed both before and after publication.

For scholars, there remain two cautionary notes. University screening committees are tasked with screening submissions, but they are not obliged to accept those submissions, even if they appear to meet all of the DHET requirements – and the ASSAf/DHET teams are not bound to agree with the institutional committees. Most critically, the pool of publication award funds does not increase each year. So as long as the pie's circumference remains more or less constant, the award slices will diminish in relation to the increase in the number of successful submissions. Today's unit value of ZAR113 000 will, inevitably, become tomorrow's ZAR90 000 or less.

References

1. Sheppard C. Research output of South African universities [HEMIS database]. Cape Town: Centre for Higher Education Transformation; 2013.
2. Bawa AC, Mouton J. Research. In: Cloete N, Maassen P, Fehnel R, Moja T, Gibbon T, Perold H, editors. Transformation in higher education: Global pressures and local realities in South Africa. 2nd ed. Dordrecht: Springer; 2006. p. 195–225. http://dx.doi.org/10.1007/1-4020-4006-7_15

Plantation forestry and invasive pines in the Cape Floristic Region: Towards conflict resolution

AUTHOR:
Brian W. van Wilgen[1]

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

CORRESPONDENCE TO:
Brian van Wilgen

EMAIL:
bvanwilgen@sun.ac.za

POSTAL ADDRESS:
Centre for Invasion Biology,
Department of Botany and
Zoology, Stellenbosch University,
Private Bag X1, Matieland 7602,
South Africa

KEYWORDS:
invasive trees; ecosystem services; water resources; policymaking; land rehabilitation

HOW TO CITE:
Van Wilgen BW. Plantation forestry and invasive pines in the Cape Floristic Region: Towards conflict resolution. S Afr J Sci. 2015;111(7/8), Art. #a0114, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0114

Forests supply important commercial resources in the form of timber for building, furniture and packaging, and pulp for paper products. As human populations grow, the demand for these products is driving substantial growth in plantation forestry. Such plantations are necessary both to meet demand for wood-based products and to protect remaining natural forests from over-exploitation, and they are usually based on fast-growing, alien trees. Globally, forest plantations currently represent 5% of forest cover but they account for 40% of commercial wood and fibre production. As more species of alien trees are introduced and widely planted in novel environments, a proportion become invasive, spreading into adjacent landscapes where they have negative effects on biodiversity and the delivery of ecosystem services.[1] Among commercial forestry species, pines (*Pinus* species) are especially problematic, with at least 19 invasive species in the southern hemisphere, where they cause significant problems.[2] The fact that pine trees can simultaneously be useful and harmful in the same region has led to 'schizophrenia' in policy formulation and conflict between land managers.[3,4]

In South Africa, formal pine plantations cover 660 000 ha, and invasive stands of pine trees occur in a further 2.9 million ha. Pine invasions are clearly linked to pine plantations[5], and the problems associated with invasive pines are most acutely felt in the fynbos-clad mountain catchments of the Cape Floristic Region (CFR)[6]. Pine-based plantation forestry is an important economic activity in the CFR, providing direct employment and supplying resources to downstream processors. Authorities responsible for conservation in the CFR (notably Cape Nature and South African National Parks) are, however, concerned about the degree of invasion of protected areas and catchments and the ecosystem services that they deliver, setting the scene for potential conflict.

Government's policy responses to the issue have at best been confused. In June 2001, the Cabinet approved the conversion of 44 763 ha of pine plantations in the CFR to other land uses, principally conservation, signalling a policy of retreat from plantation forestry in the CFR. In 2006, the former Department of Water Affairs and Forestry commissioned a study to re-assess the conversion process and its socio-economic impact, and subsequently recommended that 22 402 ha of the plantations be retained and that the remaining 22 361 ha be included in the continuation of the conversion ('exit') strategy.[7] In 2008, Cabinet approved the proposal to retain 22 402 ha, but it was only in 2014 that the Department approached the Industrial Development Corporation (IDC) to advise it on implementation options to replant state forest land. The IDC identified a number of 'key principles' to guide the restructuring and transfer process, namely community empowerment, broad participation, transformation in the sector, and maximising future forest product outputs.

In June 2004, the government also promulgated the *National Environmental Management: Biodiversity Act* (NEM:BA), which among other things makes provision for the management of listed invasive alien species. Government was obliged, in terms of NEM:BA, to publish a national list of invasive species within 2 years. However, and despite many false starts, it was only in 2014 that the Department of Environmental Affairs published new regulations that listed (among others) eight species of pines as invasive aliens. Under these regulations, it is now necessary to obtain a permit to re-afforest new areas with listed species, as well as to continue with normal forestry practices in existing plantations. Permits will place obligations on the recipients to ensure that planted pines do not spread to adjacent areas. A great deal of confusion still exists as to how the process of legalising the remaining forestry activities in the CFR will be taken forward.

Given the confusion, and the potential for conflict, it would clearly be beneficial to bring the various stakeholders together to explore ways in which the situation could be addressed in a constructive manner. With this in mind, WWF (South Africa) convened a meeting of diverse players with interests in the management of pine trees in the CFR. The meeting, held in Stellenbosch on 4 May 2015, included representatives from forestry companies, conservation agencies, science councils and universities. The goal was to explore a range of relevant issues and to initiate a dialogue aimed at finding mutually acceptable and sustainable solutions to the problem. Most participants identified the need to develop a common understanding of issues as a key outcome of the meeting.

A wide variety of topics was introduced by speakers from different organisations. Dean Muruven of WWF (South Africa) began by outlining their concerns regarding the conservation of critical water source areas, stating that 50% of the country's water is generated by just 8% of the land,[8] much of it at risk from invasion by pine trees. David Le Maitre of the Council for Scientific and Industrial Research reviewed the impacts of alien trees on the hydrological cycle, pointing out that invasive trees could potentially reduce water run-off in the CFR by 37% (from 6765 to 4271 million m^3/year)[9] if invasions are allowed to continue unchecked. Dave Richardson, Director of the Centre for Invasion Biology at Stellenbosch University, pointed to the fact that concern about invasions by introduced trees is growing globally, often leading to conflict, and that much research remains to be done before equitable solutions can be found. Matt McConnachie of the Centre for Invasion Biology outlined innovative research that sought to establish the degree to which different sources (including ornamental plantings, windbreaks and plantations) contributed to the invasion problem,[5] estimating that plantations were probably responsible for about half of the current invaded area. Kassie Carstens from Cape Pine stated that new plantings were of different species (*Pinus elliottii*) that did not produce seed in the CFR, and that this may alleviate the problem of invasions, although whether or not this is correct, and to what degree it will help, has yet to be shown. Finally, Fiona Impson from the Plant Protection Research Institute outlined a 10-year research programme that had sought to find a suitable biological control agent that would reduce seed production in *Pinus pinaster.*[10] Although a suitable agent had been identified, concerns have been expressed that it may exacerbate the problems of pitch canker in pines. The agent has not yet been released, and an application for release that would require further public consultation is overdue.

On the practical side, John Scotcher from Forestry South Africa reiterated the forest industry's commitment to sound environmental management, but outlined concerns regarding the new regulations governing alien species under NEM:BA. The regulations would logically replace earlier legislation in terms of the *Conservation of Agricultural Resources Act* (CARA), although it is not clear if this will occur. Currently, plantation forestry in South Africa is licensed under the *National Water Act*, and the *National Water Act* licences are recognised under the CARA regulations. Forestry South Africa has reviewed the new NEM:BA regulations, and identified some uncertainties regarding the 'restricted activities' with which permitted plantations would have to comply (restricted activities include owning, transporting or selling any specimens or derivatives of listed alien species). All forestry operations will have to apply for new permits under NEM:BA to clarify these issues, and until they do so there is uncertainty regarding how they can legally conduct restricted activities.

Steve Germishuizen (representing the Forestry Stewardship Council, FSC) reviewed the processes whereby all major South African plantation forestry operations had been granted certification. FSC certification requires adherence to sound environmental management practices that follow clear principles, taking account of national laws in the country concerned. FSC certification is also necessary for operating in certain markets, and thus although voluntary, is important in terms of trading with forestry products. Principle 10 of the FSC requires explicitly that an applicant for certification 'shall only use alien species when knowledge and/or experience have shown that any invasive impacts can be controlled and effective mitigation measures are in place'[11]. Nonetheless, certification has been granted to South African plantations on the basis of applications that apparently have not adequately dealt with the problem of invasive species.

Susan Steyn from the Department of Agriculture, Forestry and Fisheries (DAFF) outlined the events that had led to a reversal of earlier decisions to exit from forestry in certain parts of the CFR. The process is ongoing, and workshop participants raised some concerns. Firstly, the assessments of the economic viability of forestry as set out in the 2006 report[7] excluded the externalities associated with invasions linked to plantations. They also did not factor in the effects of fires, which are increasing in frequency and will almost certainly impact on the viability of plantation forestry. Secondly, although the authors of the report initially consulted several stakeholders, many from conservation organisations were only consulted for information, and not given the opportunity to review the report before it was finalised. ('…[I]t was not the purpose of this study to engage in wide community consultation, and the stakeholder consultation was predominantly technical in nature…')[7]. This statement is not aligned with the principle of broad consultation. There is also a dichotomy of views on how the remaining plantations earmarked for transfer to conservation should be dealt with. Cape Nature, on the one hand, has been reluctant to assume responsibility for these areas without guarantees for the funding needed to rehabilitate them to an acceptable standard. South African National Parks, on the other hand, has accepted responsibility for some of these areas despite having no funding to rehabilitate them. As DAFF also has no funding to manage these areas, it is either stuck with them, or is handing them to an organ of state equally unable to afford their management.

Workshop participants spent time discussing possible courses of action for taking this initiative forward. One suggestion was that a collaborative management initiative, involving foresters and conservation agencies, should be implemented on a small scale, which, if successful, could lead to the derivation of useful lessons and wider implementation. Another proposition was for forest companies to lease land from the state, and a portion of the lease fees to be used to control invasions in adjacent protected areas or catchments. This was seen as an attractive proposal, although the ability of these funds to make a meaningful impact would have to be assessed. Finally, it was suggested that the problem deserved a thorough, participative, scientific assessment, as was done when the South African government was faced with the development of an acceptable policy for managing elephants.[12] Such assessments are the product of a process that translates existing scientific information into a form usable by policymakers. Assessments have three critical success factors: (1) legitimacy (the stakeholders have to accept that the process is well founded), (2) saliency (it must be relevant to an expressed need) and (3) credibility (it must be conducted by experts, to the highest standards).

Assessments are characterised by an extensive, transparent review process by both experts and stakeholders. An assessment requires the authors to provide their own expert judgements when the data are sparse or equivocal (as long as these judgements are clearly identified as opinions), but puts checks and balances in place to ensure that all reasonable viewpoints are fairly reflected. Assessments include an explicit evaluation of the uncertainties on key issues, either quantitatively in terms of probability ranges (e.g. 'near certain' is $>95\%$ confidence of being true), or qualitatively (such as 'established', 'established but incomplete', 'competing explanations' or 'speculative').

For an assessment of the holistic management of pines in the CFR to take place, it would have to be endorsed by government, to provide the necessary legitimacy. While it is still too early to predict whether this will happen, this meeting has started a process which will hopefully lead to the development of policies based on sound scientific understanding. The apparent willingness to collaborate by all who attended this meeting gives hope that this development will come about.

References

1. Richardson DM, Hui C, Nuñez MA, Pauchard A. Tree invasions: Patterns, processes, challenges and opportunities. Biol Invasions. 2014;16(3):473–481. http://dx.doi.org/10.1007/s10530-013-0606-9
2. Richardson DM. Forestry trees as invasive aliens. Conserv Biol. 1998;12:18–26. http://dx.doi.org/10.1046/j.1523-1739.1998.96392.x
3. 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
4. Van Wilgen BW, Richardson DM. Challenges and trade-offs in the management of invasive alien trees. Biol Invasions. 2014;16:721–734. http://dx.doi.org/10.1007/s10530-013-0615-8
5. McConnachie MM, Van Wilgen BW, Richardson DM, Ferraro PJ, Forsyth T. Estimating the effect of plantations on pine invasions in protected areas: A case study from South Africa. J Appl Ecol. 2015;52(1):110–118. http://dx.doi.org/10.1111/1365-2664.12366
6. Van Wilgen BW, Richardson DM. Three centuries of managing introduced conifers in South Africa: Benefits, impacts, changing perceptions and conflict resolution. J Environ Manage. 2012;106:56–68. http://dx.doi.org/10.1016/j.jenvman.2012.03.052
7. VECON Consortium. Cape conversion process. Review of the original recommendations and decisions taken about phasing out plantation forestry and state forest land in the southern and western Cape and recommendations on a decision to reverse the withdrawal strategy. Pretoria: Department of Water Affairs and Forestry; 2006 [unpublished report].
8. Colvin C, Nobula S, Imelda Haines I, Nel JL, Le Maitre DC, Smith J. An introduction to South Africa's water source areas. Cape Town: WWF (South Africa); 2013.
9. Van Wilgen BW, Reyers B, Le Maitre DC, Richardson DM, Schonegevel L. A biome-scale assessment of the impact of invasive alien plants on ecosystem services in South Africa. J Environ Manage. 2008;89:336–349. http://dx.doi.org/10.1016/j.jenvman.2007.06.015
10. Hoffmann JH, Moran VC, Van Wilgen BW. Prospects for biological control of invasive *Pinus* species (Pinaceae) in South Africa. Afr Entomol. 2011;19:393–401. http://dx.doi.org/10.4001/003.019.0209
11. Forestry Stewardship Council. FSC principles and criteria for forest stewardship. Bonn: Forestry Stewardship Council; 2014.
12. Scholes RJ, Mennel KG. Elephant management: A scientific assessment for South Africa. Johannesburg: Wits University Press; 2008.

Shakespeare, plants, and chemical analysis of early 17th century clay 'tobacco' pipes from Europe

AUTHOR:
Francis Thackeray[1]

AFFILIATION:
[1]Evolutionary Studies Institute, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Francis Thackeray

EMAIL:
Francis.Thackeray@wits.ac.za

POSTAL ADDRESS:
Evolutionary Studies Institute,
Private Bag 3, Wits 2050,
South Africa

KEYWORDS:
Walter Raleigh; Francis Drake; botany; cannabis; cocaine; tobacco

HOW TO CITE:
Thackeray F. Shakespeare, plants, and chemical analysis of early 17th century clay 'tobacco' pipes from Europe. S Afr J Sci. 2015;111(7/8), Art. #a0115, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0115

In a recent issue of *Country Life*, Mark Griffiths[1] renews interest in John Gerard's *Herbal*[2], published in 1597 as a botanical book which includes engraved images of several people in the frontispiece. One of them (cited as 'The Fourth Man') is identified by Griffiths as William Shakespeare, but this identification is by no means certain. The question arises as to whether the engraving represents Sir Francis Drake.[3] Gerard's *Herbal* refers inter alia to various kinds of 'tobacco' introduced to Europe by Drake and Sir Walter Raleigh in the days of Shakespeare in Elizabethan England. One can well imagine the scenario in which Shakespeare performed his plays in the court of Queen Elizabeth, in the company of Drake, Raleigh and others who smoked clay pipes filled with 'tobacco'. However, there were several kinds of 'tobacco' in those days, as indicated in this article.

There clearly is a strong link between Drake and plants from the New World, including corn, the potato and 'tobacco'. Furthermore, one can certainly associate Sir Walter Raleigh with the introduction of 'tobacco' to Europe from North America (notably in the context of the tobacco plant called *Nicotiana,* from Virginia and elsewhere, and from which we get nicotine).

Thackeray et al.[4] reported in the *South African Journal of Science* the results of chemical analyses of plant residues in 'tobacco pipes' from Stratford-upon-Avon and environs, dating to the early 17th century. This non-destructive chemical analysis was undertaken using state-of-the-art forensic technology at the South African Police narcotics laboratory, by three scientists (Professor Francis Thackeray, Professor Nicholas van der Merwe of the University of Cape Town, and Inspector Tommy van der Merwe). A sophisticated technique called gas chromatography mass spectrometry (GCMS) was used. The pipe bowls and stems had been obtained by Thackeray on loan from the Shakespeare Birthplace Trust in Stratford-upon-Avon. Several of the pipes had been excavated from the garden of William Shakespeare.

Results of this study (including 24 pipe fragments) indicated *Cannabis* in eight samples, nicotine (from tobacco leaves of the kind associated with Raleigh) in at least one sample, and (in two samples) definite evidence for Peruvian cocaine from coca leaves of the kind which Thackeray et al.[4] associated with Drake who had himself been to Peru before 1597.

Gerard[2] has a whole section dedicated to kinds of tobacco including 'the henbane of Peru' which can be associated with cocaine (*Erythroxylum*), recognising that Sir Francis Drake could have brought coca leaves to England after his visit to Peru in South America, just as Sir Walter Raleigh had brought 'tobacco leaves' (*Nicotiana*) from Virginia in North America.

In 2000, Thackeray consulted the first edition of Gerard's *Herbal* in Stratford-upon-Avon to check the description of various kinds of 'tobacco'. As a botanist, Gerard must have known of the coca leaf as a kind of 'tobacco' from Peru. As chemists, Thackeray et al.[4] found unquestionable evidence for the smoking of coca leaves in early 17th century England, based on chemical evidence from two pipes in the Stratford-upon-Avon area. Neither of the pipes came from the garden of Shakespeare. Four of the pipes with *Cannabis* came from Shakespeare's garden.

Shakespeare may have been aware of the deleterious effects of cocaine as a strange compound. Thackeray (unpublished manuscript) suggests that Shakespeare preferred *Cannabis* as a stimulant which had mind-stimulating properties. These suggestions are based on the following literary indications. In Sonnet 76 Shakespeare writes about 'invention in a noted weed'. This can be interpreted to mean that Shakespeare was willing to use 'weed' (*Cannabis* as a kind of tobacco) for creative writing ('invention'). In the same sonnet it appears that he would prefer not to be associated with 'compounds strange', which can be interpreted, at least potentially, to mean 'strange drugs' (possibly cocaine). Sonnet 76 may relate to complex wordplay relating in part to drugs (compounds and 'weed'), and in part to a style of writing, associated with clothing ('weeds') and literary compounds (words combined to form one, as in the case of the word 'Philsides' from Philip Sidney). The so-called 'Fourth Man' depicted on the frontispiece of Gerard's *Herbal* holds a fritillary plant in one hand and corn in the other, as identified by Griffiths[1] who claims that the man is Shakespeare. However, attention can be given to an alternative hypothesis that this individual represents Sir Francis Drake.[3] We support the view that the 'Fourth Man' in Gerard's *Herbal* may represent Drake (especially as he holds an ear of corn in one hand – why should Shakespeare be holding such a plant of a kind known to have been introduced to Europe by Drake?).

An appeal is made to the Shakespearean community to give attention to articles that were published more than a decade ago[5-12] and which were largely criticised by Shakespearean scholars at that time. Chemical analyses of residues in early 17th century clay 'tobacco pipes' have confirmed that a diversity of plants were smoked in Europe. Literary analyses and chemical science can be mutually beneficial, bringing the arts and the sciences together in an effort to better understand Shakespeare and his contemporaries.

Acknowledgements

I thank the Shakespeare Birthplace Trust for the opportunity to analyse residues from pipes from the garden of Shakespeare and elsewhere in the environs of Stratford-upon-Avon.

References

1. Griffiths M. Shakespeare: Cracking the code. Country Life. 2015; Special Historic Edition May 20;120–138.

2. Gerard J. The herbal or general history of plants. London: John Norton; 1597.

3. Ward M, Lee GF. The language of flowers speaks clearly, not in riddles [homepage on the Internet]. c2015 [cited 2015 Jun 19]. Available from: http://www.historyneedsyou.com/blog/the-language-of-flowers-speaks-clearly-not-in-riddles

4. Thackeray JF, Van der Merwe NJ, Van der Merwe TA. Chemical analysis of residues from seventeenth century clay pipes from Stratford-upon-Avon and environs. S Afr J Sci. 2001;97:19–21.

5. Thackeray JF. The tenth muse: Hemp as a source of inspiration for Shakespearean literature? Occasional Paper of the Shakespeare Society of Southern Africa. 1999:1–9.

6. Thackeray JF. Shakespeare, hallucinogens and a tenth muse [homepage on the Internet]. c2001 [cited 2015 Jun 19]. http://www.teachernet.co.za/shakespeare1.html

7. Thackeray JF. Shakespeare, smoking and a substance ‘more chargeable than cane-tobacco’ [homepage on the Internet]. c2001 [cited 2015 Jun 19]. Available from: http://www.teachernet.co.za/shakespeare6.html

8. Thackeray JF. Cannabis, appetite and Shakespearean texts [homepage on the Internet]. c2001 [cited 2015 Jun 19]. Available from: http://www.teachernet.co.za/shakespeare2.html

9. Thackeray JF. Shakespeare’s sonnets and sources of inspiration [homepage on the Internet]. c2001 [cited 2015 Jun 19]. Available from: http://www.teachernet.co.za/shakespeare3.html

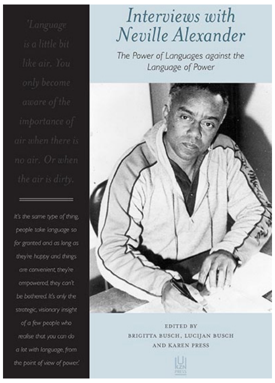
10. Thackeray JF. Portraits of Shakespeare [homepage on the Internet]. c2001 [cited 2015 Jun 19]. Available from: http://www.teachernet.co.za/shakespeare4.html

11. Thackeray JF. The dedication to Shakespeare’s sonnets [homepage on the Internet]. c2001 [cited 2015 Jun 19]. Available from: http://www.teachernet.co.za/shakespeare5.html

12. Thackeray JF. Trance, art and literature: Testing for hallucinogens. Antiquity. 2005;79:303. Available from: http://antiquity.ac.uk/projgall/thackeray/

BOOK TITLE:
Interviews with Neville Alexander: The power of languages against the language of power

BOOK COVER:


'Language is a little bit like air. You only become aware of the importance of air when there is no air. Or when the air is dirty.
It's the same type of thing, people take language so for granted and as long as they're happy and things are convenient, they're empowered, they can't be bothered. It's only the strategic, visionary insight of a few people who realise that you can do a lot with language, from the point of view of power.'
Interviews with Neville Alexander
The Power of Languages against the Language of Power
EDITED BY
BRIGITTA BUSCH, LUCIJAN BUSCH AND KAREN PRESS
UKZN PRESS

EDITORS:
Brigitta Busch, Lucijan Busch, Karen Press

ISBN:
9781869142773

PUBLISHER:
University of KwaZulu-Natal Press, Pietermaritzburg, ZAR350

PUBLISHED:
2014

REVIEW TITLE:
Neville Alexander: History, politics and the language question

REVIEWER:
Rajend Mesthrie

EMAIL:
rajend.mesthrie@uct.ac.za

AFFILIATION:
School of African and Gender Studies, Anthropology and Linguistics, University of Cape Town, Cape Town, South Africa

POSTAL ADDRESS:
School of African and Gender Studies, Anthropology and Linguistics, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa

HOW TO CITE:
Mesthrie R. Neville Alexander: History, politics and the language question. S Afr J Sci. 2015;111(7/8), Art. #a0113, 2 pages. http://dx.doi.org/10.17159/sajs.2015/a0113

Neville Alexander: History, politics and the language question

When Neville Alexander died in 2012, aged 75, after a short battle against cancer, South Africa lost its leading linguistic activist. It also lost an independent political thinker, one who had been incarcerated on Robben Island for 10 years, between 1964 and 1974, interacting there with Nelson Mandela, Walter Sisulu, Ahmed Kathrada, Eddie Daniels and others. The book under review is a fitting tribute to a great political figure and scholar. It originated in a series of interviews undertaken between 2006 and 2010, mostly by applied linguist Brigitta Busch of the University of Vienna, herself an academic-cum-language activist interested in multilingualism and the linguistic order in places like South Africa. The editorial credits include Lucijan Busch who undertook some additional interviews. The book was first published in a German edition in 2011, on the occasion of Alexander's 75th birthday. While following the structure of the German original, the South African edition is not identical. We are told in the introduction that Karen Press, the third editor, rearranged the text to suit a South African readership, highlighting Alexander's efforts in formulating a language policy for post-apartheid South Africa. The South African edition contains two parts: the first entitled 'Neville Alexander's language biography' and the second 'A selection of Neville Alexander's writings on language'.

Part 1 attests to a rich family and linguistic history. It covers Alexander's family background and language biography as a young child in Cradock in the Eastern Cape. Alexander's grandmother on his mother's side had been a slave girl from Ethiopia, freed by the British en route to a life of slavery in Saudi Arabia, and released in Port Elizabeth under the care of the London Missionary Society. Alexander recalls hearing snippets of Oromo from this grandmother. His mother spoke Afrikaans fluently but tended to favour English, while his father spoke Afrikaans, despite having an English-speaking father himself (of Scottish origin). The first essay documents further multilingual experiences with Xhosa and some Khoisan, the latter restricted to tales and isolated words from an aunt. The young boy's love of English and Afrikaans was eclipsed by his learning of German as a first foreign language at the Holy Rosary Convent, a Catholic mission school. Here began his lifelong engagement with German: at this stage '…it was the more mysterious aspect, the things you did not understand, which were interesting and which...resonated' (p.25). Alexander also pays homage to Latin, the language of mass and ritual. Later in Germany this was to ease his way into the formal study of the language and its classical authors, of whom Alexander mentions Ovid and Pliny. At the University of Cape Town (UCT) Alexander studied German, Afrikaans, English and History, and became involved in politics, joining the Teacher's League, which was affiliated to the Non-European Unity Movement. The second essay of Part 2 is an engaging account of the young Alexander's fascination with both language and literature components of the three languages he studied at UCT. It was this engagement with the local languages that led him to consider bilingual education as an option, where other activists of the time assumed the backwardness of local languages.

The most informative chapter from the point of biography is the fourth, 'University behind bars: Robben Island 1964–1974', which gives a moving account of events leading to Alexander's arrest and trial on charges of sabotage and incarceration. The story of the island days is – in common with other biographies – a very human one detailing little events that made life bearable and the future possible. The interview focuses on incarceration and labour of course; but has a great deal more on political education, studying through correspondence, visitors from outside, and relations between prisoners and warders. For his Honours essays researched and written in prison, Alexander gains close to maximum marks on the philosophy of history, essays which are – we are told – still used as a model by Unisa's history department. Political education was the bigger goal, leading to the formation of a Society for the Rewriting of South African History, comprising, inter alia, Alexander, Mandela, Sisulu and Kathrada. (Alexander's role is confirmed by Ahmed Kathrada – Robben Island prisoner for 26 years – who in a documentary film of his own life, credits Alexander with convincing Nelson Mandela and himself of the need to see history as a process, not a recent or distant past to which one could dip into and draw on selectively.) Inevitably for Alexander, memories about language are highlighted: Afrikaans as a language of command by uneducated warders, who gradually came to respect their wards; learning Xhosa, Zulu and Sotho work songs; studying Xhosa grammar; teaching English, German and other subjects to fellow prisoners studying for a high school certificate or through Unisa; and secretly translating the *Communist Manifesto* into German for his fellow prisoners. Their obvious connection with a wider world of learning was not lost on the warders, who soon asked for help with their own correspondence studies. In helping one Afrikaans warder with Afrikaans as a subject, Alexander comes to the conclusion that the young warders to a large extent were as much prisoners as they were.

The rest of Part 1 discusses Alexander's release, house arrest and his involvement in SACHED (South African Committee on Higher Education) and the NLP (National Language Project), both of which focused on alternative education for a democratic future. These led to his chairing the LANGTAG (Language Task Action Group) which was a wide-ranging committee of academics, community representatives and language activists, convened in 1994 to inform and advise the new government about language policy matters. As Alexander confirms, the final 11 languages policy was not of his or the task group's making (p.154). Alexander also gives the background to his subsequent involvement as Director of PANSALB (the Pan South African Language Board) and his resignation over the direction the board was taking as well as issues pertaining to autonomy from ministerial control. His role as leader of PRAESA (Project for Alternate Education in South Africa) at UCT is detailed with emphasis on their research and continued language activism.

The second part of the book comprises six essays on the language question, as Alexander liked to call it, prefaced by a three-page select biography of his extensive writings. The essays showcase Alexander's previous writings and academic articles previously published as PRAESA working papers. 'The national question in South Africa' is an excerpt from a book he published in 1979 under the pseudonym *No Sizwe* ('Mother of the Nation'), *One Azania, One Nation: the National Question in South Africa*. It is well chosen as an initial essay that highlights Alexander's political revolutionary thinking. It is followed by another well-chosen excerpt from Alexander's little book of 1989, *Language Policy and National Policy in South Africa/Azania*. This excerpt forms a link between the historico-political and the sociolinguistic strands of Alexander's thinking. It is the sociological-cum-educational strands that are highlighted in the last four essays: 'Majority and minority languages in South Africa', 'The African Renaissance and the use of African languages in tertiary education', 'Street and standard: Managing language in contemporary South Africa' and 'The potential role of translation as a social practice for the intellectualisation of African languages'. The essays are a good representation of Alexander's writings on language in the last two decades for those wishing to gain an acquaintance of these writings. They are complemented by a very good index.

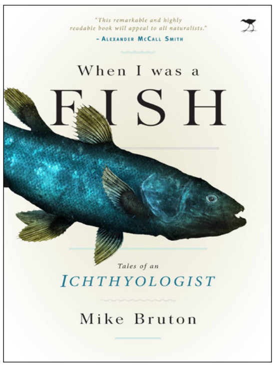
The insights into Alexander's life and thinking are many, even for those who knew Alexander as a colleague. For – while he was a warm and engaging person – he was not one to dwell on himself. He declined to become part of the new post-apartheid political elite, and refused to engage in mainstream ANC politics. Inevitably, there are some aspects of his activism that are downplayed in this memoir: the failure at the 1994 polls of the political party WOSA (Workers of South Africa) that Alexander launched in 1990; the lack of support for his harmonisation proposals (to restandardise and unify the Nguni languages and the Sotho–Tswana complex, respectively); and Alexander's position that race as a category should be ignored in admission to higher education. Also, his calls (following Amilcar Cabral, p.265) for the middle classes to commit class suicide to enable the birth of a new society are iterated; there is a need to balance this in relation to one of Alexander's last essays (*Who can say where the dog lies buried?*), which is more moderate in its assessment. The essay, published in the *Cape Argus*, is a scathing attack on the ethics of present-day politicians and administrators, but holds up the deracialising practices of the new, young, educated middle class as one of the few success stories in an otherwise bleak and imploding society.

The editors are to be congratulated on compiling a fitting volume to a great and sometimes unsung scholar-cum-activist. The book will be of great interest to students of language as well as to a more general readership concerned with South African political and intellectual history. In the spirit of Neville Alexander's commitment to multilingualism, it would be good to see the book appear in translation in more South African languages.

Memoirs of an ichthyologist

BOOK TITLE:
When I was a fish: Tales of an ichthyologist

BOOK COVER:



AUTHOR:
Mike Bruton

ISBN:
9781431420575 (softcover)

PUBLISHER:
Jacana Media, Johannesburg, ZAR246

PUBLISHED:
2015

REVIEW TITLE:
Memoirs of an ichthyologist

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:
Centre for Invasion Biology,
Department of Botany and
Zoology, Stellenbosch University,
Private Bag X1, Matieland 7602,
South Africa

HOW TO CITE:
Van Wilgen BW. Memoirs of an ichthyologist. S Afr J Sci. 2015;111(7/8), Art. #a0116, 1 page. http://dx.doi.org/10.17159/sajs.2015/a0116

With nearly 10% of all known species of birds, fish and plants, and 6% of mammal and reptile species, South Africa is one of the most biodiverse countries in the world. Any biologist, ecologist or conservationist who sets out to work in this country in one of these fields is fortunate indeed, as this diversity offers a multitude of opportunities for a rewarding, eventful, influential and memorable career. Mike Bruton is one such person. He started a career as a freshwater ichthyologist studying the fishes in the lakes of northern KwaZulu-Natal, and moved on to being an academic, research director and science educator, among others. His career has been based in southern Africa, but has provided many opportunities for travel to exotic destinations. Rather than write an autobiography, Bruton has penned a 'memoir', which in his own words is 'simply a story [or rather, a collection of stories] of how a boy grew up to become a man who had a passion for fishes'.

As a student of zoology (and later an academic) specialising in ichthyology at Rhodes University, Bruton had a close relationship with (and for a while served as director of) the J.L.B. Smith Institute of Ichthyology. The institute was founded by the legendary James Leonard Brierley Smith, author of the well-known book *The Sea Fishes of Southern Africa* (1949), but arguably better known as the person who described the first known modern coelacanth. The discovery of this ancient fish, previously known only from fossils that dated back 80 to 350 million years, was one of the most remarkable scientific events of the 20th century. Smith's description of the first specimen was published in 1939 in the journal *Nature*, and his discovery (after a long search) of a second specimen in 1952, caught the world's attention in a big way. This story is recounted in Smith's book *Old Fourlegs*, in my view one of the most exciting works of non-fiction ever written. Bruton is of the opinion that, had these events not occurred, J.L.B. Smith would in all likelihood have remained an amateur ichthyologist, and the centre of southern African ichthyology would probably have developed in Cape Town or Durban, and not at Rhodes University in Grahamstown.

As a result of this association, a good deal of Bruton's book is focused on subsequent work on the ecology and conservation of coelocanths, regarded as an icon for marine conservation in the same way that the giant panda symbolises terrestrial conservation. Bruton is one of the few people lucky enough to have observed living coelocanths, an experience he describes in the book, and his passion for the topic is quite apparent. He describes in some detail the documentation of all known specimens, the study of their distribution, ecology and breeding strategies, and the ethical question of whether or not coelocanths should be caught for public display in aquaria.

Bruton's book, however, is not just about coelocanths. It contains a myriad of snippets from a long and eventful career, many personal, quite a few informative, and almost all very entertaining. These stories range from those about sharing his working environment with crocodiles and hippos at Lake Sibiya, to meeting a range of dignitaries including Queen Elizabeth II. Two chapters are devoted to the study and documentation of life-history strategies in fishes. Biologists and non-biologists alike should benefit from the insights into the life and times of fishes that are described in this book, as well as the way in which they are managed (or mismanaged). In some places, I would have liked to have seen more detail, but there is only so much one can fit into a memoir like this. For example, it is mentioned that the stocking of the newly formed Lake Kariba with alien fish to create a new fishery was 'epic and highly controversial', and that 'to our inexperienced and biased eyes the Kariba experiment appeared to be a huge success', but very little further detail is given.

Bruton also pays tribute to the many colleagues with whom he has worked over the years. It is the lot of science administrators that many of their research colleagues will end up better known than their managers. South Africa is a 'fish-crazy nation', and those who publish the detailed guides to fishes are more likely to become known to anglers, naturalists, ecotourists, aquarists and food gourmets than those who nurture the environment that makes these publications possible. Besides J.L.B. and Margaret M. Smith (J.L.B.'s wife, lifelong scientific collaborator and successor), specific mention is made of (for example) Paul Skelton, Phil Heemstra and Ofer Gon, who between them have produced the next generation of definitive books documenting regional fish faunas.[1-3]

Bruton provides an interesting perspective on conservation in his final chapter. This perspective was again obviously influenced by his long study of coelocanths, a species that has survived in a more-or-less unchanged form for hundreds of millions of years. During this time, many other life forms have come and gone, and new forms of life have evolved. Bruton argues that extinction is neither good nor bad, and that species extinctions simply make space for new species to evolve. Humanity's current concern over the conservation of endangered species may therefore be misplaced. We are simply modifying the planet in a way that will make the evolution of new life forms possible, but it will also render the world uninhabitable for ourselves. Given that life will continue with newly evolved forms long after humans are gone, our concern should rather be that humans are an endangered species. With that in mind, Bruton has devoted the closing phases of his career to science education.

This book is a mixture of adventure stories, entertaining anecdotes, scientific facts and interesting interpretations. The book should be of wide interest to many people from a range of backgrounds, and I would recommend it to anyone with an interest in biology, conservation, education or a career in the sciences.

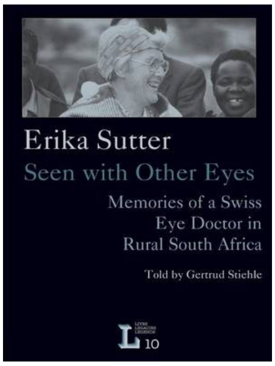
References

1. Smith MM, Heemstra PC. Smith's sea fishes. Johannesburg: Macmillan; 1986. http://dx.doi.org/10.1007/978-3-642-82858-4
2. Skelton PH. A complete guide to the freshwater fishes of southern Africa. Johannesburg: Southern Book Publishers; 1993.
3. Gon O, Heemstra PC. Fishes of the Southern Ocean. Grahamstown: J.L.B. Smith Institute of Ichthyology; 1990.

Lessons for eye health practitioners

BOOK TITLE:
Seen with other eyes: Memories of a Swiss eye doctor in rural South Africa

BOOK COVER:



AUTHOR:
Gertrud Stiehle

ISBN:
9783905758337 (softcover)

PUBLISHER:
Basler Afrika Bibliographien, Basel, ZAR460

PUBLISHED:
2014

REVIEW TITLE:
Lessons for eye health practitioners

REVIEWER:
Kovin Naidoo[1,2]

EMAIL:
K.Naidoo@brienholdenvision.org

AFFILIATIONS:
[1]Brien Holden Vision Institute, Durban, South Africa
[2]African Vision Research Institute, University of KwaZulu-Natal, Durban, South Africa

POSTAL ADDRESS:
172 Umbilo Road, Durban 4000, South Africa

HOW TO CITE:
Naidoo K. Lessons for eye health practitioners. S Afr J Sci. 2015;111(7/8), Art. #a0112, 1 page. http://dx.doi.org/10.17159/sajs.2015/a0112

The relative success of eye care programmes with quick outcomes, from cataract surgery to providing eye glasses, has often mitigated against comprehensive eye care programmes. Success comes ‘easily’, so why focus on complex solutions? However, the need to upscale eye care programmes beyond the individual programme success and the need to address sustainability and community ownership of such programmes has forced, and is still forcing, many to rethink strategies. Given the lack of eye care practitioners with public health training and the predominance of curative approaches to eye care, there is a shortage of skills and exposure to adequately think or act beyond the curative paradigm.

In the context of the above, it is indeed opportune to have a book that adds to this process in a very unique way – not through the usual theory presentation but through the real-life experiences of an eye doctor. This is particularly useful in a context in which clinicians often see public health and community engagement as a responsibility of a small group rather than an intrinsic aspect of the practice of ophthalmology. This identification with the book is aided by several aspects:

- The life of Dr Erika Sutter is presented in a frank and undramatic way. She is presented as a real person with whom many can relate. Her own, very realistic transformation from a lab scientist to a clinician and eventually a community health and development activist, makes one believe that they too can achieve the same. She also comes across as a person who has the ability to balance her faith, interest in music, her professional career, her social consciousness, her family relationships and friendships, while being an asset to society. This achievement is often one that eludes many social activists and makes Dr Sutter fairly unique.
- Dr Sutter’s sensitivity to the rights of the local community and the need to recognise and respect them as well as acknowledge them, is a valuable lesson in how locals and outsiders should relate to each other. This addresses one of the major issues in development, i.e. do we work for people, or with them, and how do we appropriately empower communities and develop the local leadership needed?
- Key development concepts are presented through the practical experiences of the book’s key subject. The achievement of development objectives are presented in a realistic way, giving the reader a sense of both the negative and positive aspects and not glorifying outcomes. It is particularly useful how the community’s perceptions of efforts, failures and successes, and how these were addressed, is presented.
- Throughout the narrative, some useful tips about leadership are presented through the experiences of Dr Sutter. These include the reluctant leader, the cautious leader, the engaging leader, the respectful leader and the leader who acknowledges others and motivates them.

The book also provides an enlightening look at life in South Africa and insight on a less written aspect of life under apartheid – the effects of discrimination on the health and access to health care of the majority. It highlights the resilience of communities under such conditions and the capacity of communities to take control of and change their own lives despite the odds being stacked against them. For health professionals, the issues of choices in unjust conditions is magnified. The incidents of how health professionals dealt with the segregation and unequal nature of health care and the dilemmas it created, are useful issues for us to explore and debate, even in the current context. It is unfortunate that limited discourse exists around this issue, but the book does play a valuable role in flagging this issue as one for health professionals to constantly consider.

It is very useful that the changes and contributions of Dr Sutter occurred within a mainstream missionary organisation and not through a progressive NGO, which was often the norm during this era. There are strong lessons in the life story of Dr Sutter that relate to how one organises and influences others in religious and mainstream organisations that have a more charitable than development focus. Dr Sutter’s ability to influence and expand the agenda of the eye hospital, and thus the missionary work, is a great lesson for clinicians who often find themselves in facilities that are less prone to adopting broader development approaches.

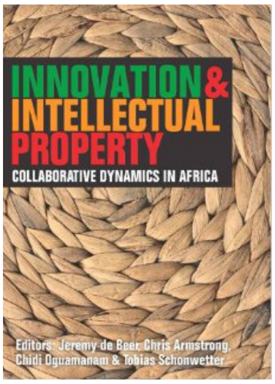
As a relative newcomer to the field of public health ophthalmology and optometry, I have been faced with the dominance of NGOs from the developed world which lead and secure much of the funding for the management of trachoma in Africa and the rest of the developing world. The irony is that the solution they promote is very much the African solution that Dr Sutter and colleagues pioneered in their work at Elim and promoted through their books and lectures in London and elsewhere – the concepts of community engagement and a promotive as well as curative approach to eye health. There is a strong lesson here for African health professionals. We need to harness the intellectual wealth to not only serve our people but also to command leadership at an international level by exporting knowledge and taking leadership in addressing social and health problems, not just for Africa but for other developing countries with similar problems.

The book could have been aided by more input or quotes from local role players; however, it is an excellent read for health professionals seeking to practice in a manner that recognises that curative care alone is insufficient in improving the health of communities. Most importantly, it is a lesson on how ordinary people can do amazing things with perseverance and commitment to serving others.

Intellectual property rights in Africa: The way ahead

BOOK TITLE:
Innovation & intellectual property: Collaborative dynamics in Africa

BOOK COVER:



INNOVATION &
INTELLECTUAL
PROPERTY
COLLABORATIVE DYNAMICS IN AFRICA
Editors: Jeremy de Beer, Chris Armstrong, Chidi Oguamanam & Tobias Schonwetter

EDITORS:
Jeremy de Beer, Chris Armstrong, Chidi Oguamanam, Tobias Schonwetter

ISBN:
9781919895994 (softcover)

PUBLISHER:
Open A-I-R, UCT Press, Cape Town, ZAR350 (the electronic version is licensed under a Creative Commons licence)

PUBLISHED:
2014

REVIEW TITLE:
Intellectual property rights in Africa: The way ahead

REVIEWER:
Anastassios Pouris

EMAIL:
Anastassios.Pouris@up.ac.za

AFFILIATION:
Institute for Technological Innovation, University of Pretoria, Pretoria, South Africa

POSTAL ADDRESS:
Institute for Technological Innovation, University of Pretoria,
Private Bag X20, Hatfield 0028,
South Africa

HOW TO CITE:
Pouris A. Intellectual property rights in Africa: The way ahead. S Afr J Sci. 2015;111(7/8), Art. #a0111, 1 page. http://dx.doi.org/10.17159/sajs.2015/a0111

This book was produced by the Open African Innovation Research and Training Project which is a research and training network aimed at raising intellectual property (IP) awareness in Africa and facilitates critical policy engagements including the identification of relevant bottlenecks and the interrogation of innovation metrics and structures. The Project is supported financially by Canada's IDRC and Germany's BMZ.

The editors of the book are associated with the Universities of Ottawa, the Witwatersrand and Cape Town. The contributing authors responded to an open public call to investigate issues related to the question: How can existing or potential IP systems be harnessed to appropriately value and facilitate innovation and creativity for open development in Africa? The themes covered by the book include patents, copyrights, publicly funded research, trademarks, geographic indications, traditional knowledge and informal appropriations.

The book – 408 pages – covers mainly case studies in nine African countries: Botswana, Egypt, Ethiopia, Ghana, Kenya, Nigeria, Mozambique, South Africa and Uganda. Most of the case studies are localised in nature. Chapter 10 (African Patent Offices not fit for purpose) is the only paper covering the continent extensively. It reports on the capacities of patent offices in 44 countries in Africa. Professor Mgbeoji found that most of the national patent offices on the continent were ill equipped to discharge their two main functions: examining patent applications and collating patent information so that it can be made publicly available for public and inventor follow-on use.

The authors state that 'the book makes no claim to be comprehensive' (p.15). The final chapter is a summary of the findings of collaborative innovation and creativity, openness and intellectual property. There are 'vibrant collaborative models at play in relation to innovation and livelihood development' (p.388), but they range from the extremely informal to considerably more formal. Similarly, the case studies show that open development cannot be conceived as a binary proposition – either open or closed.

As far as IP is concerned, the editors suggest that:

> *Policymakers and actors will need to move away from dominant preconceptions of IP involving mainly patent, copyright and trademark protections. Informal and flexible protections such as trade secrets seem much better suited to the informal sector.* (p.389)

Finally, the editors identify that although IP laws are in place, their impact is minimal because of shortcomings in the administrative structure needed to implement and enforce these laws.

The book develops three recommendations for African policymakers. The first recommendation is to avoid mistakes. The editors suggest that not having an IP policy may be better than entrenching the wrong IP policy. The second recommendation is to broaden conceptions of relevant and valuable IP practices. The editors suggest that 'patent systems are irrelevant to many of the modes of innovation and creativity happening in Africa' (p.392). They suggest that informal modes of IP protection, such as trade secrecy, may be better suited to Africa's environment. Similarly, they suggest that branding may be a useful form of IP in many instances. Finally, the suggestion is developed that African policymakers need to look forwards not backwards. It is stated that 'going forward, African policymakers, as with the innovators and creators whom the policymakers are supposed to serve, must seek to harness IP rights on their own terms'.

It is important to notice that the book aimed to answer certain questions by interrogating mainly African experts and practitioners. However, a broader view could provide additional useful information for advice. For example, IP laws are imposed on Africa and other developing countries by outside forces. Starting in the 1980s, the USA imposed strong pressure on developing countries with weak IP laws and institutions through its 'Special 301' provision of the US trade law. Countries threatened with sanctions are forced to strengthen their patent laws. Countries like China, Indonesia, Taiwan and Malaysia were forced in the 1980s and 1990s to introduce patent laws related to pharmaceuticals.

It should also be emphasised that the non-examining approaches followed in Africa are facilitating their exploitation by foreigners who are able to protect their ideas even when they are not patentable. Pouris[1], investigating the South African patent system, argued that 'granting patents for inventions that are not new or useful or that are obvious, unjustly rewards the patent holder at the expense of real inventors, the consumer and social welfare'. African countries need to get out of the current method of non-examining approaches.

Finally, the suggestion that policymakers should look forward may be misunderstood. History usually provides valuable lessons for the future. For example, Chang[2] suggests that the developed countries refused to institutionalise patent regimes when they were in the development stage.

The book makes a valuable contribution in a neglected field in Africa and undoubtedly will attract considerable debate.

References

1. Pouris A, Pouris A. Patents and economic development in South Africa: Managing intellectual property rights. S Afr J Sci. 2011;107(11/12):24–33. http://dx.doi.org/10.4102/sajs.v107i11/12.355
2. Chang HJ. Bad samaritans: The myth of free trade and the secret history of capitalism. New York: Bloomsbury Press; 2008.

A review of quantitative studies of South African youth resilience: Some gaps

AUTHORS:
Angelique van Rensburg[1]
Linda Theron[1]
Sebastiaan Rothmann[1]

AFFILIATION:
[1]Optentia Research Focus Area, North-West University, Vaal Triangle Campus, Vanderbijlpark, South Africa

CORRESPONDENCE TO:
Angelique van Rensburg

EMAIL:
angelique.vanrensburg@nwu.ac.za

POSTAL ADDRESS:
Optentia Research Focus Area, North-West University, Vaal Triangle Campus, PO Box 1174, Vanderbijlpark 1900, South Africa

DATES:
Received: 15 May 2014
Revised: 31 Oct. 2014
Accepted: 16 Jan. 2015

KEYWORDS:
quantitative studies; resilience; systematic review; social ecology; South Africa; youth

HOW TO CITE:
Van Rensburg A, Theron L, Rothmann S. A review of quantitative studies of South African youth resilience: Some gaps. S Afr J Sci. 2015;111(7/8), Art. #2014-0164, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140164

Resilience (positive adjustment to hardship) relies on a socioecologically facilitated process in which individuals navigate towards, and negotiate for, health-promoting resources, and their social ecology, in return, provides support in culturally aligned ways (Ungar, Trauma Violence & Abuse 2013;14(3):255–266). In the light of international critiques of the conceptualisation and measurement of resilience, the aim of this study was to systematically review quantitative studies of South African youth resilience in order to consider to what extent such studies failed to address documented critique (Luthar et al., Child Development 2000;71(3):543–562). We argue that, for the most part, quantitative studies of South African youth resilience did not mirror international developments of understanding resilience as a complex socioecologically facilitated process. Furthermore, the majority of reviewed studies lacked a culturally or contextually sound measurement and contained conflicting operationalisations of resilience-related constructs. Essentially, the results of this study call for quantitative studies that will statistically explain the complex dynamic resilience-supporting transactions between South African youth and their contexts and guide mental health practitioners and service providers towards more precise explanations and promotion of resilience in South African youth.

Introduction

Resilience, or positive adjustment to hardship, relies on a complex transactional process between individuals and their social ecology in which the individual navigates towards, and negotiates for, health-promoting resources, and the social ecology reciprocates by providing support in culturally aligned ways.[1] It is important to note that a precondition of resilience is a lived experience of risk – in other words, an experience of adversity as personally threatening. Risks typically heighten the chances of negative developmental outcomes.[2] Risks include challenging social ecologies (e.g. a violent community or ineffective school), specific negative life events (e.g. the death of a parent), compound sociodemographic risks (e.g. growing up in a single-parent, impoverished family in a violent community), as well as the impact of biological vulnerabilities (e.g. genetic predispositions or premature births).[1,3]

As explained by Masten and colleagues, the interpretation of resilience as a complex process evolved over decades.[2,4-8] In the early 1970s, researchers focused on the elementary principles involved in resilience: a considerable amount of research emphasised the definition and measurement of resilience. What emerged was a list of protective factors (i.e. attributes of an individual that could result in better outcomes under high levels of adversity) supporting resilience.[2] It was thought that these protective resources were embedded within the individual as personality traits, skills and genetic predispositions.[4,5] As a result, *person-focused models* of resilience emerged, in which the emphasis was on differences between resilient and non-resilient individuals researched in the form of single case studies. However, this model did not allow researchers to identify the processes that underpinned resilience.[8] This limitation led to a shift in researchers exploring the mechanisms of resilience and conceptualising how these might inform processes of positive adjustment to hardship. This 1980s shift manifested as *variable-focused models* that relied on analysis of the relationship between resilience and a person's characteristics and aspects relating to their ecologies (e.g. violence, divorce, supportive families, religion).[4,6] Researchers subsequently focused on testing and promoting these models of resilience processes through prevention, interventions and policymaking.[5,7] This then prompted questions about how adaptive and maladaptive pathways differed in individuals who experienced adversity, and generated *pathway models* (from the 1990s onwards). However, not enough was known about how resilience processes differed across contexts and cultures.[2,6] Studies of contextual and cultural influences on resilience led researchers to acknowledge the complexity and cultural relativity involved in processes of positive adaptation. Consequently, resilience is now seen as a *culturally aligned transaction* that is facilitated by actions that social ecologies and young people reciprocally take.[1,4]

As detailed below, how resilience (particularly quantitative studies of resilience) was studied in the course of this evolution has received much criticism. Studies of South African youth resilience followed a similar evolution.[9] The purpose of this article was to conduct a systematic review of quantitative studies of South African youth resilience in order to evaluate how well these studies have avoided the pitfalls made public in the international critiques of resilience studies. In 2010, Theron and Theron[9] published a review of published studies of South African youth resilience. Although their review did raise some criticisms of prior studies, it did not evaluate the quantitative studies of South African youth resilience against internationally voiced concerns. The current review was guided by the following questions: how does quantitative South African youth resilience research measure up to international critiques?, and which subsequent gaps necessitate future investigation?

International critiques of quantitative resilience research

In order to weigh quantitative South African youth resilience research against international commentaries, a systematic review of critical commentaries on international youth resilience studies was conducted. The inclusion criteria were (1) internationally indexed, scholarly, peer-reviewed articles and/or book chapters, (2) with titles, keywords, or abstracts that included one of the following terms: 'review', 'issues', 'critique', 'commentary', 'evaluation', 'frameworks', 'future directions', 'research development' and 'resilience'. Relevant commentaries were sourced through a database search (EBSCOhost, JSTOR, ScienceDirect), perusal of published reference lists, and/or recommended by authors' resilience-focused networks. Only one[10] commentary (included in a resilience-focused volume) did not define resilience explicitly. In this instance, the commentary aligned with how resilience was defined in the volume in which it was included. We excluded commentaries not reported in English, and/or those that were not resilience specific (e.g. coping focused), or not youth specific (e.g. adult/geriatric resilience). We further excluded commentaries on resilience-supporting interventions. Applying these criteria resulted in the inclusion of 26 documents.[1-4,8,10-30]

When conducting research, scholars make use of scientific approaches of investigation, which Creswell[31] refers to as a 'process of research' (p.7). The process consists of six steps: identifying a research problem, reviewing the literature, specifying a purpose for research, collecting data, analysing and interpreting the data, and reporting and evaluating research. Because the reviewed documents typically addressed problems with the researching of resilience, these steps were used to structure the synthesis of international critiques on resilience research. No critiques were levelled at three of Creswell's[31] steps – reviewing the literature, specifying a purpose for research, and reporting and evaluating research.

Identifying a research problem

Part of identifying a research problem is conceptualising the focus of the research.[31] A number of criticisms was aimed at how researchers conceptualised resilience and related constructs such as risk and protective factors.[11,27]

Conflicting conceptualisation of resilience

Masten[8] and Werner[25], among others, affirm that disagreement exists among researchers with regard to how to conceptualise resilience and that such disagreement confounds the study of resilience. Resilience was originally thought of as a person-centred construct (see the work of Anthony and Cohler[32]). Person-centred conceptualisations of resilience meant that what was needed to be resilient lay within the individual.[4] However, Lerner[19] and Luthar et al.[11], among others,[21,26] became critical of a person-focused conceptualisation of resilience and encouraged understandings of resilience as a person ↔ context transaction[19] (i.e. a dynamic transaction between the environment and the individual that supports access to, and use of, resilience-promoting resources). A danger in person-focused definitions is how it accentuates youths' responsibility to be resilient.[1] In spite of this danger, some recent studies of resilience have continued to interpret resilience as an individual-centred concept.[1,15] For example, researchers use terms such as 'psychological resilience' and 'resiliency'.[33] These terms imply personality characteristics or individual skills in explanations of resilience and downplay the importance of pro-active, supportive socioecological contributions to resilience.[34] Thus, when some researchers define resilience as person-focused and others as a construct supported by transactions between youths and their context, resilience is inconsistently conceptualised and the importance of socioecological contributions to resilience marginalised.[19]

Moreover, in their understanding of resilience, there is some disagreement among researchers about the exclusivity of positive adjustment. For example, Rutter[35] stated that one ought not to assume that everyone could be/become resilient. Masten[8], however, referred to resilience-promoting resources as 'common phenomena' and to resilience as 'ordinary magic' (p.227). Likewise, Windle[27] stated that the capacity for resilience is widespread and possible for anyone.

Additionally, some scholars have discouraged the conceptualisation of vulnerability as the opposite of resilience – vulnerability refers to susceptibility to adverse outcomes.[2,11] Early studies described resilient individuals as being invulnerable[2,11], implying that resilience is the opposite of being vulnerable[36,37]. This is problematic because vulnerability and resilience co-exist, and resilience does not imply an absence of vulnerability.[11]

Varying/absent conceptualisations of key terms

Critiques of prior resilience studies reported that key terms used to describe resilience-related phenomena were used conflictingly. For example, Luthar et al.[11] reported that researchers used resilience-related terms such as *protective factors* and *risk factors* inconsistently. *Protective factors* (e.g. good parenting, personal agency, supportive teachers or effective schools) are factors that heighten the chances of constructive developmental outcomes.[2] However, different connotations for the term *protective factors* are seen in resilience literature. For example, protective factors were used to explain main-effects models – referring to protective factors that have a single or direct effect on positive adaptation (e.g. good parenting might result in good coping skills).[11,38] In contrast, other studies conceptualised protective factors as interactive[11] or bidirectional[19]. From this perspective, multiple protective factors work in tandem to support functional outcomes, often as part of a give-and-take process (e.g. a learner's personal agency in securing support from her teacher when experiencing difficulties and her teacher's supportive response). Furthermore, what one community/context might consider as a protective factor/process might not be relevant to another.[1,15] For example, in Africentric contexts, youths are taught to value ancestral bonds as protective[39], whereas youths in Eurocentric contexts value different relational bonds[40]. Culturally sensitive conceptualisations of protective factors potentially protect highly mobile youth who must negotiate non-familiar contexts.

In the absence of risk, researchers would be observing coping rather than resilience.[1] For this reason, researchers are compelled to explain how study participants are at risk and to define such risks.[1] The heterogeneity of the source of risks (i.e. negative influences, experiences, specific life events, etc.) calls for researchers to clarify the types of risk that make youth participants vulnerable as a result of the varying processes involved in each source of risk.[23] Similarly, a distinction should be made between single occurrences of risk and compound/chronic risks, given that compound/chronic risks are known to heighten vulnerability.[6,23] Moreover, the impact of any given risk is not homogeneous across individuals and sociocultural contexts: even though individuals, families and communities share similar adverse experiences, one cannot assume that all individuals interpret these experiences as equally threatening.[1,2]

Collecting data

How data are collected is influenced by theoretical frameworks, research designs and instruments used.[31] Commentaries on the study of resilience included concerns about all of these.

Undeclared or outdated theoretical frameworks

Theoretical frameworks shape how resilience and related constructs (e.g. risks and protective factors) are defined, operationalised and subsequently measured.[11] Theoretical frameworks must be made explicit.[4] If the theoretical framework were not declared, it would make little sense why resilience would be measured in terms of individual, socioecological processes or otherwise. Additionally, if earlier theoretical frameworks (e.g. person-focused or variable-focused theories) were used, the data collected would contribute minimally to the evolved discourse of resilience.[11]

Over-reliance on cross-sectional research designs

A number of research reviews noted a preference for cross-sectional designs in resilience studies and emphasised that cross-sectional designs limited understandings of long-term pathways individuals took towards resilience.[13,14,16,20,22] The repeated choice of cross-sectional designs is problematic because these studies do not identify cause-and-effect relationships associated with resilience.[41] They also cannot establish the direction or magnitude of resilience processes, which is required to determine lifespan pathways individuals take towards resilience.[14,16] As a result, there have been calls for longitudinal research designs to be used in studies of youth resilience. Provided the theoretical framework is socioecological, and measures are chosen accordingly, long-term designs will allow researchers to observe individual and socioecological change. Long-term designs are integral to examining, explaining and predicting the causality, direction or magnitude of factors involved in resilience processes.[14,41]

Problematic measurement of resilience

Several resilience scales have been developed over the decades; however, there has been little consistency in how these instruments have been constructed and/or the cultural and contextual equivalence of these scales, resulting in possible construct, item and sampling biases.[1,10,28] Also, different conceptualisations of resilience have informed these multiple scales. Subsequent measurement of key concepts (e.g. resilience, risk and protective factors) is variable, potentially rendering data biased.[10] Moreover, despite current consensus that resilience is a transaction between an individual and his/her social ecology, most resilience measures do not mirror this view.[18,24,28] Gartland et al.[17] and Tol et al.[22] note that many resilience measures have a limited focus and scope because they address individual characteristics and not the dynamic socioecological transaction.

The proliferation of resilience scales might relate to resilience and risk being culturally and contextually specific constructs, which are not similarly defined universally. Ungar[1] is critical of notions of a universal measuring instrument. Measurements of resilience are flawed if the measure used is not contextually and culturally appropriate to the population to which it will be administrated.[28] Analysis of biased data could cause inaccurate assumptions about cultural or other differences in resilience, resulting in faulty theories.[10]

The measurement of resilience is also problematic when measurements are conducted on non-representative samples.[1] Of concern is that resilience theories currently reflect measurements that were predominantly taken from white, Western participants[1,11], which essentially translates into sampling bias[10]. Subsequent theories of resilience will be limited by the narrow sampling that informed them. For example, when researchers work with one narrowly defined group, such as substance-abusing youth attending private schools, they exclude substance-abusing youth who are not at school or who are in government schools. Another example pertains to youths who are routinely excluded from resilience studies: youth with disabilities, life-limiting conditions, and/or terminal illnesses are under-represented in resilience studies, resulting in a poor understanding of their specific resilience processes.[42]

Then, discrepancies exist in how constructs related to resilience are measured. For example, Luthar et al.[11] stated that risk measurement was not uniform across resilience studies. Individuals experience various levels of adversity (e.g. some individuals experience shorter, longer, single or multiple risks).[29] Nevertheless, the chronicity and/or multiplicity of risk is/are not always assessed. Moreover, being exposed to contexts characterised by adversity does not prove that risk was experienced. For example, some people might live in a risk-laden context, but might not experience risk as personally threatening. Vanderbilt-Adriance and Shaw[23] indicated that researchers seldom measure personal experiences of risks specifically, but rely on available national and regional statistics (i.e. sociodemographic statistics) to prove adversity. Including individuals who do not experience risks as personally threatening in studies of resilience because of their membership in risk-saturated life worlds makes the measurement of their 'resilience' questionable.[4,11]

Inadequate information about psychometric properties of resilience scales

There appears to be inadequate publication of the psychometric development and evaluation (e.g. validation of instruments) of resilience scales.[12,28] In the absence of such public knowledge, resilience researchers' use of existing scales (also across sociocultural contexts) is restricted.

Analysing and interpreting data

Accurate analyses and interpretations of data collected are vital.[31] The reviewed literature included several critiques concerning statistical analysis, the accuracy of analyses/interpretations and possible biases.

Unsophisticated statistical analysis

Masten[4,30] was unambiguous about the lack of sophisticated statistical methods across resilience studies. Her critique probably relates to the univariate (i.e. frequency analyses or comparisons of means) and bivariate (i.e. correlations, simple regression analyses or discriminant analyses) analyses most typically used in resilience studies.[43] The statistical innovations of recent years have made more sophisticated, multivariate analyses (i.e. structural equation modelling and multilevel analyses) possible. Without these, the influence of context on youths' resilience cannot be determined, and the study of resilience will be impeded.[2] Sophisticated statistical analyses are, however, limited by small samples (<200).[31] It is, therefore, possible that criticism of unsophisticated techniques relates to design and/or sampling issues.[28]

Arbitrary decisions influence analysis and interpretations

Resilience assumes experiences of severe hardship and functional outcomes. Therefore, to be deemed resilient, individuals need to fulfil both criteria (i.e. evidence of hardship and functional outcomes). Both hardship and functional outcomes can be continuous (e.g. parental conflict can range from mild to severe) or dichotomous (e.g. either having a single parent or not). With regard to continuous data, Luthar et al.[11] suggested that resilience deals with 'two tails of continua' (p.551), i.e. severe and mild conflict. In the analyses of data, this means that researchers need to make decisions on cut-off scores that prove hardship and positive adaptation. Depending on their decisions, researchers could end up with smaller or larger numbers of 'resilient' individuals (if resilient at all), which would, in turn, influence their interpretations of data collected. Hence, when researchers analyse continuous data, their choice (i.e. either severe or mild conflict) could be arbitrary, and this arbitrariness could influence their analysis and interpretation.

Conclusion to international critiques

In summary, from the critiques synthesised above, it is possible to conclude that studies of resilience can be limited, among others, by design faults. These faults include conflicting conceptualisations of resilience, undeclared or outdated theoretical frameworks, problematic measurement of resilience and unsophisticated statistical analysis. Studies to which these critiques apply offer questionable conclusions about how some youths adjust well to significant adversities.

A critical review of South African quantitative studies of youth resilience

In this phase of the review, we used the critiques listed above to comment critically on quantitative resilience studies of South African youth. To select relevant studies, we included only peer-reviewed, South African quantitative studies with 'resilience/resiliency/resilient' (as opposed to coping) in their titles, abstracts and keywords. We excluded quantitative sections of mixed-method studies, as our focus fell strictly on quantitative studies. In addition, we only included studies of children (0–18 years) and youth (15–24 years), as defined by UNESCO[44] and the UN[45]. We acknowledge the possibility of sampling bias resulting from our use of the above-mentioned inclusion criteria. Nevertheless, using these criteria, we included 13 studies.[46-58] The first study appeared in 1996 and the last in 2012. We evaluated these studies against the concerns that flowed from our synthesis of international critiques. The findings are summarised in Table 1 and are detailed below.

Table 1: Summary of South African quantitative resilience studies

	Conceptualisation of resilience	Conceptualisation of risk	Conceptualisation of protective factors	Operationalisation of resilience	Measurement scale and sample size	Potential data collection problems	Measurement of key components associated with resilience	Statistical analysis	Theoretical or conceptual framework used to interpret data
Bloemhoff[46]	Person-focused construct	Negative environments and/ or lack of skills	Assets that buffer the impact of environmental/ biological stressors	Product of various protective factors	Shortened Protective Factors Scale $N = 46$	Potential construct/ item bias Cultural appropriateness of scale not mentioned Validation of scale not reported	PF	*t*-test ANCOVA	Theorised five interrelated conceptual domains of risk and related protective factors
Bloemhoff[47]	Person-focused construct	Environmental stressors	Factors that buffer the impact of environmental/ biological stressors	Product of various protective factors	Shortened Protective Factors Scale $N = 47$	Potential construct/item bias Cultural appropriateness of scale not mentioned Validation of scale not reported	PF	*t* -test ANCOVA	Theorised five interrelated conceptual domains of risk and related protective factors
Bloemhoff[48]	Person-focused construct	Environmental stressors/lack of skills	Factors that buffer the impact of environmental/ biological stressors	Product of various protective factors	Shortened Protective Factors Scale Measure translated/back-translated $N = 29$	Potential construct/item/ administration bias Cultural appropriateness of scale not mentioned Validation of scale not reported	PF	*t* -test ANCOVA	Theorised five interrelated conceptual domains of risk and related protective factors
Choe et al.[49]	Process-oriented	Factors that increase the odds of negative outcomes (i.e. violence)	Factors associated with positive outcomes	Interaction between multiple risk and protective factors	Cumulative Measure for Adult Involvement $N = 424$	Potential construct/item/ sampling/ administration bias Cultural appropriateness of scale not mentioned Validation of scale not reported	R PF	SEM	Cumulative measures of risk and protective factors
De Villiers and Van den Berg[52]	Process-oriented	Threats to well-being (i.e. violence or poverty)	Serve a protective role	Product of various protective factors	Behavioural and Emotional Rating Scale, Resiliency Scale and Fortitude Questionnaire $N = 161$	Potential construct/item bias Cultural appropriateness of scales not mentioned Validation of scales not reported	PF	ANOVA	X
Ebersöhn[53]	Process-oriented	Threats to youth (i.e. crime)	Protective factors are seen as assets	Interaction between multiple risk and protective factors	Self-developed Questionnaire (five closed-ended questions, asking whether youth felt safe at home, etc.) Measure translated $N = 2391$	Potential sampling bias Cultural appropriateness of scale not mentioned Validation of scale not reported	R PF	Frequency analysis	Asset-focused resilience
Jorgensen and Seedat[50](publication of psychometric results)	Person-focused construct	Threats to homeostasis	X	Set of individual characteristics	Connor–Davidson Resilience Scale $N = 701$	Potential sampling bias Cultural appropriateness of scale not mentioned	R	CFA EFA	X
Kritzas and Grobler[54]	Process-oriented	X	X	Individuals' capacity to cope	Orientation to Life Questionnaire, COPE Scale and Parental Authority Questionnaire $N = 360$	Potential construct/item/ sampling bias Cultural appropriateness of scales not mentioned Validation of scales not reported	X	Hierarchical regression analysis	X
MacDonald et al.[51]	Person-focused construct	X	X	Individuals' capacity to cope	Adolescent Coping Orientation for Problem Experiences and High School Personality Questionnaire $N = 42$	Potential construct/item/ sampling bias Cultural appropriateness of scales not mentioned Validation of scales not reported Reported use of cut-off scores	X	Correlations	X

Continues on next page

	Conceptualisation of resilience	Conceptualisation of risk	Conceptualisation of protective factors	Operationalisation of resilience	Measurement scale and sample size	Potential data collection problems	Measurement of key components associated with resilience	Statistical analysis	Theoretical or conceptual framework used to interpret data
Mampane[56] (publication of psychometric results)	Person-ecological transaction	Negative environments	Protective factors are seen as strengths	Socioecological transaction	Resilience Questionnaire for Middle-adolescents in Township Schools $N = 231$	X	R PF	EFA	X
Ward et al.[57]	Process-oriented	Negative life events	Protect against negative outcomes	Multidimensional construct	Social and Health Assessment, Substance Abuse Report Scale, Survey of Exposure to Community Violence, Anxiety and Depression Subscales of the Behavioural Assessment System for Children, Peer Risk Behaviours (self-developed) and Perceived Competence Scale for Children $N = 377$	Potential construct/item/ sampling bias Cultural appropriateness of scales not mentioned Validation of scales not reported	R PF	SEM	X
Wild et al.[55]	Process-oriented	Death of a parent	Assets or resources are uni-directional	Interaction between multiple risk and protective factors	• Life Events Questionnaire for Adolescents • 10-item Acceptance Subscale from the Revised Child Report of Parent Behaviour Inventory • Five-item Monitoring Scale • Eight-item Psychological Control Scale • One item measuring peer connection (self-developed) • Adaptation of the 11-item Measure of Peer Delinquency • One item measuring respect for individuality in the adolescent's peer relationships • Four items measuring how often the adolescent had spent time during the last six months with neighbours, parents of friends, community leaders and church leaders • Five items measuring the presence of social disorganisation • Children's Depression Inventory • Children's Manifest Anxiety Scale – Revised • Seven items from the Global Self-worth Subscale of the Self-esteem Questionnaire Measure translated/back-translated $N = 159$	Potential construct/item/ sampling/ administration bias Cultural appropriateness of scales not mentioned Validation of scales not reported	R PF	Hierarchical regression analysis Correlations ANOVA	Variable-based model of resilience
Fincham et al.[58]	Process-oriented	Exposure to community violence, trauma, stress, and childhood abuse/ neglect	Protective factors that buffer the effects of risks	Product of various protective factors	Child PTSD Checklist, Child Exposure to Community Violence, Childhood Trauma Questionnaire, Perceived Stress Scale and Connor-Davidson Resilience Scale $N = 787$	Potential construct/ item bias Cultural appropriateness of scales not mentioned Validation of scales not reported	R PF	Correlations Hierarchical multiple regression analysis	X

PF, protective factors; R, risk; X, absent; ANCOVA, analysis of covariance; SEM, structural equation modelling; ANOVA, analysis of variance; EFA, exploratory factor analysis; CFA, confirmatory factor analysis

Identifying a research problem

Conflicting conceptualisations of resilience

South African quantitative studies of resilience did not conceptualise resilience uniformly. In 5 of the 13 included studies, resilience was perceived as a person-focused construct and mostly explained as a personality trait.[46-48,50,51] Furthermore, the person-focused nature of resilience was conceptualised variably as an individual's capacity to overcome or escape from risk and/or avoid negative outcomes[46-48,51] or as an individual's ability to bounce back after experiencing hardship[50]. In contrast, seven studies explained resilience as a process in which protective factors alleviate, buffer, or compensate for the effects of risks or negative outcomes.[49,52-55,57,58] One study[56] conceptualised resilience as a person ↔ ecology transaction that is sensitive to contextual factors: 'Resilience demonstrated by youths and children is not purely the result of their intrinsic characteristics; it can partly be attributed to supportive contextual and normative factors' (p.405). None of the 13 studies defined resilience as either exclusive or ordinary, as debated by Masten[8], Windle[27] and Rutter[35]. More recent studies did not reflect the evolution of resilience conceptualisations, as described by Masten[4].

Varying or absent conceptualisations of key terms

Only 11[46-50,52,53,55-58] studies specifically defined risk. In nine of these, risks were defined as witnessing conflict or violence, growing up in negative environments (e.g. poor parental supervision, parental alcoholism), or knowing environmental stressors (e.g. unemployment or poverty).[46-49,52,53,56-58] In the 10th study,[55] risks were defined as the death of a parent, which pointed to specific life events that led to negative outcomes. The 11th study[50] described risks as threats to an individual's intrinsic stability. The remaining two studies[51,54] did not clarify what risks threatened study participants. Moreover, only six studies[46,49,55-58] considered the compound nature of risks (i.e. the presence of multiple or chronic rather than single risks that left youths vulnerable).

Ten studies[46-49,52,53,55-58] explicitly defined protective factors. Of these, eight[46-49,52,53,56,58] described protective factors as interactive assets that worked together to support youths in adjusting well. The ninth[55] described protective factors as unidirectional assets aligned with a main-effects model. One study[57] implicitly conceptualised protective factors by suggesting that an internal locus of control might be protective against negative outcomes. Three studies[50,51,54] did not define protective factors at all.

Collecting data

Over-reliance on cross-sectional research designs

All 13 studies (see Table 1) followed a cross-sectional research design.

Undeclared or outdated theoretical frameworks

Only six studies[46-49,53,55] specified the theoretical framework on which they were based, and none of these frameworks reoccurred across these studies. Moreover, all six reported variable-focused theoretical frameworks that align with Wave 2 of resilience development.[4,6] Alignment with Wave 2 suggested the use of outdated frameworks, given that these studies were conducted from 2006 to 2012. During this period, international resilience research had already shifted on to Wave 3 or a pathway model approach.[4,6]

Problematic measurement of resilience

There was inconsistent measurement of resilience. To assess resilience, eight studies[46-49,52,53,55,57] measured the interaction of resources within the individual, peers, family, school and community risks. All of these studies were between 2006 and 2012. Four studies[50,51,54,58] measured indicators associated with individual characteristics of resilience. For example, measurements were taken of individual coping ability[51,54] and individual traits (i.e. personal competence and spirituality)[50,58]. A single study[56] measured resilience as a person ↔ ecology transaction. Seven[49,50,53,55-58] studies clarified which risks were measured (i.e. individual and environmental risks, negative life events, exposure to violence, and violent attitudes and behaviours). Only three[50,55,58] of these seven reported that the risks measured were personally experienced by study participants. In the remaining six studies, risk appeared to be assumed from participants' demographics (e.g. living in a socio-economically impoverished community). No justification of risk was provided in the six remaining studies.[46-48,51,52,54] For example, one[54] study gathered data from Grade-12 learners in English-medium schools that were racially integrated. It was not apparent how attending a middle-class school or a racially integrated school placed learners at risk. It would seem, therefore, that sampling bias[10] was present in the majority of studies reviewed because they appeared to include youths who were not truly at risk (as per the international definition of risk).[1,2] None included youth with disabilities or life-threatening illnesses. However, one study[55] investigated youth living in an AIDS-affected community. In addition, 6[46-48,51,52,55] of the 13 studies had sample sizes smaller than 200.

Of the 13 studies, 10[46-49,51,52,54,55,57,58] used Western scales to measure South African youth resilience without considering the cultural and contextual equivalence (or inequivalence) of the scales and their related constructs and/or items/questions to the populations of the studies; therefore, scales were invalid. Invalid scales used to measure resilience might potentiate construct or item biases.[10] One study[54] reported that the scale used (i.e. the COPE scale) was designed to be 'culture-free' (p.4). A single study[56] was sensitive to how culture and context shaped resilience and factored this into the measurement of participants' resilience by developing and validating a scale that measured risks and protective factors relevant to South African township contexts.

Inadequate information about psychometric properties of resilience scales

Two studies[50,56] reported on the development and ongoing validation of the scales used. The first[56] referred to the R-MATS (Resilience Questionnaire for Middle-adolescents in Township Schools), a multidimensional scale containing four factors: confidence and internal locus of control, social support, toughness and commitment, and achievement orientation. The R-MATS was described as a valid resilience measure in low-income, township school contexts in Mamelodi, South Africa. However, readers of the study were cautioned that the R-MATS needed to be administered to a nationally representative sample and the related psychometric property needed to be determined before it could be deemed valid for use in other South Africa populations. The second[50] detailed the five-factor structure that made up the Connor–Davidson Resilience Scale (CD-RISC), but cautioned that the CD-RISC had not been sufficiently validated for cultural groups in South Africa and that the factor structure needed to be re-examined. The remaining 11 studies[46-49,51-55,57,58] did not report any validation of their chosen resilience scales or how appropriate they were for use with South African youths.

Analysing and interpreting data

Unsophisticated statistical analysis

Quantitative studies of South African youth resilience used a variety of methods to analyse data – some more complex than others. From 1996 to 2012, six studies relied on univariate[53] and bivariate analysis[46-48,51,52], e.g. frequency analysis, correlations, analysis of variance (ANOVA) and covariance (ANCOVA). From 2007 to 2012, seven studies[49,50,54-58] employed more advanced statistical analysis (i.e. structural equation modelling, multiple regression analysis, and exploratory and confirmatory factor analysis). Most of these seven studies[49,50,54-58] reported variable-focused methodologies (Wave 2). The abundance of multivariate, variable-focused studies highlights our limited knowledge of pathways youths take towards resilience.

Arbitrary decisions influence analysis and interpretations

Only one[51] study reported cut-off scores for risks and/or functional outcomes when analysing youth resilience. However the study did not explain the rationale for these cut-off scores. Another study[55] specifically reported that because of the lack of standardisation of the scale (Depression Inventory, Children's Manifest Anxiety Scale – Revised), no cut-off scores were available and, therefore, none was made use of. The remaining 11[46-50,52-54,56-58] studies did not report the cut-off scores used to analyse risk and resilience. The lack of indicators (i.e. cut-off scores) used to identify resilience point out arbitrary decisions, and so resilience might have been overestimated and could have led to an overestimated number of youths labelled 'resilient'.

Discussion

We undertook a review of quantitative studies of South African youth resilience to comment on whether and how local studies are compromised in light of the public critiques of international studies of resilience. We have shown that the majority of published studies contributed marginally to our knowledge regarding person ↔ ecological transactions of South African youth resilience for the reasons discussed below. In spite of this finding, there were some steps in the right direction. The recent work of one study[56] demonstrates the importance of individual, contextual and cultural influence for the transactional processes of resilience in South African youth. Nonetheless, because most quantitative studies defined resilience as either a simple process or a person-focused construct (see Table 1), these positive steps are nascent and require follow-up studies to scrutinise the transactional nature of resilience processes of at-risk South African youth.

The problems inherent in the reviewed quantitative studies of South African youth resilience are mostly related to the use of outdated and/or undeclared theoretical frameworks informing conceptualisation and operationalisation, an abundance of cross-sectional studies, as well as overreliance on univariate and bivariate analyses. The majority of South African youth resilience studies explained resilience as too simple a process and failed to report the complexity and/or culturally aligned transactional nature that characterises resilience processes.[1] The lack of longitudinal studies also restricts our understanding of the long-term pathways towards resilience that South African youths take. Moreover, some methodological flaws limited how resilience was measured. International critiques revealed that individual, contextual and cultural influences shape resilience; therefore, one universal resilience measure was unlikely.[1] Our findings, however, indicated that the majority of reviewed studies made use of invalid Western scales, suggesting possible biased findings, which potentially invalidates results.[10] The lack of published psychometric results limits decisions about the validity of scales available for use with South African youth and, in so doing, restricts researchers' repertoire of culturally appropriate instruments to measure resilience. Another limitation was the lack of direct measures of risks. Resilience implies functional outcomes despite adversity, and functional outcomes outside the contexts of adversity are conceptualised as coping, not resilience.[1] Therefore, it is possible that the studies that excluded or did not specify measurement of risks[46-48,51,52,54] produced findings relating to coping rather than resilience.[1] Likewise, the absence of cut-off scores (i.e. scores used to denote risks and functional outcomes) and presence of sampling biases might have resulted in a greater number of youths being deemed resilient, without scientific proof of this. Implicit in the failure to pinpoint and measure risks and protective factors is uncertainty about what is informing young people's resilience processes.[11] In particular, the exclusion of youth with disabilities or life-threatening illnesses translates into an inadequate understanding of their resilience.

The South African youth resilience studies reviewed did not replicate international progress in the conceptualisation and measurement of resilience as a complex transactional process. Person-focused, variable-focused and pathway-model-focused studies were sporadically published from 1996 to 2012, whereas complex pathway model designs were less frequently researched. As a result, there still is little known about how South African youths transact with their ecologies towards resilience.

The above-listed caveats have implications for future quantitative youth resilience studies. To contribute meaningfully to prevailing person ↔ ecological conceptualisations of resilience[30] and to offer complex theories of youth ↔ context resilience processes among South African youths, researchers need to ground their quantitative research designs in up-to-date theoretical frameworks in ways that respect the sociocultural life worlds of South African youths.[4,11] Doing so would encourage conceptualisations and operationalisations of resilience as well as the choice of resilience scales that fit with theoretical and methodological progress made in resilience studies elsewhere[1] and that offer more socioculturally sensitive explanations of South African youths' resilience. Concomitant with this is that researchers take advantage of the statistical strengths of multivariate analysis. Univariate and bivariate analyses will not allow researchers to make complex, culturally congruent inferences regarding the transactional, contextually relevant dynamics of resilience processes.[30,43]

The lack of validated tools and/or evidence regarding validated tools does not imply that no resilience studies should be done. Rather, resilience researchers are encouraged to conduct studies using available scales, while employing various methods of multivariate analysis to establish contextual and cultural equivalence (i.e. testing for construct, metric and scalar equivalence) and to avoid potentially biased findings (i.e. exploratory and confirmatory factor analysis, target rotations and differential item functioning).[10] In addition, researchers need to prioritise the development and validation of contextually and culturally suitable instruments. Researchers are, therefore, encouraged to publish the psychometric results of scales to further stimulate the development, validation and use of contextually and culturally appropriate resilience scales with South African youth. Likewise, careful consideration should be given to how experiences of risk (lived versus exposed) and functional outcomes are chosen, measured and reported. The use of validated risk and protective factor measures, as well as culturally and contextually appropriate cut-off scores, could ensure that actual at-risk, resilient youths are being investigated, potentially evading sampling biases.[10] Finally, longitudinal studies of South African youth resilience are overdue. A continued absence of longitudinal studies will impede understanding of the long-term wellness of South African youths who are at risk.

Conclusion

We considered concerns relating to reviewed quantitative studies of South African youth resilience. The concerns are numerous and dictate sophisticated, multivariate-driven future investigations. What emerged urges future studies of South African youth resilience that are grounded in complex, person ↔ ecological conceptualisations of resilience and that employ culturally relevant measures and sophisticated statistical analyses to generate theories that illuminate the complex, culturally relevant transactions that inform the resilience processes of South African youth. Because many South African youths remain at risk, it is imperative for researchers to offer compelling evidence of how and why youths cope well with these risks.[59] Until tangible evidence is offered, mental health practitioners, service providers, educators and policymakers will not be able to intervene with confidence that their resilience-related interventions are based on sound and culturally specific scientific evidence.

Acknowledgement

We are grateful for the blinded peer reviews which supported the final version of this article.

Authors' contributions

A.v.R. was the project leader and was responsible for conceptualisation, data analysis and writing of the manuscript as fulfilment of her PhD studies. L.T. was the promoter and made conceptual contributions. S.R. was co-promoter and provided conceptual contributions.

References

1. Ungar M. Resilience, trauma, context, and culture. Trauma Violence Abus. 2013;14(3):255–266. http://dx.doi.org/10.1177/1524838013487805

2. Wright MO, Masten AS, Narayan AJ. Resilience processes in development: Four waves of research on positive adaptation in the context of adversity. In: Goldstein S, Brooks RB, editors. Handbook of resilience in children. New York: Springer; 2013. p. 15–37. http://dx.doi.org/10.1007/978-1-4614-3661-4_2

3. Luthar SS, Cushing G. Measurement issues in the empirical study of resilience: An overview. In: Glantz MD, Johnson JL, editors. Resilience and development: Positive life adaptations. New York: Plenum; 1999. p. 129–160.

4. Masten AS. Resilience in children threatened by extreme adversity: Frameworks for research, practice, and translational synergy. Dev Psychopathol. 2011;23(2):493–506. http://dx.doi.org/10.1017/S0954579411000198

5. Masten AS, Obradović J. Competence in resilience in development. Ann N Y Acad Sci. 2006;1094:13–27. http://dx.doi.org/10.1196/annals.1376.003

6. Masten AS, Wright MO. Resilience over the lifespan: Developmental perspectives on resistance, recovery, and transformation. In: Reich JW, editor. Handbook of adult resilience. New York: Guilford Press; 2010. p. 213–237.

7. Masten AS. Resilience in developing systems: Progress and promise as the fourth wave rises. Dev Psychopathol. 2007;19(3):921–930. http://dx.doi.org/10.1017/S0954579407000442

8. Masten AS. Ordinary magic: Resilience processes in development. Am Psychol. 2001;56(3):227–238. http://dx.doi.org/10.1037/0003-066X.56.3.227

9. Theron L, Theron AMC. A critical review of studies of South African youth resilience, 1990–2008. S Afr J Sci. 2010;106(7/8):11–18. http://dx.doi.org/10.4102/sajs.v106i7/8.252

10. He J, Van de Vijver FJR. The value of keeping an open eye for methodological issues in research on resilience and culture. In: Theron L, Liebenberg L, Ungar M, editors. Youth resilience and culture: Complexities and commonalities. Dordrecht: Springer; 2015. p. 189–202. http://dx.doi.org/10.1007/978-94-017-9415-2_14

11. Luthar SS, Cicchetti D, Becker B. The construct of resilience: A critical evaluation and guidelines for future work. Child Dev. 2000;71(3):543–562. http://dx.doi.org/10.1111/1467-8624.00164

12. Ahern NR, Kiehl EM, Lou Sole M, Byers J. Review of instruments measuring resilience. Issues Compr Pediatr Nurs. 2006;29(2):103–125. http://dx.doi.org/10.1080/01460860600677643

13. Barber BK. Annual research review: The experience of youth with political conflict – Challenging notions of resilience and encouraging research refinement. J Child Psychol Psychiatry. 2013;54(4):461–473. http://dx.doi.org/10.1111/jcpp.12056

14. Betancourt TS, Meyers-Ohki SE, Charrow A, Hansen N. Annual research review: Mental health and resilience in HIV/AIDS-affected children – A review of the literature and recommendations for future research. J Child Psychol Psychiatry. 2013;54(4):423–444. http://dx.doi.org/10.1111/j.1469-7610.2012.02613.x

15. Bottrell D. Understanding 'marginal' perspectives: Towards a social theory of resilience. Qual Soc Work. 2009;8(3):321–339. http://dx.doi.org/10.1177/1473325009337840

16. Cicchetti D. Annual research review: Resilient functioning in maltreated children – Past, present, and future perspectives. J Child Psychol Psychiatry. 2013;54(4):402–422. http://dx.doi.org/10.1111/j.1469-7610.2012.02608.x

17. Gartland D, Bond L, Olsson CA, Buzwell S, Sawyer SM. Development of a multi-dimensional measure of resilience in adolescents: The adolescent resilience questionnaire. BMC Med Res Methodol. 2011;11(1):134. http://dx.doi.org/10.1186/1471-2288-11-134

18. Klika JB, Herrenkohl TI. Review of developmental research on resilience in maltreated children. Trauma Violence Abus. 2013;14(3):222–234. http://dx.doi.org/10.1177/1524838013487808

19. Lerner RM. Resilience as an attribute of the developmental system: Comments on the papers of Professors Masten & Wachs. In: Lester BM, Masten MS, McEwen B, editors. Resilience in children. Boston: Blackwell; 2006. p. 40–51. http://dx.doi.org/10.1196/annals.1376.005

20. Panter-Brick C, Leckman JF. Editorial commentary: Resilience in child development – Interconnected pathways to wellbeing. J Child Psychol Psychiatry. 2013;54(4):333–336. http://dx.doi.org/10.1111/jcpp.12057

21. Rutter M. Psychosocial resilience and protective mechanisms. Am J Orthopsychiatry. 1987;57(3):316–331. http://dx.doi.org/10.1111/j.1939-0025.1987.tb03541.x

22. Tol A, Song S, Jordans MJD. Annual research review: Resilience and mental health in children and adolescents living in areas of armed conflict – A systematic review of findings in low- and middle-income countries. J Child Psychol Psychiatry. 2013;54(4):445–460. http://dx.doi.org/10.1111/jcpp.12053

23. Vanderbilt-Adriance E, Shaw DS. Conceptualizing and re-evaluating resilience across levels of risk, time, and domains of competence. Clin Child Fam Psychol Rev. 2008;11(1–2):30–58. http://dx.doi.org/10.1007/s10567-008-0031-2

24. Walsh VA, Dawson J, Mattingly MJ. How are we measuring resilience following childhood maltreatment? Is the research adequate and consistent? What is the impact on research, practice, and policy? Trauma Violence Abus. 2010;11(1):27–41. http://dx.doi.org/10.1177/1524838009358892

25. Werner EE. Vulnerability and resiliency in children at risk for delinquency: A longitudinal study from birth to young adulthood. In: Burchard JD, Burchard SN, editors. Prevention of delinquent behavior: Vermont conference on the primary prevention of psychopathology. Thousand Oaks, CA: Sage; 1987. p. 16–43.

26. Werner EE, Smith RS. Vulnerable but invincible: A longitudinal study of resilient children and youth. New York: McGraw-Hill; 1982.

27. Windle G. What is resilience? A review and concept analysis. Rev Clin Gerontol. 2011;21(2):152–169. http://dx.doi.org/10.1017/S0959259810000420

28. Windle G, Bennett K, Noyes J. A methodological review of resilience measurement scales. Health Qual Life Outcomes [serial on the Internet]. 2011 [cited 2014 Oct 03];9(8). Available from: http://www.hqlo.com/content/9/1/8.

29. Zolkoski SM, Bullock L. Resilience in children and youth: A review. Child Youth Serv Rev. 2012;34(12):2295–2303. http://dx.doi.org/10.1016/j.childyouth.2012.08.009

30. Masten AS. Resilience in children: Vintage Rutter and beyond. In: Slater A, Quinn PC, editors. Developmental psychology: Revisiting the classic studies. London: Sage; 2012. p. 204–221.

31. Creswell J. Educational research: Planning, conducting, and evaluating quantitative and qualitative research: International edition. 3rd ed. Boston: Pearson; 2012.

32. Anthony EJ, Cohler BJ. The invulnerable child. New York: Guilford Press; 1987.

33. Fletcher D, Sarkar M. A grounded theory of psychological resilience in Olympic champions. Psychol Sport Exerc. 2012;13(5):669–678. http://dx.doi.org/10.1016/j.psychsport.2012.04.007

34. Gooding PA, Hurst A, Johnson J, Tarrier N. Psychological resilience in young and older adults. Int J Geriatr Psychiatry. 2012;27(3):262–270. http://dx.doi.org/10.1002/gps.2712

35. Rutter M. Resilience: Causal pathways and social ecology. In: Ungar M, editor. The social ecology of resilience: A handbook of theory and practice. New York: Springer; 2012. p. 33–42. http://dx.doi.org/10.1007/978-1-4614-0586-3_3

36. Engle PL, Castel S, Menon P. Child development: Vulnerability and resilience. Soc Sci Med. 1996;43(5):621–635. http://dx.doi.org/10.1016/0277-9536(96)00110-4

37. Pines M. In praise of 'invulnerables'. APA Monitor. 1975 December. p. 7.

38. Beasley M, Thompson T, Davidson J. Resilience in response to life stress: The effects of coping style and cognitive hardiness. Pers Individ Dif. 2003;34:77–95. http://dx.doi.org/10.1016/S0191-8869(02)00027-2

39. Theron L, Theron AMC, Malindi MJ. Toward an African definition of resilience: A rural South African community's view of resilient Basotho youth. J Black Psychol. 2013;39(1):63–87. http://dx.doi.org/10.1177/0095798412454675

40. Werner EE. What can we learn about resilience from large-scale longitudinal studies? In: Goldstein S, Brooks RB, editors. Handbook of resilience in children. New York: Springer; 2006. p. 91–105.

41. Kail RV, Cavanaugh JC. Human development: A life-span view. 6th ed. Belmont: Wadsworth Cengage Learning; 2011.

42. Hart A, Heaver B, Brunnberg E, Sandberg A, Macpherson H, Coombe S, et al. Resilience-building with disabled children and young people: A review and critique of the academic evidence base. Int J Child Youth Fam Stud [serial on the Internet]. 2014 [cited 2014 Oct 03];5(3):394–422. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-24652

43. Monette DR, Sullivan TJ, DeJong CR. Applied social research. 8th ed. Belmont: Brooks/Cole; 2011.

44. UNESCO. What do we mean by "youth"? [homepage on the Internet]. c2014 [cited 2014 Oct 29]. Available from: http://www.unesco.org/new/en/social-and-human-sciences/themes/youth/youth-definition/.

45. United Nations Human Rights. Convention on the rights of the child [homepage on the Internet]. c1989 [cited 2014 Oct 29]. Available from: http://www.ohchr.org/en/professionalinterest/pages/crc.aspx.

46. Bloemhoff HJ. The effect of an adventure-based recreation programme (ropes course) on the development of resiliency in at-risk adolescent boys confined to a rehabilitation centre. S Afr J Res Sport Ph. 2006;28(1):1-11.

47. Bloemhoff HJ. Impact of facilitation of the effectiveness of an adventure programme for the development of resiliency in at-risk adolescent boys confined to a rehabilitation centre. Afr J Phys Health Educ Recr Dance. 2006;12(2):138–151. http://dx.doi.org/10.4314/ajpherd.v12i2.24715

48. Bloemhoff HJ. High-risk adolescent girls, resiliency and ropes course. Afr J Phys Health Educ Recr Dance [serial on the Internet]. 2012 [cited 2014 Oct 03];18:128–139. Available from: http://hdl.handle.net/10520/EJC135177

49. Choe DE, Zimmerman MA, Devnarain B. Youth violence in South Africa: Exposure, attitudes, and resilience in Zulu adolescents. Violence Vict. 2012;27(2):166–181. http://dx.doi.org/10.1891/0886-6708.27.2.166

50. Jorgensən IR, Seedat S. Factor structure of the Connor–Davidson Resilience Scale in South African adolescents. Int J Adolesc Med Health. 2008;20(1):23–32. http://dx.doi.org/10.1515/ijamh.2008.20.1.23

51. MacDonald E, Gillmer BT, Collings SJ. Adolescents in residential care: A study of personality, coping styles, and resilience. Soc Work Pract Res. 1996;9(3):233–240.

52. De Villiers M, Van den Berg H. The implementation and evaluation of a resiliency programme for children. S Afr J Psychol. 2012;42(1):93–102. http://dx.doi.org/10.1177/008124631204200110

53. Ebersöhn L. Children's resilience as assets for safe schools. J Psychol Afr. 2008;18(1):11–18.

54. Kritzas N, Grobler AA. The relationship between perceived parenting styles and resilience during adolescence. J Child Adolesc Ment Health. 2005;17(1):1–12. http://dx.doi.org/10.2989/17280580509486586

55. Wild LG, Flisher AJ, Robertson BA. Risk and resilience in orphaned adolescents living in a community affected by AIDS. Youth Soc. 2011;45(1):140–162. http://dx.doi.org/10.1177/0044118X11409256

56. Mampane R. Psychometric properties of a measure of resilience among middle-adolescents in a South African setting. J Psychol Afr. 2012;22(3):405–408.

57. Ward CL, Martin E, Theron C. Factors affecting resilience in children exposed to violence. S Afr J Psychol. 2007;37(1):165–187. http://dx.doi.org/10.1177/008124630703700112

58. Fincham DS, Altes LK, Stein DJ, Seedat S. Posttraumatic stress disorder symptoms in adolescents: Risk factors versus resilience moderation. Compr Psychiatry. 2009;50(3):193–199. http://dx.doi.org/10.1016/j.comppsych.2008.09.001

59. Berry L, Biersteker L, Dawes H, Lake L, Smith C. South African child gauge 2013. Cape Town: University of Cape Town; 2013.

Pancreatic islet regeneration: Therapeutic potential, unknowns and controversy

AUTHORS:
Ingrid L. Cockburn[1]
William F. Ferris[1]

AFFILIATION:
[1]Division of Endocrinology, Department of Medicine, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa

CORRESPONDENCE TO:
William F. Ferris

EMAIL:
wferris@sun.ac.za

POSTAL ADDRESS:
Division of Endocrinology,
Department of Medicine,
Faculty of Medicine and
Health Sciences, Stellenbosch
University, PO Box 19063,
Tygerberg 7505, South Africa

DATES:
Received: 22 Oct. 2014
Revised: 10 Nov. 2014
Accepted: 15 Nov. 2014

KEYWORDS:
diabetes; islets of Langerhans; β-cells; neogenesis; transdifferentiation

HOW TO CITE:
Cockburn IL, Ferris WF. Pancreatic islet regeneration: Therapeutic potential, unknowns and controversy. S Afr J Sci. 2015;111(7/8), Art. #2014-0369, 5 pages. http://dx.doi.org/10.17159/sajs.2015/20140369

Glucose homeostasis in mammals is primarily maintained by the insulin-secreting β-cells contained within pancreas-resident islets of Langerhans. Gross disruption of this glucose regulation as a result of pancreatic dysfunction frequently results in diabetes, which is currently a major health concern in South Africa, as well as globally. For many years, researchers have realised that the pancreas, and specifically the islets of Langerhans, have a regenerative capacity, as islet mass has frequently been shown to increase following induced pancreatic injury. Given that gross β-cell loss contributes significantly to the pathogenesis of both type 1 and type 2 diabetes, endogenous pancreatic islet regeneration has been investigated extensively as a potential β-cell replacement therapy for diabetes. From the extensive research conducted on pancreatic regeneration, opposing findings and opinions have arisen as to *how,* and more recently even *if,* pancreatic regeneration occurs following induced injury. In this review, we outline and discuss the three primary mechanisms by which pancreatic regeneration is proposed to occur: neogenesis, β-cell replication and transdifferentiation. We further explain some of the advanced techniques used in pancreatic regeneration research, and conclude that despite the technologically advanced research tools available to researchers today, the mechanisms governing pancreatic regeneration may remain elusive until more powerful techniques are developed to allow for real-time, live-cell assessment of morphology and gene expression within the pancreas.

Diabetes: Therapies and challenges

The prevalence of diabetes and its comorbidities, such as cardiovascular disease, is increasing rapidly both globally and in South Africa.[1] It is currently estimated that 347 million people worldwide suffer from the disease and the prevalence of diabetes is predicted to double between 2005 and 2030.[2] Indeed, in a recent comprehensive survey on health and nutrition in South Africa, diabetes was diagnosed in 9.6% of the survey participants (aged ≥15 years),[3] which, based on South Africa's current population, equates to ~5 million people living with the disease. Of particular concern is that the prevalence in some demographic groups far exceeds the national average: diabetes was diagnosed in as much as 30.7% of the Asian / Indian study participants.[3] This malady exerts a considerable burden of disease, which will increase with its rapidly escalating prevalence.

Diabetes is commonly subdivided into two types, with type 1 diabetes (T1D) believed to account for only 5–10% of all diabetes cases[4], although little data are currently available regarding the prevalence of T1D in South Africa[5]. The insulin deficiency associated with T1D is caused by an autoimmune destruction of insulin-producing pancreatic β-cells.[6] The hyperglycaemia and ketosis resulting from gross β-cell depletion in T1D patients are currently treated with insulin replacement,[7] and research into potential T1D therapeutics is therefore commonly focused on developing strategies to eliminate the dependence of patients on exogenous insulin. One such approach, which represents a major advancement in diabetes therapy, is islet or pancreas transplantation. Transplantation of either whole pancreata or isolated islets as a means to regain pancreatic endocrine function has been successfully used to reverse T1D; however, limitations – including an insufficient number of donor organs, poor cell viability and undesirable effects of the accompanying immunosuppressive drugs – have meant that transplantation is currently not a feasible and sustainable solution.[8,9] Although type 2 diabetes (T2D) is generally characterised by hyperglycaemia primarily resulting from insulin resistance, insufficient insulin production as a result of the loss of β-cells, as in T1D, is also important in the aetiology of the disease[10], as β-cell function and mass are known to be reduced in T2D patients[11].

To address the limitations associated with the management of diabetes, such as the exogenous insulin-dependence of patients as well as the shortcomings of islet or pancreas transplantation, possible alternative sources of β-cells for replacement therapy have been investigated. These alternatives include the stimulation of the pancreas to promote the endogenous regeneration of viable β-cells. With a view to future pharmacological or cell therapy interventions, researchers have looked to in-utero pancreatic organogenesis to ascertain which molecular pathways may be important for generating increased islet mass. In particular, the temporal expression of transcription factors involved in pancreatic cell fate determination has been investigated and characterised.

In-utero pancreatic development

Knowledge of the in-utero development of the mammalian pancreas, and in particular of the origin and development of the islets of Langerhans and β-cells, provides a starting point for investigations into potential regenerative processes in the adult pancreas. All exocrine and endocrine cell types of the pancreas originate from a common pool of progenitor cells in the gut endoderm of the embryo. A pancreatic bud forms from the endoderm, and subsequently expands and forms branched structures which eventually form the pancreatic ducts.[12] During the extension of these branches, clusters of endocrine cells bud off and aggregate to form the islets of Langerhans.[13] A brief overview of the in-utero development of islet cells and the transcription factors involved in this process is depicted in Figure 1.

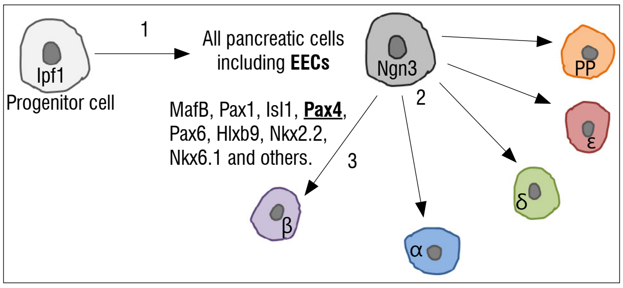

Ipf1
Progenitor cell
1
All pancreatic cells including **EECs**
Ngn3
2
MafB, Pax1, Isl1, **Pax4**, Pax6, Hlxb9, Nkx2.2, Nkx6.1 and others.
3
PP
ε
δ
β
α

Figure 1: Brief overview of the regulation of in-utero β-cell development by transcription factors during pancreatic organogenesis. (1) Mesenchymal Ipf1-expressing cells in the gut give rise to all pancreatic cells including exocrine endothelial cells (EECs). (2) A subset of EECs which express Ngn3 then differentiate into the five cell types constituting the islets of Langerhans. (3) The development of β-cells from Ngn3+progenitor cells requires the expression of numerous transcription factors including MafB, Pax1, Isl1, Pax4, Pax6, Hlxb9, Nkx2.2 and Nkx6.1.

The transcription factor Ipf1 (insulin promoter factor 1, Pdx1 in rodents) is known to play a major role in pancreatic specification and growth in early embryonic development, and indeed, lineage-tracing experiments have shown that all pancreatic cells, both exocrine and endocrine, arise from Ipf1-expressing progenitor cells.[14] Further along in pancreatic development, the endocrine portion of the pancreas (islets of Langerhans) is formed from a subset of pancreatic duct endothelial cells expressing a second transcription factor central to islet development, namely neurogenin 3 (Ngn3): Ngn3-expressing progenitors differentiate into the five endocrine cell types (α-, β-, δ-, ε- and PP-cells), which subsequently separate from the endothelium and cluster to form the islets of Langerhans.[15,16] Of the five islet cell types, α- and β-cells are the most abundant[17]; and while murine islets consist of a β-cell core surrounded by a mantle of α- and δ-cells, the α-, β- and δ-cells are dispersed throughout human and non-human primate islets[18].

The development of β-cells, in particular from Ngn3$^+$ cells, is regulated by the Pax4 gene[19], a member of the homeobox transcription factor gene family which comprises a large and diverse group of genes that play an important role in embryonic development[20]. Pax4 expression is known to peak during in-utero β-cell development[21], and, in a knockout study, Pax4-deficient mice showed significantly diminished β-cell development[19]. The absence of Pax4 expression in mature islet cells is indicative of the importance of Pax4 expression to β-cell development, in particular, rather than β-cell function or maintenance.[21]

Pancreatic regeneration

Discovery and models

Since the 1960s, many reports have demonstrated in-vivo manipulation of the pancreas by mechanical stress (injury) to stimulate the regeneration of damaged tissue and, importantly, also an increase in islet mass in the pancreas. Ligation or partial occlusion of the main pancreatic duct[22–25] as well as incomplete pancreatectomy[26,27] have been shown to result in pancreatic regeneration – observed as increased mass of the endocrine portion (islets) of pancreata following induced injury. In the 1970 study by Boquist and Edström[22], ligation of the main pancreatic duct in rats resulted in degeneration of acinar cells, with no signs of regeneration of this exocrine moiety of the pancreas. There was, however, an increase in endocrine cells following duct ligation, specifically by endocrine clusters seen to bud off from proliferating ductules and develop into islets. Similarly, Rosenberg and colleagues[23] observed the formation of new islets from hyperplasic (proliferating) ductules after partial occlusion of the pancreatic ducts of hamsters by wrapping a thin cellophane strip around the head of the pancreas. In more significant pancreatic injury (90% pancreatectomy in rats), pronounced regeneration of both the endocrine and exocrine pancreatic tissues was observed 8 weeks subsequent to surgeries.[26] All these studies involved subjecting the pancreata of study animals to prolonged mechanical stress; however, it was later shown that even brief occlusion of the main pancreatic duct (by gently squeezing the pancreas for 60 s) in rats results in duct cell proliferation and signs of islet regeneration[24], as well as overall increased endocrine mass (by 80%) 56 days after surgery, suggestive of islet neogenesis[25]. The fact that apparent islet neogenesis was observed even after only very brief occlusion of the pancreatic duct suggested that the signalling events triggering these regenerative processes in response to pancreatic injury occur immediately at the initiation of injury.

Proposed mechanisms of regeneration

The observations described above lead into key and currently unresolved questions regarding the mechanisms and processes governing islet regeneration. Islet regeneration, and specifically regeneration of β-cells following pancreatic injury, has generally been attributed to one of three mechanisms: transdifferentiation of non-endocrine cells into β-cells, neogenesis of β-cells from progenitor cells, or replication of existing β-cells (Figure 2); however, contradictory opinions and data surrounding these theories are plentiful.

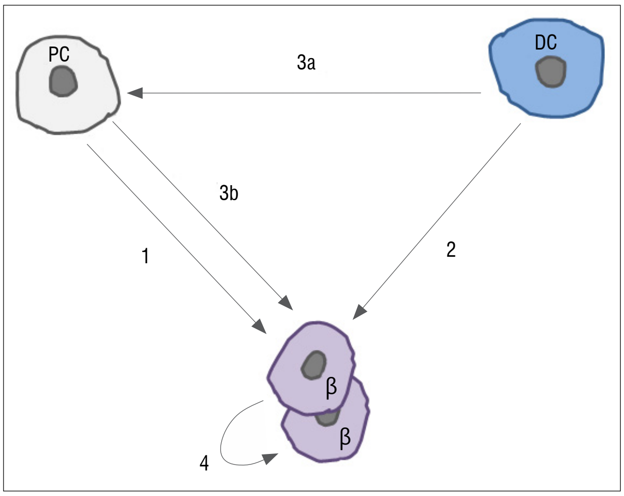


Figure 2: Proposed mechanisms of β-cell regeneration. β-Cells (β) have been proposed to regenerate via one of several mechanisms: (1) neogenesis via differentiation of specialised progenitor cells (PC); (2) transdifferentiation of differentiated non-β-cells (DC) to β-cells, in which DCs are either exocrine (acinar or ductal) or endocrine (α-cells) non-β-cells undergoing either exocrine-to-endocrine or inter-endocrine transdifferentiation, respectively; (3) de-differentiation of differentiated non-β-cells to progenitor-type cells (3a) and subsequent re-differentiation of these progenitor-type cells to β-cells (3b); or (4) replication or self-renewal of existing, differentiated β-cells.

Transdifferentiation

As islet cells originate from a subset of duct cells during pancreatic organogenesis (as described above), transdifferentiation has been investigated as a possible mechanism by which islet regeneration occurs in the adult pancreas. In adult rats subjected to pancreatic duct ligation, transdifferentiation has been identified in the pancreas in the form of cells co-expressing markers from more than one terminally differentiated cell type, indicating a mixed lineage. These cells include those expressing both epithelial and β-cell markers; cells co-expressing epithelial and α-cell markers; duct cells expressing GLUT-2, the β-cell-specific glucose transporter protein[28]; or intra- and extra-islet cells co-expressing acinar cell and β-cell markers[29]. In these and other studies,[30–32] increased islet cell number following pancreatic injury was thus attributed mainly to the transdifferentiation of non-endocrine (acinar or duct) cells to endocrine (α- or β-) cells. In an in-vitro study, AR42J acinar-derived amylase-secreting cells were converted to insulin-secreting cells by treatment with the growth factors betacellulin and activin A,[33] and although insulin secretion by these cells does not confirm their identity as β-cells, the observed conversion is suggestive of exocrine-to-endocrine transdifferentiation. In a more recent lineage-tracing study on transgenic mice subjected to dipthera toxin-induced β-cell ablation, inter-endocrine transdifferentiation was also suggested to occur: new β-cells were identified as arising from α-cells.[34]

Desai and colleagues[35] refuted the hypothesis that exocrine acinar cells transdifferentiate into endocrine β-cells based on the findings of their in vivo lineage-tracing experiments in which the acinar cells of transgenic mice were genetically labelled. This labelling strategy allowed for the progeny of these cells to be identified, thereby enabling the investigators to identify cells of acinar origin. Various models of pancreatic injury were used to induce regeneration, after which a lack of labelled endocrine cells was observed, leading the authors to conclude that 'acinar cells do not normally transdifferentiate into islet beta cells in vivo in adult mice'[35]. The findings of another genetic labelling-based study further countered the notion of transdifferentiation being the mechanism by which islet regeneration occurs by indicating that β-cells only arise from duct epithelial cells during embryogenesis, and that these epithelial cells do not significantly contribute to endocrine or acinar cell populations after birth.[36]

The findings described here clearly demonstrate that reports regarding transdifferentiation as a mechanism by which pancreatic regeneration occurs are as contradictory as they are plentiful.

Neogenesis from specialised progenitors

A population of specialised progenitor cells in the pancreas that can give rise to new β-cells would be extremely valuable for the development of endogenous cell therapies for the treatment of diabetes-associated β-cell depletion. Such a cell population could conceivably either be isolated, expanded ex vivo and used for transplantation or alternatively be stimulated to produce new β-cells in vivo. Thus far, however, whether or not such a specialised progenitor cell population exists in the pancreas remains unclear despite a number of reports in support of the existence of such cells. In corroboration with the existence of progenitor cells within the pancreas, flow cytometry has been used to identify a side population of cells in the murine pancreas which has been described as a putative stem cell population. These cells, identified as stem cells based on their ability to expel the Hoechst 33342 DNA-binding dye, were shown to undergo hyperplasia in vivo after β-cell or pancreas injury and, upon induction, the cells proliferated and differentiated in vitro giving rise to endocrine cells that exhibited glucose-stimulated insulin secretion.[37] These findings led the authors to conclude that progenitor cells that give rise to endocrine cells exist within the pancreas, a conclusion also reached by other researchers who identified endogenous β-cell progenitors on the basis of the expression of the islet cell-specific transcription factor Ngn3: Ngn3-expressing cells located in the ductal lining of the pancreas were shown to be multi-potent progenitor cells with the ability to give rise to new glucose-responsive β-cells both in situ and in vitro.[38] Very small embryonic-like stem cells (VSELs), a novel type of pluripotent stem cell reported to exist in various adult murine organs including the pancreas,[39] were recently reported to mobilise to the pancreas following partial pancreatectomy in mice. These VSELs reportedly differentiated into progenitor cells expressing Pdx1 (the transcription factor expressed by progenitor cells that differentiate into all pancreatic cell types during organogenesis), potentially giving rise to new acinar and islet cells.[40]

Based on the findings of another study using a partial pancreatectomy model in rats, islet regeneration has been suggested to occur via initial de-differentiation of duct cells to progenitor-type cells, followed by re-differentiation of these cells, which 'recapitulate aspects of embryonic pancreas differentiation' to facilitate pancreatic regeneration[41] – a mechanism which can be classed as neogenesis or transdifferentiation, or indeed a combination of the two.

A report opposing the hypothesis that β-cell regeneration following pancreatic injury occurs via neogenesis was recently published: the results of lineage-tracing experiments carried out using various models of β-cell loss in the adult murine pancreas led the authors to conclude that little to no β-cell neogenesis occurs in the adult pancreas under normal and pathological (partial pancreatectomy, duct ligation or treatment with β-cell-specific toxins) conditions.[42]

β-cell replication

The controversies and conflicting opinions within the field of islet regeneration are clearly demonstrated by the literature on β-cell replication as the mechanism driving β-cell maintenance and/or regeneration: 'Adult pancreatic beta-cells are formed by self-duplication rather than stem-cell differentiation'[43]; 'β-cell replication is the primary mechanism subserving the postnatal expansion of β-cell mass in humans'[44] and 'β-cell growth and regeneration: Replication is only part of the story'[45] are the titles of just some of these publications.

It was initially believed that significant β-cell proliferation does not occur after the initial periods of β-cell mass growth during the neonatal and infancy stages of development, and that β-cell mass is not replenished or maintained by β-cell proliferation.[46-48] Studies on β-cell replication have thus generally been carried out to assess not only β-cell replenishment following pancreatic injury, but also increases and maintenance of β-cell mass under normal physiological conditions. Similar to the lineage-tracing experiments carried out by Desai and colleagues[35] (described previously), Dor et al.[43] made use of a transgenic mouse model and a tamoxifen-inducible Cre/lox system to integrate genes into the genome by which β-cells were specifically labelled with a histochemically detectable, heritable label. Subsequent to removal of the initial 'pulse' (tamoxifen treatment), only pre-labelled β-cells and progeny of such cells would carry the detectable label, allowing newly formed β-cells to be identified as being β-cell-derived or non-β-cell-derived. These genetic labelling experiments led the authors to conclude that, during normal adult life or following pancreatectomy in the murine pancreas, adult β-cells are formed or replenished by replication rather than by islet neogenesis or differentiation of stem cells.[43] In agreement with the conclusion reached by Dor and colleagues, a later human study in which β-cell mass was assessed using computer tomography techniques revealed that β-cell mass expansion during infancy, the period during which β-cell growth rates were found to be highest, occurs primarily via β-cell replication[44]; and in a transgenic mouse model, near-complete β-cell ablation was followed by full recovery of the pancreas, which was described to occur via β-cell replication within existing islets rather than islet neogenesis[49]. In an extensive review by Bonner-Weir and colleagues[45], β-cell replication and neogenesis as mechanisms of pancreatic regeneration or postnatal islet growth are described as not being mutually exclusive. The authors go on to review the many reports on pancreatic regeneration and conclude that both neogenesis and β-cell replication contribute to β-cell mass maintenance, and that both these mechanisms have the potential to be harnessed for therapeutic applications.[45]

Doubt shed on regeneration

Controversy surrounding β-cell regeneration has recently extended from differing opinions on the mechanisms allowing for regeneration, to doubts being raised as to whether regeneration following pancreatic injury does in fact occur at all: Rankin et al.[50] recently reported on extensive experiments carried out in an adult mouse model which, in agreement with the report by Xiao and colleagues[42], show that adult β-cells do not develop from specialised pancreas-resident progenitor cells following pancreatic injury and, importantly, also that new β-cells are not generated following pancreatic duct ligation.[50] The authors of this report attribute the apparent regeneration that has frequently been described for similar pancreatic injury models to quantitative artefacts and variable recovery of pancreatic tissue; an extensive morphometric assessment of the entire murine pancreas in Rankin et al.'s study indicates that β-cell mass is unaltered in the ligated pancreas compared with sham-operated tissue and therefore that injury of the adult murine pancreas does not induce β-cell regeneration.[50]

Both Xiao et al.[42] and Rankin et al.[50] conclude that β-cell neogenesis does not occur following pancreatic injury; however, in both these studies, the expression of the islet-specific transcription factor Ngn3 was shown to be increased following pancreatic injury. Kushner and colleagues attributed the observed induction of Ngn3 expression by ligation in their study to an 'artefact due to differences in RNA recovery from injured compared with uninjured pancreas'[50], while damage to the injured portion of the pancreas (specifically exocrine cell contents and various inflammatory factors) was proposed to induce an up-regulation of Ngn3 expression in existing β-cells by Xiao and colleagues[42].

Perspective

In this brief review we have presented the diverse and often contradictory findings of some of the many investigations into pancreatic regeneration carried out over the last five decades. It is clear from these studies that putative islet regeneration is a complex, poorly understood and controversial research area, but the potential benefits of understanding and possibly harnessing the processes involved are immense. Despite recent reports contesting the existence of β-cell regeneration, a lot of unknowns in this field remain to be clarified: Why do different studies obtain vastly different results when investigating the same models and systems? If pancreatic regeneration does not take place, why is Ngn3 up-regulated following pancreatic injury?

Although sophisticated techniques available to us today, such as genetic lineage-tracing technology, are extremely powerful, they too are limited. In the case of investigations into pancreatic injury-induced events, the ideal scenario would be one in which cells can be monitored in real time within the pancreas, simultaneously assessing both morphological changes and gene expression. Although three-dimensional microscopy and the culturing of whole or partial fragments of tissues have advanced considerably in recent years, this ideal is still beyond our capabilities. Until such time that technological advances will allow for such assessments to be carried out, definitive mechanisms that potentially stimulate an increase in β-cell mass triggered by pancreatic injury remain elusive.

Acknowledgements

We thank the Harry Crossley Foundation for funding our research on pancreatic islet neogenesis.

Authors' contributions

I.C. researched and wrote the review; W.F. was involved in planning the review outline and edited drafts of the manuscript.

References

1. Danaei G, Finucane MM, Lu Y, Singh GM, Cowan MJ, Paciorek CJ, et al. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: Systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2.7 million participants. Lancet. 2011;378(9785):31–40. http://dx.doi.org/10.1016/S0140-6736(11)60679-X

2. World Health Organization. Diabetes programme [homepage on the Internet]. No date [cited 2014 Oct 21]. Available from: http://www.who.int/diabetes/en/

3. Shisana O, Labadarios D, Rehle T, Simbayi L, Zuma K, Dhansay A, et al. South African National Health and Nutrition Examination Survey (SANHANES-1). Cape Town: HSRC Press; 2013.

4. American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2009;32(Suppl 1):S62–S67.

5. Mbanya JCN, Motala AA, Sobngwi E, Assah FK, Enoru ST. Diabetes in sub-Saharan Africa. Lancet. 2010;375(9733):2254–2266. http://dx.doi.org/10.1016/S0140-6736(10)60550-8

6. Atkinson MA, Maclaren NK. The pathogenesis of insulin-dependent diabetes mellitus. N Engl J Med. 1994;331(21):1428–1436. http://dx.doi.org/10.1056/NEJM199411243312107

7. Van Belle TL, Coppieters KT, Von Herrath MG. Type 1 diabetes: Etiology, immunology, and therapeutic strategies. Physiol Rev. 2011;91(1):79–118. http://dx.doi.org/10.1152/physrev.00003.2010

8. Noguchi H. Stem cells for the treatment of diabetes. Endocr J. 2007;54(1):7–16. http://dx.doi.org/10.1507/endocrj.KR-86

9. Haller MJ, Viener H-L, Wasserfall C, Brusko T, Atkinson MA, Schatz DA. Autologous umbilical cord blood infusion for type 1 diabetes. Exp Hematol. 2008;36(6):710–715. http://dx.doi.org/10.1016/j.exphem.2008.01.009

10. Ahmad LA, Crandall JP. Type 2 diabetes prevention: A review. Clin Diabetes. 2010;28:53–59. http://dx.doi.org/10.2337/diaclin.28.2.53

11. Maedler K, Donath MY. Beta-cells in type 2 diabetes: A loss of function and mass. Horm Res. 2004;62(Suppl 3):67–73. http://dx.doi.org/10.1159/000080503

12. Shih HP, Wang A, Sander M. Pancreas organogenesis: From lineage determination to morphogenesis. Annu Rev Cell Dev Biol. 2013;29:81–105. http://dx.doi.org/10.1146/annurev-cellbio-101512-122405

13. Suckale J, Solimena M. Pancreas islets in metabolic signaling – Focus on the beta-cell. Front Biosci J Virtual Libr. 2008;13:7156–7171. http://dx.doi.org/10.2741/3218

14. Murtaugh LC. Pancreas and beta-cell development: From the actual to the possible. Dev Camb Engl. 2007;134(3):427–438. http://dx.doi.org/10.1242/dev.02770

15. Desgraz R, Herrera PL. Pancreatic neurogenin 3-expressing cells are unipotent islet precursors. Development. 2009;136(21):3567–3574. http://dx.doi.org/10.1242/dev.039214

16. Benitez CM, Goodyer WR, Kim SK. Deconstructing pancreas developmental biology. Cold Spring Harb Perspect Biol. 2012;4(6):a012401. http://dx.doi.org/10.1101/cshperspect.a012401

17. Bosco D, Armanet M, Morel P, Niclauss N, Sgroi A, Muller YD, et al. Unique arrangement of α- and β-cells in human islets of Langerhans. Diabetes. 2010;59(5):1202–1210. http://dx.doi.org/10.2337/db09-1177

18. Brissova M, Fowler MJ, Nicholson WE, Chu A, Hirshberg B, Harlan DM, et al. Assessment of human pancreatic islet architecture and composition by laser scanning confocal microscopy. J Histochem Cytochem. 2005;53(9):1087–1097. http://dx.doi.org/10.1369/jhc.5C6684.2005

19. Sosa-Pineda B, Chowdhury K, Torres M, Oliver G, Gruss P. The Pax4 gene is essential for differentiation of insulin-producing beta cells in the mammalian pancreas. Nature. 1997;386(6623):399–402. http://dx.doi.org/10.1038/386399a0

20. Holland PWH, Booth HAF, Bruford EA. Classification and nomenclature of all human homeobox genes. BMC Biol. 2007;5:47. http://dx.doi.org/10.1186/1741-7007-5-47

21. Smith SB, Ee HC, Conners JR, German MS. Paired-homeodomain transcription factor PAX4 acts as a transcriptional repressor in early pancreatic development. Mol Cell Biol. 1999;19(12):8272–8280.

22. Boquist L, Edström C. Ultrastructure of pancreatic acinar and islet parenchyma in rats at various intervals after duct ligation. Virchows Arch A. 1970;349(1):69–79. http://dx.doi.org/10.1007/BF00548522

23. Rosenberg L, Brown RA, Duguid WP. A new approach to the induction of duct epithelial hyperplasia and nesidioblastosis by cellophane wrapping of the hamster pancreas. J Surg Res. 1983;35(1):63–72. http://dx.doi.org/10.1016/0022-4804(83)90127-0

24. Ferris WF, Woodroof CW, Louw J, Wolfe-Coote SA. Brief occlusion of the main pancreatic duct rapidly initiates signals which lead to increased duct cell proliferation in the rat. Cell Biol Int. 2001;25(1):113–117. http://dx.doi.org/10.1006/cbir.2000.0683

25. Woodroof CW, De Villiers C, Page BJ, Van der Merwe L, Ferris WF. Islet neogenesis is stimulated by brief occlusion of the main pancreatic duct. S Afr Med J. 2004;94(1):54–57. http://dx.doi.org/10.1080/22201009.2004.10872330

26. Brockenbrough JS, Weir GC, Bonner-Weir S. Discordance of exocrine and endocrine growth after 90% pancreatectomy in rats. Diabetes. 1988;37(2):232–236. http://dx.doi.org/10.2337/diab.37.2.232

27. Bonner-Weir S, Baxter LA, Schuppin GT, Smith FE. A second pathway for regeneration of adult exocrine and endocrine pancreas. A possible recapitulation of embryonic development. Diabetes. 1993;42(12):1715–1720. http://dx.doi.org/10.2337/diab.42.12.1715

28. Wang RN, Klöppel G, Bouwens L. Duct- to islet-cell differentiation and islet growth in the pancreas of duct-ligated adult rats. Diabetologia. 1995;38(12):1405–1411. http://dx.doi.org/10.1007/BF00400600

29. Bertelli E, Bendayan M. Intermediate endocrine-acinar pancreatic cells in duct ligation conditions. Am J Physiol. 1997;273(5 Pt 1):C1641–C1649.

30. Gu D, Lee MS, Krahl T, Sarvetnick N. Transitional cells in the regenerating pancreas. Dev Camb Engl. 1994;120(7):1873–1881.

31. Gu D, Arnush M, Sarvetnick N. Endocrine/exocrine intermediate cells in streptozotocin-treated Ins-IFN-gamma transgenic mice. Pancreas. 1997;15(3):246–250. http://dx.doi.org/10.1097/00006676-199710000-00005

32. Lardon J, Huyens N, Rooman I, Bouwens L. Exocrine cell transdifferentiation in dexamethasone-treated rat pancreas. Virchows Arch Int J Pathol. 2004;444(1):61–65. http://dx.doi.org/10.1007/s00428-003-0930-z

33. Mashima H, Ohnishi H, Wakabayashi K, Mine T, Miyagawa J, Hanafusa T, et al. Betacellulin and activin A coordinately convert amylase-secreting pancreatic AR42J cells into insulin-secreting cells. J Clin Invest. 1996;97(7):1647–1654. http://dx.doi.org/10.1172/JCI118591

34. Thorel F, Népote V, Avril I, Kohno K, Desgraz R, Chera S, et al. Conversion of adult pancreatic α-cells to β-cells after extreme β-cell loss. Nature. 2010;464(7292):1149–1154. http://dx.doi.org/10.1038/nature08894

35. Desai BM, Oliver-Krasinski J, De Leon DD, Farzad C, Hong N, Leach SD, et al. Preexisting pancreatic acinar cells contribute to acinar cell, but not islet beta cell, regeneration. J Clin Invest. 2007;117(4):971–977. http://dx.doi.org/10.1172/JCI29988

36. Solar M, Cardalda C, Houbracken I, Martín M, Maestro MA, De Medts N, et al. Pancreatic exocrine duct cells give rise to insulin-producing beta cells during embryogenesis but not after birth. Dev Cell. 2009;17(6):849–860. http://dx.doi.org/10.1016/j.devcel.2009.11.003

37. Banakh I, Gonez LJ, Sutherland RM, Naselli G, Harrison LC. Adult pancreas side population cells expand after β cell injury and are a source of insulin-secreting cells. PLoS ONE. 2012;7(11):e48977. http://dx.doi.org/10.1371/journal.pone.0048977

38. Xu X, D'Hoker J, Stangé G, Bonné S, De Leu N, Xiao X, et al. β cells can be generated from endogenous progenitors in injured adult mouse pancreas. Cell. 2008;132(2):197–207. http://dx.doi.org/10.1016/j.cell.2007.12.015

39. Zuba-Surma EK, Kucia M, Wu W, Klich I, Lillard JW, Ratajczak J, et al. Very small embryonic-like stem cells are present in adult murine organs: ImageStream-based morphological analysis and distribution studies. Cytom Part J Int Soc Anal Cytol. 2008;73A(12):1116–1127. http://dx.doi.org/10.1002/cyto.a.20667

40. Bhartiya D, Mundekar A, Mahale V, Patel H. Very small embryonic-like stem cells are involved in regeneration of mouse pancreas post-pancreatectomy. Stem Cell Res Ther. 2014;5(5):106. http://dx.doi.org/10.1186/scrt494

41. Li W-C, Rukstalis JM, Nishimura W, Tchipashvili V, Habener JF, Sharma A, et al. Activation of pancreatic-duct-derived progenitor cells during pancreas regeneration in adult rats. J Cell Sci. 2010;123(16):2792–2802. http://dx.doi.org/10.1242/jcs.065268

42. Xiao X, Chen Z, Shiota C, Prasadan K, Guo P, El-Gohary Y, et al. No evidence for β cell neogenesis in murine adult pancreas. J Clin Invest. 2013;123(5):2207–2217. http://dx.doi.org/10.1172/JCI66323

43. Dor Y, Brown J, Martinez OI, Melton DA. Adult pancreatic β-cells are formed by self-duplication rather than stem-cell differentiation. Nature. 2004;429(6987):41–46. http://dx.doi.org/10.1038/nature02520

44. Meier JJ, Butler AE, Saisho Y, Monchamp T, Galasso R, Bhushan A, et al. β-cell replication is the primary mechanism subserving the postnatal expansion of β-cell mass in humans. Diabetes. 2008;57(6):1584–1594. http://dx.doi.org/10.2337/db07-1369

45. Bonner-Weir S, Li W-C, Ouziel-Yahalom L, Guo L, Weir GC, Sharma A. β-cell growth and regeneration: Replication is only part of the story. Diabetes. 2010;59(10):2340–2348. http://dx.doi.org/10.2337/db10-0084

46. Bonner-Weir S. Life and death of the pancreatic beta cells. Trends Endocrinol Metab. 2000;11(9):375–378. http://dx.doi.org/10.1016/S1043-2760(00)00305-2

47. Brennand K, Melton D. Slow and steady is the key to β-cell replication. J Cell Mol Med. 2009;13(3):472. http://dx.doi.org/10.1111/j.1582-4934.2008.00635.x

48. Granger A, Kushner JA. Cellular origins of β-cell regeneration: A legacy view of historical controversies. J Intern Med. 2009;266(4):325–338. http://dx.doi.org/10.1111/j.1365-2796.2009.02156.x

49. Cano DA, Rulifson IC, Heiser PW, Swigart LB, Pelengaris S, German M, et al. Regulated β-cell regeneration in the adult mouse pancreas. Diabetes. 2007;57(4):958–966. http://dx.doi.org/10.2337/db07-0913

50. Rankin MM, Wilbur CJ, Rak K, Shields EJ, Granger A, Kushner JA. β-cells are not generated in pancreatic duct ligation-induced injury in adult mice. Diabetes. 2013;62(5):1634–1645. http://dx.doi.org/10.2337/db12-0848

Fly ash composting to improve fertiliser value – A review

AUTHORS:
Hupenyu A. Mupambwa[1]
Ernest Dube[2]
Pearson N.S. Mnkeni[1]

AFFILIATIONS:
[1]Department of Agronomy, University of Fort Hare, Alice, South Africa
[2]Agricultural Research Council Small Grain Institute – Production Systems, Bethlehem, South Africa

CORRESPONDENCE TO:
Hupenyu Mupambwa

EMAIL:
hmupambwa@ufh.ac.za

POSTAL ADDRESS:
Department of Agronomy, University of Fort Hare, Private Bag X1314, Alice 5700, South Africa

DATES:
Received: 26 Mar. 2014
Revised: 17 Aug. 2014
Accepted: 01 Dec. 2014

KEYWORDS:
bioremediation; coal waste management; soil fertility; heavy metals; vermicomposting

HOW TO CITE:
Mupambwa HA, Dube E, Mnkeni PNS. Fly ash composting to improve fertiliser value – A review. S Afr J Sci. 2015;111(7/8), Art. #2014-0103, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20140103

South Africa is increasingly reliant upon coal-fired power stations for electricity generation. Fly ash, a by-product of coal combustion, contains a high total content of essential plant nutrients such as phosphorus, as well as heavy metals. If the plant nutrient bio-availability in fly ash could be improved, and the toxic element content reduced, fly ash could contribute significantly as a fertiliser source in South African agriculture. In this review, we summarise up-to-date information on the soil fertility and detoxification benefits of fly ash composting, and identify information gaps in this regard. We discuss scientific studies on the potential of fly ash based composts to supply plant nutrients and to contaminate the environment. We also explore the roles of earthworms and microorganisms in improving the decomposition process, and hence the fertiliser value of fly ash composts. Although much progress has been made, further research efforts are required to optimise microbial and earthworm activity in the decomposition process, which could further enhance nutrient supply benefits and reduce toxic elements at higher fly ash incorporation rates.

Introduction

There has been a rapid expansion in population growth, urbanisation, agricultural production and industrialisation in South Africa over the past two decades. This expansion has greatly increased electricity demand, and South Africa generates more than 90% of its electricity through coal combustion.[1] South Africa has vast coal deposits, mainly in the Central Basin, covering the Witbank, Highveld and Ermelo areas.[2] Coal seams that are relatively thick and close to the surface (15–50 m) allow low-cost coal mining, making thermal electricity power stations a much cheaper option than hydro- and nuclear power stations in South Africa.[2] Coal-fired power stations will remain the principal power generation source for South Africa in the foreseeable future as evidenced by the construction of two new coal-fired stations (the 4764 MW Medupi and 4800 MW Kusile plants).[2,3] During coal combustion, fly ash – the airborne, fine, solid residue captured from exhausts through electrostatic precipitators – is obtained, and constitutes more than 70% of the solid waste.[4] Fly ash is composed of oxidised, non-combustible materials which are very fine in size (0.01–1000 μm) and is generally greyish in colour.[1] Most fly ash particles are spherical in shape and the average diameter is 10 μm.[5] The chemical characteristics of fly ash vary widely and are influenced by coal combustion processes, age of the ash and, most importantly, coal characteristics.[6,7]

Coal is classified into three broad groups based on organic maturity: lignite, bituminous and anthracite coal.[1] Bituminous and sub-bituminous coals constitute more than 90% of South African coal and they are characterised by higher contents of CaO (5–40%), MgO (1–10%) and SO_3 (0–10%) than the higher-grade anthracite coals.[1,8,9] These different groups of coal also yield different classes of fly ash. Class F fly ash has a low total calcium (Ca) content which ranges from 1% to 12%, and is derived mainly from bituminous and anthracite coals, whereas Class C fly ash has a Ca content as high as 30–40% and is derived from lignite and sub-bituminous coals.[1,10] Class F fly ash is mostly used in the construction industry for cement making, brick making and as road bed material because of its high pozzolanic (cementing) effect.[1] Class C fly ash has a high Ca content, which makes it a potential neutralising agent in acid mine drainage and acidic soils.[8]

The physico-chemical properties of fly ash are determined primarily by the type of coal burned to produce the fly ash, hence there is significant variation in fly ash quality among and even within regions of production. The general chemical composition of fly ash consists of metal oxides that occur in the order $SiO_2 > Al_2O_3 > CaO > MgO > K_2O > NaO > TiO_2$, as highlighted in Table 1.[9] Using X-ray diffraction, Gitari et al.[11] showed that the major crystalline mineral phase in typical fly ashes are quartz (SiO_2) and mullite ($3Al_2O_3.2SiO_2$), with lower amounts of magnetite and maghemite, together with lime and calcite, which give fly ash its alkaline pH. The lime occurs as particles on the surface of the fly ash spheres and is thought to originate from decarbonation of dolomite or limestone impurities during coal formation.[8] Fly ash also contains toxic heavy metals which originate from rock weathering into coal basins (Table 2). From an agronomic point of view, fly ash is a potential fertiliser for crop production as it contains essential elements such as phosphorus (P), potassium (K), sulphur (S), sodium (Na) and magnesium (Mg) that are potentially beneficial to crop growth. The various concentrations of plant available and easily available (water soluble) nutrients, heavy metals and metalloids in fly ash are presented in Table 2.[11,12]

Agricultural application of fly ash

Fly ash is an abundant waste material which has a high total concentration of essential plant nutrients, but low bioavailability of the nutrients greatly limits its direct use in agriculture.[15,16] The low nutrient bioavailability is apparent with essential nutrients such as P, as highlighted in Table 2. This low bioavailability is partly a result of the low microbial activity in fly ash, which limits its mineralisation. Even when applied to the soil, fly ash has been reported to severely inhibit microbial respiration, enzyme activity and nitrogen (N) cycling processes.[17,18] The inhibitory effects of fly ash when applied to soil have been mainly observed in alkaline fly ashes such as the ones in South Africa, which is mainly attributed to the high salinity, pH, boron (B) toxicity and lack of substrate carbon (C) and N.[4] Schumann and Sumner[19] also highlighted that the major pitfalls in direct use of fly ash include low supply of major plant nutrients, nutrient deficiencies caused by unfavourable pH, slow nutrient release and fixation of other nutrients already present in the soil solution, such as P.

Table 1: Typical chemical concentrations of major elements as oxides in different fly ashes as analysed using X-ray fluorescence

Component	Range (mass %)			
	Europe[a]	China[a]	India[a]	South Africa[b]
SiO_2	28.5 – 59.7	35.6 – 57.2	50.2 – 59.7	50.1 – 67.0
Al_2O_3	12.5 – 33.6	18.8 – 55.0	14.0 – 32.4	23.4 – 27.0
Fe_2O_3	2.6 – 21.2	2.3 – 19.3	2.7 – 14.4	2.7 – 4.7
CaO	0.5 – 28.9	1.1 – 7.0	0.6 – 2.6	6.4 – 8.7
MgO	0.6 – 3.8	0.7 – 4.8	0.1 – 2.1	1.9 – 2.7
Na_2O	0.1 – 1.9	0.6 – 1.3	0.5 – 1.2	0 – 1.3
K_2O	0.4 – 4.0	0.8 – 0.9	0.8 – 4.7	0.5 – 0.9
P_2O_5	0.1 – 1.7	1.1 – 1.5	0.1 – 0.6	0.3 – 0.89
TiO_2	0.5 – 2.6	0.2 – 0.7	1.0 – 2.7	1.3 – 1.6
MnO	0.03 – 0.2	nd	0.5 – 1.4	0.04 – 0.5

Sources: [a]Blissett and Rowson[9]; [b]Gitari et al.[8,11,13] nd, no data

Table 2: Typical elemental concentrations of total plant available fraction (mg/kg) and easily available fraction (%) of selected fly ash samples

Element	Total concentration (mg/kg)	Plant available fraction (mg/kg)	Easily soluble fraction (%)[d]
P	553.3 – 1197.3[a]	130.0 – 256.2[a]	nd
K	0.15 – 3.5[b]	nd	0.23 – 0.25
Ca	0.11 – 22.2[b]	nd	15.84 – 24.23
Mg	0.04 – 7.6[b]	nd	0.0047 – 0.0062
Na	0.01 – 2.03[b]	nd	0.76 – 0.82
Al	0.1 – 17.3[b]	nd	0.0005 – 0.0019
Fe	3000 – 6111[a]	4.83 – 136.0[a]	0.00049 – 0.001
Mn	500 – 750[c]	0.9 – 1.5[c]	BDL
Zn	9.7 – 23.7[a]	0.6 – 0.7[a]	0 – 0.12
Cu	32 – 54[a]	0.2 – 0.9[a]	0.17 – 0.92
B	17 – 38[c]	0.5 – 0.8[c]	nd
As	1.0 – 4.0[c]	BDL[c]	BDL
Cd	5 – 10[c]	0.03 – 0.07[c]	BDL
Cr	143.7 – 488.3[a]	0.36 – 1.0[a]	0.22 – 0.54
Ni	33.3 – 69.8[a]	0.2 – 0.3[a]	BDL
Pb	26.5 – 121.3[a]	0.17 – 0.42[a]	BDL
Co	10 – 50[c]	0.05 – 0.15[c]	BDL
Se	0.6 – 2.6[c]	0.1 – 0.4[c]	2.17 – 4.83

Sources: [a]Mupambwa and Mnkeni[14]; [b]Basu et al.[6]; [c]Ram and Masto[7]; [d]Gitari et al.[11] nd, no data; BDL, below detectable limits

Globally, the utilisation of the various classes of fly ash falls within the 0–30% range, with most developing countries, including South Africa, utilising less than 5% of the fly ash that they produce, mostly in the construction industry.[4,10] In the USA, China and India, fly ash generation is in the range of 30–130 million t/year; for South Africa, it is more than 28 million t/year.[4,10] Enormous quantities of fly ash remain unused in South Africa, and they are deposited into fly ash heaps or dams close to power stations. They are not only an eyesore, but also a public health and environmental hazard because of fly ash erosion and leachate generation, which may result in sub-soil siltation and heavy metal pollution.[20] Information on leachate chemistry and contaminants attenuation in acid mine drainage by fly ash and its derivatives in South Africa is available from Gitari[8].

Much research has been carried out to demonstrate that direct application of fly ash to the soil increases the heavy metal concentration in crops and, sometimes, in soil. For example, Pandey et al.[21] mixed fly ash with garden soil at various ratios (0%, 25%, 50% and 100%) and used it as a planting medium for *Cajanus cajan*. They observed that heavy metal (Fe, Zn, Cu, Cr, Cd) accumulation in the crop was highly responsive to increases in the fly ash application rate. Similarly, Bilski et al.[22] observed higher concentrations of all heavy metals in fly ash treatments compared to the soil alone when they evaluated the germination and subsequent heavy metal accumulation during early growth of selected cereal crops. This higher bioaccumulation of heavy metals in fly ash amended treatments has been reported by several other researchers.[23,24] Apart from the low nutrient bioavailability, it appears that another major concern from direct application of fly ash to the soil in crop production is the potential accumulation of toxic heavy metals in crops. Direct application of fly ash to the soil has some positive effects, but these tend to be outweighed by the negative effects as summarised in Table 3.

It is generally agreed that addition of large quantities of fly ash to soils should be done with special consideration of pH and intensive monitoring of heavy metals.[5] There is much concern about the possible loading effects of these heavy metals through continuous soil application of fly ash and the possible leaching of the metals into groundwater.[29] These concerns limit direct utilisation and approval of fly ash as a source of plant nutrients for most edible crops. In order to address this challenge, much research has since been dedicated towards bio-remediation strategies for fly ash, such as composting.

Problem statement

If the plant nutrient bioavailability in fly ash could be improved, and the toxic element content reduced, or bio-absorbed, fly ash could contribute significantly as a nutrient source in South African agriculture. As a potential solution to this problem, there is interest in research to refine the fly ash composting strategies in a cost-effective and environmentally sustainable way. We summarise up-to-date information on the effects of fly ash composting and identify information gaps in regard to fly ash composting science, with the aim of guiding future research programmes for use of fly ash as a nutrient source in agriculture. We were guided by the following questions:

1. Can the bioavailability of plant nutrients from fly ash be improved significantly through refining the composting strategy?
2. Can the plant available fraction of heavy metals from fly ash be managed through refining the composting strategy?

Improving nutrient mineralisation in fly ash based composts

The soil nutrition improvement capacity of fly ash composts is highly variable and largely depends on the chemical characteristics of the fly ash, incorporation ratio of the fly ash and the composting technique. The variations in fly ash elemental content (total and plant available nutrients) are presented in Table 2. The most abundant primary fertiliser nutrient in fly ash is P, and it is also a major limiting nutrient to crop production.[15] Therefore, although fly ash composts supply other important plant nutrients, in this review, we focus mainly on P supply.

Traditional composting of fly ash

Traditional composting, known scientifically as thermophilic composting, is probably the oldest and most widely applied method of enhancing the fertiliser value of waste materials. It can be described as:

> *The accelerated degradation of organic matter by microorganisms under controlled conditions, during which the organic material undergoes a characteristic thermophilic phase (45°C–65°C), which allows sanitization of the waste by the elimination of pathogenic microorganisms.*[30]

During the thermophilic stage, high microbial activity increases respiration and C loss, resulting in a lower C: N ratio of the compost. A lower C: N ratio is one of the important determinants of a mature compost.[31] The end product of thermophilic composting is a stabilised and well-humified compost which should have a higher fertiliser value than the constituent materials, and no pathogens. A major disadvantage of thermophilic composting is the loss of N through volatilisation of ammonia during the thermophilic stage.[32]

Fly ash contains 0–0.2% N and 0–0.34% C, making it an inorganic by-product.[6,17] It cannot support microbial activity and it is not possible to decompose fly ash biologically, unless a rich and balanced C and N source is added to the compost.[33] Hence, numerous studies have been carried out to evaluate various organic substrates as additives in biological decomposition of fly ash. Fang et al.[34] tested the decomposition characteristics of alkaline fly ash and sewage sludge mixtures (C: N ratio of 25) and reported that fly ash incorporation rate for sewage sludge composts should not exceed 35% because the decomposition index, used to evaluate compost maturity, would be significantly decreased. This decrease was attributed to the inhibitory effect of alkaline fly ash on thermophilic microorganisms during decomposition. The high pH also causes loss of essential N through ammonification and volatilisation, thus greatly reducing microbial activity. A progressive decrease in thermophilic bacterial population and diversity was also observed when municipal green waste was amended with fly ash at 0%, 25%, 50%, 75% and 100% (w/w).[35] The amended treatments did not reach the thermophilic phase during composting, with no or little self-heating observed beyond 75% fly ash incorporation rates. A major challenge in thermophilic composting is therefore the substantial reduction of microbial activity and decomposition rate.

Microbial activity during biological decomposition produces organic acids that can solubilise minerals associated with phosphates in fly ash, resulting in increased availability of P and other essential plant nutrients.[36] Evidence of the occurrence of phosphate-solubilising microbes (PSMs) has existed since the early 20th century and PSMs have been used as bio-fertilisers since the 1950s.[37] Within PSMs, the phosphate-solubilising bacteria (PSB) are more effective than the phosphate-solubilising fungi, and they generally constitute 1–50% of the soil microbial population.[37] These PSMs release low molecular weight organic acids that bind to cations attached to the mineral phosphate, thus converting the phosphate into plant available forms.[38] Much of the research to show the benefits of PSMs has focused on solubilisation of

Table 3: Some liming and crop nutrient supply effects from direct soil application of different fly ashes under various agricultural systems

Experimental objective, fly ash and soil characteristics	Experimental conditions and treatments	Observations	References
To evaluate the effects of soil (Ultic and Typic Hapoxeralf) application of aged alkaline fly ash from two power stations on pH, salinity, available B and P, growth and uptake of B and P by rye grass. Soils had a pH of 4.7 and 5.8. The fly ash was 6 months old and had a pH of 8.9.	Pot experiments were carried out with fly ash being incorporated at 0, 5, 20 and 50 g/kg soil. Rye grass was grown in the pots for 300 days and well watered as well as fertilised. Plant samples were harvested five times for nutrient determination.	Direct fly ash addition to the soil increased pH to an average of 7.03 compared to 5.25 for the control. Originally, all the fly ashes had very low plant available P and B, hence the application did not result in any significant increase in soil P and B. However, the application of fly ash did significantly increase plant P and B. This observation highlights the need to consider the effects of fly ash on toxic nutrient concentrations in plants, even when it may not have apparent effects on soil concentration. B is toxic to plants at very low concentrations.	Matsi and Keramidas[25]
To determine the impact of fly ash from Western Australia on soil physical and chemical properties, heavy metals and subsequent growth of turf grass. Fly ash had a pH of 5.5 to 7.9 and the sandy soil had a pH of 4.7 ($CaCl_2$).	Fly ash was applied at 0, 73, 150 and 300 t/ha. It was incorporated into the soil and 7 days later turf grass (*Cynodon dactylon*) was planted.	Direct fly ash incorporation into the soil at all levels resulted in a significant increase of soil extractable P (18.5, 42.6, 46.1 and 51.2 mg/kg, respectively) but not leaf tissue P. However, a significant increase in heavy metals was also realised for Cd, Mn, Se and Zn.	Pathan et al.[26]
To clarify the differences among plant species in their response to fly ash amendment. Two types of fly ash that were used were derived from sub-bituminous and alkaline coal, and had a pH of 10.8 and 9.0, respectively. The potting mixture (50:50 sand/peat mixture) had a pH of 6.5.	The test crops were canola, radish, field peas, lucerne, barley and rye. Radish and rye grass were planted in potting mixture and the other crops were planted in soil. Ashes were applied to the pots at 0, 2.5, 5.0, 10 and 25 t/ha and fertiliser (8:3:8) applied at 20 days after planting.	Both types of fly ash significantly increased growth rates and concentrations of chlorophyll *a* and *b* at application levels of 5 t/ha, but reduced carotenoid concentrations. Addition of ash at all rates increased CO_2 assimilation of barley and radish. Application of ashes up to 5 t/ha also increased transpiration in barley. In this study, all crops showed a general difference in response to fly ash application rates, highlighting the need for crop specific recommendations for field application of fly ash.	Yunusa et al.[27]
To investigate the impact of fly ash amendment of soil on microbial responses, extent of heavy metal accumulation in the soil and rice crop growth. Unweathered fly ash with a pH of 7.7 and soil (Inceptsol) with a pH of 5.8 (in water) were used.	Pot experiments were carried out using 10 kg soil mass after fly ash amendment at 0, 5, 10, 20, 40 and 100% on a volume basis. Each pot was planted with 25-day-old rice seedlings and fertiliser (20:40:20) was applied to each pot. Destructive sampling was done at panicle initiation and at harvest.	Significant increases in crop growth parameters (chlorophyll content, plant height, leaf area index, number of panicles) were observed at fly ash application rates of 5–20%. Beyond 20% direct fly ash incorporation, significant differences were observed in heavy metals (Fe, Mn, Zn, Cu, Pb, Cr and Cd). Application of fly ash above 40% significantly influenced the microbial population dynamics and enzyme activity. These results highlighted the potential heavy metal toxicity effects of fly ash which are likely to be greater under repeated applications.	Nayak et al.[28]

P in rock phosphate,[39,40] and limited information is available on the role of PSM strains in enhancing solubilisation of fly ash P.

It would be interesting to test the effects of special microbial cocktails such as Effective Micro-organisms (EM) on the decomposition rate in fly ash based composts. According to the Japanese inventors of the technology, EM is a mixed culture of natural and beneficial microorganisms, which form clusters to make a food chain, living in a symbiotic relationship.[41,42] Effective microorganisms include predominant populations of lactic acid bacteria, yeasts, actinomycetes and photosynthetic bacteria.[41] Anecdotal evidence suggests that the use of EM as an activator can bring down the traditional composting period from 12 weeks to 4 weeks.[43] At present, in reference to traditional composting of coal fly ash mixed with different organic wastes, there is a paucity of information on the effects of EM on composting processes. It is necessary to determine if the groups of microorganisms within the EM cocktail are sufficiently resilient to remain active during composting of fly ash mixtures, and, if so, to identify the optimal EM inoculation level.

Vermicomposting of fly ash

Earthworms have an important role to play in enhancing bio-degradation and stabilisation of organic wastes. Vermicomposting has been defined as a process in which earthworms interact with microorganisms and soil invertebrates within the decomposer community, strongly affecting decomposition processes, accelerating the stabilisation of organic matter, and greatly modifying its physical and biochemical properties.[44] Earthworms are the crucial drivers of the process, as they mechanically fragment the waste with their gizzards and increase substrate surface area, thus altering micro-flora activity.[45] Earthworms significantly increase conversion of micronutrients into plant available forms in fly ash and cow dung compost mixtures.[15] However, as indicated previously, fly ash is devoid of C and N, which are essential components for any biological process, therefore, for vermicomposting, fly ash should also be enriched with a C and N source. There is theoretical evidence suggesting that microbial activity in fly ash based composts can be enhanced by adding earthworms to the composts.[15] Earthworms modulate the microbial community and tend to selectively feed more on fungi than bacteria.[44] Earthworms carry microbes in their digestive system, possibly shielding them from the direct adverse environment brought about by fly ash addition. Thus, better results at higher incorporation rates of fly ash (up to 50% to organic waste) have been reported when earthworms were added to the compost.[15,46] During vermicomposting, earthworms secrete mucus which moistens the waste, and also provide a more habitable environment for waste biodegradation through their gut micro-organisms. Research is, however, required to determine the interactions of various earthworm species with combinations of EM.

There are more than 3000 known species of earthworms, which can be divided into three categories based on their feeding behaviour, burrowing habit, habitat, body size, fecundity, casting activity and mobility.[32] Surface feeding earthworms, known as 'epigeic earthworms' have an important role to play in organic waste bio-degradation and stabilisation.[28,34] This group of earthworms is widely used for vermicomposting and includes *Eisenia fetida, Eisenia andrei* and *Eudrilus eugeniae*.[27,34,35] *Eisenia* species could be the most effective organic waste decomposers during vermicomposting.[26] This species is ubiquitous and resilient, and can feed on a wide range of organic materials and has good tolerance to a wide temperature and moisture range.[26,41]

Cow dung appears to be one of the most commonly preferred substrates for enriching fly ash with C during vermicomposting. Several studies have evaluated the transformation of nutrients during fly ash vermicomposting using various species of earthworms, mixed at various ratios with cow dung and other waste materials. Using a non-specified earthworm species, Bhattacharjee et al.[47] evaluated cow dung, soil and fly ash mixtures. The cow dung was first mixed with soil at a ratio of 2:5 and then fly ash was incorporated at six levels (5%, 10%, 15%, 25%, 40% and 50% w/w) to achieve a final weight of 700 g. These mixtures were moistened to 40–45% and then inoculated with 25 earthworms. The earthworms not only survived in the cow dung–soil mixture amended with fly ash up to 25%, but they also bio-accumulated Pb in their bodies. However, the moisture content used in this study (40–45%) may not have been the optimum for maximum activity of the earthworms, which prefer moisture contents of 50–90%.[30] The pH of the cow dung–soil–fly ash mixtures was not optimised in this study, which could have affected the vermicomposting process. With an optimal moisture content and pH level, it is possible that earthworms can tolerate a higher fly ash amendment ratio than the 25% reported in this study.

In another cow dung–fly ash vermicomposting study, Bhattacharya and Chattopadhyay[46] evaluated the potential of *E. fetida* for improving compost plant available P levels at 25%, 50% and 75% fly ash to cow dung mixing ratios. Earthworms proved superior in increasing the phosphate-utilising bacteria responsible for conversion of P to plant available forms compared to the control with no earthworms. The fly ash incorporation ratios of 25%, 50% and 75% contributed 10.8 mg/kg, 42.8 mg/kg and 12.7 mg/kg of P, respectively, after 50 days. However, it also appears that in this study, the C: N ratio and earthworm stocking density were not optimised for effective vermicomposting. It is possible that even better results could have been obtained with an optimal substrate C: N ratio and earthworm stocking density. Other fly ash vermicomposting studies, e.g. Ananthakrishnasamy et al.'s[48], did not report plant available nutrients in the composts, but rather measured the total nutrients which are most likely to increase as a result of the concentration effect from weight loss during composting, rather than exclusively earthworm activity. Substrate C: N ratio and earthworm stocking density strongly influence the vermicomposting process.[49] There is also a lack of information on the types of microbes that flourish under different fly ash incorporation ratios from these studies. Such information is required as it could form a basis for development of specialised microbial cocktails for effective bio-conversion of fly ash. Recent studies at the University of Fort Hare in South Africa, using *E. fetida* and fly ash, cow dung and waste paper mixtures indicated that a 2: 1 (cow dung–waste paper: fly ash) ratio, which gives a C: N ratio of approximately 30, may be the most appropriate, as reflected by rapid decomposition and the increase in extractable P.[50]

Fly ash–cow dung compost mixtures may sometimes have less extractable P than cow dung alone.[15] This problem is attributed to microbial community modification as evidenced by very low levels of PSB in fly ash composts.[15] In India, Bhattacharya and Chattopadhyay[46] composted cow dung and fly ash at various ratios (1:1; 1:3 and 3:1) for 50 days at room temperature and observed an average occurrence of PSB of 0.067×10^8/g for the fly ash treatments compared with 4.63×10^8/g for the cow dung alone. This microbial modification can be corrected by introducing earthworms, as shown in the follow-up study, in which the average occurrence of PSB significantly increased to 30.3×10^8/g compared with 33×10^8/g for cow dung alone. This finding also corresponded with the fly ash modified treatments which yielded 54.7% (79.9 mg/kg) more extractable P under vermicomposting than the same treatments without earthworms, which had 51.6 mg/kg.[46]

Reducing the content of toxic heavy metals in fly ash composts

Toxic heavy metals in fly ash limit its use as a direct source of nutrients in agriculture. A high soil concentration of toxic heavy metals hinders soil microbial activity,[18] thus affecting vital soil processes such as nutrient mineralisation, and effectively sterilising the soil. The plant availability of heavy metals following fly ash addition to soil tends to be variable and is controlled by the presence of Mn, Al and Fe oxides, carbonates, pH and other anions.[1,51] For example, above pH 6, an increase in surface charge on oxides of Fe, Al and Mn, which are pH dependent, coupled with binding by organic matter, greatly lowers metal availability in soil.[51] Whilst a once-off application of fly ash compost to the soil at moderate levels does not seem to present much of a heavy metal problem, the potential heavy metal load increase over time as a result of continuous application of fly ash compost is a cause for concern. Hence, any activity that will further reduce the level of heavy metals in fly ash composts is important as it will lower the risk associated with continuous fly ash compost application to soil. There is, however, a lack of information on the heavy metal dynamics in soil under continuous application of fly ash composts in the current literature, and it may be necessary to establish or model the cumulative heavy metal load associated with such.

Earthworms have the capacity to bio-accumulate heavy metals,[52-55] suggesting that earthworm harvests from the composts can be used to reduce the heavy metal load. The effects of vermicomposting as a possible way of reducing the heavy metal concentrations of fly ash and cow dung mixtures have been investigated.[53] Gupta et al.[53] started with 2 kg of feed consisting of varying proportions (20%, 40%, 60% and 80% w/w) of fly ash in cow dung with 125 mature earthworms (*E. fetida*). After 30 days, the earthworms and casting were separated and the reactor contents discarded, and a new set of 2 kg material added to which the earthworms from the previous 30-day period were added. A total of six runs of 30 days each were done following which various parameters were determined. Reductions of 85%, 77. 2%, 68.8% and 33.5% for Cr and 78.8%, 69.4%, 83.7% and 25.3% for Pb when fly ash was incorporated at rates of 20%, 40%, 60% and 80%, respectively, were reported from this study. Gupta et al.[53] reported that earthworms bio-accumulated on average 58.1 mg/kg Cr and 42.8 mg/kg Pb after 180 days of vermicomposting fly ash and cow dung mixtures which had on average 52.5 mg/kg Cr and 43.5 mg/kg Pb.[53] Bhattacharya and Chattopadhyay[15] also reported a decrease in levels of easily extractable Cr, Cd and Pb in all treatments as a result of vermicomposting after 50 days compared to the respective treatments without earthworms.

A refinement of the composting strategy will indeed improve the nutrient bioavailability and reduce the heavy metals in the large quantities of fly ash produced by coal-fired power stations. Although there is currently limited scientific information on the economics of composting, and more especially fly ash composting, the proposed technologies for fly ash composting make use of cheap and abundantly available waste materials such as cow dung, food waste, saw dust and waste paper. The earthworms and microbial populations do not require specialised, artificial conditions. As such, the cost associated with the composting of huge fly ash quantities should be minimal and the composting can be done at subsistence or commercial scale, enabling the production of cheaper fertiliser.

Conclusions

A comprehensive, up-to-date review of research on improving the fertiliser value of fly ash based composts has hitherto been unavailable. In this review, scientific studies on fly ash composting have been discussed to explore information gaps towards refining fly ash composting science. Sewage sludge, cow dung, paper and food waste are the organic substrates that are most commonly tested as sources of C and N in fly ash composting. In this case, decomposition rate, and hence nutrient release, is strongly influenced by the fly ash: organic waste mixing ratio, as well as the C: N ratio of the organic waste. The fly ash composts show great potential to supply the major elements, especially P, in crop production. A major drawback to biological decomposition of fly ash appears to be the reduction of microbial activity, population and diversity. Earthworms and special microbial cocktails such as EM, PSMs and other bio-inoculants are a potential solution to this problem. Research is required to identify the microbes that tolerate high concentrations of fly ash modification during composting. Fly ash composting appears viable mostly at low incorporation rates ranging from 5% to 25%; and at these low application rates, the heavy metals emanating from fly ash composting may not be a serious challenge as they fall within permissible limits outlined for other wastes, such as sewage sludge, in South Africa. However, repeated applications of fly ash composts to the soil over time may increase the heavy metal load to toxic levels. In this regard, research efforts aimed at further reducing the heavy metal load in fly ash composts are required.

Authors' contributions

P.N.S.M. was the project leader who initiated the project under the 'Soil Health and Sustainable Waste Management' project and provided significant intellectual input as the team leader. H.A.M. was responsible for developing the first draft from the concepts discussed at the initiation of this review and contributed significantly during revision of the manuscript. E.D. gave significant scientific input on context and relevance and also revised and refined the manuscript to its current format.

References

1. Seshadri B, Bolan NS, Naidu R, Brodie K. The role of coal combustion products in managing the bioavailability of nutrients and heavy metals in soils. J Soil Sci Plant Nutr. 2010;10:378–398. http://dx.doi.org/10.4067/S0718-95162010000100011

2. Eberhard A. The future of South African coal: Market, investment, and policy challenges. Program on Energy and Sustainable Development Working paper #100.2001. Stanford, CA: Stanford University; 2011.

3. Baker L. Governing electricity in South Africa: Wind, coal and power struggles. Working Paper 015 of The Governance of Clean Development Working Paper Series. Norwich: School of International Development, University of East Anglia; 2011.

4. Haynes RJ. Reclamation and re-vegetation of fly ash disposal sites – Challenges and research needs. J Environ Manage. 2009;90:43–53. http://dx.doi.org/10.1016/j.jenvman.2008.07.003

5. Ukwattage NL, Ranjith PG, Bouazza M. The use of coal combustion fly ash as a soil amendment in agricultural lands (with comments on its potential to improve food security and sequester carbon). Fuel. 2013;109:400–408. http://dx.doi.org/10.1016/j.fuel.2013.02.016

6. Basu M, Pande M, Bhadoria PBS, Mahapatra SC. Potential fly ash utilization in agriculture. Prog Nat Sci. 2009;19:1173–1186. http://dx.doi.org/10.1016/j.pnsc.2008.12.006

7. Ram LC, Masto RE. An appraisal of the potential use of fly ash for reclaiming coal mine spoil. J Environ Manage. 2010;91:603–617. http://dx.doi.org/10.1016/j.jenvman.2009.10.004

8. Gitari MW. Evaluation of the leachate chemistry and contaminants attenuation in acid mine drainage by fly ash and its derivatives [PhD thesis]. Cape Town: University of the Western Cape; 2006.

9. Blissett RS, Rowson NA. A review of the multi-component utilisation of coal fly ash. Fuel. 2012;97:1–23. http://dx.doi.org/10.1016/j.fuel.2012.03.024

10. Gitari WM, Petrik LF, Key DL, Okujeni C. Partitioning of major and trace inorganic contaminants in fly ash acid mine drainage derived solid residues. Int J Environ Sci Tech. 2010;7:519–534. http://dx.doi.org/10.1007/BF03326161

11. Gitari WM, Fatoba OO, Petrik LF, Vadapalli VRK. Leaching characteristics of selected South African fly ashes: Effect of pH on the release of major and trace species. J Environ Sci Heal A. 2009;44:206–220. http://dx.doi.org/10.1080/10934520802539897

12. Izquierdo M, Querol X. Leaching behaviour of elements from coal combustion fly ash: An overview. Int J Coal Geol. 2012;94:54–66. http://dx.doi.org/10.1016/j.coal.2011.10.006

13. Gitari MW, Petrik LF, Key D, Etchebers O, Okujeni C. Mineralogy and trace element partitioning in coal fly ash/acid mine drainage co-disposed solid residues. Paper presented at: World of Coal Ash (WOCA) Conference; 2005 April 11–15; Lexington, KY, USA.

14. Mupambwa HA, Mnkeni PNS. Elemental composition and release characteristics of some South African fly ashes and their potential for land application. Arch Agron Soil Sci. Forthcoming 2015. http://dx.doi.org/10.1080/03650340.2015.1017567

15. Bhattacharya SS, Chattopadhyay GN. Effect of vermicomposting on the transformation of some trace elements in fly ash. Nutr Cycl Agroecosys. 2006;75:223–231. http://dx.doi.org/10.1007/s10705-006-9029-7

16. Yadav A, Garg VK. Industrial wastes and sludges management by vermicomposting. Rev Environ Sci Biotech. 2011;10:243–276. http://dx.doi.org/10.1007/s11157-011-9242-y

17. Jala S, Goyal D. Fly ash as a soil ameliorant for improving crop production – A review. Bioresource Technol. 2006;97:1136–1147. http://dx.doi.org/10.1016/j.biortech.2004.09.004

18. Pandey VC, Singh N. Impact of fly ash incorporation in soil systems. Agric Ecosyst Environ. 2010;136:16–27. http://dx.doi.org/10.1016/j.agee.2009.11.013

19. Schumann AW, Sumner ME. Chemical evaluation of nutrient supply from fly ash–bio-solids mixtures. Soil Sci Soc Am J. 2000;64:419–426. http://dx.doi.org/10.2136/sssaj2000.641419x

20. Malik A, Thapliyal A. Eco-friendly fly ash utilization: Potential for land application. Crit Rev Env Sci Tec. 2009;39:333–366. http://dx.doi.org/10.1080/10643380701413690

21. Pandey VC, Abhilash PC, Upadhyay RN, Tewari DD. Application of fly ash on the growth performance and translocation of toxic heavy metals within *Cajanus cajan* L.: Implication for safe utilization of fly ash for agricultural production. J Hazard Mater. 2009;166:255–259. http://dx.doi.org/10.1016/j.jhazmat.2008.11.016

22. Bilski J, Jacob D, Mclean K, McLean E, Soumaila F, Lander M. Agro-toxicological aspects of coal fly ash (FA) phytoremediation by cereal crops: Effects on plant germination, growth and trace elements accumulation. Adv Bioresearch. 2012;3:121–129.

23. Pandey VC, Singh JS, Kumar A, Tewari DD. Accumulation of heavy metals by chickpea grown in fly ash treated soil: Effect on antioxidants. Clean Soil Air Water. 2010;38:1116–1123. http://dx.doi.org/10.1002/clen.201000178

24. Gautam S, Singh A, Singh J, Shikha. Effect of fly ash amended soil on growth and yield of Indian mustard (*Brassica juncea*). Adv Bioresearch. 2012;3:39–45.

25. Matsi T, Keramidas VZ. Fly ash application on two acid soils and its effect on soil salinity, pH, B, P and on ryegrass growth and composition. Environ Pollut. 1999;104:107–112. http://dx.doi.org/10.1016/S0269-7491(98)00145-6

26. Pathan SM, Aylmore LA, Colmer TD. Fly ash amendment of sandy soil to improve water and nutrient use efficiency in turf culture. Int Turfgrass Soc Res J. 2001;9:33–39.

27. Yunusa IA, Burchett MD, Manoharan V, DeSilva DL, Eamus D, Skilbeck CG. Photosynthetic pigment concentrations, gas exchange and vegetative growth for selected monocots and dicots treated with two contrasting coal fly ashes. J Environ Qual. 2009;38:1466–1472. http://dx.doi.org/10.2134/jeq2008.0285

28. Nayak AK, Raja R, Rao KS, Shukla AK, Mohanty S, Shahid M, et al. Effect of fly ash application on soil microbial response and heavy metal accumulation in soil and rice plant. Ecotoxicol Environ Saf. 2015;114:257–262. http://dx.doi.org/10.1016/j.ecoenv.2014.03.033

29. Yunusa IAM, Eamus D, DeSilva DL, Murray BR, Burchett MD, Skilbeck GC, et al. Fly-ash: An exploitable resource for management of Australian agricultural soils. Fuel. 2006;85:2337–2344. http://dx.doi.org/10.1016/j.fuel.2006.01.033

30. Dominguez J, Edwards CA. Relationships between composting and vermicomposting. In: Edwards CA, Arancon NQ, Sherman R, editors. Vermiculture technology: Earthworms, organic wastes and environmental management. New York: CRC Press; 2011. p. 11–26.

31. Raj D, Antil RS. Evaluation of maturity and stability parameters of composts prepared from agro-industrial wastes. Bioresource Technol. 2011;102:2868–2873. http://dx.doi.org/10.1016/j.biortech.2010.10.077

32. Lazcano C, Arnold J, Tato A, Zaller JG, Dominguez J. Compost and vermicompost as nursery pot components: Effects on tomato plant growth and morphology. Span J Agric Res. 2009;7:944–951. http://dx.doi.org/10.5424/sjar/2009074-1107

33. Anbalagan M, Manivannan S. Capacity of fly ash and organic additives to support adequate earthworm biomass for large scale vermicompost production. J Res Ecol. 2012;1:001–005.

34. Fang M, Wong JWC, Ma KK, Wong MH. Co-composting of sewage sludge and coal fly ash: Nutrient transformations. Bioresource Technol. 1999;67:19–24. http://dx.doi.org/10.1016/S0960-8524(99)00095-4

35. Belyaeva ON, Haynes RJ. Chemical, microbial and physical properties of manufactured soils produced by co-composting municipal green waste with coal fly ash. Bioresource Technol. 2009;100:5203–5209. http://dx.doi.org/10.1016/j.biortech.2009.05.032

36. Imran M, Waqas R, Nazli ZIH, Shaharoona B, Arshad M. Effect of recycled and value-added organic waste on solubilization of rock phosphate in soil and its influence on maize growth. Int J Agric Biol. 2011;13:751–755.

37. Khan AA, Jilani G, Akhtar MS, Naqvi SMS, Rasheed M. Phosphorus solubilizing bacterial: Occurrence, mechanisms and their role in crop production. J Agric Biol Sci. 2009;1:48–58.

38. Chen SH, Rekha PD, Arun AB, Shen FT, Lai WA, Young CC. Phosphate solubilizing bacteria from subtropical soil and their tricalcium phosphate solubizing abilities. Appl Soil Ecol. 2006;34:33–41. http://dx.doi.org/10.1016/j.apsoil.2005.12.002

39. Sibi G. Role of phosphate solubilizing fungi during phosphocompost production and their effect on the growth of tomato (*Lycopersicon esculentum* L.) plants. J Appl Natur Sci. 2011;3:287–290.

40. Aria MM, Lakzian A, Haghnia GH, Berenji AR, Besharati H, Fotovat A. Effect of Thiobacillus, sulphur and vermicompost on the water-soluble phosphorus of hard rock phosphate. Bioresource Technol. 2010;101:551–554. http://dx.doi.org/10.1016/j.biortech.2009.07.093

41. Yamada K, Xu H. Properties and applications of an organic fertilizer inoculated with effective microorganisms. J Crop Prod. 2001;3(1):255–268. http://dx.doi.org/10.1300/J144v03n01_21

42. Mupondi LT, Mnkeni PNS, Brutsch MO. The effects of goat manure, sewage sludge and effective microorganisms on the composting of pine park. Compost Sci Util. 2006;14:201–210. http://dx.doi.org/10.1080/1065657X.2006.10702284

43. Freitag DG. The use of effective microorganisms (EM) in organic waste management [document on the Internet]. c2000 [cited 2013 Dec 31]. Available from: http://www.envismadrasuniv.org/pdf.

44. Dominguez J. The microbiology of vermicomposting. In: Edwards CA, Arancon NQ, Sherman R, editors. Vermiculture technology: Earthworms, organic wastes and environmental management. New York: CRC Press; 2011. p. 53–66.

45. Lazcano C, Gomez-Brandon M, Dominguez J. Comparison of the effectiveness of composting and vermicomposting for the biological stabilization of cattle manure. Chemosphere. 2008;72:1013–1019. http://dx.doi.org/10.1016/j.chemosphere.2008.04.016

46. Bhattacharya SS, Chattopadhyay GN. Increasing bioavailability of phosphorus from fly ash through vermicomposting. J Environ Qual. 2002;31:2116–2119. http://dx.doi.org/10.2134/jeq2002.2116

47. Bhattacharjee S, Bhattacharyya G, Gupta P, Mitra A. Amendment of fly ash by vermitechnology for agricultural application. In: Chatterjee B, Singh KK, Goswanzi NG, editors. Fly ash utilisation for value added products. Jamshedpur: National Metallurgical Laboratory; 1999. p. 142–147.

48. Ananthakrishnasamy S, Sarojini S, Gunasekaran G, Manimegala G. Flyash – A lignite waste management through vermicomposting by indigenous earthworms Lampito mauritii. Am Eurasian J Agric Environ Sci. 2009;5:720–724.

49. Ndegwa PM, Thompson SA. Effects of C-to-N ratio on vermicomposting of bio solids. Bioresource Technol. 2000;75:7–12. http://dx.doi.org/10.1016/S0960-8524(00)00038-9

50. Mupambwa HA, Mnkeni PNS. Optimization of fly ash incorporation into cow dung–waste paper mixtures for enhanced vermi-degradation and nutrient release. J Environ Qual. 2015;44:972–981. http://dx.doi.org/10.2134/jeq2014.10.0446

51. Adriano DC, Bolan NS, Koo B, Naidu R, Lelie D, Vangronsveld J, et al. Natural remediation process: Bioavailability interactions in contaminated soils. Paper presented at: The 17th WCSS Conference; 2002 Aug 14–21; Bangkok, Thailand.

52. Mupondi LT. Improving sanitization and fertiliser value of dairy manure and waste paper mixtures enriched with rock phosphate through combined thermophilic composting and vermicomposting [PhD thesis]. Alice: University of Fort Hare; 2010.

53. Gupta SK, Tewari A, Srivastava R, Murthy RC, Chandra S. Potential of Eisenia fetida for sustainable and efficient vermicomposting of fly ash. Water Air Soil Poll. 2005;163:293–302. http://dx.doi.org/10.1007/s11270-005-0722-y

54. Neuhauser EF, Cukic ZV, Malecki MR, Loehr RC, Durkin PR. Bio-concentration and bio-kinetics of heavy metals in the earthworm. Environ Pollut. 1995;89:293–301. http://dx.doi.org/10.1016/0269-7491(94)00072-L

55. Li L, Xu Z, Wu J, Tian G. Bioaccumulation of heavy metals in the earthworm Eisenia fetida in relation to bioavailable metal concentrations in pig manure. Bioresource Technol. 2010;101:3430–3436. http://dx.doi.org/10.1016/j.biortech.2009.12.085

Predictive modelling of wetland occurrence in KwaZulu-Natal, South Africa

AUTHORS:
Jens Hiestermann[1]
Nick Rivers-Moore[2]

AFFILIATIONS:
[1]GeoTerraImage, Pretoria, South Africa
[2]Centre for Water Resources Research, University of KwaZulu-Natal, Pietermaritzburg, South Africa

CORRESPONDENCE TO:
Nick Rivers-Moore

EMAIL:
blackfly1@vodamail.co.za

POSTAL ADDRESS:
Centre for Water Resources Research, University of KwaZulu-Natal, Private Bag X01, Scottsville 3209, South Africa

DATES:
Received: 30 May 2014
Revised: 26 Aug. 2014
Accepted: 16 Sep. 2014

KEYWORDS:
ancillary data; Bayesian network; logistic regression; probability; wetland mapping

HOW TO CITE:
Hiestermann J, Rivers-Moore N. Predictive modelling of wetland occurrence in KwaZulu-Natal, South Africa. S Afr J Sci. 2015;111(7/8), Art. #2014-0179, 10 pages. http://dx.doi.org/10.17159/sajs.2015/20140179

The global trend of transformation and loss of wetlands through conversion to other land uses has deleterious effects on surrounding ecosystems, and there is a resultant increasing need for the conservation and preservation of wetlands. Improved mapping of wetland locations is critical to achieving objective regional conservation goals, which depends on accurate spatial knowledge. Current approaches to mapping wetlands through the classification of satellite imagery typically under-represents actual wetland area; the importance of ancillary data in improving accuracy in mapping wetlands is therefore recognised. In this study, we compared two approaches – Bayesian networks and logistic regression – to predict the likelihood of wetland occurrence in KwaZulu-Natal, South Africa. Both approaches were developed using the same data set of environmental surrogate predictors. We compared and verified model outputs using an independent test data set, with analyses including receiver operating characteristic curves and area under the curve (AUC). Both models performed similarly (AUC>0.84), indicating the suitability of a likelihood approach for ancillary data for wetland mapping. Results indicated that high wetland probability areas in the final model outputs correlated well with known wetland systems and wetland-rich areas in KwaZulu-Natal. We conclude that predictive models have the potential to improve the accuracy of wetland mapping in South Africa by serving as valuable ancillary data.

Introduction

There has been extensive loss of wetland areas globally through the combined effects of habitat loss and fragmentation, ecosystem disruption and global warming.[1-3] This loss is problematic because wetlands are highly productive environments that support unique fauna and flora[4]; in addition, these environments can be called the 'kidneys of the landscape' because of the environmental services they provide. Their hydrological and chemical cycles cleanse polluted waters, prevent floods, protect shorelines and recharge groundwater aquifers.[1,5] Wetland services include provisioning services, regulating services, cultural services and supporting services.[1,6,7] The loss of wetland ecosystems has adverse effects on the surrounding ecosystems, and, as a result, wetlands have gained considerable recognition over the past 20 years as society realises the importance of managing them.[8]

To prevent further loss and to conserve existing wetland ecosystems for their biodiversity value and ecosystem goods and services, it is important to develop an inventory of wetlands.[7,9] An inventory of wetlands forms a baseline data layer which can be used for many purposes, including comprehensive resource management plans, environmental impact assessments, natural resource inventories, habitat surveys, and the trend analysis of wetland status.[9-11] Critical to building a wetland inventory is mapping wetlands and gathering necessary information such as the wetland type, location and size. However, mapping of wetlands is notoriously difficult because the distribution of wetlands across a landscape are unique actualisations of many abiotic and biotic factors, including geological and geomorphic history, topography, connections to the local and regional hydrological system, connections to local and regional ecosystems, time since formation, and disturbance history.[12] Nevertheless, in spite of the high heterogeneity of wetlands that complicates their return signal, it is possible to generalise that three key factors – climate, topography and geology – are necessary in the formation of the hydrological conditions found in wetlands.[12] Based on this generalisation, it is consequently possible to map wetlands at a regional level.

At the broad scale, early wetland mapping exercises relied on interpretation of aerial photographs, which was time consuming and limited by the extent and resolution of the imagery available. In recent decades, the classification of satellite remote sensing has been the common approach in mapping wetlands globally.[13] Remote sensing has proved to be cost effective and a less time-consuming method of mapping wetlands over large geographical areas.[9,14,15] While there is an abundance of literature reviewing approaches to mapping wetlands using satellite imagery,[5,10,14,16-20] limitations associated with satellite image classification of wetlands exist. These limitations include spectral confusion and the misclassification of satellite imagery, which can be caused by fluctuating water levels which alter the spectral reflectance of the vegetation, or fire scars and hill shading which are often misclassified as open water on satellite imagery.[9,10]

The literature therefore also highlights the importance of ancillary data to increase the accuracy of wetlands mapped.[14] Ancillary data can be in the form of topological variables (e.g. slope, elevation, flow accumulation), environmental characteristics (e.g. soil characteristics, geology, rainfall, evaporation) and predictive models (e.g. a terrain-based hydrological model). These data have been used to improve the accuracy of many satellite image classification techniques, including wetland mapping approaches.[14,21] Ancillary data may take the form of probability surfaces, in which estimates of those parameters corresponding with identified wetland areas are used to guide the ground truthing exercise of wetland spatial images, to investigate regions where wetlands are under-represented (i.e. likely to be more prevalent than their current mapped status reflects), and to assess whether seemingly separate wetland polygons are in fact fragments of single larger wetland systems.

Predictive models have advantages that include that their outputs are readily interpretable (values range between 0 and 1, or as a percentage), that their outputs can be treated as ratios (a probability of 0.6 or 60% is twice as high as a probability of 0.3 or 30%), and that their accuracy can be tested with sample data.[22] Such models may make use of continuous frequency data (for example, logistic regression models) or continuous data which are discretised into states associated with conditional probabilities, as is the case with Bayesian network models. While both approaches essentially produce the same end product, each method offers advantages and disadvantages. In this study, we compared the probability of wetland occurrence surfaces derived from a Bayesian network (BN) with those derived from a logistic regression (LR) model. We also assessed the use of probabilistic models as a method for deriving ancillary data to supplement an existing regional wetland coverage and improve its reliability and accuracy.

Methods

Study area

The study area covered the entire province of KwaZulu-Natal (KZN), which is located in the eastern central part of South Africa (Figure 1). The western boundary of the province is marked by the Drakensberg escarpment, which reaches over 3000 m amsl in places. The escarpment in the west and the warm Mozambique current in the east account for much of the large annual variation in temperature and rainfall experienced in the province.[22,23] Partly as a result of the varied geology, topography and climate (high mean annual precipitation and relatively low potential evapotranspiration) of the province[24], wetlands are well represented in this province, covering an area of at least 4200 km^2 (approximately 5% of KZN)[25]. The hydrological regimes of wetlands in KZN are generally not only supplied by precipitation, but are also driven by a mixture of precipitation, groundwater (including infiltration, percolation and interflow) and streamflow. For example, the wetlands on the coastal plains owe their existence to the high rainfall averages and subtropical conditions as well as a series of marine regressions and transgressions that took place from 120 000 to 20 000 years BP.[26] Conversely, high rainfall, gradual sediment trapping slopes and Karoo dolerite key points make areas climatically and geologically conducive to wetland formation inland. In KZN's escarpment areas, there is a high run-off of water which is collected in the topography of the landscape, and in cases in which the mean annual precipitation exceeds the potential evapotranspiration, the saturated soils remain wet – forming wetland rich areas. A second reason KZN was selected as a study area was that the province's array of diverse natural resources makes KZN suitable for varied agricultural production (mainly sugarcane, forestry and maize), mining activities and a variety of different domestic and industrial uses.[27] These activities are increasingly exerting pressure on the province's natural resources, including wetlands.

Model data

The initial step was to construct a data set of wetland presence/absence and associated environmental (landscape and climatic) variables[1,12] (Table 1). The existing wetland layer for KZN compiled by Scott-Shaw and Escott[33] represents the best available and most comprehensive wetland data set for the province, and was used as the basis for establishing wetland presence/absence. The provincial wetland layer is a compilation product that has been ongoing over the past decade, drawing from multiple sources, including the 1:250 000 geological map, the priority wetlands coverage identified by Begg[34], manual mapping using a wide variety of aerial and satellite imagery, and information collated from private sector sources. The wetland layer was split randomly using Hawth's tools[35] into training and test data sets to avoid over-fitting of the model[36]. This wetland layer was used as the template for extracting environmental parameter statistics correlating with known wetland areas (training wetland data set), and to assess and validate the final probability layer output (test wetland data set). A key assumption was that the KZN wetland layer was accurate in terms of wetland extent area and location. To provide some level of confidence in the KZN wetland layer, an accuracy assessment of the layer was first completed before modelling began. Using 239 wetland sites from referenced aerial photographs,[37] the 2011 KZN wetland layer was visually assessed to determine if the coverage had captured their location and the extent of the wetland sites. The wetland layer correctly identified 82% (196 out of 239) of the photo reference wetland sites, providing a degree of confidence in the input wetland layer used in building and assessing the model. The 239 sites were identified from georeferenced large-scale aerial photographs, clearly identifying different wetland systems spread broadly across the entire province. These sites formed a part of a broader land-cover mapping field verification exercise, and therefore had no bias to the existing KZN wetland layer and were suitable for assessing and validating the wetland layer.

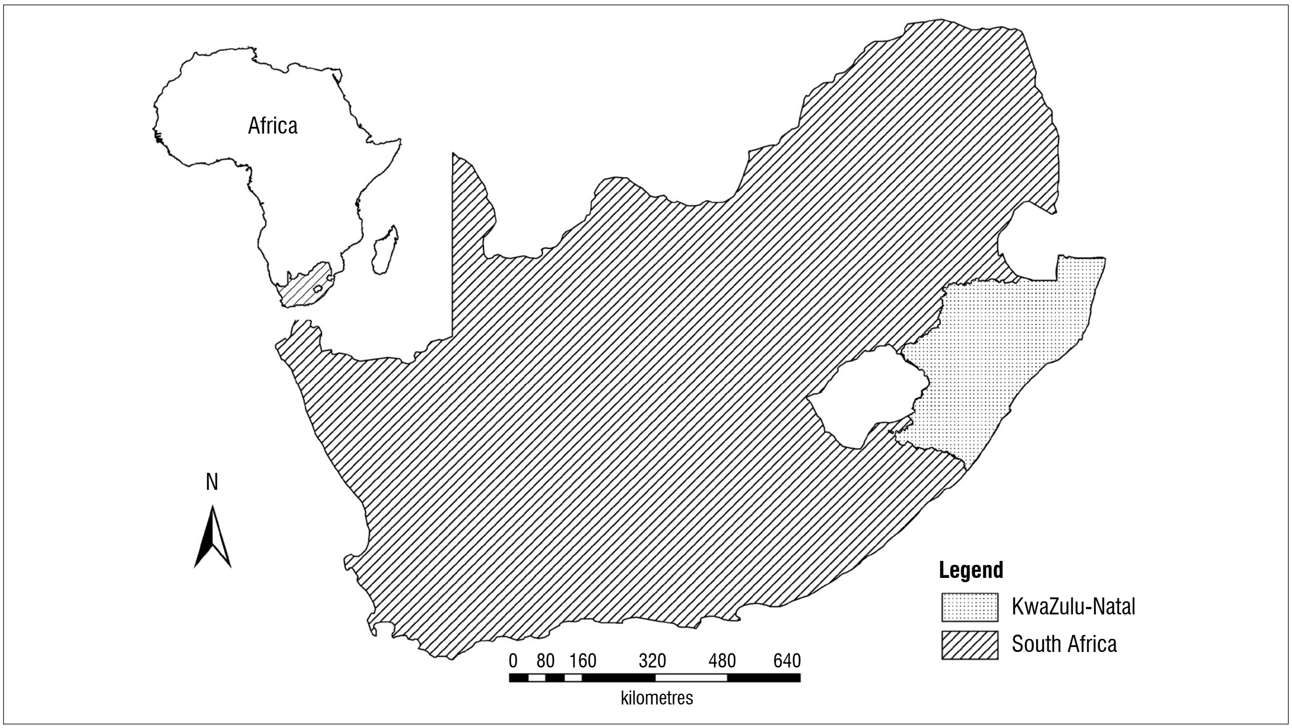


Figure 1: The study area of KwaZulu-Natal, South Africa. Note that provincial boundaries reflect those of 2008, as these corresponded with the wetland coverage used.

Table 1: Maximal list of input variables used to develop predictive wetland models

Variable type	Variable	Units
Climatic	Solar radiation[28]	MJ/m² per day
	Mean annual temperature[28]	°C
	Summer heat units[28]	° days
	Winter heat units[28]	° days
Hydrologic	Mean annual precipitation[28]	mm
	Mean annual potential evaporation[28]	mm
	Mean annual evapotranspiration[28]	mm
	Groundwater depth[29]	m
Geologic, soil and topographic	†Landform[30]	
	†Clay content[31]	–
	†Soil depth[31]	–
	†Hydromorphic soil[31]	–
	†Soil moisture[31]	–
	†Terrain units[31]	–
Digital elevation model derived	Altitude[32]	m amsl
	Slope[32]	Degrees
	Aspect[32]	Degrees
	Flow accumulation[32]	–
	Flow direction[32]	Degrees

†Ordinal variables (others are continuous)
–, no units

As the environmental variables associated with the wetlands layer differed in data format, spatial resolution, projection and extent because they were sourced from different organisations and institutions, standardisation of all input variables was required. Input variables that were model-derived had their own limitations and errors, and standardisation of these layers may have compounded these limitations; however, this possibility was unavoidable in this study. The initial step was to standardise the input variable layers in terms of projection, extent and format. The standardisation included the transformation of all layers to a common projection system (Transverse Mercator, WGS 84 datum), and resampling to a resolution of 20 m and an extent according to the digital elevation model (DEM)[32] used in the model. Most layers were resampled from a coarser (± 30 m to 1600 m) to a finer (20 m) resolution using the nearest-neighbour resampling technique. This technique was chosen because it does not alter the cell values in the categorical variables during the resampling process.

Model development

The modelling process of the study was broken into a number of steps to derive the final raster layer representing the probability of wetland occurrence in KZN. The process made use of the geographical information system software package ArcGIS 9.3[38], statistical package R[39], the multivariate statistical package MVSP[40], NETICA[41], and Medcalc[42], with additional data manipulations performed in a spreadsheet. The basis for the modelling was a spreadsheet we generated of wetland presence and absence versus associated environmental variable values. This spreadsheet consisted of approximately 45 000 statistical extraction points (Hawth's tools was used to generate extraction points for non-wetland areas) to build the database, of which 25 000 were records for wetland presence and 20 000 for wetland absence.

Next, we investigated whether there were high levels of inter-correlation between the input predictor variables. From this analysis, we wanted to derive an optimal predictor data set by eliminating redundant variables for the original list of 19 possible variables. Principal component analysis (PCA)[40] was used to select an optimal set of input variables representing greater predictive power with regard to where wetlands are likely to occur. This approach of eliminating variables using PCA could only be processed using ordinal data, therefore excluding the nominal variables in this step, namely hydromorphic soils, geology and soil association. The process of elimination of redundancy involved the stepwise analysis of the biplots and variable loadings of each input variable. Correlation of two variables resulted in the elimination of the variable with the smallest variable loading. Preference was given to variables with higher spatial resolution. The co-linearity of the data set was tested following each rerun of the PCA, until the co-linearity of the data set was below the critical threshold value of ten.[43]

Calculating probabilities for the BN required the data to be reduced to a finite set of mutually exclusive states (e.g. high, medium and low; yes or no).[44] Following this assumption, the refined pool of input variables was translated from continuous values to qualitative states of high, medium and low. Two approaches were used in discretising the data. Firstly, continuous data were reclassified into states using the Jenks natural break algorithm,[45] in which class breaks are identified that best group similar values and that maximise the differences between classes. Secondly, data in qualitative states were reclassified into states that best represented the respective variable characteristics; for example, the qualitative variable 'terrain units' discretised the states 'foot slope' and 'valley bottom' as low, 'mid-slope convex' and 'mid-slope concave' as medium and 'crest' as high. This reclassification was done for both the database and the corresponding spatial layers. Nominal variables that could not be quantified into qualitative states as required for the BN were eliminated from the model. Following the PCA elimination process and defining variable states, 'Hydromorphic Soils' was the only nominal variable added post PCA, for both models, because it could be discretised into 'yes' or 'no' states, signifying areas presumably well saturated under normal conditions, and thereby providing an additional predictor of wetland areas. Nominal variables made up of a number of nominal classes, such as soil association and geology, could not be discretised into states (high, medium or low). The discretising of variables into states with many nominal classes is a limitation of this method, and is a challenge of BNs in general.[46]

We used NETICA[41], which is the most popular BN software used in environmental modelling[36], to construct our BN model. The BN was structured using network nodes, each with a finite set of mutually exclusive states. Cain[44] explains that the links between these nodes represent their causal relationship, and each node has a set of probabilities specifying the likelihood that a node will be in a particular state given the states of those nodes that affect it directly.

Following the transformation of the input variables into states, the records in the database became cases in the case learning file for the BN, which produced conditional probabilities in a final conditional probability table (CPT), which in turn formed the basis in creating a wetland probability layer. BNs are informed through a set of cases (case learning file); in this instance, the training database represented the set of cases, and the number of cases represented the sample size.[36] Because BNs are not spatially explicit[47], a wetland probability layer had to be generated using the probabilities in the BN's CPT. The probability layer was derived by logarithmically coding the cases in both the CPT and the spatial input raster layers, so that values from the Bayesian CPT matched the equivalent logarithmic coded variable raster layers. The aggregation of the coded spatial layers created a single raster layer with unique logarithmic values as raster values. The aggregation of cases in the CPT created matching unique logarithmic values corresponding to probabilities. The unique logarithmic values with probabilities in the CPT formed the reclassification file, which was used to reclassify all the raster values in the raster layer to probabilities found in the CPT. This step created the probability layer, spatially indicating the probability of wetland occurrence in KZN. Probability values were then extracted using the test wetland and non-wetland point data set, which provided the necessary data for the receiver operating characteristic (ROC) analysis.

For a comparative approach, we used the same refined training wetland data set to fit a logit binomial model in the form of Equation 1 using a stepwise regression process.[48] The constant and variable coefficients calculated here were used to determine the probability of wetland occurrence at each wetland and non-wetland point in the binary test wetland data set. Model fitting was undertaken with the statistical package R[39] using the binary condition values (0, 1) as the response variable (generalised linear model, binomial distribution, logit link function[39]) to estimate the probability of wetland occurrence. The LR model was made spatially explicit by multiplying the variable raster layers by their determined coefficients, and then adding all raster layers together into a single layer with the model constant added to the final layer.

$$p = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}, \qquad \text{Equation 1}$$

where α is a constant and β is a coefficient for variable x, and where x could be, for example, the Landform type.

Model validation and assessment

The test wetland data set was used to compare the modelled probabilities derived from the LR model and the probabilities derived through a BN. ROC curves and the area under the curve (AUC) were used to compare the two models, and were calculated using suitable software.[42] ROC analysis is a useful technique in visualising the performance of a binary classifier system as its discrimination threshold is varied. This variability provides a richer measure of classification performance than scalar measures such as accuracy, error rate and error cost.[49] In this instance, probabilities from the BN and LR model were assessed in approximately 8600 test wetland sites (derived from the test wetland data set), and in 8600 non-wetland sites (derived randomly using Hawth's tools). The 17 200 wetland and non-wetland sites formed the binary data necessary for the ROC analysis. The predicted probabilities from both models were plotted against each other to assess their degree of correlation. Key to this analysis was determining the sensitivity and specificity of the final probability layers at a criterion threshold as well as the AUC. The sensitivity defined how many positive results occurred among the 8600 test wetland sites (Equation 2) and the specificity defined how many correct negative results occurred among the 8600 non-wetland test sites (Equation 3).

$$\text{Sensitivity} = \text{TP/P} = \text{TP/(TP + FN)} \qquad \text{Equation 2}$$

$$\text{Specificity} = \text{TN/N} = \text{TN/(FP+TN)} = \text{1-sensitivity}, \qquad \text{Equation 3}$$

where P is positive and N is negative, TP is true positive and FP is false positive, and TN is true negative and FN is false negative.

To add to the comparison of the two models, we plotted a correlation between probabilities as outputs from each approach. This was further complemented and quantified using Cohen's kappa statistic to provide a statistical measure of agreement between the BN probability layer and the LR probability layer.[50] Using a determined threshold, the BN and LR probability outputs were transformed into two binary (wetland versus non-wetland) data sets for an agreement comparison.

To assess model usefulness for predicting wetland occurrence and extent, we generated 1000 random points across 20 classes of probability (i.e. 5–10%, 10–15%, 15–20%, etc.) and used Hawth's tools[35] to analyse the trend in accuracy in predicting wetland occurrence and extent. Wetland occurrence accuracy was determined by assessing whether a point occurred in a wetland area and confirmed using satellite imagery (Spot5 2009; Google Earth™). Here, each point was recorded as either falling within a wetland clearly identifiable on the imagery, or not, drawing on previous experience gained in desktop wetland delineation. The wetland extent accuracy was determined using the test wetland data set to obtain a probability value versus wetland extent area curve i.e. wetland extent area covered at increasing probability value intervals (i.e. 0–100%, 5–100%, 10–100%, etc.).

Results

The maximal PCA accounted for 53.66% of the cumulative variance in axes 1, 2 and 3, with high levels of correlation (i.e. redundancy) between variables originally considered. The co-linearity condition number of the first PCA iteration was 37.9, exceeding the critical value of 10.0.[37] An example of this redundancy was the high correlation (r^2<-0.9) between the clay and altitude variables. The clay variable was eliminated from the PCA as a result of this correlation, and because the altitude variable had a higher spatial resolution and accounted for more variation. Following the first iteration of the PCA, flow direction, flow accumulation, soil moisture and aspect (modified to ordinal variables) were eliminated because of their short vector length, which signifies only a small influence on the determination of wetland probability. In the second PCA iteration, the temperature variables (winter heat units, summer heat units, and mean annual temperature) were highly negatively correlated (r^2<-0.9) with altitude, indicated by the ±180° angle separating the vectors. Because mean annual temperature accounted for more variation and had a higher spatial resolution, the additional temperature variables were eliminated. In the third PCA iteration, the biplot vectors of the groundwater and slope (DEM-derived[32]) variables displayed virtually no angle between them (i.e. they were correlated), resulting in the elimination of groundwater because it was the variable with the shortest vector and smaller variable PCA loading. In the final PCA iteration, evaporation was highly correlated (r^2=0.895) with the evapotranspiration variable. Evaporation was eliminated because it had the lower variable PCA loading of the two. Following the elimination of the evaporation variable, the data set's co-linearity condition number fell below the recommended critical threshold value.[43] After these iterations, the maximal data set was reduced to eight ordinal variables, with a resultant co-linearity condition number of 5.3. The final iteration of the PCA accounted for 69.15% of the cumulative variance in axes 1, 2 and 3 (Figure 2; Table 2), with the remaining input variables into the models being Mean Annual Precipitation, Slope (degrees), the 20-m DEM (hereafter referred to as 'Altitude'), Mean Annual Solar Radiation (hereafter referred to as Solar Radiation), Soil Depth, Evapotranspiration, Terrain Units and Landform. Following the PCA elimination process, Hydromorphic Soils was the only nominal variable added post-PCA. The final spreadsheet for both models was therefore based on a common predictor data set for nine variables (eight quantitative variables and one nominal variable).

Table 2: Eigenvector scores for the remaining ordinal input variables for Axes 1 and 2

PCA variable loadings	Axis 1	Axis 2
Soil depth	0.483	-0.234
Terrain units	-0.369	0.222
Solar radiation	-0.148	-0.650
Mean annual precipitation	-0.181	0.271
Evapotranspiration	0.447	0.154
Altitude	-0.436	-0.509
Slope	-0.398	0.271
Landform	-0.167	0.212

PCA, principal component analysis

The BN model was informed by learning the cases in the database (case learning file) of the remaining input variables (Figure 3). The case learning file formed the central input for calculating the prior probabilities of all the parent node variables in the BN, as well as the final conditional probabilities. The output of the BN was a table with the conditional probabilities of wetland probability given the state of each input variable. The input variables reduced to qualitative states (high, medium, low) formed the parent nodes and the 'probability of wetland occurrence' formed the child node. The LR model to estimate probability of wetland occurrence (i.e. wetland = yes or no) was significant for all nine variables (p<0.05) (Table 3).

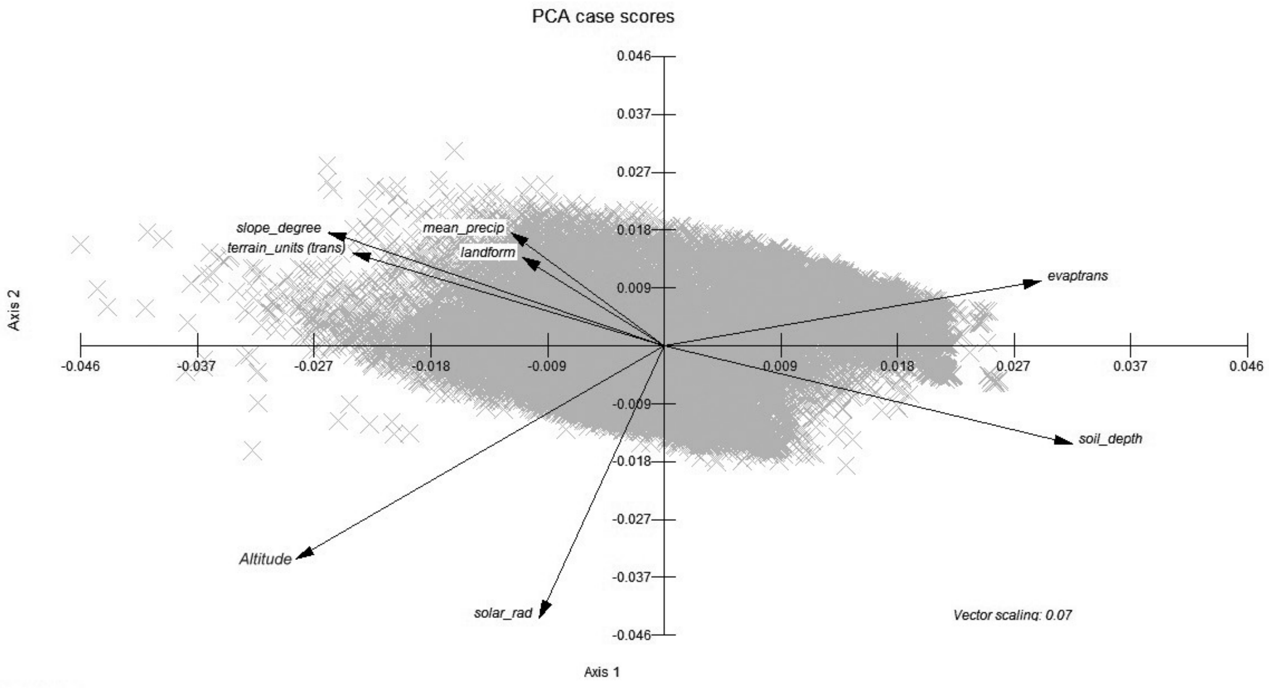


Figure 2: Biplot of the final principal component analysis (PCA) showing eight variables and with the co-linearity coefficient reduced to 5.3. Eigenvalues for Axes 1 and 2 (cumulative percentages of variation accounted for in brackets) are 2.51 (31.43%) and 1.60 (51.43%), respectively.

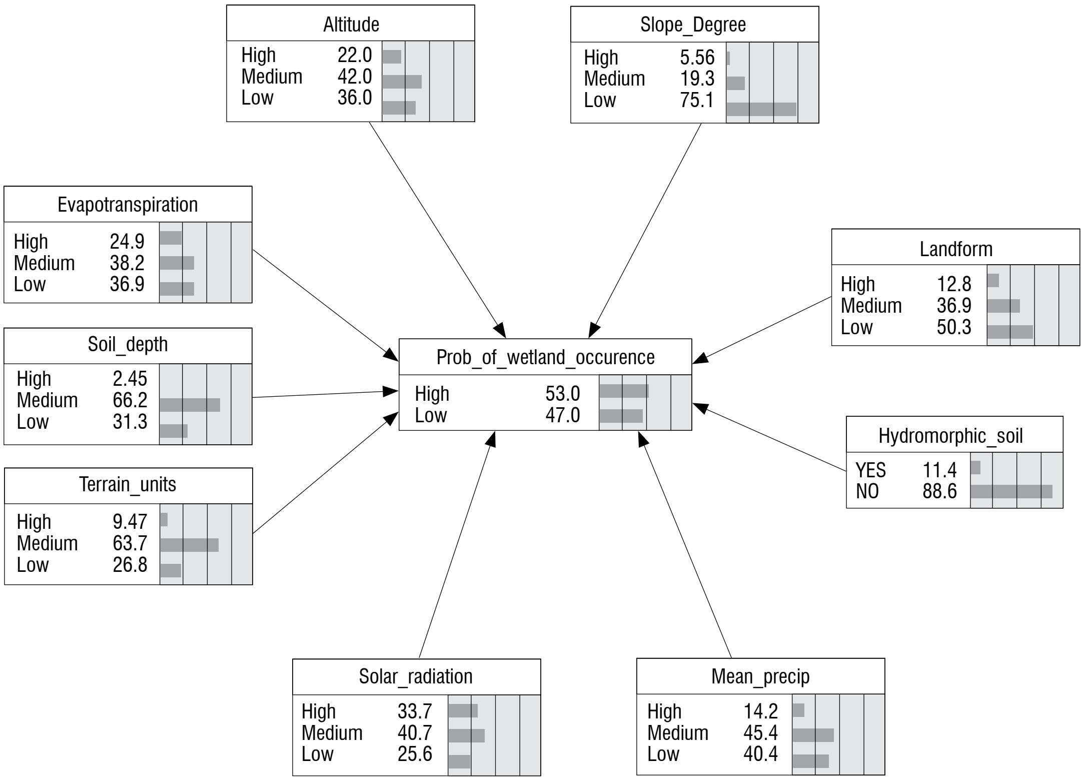


Figure 3: Diagram illustrating the Bayesian network structure used in calculating the conditional probabilities. The outer parent nodes are the remaining input variables and the child node is the 'probability of wetland occurrence'.

Table 3: Coefficients and standard errors for variables used in the logistic regression model

Variable	Coefficient	Standard error
Constant	4.347	0.001
Altitude	0.001	0.001
Evapotranspiration	0.018	0.002
Hydromorphic soil	0.124	0.043
Landform	-0.113	0.007
Mean annual precipitation	0.002	0.001
Slope	-0.129	0.003
Soil depth	0.382	0.021
Solar radiation	-0.451	0.017
Terrain units	0.668	0.013

p*<0.005 for all variables*

The final outputs of the BN and LR models were two raster layers with the pixel values representing probabilities of being a wetland at a spatial resolution of 20 m (Figure 4; Table 4). Probability values below 0.50 accounted for over 60% (~57 000 km^2) of the total KZN area, while probability values of 0.80 and above accounted for only approximately 4–6% (3700–5500 km^2) of the total KZN area (Table 4). The BN probability map appeared to be more conservative in estimating area than the LR map, for which, at a threshold of 0.6, the percentage area covered by the LR model was 0.88% more than the area covered by the BN model, even though probabilities from the BN and LR models were generally strongly correlated (r^2=0.79). For Cohen's kappa statistic, both LR and BN layers were transformed into two binary data sets using a 0.6 probability cut-off. Based on Cohen's kappa result, observed and predicted wetland occurrences were more than 78% similar in all cases, with a 91% agreement between the LR- and the BN-derived probability layers (Figure 5).

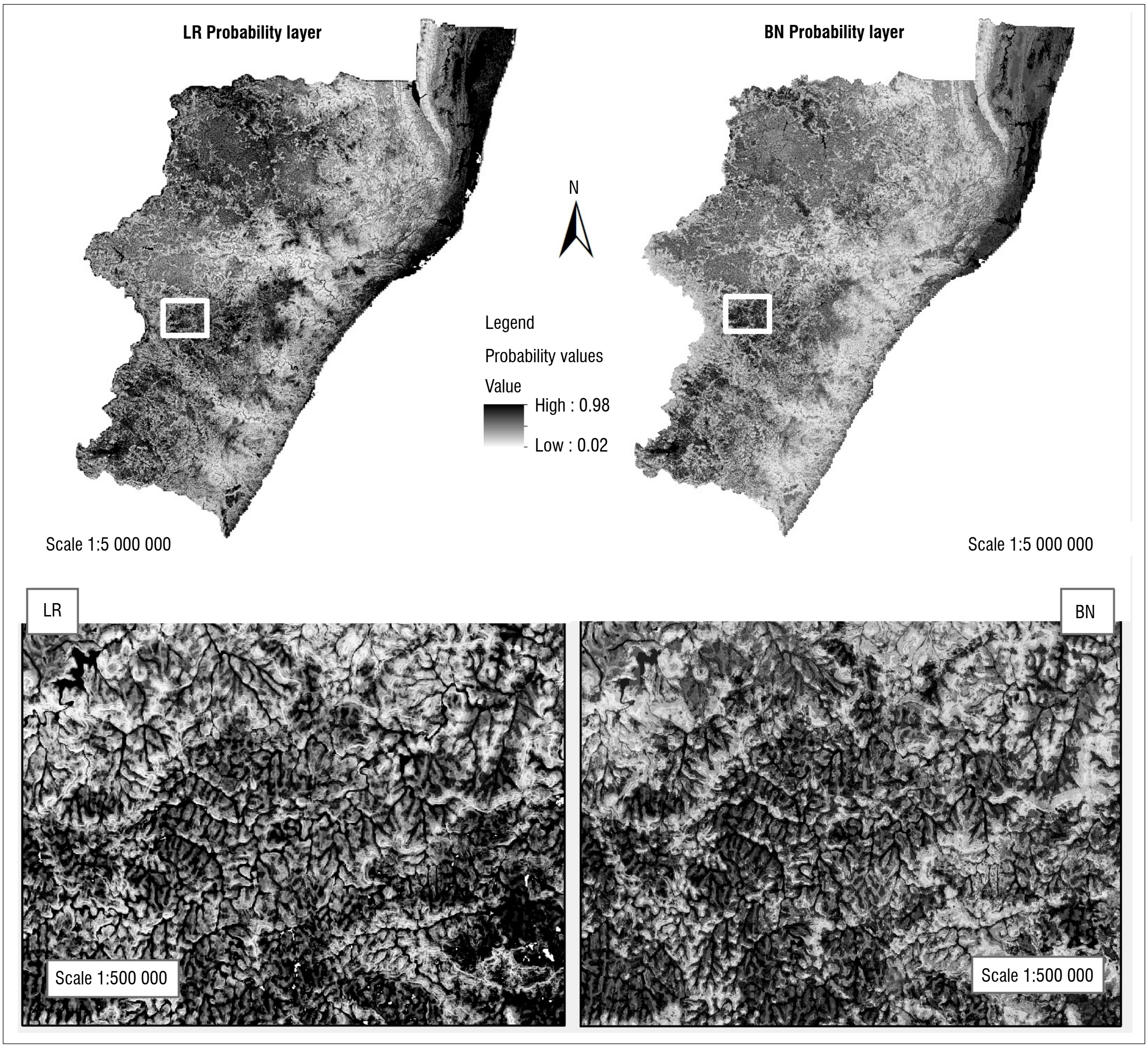


Figure 4: Comparison of the logistic regression (LR) and Bayesian network (BN) probability outputs ranging in scale from 1:5 000 000 (top) to 1:500 000 (bottom). Pixel values range from 0.02 to 0.98, in which the higher the probability values of a pixel, the greater the likelihood of wetland occurrence.

Table 4: Percentage area of KwaZulu-Natal covered by wetlands at different probability value thresholds as determined by the logistic regression and Bayesian network models

Probability threshold	Area covered (%)	
	Logistic regression	Bayesian network
0	100	100
0.1	40.72	42.34
0.2	34.26	34.21
0.3	28.04	28.39
0.4	22.90	22.38
0.5	16.44	16.19
0.6	11.65	9.96
0.7	7.51	5.78
0.8	4.23	2.79
0.9	1.68	0.98
1	0	0

Model validation and assessment

Results from the ROC curves indicated that the AUC for the predicted probabilities of the BN model and LR model were 0.853 (SE=0.00287; 95% confidence level = 0.847–0.858) and 0.840 (SE=0.00301; 95% confidence level = 0.835–0.846), respectively (Figure 6). An AUC of 1 indicates perfect prediction, whereas an AUC of 0.5 indicates completely random binary prediction. Although there was a marginal difference in the AUC results for the BN and LR models, the difference was not pronounced enough to conclusively state that one model has outperformed the other in predicting wetland occurrence. ROC analyses were performed using the test wetland data set to compare the predicted probabilities derived from a simple binary LR to the derived and predicted probabilities from the BN. The ROC analyses performed on the BN probability layer and the LR probability layer indicated that both models predicted the occurrence of wetlands relatively similarly[51] (Figure 6). The ROC analysis determined that the criterion model probability threshold for both models (at which probability both the sensitivity (71.6) and specificity (81.2) are the highest as a pair) was greater than 0.60; i.e. if the probability values in both layers were split into binary classes of wetland and non-wetland areas, then 0.60 would be the ideal split to maintain good predictability of wetland and non-wetland occurrence.

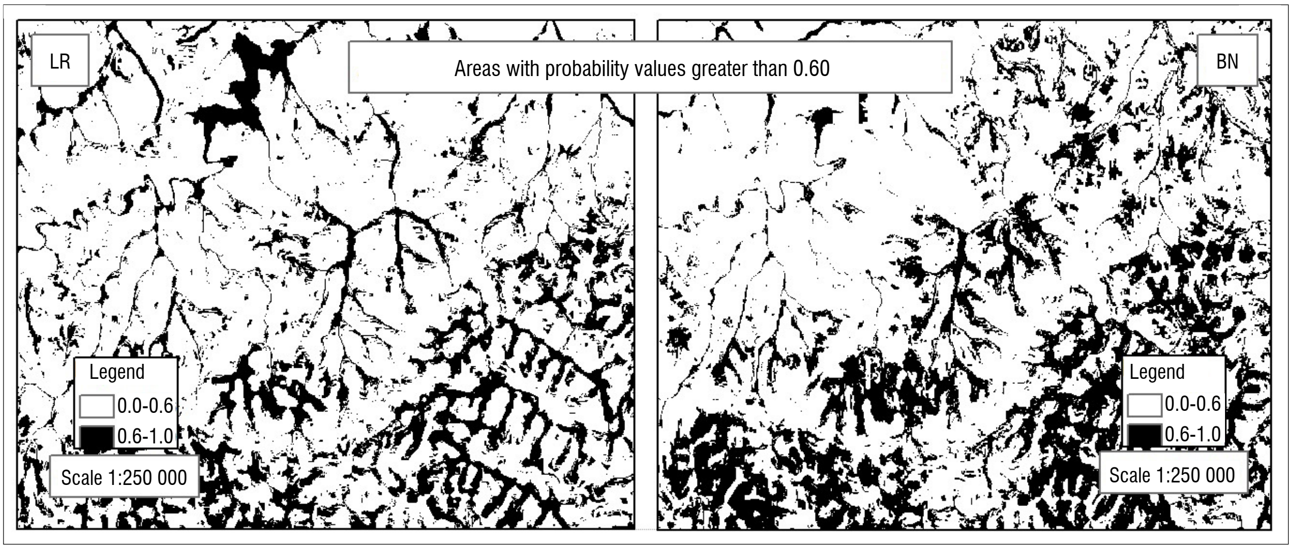


Figure 5: Comparison of likelihood of wetland occurrences for probability > 0.6 for logistic regression (LR) and Bayesian network (BN) models.

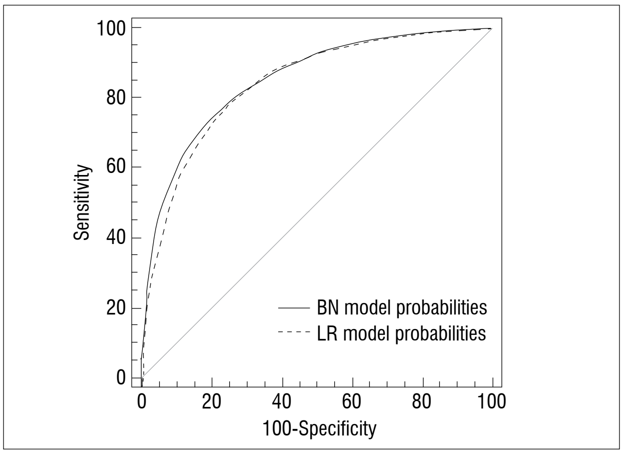


Figure 6: Receiver operating characteristic curves comparing the prediction accuracy of the Bayesian network (BN) model with the logistic regression (LR) predictive model.

The trend analysis results indicated that there was a decreasing trend in accuracy of mean wetland extent area covered with an increase in probability of wetland occurrence (Figure 7). However, with the increase in probability, there was an increasing trend in accuracy of correctly predicting the presence of a wetland. The probable explanation for this trend is that the average wetland extent is made up of a mosaic of probability values, with the core being predicted by high probability values and the outer extents by lower probability values. Therefore lower probability ranges will occupy larger area extents but with a lower accuracy in identifying a single wetland area, whereas higher probability values will more likely identify wetland location but at the cost of accurately identifying the wetland's extent (Table 4).

Discussion

Current wetland mapping approaches are typically based either on aerial photography and/or satellite imagery interpretation or classification of satellite imagery, the latter involving complex methodologies utilising spectral ratios, indexes and values which are classified to identify wetland areas. It is common for wetland maps to under-represent certain areas for various reasons (errors, lack of resources, misclassification), and therefore these layers could help minimise these areas of under-representation. Consequently, resultant maps may suffer from pitfalls that include not linking wetland polygons that are fragments of larger wetland systems, and wetland omissions because of seasonal effects on satellite images. To compensate for these pitfalls, many methodologies draw support from ancillary data to improve the accuracy of wetlands mapped and classified.[14,18] Topological entities are commonly used as ancillary data and have been valuable in the success of many other wetland mapping approaches.[5,14,18,52-55]

In this study, we assessed two methods that use topological and climatic variables as the basis for predicting the probability of wetland occurrence over a large spatial domain. The final wetland probability layers had good agreement with the current regional wetland layer and literature highlighting regional priority wetland areas.[26,34] Output layers indicated new areas of wetlands and how wetland fragments are likely to be part of larger but fragmented wetland systems. We acknowledge that using the PCA approach resulted in variables selected for the correlative rather than causative link to wetland presence. This was a necessary trade-off that ignores the complexities and subtleties of topographic influence on wetland presence, but allows for greater objectivity and repeatability in model development; both methods do not require an understanding of the complex interactions and relationships of the environmental components that drive wetland formation and function. It is also to be expected that the models will predict different hydrogeomorphic wetland types with different levels of accuracy because of the different ways in which such environmental drivers interact, and at different scales, within the landscape. Moreover, the variables relevant to this model may not be entirely transferrable across different regions of South Africa because of possible differences in wetland ontology (geology, climate) between regions. A clear identified research need is therefore to establish suitable predictor variable sets across different climatic regions of South Africa, as the basis for developing regional probabilistic models.

There are advantages and limitations inherent in each approach. Uusitalo[46] reports that a BN can only deal with continuous data in a limited manner, and that there is no satisfactory automatic discretisation technique or method for data translation to qualitative states in BNs. The Jenks natural break interval method[45] was used to discretise the data into qualitative states in this study, but it is important that future researchers adopting this approach should pay careful attention when defining value ranges of continuous data to ensure that the intervals signify important breaks within data, so that the generalisation caused by discretisation is minimised. Conversely, the LR approach did not require the data to be discretised, and this method provided simpler and more integrated handling of continuous data than the BN approach. Given such trade-offs, perhaps the most promising approach would be to use 'ensemble' or consensus modelling, in which the outputs of both models are combined such that the probabilities of occurrence of both algorithms are used to provide a combined output with a lower mean error.[56]

The probability layers from either method have the potential to not only identify new wetland areas, but to guide the classification of satellite imagery by showing highly probable wetland areas, thereby avoiding the misclassification of pixels or the high errors associated with spectral confusion. However, we note that there will inevitably be trade-offs between the accuracy in the model's prediction of wetland extent and wetland occurrence. While the cut-off point for creating a useable wetland map is dependent on the user, we recommend 0.6 as a threshold for mapping probability only, but 0.8 when mapping extent and probability. For example, if the user is interested in using the model to identify new wetland areas, and is not concerned with the model's ability to predict wetland extent, the user would opt for a higher cut-off probability value, which produces a wetland map with a higher accuracy in predicting the wetland occurrence and a low accuracy in modelling the correct wetland extent. Using such ancillary data to support wetland mapping efforts has the potential not only to improve general land-cover assessments but also to better establish important spatial priorities for wetland conservation and management through improved conservation target estimation when measuring the current status of wetlands in an area in terms of wetland loss and current state of integrity. The final output has further applicability in already modified areas because the final model output predicts the likelihood of wetland occurrence regardless of any land-cover transformation. Predicted occurrence of

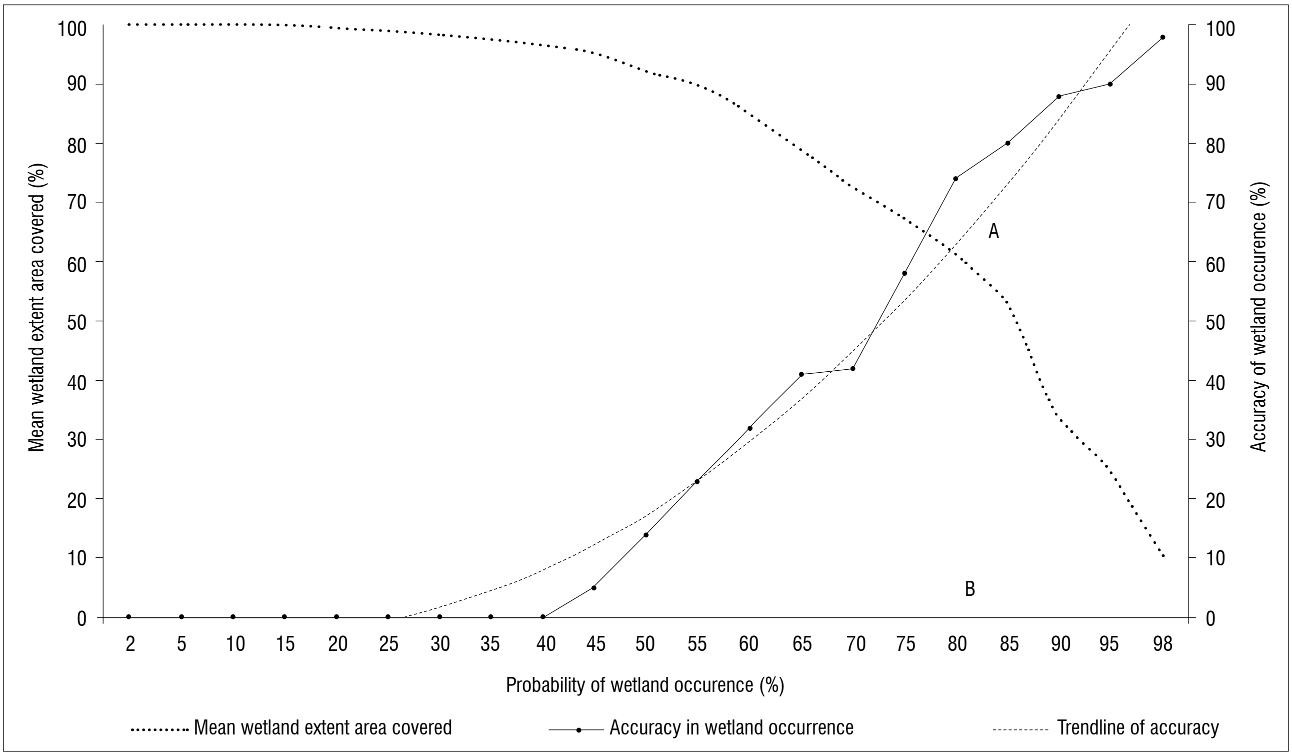


Figure 7: Accuracy in terms of model ability to predict wetland extent and occurrence with the increase in probability percentage. An optimal trade-off value exists at the intersection of extent (A) and accuracy (B) to produce a useable wetland map output.

wetlands without the effects of land transformation has implications both in establishing the historical extent of wetlands in regions of extensive land-cover transformation, as well as in establishing to what extent seemingly unrelated wetland polygons are in fact components of single larger systems now fragmented. We conclude that the methods assessed in this study have the potential to generate useful ancillary data to improve wetland mapping accuracy by identifying new wetland areas and providing insights on linkages between wetland fragments, but we recommend further ground truthing to assess such layers. From a pragmatic and computational perspective, our preference would be to use the LR approach as the basis for developing regional wetland probability maps for additional regions in South Africa.

Acknowledgements

We thank the Mondi Wetland Programme for funding this research, and Dr Boyd Escott (Ezemvelo KZN Wildlife) and the Agricultural Research Council: Institute for Soil, Climate and Water for providing spatial data relating to wetlands in KwaZulu-Natal. We sincerely thank the two reviewers for their comments and inputs into the manuscript.

Authors' contributions

N.R.M. was the project supervisor and conceptualised the study. J.H. performed the majority of the data analyses and all of the GIS analyses as part of his MSc study at UKZN. N.R.M. and J.H. wrote the manuscript.

References

1. Mitsch WJ, Gosselink JG. Wetlands. 3rd ed. New York: John Wiley & Sons Inc.; 2000.
2. Woodhouse S, Lovett A, Dolman P, Fuller R. Using GIS to select priority areas for conservation. Comput Environ Urban Sys. 2000;24:79–93. http://dx.doi.org/10.1016/S0198-9715(99)00046-0
3. May D, Wang J, Kovacs J, Mutter M. Mapping wetland extent using IKONOS satellite imagery of the O'donell point region, Georgian Bay, Ontario. London, Ontario: University of Western Ontario; 2002.
4. Chhokar KB, Pandya M, Raghunathan M. Understanding environment. New Delhi: Sage Publications India Pvt; 2004.
5. Islam MA, Thenkabail PS, Kulawardana RW, Alankara R, Gunasinghe S, Edussriya C, et al. Semi-automated methods for mapping wetlands using Landsat ETM+ and SRTM data. Int J Remote Sens. 2008;29:7077–7106. http://dx.doi.org/10.1080/01431160802235878
6. Millennium Ecosystem Assessment (MEA). Ecosystem and human well-being: Wetlands and water synthesis. Washington DC: World Resources Institute; 2005.
7. Ramsar Convention Secretariat. Ramsar handbooks for the wise use of wetlands. 4th ed. Gland: Ramsar Convention Secretariat; 2010.
8. Olhan E, Gün S, Ataseven Y, Arisoy H. Effects of agricultural activities in Seyfe wetland. Sci Res Essays. 2010;5:9–14.
9. Ozesmi SL, Bauer ME. Satellite remote sensing of wetlands. Wetl Ecol Manag. 2002;10:381–402. http://dx.doi.org/10.1023/A:1020908432489
10. Adam E, Mutanga O, Rugege D. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: A review. Wetl Ecol Manag. 2010;18:281–296. http://dx.doi.org/10.1007/s11273-009-9169-z
11. Wilen BO, Carter V, Jones JR. Wetland management and research: Wetland mapping and inventory. National Water Summary on Wetland Resources, US Geological Survey Water Supply Paper 2425 [homepage on the Internet]. c2002 [cited 2015 June 17]. Available from: https://water.usgs.gov/nwsum/WSP2425/mapping.html
12. Batzer, DP, Sharitz, RR, editors. Ecology of freshwater and estuarine wetlands. Berkeley, CA: University of California Press; 2006.
13. Yu H, Zhang S. Application of high resolution satellite imagery for wetlands cover classification using object-oriented method. Int Arch Photogramm Remote Sens. 2008;XXXVII(B7):521–526.
14. Kulawardhana RW, Thenkabail PS, Vithanage J, Biradar C, Islam MA, Gunasinghe S, et al. Evaluation of the wetland mapping methods using Landsat ETM+ and SRTM data. J Spat Hydrol. 2007;7:62–96.
15. Rebelo LM, Finlayson CM, Nagabhatla N. Remote sensing and GIS for wetland inventory, mapping and change analysis. J Environ Manage. 2009;90:2144–2153. http://dx.doi.org/10.1016/j.jenvman.2007.06.027
16. Knight AW, Tindall DR, Wilson BA. A multitemporal multiple density slice method for wetland mapping across the state of Queensland, Australia. Int J Remote Sens. 2009;30:3365–3392. http://dx.doi.org/10.1080/01431160802562180
17. Landmann T, Schramm M, Colditz RR, Dietz A, Dech S. Wide area wetland mapping in semi-arid Africa using 250-meter MODIS metrics and topographic variables. Remote Sens. 2010;2:1751–1766. http://dx.doi.org/10.3390/rs2071751
18. Li J, Chen W. A rule-based method for mapping Canada's wetlands using optical, radar and DEM data. Int J Remote Sens. 2005;26:5051–5069. http://dx.doi.org/10.1080/01431160500166516
19. Lunetta RS, Balogh ME, Merchant JW. Application of multi-temporal Landsat 5 TM imagery for wetland identification. Photogramm Eng Remote Sens. 1999;65:1303–1310.
20. Ryo M, Peng G, Bing X. Spectral mixture analysis for bi-sensor wetland mapping using Landsat TM and Terra MODIS data. Int J Remote Sens. 2012;30:3373–3401.
21. Ricchetti E. Multispectral satellite image and ancillary data integration for geological classification. Photogramm Eng Remote Sens. 2000;66:429–435.
22. Allen, KM, Green SW, Zubrow EB. Interpreting space: GIS and archaeology. London: Taylor and Francis; 1990.
23. Eeley HAC, Lawes MJ, Piper SE. The influence of climate change on the distribution of indigenous forest in KwaZulu-Natal, South Africa. J Biogeogr. 1999;26:595–617. http://dx.doi.org/10.1046/j.1365-2699.1999.00307.x
24. King L. A geomorphology of central and southern Africa. Biogeography and ecology of southern Africa. Monogr Biol. 1978;31:1–17. http://dx.doi.org/10.1007/978-94-009-9951-0_1
25. Rivers-Moore NA, Cowden C. Regional prediction of wetland degradation in South Africa. Wetl Ecol Manag. 2012;20:1–14. http://dx.doi.org/10.1007/s11273-012-9271-5
26. Patrick MJ, Ellery WN. Plant community and landscape patterns of a floodplain wetland in Maputaland, Northern KwaZulu-Natal, South Africa. Afr J Ecol. 2006;45:175–183. http://dx.doi.org/10.1111/j.1365-2028.2006.00694.x
27. KwaZulu-Natal Provincial Planning Commission (KZNPPC). Provincial growth and development strategy. Pietermaritzburg: KZNPPC; 2011.
28. Schulze RE. South African atlas of agrohydrology and climatology. Report TT82/96. Pretoria: Water Research Commission; 1997.
29. Colvin C, Le Maitre D, Saayman I, Hughes S. Introduction to aquifer dependent ecosystems in South Africa. Pretoria: Natural Resources and the Environment, CSIR; 2007.
30. Escott B. Landform map for KZN based on the 90m SRTM DEM (v4 edited). Pietermaritzburg: Ezemvelo KZN Wildlife; 2011.
31. Van den Berg HM, Weepener HL, Metz M. Spatial modeling for semi-detailed soil mapping in KwaZulu-Natal. Report No: GW/A/2009/48. Pretoria: Agricultural Research Council – Institute for Soil, Climate and Water; 2009.
32. GISCOE 20m GISCOE DTM Data. Pretoria: GISCOE Pty Ltd; 2001.
33. Scott-Shaw CR, Escott BJ. KwaZulu-Natal provincial pre-transformation vegetation type map 2011. Pietermaritzburg: Biodiversity Conservation Planning Division, Ezemvelo KZN Wildlife; 2011 [unpublished].
34. Begg GW. The wetlands of Natal part 3: The location, status and function of the priority wetlands of Natal, report 73. Pietermaritzburg: Natal Town and Regional Planning Commission; 1989.
35. Beyer HL. Hawth's analysis tools for ArcGIS [homepage on the Internet]. c2004 [cited 2011 May 22]. Available from: http://www.spatialecology.com/htools/
36. Aguilera PA, Fernandez A, Fernandez R, Rumi R, Salmeron A. Bayesian networks in environmental modelling. Environ Model Softw. 2011;26:1376–1388. http://dx.doi.org/10.1016/j.envsoft.2011.06.004
37. Jewitt D. Landcover accuracy assessment photography 2011. Pietermaritzburg: Biodiversity Research and Assessment Division, Ezemvelo KZN Wildlife; 2011 [unpublished].

38. Environmental Systems Research Institute (ESRI). ArcGIS desktop release 9.3. Redlands, CA: ESRI; 2007.

39. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2009. Available from: http://www.R-project.org

40. Kovach WL. MVSP: A multivariate statistical package for Windows version 3.2. Pentraeth, Wales: Kovach Computing Services; 1999.

41. Netica version 4.10. Vancouver: Norsys; 2010. Available from: www.norsys.com

42. MedCalc for Windows version 12.5. Ostend, Belgium: MedCalc Software; 2013. Available from: www.medcalc.org

43. Fry JC. Biological data analysis. New York: Oxford University Press; 1993.

44. Cain J. Planning improvements in natural resources management: Guidelines for using Bayesian networks to support the planning and management of development programmes in the water sector and beyond. Wallingford, UK: Centre for Ecology & Hydrology; 2001.

45. Jenks GF, Caspall FC. Error on chloroplethic maps: Definition, measurement, reduction. Ann Amer Geogr. 1971;61:217–244. http://dx.doi.org/10.1111/j.1467-8306.1971.tb00779.x

46. Uusitalo L. Advantages and challenges of Bayesian networks in environmental modelling. Ecol Model. 2007;203:312–318. http://dx.doi.org/10.1016/j.ecolmodel.2006.11.033

47. Grêt-Regmey A, Straub D. Spatially explicit avalanche risk assessment linking Bayesian networks to GIS. Nat Hazards Earth Syst Sci. 2006;6:911–926. http://dx.doi.org/10.5194/nhess-6-911-2006

48. Crawley MJ. The R book. Chichester: John Wiley & Sons Ltd; 2007.

49. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27:861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010

50. Carletta J. Assessing agreement on classification tasks: The kappa statistic. Comput Ling. 1996;22:249–254.

51. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: Wiley; 2002. http://dx.doi.org/10.1002/9780470317082

52. Ibrahim K. Assessment of wetlands in Kuala Terengganu District using Landsat TM. J Geogr Geol. 2009;1:33–40. http://dx.doi.org/10.5539/jgg.v1n2p33

53. Bwangoy JB, Hansen MC, Roy DP, Grandi G, Justice CO. Wetland mapping in the Congo Basin using optical and radar remotely sensed data and topographical indices. Remote Sens Environ. 2010;114:73–86. http://dx.doi.org/10.1016/j.rse.2009.08.004

54. Pantaleoni E, Wynne RH, Galbraith JM, Campbell JB. A logit model for predicting wetland location using RASTER and GIS. Int J Remote Sens. 2009;30:2215–2236. http://dx.doi.org/10.1080/01431160802549310

55. Wright C, Gallant A. Improved wetland remote sensing in Yellowstone National Park using classification trees to combine TM imagery and ancillary environmental data. Remote Sens Environ. 2007;107:582–605. http://dx.doi.org/10.1016/j.rse.2006.10.019

56. Araújo MB, New M. Ensemble forecasting of species distributions. Trends Ecol Evol. 2007;22:42–47. http://dx.doi.org/10.1016/j.tree.2006.09.010

Evidence for climate-induced range shift in *Brachystegia* (miombo) woodland

AUTHORS:
Brenden Pienaar[1]
Dave I. Thompson[2,3]
Barend F.N. Erasmus[1*]
Trevor R. Hill[4]
Ed T.F. Witkowski[1]

AFFILIATIONS:
[1]School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2]South African Environmental Observation Network, Phalaborwa, South Africa
[3]School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa
[4]School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa
*Current address: Global Change and Sustainability Research Institute, University of the Witwatersrand, Johannesburg, South Africa

CORRESPONDENCE TO:
Ed Witkowski

EMAIL:
Ed.witkowski@wits.ac.za

POSTAL ADDRESS:
School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Private Bag 3, Wits 2050, South Africa

DATES:
Received: 13 Aug. 2014
Revised: 06 Oct. 2014
Accepted: 16 Oct. 2014

KEYWORDS:
Brachystegia spiciformis; climate change; ecological niche model; MaxEnt; refugia

HOW TO CITE:
Pienaar B, Thompson DI, Erasmus BFN, Hill TR, Witkowski ETF. Evidence for climate-induced range shift in *Brachystegia* (miombo) woodland S Afr J Sci. 2015;111(7/8), Art. #2014-0280, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140280

Brachystegia spiciformis Benth. is the dominant component of miombo, the sub-tropical woodlands which cover 2.7 million km^2 of south-central Africa and which is coincident with the largest regional centre of endemism in Africa. However, pollen records from the genus *Brachystegia* suggest that miombo has experienced rapid range retraction (∼450 km) from its southernmost distributional limit over the past 6000 years. This abrupt biological response created an isolated (by ∼200 km) and incomparable relict at the trailing population edge in northeast South Africa. These changes in miombo population dynamics may have been triggered by minor natural shifts in temperature and moisture regimes. If so, *B. spiciformis* is likely to be especially responsive to present and future anthropogenic climate change. This rare situation offers a unique opportunity to investigate climatic determinants of range shift at the trailing edge of a savanna species. A niche modelling approach was used to produce present-day and select future *B. spiciformis* woodland ecological niche models. In keeping with recent historical range shifts, further ecological niche retraction of between 30.6% and 47.3% of the continuous miombo woodland in Zimbabwe and southern Mozambique is predicted by 2050. Persistence of the existing relict under future climate change is plausible, but range expansion to fragmented refugia in northeast South Africa is unlikely. As *Brachystegia* woodland and associated biota form crucial socio-economic and biodiversity components of savannas in southern Africa, their predicted further range retraction is of concern.

Introduction

At regional and global scales, climate broadly limits the distribution of plant taxa,[1,2] and the response of species to changing environments is consequently likely to be largely determined by population responses at range margins.[3] However, few studies have examined climatic determinants of range shift at the trailing margin of a savanna species in the southern hemisphere. Understanding the historical and present-day spatial dynamics of vegetation plays a crucial role in our ability to predict the likely community responses and biodiversity consequences of future global change.[4]

Brachystegia spiciformis Benth. is the dominant component of miombo, the colloquial term used to describe sub-tropical woodlands dominated by *Brachystegia, Julbernardia* and *Isoberlinia* – three closely related genera in the family Fabaceae and subfamily Caesalpinioideae.[5] Miombo encompasses the woodland-dominated savanna ecosystems[6,7] which cover (Figure 1) an estimated 2.7 million km^2 of south-central Africa[8] and is coincident with White's[9] Zambezian Phytochorion, the largest regional centre of endemism in Africa[10]. The dynamics of miombo woodland are largely determined by the woody component which, apart from climate, is predominantly influenced by people and fire. An estimated 75 million people inhabit areas covered by, or formerly covered by, miombo woodland; an additional 25 million urban dwellers rely on miombo wood or charcoal as a source of energy.[11] Much of the woodland has been, and continues to be, modified by people. Changes in vegetation structure occur both directly as a result of woodland cover removal, and indirectly from changes in fire regime brought about by higher grass production under a more open tree canopy.[12] Most of the dominant miombo canopy species, including *B. spiciformis*, are considered to be fire-tender species, which decline in abundance under regular burning and increase under complete fire protection.[5]

Mean annual precipitation and mean annual temperature throughout the miombo region range from 650 mm to 1400 mm and 15 °C to 25 °C, respectively. The majority (>95%) of precipitation falls during the summer season which prevails from October to March.[5] However, sporadic dry periods during the onset of the precipitation season may cause large fluctuations in soil moisture and temperature. The resulting water stress during this concurrent and limited germination phase accounts for high seedling mortality in miombo woodlands.[13] Once established, the influence of climate, and of precipitation in particular, on *B. spiciformis* growth performance is strongest during the core of the precipitation season (December to February).[14]

Brachystegia are likely under-represented in pollen spectra because of the relatively low pollen production typical of entomophily.[5] Nevertheless, sediment cores containing pollen from the genus, dated to 38 000 years BP, have been recovered from the south-central African plateau.[5,15] Evidence suggests that the central plateau vegetation has undergone significant flux over this period, ranging from cool upland grassland during the Last Glacial Maximum (LGM, 19 000 years BP), to *Brachystegia*-dominated savanna during the warm interglacial period of the Mid-Holocene (6000 years BP).[5,16] Although literature regarding climate-induced range shifts at the genus level is limited for the subcontinent, Eeley et al.[17] concluded that the distribution of the forest biome in KwaZulu-Natal, South Africa, receded during the colder and drier conditions of the LGM, while warmer and wetter conditions during the Mid-Holocene were conducive to forest expansion. Correspondingly, *Brachystegia* woodland likely expanded to occupy a widespread historical range across south-central Africa prior to the LGM and during the warmer and wetter conditions of the Mid-Holocene.[15,18,19] Sediment cores containing pollen from the genus *Brachystegia* have been recovered from Pretoria, Mookgophong and Tate Vondo (Figure 2) in South Africa,[15] the dates of which stand as evidence of a much more widespread distribution during the recent past (~6000 to less than 1000 years BP). These records support the presence of the genus at least up to 450 km southwest of the present-day distribution limit. This most recent and abrupt range retraction suggests either a sudden change in climate, for which there is no evidence,[19] or that minor shifts in temperature and moisture regimes have triggered marked changes in *Brachystegia* population dynamics.[5]

A change in geographical distribution from the simultaneous migration of populations throughout their range is unlikely. Instead, change is generated by the establishment of new, often discontinuous, populations at the leading edge of a species distribution and the coincident death of individuals and extirpation of populations at the trailing edge.[20] These retractions are often not complete, but instead leave behind fragmented populations that persist as relicts, in isolated enclaves of favourable environmental conditions within an inhospitable regional climate.[21] Globally, numerous relict populations have resulted from species range shifts experienced after the LGM.[21] For example, the present-day distribution of Neotropical seasonal dry forest formations is considered to comprise fragmentary remnants of the once extensive forests that characterised the dry climatic maxima of the Pleistocene.[22]

Rutherford et al.[23] suggested that *B. spiciformis,* which at the time was known only from north of the South African border, could find a suitable ecological niche under future global climate change conditions along the high rainfall savanna–grassland biome interface of northeast South Africa. However, they concluded that range expansion through long-distance dispersal of the species into South Africa, across the Limpopo River Valley, was unlikely to occur unassisted given that seed dispersal distance is limited to less than 6 m,[13] and that regeneration takes place predominantly through coppice regrowth and root suckers, rather than seed.[24] Startlingly, the discovery of the isolated *B. spiciformis* woodland (~15 ha, Figure 2) in the eastern Soutpansberg of South Africa[25] placed the species within the region predicted by Rutherford et al.[23] This isolated population, occurring some 200 km south of the continuous miombo woodlands of the subcontinent, suggests a trailing edge refugium persisting from a previously wider historical range dating prior to the LGM or during the Mid-Holocene. This woodland relict (known locally as Gundani) shares characteristics of the extensive *B. spiciformis* woodlands of Zimbabwe and southern Mozambique, including the open structure, medium canopy height, poorly developed lower strata and high dependence on root suckering and coppicing for regeneration.[26] The presence of a second *Brachystegia* species within the Gundani woodland, limited to a single *B. utilis* individual, further suggests that this is likely a climate relict of a vegetation type from a time when, at least, the genus *Brachystegia* dominated the Soutpansberg massif in northeastern South Africa.[27]

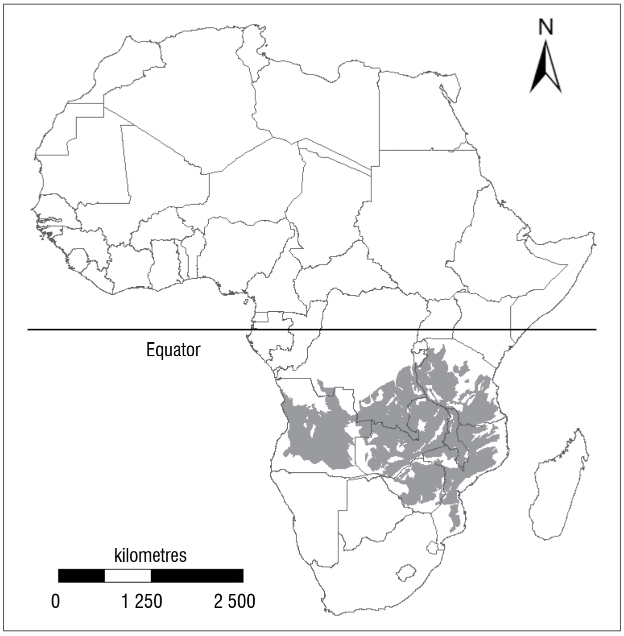
Trailing range edge studies reflect a bias (86%) towards high-latitude range margins and temperate vegetation communities,[3] which have minimum temperature-related constraints. Conversely, this rare situation provides a unique opportunity to explore climatic constraint at low-latitude range margins for an ecologically significant savanna species.

In this climate specific study, we used a predictive modelling approach to determine (1) the likely geographical footprint of the *B. spiciformis* woodland ecological niche in southern Africa under present-day and selected future (2050) global climate change scenarios, (2) the specific ecological niche dimensions of *B. spiciformis* woodland in southern Africa and (3) whether bioclimatic variables at palaeohistorical distribution sites for the genus could have supported *B. spiciformis*, the dominant component of present-day miombo woodland in southern Africa, during the Mid-Holocene.

Methods

Species data

A total of 914 mature *B. spiciformis* specimens were considered (717 in-situ observations and 197 from regional and online herbaria). Their locations were subsequently converted to quarter-degree square (QDS) centroid points to accommodate numerous herbaria records. After duplicates were excluded, a total of 76 QDS centroid points (n=76) remained. QDS is a commonly used format for general species distribution mapping in South Africa.[28] These species data include the vicariant *B. spiciformis* population from South Africa, as the inclusion of relict populations has been shown to improve the performance of model-based projections.[20]



Source: Adapted from White's[9] map of African vegetation.

Figure 1: Distribution of miombo woodland in Africa.

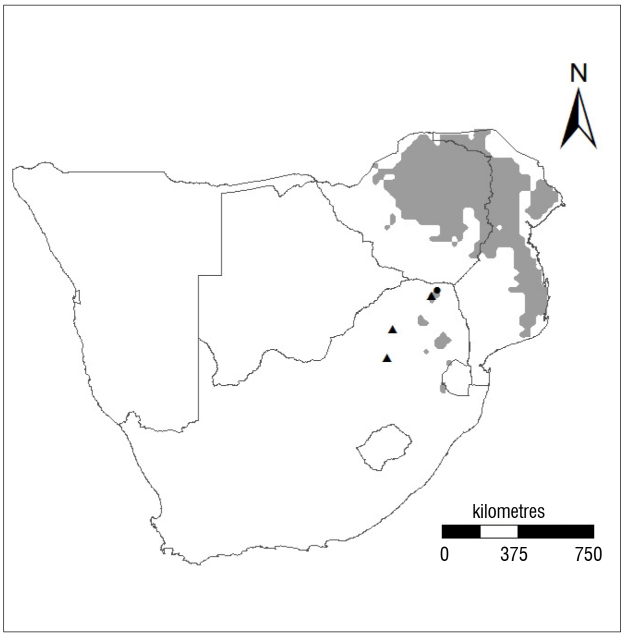


Figure 2: Ecological niche of *Brachystegia spiciformis* woodlands in southern Africa under present-day climate conditions. Location of the Gundani climate relict population (black circle) and three Mid-Holocene pollen records (black triangles) are included.

Predictor modelling

MaxEnt v.3.3.3k[29] was used to project suitable ecological niche models (ENMs) for *B. spiciformis* woodland in southern Africa as it has been found to perform best among many different modelling approaches[30]. Functionality relies on a list of presence-only locations and a set of environmental predictors across a user-defined landscape, which is divided into grid cells, to generate species presence probability.

Predictor variables

Nineteen regional bioclimatic variables and altitude ('BioClim') grid data were downloaded from the 'WorldClim'[31] database. The bioclimatic variables are derived from monthly temperature and rainfall recorded worldwide (period 1950–2000) and are often used in ecological niche modelling (for examples see Blach-Overgaard et al.[32] and Penman et al.[33]). A soil variable (dominant soil group) was added from the Harmonised World Soil Database.[34] ENM Tools v.1.3[35] was used to produce pairwise Pearson correlation coefficients to eliminate spatially correlated variables ($r > 0.85$) from the modelling process, with preference given to variables with higher influence on species presence probability. Summer precipitation variables were also prioritised during the elimination process as Trouet et al.[14] suggest they have a strong influence on *B. spiciformis* growth performance. The 14 variables used in all models were Bio2–6, Bio8–9, Bio15–19, altitude and soil (see the Appendix for a description of the variables).

There are currently at least 24 coupled atmosphere–ocean general circulation models (AOGCMs) used to project climatic changes for more than 10 different greenhouse-gas emission scenarios.[36] We selected three commonly used AOGCMs[37,38] – HadCM3, CGCM2 and CSIRO-MK2 – to forecast the impact of climate change on *B. spiciformis* distribution. The A2a emission scenario, an intermediate scenario representing regional development and slow economic growth, was used across all three AOGCMs.

Model settings

The spatial resolution of all variables and landscape grid cells was resampled to QDS resolution. MaxEnt iterations were set to 5000 and accuracy was evaluated by constructing the model using 75% of presence records as training points, with the remaining 25% used in validation. The accuracy of the present-day ecological niche model was inferred from the area under the receiving operating characteristic curve (AUC) score that varies from 0 to 1, with 0 being the lowest and 1 being the highest probability of matching the species distribution.[31] Subsequently, the present-day ecological niche model was applied to the three selected future (2050) AOGCMs. The increase in AUC score was used as a test metric to determine the most important bioclimatic variables explaining *B. spiciformis* woodland distribution, when each variable was used in isolation.

Ecological niche comparison

The fundamental niche[39] was represented by a box plot, with the interquartile range (box with median present) representing 50% of the total data values and the upper and lower whiskers representing 25% of the values, respectively. The interquartile space is interpreted as the 'most suitable' range of climatic or environmental conditions for the species, while the lower and upper whiskers span less suitable conditions, reflecting lower and upper tolerance limits, respectively. Extreme values, beyond the 1.5 coefficient value, were plotted as individual open circles.

Geospatial Modelling Environment v.0.7.2.1[40] was used to intersect *B. spiciformis* distribution records ($n = 76$*)* from selected (highest AUC scores and limiting factors) bioclimatic layers (Bio4, 5, 6, 15 and 16). The grid value at each point was used to create box plots representing the present-day ecological niche (Figure 3).

Seasonal data (precipitation and temperature) from Mid-Holocene ocean–atmosphere–vegetation models were used to intersect three palaeohistorical *Brachystegia* distribution points using Geospatial Modelling Environment. These grid values were subsequently used to construct Bio16 (precipitation of the wettest quarter) and Bio4 (temperature seasonality, the standard deviation of mean monthly temperature in degrees Celsius) as per BioClim. These two bioclimatic variables were selected as they achieved the highest increase in AUC score. This methodology was repeated across all available Mid-Holocene ocean–atmosphere–vegetation models (ECBILTCLIOVECODE, ECHAM53-MPIOM127-LPJ, FOAM, MRI-CGCM2.3.4fa, MRI-CGCM2.3.4nfa and UBRIS-HadCM3M2) from the Palaeoclimate Modelling Intercomparison Project Phase 2 database[41] (http://www-lscedods.cea.fr/pmip2_dbext/pmip2_6k_oav/atm/se/). Mid-Holocene Bio16 and Bio4 values, averaged across all six models, were evaluated against the present-day ecological niche.

Limiting factors map

A limiting factor map (Figure 4) was produced as per Elith et al.[30] Using Geospatial Modelling Environment and implementing the method described above, box plots were created from randomly selected QDS centroid points ($n = 76$*)* within the area immediately adjacent to the current distribution of the species, for which a specific bioclimatic variable was shown as limiting. Box plots and paired sample *t*-tests were used to compare *B. spiciformis* ecological niche and limiting factor dimensions per selected bioclimatic variable (Bio4, 5, 6, and 16). Comparison of means was done using R v.3.0.3.[42]

Results

Model performance and variable contribution

The present-day ecological niche model for *B. spiciformis* achieved an AUC value of 0.923, which is considered very good.[43] The model was subsequently applied to three future AOGCMs. Precipitation of the wettest quarter (Bio16) and temperature seasonality (Bio4) were identified as the two most important bioclimatic variables explaining *B. spiciformis* woodland distribution in southern Africa (Table 1). This result was consistent across the present-day and three future models.

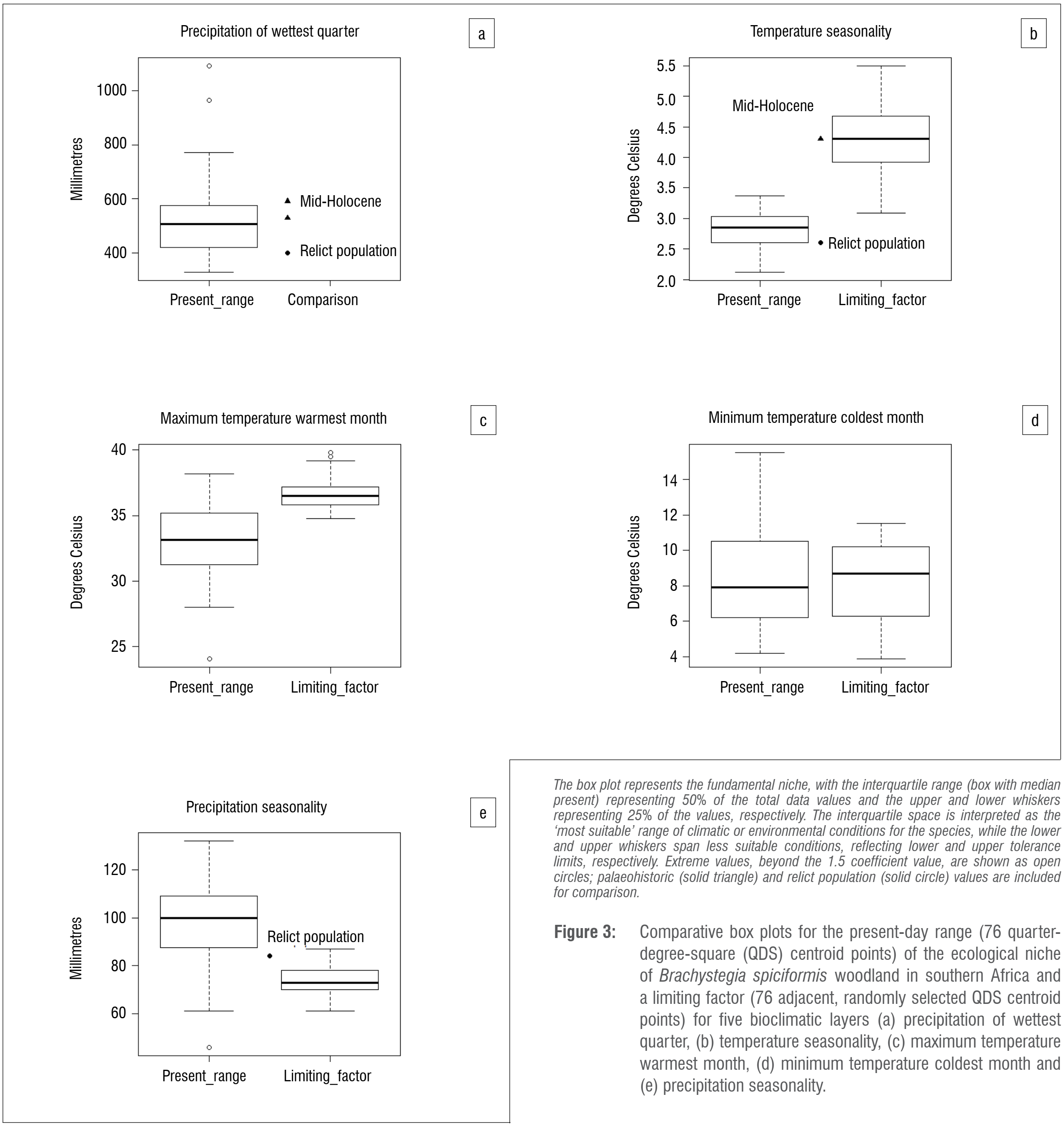
Present-day ecological niche

The ENM (Figure 2) predicts the presence of isolated *B. spiciformis* populations in South Africa, one of which is coincident with the location of the Gundani relict discovered in 2001. The model further indicates suitable ecological conditions for the species elsewhere at disjunct locations along the high rainfall savanna–grassland interface of northeastern South Africa. However, the species does not presently occur outside of the relict population.[44] Besides the fact that White's map of African vegetation[9] does not indicate miombo woodland presence in South Africa, the distribution is coincident with the present-day ENM projection for *B. spiciformis* in southern Africa.

Precipitation of the wettest quarter (Figure 3) suggests a relatively narrow range of optimal suitability of 422–576 mm, with a median of 507 mm. Similarly, the 'most suitable' range for temperature seasonality (Figure 3) is constrained between 2.6 °C and 3.0 °C, with a median of 2.8 °C. The value of the former variable at the Gundani relict population (401 mm) falls within the lower tolerance limit of conditions experienced by the species, whereas the value for the latter variable at the relict population (2.6 °C) falls within the 'most suitable' range for *B. spiciformis* in southern Africa.

Past ecological niche

A crude palaeohistoric distribution of *Brachystegia*, as inferred from three Mid-Holocene sediment core pollen records from South Africa (Figure 2), supports past average climate values of 529 mm and 591 mm (two records) for precipitation of the wettest quarter, and temperature seasonality averages of 4.3 °C and 4.8 °C (two records) for that period. For precipitation (Figure 3), these averages fall within (or nearly within) the 'most suitable' range identified under present-day conditions (422–576 mm). Alternatively, the historical temperature seasonality values exceed even the upper tolerance limit (a maximum of 3.4 °C) for this species across its present-day southern African distribution (Figure 3).



The box plot represents the fundamental niche, with the interquartile range (box with median present) representing 50% of the total data values and the upper and lower whiskers representing 25% of the values, respectively. The interquartile space is interpreted as the 'most suitable' range of climatic or environmental conditions for the species, while the lower and upper whiskers span less suitable conditions, reflecting lower and upper tolerance limits, respectively. Extreme values, beyond the 1.5 coefficient value, are shown as open circles; palaeohistoric (solid triangle) and relict population (solid circle) values are included for comparison.

Figure 3: Comparative box plots for the present-day range (76 quarter-degree-square (QDS) centroid points) of the ecological niche of *Brachystegia spiciformis* woodland in southern Africa and a limiting factor (76 adjacent, randomly selected QDS centroid points) for five bioclimatic layers (a) precipitation of wettest quarter, (b) temperature seasonality, (c) maximum temperature warmest month, (d) minimum temperature coldest month and (e) precipitation seasonality.

Future ecological niche

The future ecological niches for all three models (Figure 5) are consistent with a predicted decrease in the continuous *B. spiciformis* woodland distribution of Zimbabwe and southern Mozambique (30.6% to 47.3%) and a decrease in the overall ecological niche (including South Africa) of between 16.9% and 32.2% by 2050 (Table 2). These figures reflect net change, with a decreased niche potentially resulting from larger retraction at present-day distributions combined with comparatively smaller range increases at newly favourable distributions elsewhere.

Two models (HadCM3 and CGCM2) suggest extensive southward ecological niche expansion in South Africa by approximately 200% and 400%, respectively, at the savanna–grassland biome interface, into areas south (29°S) of the present-day actual (22.5°S) and predicted (~27°S) ranges. Models HadCM3 and CSIRO-MK2 indicate that a suitable ecological niche at the location of the South African *B. spiciformis* woodland relict will not persist. Under no climatic scenario does the predicted future distribution mirror the southwestern range historically occupied by the species during the Mid-Holocene.

Present-day limiting factors

Climate values falling outside of the present-day fundamental niche or suitability range of *B. spiciformis* woodland in southern Africa place a climatic constraint on the species, and can therefore be considered as limiting factors to its geographical footprint. Three bioclimatic variables were identified as potential limiting factors for *B. spiciformis*.

Temperature seasonality (Bio4) and maximum temperature of the warmest month (Bio5) are suggested to be responsible for restricting distribution of the continuous *B. spiciformis* woodlands of Zimbabwe and southern Mozambique (Figure 4). The temperature seasonality range in areas to the southwest of the present-day distribution in Zimbabwe is broad

Table 1: Area under the curve (AUC) scores for top-performing bioclimatic variables explaining current and three future distributions for *Brachystegia spiciformis* in southern Africa

	AUC score			
Bioclimatic variable	**Present**	**HadCM3**	**CGCM2**	**CSIRO-MK2**
Precipitation of wettest quarter (Bio16)	0.916	0.911	0.901	0.913
Temperature seasonality (Bio4)	0.854	0.856	0.848	0.849

Table 2: Percentage ecological niche change by 2050 for *Brachystegia spiciformis* woodland in southern Africa according to three different climatic models under the A2a scenario

	Scenario A2a		
Ecological niche	**HadCM3**	**CGCM2**	**CSIRO-MK2**
South Africa	+170.6	+376.5	-70.6
Continuous	-31.3	-47.3	-30.6
Overall	-16.9	-32.2	-32.0

(3.9–4.7 °C), but exceeds the maximum 'most suitable' temperature seasonality of 3.0 °C experienced by the species across its present-day range (Figure 3). Although there is overlap between the lower quartile of the limiting factor range and the upper tolerance limit of the current distribution, the mean of the limiting factor range (4.2 °C) is significantly higher ($p<0.01$) than the mean of the current distribution (2.8 °C).

Maximum temperature of the warmest month in areas to the north and northwest of the Zimbabwean population, and to the south and east of the Mozambican population (Figure 4), is tightly constrained between 35.9 °C and 37.2 °C (Figure 3). This maximum exceeds the upper limit of the 'most suitable' range of 35.2 °C for this variable established from the present-day distribution of the species. Despite substantial overlap between the upper tolerance of the current distribution range and the upper- and even the interquartiles for the limiting factor range, the means (36.6 °C versus 33.1 °C) differed significantly ($p<0.01$), with temperatures too high to support *B. spiciformis* woodland.

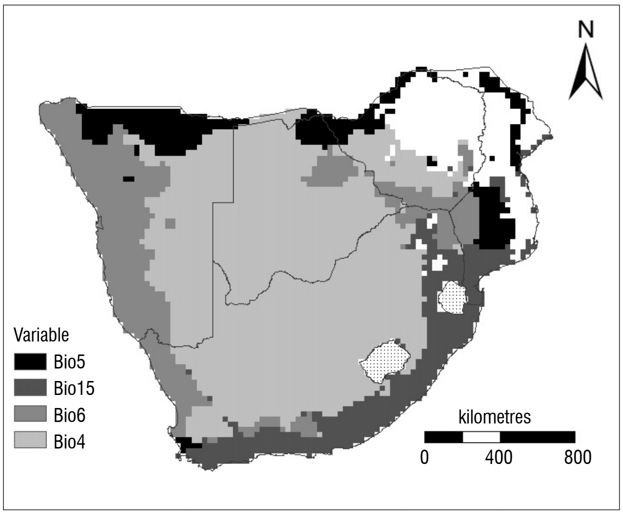
Correspondingly, precipitation seasonality (Bio15) and minimum temperature of the coldest month (Bio6) may restrict present-day distribution of the fragmented *B. spiciformis* ecological niche in South Africa (Figure 4). However, there is no significant difference ($p=0.5499$) between the limiting factor mean of adjacent minimum temperature of the coldest month (8.4 °C) and that of the 'most suitable' range (8.7 °C) for *B. spiciformis* (Figure 3). Consequently we rejected minimum temperature of the coldest month as a valid limiting factor for this species. Alternatively, precipitation seasonality's range is tightly constrained (70–78 mm) and falls short of the minimum precipitation seasonality value of 87.8 mm which bounds the present-day 'most suitable' range of the species (Figure 3). Although there is overlap between the interquartile range of the limiting factor and the lower tolerance limit of the current distribution range, the mean of the limiting factor range (74.3 mm) is significantly lower ($p<0.01$) than that of the current distribution (98.2 mm).

Discussion

Present-day ecological niche

The fragmented ecological niche suggested for *B. spiciformis* in South Africa is a refinement of the Rutherford et al.[23] model, which predicted a similar distribution for this species under future global change conditions. Moreover, the presence of a population within one of the modelled ecological niche fragments suggests a refugium. Defined on climatic grounds, refugia are physiographical settings that can support a population once prevalent regional climates have been lost (or are being lost) as a result of climate shifts.[45] The isolated *B. spiciformis* woodland relict at Gundani has therefore likely persisted in a refugium created by the varied topography of the Soutpansberg massif in South Africa. This topography has led to the formation of an isolated pocket of habitat that experiences the suitable climatic conditions (precipitation of the wettest quarter and temperature seasonality) that define the distribution of the species across the remainder of its current southern African range, and which prevailed at sites where *Brachystegia* occurred prior to the LGM and during the warmer and wetter conditions of the Mid-Holocene in the recent past.

Precipitation seasonality values in areas immediately adjacent to this isolated refugium are significantly lower than those for the ecological niche dimension of *B. spiciformis* woodland in the rest of southern Africa (Figure 3). Consequently, this bioclimatic variable reflects climatic constraints on the present-day ecological niche and isolated relict population in South Africa, potentially preventing the distribution of the species outside of the relict population. Similarly, high temperature



Bioclimatic variables: Bio5, maximum temperature of the warmest month; Bio15, precipitation seasonality; Bio6, minimum temperature of the coldest month; Bio4, temperature seasonality.

Figure 4: Limiting factors of the present-day ecological niche of *Brachystegia spiciformis* woodland in southern Africa (excluding Lesotho and Swaziland).

seasonality and maximum temperature of the warmest month (Figure 3) place climatic constraints on the extremities (southwest and northeast, respectively) of the continuous *B. spiciformis* woodlands of Zimbabwe and southern Mozambique (Figure 4). Although most savanna plants can tolerate maximum temperature extremes, Chidumayo[46] showed that some savanna trees are sensitive to seasonal highs.

Past ecological niche

According to the precipitation of the wettest quarter bioclimatic variable, conditions were considered suitable for the presence of *B. spiciformis* at the three *Brachystegia* pollen sites during the Mid-Holocene. However, temperature seasonality 6000 years BP was unfavourably high according to the present-day ecological niche dimensions of *B. spiciformis* woodland in southern Africa, casting doubt on the value of this climatic metric in controlling the distribution of *B. spiciformis* in its current range (from ENM), and in limiting its distribution north of South Africa's border (from limiting factors). However, two assumptions underlie most analyses of past climate using proxies and models. The first is that climate sets the boundaries to vegetation types, and therefore vegetation types are in equilibrium with climate except during the most rapid periods of climate change. Under this assumption, pollen–climate transfer functions can therefore provide reliable estimates of past climate. The second assumption is that, on long time scales, climate changes are driven by solar insolation changes.[47]

It should therefore be considered that precessional insolation in the southern hemisphere reached a minimum during the Early Holocene ~9000 years BP.[48] Hence, there were reduced seasonal cycles and lower temperature seasonality. Subsequently, a steady increase in summer insolation, and therefore temperature seasonality, between 9000 and 6000 years BP may already have reached an unfavourable range for *B. spiciformis* as the Mid-Holocene bioclimatic layer would suggest. The presence of pollen at these sites indicates that vegetation type boundaries are not in equilibrium with climate during rapid response phases to climate change. Adult trees may persist in an area of previously suitable climate during extended periods of climatic constraint (the storage effect), particularly if populations are capable of adaptive dynamics such as clonal regeneration. *Brachystegia* may therefore be subject to an extensive period of persistence between vegetation–climate equilibrium and local extirpation.

Future ecological niche

In keeping with the approximate northeastward retraction of the species distribution over the past 6000 years, a further ecological niche retraction of between 30.6% and 47.3% of the continuous *B. spiciformis* woodland in Zimbabwe and southern Mozambique is predicted by 2050. The decreasing suitability of habitat at the periphery of the continuous miombo population is supported by climate projections for the subcontinent.

It has been suggested that southern Africa will become hotter (temperature will increase by up to 3 °C) and drier (precipitation will decrease by ~10%) by 2060 and 2100, respectively.[49,50] The projected increase in heat waves and the number of days above 35 °C[50] over the continuous miombo woodland region will subsequently amplify variability in mean monthly temperatures. This temperature variability is likely to stress the tolerance of *B. spiciformis* further by pushing maximum temperature of the warmest month and temperature seasonality into the upper tolerance limit of the species, and closer to the limiting factor mean (Figure 3).

The future predicted decline in rainfall indicates a relatively strong drying signal for Zimbabwe and central Mozambique.[50-52] The associated decrease in precipitation of the wettest quarter over central Zimbabwe[50] may therefore place climatic constraint on the future ecological niche as this variable, which is considered the most important in determining *B. spiciformis* distribution, will become increasingly unsuitable.

In contrast, the suitable *B. spiciformis* niche in South Africa is modelled to experience very large and inconsistent spatial shifts by 2050, varying between range retractions of 70.6% to a range expansion of 376.5% (Table 2). The large discrepancy between models likely relates to the coincidence with a topographical heterogenous area on the northeastern escarpment, where the central highlands of South Africa abruptly give way to low-lying coastal plains. Fine-scale topographical effects on climate, which seem necessary for refugia, are not well captured within downscaled AOGCMs, contributing to model uncertainty.

Nevertheless, future climate change projections for northeast South Africa, which include the *B. spiciformis* woodland relict, suggest hotter (temperature increase of ~0.9 °C) and wetter (precipitation increase of ~11%) conditions by 2100.[53] Although subcontinental projections indicate increases in temperature seasonality,[50] we suggest that future change will be buffered by the varied topography of the Soutpansberg massif, as has occurred historically. We suggest that the subsequent increase in precipitation of the wettest quarter (~8%[53]) will shift local climatic conditions at Gundani, from the present-day lower tolerance limit, into the 'most suitable' range identified for the species (Figure 3). However, the concurrent regional decrease (~6%[53]) in precipitation seasonality may cause conditions to become less tolerable as rainfall shifts away from the 'most suitable' range of *B. spiciformis* woodland (Figure 3).

Considering that there may be greater uncertainty with regard to the impact of climate on the ecological niche of *B. spiciformis* at the distribution edge, as compared to the continuous range, we should consider all three possible responses of populations to climate change: migration, adaptation and extinction.

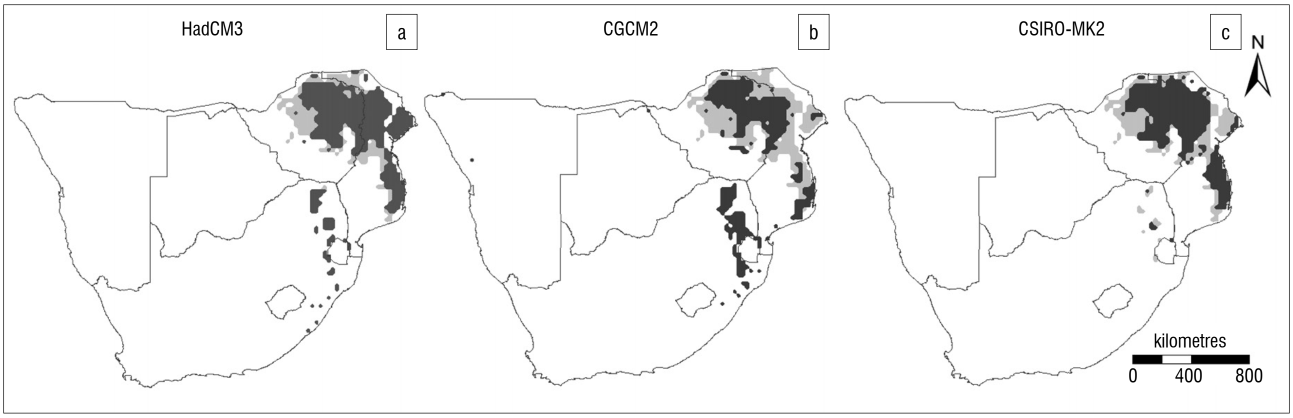


Figure 5: Predicted future (2050) ecological niche for *Brachystegia spiciformis* woodland in southern Africa (dark grey) under the A2a scenario across three general circulation models: (a) HadCM3, (b) CGCM2 and (c) CSIRO-MK2. The present-day ecological niche (light grey) is included for comparison.

Migration

It is unlikely that *B. spiciformis* woodland is capable of unaided range expansion through long-distance dispersal events, particularly given the very limited dispersal distances reported for the genus. There are limited untransformed or protected areas that could facilitate the establishment of new populations within the suggested ecological niche. In addition, current land-use practices have led to large-scale fragmentation which does not allow natural corridors for population expansion through clonal regeneration. Subsequently, any predicted range expansion of *B. spiciformis* woodland in South Africa along the eastern escarpment is considered unlikely as a result of anthropogenic constraints.

Adaptation

The persistence of populations as relicts may result from pockets of environmental suitability within the landscape, or may be the consequence of some adaptive dynamics by which species manage to overcome the climatic constraints posed by a changing climate. For example, plant ecology studies indicate climatic constraints on seed production.[54] Consequently, rather than relying on pollination and seed production, many climate relicts rely more strongly on vegetative or clonal reproduction. The well-documented reliance of *B. spiciformis* on coppice regrowth and root suckering, rather than seed germination, both at the Gundani relict[27] and in the continuous woodland[6] suggests that even in the event of failed recruitment through climatic intolerance, the populations would persist though vegetative reproduction and the longevity of genets. Relict populations can furthermore improve their survival prospects by enlarging their climatic tolerance through phenotypic plasticity or micro-evolutionary adaptation.[55] Once the capacity for phenotypic adjustment is exceeded, genetic adaptation remains the only option.[22]

However, several characteristics of relict populations imply that their potential for micro-evolutionary adaptation may be limited. Firstly, strong phenotypic divergence, compared with conspecifics from other parts of the range, does not appear to be a common phenomenon despite the presumed long-term exposure of relict populations to relatively strong selective pressures.[56] Secondly, within-population genetic variation tends to be low in many climate relicts as a consequence of small population size, past bottlenecks[3] and the occurrence of clonality. Thirdly, the small size of many relict populations, in combination with strong selection pressure, is likely to substantially elevate their risk of extinction as a result of demographic or environmental stochasticity before they can effectively adapt.[56]

Extinction

Although broadscale changes in species distribution can reasonably be forecast,[21] our understanding of the environment and ecology of most climate relicts remains too poor to adequately anticipate their persistence or demise.[22] The historical resilience of such populations demonstrates that we cannot simply assume that they will be extirpated rapidly after climate change.[22] The storage effect suggests that individuals of the species may be present in an area long after the climate has become marginally tolerable, or completely unsuitable, depending on the life stage on which the climate control may act. The isolated *B. spiciformis* woodland plays an important role in the natural heritage of the local people and has been placed under a traditional woodland management regime. The tree is not currently used ethnobotanically and the direct threat of anthropogenic extinction is considered unlikely.

With future climate-induced migration ruled out and opportunities for adaptation beyond phenotypic adjustment limited, what then is the fate of South Africa's only miombo woodland? We suggest that the medium-term persistence of the relict is plausible based on (1) future climate change projections (2100), which suggest that local climatic variables will remain within (or nearly so) the 'most suitable' range of *B. spiciformis*, (2) the longevity of genets and the species' ability to regenerate vegetatively and (3) its historical resilience.

The suggested ecological niche retraction of the continuous *B. spiciformis* woodland in Zimbabwe and southern Mozambique may, however, be less exaggerated. Bioclimatic models are limited in their ability to determine the full extent of ecological interactions, especially in savannas in which the direct impact of climate on species distribution may be variable. One such example is that of future increases in atmospheric CO_2. Species with large below-ground carbon sinks, like miombo, benefit from CO_2 fertilisation[57] which increases tree water use efficiency.[58] This tree water use efficiency is consequently likely to contribute towards a greater tolerance of suggested decreased precipitation of the wettest quarter.

Conclusion

Through this unique study, we have identified the likely climatic determinants responsible for range shift at the trailing distribution margin of *B. spiciformis* in southern Africa.

Given the suggested divergent future responses of the continuous woodlands of Zimbabwe and Mozambique relative to that of the South African relict, an understanding of population dynamics and demographics across the subcontinent will prove critical in validating our predictions concerning the climate response phases (migration, adaptation, persistence) of *B. spiciformis* and elucidating the life-history stages during which climatic constraints may be imposed. Understanding both historical and contemporary climatic determinants of a species range is the first step towards assessing the vulnerability of extant climate relicts under global climate change. Theoretical responses must then be supported by in-situ monitoring of the populations.

Climate relicts have value as instructive models and natural laboratories for investigating how populations react to ongoing climatic change.[59] Beyond the scope of this study exists an opportunity to explore genetic variation within the *B. spiciformis* relict population and investigate potential micro-evolutionary adaptation since isolation. Furthermore, development of long *B. spiciformis* and *B. utilis* tree-ring chronologies from the relict would allow for the investigation of temporal El Niño Southern Oscillation (ENSO) variability and predictions of future regional effect. Considering that ENSO effects on precipitation variability are strongest in southern Africa during the wettest quarter,[14] the investigation may assist to adequately anticipate the persistence or demise of the relict population.

Acknowledgements

B.P. received funding from the Open Society Foundation through the Global Change and Sustainability Research Institute at the University of the Witwatersrand. The South African Environmental Observation Network provided partial financial support.

Authors' contributions

B.P. and D.I.T. conceived the study and the other authors helped to develop it; B.P and B.F.N.E. performed the analyses; B.P. and D.I.T. drafted the first complete version of the manuscript; and all co-authors read and improved the manuscript.

References

1. Woodward FI. Stomatal numbers are sensitive to increases in CO_2 from pre-industrial levels. Nature. 1987;327:617–618. http://dx.doi.org/10.1038/327617a0

2. Prentice C, Cramer W, Harrison SP, Leemans R, Monserud RA, Solomon AM. A global biome model based on plant physiology and dominance, soil properties and climate. J Biogeogr. 1992;19:117–134. http://dx.doi.org/10.2307/2845499

3. Hampe A, Petit RJ. Conserving biodiversity under climate change: The rear edge matters. Ecol Lett. 2005;8:461–467. http://dx.doi.org/10.1111/j.1461-0248.2005.00739.x

4. Graham CH, Moritz C, Williams SE. Habitat history improves prediction of biodiversity in rainforest fauna. Proc Natl Acad Sci USA. 2006;103:632–636. http://dx.doi.org/10.1073/pnas.0505754103

5. Campbell BM. The miombo in transition: Woodlands and welfare in Africa. Bogor: Centre for International Forestry Research; 1996.

6. Walker BH. Is succession a viable concept in African savanna ecosystems? In: West DC, Shugart HH, Botkin DB, editors. Forest succession: Concepts and applications. New York: Springer-Verlag; 1981. p. 431–447. http://dx.doi.org/10.1007/978-1-4612-5950-3_25

7. Huntley BJ. Southern African savannas. In: Huntley BJ, Walker BH, editors. Ecology of tropical savannas. Berlin: Springer-Verlag; 1982. p. 101–109. http://dx.doi.org/10.1007/978-3-642-68786-0_6

8. Millington AC, Critchley RW, Douglas TD, Ryan P. Estimating woody biomass in sub-Saharan Africa. Washington DC: The World Bank; 1994.

9. White F. Vegetation of Africa – A descriptive memoir to accompany the Unesco/AETFAT/UNSO vegetation map of Africa, Natural Resources Research Report XX. Paris: UN Educational, Scientific and Cultural Organization; 1983.

10. Chidumayo EN. Miombo ecology and management: An introduction. London: Southampton Raw; 1997.

11. Campbell BM, Angelson A, Cunningham A, Katerere Y, Sitoe A, Wunder S. Miombo woodland – Opportunities and barriers to sustainable forest management. Bogor: Centre for International Forestry Research; 2007.

12. Chidumayo EN. Responses of miombo to harvesting: Ecology and management. Stockholm: Stockholm Environmental Institute; 1993.

13. Ernst W. Seed and seedling ecology of *Brachystegia spiciformis*, a predominant tree component in miombo woodland in south central Africa. Forest Ecol Manag. 1988;25:195–210. http://dx.doi.org/10.1016/0378-1127(88)90087-4

14. Trouet V, Esper J, Beeckman H. Climate/growth relationships of *Brachystegia spiciformis* from the miombo woodland in south central Africa. Dendrochronologia. 2010;28:161–171. http://dx.doi.org/10.1016/j.dendro.2009.10.002

15. Scott L. A late Quaternary pollen record from the Transvaal Bushveld, South Africa. Quaternary Res. 1982;17:339–370. http://dx.doi.org/10.1016/0033-5894(82)90028-X

16. Cowling RM, Richardson DM, Pierce SM. Vegetation of southern Africa. Cambridge: Cambridge University Press; 2004.

17. Eeley HAC, Lawes MJ, Piper SE. The influence of climate change on the distribution of indigenous forest in KwaZulu-Natal, South Africa. J Biogeogr. 1999;26:595–617. http://dx.doi.org/10.1046/j.1365-2699.1999.00307.x

18. Scott L. Late Quaternary palaeoenvironments in the Transvaal on the basis of palynological evidence. In: Vogel JC, editor. Late Cainozoic palaeoclimates of the southern hemisphere. Rotterdam: Balkema; 1984. p. 317–327.

19. Scott L. Palynological evidence for Quaternary palaeoenvironments in southern Africa. In: Klein RG, editor. Southern African prehistory and palaeoenvironments. Rotterdam: Balkema; 1984.

20. Thuiller W, Albert C, Araujo MB, Berry PM, Cabeza M. Predicting global change impacts on plant species' distributions: Future challenges. Perspect Plant Ecol. 2008;9:137–152. http://dx.doi.org/10.1016/j.ppees.2007.09.004

21. Hampe A, Jump AS. Climate relicts: Past, present, future. Annu Rev Ecol Evol S. 2011;42:313–333. http://dx.doi.org/10.1146/annurev-ecolsys-102710-145015

22. Prado DE, Gibbs PE. Patterns of species' distributions in the dry seasonal forests of South America. Ann Mo Bot Gard. 1993;80(4):902–927. http://dx.doi.org/10.2307/2399937

23. Rutherford MC, Midgley GF, Bond WJ, Powrie LW, Roberts R, Allsopp J. Plant biodiversity. In: Kiker G. Climate change impacts in southern Africa. Report to the National Climate Change Committee, Department of Environmental Affairs and Tourism. Pretoria: Department of Environmental Affairs and Tourism; 2000.

24. Luoga EJ, Witkowski ETF, Balkwill K. Regeneration and coppicing of miombo trees in relation to land use. Forest Ecol Manag. 2004;189:23–35. http://dx.doi.org/10.1016/j.foreco.2003.02.001

25. Hurter PJH, Van Wyk E. First distribution record for *Brachystegia spiciformis* in South Africa. Bothalia. 2001;31:43–44.

26. Saidi TA, Tshipala-Ramatshimbila TV. Ecology and management of a remnant *Brachystegia spiciformis* (miombo) woodland in north eastern Soutpansberg, Limpopo Province. S Afr Geogr J. 2006;88(2):205–212. http://dx.doi.org/10.1080/03736245.2006.9713862

27. Burrows J, Lötter M, Hahn N. A checklist of the plants found in the Venda *Brachystegia* sites and some comments on their future. PlantLife. 2003;28:5–10.

28. Erasmus BFN, Van Jaarsveld AS, Chown SL, Kshatriya M, Wessels K. Vulnerability of South African animal taxa to climate change. Global Change Biol. 2002;8:679–693 http://dx.doi.org/10.1046/j.1365-2486.2002.00502.x

29. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modelling of species' geographic distributions. Ecol Model. 2006;190:231–259. http://dx.doi.org/10.1016/j.ecolmodel.2005.03.026

30. Elith J, Graham CH, Anderson RP, Dudil M, Ferrier S, Guisan A, et al. Novel methods improve prediction of species' distributions from occurrence data. Ecography. 2006;29:129–151. http://dx.doi.org/10.1111/j.2006.0906-7590.04596.x

31. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965–1978. http://dx.doi.org/10.1002/joc.1276

32. Blach-Overgaard A, Svenning JC, Dransfield J, Greve M, Balslev H. Determinants of palm species distributions across Africa: The relative roles of climate, non-climatic environmental factors, and spatial constraints. Ecography. 2010;33;380–391. http://dx.doi.org/10.1111/j.1600-0587.2010.06273.x

33. Penman TD, Pike DA, Webb JK, Shine R. Predicting the impact of climate change on Australia's most endangered snake, *Hoplocephalus bungaroides*. Divers Distrib. 2010;16:109–118. http://dx.doi.org/10.1111/j.1472-4642.2009.00619.x

34. FAO / IIASA / ISRIC / ISSCAS / JRC. Harmonized World Soil Database (version 1.0). Rome and Luxemburg: FAO & IIASA; 2008.

35. Warren DL, Glor RE, Turelli M. ENM Tools: A toolbox for comparative studies of environmental niche models. Ecography. 2010;33:607–611.

36. PCMDI (Program for Climate Model Diagnosis and Intercomparison) [software]. c2007 [cited 2013 May 04]. Available from: http://www.pcmdi.llnl.gov/ipcc/about_ipcc.php.

37. Mika AM, Weiss RM, Olfert O, Hallett RH, Newman JA. Will climate change be beneficial or detrimental to the invasive swede midge in North America? Contrasting predictions using climate projections from different general circulation models. Global Change Biol. 2008;14:1721–1733. http://dx.doi.org/10.1111/j.1365-2486.2008.01620.x

38. Buisson L, Thuiller W, Casajus N, Lek S, Grenouillet G. Uncertainty in ensemble forecasting of species distribution. Global Change Biol. 2010;16:1145–1157. http://dx.doi.org/10.1111/j.1365-2486.2009.02000.x

39. Peterson AT, Soberón J, Anderson RP, Pearson RG, Martínez-Meyer E, Nakamura M, et al. Ecological niches and geographic distributions: A modelling perspective. Princeton, NJ: Princeton University Press; 2012.

40. Beyer HL. Geospatial Modelling Environment 0.7.2.1 [software]. c2012 [cited 2013 Apr 24]. Available from: http://www.spatialecology.com/gme.

41. Braconnot P, Otto-Bliesner B, Harrison S, Joussaume S, Peterchmitt JY, Abe-Ouche A, et al. Results of PMIP2 coupled simulations of the Mid-Holocene and Last Glacial Maximum – Part 1: Experiments and large-scale features. Clim Past. 2007;3:261–277. http://dx.doi.org/10.5194/cp-3-261-2007

42. R Development Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2012. Available from: http://www.R-project.org/

43. Baldwin RA. Use of maximum entropy modelling in wildlife research. Entropy. 2009;11:854–866. http://dx.doi.org/10.3390/e11040854

44. Kamundi DA, Victor JE. *Brachystegia spiciformis* Benth. National assessment: Red list of South African plants version 2014.1 [homepage on the Internet]. c2005 [cited 2015 June 17]. Available from: http://redlist.sanbi.org/species.php?species=240-36

45. Dobrowski SZ. A climatic basis for microrefugia: The influence of terrain on climate. Global Change Biol. 2011;17:1022–1035. http://dx.doi.org/10.1111/j.1365-2486.2010.02263.x

46. Chidumayo EN. Climate and phenology of savanna vegetation in southern Africa. J Veg Sci. 2001;12:347–354. http://dx.doi.org/10.2307/3236848

47. Wassen RJ, Claussen M. Earth system models: A test using the mid-Holocene in the southern hemisphere. Quaternary Sci Rev. 2001;21:819–824. http://dx.doi.org/10.1016/S0277-3791(01)00130-5

48. Wright HE. Global climates since the last glacial maximum. Minneapolis, MN: University of Minnesota Press; 1993.

49. Engelbrecht FA, McGregor JL, Engelbrecht CJ. Dynamics of the Conformal-Cubic Atmospheric Model projected climate-change signal over southern Africa. Int J Climatol. 2009;29:1013–1033. http://dx.doi.org/10.1002/joc.1742

50. Tadross M, Davis C, Engelbrecht F, Joubert A, Archer van Garderen E. Regional scenarios of future climate change over southern Africa. In: Davis C, editor. Climate risk and vulnerability, a handbook for southern Africa. Pretoria: Council for Scientific and Industrial Research; 2011.

51. Hulme M, Doherty R, Ngara T, New M, Lister D. African climate change: 1900–2100. Climate Res. 2001;17:145–168. http://dx.doi.org/10.3354/cr017145

52. Engelbrecht FA, Landman WA, Engelbrecht CJ, Landman S, Bopape MM, Roux B, et al. Multi-scale climate modelling over southern Africa using a variable-resolution global model. Water SA. 2011;37:5. http://dx.doi.org/10.4314/wsa.v37i5.2

53. Davis CL. A climate change handbook for north-eastern South Africa. Pretoria: Council for Scientific and Industrial Research; 2010.

54. Montesinos D, Garcia-Fayos P, Verdu M. Relictual distribution reaches the top: Elevation constrains fertility and leaf longevity in *Juniperus thurifera.* Acta Oecol. 2010;36:120–125. http://dx.doi.org/10.1016/j.actao.2009.10.010

55. Reed TE, Schindler DE, Waples RS. Interacting effects of phenotypic plasticity and evolution on population persistence in a changing climate. Conserv Biol. 2011;25:56–63. http://dx.doi.org/10.1111/j.1523-1739.2010.01552.x

56. Pfenning DW, Wund MA, Snell-Rood EC, Cruickshank T, Schlichting CD, Moczek AP. Phenotypic plasticity's impacts on diversification and speciation. Trends Ecol Evol. 2010;25:459–467. http://dx.doi.org/10.1016/j.tree.2010.05.006

57. Hoffmann WA. Post-establishment seedling success in the Brazilian cerrado: A comparison of savanna and forest species. Biotropica. 2000;32:62–69. http://dx.doi.org/10.1111/j.1744-7429.2000.tb00448.x

58. Polley HW, Tischler CW, Johnson HB. Growth, water relations, and survival of drought-exposed seedlings from six maternal families of honey mesquite (*Prosopis glandulosa*): Response to CO_2 enrichment. Tree Physiol. 1999;19:359–366. http://dx.doi.org/10.1093/treephys/19.6.359

59. Petit RJ, Hampe A, Cheddadi R. Climate changes and tree phylogeography in the Mediterranean. Taxon. 2005;54(4):877–885.

Appendix:

Key to bioclimatic variables

Bio1	Annual mean temperature
Bio2	Mean diurnal range (mean of monthly (max temp – min temp))
Bio3	Isothermality (Bio2/Bio7) (*100)
Bio4	Temperature seasonality (standard deviation *100)
Bio5	Maximum temperature of warmest month
Bio6	Minimum temperature of coldest month
Bio7	Temperature annual range (Bio5–Bio6)
Bio8	Mean temperature of wettest quarter
Bio9	Mean temperature of driest quarter
Bio10	Mean temperature of warmest quarter
Bio11	Mean temperature of coldest quarter
Bio12	Annual precipitation
Bio13	Precipitation of wettest month
Bio14	Precipitation of driest month
Bio15	Precipitation seasonality (coefficient of variation)
Bio16	Precipitation of wettest quarter
Bio17	Precipitation of driest quarter
Bio18	Precipitation of warmest quarter
Bio19	Precipitation of coldest quarter

Using satellite data to identify and track intense thunderstorms in South and southern Africa

AUTHORS:
Estelle de Coning[1]
Morne Gijben[1]
Bathobile Maseko[1]
Louis van Hemert[1]

AFFILIATION
[1]South African Weather Service – Research, Pretoria, South Africa

CORRESPONDENCE TO:
Estelle de Coning

EMAIL:
estelle.deconing@weathersa.co.za

POSTAL ADDRESS:
South African Weather Service – Research, Private Bag X097, Pretoria 0001, South Africa

DATES:
Received: 18 Nov. 2014
Revised: 02 Feb. 2015
Accepted: 15 Feb. 2015

KEYWORDS:
geostationary satellite; remote sensing; lightning; data sparse regions

HOW TO CITE:
De Coning E, Gijben M, Maseko B, Van Hemert L. Using satellite data to identify and track intense thunderstorms in South and southern Africa. S Afr J Sci. 2015;111(7/8), Art. #2014-0402, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140402

To issue warnings of thunderstorms, which have the potential for severe weather elements such as heavy rainfall and hail, is a task of all weather services. In data sparse regions, where there is no or limited access to expensive observation systems, satellite data can provide very useful information for this purpose. The Nowcasting Satellite Application Facility in Europe developed software to identify and track rapidly developing thunderstorms (RDT) using data from the geostationary Meteosat Second Generation satellite. The software was installed in South Africa and tested over the South African as well as the southern African domain. The RDT product was validated by means of 20 case studies. Over the South African region, validation was done by means of visual comparison to radar images as well as in a quantitative manner against the occurrence of lightning. Visual comparisons between the RDT product and images from satellite data as well as the occurrence of heavier rainfall were done over areas outside South Africa. Good correlations were found between the identified storms and the occurrence of lightning over South Africa. Visual comparisons indicated that the RDT software can be useful over the southern African domain, where lightning and radar data are not available. Very encouraging results were obtained in the 20 case studies. The RDT software can be a valuable tool for general and aviation forecasters to warn the public of pending severe weather, especially in areas where other data sources are absent or not adequate.

Introduction

National weather services such as the South African Weather Service (SAWS) have the responsibility to warn the public of pending hazardous weather events. Based on inputs from numerical weather prediction (NWP) models, an advisory can be issued a few days ahead of a possible severe weather event, such as heavy rainfall or strong wind. Closer to the time of a potentially dangerous weather event, warnings or watches are issued when forecasters have more certainty of such an event. Nowcasting implies the forecasting of weather events for the next 2–6 h and relies heavily on remote-sensing tools such as satellite and/or radar systems. In countries in which radar systems are available and well maintained, radar data form a crucial part of nowcasting systems as a result of the vast amount of information that can be supplied by well-calibrated radar systems. However, in many countries, especially developing and least developed countries, radar systems are too expensive to obtain and maintain.[1] In many African countries, even the basic ground-based observation systems are not adequate to provide a real-time feed of observed weather elements such as rainfall.[2] Despite the lack of enough data, the public still needs to be warned of severe weather events, which can lead to loss of life and/or property.

The World Meteorological Organization realised the need for nowcasting in countries in which advanced observation data are not available and thus initiated the Severe Weather Forecasting Demonstration Projects (SWFDP) to improve nowcasting and severe weather forecasting in data sparse regions around the globe. At the World Meteorological Organization steering group meeting for SWFDP in 2012,[3] it was recognised that there is a growing need for nowcasting tools for severe weather events in the absence of adequate real-time observation networks and radar coverage. At this meeting, it was mentioned that NWP data, together with geostationary satellite data, could be very useful for this purpose.

The Meteosat Second Generation (MSG) satellite was launched in 2002 by the European Space Agency and the European Organisation for the Exploitation of Meteorological Satellites. This satellite provides full coverage of the entire African continent with a time resolution of 15 min. Although polar orbiting satellites with microwave sensors on board can provide much more detailed information at a higher spatial resolution, the latency of these satellites' data is more than 6 h and thus are not useful in a nowcasting environment. The numerous visible (VIS), infrared (IR) and water vapour (WV) channels as well as colour (red-green-blue, RGB) combinations of these channels, can provide reliable data for nowcasting purposes over data sparse regions. When the MSG satellite was launched, an initiative was also started to task the experts in Europe to develop applications for various purposes using the 12 channels provided by the satellite. Eight so-called Satellite Application Facilities (SAF) were established, each with their own focus: (1) nowcasting and very short range forecasting, (2) ocean and sea ice, (3) climate monitoring, (4) NWP, (5) land surface analysis, (6) ozone and atmospheric chemistry, (7) radio-occultation meteorology and (8) operation hydrology.[4] Various products were developed for the nowcasting purpose, one of which is the Rapidly Developing Thunderstorms (RDT) product. The RDT uses data from the geostationary MSG satellite and NWP data to provide information on clouds related to significant convective systems. The objectives of RDT are to identify, monitor and track intense convective system clouds and to detect rapidly developing convective cells.[5]

It is becoming more and more important to improve decision-making for aviation safety. The need for more accurate decisions is expected to be even more critical with expected future increases in flight demand. 'The rapidly changing nature of convective weather places a premium on making good decisions in a short amount of time'[6]. The RDT tool can enhance nowcasting of meteorological features and could be very useful for aeronautical users.

The RDT distinguishes different phases of the thunderstorm, namely, growing, mature and decaying. The RDT has been part of the operational nowcasting tools in many European countries for a number of years. Several upgrades of the software have been done since the initial phase and the most recent operational version of the software was released in 2013. As part of the development of the RDT, extensive validation studies were done over the European region.[7] Lightning data are often used for the nowcasting and tracking of convective thunderstorms[8] because of the close relationship with convective updrafts.[9,10] Aside from its relationship to convection, lightning data can also be used as a validation tool for other nowcasting tools. The objective validation for the RDT product over the European region was done by using the European Cooperation for Lightning Detection data. The RDT product showed good skill for the full-trajectory approach using an object-orientated validation methodology. Probability of detection was 74%, probability of false detection was 2%, the false alarm ratio was 22% and the threat score was 61%.[7] Better scores were acquired when sections of trajectories or individual cloud cells were considered. In their conclusion, it was found that the RDT:

> [provides] *an accurate depiction of convective phenomena, from triggering phase to mature stage. The RDT object allows pointing out some areas of interest of a satellite image. It provides relevant information on triggering and development clouds and on mature systems.*[7]

The RDT product uses satellite and NWP data for nowcasting purposes, which could clearly be useful over data sparse regions. A project was started in 2013 to implement the software to run the Nowcasting SAF products for the southern African region. The aim of this Water Research Commission funded project was to implement and test the RDT software based on MSG and a local version of the UK Met Office Unified Model (UM) data. The purpose of the project was to provide forecasters as well as aviation meteorologists with information about the development, life cycle and dissipation of significant convection in regions where radar systems do not provide coverage (in between radars over South Africa) or where no radar systems are available (most of South Africa's neighbouring countries).

Our goal in this paper is to show the value of RDT for nowcasting processes in data sparse regions, such as Africa. Lightning and radar data over South Africa are available to validate the RDT product; thus if it can be shown that the RDT validates well over South Africa, it should be useful in other parts of Africa where much less data are available for validation purposes. Ten case studies were done with the 2012 version of the software over the South African domain. Visual as well as quantitative validation could be done; for the latter, data from the ground-based lightning detection network were used to validate the RDT product. The results of these 10 cases are discussed and some examples are shown to demonstrate the potential of the RDT product over the South African domain. Because of a lack of adequate lightning data over the African continent, a quantitative validation of the RDT could not be done for the regions outside the borders of South Africa, but some visual comparisons are shown of a few cases over countries other than South Africa.

Methods

Data

The RDT product requires the use of 11 of the 12 channels from the MSG satellite, including: high-resolution VIS, VIS 0.6, VIS 0.8, IR 1.6, IR 3.9, WV 6.2, WV 7.3, IR 8.7, IR 10.8, IR 12.0 and IR 13.4 mm. Updated satellite data are available every 15 min. Input from several NWP fields are also needed: 2-m temperature, 2-m relative humidity, 2-m dew point temperature, surface pressure, temperature and humidity at all the levels of the NWP, temperature of the tropopause, geopotential at the surface as well as a land sea mask.[5] Adjustments needed to be made to the RDT software to accommodate the UM, because in Europe most countries make use of the European Centre for Medium-Range Weather Forecasts model. At the SAWS, the local version of the UM runs operationally once a day at a horizontal resolution of 12 km and provides hourly output for 48 h ahead. Similar to the satellite data, the RDT product updates every 15 min and is available during day and night-time. For the purpose of the study, the period from 11:00 to 18:00 UTC was considered in each of the 10 case study days, as this period coincides with the time when most convection occurs – late afternoon and early evening.

For the South African domain

For the South African case studies, the domain of interest is the area covering South Africa – 22°S to 36°S and between 16°E and 33°E. The South African Lightning Detection Network (SALDN) includes 20 LS 7000 and 4 LS 7001 Vaisala sensors which detect cloud-to-ground (CG) lightning.[11]

The RDT product (version 2012 of the software) was validated against (1) radar images (visually) and (2) data from the SALDN (quantitatively). The 10 dates selected for evaluation are indicated in Table 1. These days were selected based on their convective activity, with a form of severe weather also occurring on some of the days. All the dates chosen were in the spring and summer time over South Africa (September – March), with the exception of one day in June (winter) when convective activity occurred over the central parts of South Africa.

Table 1: The 10 dates chosen for evaluation of the RDT for the South African region and those for the southern African region

South African case study dates	Southern African case study dates
31 December 2011	8 November 2013
23 June 2012	13 November 2013
6 September 2012	19 November 2013
9 October 2012	20 November 2013
17 October 2012	21 November 2013
20 October 2012	24 December 2013
8 November 2012	28 December 2013
9 November 2012	2 January 2014
10 December 2012	4 January 2014
19 January 2013	10 January 2014

For the southern African domain

For the southern African case studies, the domain of interest is the area covering 0°S to 36°S and between 12°E and 40°E. Ten case study dates were chosen over the southern African region when convective activity was evident. Validation of the RDT (version 2012) over this region could be done using (1) satellite data, using RGB images which are useful in depicting deep convective features of thunderstorms and (2) rainfall estimation from the Tropical Rainfall Measuring Mission (TRMM) data set. The satellite channels and RGB combinations that were used include: (1) the high-resolution visible (HRV)RGB, (2) convection RGB and (3) the colour-enhanced IR 10.8 channel. These RGB combinations all enhance convective features of cloud systems. In the HRVRGB, deep precipitating clouds (cumulonimbus or nimbostratus) are indicated in bright white. In the convection RGB, deep precipitating clouds with small ice particles and strong updrafts (which could include severe weather features) are indicated in bright yellow. In the colour-enhanced infrared imagery, brightness temperatures are allocated to different colours in a palette: red indicates very cold cloud top temperatures of -73 °C and colder.[12] A more complete interpretation of the MSG RGB colours can be found in the Appendix. The case dates for the southern African region are given in Table 1.

Methodology

The methodology used by the RDT algorithm

The methodology of the RDT algorithm is described in detail in the NWC SAF documentation.[5] A summary of the principles is provided here. The first step is to identify and track cloud systems and then to define satellite characteristics of these cloud systems during different phases (triggering, development and mature) of the storms. The RDT algorithm could be divided into three parts: the detection of cloud systems, the tracking of cloud systems and the discrimination of convective cloud objects.

The detection algorithm defines ‘cells’ which represent the cloud systems. In the RDT algorithm, ‘cells’ are identified by using the infrared channel (IR 10.8). An adaptive threshold that is specific to each cloud system is used, based on the brightness temperature patterns of the cell and its surroundings.

The tracking algorithm is built on the principle of overlapping between cells in two successive images. Cells are tracked according to their movement and speed in successive images. The time series of a cell’s characteristics (e.g. peripheral gradient, volume, cooling rate) serve as input variables to the discrimination algorithm.

The goal of the discrimination method is to identify the convective RDT objects among all cloud cells. The discrimination phase makes use of discrimination parameters calculated from five MSG channels – IR 10.8mm, IR 8.7mm, IR 12mm, WV 6.2mm and WV 7.3mm – as well as NWP data. Both spatial and temporal characteristics are used as discrimination parameters. The discrimination scheme is a mix between empirical rules and statistical models tuned on a learning database. The different phases of the storms are determined by using the history of the cell, the temperature trend (cooling or warming), the vertical extent, the expansion of the cell as well as whether there is convective or non-convective activity in the storm (if lightning data are used as input). NWP data are used to identify areas that are stable and where no convective development should occur, thus reducing the false alarms.

Finally, the outcome of the RDT is a map with polygons, identifying the different phases of thunderstorms: triggering, growing, mature and decaying. Each of these phases is depicted in different colours to make them easy to see if overlaid on an infrared satellite image.

Validation methodology

For the South African cases, the growing and mature phases of the RDT product were quantitatively evaluated against CG lightning data from the SALDN. CG lightning constitutes less than 50% of total lightning. During the mature phase of a thunderstorm, the maximum intra-cloud and significant amounts of CG lightning are found, while IC lightning dominates in the growing phase.[13] It should be noted that the large fraction of IC lightning in both the growing and mature phase of a thunderstorm cannot be detected by the SALDN and cannot be included in the evaluation. This limitation of the SALDN can affect the performance of the statistical evaluation scores.

Each RDT time step was evaluated against lightning that occurred in the 10 min before and 10 min after the time of the RDT. To exclude areas outside the coverage of the SALDN, a mask was applied to the evaluation domain so that only accurate lightning data were used for the validation.

The RDT algorithm is an object-orientated methodology to identify and track thunderstorms by means of polygons for different phases of the convective storms – triggering, growing and mature. To validate these objects against point measurements of lightning requires an object-orientated methodology, which was also followed by the developers.[7] Ebert[14] gives a thorough overview of the research done on spatial validation methodology. Feature-based or object-based approaches identify different attributes (such as position and size) for each individual pair of forecast observed ‘objects’.[14] The methodology behind object-orientated methods has been built into an open-source statistical software package called ‘R’.[15] This methodology identifies individual features within a field, and subsequently analyses the fields on a feature-by-feature basis. Information on errors in intensity and location can be included in the analysis. Contingency table verification statistics can be calculated using new definitions for hits, misses and false alarms according to the identified objects. The software module that can be used in R – called SpatialVx16 – currently uses the analysis methodology of Davis et al.[17,18], the merge/match algorithms of Gilleland et al.[19] and the structure, amplitude and location technique of Wernli et al.[20,21] This package was used to do an object-orientated validation of the RDT in this study.

An example of how this software matches the RDT object to a lightning object is shown in Figure 1. Figure 1a depicts the RDT product for 13:15 UTC on 9 October 2012. The RDT polygons for growing (red) and mature (purple) storms are displayed on the IR 10.8 background. Also indicated in Figure 1a is the amount of lightning which occurred in the 10 min before and after the RDT time – i.e. lightning from 13:05 to 13:25 UTC – in different colours for intensity, as indicated by the colour palette on the right. Figure 1b indicates how these features are transformed to objects and the colours indicate which RDT objects (right) match with which lightning objects (left). The analysis indicates that there were 40 hits, 10 misses, 5 false alarms and 662 correct negatives based on the objects that were identified and matched. Using the contingency table scores as defined by Wilks[22], the scores for this time stamp were: probability of detection 0.8, false alarm rate 0.111, probability of false detection 0.007 and Heidke skill score 0.831.

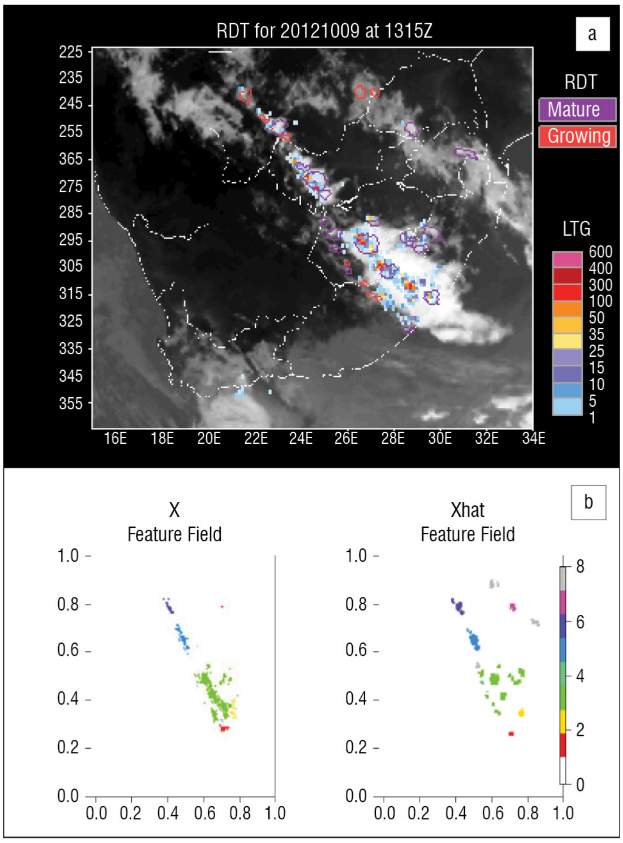


Figure 1: (a) RDT data for 9 October 2012 at 13:15 UTC, overlaid on IR 10.8 channel. Lightning intensity is given by the different colours as indicated by the palette. (b) Object-orientated comparison for the lightning occurrence (left) and RDT product (right). The colours indicate which objects match with which; grey indicates false alarms and misses.

Results

South African cases

A few examples of the 10 cases will be shown in which the lightning data surrounding the time of the RDT product (10 min before and 10 min after) are compared with radar images at the same time of the RDT. The statistical validation of the RDT against lightning data will be shown as an average of all 10 cases.

Case 1: 10 December 2012

On 10 December 2012, there was a cut-off low present over the country with a cold front ahead of it. The systems affected the country for 3 days, from 9 to 11 December 2012, and as such heavy rainfall and local floods occurred. The RDT image for 15:30 UTC (Figure 2) indicated that there was an extensive cloud shield associated with the cut-off low system over the central part of South Africa. Figure 2 includes the lightning strikes for the 10 min surrounding the RDT timestamp, i.e. 15:20 to15:40 UTC. The intensity of the lightning strikes is indicated by the colour palette on the right. Much lightning occurred along the eastern edges of the system. The most intense lightning locations were captured within the RDT storm polygons. Not all lightning is expected to be inside the RDT polygons, as this product aims to identity the more intense parts of the storms and not all thunderstorms.

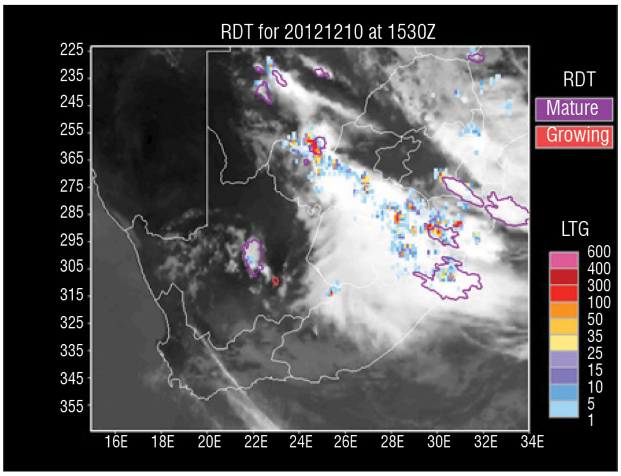


Figure 2: RDT data at 15:30 UTC overlaid with lightning strikes from 15:20 to 15:40 UTC on 10 December 2012.

Case 2: 23 June 2012

On 23 June 2012, a cut-off low system was present over the western parts of South Africa. At the eastern edge of the cut-off low system, there was a band of convective cells. These storms did not have a very high vertical extent (as would be the case with a typical summer convection). The RDT at 12:30 UTC (Figure 3a) indicated a few mature storms (purple polygons) and a number of growing storms (red polygons). This image includes the number of lightning strikes from 12:20 to 12:40 UTC, indicated by the colour palette on the right. The location of the mature storms corresponded well with the higher number of lightning strikes. Smaller amounts of lightning are noted within the growing storms. Radar reflectivity values are shown in Figure 3b, with the reflectivity scale on the right-hand side. The mature RDT storms corresponded very well with the areas of higher reflectivity in the radar image. The growing storms corresponded with the lower radar reflectivity values.

Case 3: 9 October 2012

On 9 October 2012, a cold front passed over the country. The associated trough formed a convergent region that encouraged convective activity. Convective storms started to develop around 10:00 UTC over the central and eastern parts of the country, and later developed into well-organised convective systems. The squall lines produced cirrus outflow and caused much stratiform cloud in association with the line. In Figure 4, it can be seen that areas with higher amounts of lightning corresponded well with the RDT polygons for mature storms at 15:30 UTC. The storms indicated by the RDT coincided with the most intense convective cells within the radar imagery. This case demonstrates how the RDT product can complement the radar data.

Case 4: 9 November 2012

On 9 November 2012, a surface trough extended from the northern parts of Namibia to a low-pressure system over the central interior of South Africa, with a high-pressure system east of the country. All the RDT polygons depicted on the image (Figure 5a) were associated with lightning at 11:30 UTC. The radar image (Figure 5b) indicates that the mature storms (purple polygons) correspond well with the areas of high reflectivity. The RDT identified a mature storm with lightning

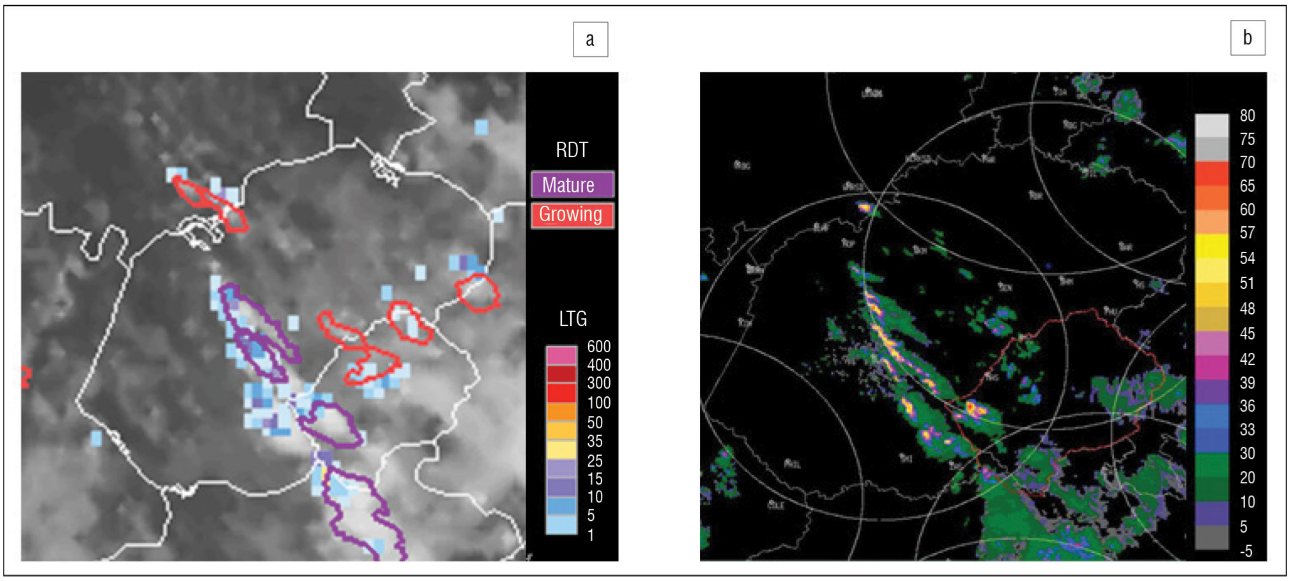


Figure 3: (a) RDT data at 12:30 UTC and lightning from 12:20 to 12:40 UTC and (b) radar reflectivity at 12:30 UTC on 23 June 2012 over the central part of South Africa.

(black arrow), where the radar image is not showing a storm (white arrow). The radar that has to cover that area was not operational on this day. It is clear that the RDT product could be beneficial in areas where radar systems are not available.

Statistical validation of the RDT

Quantitative object-orientated validation of the RDT polygons against the occurrence of all lightning strikes for all 10 cases considered is shown in Figure 6. The probability of detection, probability of false detection, false alarm rate and Heidke skill scores for all the time intervals for all 10 cases are indicated in bars and lines, respectively. The probability of detection was more than 0.6 for all time steps, with very low values for the probability of false detection. The false alarm rate was generally less than 0.2, while the Heidke skill score was more than 0.7 for most time intervals, and sometimes even close to 0.8. These scores are similar to the scores found by the developers.[7] They also used an object-orientated methodology, but other aspects of their validation scheme differ from the method that was followed in this study.

Southern African cases

A few examples of the 10 cases which were evaluated will be shown, compared to MSG RGB images and TRMM data. Because of the lack of radar and/or lightning data in this region, a quantitative validation could not be done.

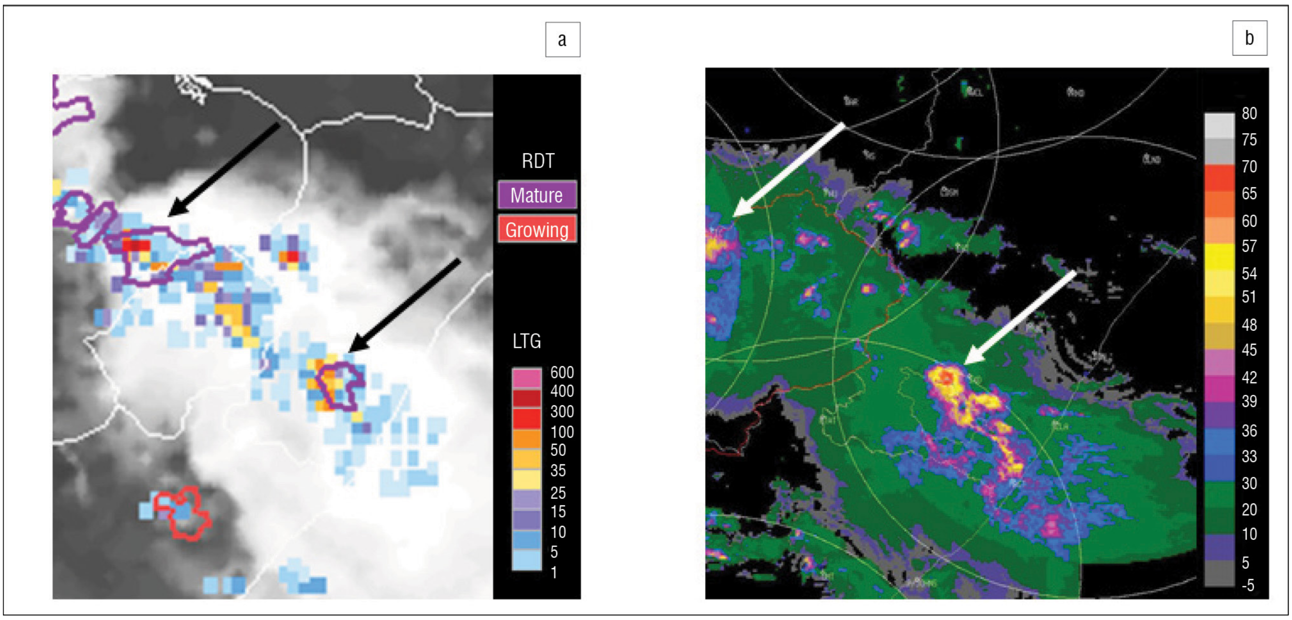


Figure 4: (a) RDT data at 15:30 UTC and lightning from 15:20 to 15:40 UTC and (b) radar reflectivity at 15:30 UTC on 9 October 2012 over the southeastern parts of South Africa.

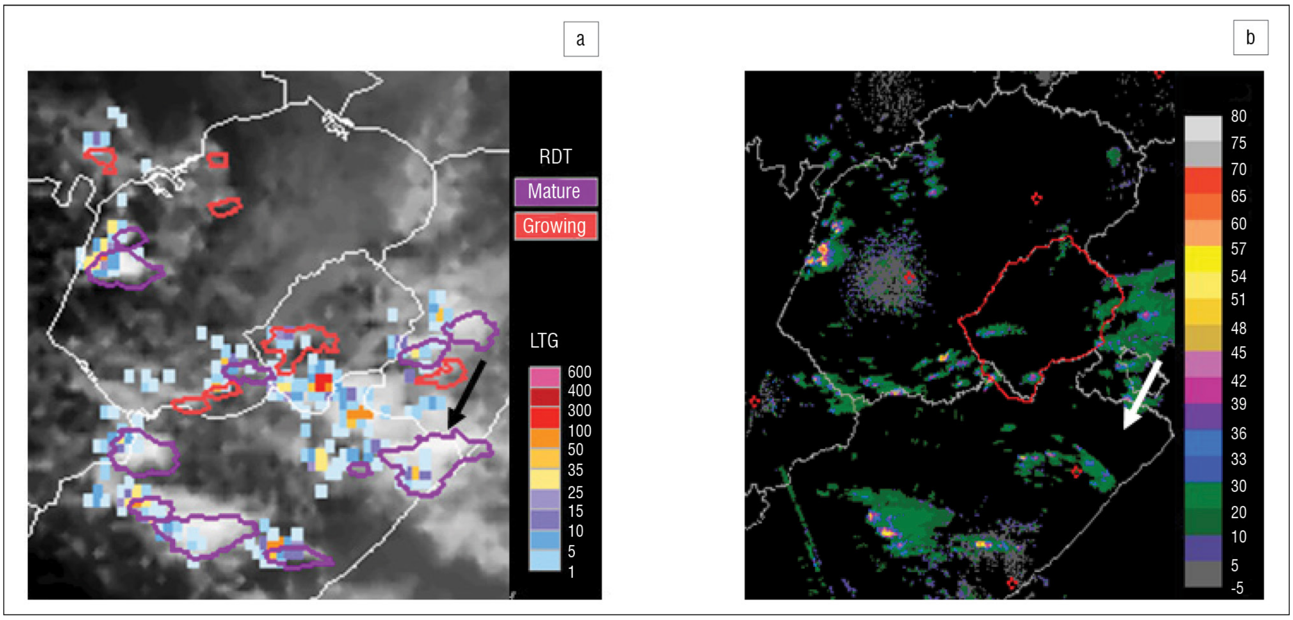


Figure 5: (a) RDT data at 11:30 UTC and lightning from 11:20 to 11:40 UTC and (b) radar reflectivity at 11:30 UTC on 9 November 2012 over the southeastern parts of South Africa.

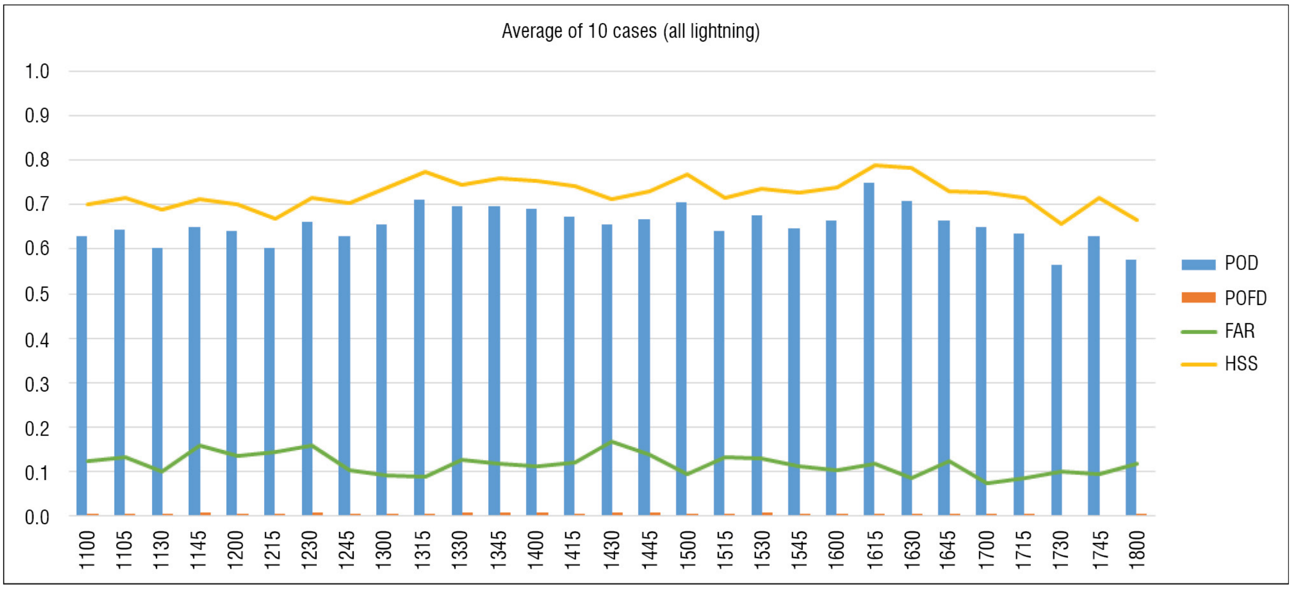


Figure 6: Probability of detection (POD, blue), probability of false detection (POFD, orange), false alarm rate (FAR, green) and Heidke skill score (HSS, yellow) for the validation of the growing and mature RDT polygons against the lightning occurrence from 11:00 to 18:00 UTC for all 10 cases.

Case 1: 10 January 2014

On 10 January 2014 a trough was located over the southern part of the continent that extended to the northeastern part of South Africa (Figure 7). The convection RGB (Figure 7b) showed that convection occurred over a large area (convection is indicated by red to bright yellow colours) at 12:00 UTC. The RDT product (Figure 7a) identified mature and growing storms in this region. The biggest storm – in the north of Angola (red circle) – was bright yellow, which indicates strong convection with the possibility of severe weather; the core of this storm was identified as a mature storm by the RDT product.

Case 2: 21 November 2013

An upper air trough was situated over the Southern African Development Community (SADC) region and – together with the typical heat effects of the tropics – resulted in large storms over Zambia and Zimbabwe on 21 November 2013. Multiple small storm cells formed over Angola and a squall line formed over the Democratic Republic of the Congo. The RDT (Figure 8a) identified a number of mature as well as smaller, growing storms at 15:00 UTC. A significant mature storm with large aerial extent was evident as indicated by the black arrow. In the HRV imagery (Figure 8b), the convection RGB (Figure 8c) and the colour-enhanced IR image (Figure 8d), the same storm seemed to be significant in size, thickness, intensity as well as cloud top temperature (-78 °C). Another large storm was also identified as a mature storm to the east of the one indicated by the arrows, which also corresponded well with the satellite imagery.

Case 3: 28 December 2013

The RDT product identified mature and growing storms over the SADC region (Figure 9a) at 12:00 UTC. The 3-h sum of rainfall as estimated by TRMM (Figure 9b) indicates that rainfall of up to 15 mm was recorded between 10:30 UTC and 13:30 UTC on 28 December 2013. The areas in which heavier rainfall was estimated by the TRMM algorithm corresponded well with the areas of mature storms identified by the RDT. The convection RGB (Figure 9c) and colour-enhanced IR (Figure 9d) both complemented the identification of mature storms by the RDT and the higher rainfall amounts over the southern parts of Zimbabwe. The RDT polygons enhance the satellite images and provide focus to the core elements of the storms.

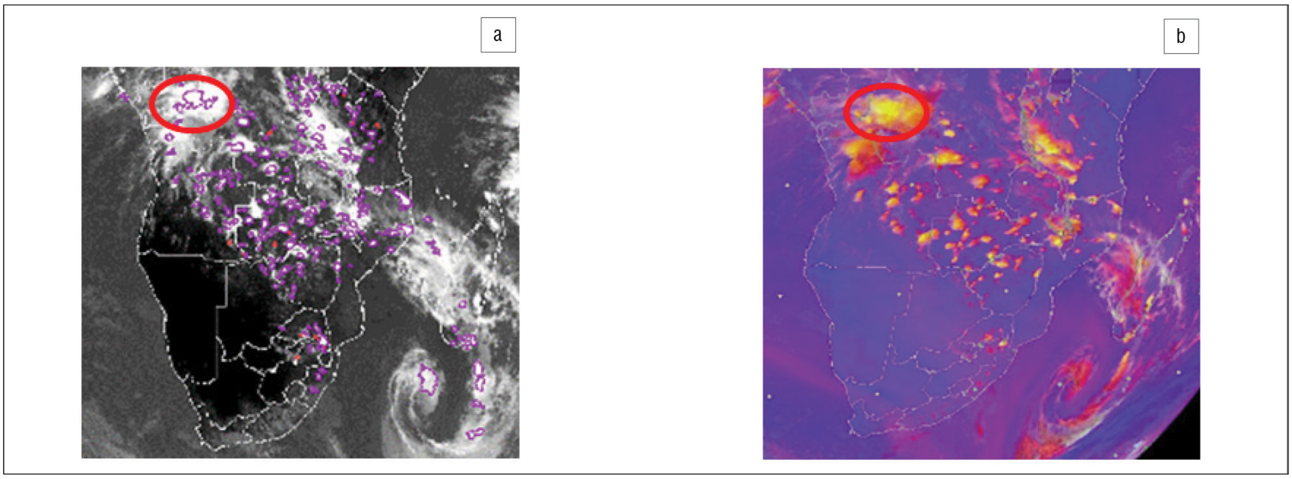


Figure 7: (a) The RDT product and (b) the convection RGB on 10 January 2014 at 12:00 UTC.

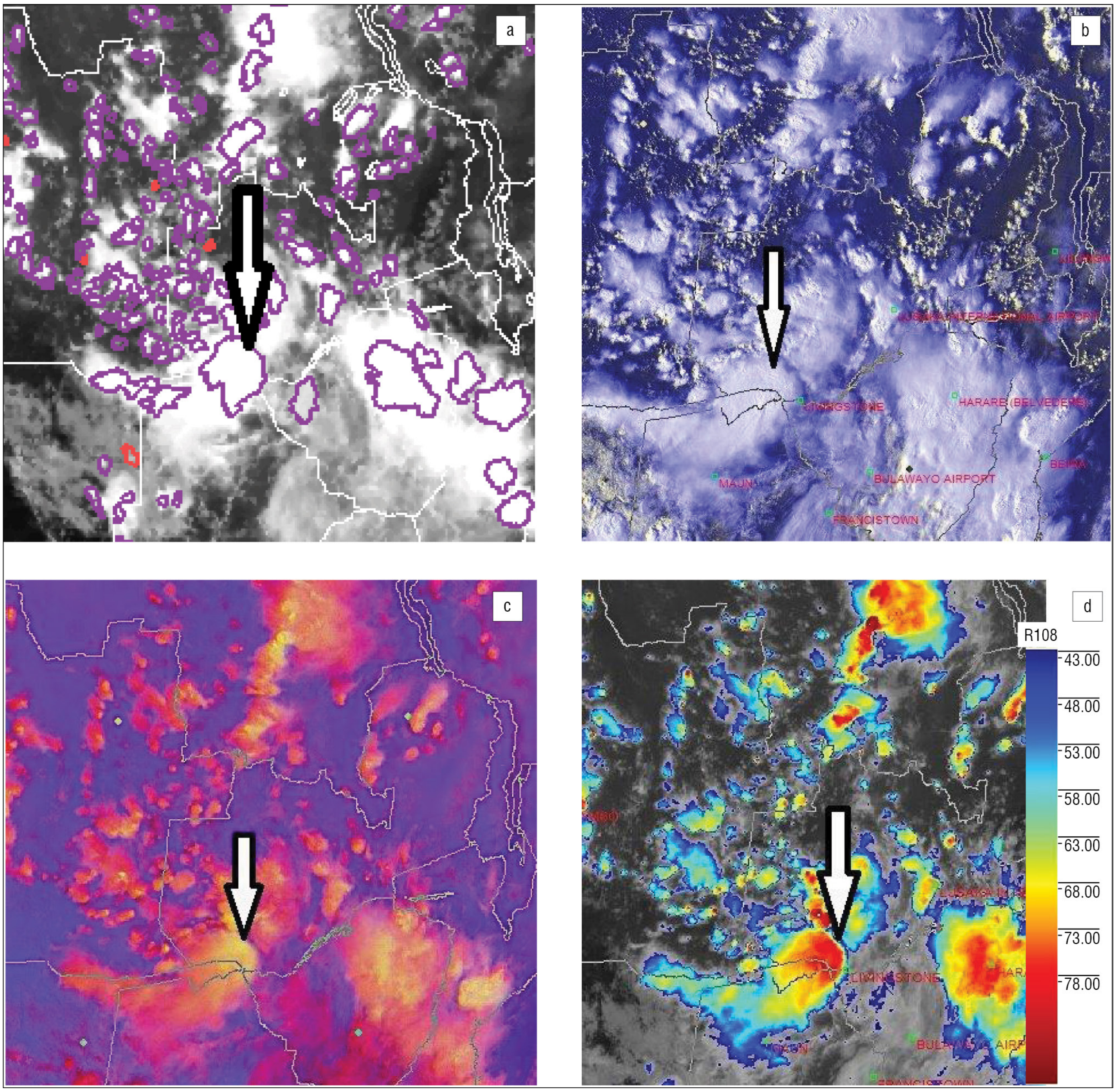


Figure 8: Images for (a) RDT, (b) HRV RGB, (c) convective storms RGB and (d) IR 10.8 colour-enhanced imagery on 21 November 2013 at 15:00 UTC over the central parts of Africa.

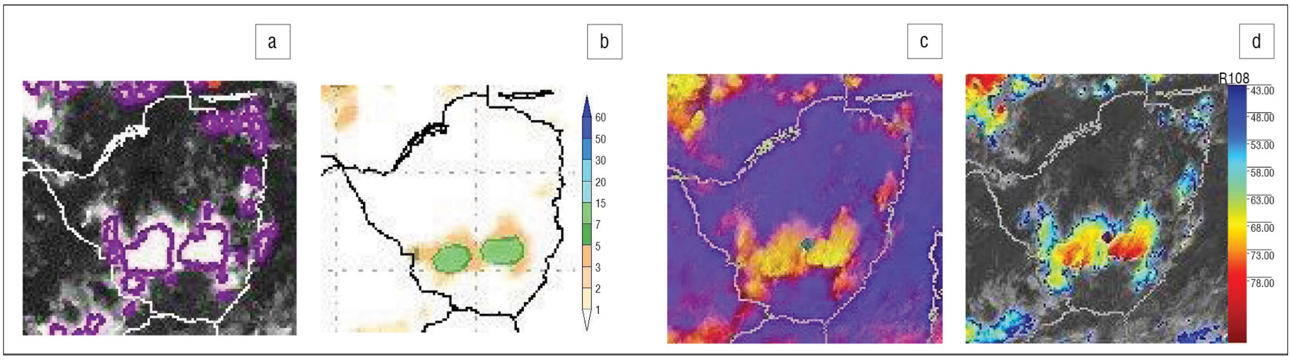


Figure 9: Images for (a) RDT, (b) 3-h TRMM rainfall, (c) convection RGB and (d) IR 10.8 colour-enhanced imagery on 28 December 2013 at 12:00 UTC over the central parts of Africa.

Conclusions

Warning the public about pending severe weather events saves lives as well as property – and is the mandate of all operational weather services. The identification and tracking of thunderstorms is a bigger challenge in data sparse regions than in areas in which expensive and extensive ground and remote observation networks exist. The RDT software, which was developed by the NWC SAF, was installed in South Africa during 2014. Initial tests of 10 cases over the South African domain and 10 cases over the southern African domain show encouraging results indicating that the RDT product can be used to identify the more intense part of thunderstorms using mainly data from the geostationary MSG satellite.

Over the South African domain, the RDT product could be compared to radar images and it was shown that the two products complemented one another in regions where both satellite and radar data are available. Lightning data from the SALDN were used to do a quantitative object-orientated validation to measure how well the RDT polygons for mature and growing storms match the occurrence of lightning. Contingency table scores indicated very good results when RDT polygon-objects were matched with lightning objects. The Heidke skill score often exceeded 0.6 and even 0.7 between 13:00 and 17:00 UTC, which is the most convective part of the day.

Over the southern African domain, adequate lightning data to do quantitative validation are not readily available. Some global lightning networks that cover the southern African region exist, but have a very low detection efficiency, which would not be useful in quantitative validation studies. Some examples of visual comparisons are shown in which the RGB combinations of satellite channels (which highlight convective activity) were used to validate the RDT storms. TRMM rainfall estimates were also used to show that higher rainfall amounts occurred in areas designated as mature storms by the RDT product. Over the southern African domain, it was shown that the RDT could add value to the satellite images to identify the more intense parts of the thunderstorms over regions where very little other data exist.

The 2013 version of the RDT product has been running operationally at SAWS since September 2014. Although this software is used extensively in European weather services, South Africa is, to our knowledge, the only African country to run this software locally and operationally. RDT products are updated every 15 min (similar to the satellite data) and are now available to operational forecasters in South Africa as well as southern African through websites designated for the purpose of nowcasting and forecasting. Initial workshops to demonstrate the new tool to general and aviation forecasters resulted in a very positive response, especially for regions in South Africa that are not covered by the radar network and for the southern African region.[23-26] Validation of the RDT will be an ongoing project, and statistics will be gathered over a longer period, such as months and seasons. The NWC SAF regularly updates their software to improve on the products and to address the needs expressed by the users. The next upgrade is planned for 2015. The latest release of the software will be downloaded and installed when it is available. SAWS is in the process of updating and upgrading their local version of the UM. This upgrade will include not only better physics, but also higher resolution. The plan is to run a 4-km resolution version of the UM over the southern African domain and perhaps even a 1.5-km resolution window over South Africa. Improvements in the NWP input will also benefit the RDT product. During 2015, we also plan to add lightning data as an additional input to the RDT product over South Africa, which, according to the developers, will lead to an improvement in the RDT product. We are convinced that this geostationary satellite-based methodology can play a role in assisting in the identification and tracking of rapidly developing and intense thunderstorms in an operational environment, especially over data sparse regions (outside South Africa). This product could benefit not only nowcasting procedures, but also clients in the aviation industry.

Acknowledgements

We thank the Water Research Commission for funding the research done to implement and validate the RDT product presented in this paper. Cassandra Pringle was part of the project team until March 2013, when she resigned from SAWS; her contribution to the case studies is appreciated. Regular interaction with the help desk of the NWC SAF contributed to the successful installation of the software. We are also grateful for the interaction with Frederic Autones from Meteo-France, one of the developers of the RDT product, which guided us in the installation of the RDT product. Guidance on the use of object-orientated validation for the SpatialVx package was provided by Beth Ebert (Bureau of Meteorology, Australia) and Eric Gilleland (NCAR, USA).

Authors' contributions

E.d.C. was the project leader, M.G. was responsible for the display and validation methodology, B.M. worked on the case studies and L.v.H. was responsible for implementing the software on local servers. E.d.C. wrote the manuscript.

References

1. De Coning E. Satellite applications for very short range weather forecasting systems in southern African developing countries. In: Gardiner S, Olsen KP, editors. Recent advances in satellite research and development. New York: Nova Science Publishers; 2013. p. 67–92.

2. Global Humanitarian Forum. Weather info for all initiative 2008-2012 [document on the Internet]. No date [cited 2014 Nov 13]. Available from: http://publicintelligence.info/WIFA_Project_Outline_Executive_Summary.pdf.

3. WMO CBS Steering Group SWFDP. Fourth meeting, Geneva, Switzerland, February 2012 [document on the Internet]. c2012 [cited 2014 Nov 07]. Available from: https://www/wmo/int/pages/prog/www/CBS-Reports/documents/SG-SWFDP-4_Final_Report.pdf.

4. Nowcasting Satellite Application Facilities [homepage on the Internet]. No date [cited 2014 Nov 07]. Available from: http://www.eumetsat.int/website/home/Satellites/GroundSegment/Safs/index.html.

5. Algorithm theoretical basis document for 'Rapid Development Thunderstorms' (RDT-PGE11 v2.3) SAF/NWC/CDOP/MFT/SCI/ATBD/11, Issue 2, Rev. 3, Applicable to SAFNWC/MSG version 2012 [homepage on the Internet]. No date [cited 2014 Nov 07]. Available from: https://www.nwcsaf.org/HD/Main.jsp.

6. Weber ME, Evans JE, Moser WR, Newell OJ. Air traffic management decision support during convective weather. Linc Lab J. 2007;16 (2):263–275.

7. Validation report for Rapidly Developing Thunderstorms (RDT-PGE11 v3.0). SAF/NWC/CDOP/MFT/SCI/ATBD/11, Issue 3, Rev. 0, Applicable to SAFNWC/MSG version 2013 [homepage in the Internet]. No date [cited 2014 Nov 07]. Available from: https://www.nwcsaf.org/HD/Main.jsp.

8. Price C. Lightning sensors for observing, tracking and nowcasting severe weather. Sensors. 2008;8:157–170. http://dx.doi.org/10.3390/s8010157

9. Weiss SA, MacGorman DR, Calhoun KM. Lightning in the anvils of supercell thunderstorms. Mon Wea Rev. 2012;140:2064–2079. http://dx.doi.org/10.1175/MWR-D-11-00312.1

10. Carey LD, Buffalo KM. Environmental control of cloud-to-ground lightning polarity in severe storms. Mon Wea Rev. 2007;135:1327–1353. http://dx.doi.org/10.1175/MWR3361.1

11. Gijben M. The lightning climatology of South Africa. S Afr J Sci. 2012;108(3/4):44–53. http://dx.doi.org/10.4102/sajs.v108i3/4.740

12. MSG channels interpretation guide: Weather, surface conditions and atmospheric constituents [homepage on the Internet]. No date [cited 2014 Nov 07]. Available from: http://eumetrain.org/IntGuide/.

13. Price CG. Lightning applications in weather and climate research. Surv Geophys. 2013;34(6):755–767. http://dx.doi.org/10.1007/s10712-012-9218-7

14. Ebert EE. Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. Meteorol Appl. 2008;15:51–64. http://dx.doi.org/10.1002/met.25

15. The R project for statistical computing [homepage on the Internet]. No date [cited 2014 Nov 14]. Available from: http://www.r-project.org/.

16. Package 'Spatial Vx' [document on the Internet]. c2014 [cited 2014 Nov 14]. Available from: http://cran.r-project.org/web/packages/SpatialVx/SpatialVx.pdf

17. Davis CA, Brown BG, Bullock RG. Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. Mon Wea Rev. 2006;134:1772–1784. http://dx.doi.org/10.1175/MWR3145.1

18. Davis CA, Brown BG, Bullock RG, Halley Gotway J. The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC Spring Program. Wea Forecsting. 2009;24:1252–1267. http://dx.doi.org/10.1175/2009WAF2222241.1

19. Gilleland E, Lee TCM, Halley Gotway J, Bullock RG, Brown BG. Computationally efficient spatial forecast verification using Baddeley's delta image metric. Mon Wea Rev. 2008;136:1747–1757. http://dx.doi.org/10.1175/2007MWR2274.1

20. Wernli H, Paulat M, Hagen M, Frei C. SAL – A novel quality measure for the verification of quantitative precipitation forecasts. Mon Wea Rev. 2008;136:4470–4487. http://dx.doi.org/10.1175/2008MWR2415.1

21. Wernli H, Hofmann C, Zimmer M. Spatial forecast verification methods intercomparison project: Application of the SAL technique. Wea Forecasting. 2009;24:1472–1484. http://dx.doi.org/10.1175/2009WAF2222271.1

22. Wilks DS. Statistical methods in the atmospheric sciences. London: Academic Press; 1995.

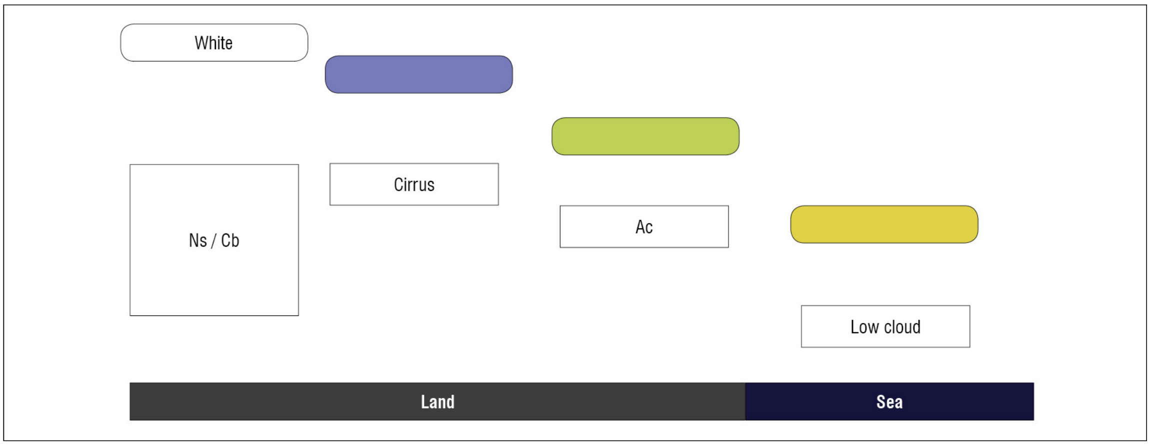
23. Maseko B, Pringle C, Gijben M. Preliminary results of the rapidly developing thunderstorm product in South Africa. In: Proceedings of the 29th South African Society for Atmospheric Scientists; 2013 Sept 26–27; Shaka's Rock, KwaZulu-Natal, South Africa. p. 44.

24. Gijben M, De Coning E, Van Hemert L, Maseko B, Pringle C. The Rapidly Developing Thunderstorm product – Results of case studies and future plans. In: Proceedings of the 30th South African Society for Atmospheric Scientists; 2014 Oct 8–10; Potchefstroom, South Africa. p. 33.

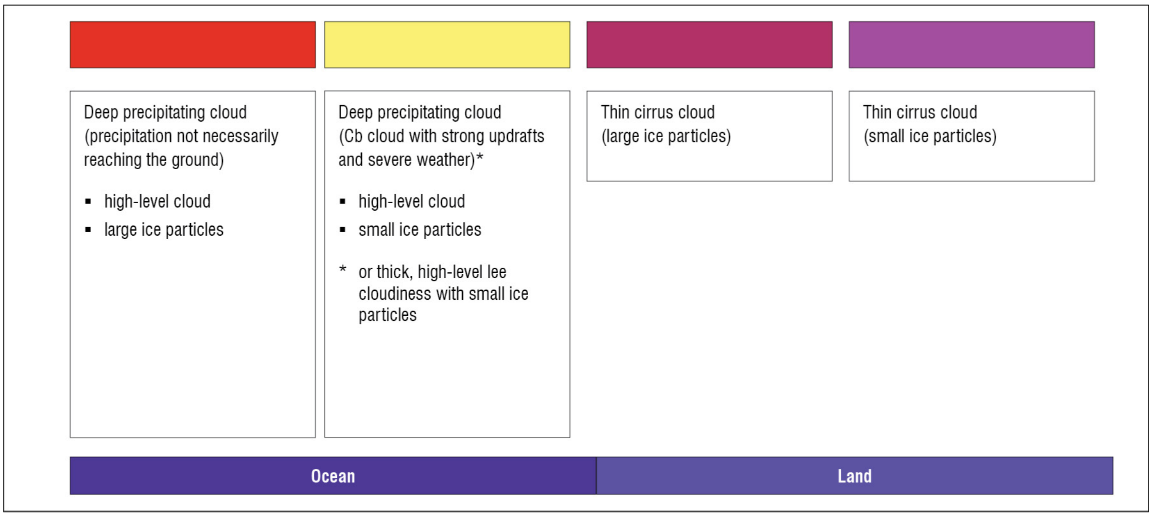
25. De Coning E, Gijben M, Maseko B, Pringle C, Van Hemert L. Using the Nowcasting SAF products over South Africa and southern Africa to enhance nowcasting capabilities in data sparse regions. In: Proceedings of the 11th EUMETSAT African User Forum; 2014 Aug 8–12; Johannesburg, South Africa.

26. De Coning E. Improving nowcasting techniques in data sparse regions using the Nowcasting SAF products. Paper presented at: Severe Weather Forecasting Demonstration Project Southern Africa – Training workshop on severe weather forecasting; 2014 Nov 3–7; Pretoria, South Africa.

Appendix



Description of colour interpretation for the high-resolution visible red-green-blue (HRVRGB) imagery.[12]


Deep precipitating cloud (precipitation not necessarily reaching the ground)
high-level cloud
large ice particles
Deep precipitating cloud (Cb cloud with strong updrafts and severe weather)*
high-level cloud
small ice particles
* or thick, high-level lee cloudiness with small ice particles
Thin cirrus cloud (large ice particles)
Thin cirrus cloud (small ice particles)
Ocean
Land

Description of the colour interpretation for the convection red-green-blue colour palette.[12]

Multiplexed CRISPR/Cas9 genome editing increases the efficacy of homologous-dependent repair of donor sequences in mammalian cells

AUTHORS:
Ezio T. Fok[1]
Clement B. Penny[1]
Musa M. Mhlanga[2,3]
Marc S. Weinberg[4,5,6]

AFFILIATIONS:
[1]Medical Oncology Research Unit, Department of Internal Medicine, School of Clinical Medicine, University of the Witwatersrand, Johannesburg, South Africa

[2]Gene Expression and Biophysics Group, Synthetic Biology–Emerging Research Area, Council for Scientific and Industrial Research, Pretoria, South Africa

[3]Unit of Biophysics and Gene Expression, Institute of Molecular Medicine, Faculty of Medicine, University of Lisbon, Lisbon, Portugal

[4]Antiviral Gene Therapy Research Unit, Department of Molecular Medicine and Haematology, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa

[5]HIV Pathogenesis Research Unit, Department of Molecular Medicine and Haematology, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa

[6]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, California, USA

CORRESPONDENCE TO:
Marc Weinberg

EMAIL:
marc.weinberg@wits.ac.za

POSTAL ADDRESS:
Antiviral Gene Therapy Research Unit, Department of Molecular Medicine and Haematology, School of Pathology, University of the Witwatersrand Medical School, 7 York Road, Parktown 2193, South Africa

DATES:
Received: 05 Jan. 2015
Revised: 04 Mar. 2015
Accepted: 23 Mar. 2015

KEYWORDS:
genome repair; genome targeting; sgRNA; HDR; T7 endonuclease I

Efficient and robust genome editing tools and strategies allow for specific and exact genetic changes to be captured in model systems, thereby accelerating both forward and reverse genetics studies. The development of CRISPR/Cas9 as a facile designer nuclease toolset has allowed for defined genetic modifications to be efficiently made through homology-directed repair of targeted DNA double-stranded breaks (DSBs) using exogenous repair templates. However, traditional single DSB strategies are still relatively inefficient as the short gene conversion tracts of mammalian cell systems limit the extent of achievable gene alteration from the DSB site. In order to improve on the inefficiency, we devised a dual cut strategy, which relies on reconstituting entire deleted gene fragments to precisely modify extensive gene regions of interest. Using the CRISPR/Cas9 system, we were able to introduce targeted deletions and repair of the endogenous *KRAS* gene locus in cell culture. The use of two simultaneous DSBs can be employed for efficient application of homology-directed repair with a large dsDNA donor sequence, thereby improving the efficacy of deriving cells with a desired gene editing outcome. In conclusion, a multiplexed CRISPR/Cas9 editing strategy represents an efficient tool for the editing of complex, heterologous sequence tracts.

Introduction

Precise genome editing represents a powerful new paradigm for both forward and reverse genetics studies in model systems. Sequence-specific nucleases have been used to expand our ability for precision engineering of the genome, and can be programmed to introduce targeted chromosomal double-stranded breaks (DSBs) to trigger endogenous DNA repair pathways.[1-3] The error-prone non-homologous end joining (NHEJ) repair pathway involves the re-ligation of the broken DNA ends, and in the process introduces small mutagenic insertions and deletions (indels). The homology-dependent repair (HDR) pathway seamlessly repairs the damaged site by utilising a homologous template, a process which can be exploited for targeted gene modifications.[4-6] These tools have enabled functional studies based on the systematic knockout and knockin of genes[7], the modelling of human diseases in cell- or animal-based systems[8-10], and the generation of isogenic cell lines for stable transgene expression[11,12].

The clustered regularly interspaced short palindromic repeats (CRISPR) and the CRISPR-associated (Cas) protein system constitute an adaptive immune system found in prokaryotes, which functions to silence foreign DNA by RNA-guided nucleolytic digestion.[13-15] CRISPR/Cas was adapted to target DNA by using short chimeric single-guide RNA (sgRNA) and a Cas9 nuclease reconstituted to function in cultured mammalian cells.[16-19] The Cas9 nuclease can be directed to a specific cognate DNA target via a ~20-nucleotide 'guide' sequence within the sgRNA. The ease at which the nuclease can be reprogrammed by one or many sgRNAs has made this system highly effective for multiplexed genome editing applications in different cells[20-23] and eukaryotic model systems[22-35].

Conventional genome editing strategies make use of a single cleavage event to stimulate gene conversion with either ssDNA or dsDNA homologous repair templates.[36] However, donor DNA sequences that are more distant from the DSB site, or which include heterologous sequences, are less efficiently repaired.[37-39] This situation can be partially resolved by the simultaneous introduction of two separate DSBs at a targeted locus to induce deletion and replacement events over large genomic regions.[26,27,40] However, the efficiency of this approach has not been determined and, to date, the HDR frequencies among approaches that apply single or multiplexed DSBs remain unknown. Using the CRISPR/Cas9 adopted from *Streptococcus pyogenes*, we exploited the capabilities of this system to induce DSBs to introduce an oncogenic *KRAS* (c.35G>T) mutant variant in a selectable human cell-based model system. We demonstrate that the facile introduction of two simultaneous DSBs improves the frequency, relative to a single DSB, of reconstituting dsDNA homologous donor sequences, but not ssDNA ODNs (oligodeoxynucleotides), at the desired locus. These results point to an improved approach for efficient HDR-mediated repair using the facile CRISPR/Cas9 editing system.

HOW TO CITE:
Fok ET, Penny CB, Mhlanga MM, Weinberg MS. Multiplexed CRISPR/Cas9 genome editing increases the efficacy of homologous-dependent repair of donor sequences in mammalian cells. S Afr J Sci. 2015;111(7/8), Art. #2015-0002, 7 pages. http://dx.doi.org/10.17159/sajs.2015/20150002

Materials and methods

Cell culture

HEK 293 cells were maintained in Dulbecco's Modified Eagle Medium: F12 mixture containing glucose (3.15 g/L), HEPES buffer (15 mM) and L-glutamine (2.50 mM) (Lonza, USA), supplemented with 10% heat-inactivated foetal bovine serum (Hyclone, USA) and 0.2% penicillin–streptomycin antibiotic mix (10 000 U/mL) (Lonza, USA). The cells were grown in a 75-cm^3 flask and incubated in a 37 °C humidified incubator with 5% CO_2.

sgRNA assembly

The sgRNA expression plasmid, pcDNA.H1.sgRNA, was generated by first using a H1.sgRNA gBlock sequence (Integrated DNA Technologies (IDT), Coralville, IA, USA) comprising two *BsmB*I sites for facile guide sequence cloning using annealed oligos. The H1.sgRNA gBlock was Gibson Assembly cloned into the standard vector pcDNA3.1(+) (Life Technologies, CA, USA). For each of the five CRISPR/Cas9 target sequences, a pair of complementary oligonucleotides was chemically synthesised by IDT (USA):

sgRNA F1 224 forward: 5′ - GATCCGAGTTTGTATTAAAAGGTAC - 3′,
sgRNA F1 224 reverse: 5′ - AAACGTACCTTTTAATACAAACTCG - 3′,
sgRNA F2 225 forward: 5′ - GATCCATGTGTGACATGTTCTAATA - 3′,
sgRNA F2 225 reverse: 5′ - AAACTATTAGAACATGTCACACATG - 3′,
sgRNA F3 226 forward: 5′ - GATCCGTTTGTATTAAAAGGTACTGG - 3′,
sgRNA F3 226 reverse: 5′ - AAACCCAGTACCTTTTAATACAAACG - 3′,
sgRNA R1 227 forward: 5′ - GATCCGTGAATTAGCTGTATCGTCA - 3′,
sgRNA R1 227 reverse: 5′ - AAACTGACGATACAGCTAATTCACG - 3′,
sgRNA R2 228 forward: 5′ - GATCCACAAGATTTACCTCTATTGT - 3′,
sgRNA R2 228 reverse: 5′ - AAACACAATAGAGGTAAATCTTGTG - 3′.

Each pair of oligonucleotides was composed of a 20-nucleotide guide sequence for their target and appropriate 5′ overhangs that allowed for cohesive-end ligation into the *BsmB*I (Thermo Scientific, USA) digested pcDNA.H1.sgRNA expression plasmid. Sanger sequencing was performed to validate the insertion of the guide sequence into the sgRNA expression plasmid.

T7 Endonuclease I assay

At 24 h prior to transfection, HEK 293 cells were seeded at a density of 250 000 cells/well in a six-well plate. The cells were transfected with 2 μg of the sgRNA expression plasmid and 2 μg of the Cas9 expression plasmid (pX330) (Addgene plasmid #42230)[16] using 12 μL of branched polyethylenimine (Sigma Aldrich, USA). At 48 h post-transfection, the cells were harvested and the gDNA was extracted from these cells using the QIAamp DNA Mini Kit (Qiagen, Germany). Genomic regions surrounding the CRISPR/Cas9 target sites were amplified from this gDNA with primers for the 5′ cleavage region (5′ *KRAS* forward: 5′ - CGCAGAACAGCAGTCTGGC - 3′ and 5′ *KRAS* reverse: 5′ - CTACGCCACCAGCTCCAAC - 3′), and the 3′ cleavage region (3′ *KRAS* forward: 5′ - AAGGCCTGCTGAAAATGACTG - 3′ and 3′ *KRAS* reverse: 5′ - GCACAGAGAGTGAACATCATGG - 3′). The T7 Endonuclease I (T7E1) assay was performed as previously described.[36,41] Briefly, 600 ng of DNA amplified from the genomic target region was denatured and annealed to form heteroduplexes. Amplicons (150 ng) were digested with 3 U of T7E1 (New England Biolabs, Germany) and resolved on an 8% polyacrylamide gel stained with SYBR Green I Nucleic Acid Gel Stain (Invitrogen, USA).

Micro-deletion detection

HEK 293 cells were transfected using polyethylenimine with 2 μg of the 5′ cleavage sgRNA expression plasmid, 2 μg of the 3′ sgRNA expression plasmid and 2 μg of pX330. gDNA was extracted using the QIAamp DNA Mini Kit (Qiagen, Germany) 48 h post-transfection. Polymerase chain reaction (PCR) amplification of the deletion junction was performed using Phusion Flash High-Fidelity PCR Mix (Thermo Scientific, USA) with an initial denaturation at 98 °C for 10 s, followed by 35 cycles of denaturation at 98 °C for 1 s, annealing at 65.4 °C for 5 s, extension at 72 °C for 10 s and a final extension at 72 °C for 1 min using the 5′ *KRAS* forward and 3′ *KRAS* reverse primers described in the T7E1 assay. The PCR products were then visualised on an ethidium bromide (0.5 μg/mL) stained TAE agarose gel (2% (w/v)).

Restriction fragment length polymorphism analysis for detection of HDR

HEK 293 cells were transfected with 2 μg of sgRNA expression plasmid, 2 μg of pX330 and 2 μL of the ssODN repair template (100 μM) for the single cut strategy. For the dual cut strategy, 2 μg of each sgRNA expression plasmid, 2 μg of pX330 and 2 μL of the ssODN repair template (100 μM) were used. After 72 h, the gDNA was extracted from these transfected cells and the repaired region was amplified using PCR with the 5′ *KRAS* forward and 3′ *KRAS* reverse primers (described for the T7E1 assay). The PCR products were purified and 300 ng of the DNA was digested with *EcoR*I (Thermo Scientific, USA). The products were resolved and visualised on a 8% polyacrylamide gel which was stained with SYBR Green I Nucleic Acid Gel Stain (Invitrogen, USA).

dsDNA donor sequences

The ssDNA ODN sequences used were chemically synthesised (IDT, USA) (*EcoR*I site underlined):

F1 ssODN: 5′-
TTGAAATAATTTTTCATATAAAGGTGAGTTTGTATTAAAAGGGAATTCTACTGG
TGGAGTATTTGATAGTGTATTAACCTTATGTGTGAC
R2 ssODN: 5′-CAGCTAATTCAGAATCATTTTGTGGACGAATATGATCCAACA
GAATTCATAGAGGTAAATCTTGTTTTAATATGCATATTACTGGTGCAG
F1+R2 ssODN: 5′-TTGAAATAATTTTTCATATAAAGGTGAGTTTGTATTAA\AA
GGGAATTCATAGAGGTAAATCTTGTTTTAATATGCATATTACTGGTGCAG.

The dsDNA donor construct was assembled by fusing five PCR products in an overlap extension PCR:

eGFP F: 5′ -
TCAAGTTGGCGGGAGACGTCGAGTCCAACCCTGGGCCCATGGTGAGCAAG
GGCGAGGAGC - 3′; eGFP R: 5′ -
CACACAAAAAACCAACACACAGATGTAATGAAAATAAAGATATTTTATTTCTAG
AGTATACGGACCGTTACTTG - 3′; 5′ HA F 5′ -
GTTTGTATTAAAAGGTACTGGTGGAGTATTTGATAGTGTATTAACCTTATCACA
CAAAAAACCAACACACAGATG - 3′; NeoR F 5′ -
TAATGAAAATGTGACTATATTAGAACATGTCACACGCGC
AGCACCATGGCCT GAAATAAC - 3′; NeoR 5′ -
ACGTCTCCCGCCAAC
TTGAGAAGGTCAAAATTCAAAGTCTGTTTCACGAAG AACTCGTCAAGAAGG
CGATAGAAG - 3′; *KRAS* F: 5′ -
GTGTGACATGTTCTAATATAGTCACATTTTCATTATTTTTATTATAAGGCCTGCT
GAAAATGACTGAATATAAACTTGTGGTAGTTGGAG - 3′; *KRAS* R1: 5′ -
GTCTACAAAATGATTCTGAATTAGCTGTATCGTCAAGGCACTCTTGCCTACGC
CAACAGCTCCAACTACCA CAAGTTT - 3′; *KRAS* R2: 5′ -
GCATATTAAAACAAGATTTACCTCTATTG TTGGATCATATTCGTCTACAAAATG
ATTCTG - 3′; 5′ HA F2: 5′ -
TAGCTGTTGCATATTGACTTCTAACACTTAGAGG - 3′; 5′ HA R2: 5′ -
ATAAGGTTAATACACTATCAAATACTCCACCAGTACC - 3′; 3′ HA F2: 5′ -
AATCATTTTGTGGACGAATATGATCCCACAAT - 3′; 3′ HA R2: 5′ -
CCATCAAACAATTATATTTCACTAGTACAATTAAATCTAACCTTT - 3′.

The plasmid pCI-NeoR-EGFP was used as a template for SV40 driven NeoR. Amplicons comprising the 5′ and 3′ homology arms, *KRAS* sequence and SV40-Neo-2A-eGFP sequences were assembled by overlap-extension PCR. The dsDNA donor was ligated into the pJet1.2/blunt Cloning Vector from the CloneJet PCR Cloning Kit (Thermo Scientific, USA) and subsequently sequenced. The dsDNA donor was linearised by PCR amplification using Phusion Flash High-Fidelity PCR Mix (Thermo Scientific, USA) and the 5′ HA F2 and 3′ HA R2 primers, with an initial denaturation step at 98 °C for 10 s, followed by 35 cycles of denaturation at 98 °C for 1 s, annealing at 62.7 °C for 5 s, extension at 72 °C for 45 s and a final extension at 72 °C for 1 min. This linear dsDNA donor construct was column purified with the GeneJet PCR Purification Kit (Thermo Scientific, Lithuania).

Quantitative real-time PCR

HEK 293 cells were transfected with 2 μg of each sgRNA plasmid, 2 μg of pX330 and 1 μg of the linear dsDNA donor PCR product. Half of the transfected cells in each well were harvested for gDNA extraction 72 h post-transfection, and the remaining cells were maintained for another 14 days. Quantitative real-time PCR (qPCR) was performed using SYBR Green PCR Mastermix (Applied Biosystems, USA) on the Applied Biosystems 7500 RT PCR System, which was programmed for an initial hold stage at 50 °C for 2 min, followed by denaturation at 95 °C for 10 min and 50 cycles of denaturation at 95 °C for 15 s and annealing and extending either at 55 °C (for primer set A, B and β-actin) or at 60 °C (for primer set C) for 1 min. An inter-plate calibrator reaction was included for each primer set to control for any technical variations between the 72-h and 14-day samples. The inter-plate calibrator template mixture was made up by adding 10 μL of each gDNA sample (16 ng/μL) obtained 72 h post-transfection. The primer sets were:

Primer set A F: 5′ - CCAAGAGAACTACTGCCATGATGC - 3′

Primer set A R: 5′ - GCATGGACGAGCTGTACAAGT - 3′

Primer set B F: 5′ - TGATATTCGGCAAGCAGGCA - 3′

Primer set B R: 5′ - GACCACCAAGCGAAACATCG - 3′

Primer set C F: 5′ - TTATTTGGGCGGAAGGCTGA - 3′

Primer set C R: 5′ - GTCAGGGACCGTCAGTTTCA - 3′

β-actin F: 5′ - ACCAACTGGGACGACATGGAGAAA - 3′

β-actin R: 5′ - TAGCACAGCCTGGATAGCAACGTA - 3′

The data of each primer set were mathematically corrected according to the changes seen in the threshold (Cq) values as recommended by the TATAA Interplate Calibrator SYBR protocol (TATAA Biocenter). The corrected data are represented as a fold change relative to β-actin.

Results

CRISPR/Cas9 can facilitate KRAS *mutagenesis*

The *KRAS* proto-oncogene is frequently found activated and facilitates a variety of cancers by acquiring point mutations at codons 12, 13 and 61, which code for a constitutively active Kras protein.[42-44] In order to model an oncogenic c.35G>T variation in codon 12 of *KRAS*, we identified suitable sgRNA target sites within a 250-bp region around the point mutation.[45] The sgRNA guide sequences were chosen on the basis of the best returned score of each target site to minimise off-target DSBs. Two of the target sites were upstream of the c.35G>T mutation (5′ cleavage sites) and two were downstream of the mutation (3′ cleavage sites) (Figure 1a). A fifth sgRNA that made use of a degenerate NAG protospacer adjacent motif (PAM) was also included (sgRNA F2 225). The H1 Pol-III promoter drove the expression of the inserted guide sequence and the RNA scaffold, as a single chimeric sgRNA molecule comprising an optimised sgRNA architecture[22] (Figure 1b). The distribution of sgRNAs allowed for potential gene deletions to span the mutation site. To determine the functionality of each sgRNA, we used a T7E1 assay to detect the indel frequency at the targeted *KRAS* locus. All the sgRNAs introduced indels at the DSB repair sites at varying efficiencies (1.28–6.01%) (Figure 1c). Guide RNAs sgRNA F1 224 and sgRNA R2 228 were the most effective, and ensured that effective targeting occurred at positions 5′ and 3′ of the *KRAS* mutation site. These sgRNAs were applied simultaneously, resulting in the deletion of 199 bp from the *KRAS* locus as determined by PCR amplification with primers that flank the deletion junction (Figure 1a and 1d).

Micro-deletions can be repaired precisely by HDR

In order to precisely modify genes, the CRISPR/Cas9 system needs to be able to stimulate the HDR pathway for the repair of DSBs according to specific exogenous repair templates. ssODNs have been shown to be able to seamlessly repair and modify targeted single DSB sites via HDR.[46] These are convenient repair templates that allow for small, targeted gene modifications to be made without the need for constructing large

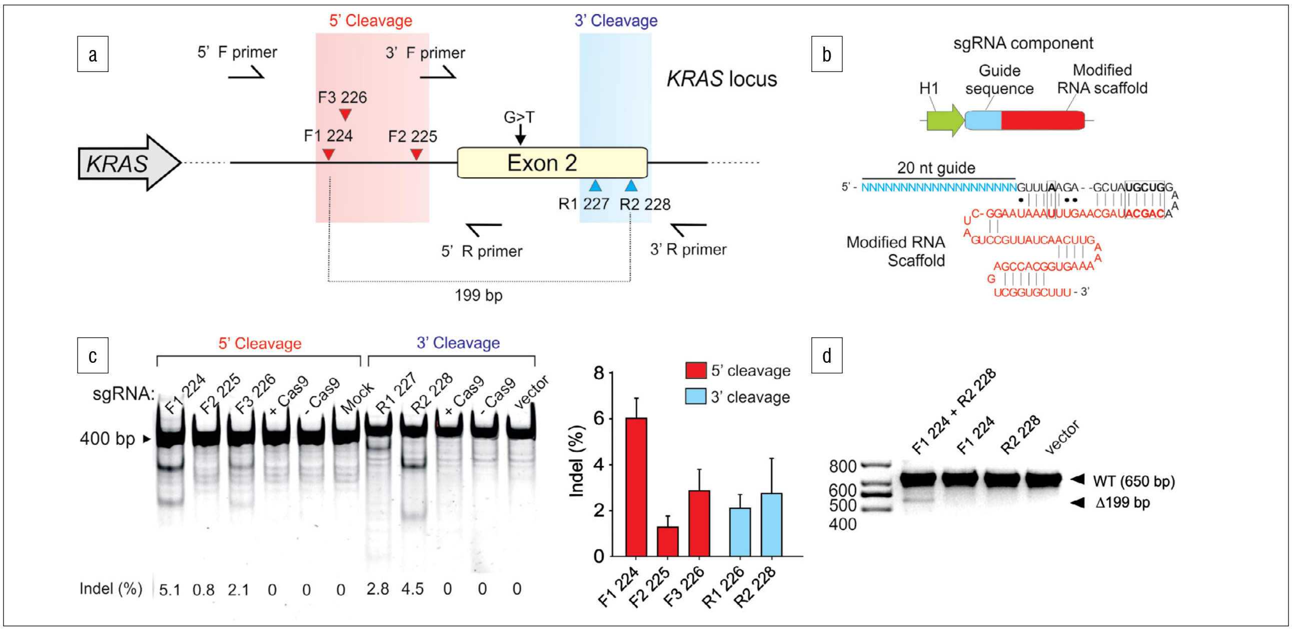


Figure 1: CRISPR/Cas9-mediated targeting of the human *KRAS* gene. (a) A schematic of the *KRAS* gene locus chosen for CRISPR/Cas9 targeting. The sgRNA target sites were designed around an oncogenic c.35G>T mutation in the *KRAS* gene, with the cleavage sites depicted as arrowheads. Two primer sets were used to amplify cleavage regions that were 5′ and 3′ of the point mutation. (b) The sgRNA construct was driven by an H1 Pol-III promoter for the expression of the guide sequence and the RNA scaffold as a single chimeric molecule. The sgRNA architecture used was previously optimised to improve sgRNA expression and Cas9 loading.[22] (c) A T7E1 assay showed the performance of each sgRNA at generating indels (n=3; mean±SD, *p<0.05, **p<0.01, unpaired Student's t-test). (d) The simultaneous application of sgRNA F1 224 and sgRNA R2 228 led to the deletion of 199 bp in the *KRAS* gene, which was detected with PCR.

donor DNA constructs. Based on an optimised design from the Church laboratory[38], ssODNs harbouring the *EcoR*I restriction site were designed to be homologous to the sense strand of the *KRAS* target region and serve as repair templates. Two 90-mer ssODNs were generated to directly flank and repair individual cleavage events generated by sgRNA F1 224 and R2 228 (Figure 2a). To demonstrate the HDR in the presence of two DSBs, a ssODN was designed to have distal homology arms, which served to bridge the resultant gap of the micro-deletion when sgRNA F1 224 and R2 228 were used concurrently (Figure 2a). The ability of the CRISPR/Cas9 to stimulate the HDR of this targeted DNA damage was assessed by restriction fragment length polymorphism analysis (Figure 2b). Guide sgRNA R2 228 facilitated HDR with an efficiency of 9.27%. The micro-deletion introduced by sgRNA F1 224 and R2 228 was efficiently repaired as was determined by the detection of a 451-bp deletion allele by PCR (Figure 2c). Interestingly, an increased abundance of the deletion allele was observed in the presence of the F1/R2 ssODN, suggesting that an oligo bridging the deletion gap facilitated NHEJ repair. The micro-deletion repair predominantly favoured the NHEJ and not the HDR pathway.

Dual cleavage generates selectable c.35G>T KRAS *mutant cells*

The application of ssODNs in tandem with targeted endonucleases to HDR-mediated genome editing is limited, even though complex and long-distance editing functions have been described.[39] However, to generate selectable recombinant cells, larger repair constructs were needed to insert expression cassettes that will allow for the selection of the modified cells. Here, we employed the CRISPR/Cas9 system in a dual cleavage strategy to introduce a *de novo* oncogenic c.35G>T point mutation in the human *KRAS* gene for positive selection. In order to do this without disrupting the gene function, we generated a knockin selection cassette for insertion into an intronic site 94 bp away from the mutation site in exon 2. Conventional strategies make use of a single DSB to stimulate HDR with dsDNA donor construct that is homologous to the sequences flanking a single DSB (Figure 3a). To improve on this approach, a dual cut repair strategy with a large dsDNA donor, using the CRISPR/Cas9 system, was predicted to efficiently facilitate the simultaneous c.35G>T transversion in exon 2 and the insertion of a selectable expression cassette into intron 1 of the *KRAS* gene. We used sgRNA F1 224 and R2 228 to introduce two concurrent DSBs to delete the intron 1–exon 2 junction, which was then reconstituted with the dsDNA donor (Figure 3a and 3b). To test the performance of the dual cleavage strategy (Figure 3b), a qPCR strategy was devised to quantify the modified *KRAS* DNA. Genomic DNA was extracted from the entire cell population without marker selection to minimise the enrichment of illegitimate recombinants. Various DNA species were quantified in the cell population 3 and 14 days post-transfection, using primer sets A–D. This quantification allowed for stable genomic integrants to be detected and monitored over time, while the residual episomal DNA from the transient transfection was diluted through continued cell proliferation. Approximately 1.9–2.5-fold more of the target DNA was detected with the dual cleavage strategy using primer set A, which spanned the gene–donor junction (Figure 3b and 3c) and ensured detection of the integrated dsDNA donor in the *KRAS* gene locus. Transfections that were sgRNA deficient showed no detectable dsDNA donor integration, confirming that donor DNA integration did not occur spontaneously. A mock donor failed to be detected, pointing to the specificity of the

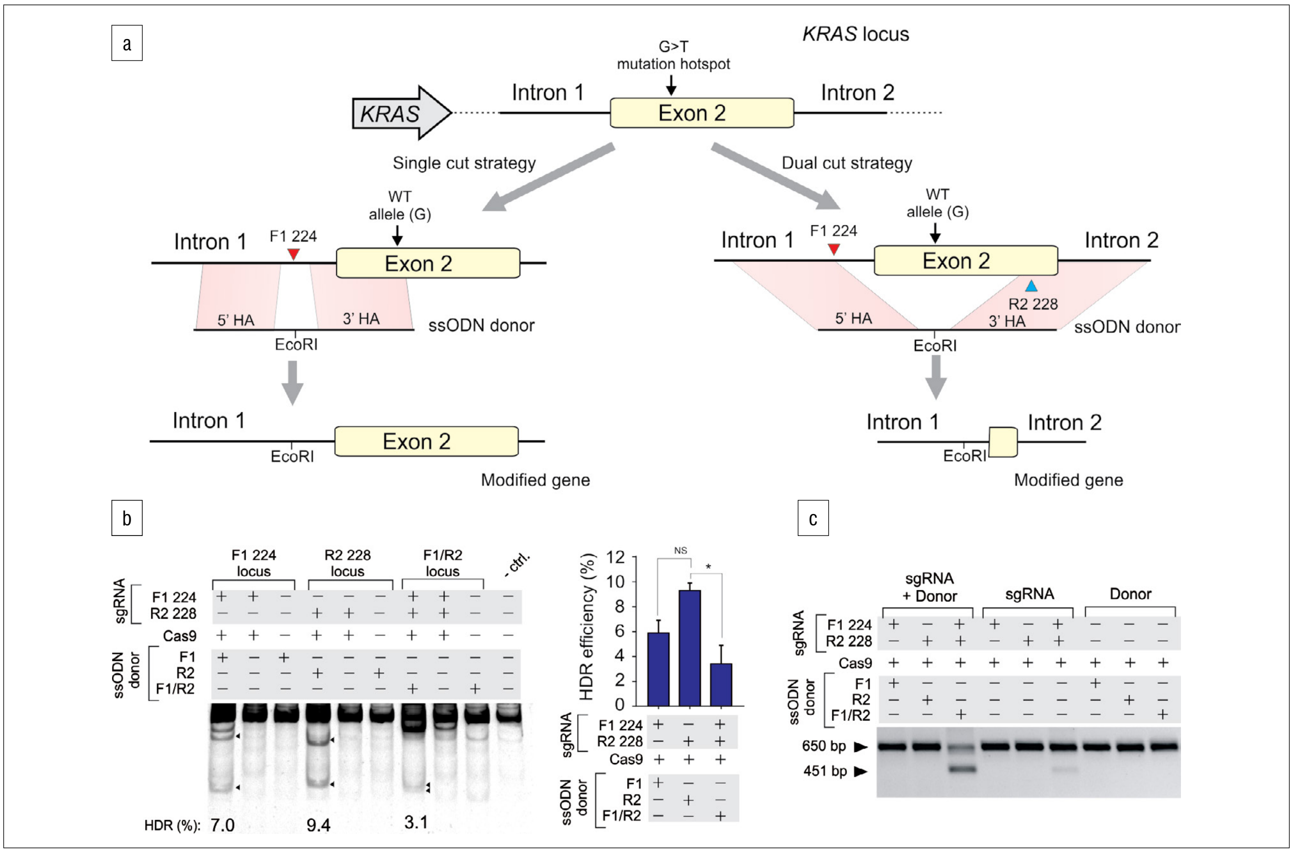


Figure 2: HDR of the individual DSBs and the micro-deletion with ssODNs. (a) 90-mer ssODNs were designed to repair DSBs from a single cut (sgRNA F1 224 or sgRNA R2 228) (left) as well as a micro-deletion from the use of a dual cut (both sgRNA F1 224 and sgRNA 228) (right). The homology arms of the ssODNs for the repair of single cuts flanked sequences directly adjacent to their respective DSBs. The micro-deletion repair ssODN was designed to flank distal sequences that were outside of the F1 224 and R2 228 DSBs. Each ssODN carried the *EcoRI* recognition site that was targeted for integration into the repair site. (b) Restriction fragment length polymorphism analysis of the PCR products demonstrated the HDR of single DSBs and the micro-deletion as the genomic integration of the *EcoRI* restriction site (n=3; mean±SD, *p<0.05, unpaired Student's t-test). (c) PCR products of the repaired region amplified from gDNA, showing the full length (650 bp) and the deletion amplicons (451 bp).

dsDNA donor for HDR. The dilution of the episomal DNA was apparent after 14 days, with a decrease of approximately 99% in each sample (primer set B; Figure 3b and 3c). Despite this dilution, readily detectable amounts of episomal DNA were still present in the transfected cells after 14 days of sub-culture (600–1000-fold relative to β-actin). The amount of detected dsDNA donor integrated into the *KRAS* gene (data from primer set A), taken as a percentage of the total *KRAS* DNA detected (data from primer set C; Figure 3b and 3c), was used as a metric for site-specific donor integration efficiency (Figure 3d). The simultaneous use of the two sgRNAs proved to be the most efficient at facilitating the integration of the heterologous donor DNA sequences, and occurred at a frequency of 0.028% and 0.123% in transfected cells, after 3 and 14 days of selection-free culture, respectively. The c.35G>T *KRAS* mutation confers a slight growth advantage in the modified cells.[43] We observed an enrichment of modified *KRAS* mutants as the amount of integrated donor DNA increased after 14 days of selection-free culture of the transfected cell population (Figure 3c and 3d).

Discussion

We used the CRISPR/Cas9 system for the improved efficacy of donor-led HDR using a multiple editing approach which allows for extensive and wide-ranging gene modifications to be achieved while maintaining gene function. As an example, the introduction of a putative oncogenic point mutation in the *KRAS* gene in a cell-based model was adopted in this study. We initially demonstrated the efficiency of introducing single cleavage events in the *KRAS* gene sequence using RNA-guided CRISPR/Cas9. These single cleavage events were repaired either by the NHEJ pathway, resulting in the introduction of mutagenic indels, or by the HDR pathway, which seamlessly repaired the DSB by using homologous ssODN templates. The NHEJ and HDR activity induced by

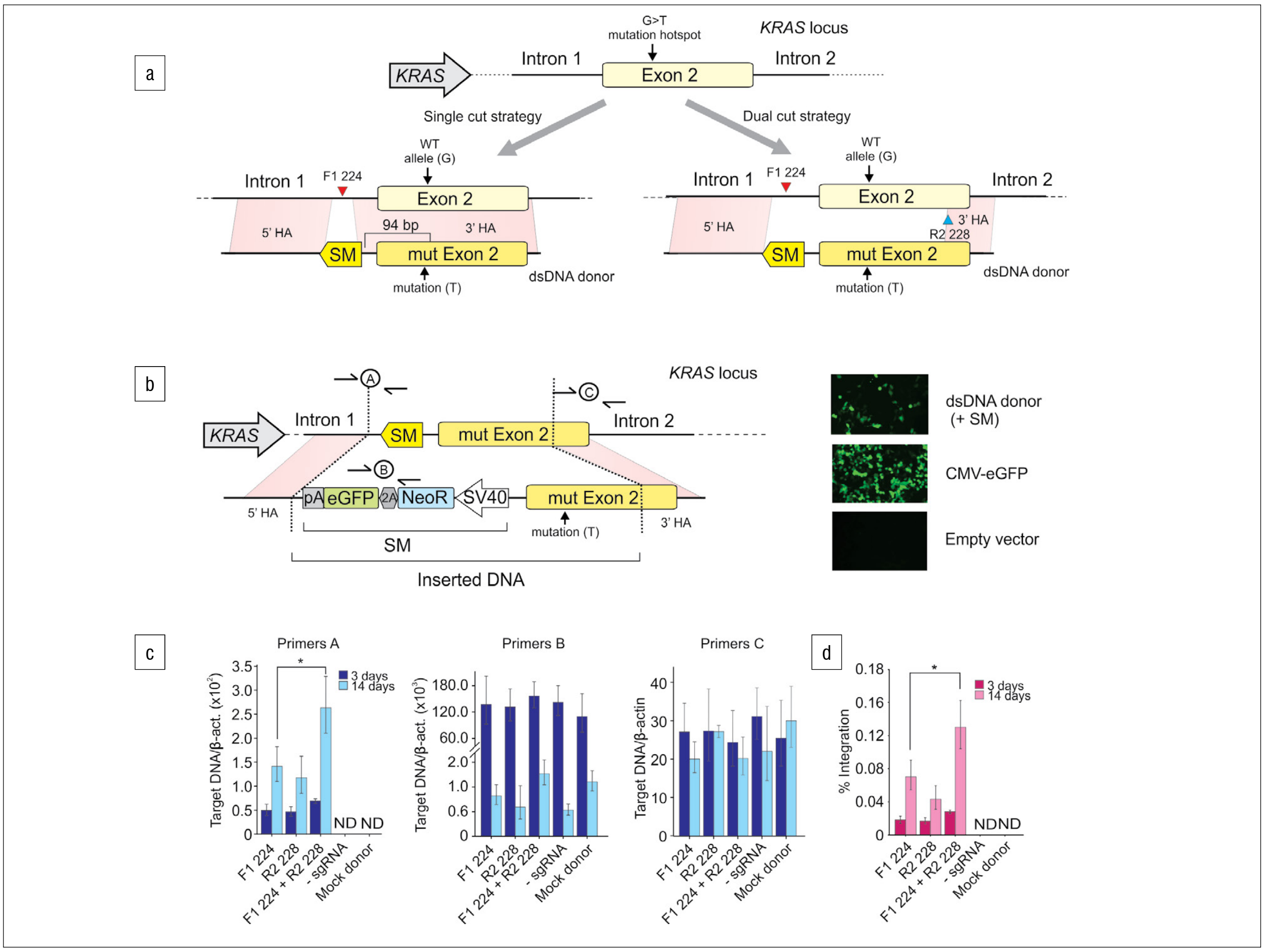


Figure 3: The envisioned dual cut strategy versus the traditional single cut strategy to modify the *KRAS* gene. (a) The extent of gene conversion is limited with the traditional single cut strategy, which uses a dsDNA donor design that has two homology arms (HAs) that directly flank the DSB site. The successful co-transfer of both the gene mutation and the selection marker cassette (SM) into sites that are 94 bp apart will be very inefficient with a single cut, as a result of short gene conversion tracts. The dual cut strategy should overcome this inefficiency by deleting the entire region between the two sites and then reconstituting this region with a modified gene sequence that contains the mutation and the selection marker. (b) The dsDNA donor was designed to functionally separate the intended gene modification elements (the selection cassette and the point mutation) from the homology arms. This separation ensured that the homology arms maintained complete homology with the target sequence and functioned solely to direct the construct to the target region. The selection cassette was designed to be expressed in the opposite direction of *KRAS* gene transcription to produce a chimeric eGFP-NeoR protein that is cleaved by an embedded 2A self-cleaving peptide, allowing for a fluorescence or antibiotic-based approach for selection. Fluorescence micrographs showed the expression of eGFP in HEK293 cells that were transfected with this construct (right). A qPCR strategy was used to determine the integration efficiency of the dsDNA donor into the *KRAS* gene by measuring the amount of various DNA components (left). Quantification was done on extracted gDNA 3 and 14 days post-transfection. (c) Relative amounts of detected target DNA of each primer set, normalised to β-actin. (d) The percentage of *KRAS* alleles that successfully underwent integration of the selection cassette at the *KRAS* gene locus, as calculated by taking the data from primer set A (integration events) as a percentage of primer set C (total *KRAS* DNA) (mean±SD, ND denotes no detected amplification).

sgRNA F1 224 showed similar levels of stimulation of each pathway. However, DSBs mediated by sgRNA R2 228 clearly favoured the HDR pathway for DSB resolution. Furthermore, HDR of the micro-deletion, with a ssODN, occurred at relatively low levels compared to the deletion events detected, indicating that NHEJ repair was more favourable. These repair pathway preferences of DNA damage could be related to cellular mechanisms that prioritise the repair of certain genetic elements.[47] Nevertheless, such repair preferences could potentially alter the efficiency of the desired gene editing outcome. Further observation of gene editing at different genomic sites, on a more high-throughput scale, may provide further insight.

We demonstrated that the CRISPR/Cas9 system can be used in multiplex to improve the efficiency of HDR for precise modifications in the *KRAS* gene. We predicted that by applying two cleavage events, the gene region of interest could be deleted and efficiently reconstituted with our mutant-containing dsDNA donor, which would allow for a more expansive genomic region to be precisely modified, without being limited by the short gene conversion tracts that branch out from a single break site. Furthermore, the design of the dsDNA donor for this strategy removes the point mutation from the homology arms and relocates it into the reconstitution sequence, allowing for complete homology with the target sequence to be maintained. This strategy should improve the efficiency of gene conversion for single, as well as multiple, point mutations, as small levels of sequence divergence in the homology arms have been shown to greatly decrease the levels of recombination.[37] To illustrate this, we targeted two distal sites within the *KRAS* gene locus for concurrent modification, using a single donor. This approach allowed for the concurrent insertion of a selection cassette into a nearby intron of the oncogenic point mutation site in the *KRAS* protein-coding sequence, preserving the integrity and function of the gene and eliminating the need for any post-modification gene restoration to rescue function.

Conclusions

A dual cut gene editing strategy allows for distally located sites to be targeted simultaneously, enabling for extensive and expansive gene modifications to be achieved efficiently. We have shown that this method is particularly useful for maintaining the native function of the gene, in which selection cassettes used to identify cells with the modifications of interest can be inserted in neighbouring introns. We significantly improved on current HDR-based gene editing strategies, which can only modify regions within close proximity to the DSB site. Furthermore, the CRISPR/Cas9 system is well suited for this dual cleavage approach, as it is simple and cost effective to employ in a multiplex fashion. This method should improve the success rate of producing recombinant cell lines, and thus expedite efforts to gain a better understanding of causal genetic elements in biological systems.

Acknowledgements

Research reported in this publication was supported by the Strategic Health Innovation Partnerships (SHIP) Unit of the South African Medical Research Council with funds received from the South African Department of Science and Technology. Funding from the National Research Foundation is also acknowledged.

Authors' contributions

E.T.F. and M.W.S. were the project leaders; E.T.F., M.S.W., C.B.P., and M.M.M. made conceptual contributions and helped design the study; E.T.F. and M.S.W. performed the experiments and wrote the manuscript.

References

1. Gaj T, Gersbach CA, Barbas CF. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. 2013;31(7):397–405. http://dx.doi.org/10.1016/j.tibtech.2013.04.004

2. Kim H, Kim JS. A guide to genome engineering with programmable nucleases. Nature Rev Genet. 2014;15(5):321–334. http://dx.doi.org/10.1038/nrg3686

3. Perez-Pinera P, Ousterout DG, Gersbach CA. Advances in targeted genome editing. Curr Opin Chem Biol. 2012;16(3–4):268–277.

4. Rouet P, Smih F, Jasin M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. Mol Cell Biol. 1994;14(12):8096–8106.

5. Choulika A, Perrin A, Dujon B, Nicolas JF. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of *Saccharomyces cerevisiae*. Mol Cell Biol. 1995;15(4):1968–1973.

6. Porteus MH, Baltimore D. Chimeric nucleases stimulate gene targeting in human cells. Science. 2003;300(5620):763. http://dx.doi.org/10.1126/science.1078395

7. Stroud DA, Formosa LE, Wijeyeratne XW, Nguyen TN, Ryan MT. Gene knockout using transcription activator-like effector nucleases (TALENs) reveals that human NDUFA9 protein is essential for stabilizing the junction between membrane and matrix arms of complex I. J Biol Chem. 2013;288(3):1685–1690. http://dx.doi.org/10.1074/jbc.C112.436766

8. Soldner F, Laganiere J, Cheng AW, Hockemeyer D, Gao Q, Alagappan R, et al. Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. Cell. 2011;146(2):318–331. http://dx.doi.org/10.1016/j.cell.2011.06.019

9. Musunuru K. Genome editing of human pluripotent stem cells to generate human cellular disease models. Dis Models Mech. 2013;6(4):896–904. http://dx.doi.org/10.1242/dmm.012054

10. Dow LE, Lowe SW. Life in the fast lane: Mammalian disease models in the genomics era. Cell. 2012;148(6):1099–1109. http://dx.doi.org/10.1016/j.cell.2012.02.023

11. DeKelver RC, Choi VM, Moehle EA, Paschon DE, Hockemeyer D, Meijsing SH, et al. Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. Genome Res. 2010;20(8):1133–1142. http://dx.doi.org/10.1101/gr.106773.110

12. Perez-Pinera P, Ousterout DG, Brown MT, Gersbach CA. Gene targeting to the ROSA26 locus directed by engineered zinc finger nucleases. Nucleic Acids Res. 2012;40(8):3741–3752. http://dx.doi.org/10.1093/nar/gkr1214

13. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science. 2007;315(5819):1709–1712. http://dx.doi.org/10.1126/science.1138140

14. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science. 2008;321(5891):960–964. http://dx.doi.org/10.1126/science.1159689

15. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008;322(5909):1843–1845. http://dx.doi.org/10.1126/science.1165771

16. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121):819–823. http://dx.doi.org/10.1126/science.1231143

17. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. eLife. 2013;2:e00471. http://dx.doi.org/10.7554/eLife.00471

18. Cho SW, Kim S, Kim JM, Kim JS. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol. 2013;31(3):230–232. http://dx.doi.org/10.1038/nbt.2507

19. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339(6121):823–826. http://dx.doi.org/10.1126/science.1232033

20. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014;343(6166):80–84. http://dx.doi.org/10.1126/science.1246981

21. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343(6166):84–87. http://dx.doi.org/10.1126/science.1247005

22. Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell. 2013;153(4):910–918. http://dx.doi.org/10.1016/j.cell.2013.04.025

23. Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. One-step generation of mice carrying reporter and conditional alleles by CRISPR/Cas-mediated genome engineering. Cell. 2013;154(6):1370–1379. http://dx.doi.org/10.1016/j.cell.2013.08.022

24. Dickinson DJ, Ward JD, Reiner DJ, Goldstein B. Engineering the *Caenorhabditis elegans* genome using Cas9-triggered homologous recombination. Nat Methods. 2013;10(10):1028–1034. http://dx.doi.org/10.1038/nmeth.2641

25. Paix A, Wang Y, Smith HE, Lee CY, Calidas D, Lu T, et al. Scalable and versatile genome editing using linear DNAs with microhomology to Cas9 sites in *Caenorhabditis elegans*. Genetics. 2014;198:1347–1356. http://dx.doi.org/10.1534/genetics.114.170423

26. Gratz SJ, Cummings AM, Nguyen JN, Hamm DC, Donohue LK, Harrison MM, et al. Genome engineering of *Drosophila* with the CRISPR RNA-guided Cas9 nuclease. Genetics. 2013;194(4):1029–1035. http://dx.doi.org/10.1534/genetics.113.152710

27. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in *Drosophila*. Genetics. 2014;196(4):961–971. http://dx.doi.org/10.1534/genetics.113.160713

28. Kondo S, Ueda R. Highly improved gene targeting by germline-specific Cas9 expression in *Drosophila*. Genetics. 2013;195(3):715–721. http://dx.doi.org/10.1534/genetics.113.156737

29. Ren X, Sun J, Housden BE, Hu Y, Roesel C, Lin S, et al. Optimized gene editing technology for *Drosophila melanogaster* using germ line-specific Cas9. Proc Natl Acad Sci USA. 2013;110(47):19012–19017. http://dx.doi.org/10.1073/pnas.1318481110

30. Yu Z, Ren M, Wang Z, Zhang B, Rong YS, Jiao R, et al. Highly efficient genome modifications mediated by CRISPR/Cas9 in *Drosophila*. Genetics. 2013;195(1):289–291. http://dx.doi.org/10.1534/genetics.113.153825

31. Gennequin B, Otte DM, Zimmer A. CRISPR/Cas-induced double-strand breaks boost the frequency of gene replacements for humanizing the mouse Cnr2 gene. Biochem Biophys Res Commun. 2013;441(4):815–819. http://dx.doi.org/10.1016/j.bbrc.2013.10.138

32. Li D, Qiu Z, Shao Y, Chen Y, Guan Y, Liu M, et al. Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. Nat Biotechnol. 2013;31(8):681–683. http://dx.doi.org/10.1038/nbt.2661

33. Niu Y, Shen B, Cui Y, Chen Y, Wang J, Wang L, et al. Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. Cell. 2014;156(4):836–843. http://dx.doi.org/10.1016/j.cell.2014.01.027

34. Platt RJ, Chen S, Zhou Y, Yim MJ, Swiech L, Kempton HR, et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell. 2014;159(2):440–455. http://dx.doi.org/10.1016/j.cell.2014.09.014

35. Xue W, Chen S, Yin H, Tammela T, Papagiannakopoulos T, Joshi NS, et al. CRISPR-mediated direct mutation of cancer genes in the mouse liver. Nature. 2014;514:380–384. http://dx.doi.org/10.1038/nature13589

36. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protocols. 2013;8(11):2281–2308. http://dx.doi.org/10.1038/nprot.2013.143

37. Elliott B, Richardson C, Winderbaum J, Nickoloff JA, Jasin M. Gene conversion tracts from double-strand break repair in mammalian cells. Mol Cell Biol. 1998;18(1):93–101.

38. Yang L, Guell M, Byrne S, Yang JL, De Los Angeles A, Mali P, et al. Optimization of scarless human stem cell genome editing. Nucleic Acids Res. 2013;41(19):9049–9061. http://dx.doi.org/10.1093/nar/gkt555

39. Chen F, Pruett-Miller SM, Huang Y, Gjoka M, Duda K, Taunton J, et al. High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. Nat Methods. 2011;8(9):753–755. http://dx.doi.org/10.1038/nmeth.1653

40. Zheng Q, Cai X, Tan MH, Schaffert S, Arnold CP, Gong X, et al. Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells. BioTechniques. 2014;57(3):115–124.

41. Mussolino C, Morbitzer R, Lutge F, Dannemann N, Lahaye T, Cathomen T. A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic Acids Res. 2011;39(21):9283–9293. http://dx.doi.org/10.1093/nar/gkr597

42. Gollob JA, Wilhelm S, Carter C, Kelley SL. Role of Raf kinase in cancer: Therapeutic potential of targeting the Raf/MEK/ERK signal transduction pathway. Semin Oncol. 2006;33(4):392–406. http://dx.doi.org/10.1053/j.seminoncol.2006.04.002

43. Loupakis F, Ruzzo A, Cremolini C, Vincenzi B, Salvatore L, Santini D, et al. *KRAS* codon 61, 146 and *BRAF* mutations predict resistance to cetuximab plus irinotecan in *KRAS* codon 12 and 13 wild-type metastatic colorectal cancer. Brit J Cancer. 2009;101(4):715–721. http://dx.doi.org/10.1038/sj.bjc.6605177

44. Heinemann V, Stintzing S, Kirchner T, Boeck S, Jung A. Clinical relevance of *EGFR*- and *KRAS*-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR. Cancer Treat Rev. 2009;35(3):262–271. http://dx.doi.org/10.1016/j.ctrv.2008.11.005

45. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31(9):827–832. http://dx.doi.org/10.1038/nbt.2647

46. Radecke F, Peter I, Radecke S, Gellhaus K, Schwarz K, Cathomen T. Targeted chromosomal gene modification in human cells by single-stranded oligodeoxynucleotides in the presence of a DNA double-strand break. Mol Ther. 2006;14(6):798–808. http://dx.doi.org/10.1016/j.ymthe.2006.06.008

47. Wang RC, Smogorzewska A, De Lange T. Homologous recombination generates T-loop-sized deletions at human telomeres. Cell. 2004;119(3):355–368. http://dx.doi.org/10.1016/j.cell.2004.10.011

Towards polarisation-encoded quantum key distribution in optical fibre networks

AUTHORS:
Sharmini Pillay[1]
Abdul R. Mirza[1,2]
Francesco Petruccione[1,2,3]

AFFILIATIONS:
[1]Quantum Research Group, School of Chemistry and Physics, University of KwaZulu-Natal, Durban, South Africa
[2]QZN Technology – Innovation Centre, University of KwaZulu-Natal, Durban, South Africa
[3]National Institute for Theoretical Physics, Durban, South Africa

CORRESPONDENCE TO:
Sharmini Pillay

EMAIL:
sharminipillay251@gmail.com

POSTAL ADDRESS:
Quantum Research Group,
School of Chemistry and Physics,
University of KwaZulu-Natal,
Private Bag X54001, Durban
4000, South Africa

DATES:
Received: 11 Dec. 2013
Revised: 25 May 2014
Accepted: 21 Sep. 2014

KEYWORDS:
QKD; cryptography; polarisation encoding; fibre network; single photon

HOW TO CITE:
Pillay S, Mirza AR, Petruccione F. Towards polarisation-encoded quantum key distribution in optical fibre networks. S Afr J Sci. 2015;111(7/8), Art. #2013-0380, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20130380

Quantum key distribution – a process that encodes digital information – often utilises fibre optic technologies for commercial applications. Fibre provides the benefit of a dark channel as well as the convenience of independence of a line-of-sight connection between the sender and receiver. In order to implement quantum key distribution protocols utilising polarisation encoding, the birefringence effects of fibre must be compensated for. Birefringence is caused by manufacturing impurities in the fibre or a change in environmental conditions and results in a rotation of the state of polarisation of light as it is propagated through the fibre. With dynamic environmental conditions, the birefringence effects should be monitored with a test signal at regular time intervals so that the polarisation of each photon can be appropriately compensated to its original state. Orthogonal states are compensated simultaneously, but most protocols, such as BB84 and B92, require non-orthogonal basis sets. Instead of using a compensator for each basis, the presented scheme fixes the polarisation controller onto the plane on the Poincaré that passes through both bases, compensating both non-orthogonal bases simultaneously.

Introduction

The reliance on information technology for global communication has highlighted the need for data security in recent years. Various applications such as online banking and government communications require a secure transmission between the transmitter and the intended recipient. Physical protection of the data, such as a lock and key transfer, is not feasible in terms of the cost and time implications. Further physical protection also involves an element of human interaction which may compromise the security of the information. It is therefore essential to develop reliable and secure cryptographic systems (cryptosystems) to encrypt sensitive data. Cryptography allows for information to be encrypted into an unintelligible state so that it may be transmitted across public networks without any risk of eavesdropping.

Quantum key distribution (QKD) encodes information into the physical properties of quantum particles, e.g. single photons of light. The Uncertainty Principle[1] and the No Cloning Theorem[2] of quantum physics, rather than the finite complexity of a mathematical algorithm, ensure the security of the key. The information shared between the transmitter (usually referred to as Alice) and the receiver (usually referred to as Bob) is carried by qubits (quantum bits).[3] The qubit is a quantum two-level system used to encode binary data. The state of the qubit, $|\Psi\rangle$, is represented as a linear superposition of two pure states:

$$|\Psi\rangle = \alpha|\,0\rangle + \beta|1\rangle. \qquad \text{Equation 1}$$

The basis set, $\{|0\rangle\,,|\,1\rangle\}$, represents the two eigenstates in which a qubit may be measured. The probability of a measurement in these respective states is given by $|\alpha|^2$ and $|\beta|^2$, such that

$$|\alpha|^2 + |\beta|^2 = 1. \qquad \text{Equation 2}$$

Information may also be encoded in any other basis set, in particular:

$$|\pm\rangle = \frac{1}{\sqrt{2}}\,(|0\rangle \pm |1\rangle\,). \qquad \text{Equation 3}$$

It is noted that $|+\rangle$ and $|-\rangle$ are orthogonal to each other but the set $\{|+\rangle\,,|\text{-}\}$ is non-orthogonal to the set $\{|0\rangle\,,|1\rangle\}$. For a practical implementation, these two bases can be represented by the rectilinear states of polarisation (SOP) and the diagonal SOPs, as applied to the BB84 and SARG04 protocols.[4,5] Because the bases are non-orthogonal to each other, any measurements carried out with the incorrect measurement basis will yield an ambiguous result.[2] The security of QKD therefore lies in the inability to gain total information from the qubit that Alice has sent to Bob without first knowing which basis to use for the measurement.

Another advantage that QKD has over conventional cryptography is the ability for the authenticated parties to detect an eavesdropper. The eavesdropper, usually referred to as Eve, may attempt to copy or measure the quantum state of the qubits. However, both these attacks violate the laws of quantum mechanics that fortify QKD. Any measurements done on the qubits during transmission will cause disturbances to the quantum states which can be observed by Alice and Bob. Therefore, if Alice and Bob observe an error rate above the accepted threshold in their transmission, they may infer the presence of Eve. As QKD exploits the physical quantum nature of particles, the principle is not vulnerable to technological advances but bound only by the laws of physics.[6]

As a consequence of the intrinsic attenuation of optical fibre, a fibre optic link is impractical for single photon transmission over 200 km.[7] However, a free space channel can support transmission over longer distances than fibre, because of a lower attenuation and a weaker dispersion property.[8] Additionally, a free space channel does not require the information and communication technology infrastructure required for a fibre network and is therefore suitable for remote locations. These advantages allow the possibility of a satellite QKD network which opens the door for a *global* QKD network. Creating an interface between a fibre channel and a free space channel will allow for fibre-based municipal networks to connect to a local free space node. The free space nodes can transmit to a satellite network and thereby connect to another metropolitan network at another location, shown in Figure 1.

Such a methodology will bring together the advantages of a fibre network, such as a higher signal-to-noise ratio, and the larger transmission distance of a free space network and would facilitate various media of communication in a combined network. Ideally, such free-space–fibre gateways within the global network must remain untrusted. This proves difficult because a fibre network is better suited for phase-encoded QKD and a free space network to polarisation-encoded QKD. A relay device that can convert the quantum systems between various encoding schemes without actually measuring the systems would require the development of quantum repeaters and quantum memories. However, these devices are still under research.[9]

It is therefore necessary to create a passive interface between the fibre channel and the free space channel. In order to achieve this interface, only one type of encoding can be used throughout the entire network. Phase encoding can be used efficiently in a fibre, plug-and-play system but in a free space system, the relative stability between the interferometers can be difficult to implement. Research is currently being done to allow for QKD to be implemented over free space using Laguerre–Gauss modes. These modes carry orbital angular momentum (OAM) which is used as the bit encoding for QKD.[10] The advantage of using OAM states is their infinite dimensional Hillbert space which can be used for encoding instead of the two-dimensional Hillbert space used in other QKD implementations. However, coupling such a system to a fibre network would pose a problem because OAM states would require multimode fibre. Because of the negative effects of modal dispersion, only single-mode fibre is used for QKD purposes. Alternatively, polarisation encoding can be used over both channels. As the atmosphere is non-birefringent, it is more suitable for polarisation encoding. However, the birefringence of a standard single-mode fibre optic cable renders it unable to maintain the SOP of the light it transmits. By cancelling birefringent effects in fibre, an untrusted interface can be developed between a fibre channel and a free space channel.

Birefringence in a fibre optic cable

Birefringence refers to the double refraction of light when transmitted through an anisotropic medium. Orthogonal components of the state of polarisation of light are transmitted through the medium at different speeds, referred to as the differential group delay.[11] The component that is perpendicular to the optical axis of the medium is the ordinary ray and the component that is parallel to the optical axis of the medium is the extraordinary ray.[12] The refractive differences between the ordinary ray and the extraordinary ray cause a decoupling of the components, resulting in the rotation of the state of polarisation as the light is transmitted through the material, as shown in Figure 2.

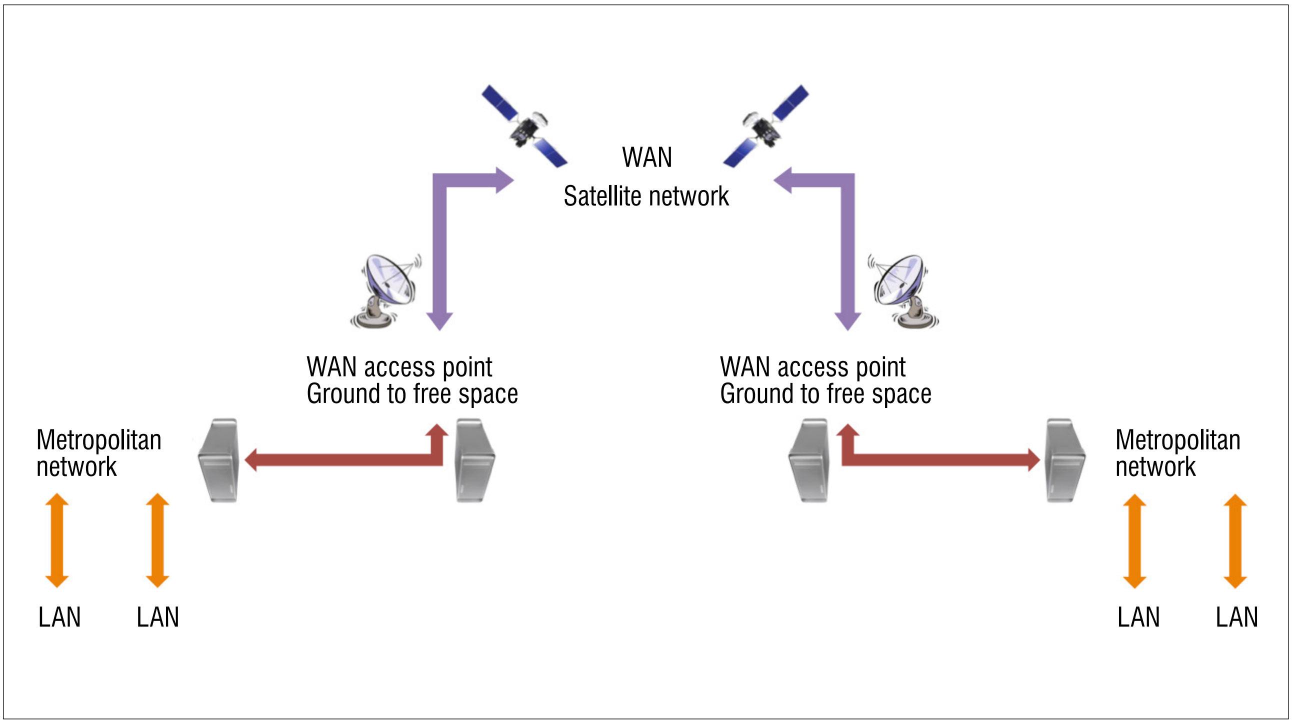
Birefringence occurs in fibre optic cables as a result of asymmetries caused by impurities in the fibre or manufacturing errors.[13] These irregularities cause a fixed rotation, *A*, of any state of polarisation, E_i, that is transmitted through the fibre. This rotation can be corrected with the use of a passive polarisation controller, represented by the inverse of *A,* such that

$$E_i = A\,A^{-1}\,E_i. \qquad \text{Equation 4}$$

The polarisation controller therefore applies the inverse of the rotation caused by the birefringence effects, returning the state of polarisation to its original form.

If the fibre is bent or subject to changing environmental stresses, such as heating or vibrations, the birefringent effects will vary randomly with time.[14] Therefore, an active polarisation controller must be used to correct for the changes in the state of polarisation of photons in real time.[15] The effects of the fibre's birefringence must be regularly tested and the polarisation controller must be adjusted each time in order to compensate for these changes. The state of polarisation of each qubit must be accurately transmitted between Alice and Bob in order for them to obtain a cryptographic key; therefore, without this active polarisation control, it would not be possible to implement polarisation-encoded QKD protocols over a fibre channel.

One method of solving the problem of birefringence in fibre would be to use polarisation-maintaining (PM) fibre in the QKD set-up. PM fibre induces a forced and fixed birefringence of any transmitted light.[14] This effect prevents the SOP from rotating as a result of any natural effects such as bends and temperature gradients. However, PM fibre is only



WAN, wide area network; LAN, local area network.

Figure 1: A diagram depicting a potential set-up for a global quantum key distribution network.

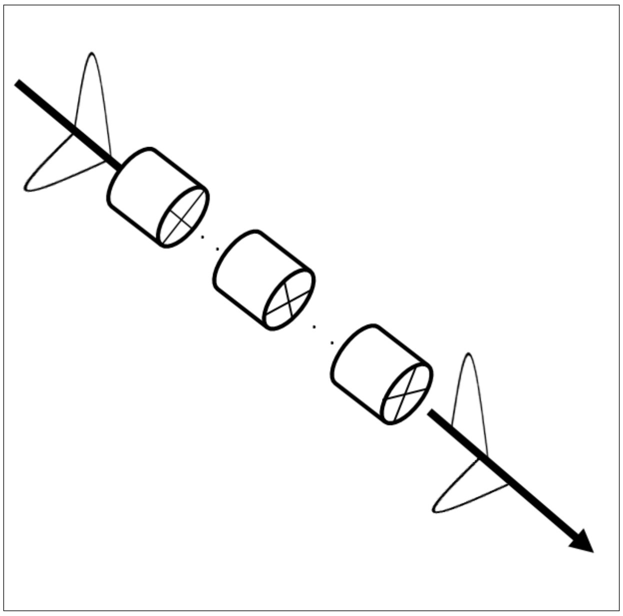


Figure 2: Diagram depicting how a state of polarisation is rotated as it is transmitted through a fibre optic cable. The cross sections of the fibre show that the orthogonal components of the state of polarisation are rotated during transmission as a result of their differing speeds. For illustrative purposes, this figure demonstrates a linear rotation, but practically, the fibre can introduce an elliptical element into the state of polarisation as well.

effective for orthogonal states because the SOPs must be aligned with the fast axis and slow axis of the PM fibre. The necessary use of non-orthogonal states in QKD means the use of PM fibre in the channel is obsolete.

The SOP of each photon in the quantum signal is assigned randomly and, therefore, the polarisation controller must adjust each photon uniquely. The SOP of each photon must remain unknown to the polarisation controller so as to maintain the security of the QKD. So as not to violate the security of QKD, no measurements can be made on the SOPs of the photons before the final detection system. The setting for the polarisation controller must be determined before the commencement of the QKD transmission and any adjustments made to the SOPs of photons must be done passively. Test pulses can be propagated through the system, and by compensating the test pulses, the same setting of the polarisation controller will passively compensate the single photon signal.

Polarisation compensation was first addressed by Breguet et al.[15] using two quarter-wave plates to minimise any ellipticity in the SOP introduced by the fibre channel. The orientation axis of the polarisation beam splitter was then adjusted in order to correct for any linear rotations. Automated polarisation compensation in a fibre optic QKD system has been addressed in recent literature. Xavier et al.[16] developed a wavelength division multiplexed (WDM) compensation system using a wavelength separation of 0.8 nm between the reference signal and the quantum channel. The reference signal was counter-propagated through a channel of 8.5 km to reduce the interference to the signal. Two piezoelectric controllers were used to compensate the SOPs in this set-up. A lithium-niobate controller was later used to improve the speed of the system.[17] To test the effectiveness of the system, a polarisation scrambler was introduced into the 16-km quantum channel and the lithium-niobate controller proved effective in controlling fast fluctuations in SOP.

Wu et al.[18] developed a one-way time division multiplexed (TDM) compensation system. In this case, the signal was periodically stopped to allow the polarisation controllers to reset. Two polarisation controllers were used; each controller compensated one of the non-orthogonal bases. The interval time required to reset the controllers varied for the different lengths of the quantum channel, i.e. 50 km, 75 km and 100 km. Chen et al.[19] designed a real-time TDM system which alternated signal pulses and reference pulses using an asymmetric Mach-Zehnder interferometer. A 50-ns time delay was introduced between the signal and the reference and a 50/50 beam splitter was used to separate these signals before reaching the detector. The polarisation compensation was executed using fibre squeezers, controlled by an electronic polarisation controller. The channel length used for this system was 50 km. Ma et al.[20] also developed TDM systems using liquid crystal retarders and piezoelectric controllers. The quantum signal was terminated every 15 min to allow the test signal through the channel. The extinction ratio of the signal was measured and if it was found to be below the accepted threshold, the compensation process was initiated. A step search was utilised to adjust the polarisation controllers and measure the extinction ratio of the signal until the correct settings were obtained. The piezoelectric controllers proved easier to operate because they did not have to be pre-aligned with the polarisation beam splitter. The system presented in this paper improves on the above-mentioned TDM systems by using just one piezoelectric polarisation controller, thereby decreasing both the size and cost of the system.

As mentioned above, a wavelength division multiplexed compensation system can be used to introduce a test signal to the QKD system, but, because of the wavelength-dependence of birefringence effects in fibre, the method of TDM compensation was chosen for this experiment. An SOP stability test was done under laboratory conditions to monitor the change in SOP as a result of standard environmental conditions. The average time for an SOP to change such that the probability of measurement in the correct basis drops by 1% was 332.75 s. Therefore, the test signal must be deployed at least every 333 s in order to ensure an acceptable quantum bit error rate. Of course each environment is different and in order to effectively use the TDM method, an SOP stability test must be done for each new environment in which the system is set up. The duration of the test signal is dependent on the resolution of the compensator used for the system.

Experimental set-up

The experimental set-up is shown in Figure 3. The components for the transmitter in this experiment comprised a pseudo-single photon source and a polarisation state generator. The pseudo-single photon source was provided by a pulsed laser source with a wavelength of 1550 nm, attenuated to simulate the power of a single photon per pulse. The laser pulses were then randomly polarised by a polarisation state generator. The photons were then transmitted through a fixed 1000 m of fibre which served as the quantum channel. The fibres used for the channel and the patchcords were single-mode fibres with a core diameter of 6 μm, unless otherwise stated as PM fibres.

The birefringence effects of the fibre on the SOP of the single photons were then corrected by a polarisation compensator. A half-wave plate was installed after the length of fibre, before the photons were transmitted to a polarisation beam splitter. The half-wave plate served as a means to change the measurement basis of the beam splitter. The fibre between the output of the polarisation compensator and the polarisation beam splitter must be PM fibre to ensure that the state of polarisation is not changed after the compensator makes the necessary corrections. After the polarisation beam splitter, the photons were directed to one of two single photon detectors.

A polarisation locker was used as the automated polarisation controller.[21] This device is a fibre-based controller which includes many internal piezoelectric polarisation controllers driven by varying voltages. Piezoelectric controllers squeeze the fibre optic cables in order to induce a controlled birefringence. An in-line polarimeter and digital signal processor form an internal feedback loop which drives the piezoelectric controllers to produce deterministic SOPs.[21] This device can be pre-programmed so that all output SOPs are adjusted to a constant SOP chosen by the user. The in-line polarimeter monitors the output SOPs and transmits the adjustment commands to the polarisation controller using the feedback loop, enabling the polarisation locker to ‘lock’ onto a predetermined SOP. The locker can also be operated in manual mode to incrementally adjust the SOP along a grid superimposed onto the Poincaré sphere.

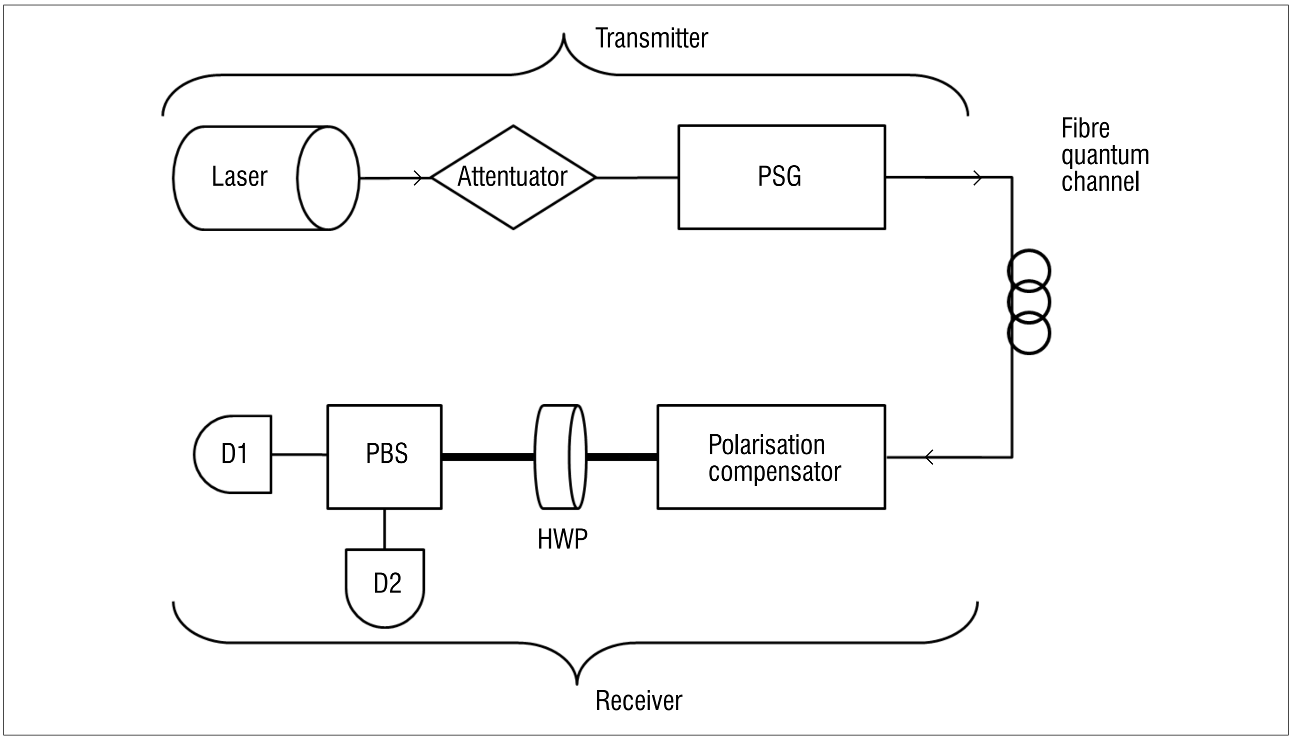


Figure 3: The proposed set-up for a polarisation-encoded quantum key distribution scheme. The laser pulses are first passed through an optical attenuator, which creates pseudo-single photons. Each photon is then assigned a state of polarisation with the polarisation state generator (PSG) and is transmitted through the quantum channel to the receiver. The receiver then uses the polarisation compensator to correct for changes in polarisation. A half-wave plate (HWP) is used to select the basis in which the receiver will measure each photon and, finally, the photons are separated at a polarisation beam splitter (PBS) to be measured at one of two detectors (D1 or D2). The bold lines in the diagram indicate polarisation-maintaining fibre.

Using these functions, the polarisation locker can be employed as an automated compensator for the experimental set-up.

Compensating a single basis

When two orthogonal SOPs undergo the same transformation using a phase retarder, the resulting vectors will also be orthogonal. As an example, a rotation matrix is applied to the Jones vector of a vertical SOP in Equation 5 and its orthogonal state, a horizontal SOP, in Equation 6:

$$\begin{bmatrix}\cos\beta & -\sin\beta \\ \sin\beta & \cos\beta\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix}=\begin{bmatrix}\cos\beta\\ \sin\beta\end{bmatrix} \qquad \text{Equation 5}$$

$$\begin{bmatrix}\cos\beta & -\sin\beta \\ \sin\beta & \cos\beta\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix}=\begin{bmatrix}-\sin\beta\\ \cos\beta\end{bmatrix} \qquad \text{Equation 6}$$

The two resulting states are still orthogonal to each other. A similar result can be achieved using the Jones matrix of a quarter-wave plate as well as other sets of orthogonal states.

It is inferred that if a polarisation controller is used to compensate for changes in the vertical SOP, the horizontal SOP will be simultaneously compensated. A similar result is obtained for the diagonal basis. Figure 4a demonstrates the compensation of a vertically polarised signal. The signal was then rotated using a free space polariser such that all linear SOPs were incident on the fibre channel. These measurements show that both the vertical and horizontal states are conserved but all states that were non-orthogonal to the rectilinear basis were not compensated. Figure 4b shows a similar result for diagonal SOPs.

Compensating non-orthogonal bases

The presented scheme, shown in Figure 3, utilises just one polarisation controller, i.e. the polarisation locker, in a TDM system. The single photon signal is periodically stopped in order to deploy a test signal through the quantum channel and apply the appropriate settings to the locker. The test signal propagates a single SOP through the fibre and the locker is set to return the test signal to its original state after its rotation in a birefringent medium. For a vertically polarised test signal, this initial setting would also compensate the horizontal SOP. The same polarisation locker must be used to compensate the diagonal SOPs as well. The locker must therefore isolate the plane on the Poincaré sphere that passes through all four SOPs used in the QKD transmission.

In order to do this, a step search may be utilised for the locker to find the plane on which all four SOPs exist. This is done by rotating the test signal to simulate all linear SOPs and measuring any ellipticity induced by the fibre channel. The locker must be incrementally adjusted in order to eliminate the ellipticity of the test signal, thus returning all SOPs back to their original linear states on the equatorial plane of the Poincaré sphere. This method has been implemented manually for a four-state protocol and the results are shown in Figure 5. In practice, any number of states that form a plane on the Poincaré sphere may be compensated through this technique.

Analysis

The polarisation locker proved effective in reversing the birefringence effects of the fibre channel. The measurements shown in Figure 5 indicate that the polarisation locker can easily be used to minimise the elliptical component of the SOP of any incoming light as well as rotate the SOP linearly so that it returns to its original state. The locker is therefore able to passively compensate all four SOPs simultaneously. Because the polarisation locker is able to lock onto any plane on the Poincaré sphere, the same method can be used to compensate the rectilinear and circular bases as well. The range of the angle of inclination obtained from the above measurements was between −0.89° and 1.69°, keeping the error rate as a result of the polarisation locker below 0.1%. The angle of inclination may deviate by 5.74° if an error of 1% is allowed. The maximum deviation of the ellipticity of the SOP may therefore be set according to the expected quantum bit error rate of the entire system.

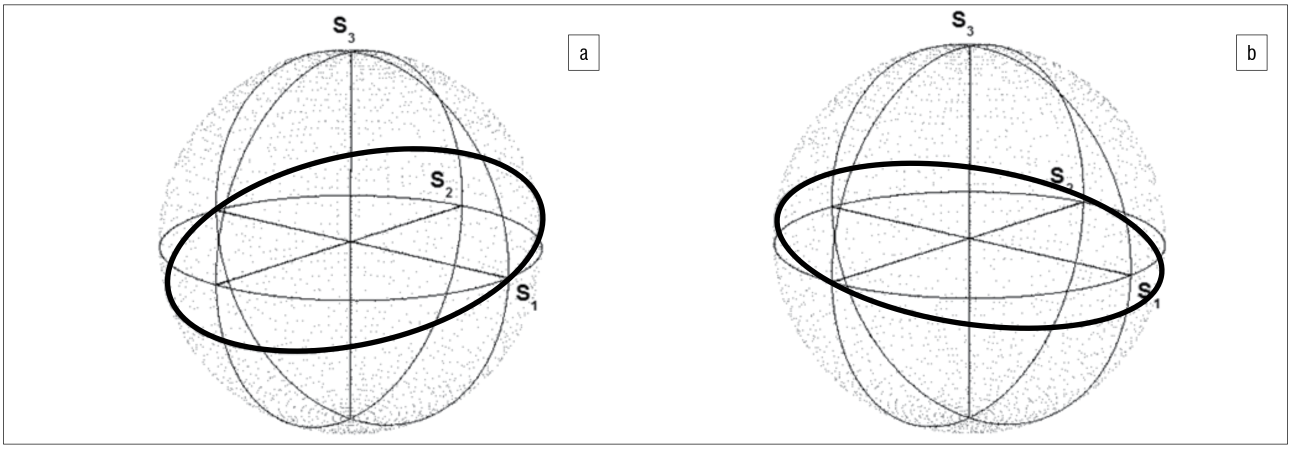


Figure 4: In (a), the bold line shows the measured states of polarisation (SOPs) of the test signal on the Poincaré sphere. The S1 axis represents the rectilinear basis, the S2 axis represents the diagonal basis and the S3 axis represents the circular basis. The measurements show that both the vertical and horizontal SOPs were returned to their original states, even though only the vertical SOP was compensated in this case. Similarly, in (b), both diagonal SOPs were corrected, even though only one of them was compensated for.

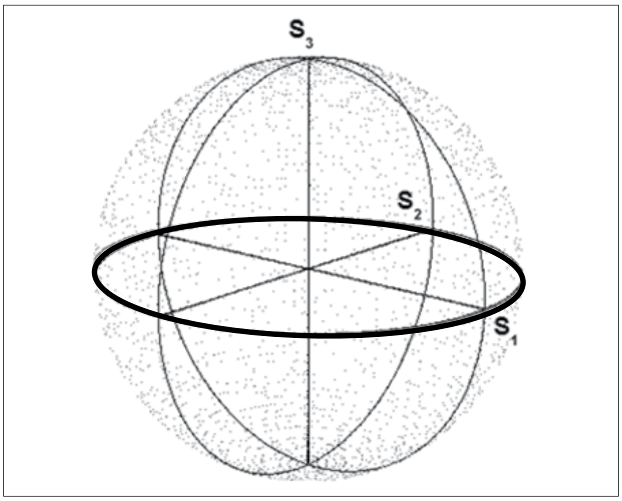


Figure 5: This figure shows the compensation of all four states used in the quantum key distribution protocol using one polarisation controller. The S1 axis represents the rectilinear basis, the S2 axis represents the diagonal basis and the S3 axis represents the circular basis. The bold line in the figure, indicated on the equator of the Poincaré sphere, shows that only linear states of polarisation were measured, thus eliminating any ellipticity induced in the fibre channel.

Conclusion

Birefringence in fibre optic cables creates a bottleneck for polarisation-encoded QKD in metropolitan fibre networks. The negative effects of birefringence were compensated for by monitoring the changes in SOP in real time using a TDM test signal. A polarisation locker was implemented to compensate the states of two non-orthogonal bases simultaneously. A search algorithm can be used to identify the plane on the Poincaré sphere on which both the non-orthogonal bases are located, thereby enabling the compensation of birefringence effects with just one device. Future work will focus on automating the search algorithm and integrating it into the QKD system so that polarisation-encoded QKD can be implemented in fibre.

Acknowledgements

This paper was developed and extended from an earlier conference proceeding.[22] This work is based on research supported by the South African Research Chair Initiative of the Department of Science and Technology and the National Research Foundation of South Africa. We thank Dr Timothy Gibbon for his assistance and guidance with the fibre optics laboratory work.

Authors' contributions

S.P. was responsible for the experimental conception and design, acquisition of data, analysis and interpretation of data and drafting of the manuscript. A.R.M. was responsible for the experimental conception and design, analysis and interpretation of data and critical review of the manuscript. F.P. was the project leader and was responsible for the study conception and design as well as the critical review of the manuscript.

References

1. Zettili N. Quantum mechanics: Concepts and applications. West Sussex, UK: John Wiley & Sons; 2009.
2. Wooters WK, Zurek WH. A single quantum cannot be cloned. Nature. 1982;299:802–803. http://dx.doi.org/10.1038/299802a0
3. Nielsen MA, Chuang IL. Quantum information processing and communication. Cambridge: Cambridge University Press; 2002.
4. Bennett C, Brassard G. Quantum cryptography: Public key distribution and coin tossing. Theor Comput Sci. 2014;560:7–11. http://dx.doi.org/10.1016/j.tcs.2014.05.025
5. Scarani V, Acin A, Ribordy G, Gisin N. Quantum cryptography protocols robust against photon number splitting attacks for weak laser pulse implementations. Phys Rev Lett. 2004;92(5):57901. http://dx.doi.org/10.1103/PhysRevLett.92.057901
6. Gisin N, Ribordy G, Tittel W, Zbinden H. Quantum cryptography. Rev Mod Phys. 2002;74:145–195. http://dx.doi.org/10.1103/RevModPhys.74.145
7. Brassard G, Lütkenhaus N, Mor T, Sanders BC. Limitations on practical quantum cryptography. Phys Rev Lett. 2000;85(6):1330–1333. http://dx.doi.org/10.1103/PhysRevLett.85.1330
8. Schmitt-Manderbach T, Weier H, Furst M, Ursin R, Tiefenbacher F, Scheidl T, et al. Experimental demonstration of free-space decoy-state quantum key distribution over 144 km. Phys Rev Lett. 2007;98(1):10504. http://dx.doi.org/10.1103/PhysRevLett.98.010504
9. Clausen C, Usmani I, Bussieres F, Sangouard N, Afzelius M, De Riedmatten H, et al. Quantum storage of photonic entanglement in a crystal. Nature. 2011;469(7331):508–511. http://dx.doi.org/10.1038/nature09662
10. Pampaloni F, Enderlein J. Gaussian, Hermite-Gaussian, and Laguerre-Gaussian beams: A primer [article on the Internet]. c2004 [cited 2013 Dec 11]. Available from: arXiv:physics/0410021v1 [physics.optics].
11. Ramaswami R, Sivarajan K. Optical networks: A practical perspective. 2nd ed. San Francisco, CA: Morgan Kaufmann; 2002.
12. Hecht E. Optics. Reading, MA: Addison-Wesley; 2001. p. 325–379.
13. EXFO. Application note – Polarization mode dispersion. Available from: documents.exfo.com/appnotes/anote047-ang.pdf.

14. OZ Optics. Application note - Polarization measurements [document on the Internet]. c1999 [cited 2013 Dec 11]. Available from: www.ozoptics.com/ALLNEW_PDF/APN0005.pdf.

15. Breguet J, Muller A, Gisin N. Quantum cryptography with polarized photons in optical fibres. J Mod Optic. 1994;41(12):2405–2412. http://dx.doi.org/10.1080/09500349414552251

16. Xavier GB, Vilela de Faria G, Temporao GP, Von der Weid JP. Full polarization control for fiber optical quantum communication systems using polarization encoding. Opt Express. 2008;16(3):1867–1873. http://dx.doi.org/10.1364/OE.16.001867

17. Xavier GB, Walenta N, Vilela de faria G, Temporao GP, Gisin N, Zbinden H, et al. Experimental polarization encoded quantum key distribution over optical fibres with real-time continuous birefringence compensation. New J Phys. 2009;11(4):045015. http://dx.doi.org/10.1088/1367-2630/11/4/045015

18. Wu G, Chen J, Li Y, Zeng H. Stable polarization-encoded quantum key distribution in fibre [article on the Internet]. c2006 [cited 2013 Dec 11]. Available from: http://arxiv.org/abs/quant-ph/0606108

19. Chen J, Wu G, Xu L, Gu X, Wu E, Zeng H. Stable quantum key distribution with active polarisation control based on time-division multiplexing. New J Phys. 2009;11(6):065004. http://dx.doi.org/10.1088/1367-2630/11/6/065004

20. Ma L, Xu H, Tang X. Polarization recovery and auto-compensation in quantum key distribution network. Gaithersburg, MD: National Institute of Standards and Technology; 2006. http://dx.doi.org/10.1117/12.679575

21. Thorlabs. Benchtop state of polarization locker [homepage on the Internet]. No date [cited 2013 Dec 11]. Available from: http://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=1769

22. Pillay S, Mirza AR, Gibbon TB, Petruccione F. Compensating birefringence effects in optical fibre for polarisation encoded QKD. In: Janse van Rensburg J, editor. Proceedings of SAIP2012: The 57th Annual Conference of the South African Institute of Physics; 2012 July 9–13; Pretoria, South Africa. Pretoria: University of Pretoria; 2014. p. 288–292. Available from: http://events.saip.org.za

A multiphysics simulation of a fluorine electrolysis cell

AUTHORS:
Ryno Pretorius[1]
Philippus L. Crouse[1]
Christiaan J. Hattingh[2]

AFFILIATIONS:
[1]Department of Chemical Engineering, University of Pretoria, Pretoria, South Africa
[2]Metallurgical Testing and Consulting (MTC) cc, Benoni, South Africa

CORRESPONDENCE TO:
Ryno Pretorius

EMAIL:
RPretorius.chemeng@gmail.com

POSTAL ADDRESS:
Department of Chemical Engineering, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

DATES:
Received: 05 Mar. 2014
Revised: 16 July 2014
Accepted: 21 Sep. 2014

KEYWORDS:
fluorine reactor; two-phase simulation; comparative study; coupled transfer process

HOW TO CITE:
Pretorius R, Crouse PL, Hattingh CJ. A multiphysics simulation of a fluorine electrolysis cell. S Afr J Sci. 2015;111(7/8), Art. #2014-0082, 5 pages. http://dx.doi.org/10.17159/sajs.2015/20140082

We modelled a laboratory-scale fluorine reactor which employed fully coupled, fundamental electron, heat, mass and momentum transfer (two-phase) equations to deliver a transient simulation. Hydrodynamic quasi-steady-state results were produced for the current density, electric field, temperature, reactive species concentration, gas and liquid velocity profiles as well as gas fraction distribution within the reactor. Simulation results were verified by modelling and comparing models from published works on similar reactors, as the laboratory-scale reactor is still in construction phase. Comparisons were favourable.

Introduction

Industrial manufacture of fluorine requires the extraction of hydrogen fluoride from fluorspar, the electrolysis of hydrogen fluoride to form fluorine gas, and, finally, purification by a separation step.[1] Moissan was the first to produce fluorine gas via electrolysis.[2] His original cell has been refined over the years, but the fundamental operating principles have not changed much.[3]

Very little is currently known, at least in the open literature, about the hydrodynamic behaviour of fluorine electrolysers. Experimental studies are difficult and dangerous, because they involve corrosive chemicals, at elevated temperatures, and high electric currents. A theoretical study was therefore commissioned to better understand typical cell operation. The hydrodynamic behaviour inside a typical two-electrode reactor was mathematically modelled using applicable computational fluid dynamics simulation software (COMSOL Multiphysics®). The reactor will be built and studied using recommendations found during the simulation procedure at a later date.

Fluorine electrolysis typically operates by subjecting molten potassium–acid–fluoride (KF·xHF, $1.8 \leq x\ 2.2$) to an electric field. The potassium–fluoride matrix is required because of the low electrical conductivity of anhydrous hydrogen fluoride. Hydrogen forms at the cathode and fluorine at the carbon anode. A separator skirt prevents explosive recombination of the gaseous components.[4]

Bubble formation severely complicates cell operation and is the major source of electrolyte flow caused by gaseous convection, which in turn causes mixing of the diluted reactant species. Furthermore the high electrical resistivity of the gaseous bubbles compared to the electrolyte results in several phenomena on the electrode.[5] The anode is particularly susceptible to bubble phenomena as a consequence of the tendency of fluorine bubbles to stick to the electrode surface and move up slowly along the anode as a result of buoyancy forces.[3]

Thermodynamic hydrogen fluoride decomposition requires a potential of 2.9 V, but an anode–cathode voltage of 8–10 V is required to maintain a current density of 10–12 A/dm^2 in industrial cells.[6] The reversible cell voltage (Φ_{RV}) or thermodynamic decomposition voltage is the minimum potential required for product formation during electrolysis. Any voltage supplied that surpasses the reversible voltage (done to achieve the desired current density) produces heat through Ohmic heating.[3,7]

As a consequence of the unavailability of experimental data at simulation completion, the modelling technique was evaluated through a comparison with published fluorine cell simulations. Published models[8,9] were compared with self-produced simulacrums that we recreated under the same conditions. Further validation was received by comparing the results with other work published.[10-12]

Modelling procedure

Four coupled transfer processes were identified as critical to ensure an accurate model: electron, heat, mass and two-phase momentum transfer. Roustan and co-workers[9] modelled uncoupled electron, heat and momentum transfer (single phase), using Flux-Expert®. They investigated the two-phase momentum transfer using Flux-Expert® and Ested Astrid code using predetermined experimental values for electron and heat transfer. We attempted to encompass all of the above-mentioned transfer processes into a single coupled model. Standard fundamental transfer equations were used where possible. Widely used empirical correlations were employed where fundamental equations were not available. A detailed description of the modelling procedure as well as the equations and variables used during modelling can be found in the Online Supplementary Material.

Because of the complex coupling (complete interdependence) between the featured transfer processes, a four-step solution method was employed. In the first step, the cell potential was slowly ramped up to its final value in a time-independent calculation. This value was then used as a starting value to do a time-dependent calculation to solve heat and momentum transfer. A third calculation was done in which transient mass transfer was calculated using the values calculated in the previous calculation as a starting value. Finally, all of the values calculated above were used to calculate the transient fully coupled electron, heat, mass and momentum transfer processes up to a point at which a hydrodynamic steady state was reached.

Two meshes were used to ensure a mesh-independent solution. The first was a very fine mesh used to calculate the first three iterations. The second mesh was as fine but utilised rectangular mesh elements in which complex variable interaction occurred. These rectangular mesh elements are extremely useful in situations in which a lot of change occurs in one direction, but very little in the perpendicular direction. The meshes were employed on the electrodes and separator skirt. Mesh-dependent solutions were further deterred by ensuring that the solutions converged to a common solution with mesh refinement.

Results and discussion

For the results shown, we assume a static homogeneous molten electrolyte at simulation initiation. Time-dependent results of interest and importance are shown at 100 s after simulation (and electrochemical reaction) commenced; this point in time was identified as a hydrodynamic steady state within the reactor.

In general, arrows represented in the results indicate direction and are proportional to the norm of the vector quantity represented, at the arrow starting point. Colours indicate values as given by the legend to the right of the image.

Published result comparison

Simulation results were not justified with experimental data; instead it was decided to simulate published fluorine electrolysers while construction and commissioning of the lab-scale electrolyser was taking place. Roustan et al.'s[9] publication was used, from which electron, heat and single-phase momentum transfer could be compared. All comparisons were favourable; one such comparison is shown in Figure 1.

We used a two-phase momentum transfer equation; a suitable model comparison was therefore required and Roustan et al.'s[9] was chosen. The two-phase momentum transfer results as found by Espinasse and co-workers[8] is shown in Figure 2a. We found similar plume shapes (as shown in Figure 2b). The predicted gas fraction was, however, significantly lower, but supported by photographic evidence.[7]

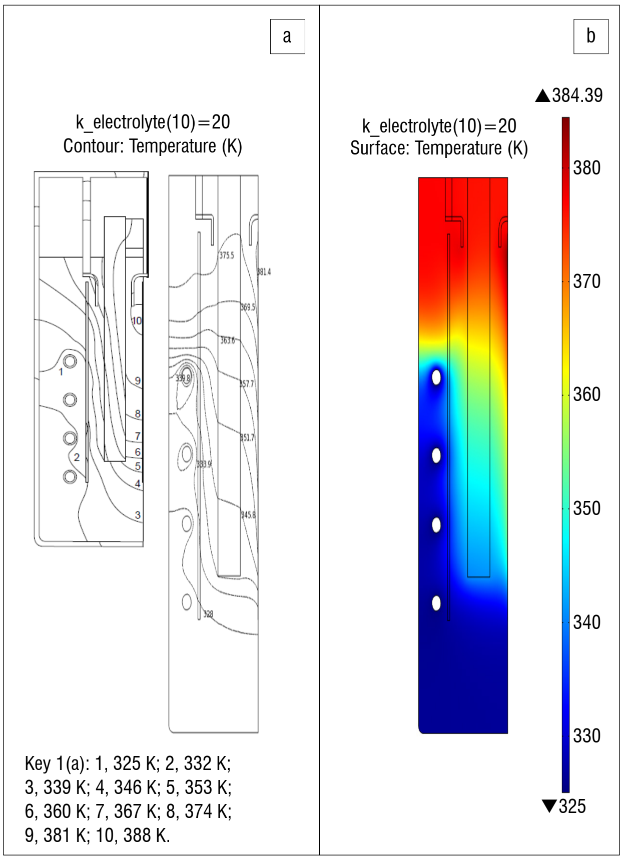


Figure 1: (a) Temperature profile within the electrolyser as simulated[1] and redrawn for this publication. (b) Temperature profiles (K) inside the reactor as simulated in this study.

Simulations

Our simulations were based on a simple lab-scale fluorine electrolyser comprised of one anode and a corresponding cathode. A cross section of the middle of the reactor was modelled and is shown in subsequent figures.

Momentum transfer

Figure 3 shows a well-developed hydrogen plume, and detachment from the cathode occurs as expected. Very little hydrogen migration into the fluorine section is observed; therefore the chance of explosive recombination of product gases is very low. This finding is also good from a productivity standpoint, as fewer product gases are lost and less purification of product streams will be required.

The electrolyte movement (Figure 4) induced by gaseous (hydrogen) movement is evidenced by the swirling liquid phase eddy between the separator skirt and the cathode at the top right of the reactor. This same eddy has the effect of causing convective stirring throughout the reactor, increasing heat and mass transfer (which in turn increases current density and reaction rate). These observations align well with what is observed in industrial and other lab-scale reactors.

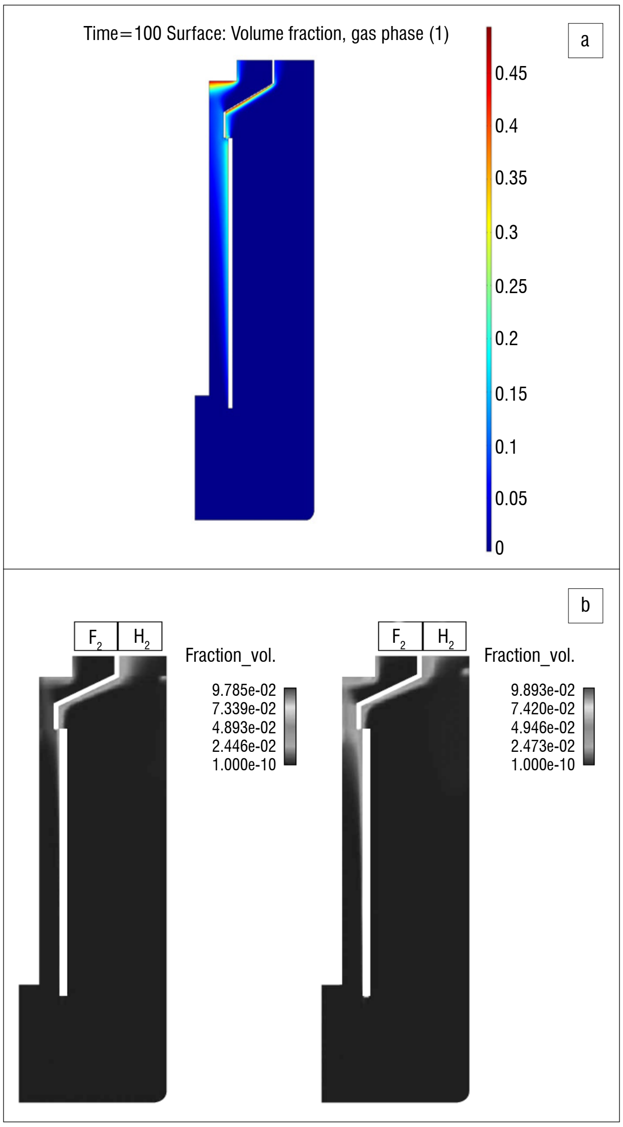


Figure 2: (a) Mean hydrogen gas distribution for two different current densities, low on the left and high on the right, reproduced from Roustan et al.[9] (b) Our simulation of the published electrolyser of Roustan et al.[9], showing relative gas fraction.

Electron transfer

Figure 5 shows the electric potential and electric potential contour lines within the electrolyte. Electric potential drops from the anode to the cathode, from 9.1 V to 0 V, as expected. This result corresponds to the potential change expected from the literature. The bending of the electric potential contour lines along the separator skirt corresponds to the electric current density field lines that bend around the skirt. Boundary effects such as bubbles cause a large electrical conductivity decrease, as a result of the low conductivity of the gaseous phase, resulting in an exponential potential drop over these bubbles on the boundaries.

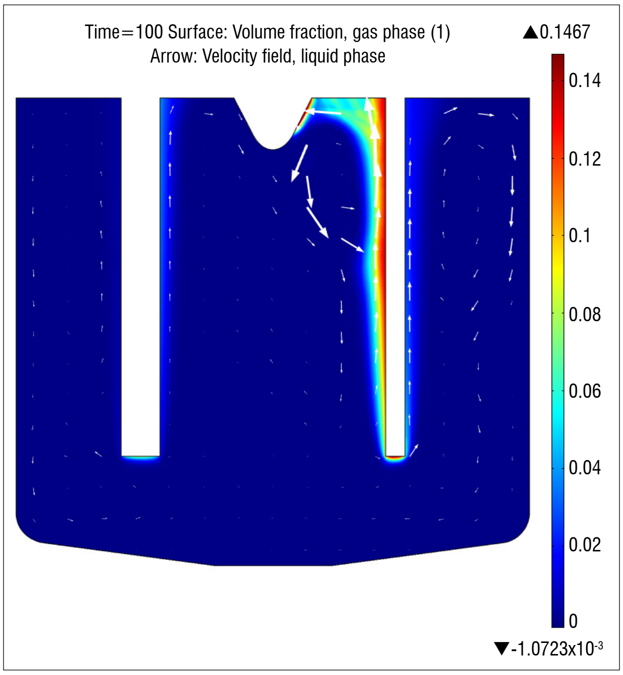


Figure 3: Hydrogen gas phase fraction in the reactor.

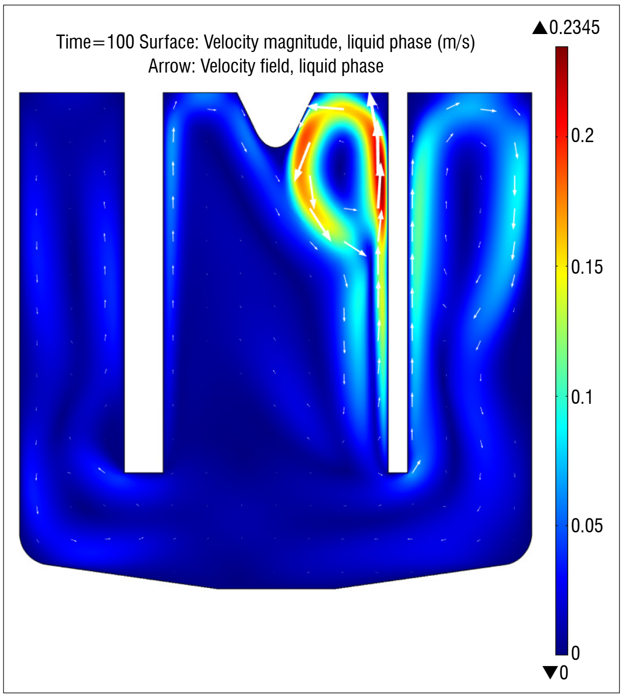


Figure 4: Liquid phase velocity inside the reactor.

Current density distribution and electric field lines within the cell are shown in Figure 6. Current density values are high at the sharp tips of the electrodes because of the small available surface area. High current density values are also visible along the separator skirt because the charged ions flow around this point to travel between electrodes.

This point is further illustrated when looking at one-dimensional current density distribution along the length of the anode, as shown in Figure 7. The points of high current density are major heat sources during electrolysis.

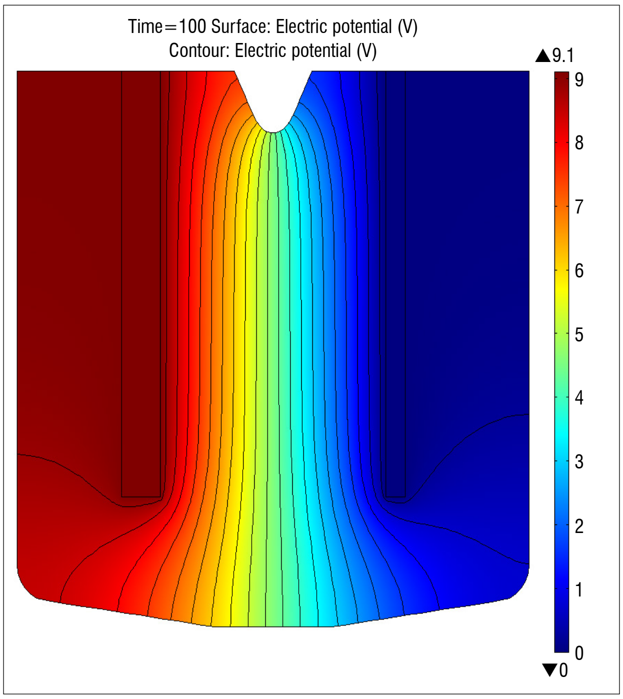


Figure 5: Inter-electrode potential variation plot.

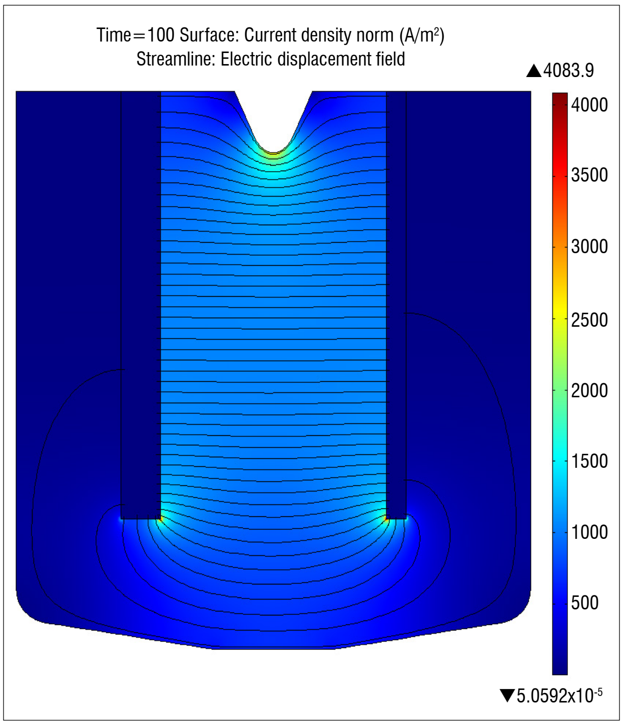


Figure 6: Current density distribution (in A/m^2) and electric field streamlines within the electrolysis cell.

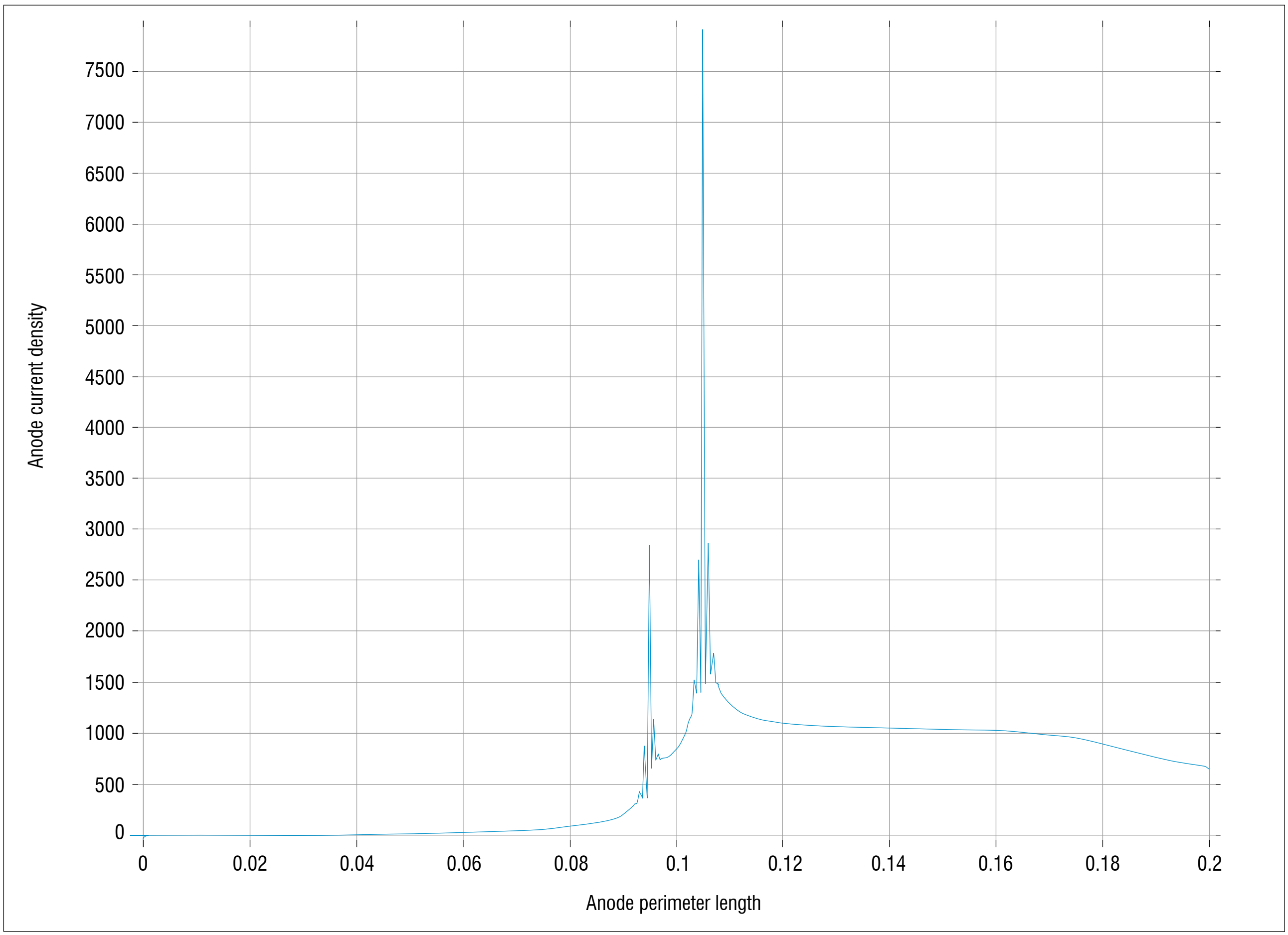


Figure 7: Current density variation (in A/m^2) along the perimeter length (in m) of the anode from the outside at electrolyte level to the inside of the anode at electrolyte level.

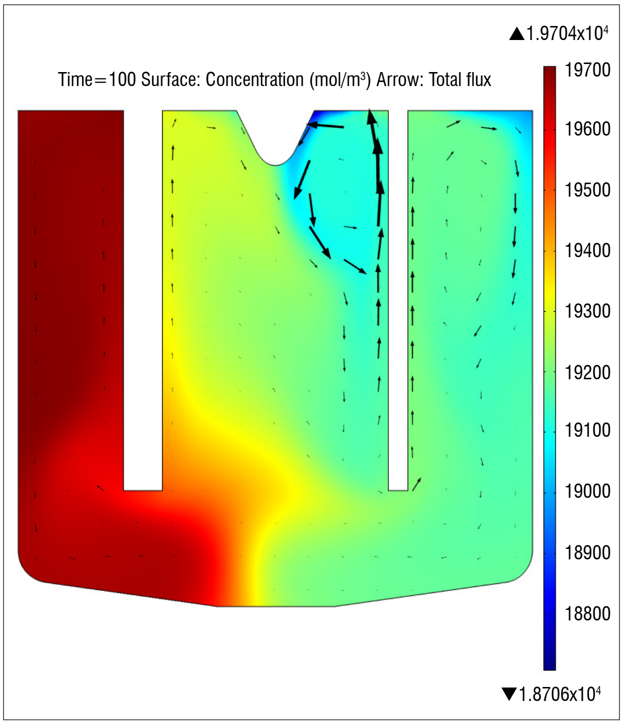


Figure 8: Dissolved hydrogen fluoride concentration and flux vectors.

Heat transfer

From the internal temperature distribution results it was found that heat flux followed the same path as the electrolyte convection path shown in Figure 4. This shows that heat convection is the dominant heat flux component that removes heat from the electrode tips and separation skirt and transfers the heat to the cooled reactor wall. A parametric study was done on the value of electrolyte thermal conductivity as it is not widely available. It was found that a thermal conductivity value of 1.25 W/(m·K) (the thermal conductivity of potassium fluoride) was more than sufficient because heat conduction is completely overshadowed by heat convection during operation. In the interests of brevity and space, heat transfer images are not shown here.

Mass transfer

HF is produced at the anode and consumed at the cathode. The concentration gradient as a result of consumption at the cathode is a contributor to flux in the form of diffusion. From the scale bar on the right, it is clear that more HF is consumed than produced, as was predicted by the electrode half reactions (Equations 10 and 11 in the Online Supplementary Material). From Figure 8, it is evident that the secondary contributors to flux are convection and migration due to electrical field.

The ion flux and concentration of HF_2^- is not shown, but is a mirror image of Figure 8. The HF_2^- ion is produced at the cathode and consumed at the anode. The concentration gradient indicates ion flux from the cathode to the anode, as expected.

Conclusions and recommendations

Results obtained from the simulations, under the quasi-steady-state assumption, are reasonable and within expectations. The simulated results show a strong correlation between the gaseous phase movement (induced by buoyancy forces) and that of the liquid phase. The gas-phase flux seen in Figure 3 shows that little or no hydrogen gas transfers to the fluorine compartment. All comparative simulations also deliver satisfactory results when compared with the published works. Current density and electric potential field line predictions correspond to expectations and match up satisfactorily with those found by Espinasse et al.[8] The shape of the gaseous plume of hydrogen that forms at the anode is the same as that reported in the literature when compared to the results from Mandin et al.[5] and Roustan et al.[9] There is, however, a difference in the qualitative gaseous fraction of the published and simulated reactors.

It is therefore recommended that the results found be used by the construction and experimental team of the physical reactor as an indication of what to expect and what to investigate. We also recommend that the physical parameters (mostly estimates) and empirical correlations, specifically the kinetics, be investigated to deliver more accurate results in future simulations.

Authors' contributions

P.L.C. was the project supervisor and academic lead at the University of Pretoria. R.P. was responsible for project design, relevant research and most of the simulation work. R.P. also presented the findings of this paper at a conference in Stuttgart, Germany. C.J.H. aided in the simulation, especially in solution finding and mathematical convergence.

References

1. Klose F. Elements and compounds, atoms and molecules, structures and bonds – Course on inorganic chemistry for the University of Magdeburg. Magdeburg: University of Magdeburg; 2004.
2. Groult H, Lantelme F, Salanne M, Simon C, Belhomme C, Morel B, et al. Role of elemental fluorine in nuclear field. J Fluor Chem. 2007;128:285–295. http://dx.doi.org/10.1016/j.jfluchem.2006.11.012
3. Rudge AJ. Production of elemental fluorine by electrolysis. In: Kuhn A, editor. Industrial electrochemical processes. Amsterdam: Elsevier; 1971. p. 1–78.
4. Shia G. Fluorine. In: Arza S, editor. Kirk-Othmer encyclopedia of chemical technology. 14th ed. New Jersey: John Wiley & Sons; 2005. p. 826–852.
5. Mandin Ph, Wüthrich R, Roustan H. Electrochemical engineering modelling of the electrodes kinetic properties during two-phase sustainable electrolysis. Paper presented at: 10th International Symposium on Process Systems Engineering; 2009 Aug 16–20; Salvador-Bahia, Brazil.
6. Groult H. Electrochemistry of fluorine production. J Fluor Chem. 2003;119:173–189. http://dx.doi.org/10.1016/S0022-1139(02)00252-X
7. Heitz E, Kreysa G. Principles of electrochemical engineering. Weinheim: VCH Verslasgesellschaft mbH; 1986.
8. Espinasse G, Peyrard M, Nicolas F, Caire JP. Effects of hydrodynamics on Faradaic current efficiency in a fluorine electrolyser. J Appl Electrochem. 2007;37:77–85. http://dx.doi.org/10.1007/s10800-006-9216-x
9. Roustan H, Caire JP, Nicolas, F, Pham P. Modelling coupled transfers in an industrial fluorine electrolyser. J Appl Electrochem. 1997;28:237–243. http://dx.doi.org/10.1023/A:1003299213119
10. Hur JS, Shin CB, Kim H, Kwonb YS. Modeling of the trajectories of the hydrogen bubbles in a fluorine production cell. J Electrochem Soc. 2003;150(3):D70–D78.
11. Newman JS. Electrochemical systems. New Jersey: Prentice Hall; 1991.
12. Nierhaus T. Two-phase flow transport phenomena in electrochemical processes [PhD thesis]. Brussels: Vrije Universiteit Brussel; 2009.

Note: This article is supplemented with online only material.

Climate change trends and environmental impacts in the Makonde Communal Lands, Zimbabwe

AUTHORS:
Ishumael Sango[1]
Nhamo Godwell[2]

AFFILIATIONS:
[1]Institute of Urban Development Studies, Ethiopian Civil Service University, Addis Ababa, Ethiopia
[2]Institute for Corporate Citizenship, University of South Africa, Pretoria, South Africa

CORRESPONDENCE TO:
Ishumael Sango

EMAIL:
ishsango@gmail.com

POSTAL ADDRESS:
Institute of Urban Development Studies, Ethiopian Civil Service University, PO Box 5648,
Addis Ababa 251, Ethiopia

DATES:
Received: 04 Aug. 2014
Revised: 30 Sep. 2014
Accepted: 22 Oct. 2014

KEYWORDS:
climate change; biophysical impacts; ecosystem services; Makonde Communal Lands; Zimbabwe

HOW TO CITE:
Sango I, Godwell N. Climate change trends and environmental impacts in the Makonde Communal Lands, Zimbabwe. S Afr J Sci. 2015;111(7/8), Art. #2014-0266, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20140266

During the last century, climate has increasingly become variable and changeable, with significant deviations from the observed normal averages, which often leads to disruptive consequences to ecosystems and livelihoods. Climate change induced environmental challenges are viewed to be particularly severe to economically challenged tropical societies including the Zimbabwean rural communities. We sought to determine local level climate change trends and associated biophysical implications in the Makonde Communal Lands of Zimbabwe. Our findings suggest that there has been significant climate change in the Makonde Communal Lands since 1962. The climate change observed has induced the deterioration of ecosystem productivity, diversity and services, to the detriment of human livelihoods. We provide insights into how to better understand local level dynamics between climate change and local ecosystem goods and services as the basis of livelihood in marginalised rural communities. Among the key reasons for concern about impacts of anthropogenic activities on climate is the fact that changing climate has direct impacts on the biophysical world, which in turn is a vital asset for human livelihoods, economies and general well-being.

Introduction

Climate implies the long-term average of the individual weather conditions that communities experience every day.[1] It is amongst the most important determinants of survival and human livelihoods.[2] Climate is a particularly strong factor for low-income rural communities whose livelihoods heavily depend on rain-fed subsistence agriculture, such as the Makonde Communal Lands of Zimbabwe – the study area and focus of this paper.

During the course of human civilisation, communities in all parts of the world have developed ways of earning livelihoods and supplying their needs for food, water, shelter and other goods and services that are adapted to benefit from the climates in which they live.[3] However, during the last century, climate has increasingly become variable and changeable, with deviations that are too far from the observed normal averages, often leading to disruptive consequences to ecosystems, livelihoods and human well-being.[4] Such major climatic deviations have become a major cause for concern in the modern world environment. In particular, impacts of anthropogenic activities on climate have become one of the most striking environmental challenges affecting current civilisations.[5]

Concerns about climate change and its associated environmental degradation are receiving increasing attention, particularly in tropical Africa, because of the large proportion of the rural population living in already ecologically vulnerable zones.[6] Evidence from the Intergovernmental Panel on Climate Change[2] suggests that sub-Saharan Africa is likely to emerge among the most vulnerable regions to climate change, with likely agricultural losses of up to 7% of the affected countries' gross domestic product.[7] Since 1900, much of southern Africa has progressively experienced warmer temperatures, rising on average 0.7 °C, and an overall decline in precipitation of 5%. If global mitigatory actions remain as weak as they currently are, many communities of the world, particularly in tropical rural Africa, are likely to experience some of the worst impacts of climate change in the current century.[8] The notable increase in the frequency and severity of drought and other weather extremes is proving to be among the biggest threats to the livelihoods of rural communities which rely heavily on climate-sensitive livelihoods. The level of vulnerability is particularly widespread given that the rural population in Zimbabwe, as in other sub-Saharan countries, comprises about 70% of the national population.

Given the growing evidence of climate change and its potential negative impacts on livelihoods, particularly those of the poor and vulnerable rural farming households and communities of Zimbabwe, our key concern in this study was to provide insights leading to the enhanced understanding and/or management of climate change related risks, thereby introducing opportunities for addressing overall livelihood vulnerability. The study is intended to highlight the adversity of the downstream externalities of global climate change. While the body of knowledge on climate change, its negative impacts, vulnerability and adaptation has grown significantly over the recent years, every local community has its own challenges associated with climate change. There is a need for more research on micro-level climate change impacts on livelihoods.

In terms of knowledge gap, until fairly recently, work investigating the impacts of and responses to climate change tended to be more prolific in the northern hemisphere. It is therefore pertinent that the long history of neglect of research on climate change and its impacts in the southern hemisphere be addressed. In this paper, we therefore seek to provide a broader view on the complex set of risks of climate change and its biophysical implications for rural, farming community livelihoods. A systems approach is applied to cover the mulitidimensional nature of climate change and its associated microscale implications. This particular focus of study is considered important for prioritising the places and people for whom adaptation intervention is required. The findings of the study are anticipated to contribute to a body of knowledge to furnish academia, global leaders, policymakers, local authorities and planners with a comprehensive understanding of the local level dynamics of climate change and its impacts.

Context of the study

Makonde Rural District lies in the Mashonaland West Region in the northwestern part of Zimbabwe. The rural district is divided into two main areas: the large-scale commercial farming area in the north and the Makonde Communal Lands which cover the southern section of the district. It is the Makonde Communal Lands in particular which constitute the area of interest in this study. Figure 1 shows the geographical location of the Makonde Communal Lands in the Makonde Rural District.

In terms of historical background, the Makonde Communal Lands as a geographical and socio-economic unit are a product of the colonial legacy which created a dual subdivision of the country into two agrarian structures. One was the commercial farming zone (Regions I to III) of the settler community in the more accessible and agro-ecologically productive regions, while the other constituted more remote reserves for the natives, on poorer soils and in hot and dry lowland regions (Regions IV and V).[9,10]

Factors such as the naturally stressed ecosystems, growing population pressure and the communal tenure system of access to and use of land resources has had a heavy toll on the state of the environment in Zimbabwe's communal lands such as the Makonde Communal Lands. According to Murombedzi[9] and Doré[11], the colonial legacy of this dual agrarian structure has prevailed through the Zimbabwean independence and continues to exist today, as evidenced by the drought-prone Makonde Communal Lands which are juxtaposed with the well-served agro-ecologically advantaged Makonde large-scale commercial farming area.

In terms of scope, therefore, the study is confined to the Makonde Communal Lands of Zimbabwe. The communal lands cover a narrow belt along the southern end of the Makonde Rural District, sharing a northern border with the Makonde commercial farming area (Figure 1). The study area is divided into six wards which are altogether divided into 10 villages, upon which the study's framework for stratified sampling is based. The subjects of the study are mostly the smallholder farmers across the 10 villages.

Theoretical background

Among the most striking environmental challenges affecting the earth is anthropogenic climate change.[4] Climate change is defined by a number of factors, including: temperature, humidity, rainfall, air pressure and wind and severe weather events.[12,13]

The Intergovernmental Panel on Climate Change[2] has reported that the average temperature of the earth's surface has risen by 0.74 °C since the late 1800s. Of the 12 warmest years in the instrumental record of global surface temperature since 1850, 11 occurred between 1995 and 2006.[3,14] The linear warming trend over the 50 years from 1956 to 2005 (0.13 [0.10 to 0.16] °C per decade) is nearly twice that for the 100 years from 1906 to 2005.[15,16]

Global warming is causing adverse impacts on the biophysical environment ranging from the melting of major glaciers and sea level rise to increased weather hazards and biodiversity loss.[17] With reference to climate change trends experienced in sub-Saharan Africa, Chishakwe[17] points out that the region has experienced a warming trend and increased climate variability over the past few decades. Smith et al.[18] suggest that temperatures in the sub-region have risen by over 0.5 °C during the last 100 years. During this period, the sub-region has also experienced a downward trend in rainfall[19]; rainfall in the region in the early 1990s was 20% lower than that in the 1970s, with significant droughts in the 1980s, early 1990s, and in 2002.[20]

Mary and Majule[21] carried out a study on the impacts of climate change and variability in Tanzania. Findings show that the local people perceived the changes in temperature and rainfall pattern and that the changes have affected crop and livestock productivity and have had a significant impact on rural livelihoods and food security. Mary and Majule's[21] study revealed that agriculture, forestry, water, coastal resources, livestock and human health were adversely affected by climate change. The specific stressors of climate change in the study manifested in the form of increasing frequency of floods, drought, erratic rains and other extreme events. The community perceptions of the most important factors undermining peasant livelihoods were, in decreasing order: increasingly unpredictable and declining amounts of rainfall, with unclear onset and ending; rising mean temperatures and pest prevalence; and increasing frequency, severity and duration of droughts.[21] Mary and Majule's[21] conclusion was that all livelihoods (including off-farm livelihoods) in the district were climate change sensitive, which implies that adaptation options to climate change in the district were not sustainable. With reference to the general climate response scenario in sub-Saharan Africa, Yohe[22] reiterates that most rural households in the region are hardly coping with the climate change impacts and as such, the current response options are not sustainable both in socio-economic and ecological terms.

Methodological approach

The objective of our study was to determine climate change trends and associated environmental implications in the Makonde Communal Lands of Zimbabwe. Based on this objective, we sought to answer the following questions:

- Is climate change occurring in the Makonde Communal Lands?
- What climate change trends are experienced in the Makonde Communal Lands?
- What climate change induced biophysical changes are occurring in the Makonde Communal Lands?

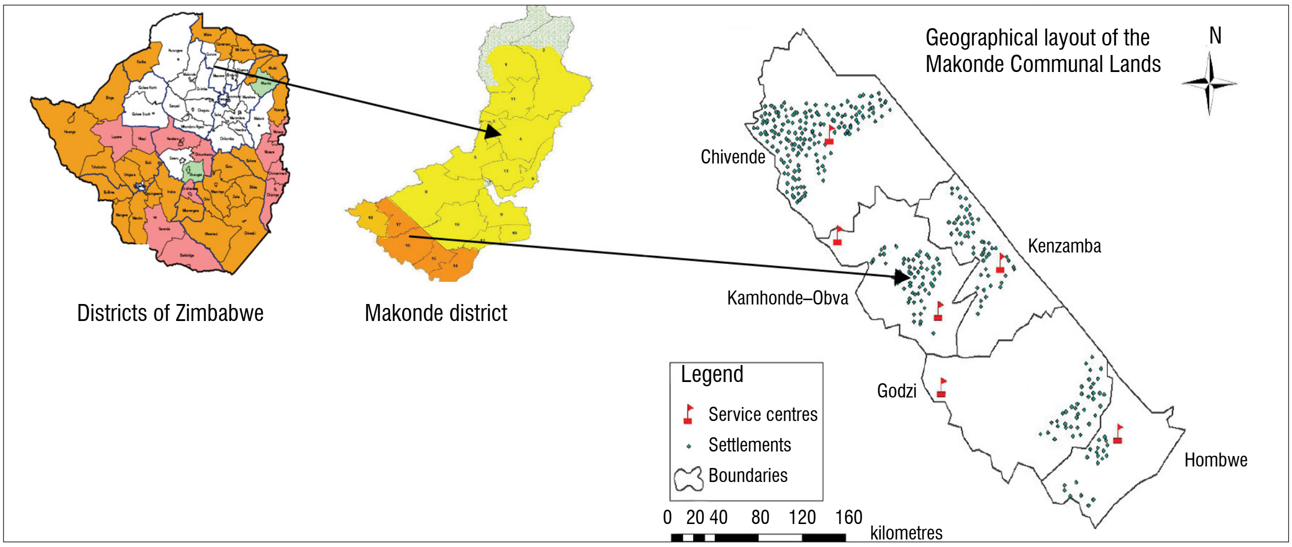


Figure 1: Geographical location of the Makonde Communal Lands within the Makonde Rural District of Zimbabwe.

Accordingly, the design for this research and the accompanying instruments were essentially determined by the research objective and corresponding research questions.[23-25]

Among the critical steps in our endeavour was the need to seek written consent and authority from the local authority for the study area (Makonde Rural District Council) and key informants in the study. The written consent in this regard was accordingly granted. With regard to the consent of the households, the Makonde Rural District Administrator, as the gatekeeper for the rural district, assured the necessary support and cooperation. Issues of confidentiality and anonymity were given serious consideration during the entire research process. The study was approved by the University of South Africa's College of Agriculture and Environmental Sciences Ethics Committee.

Design and instruments

A mixed research strategy was adopted in which both qualitative and quantitative approaches were included. Given the multidimensional nature of climate change indictors, a pluralistic approach in the survey was found to be the most appropriate. A mix of meteorological, biophysical, and sociocultural dimensions, altogether constituting a rural livelihood system, was covered. In order to have a scientific perspective of climate change trends in the case study, quantitative meteorological data recorded between 1962 and 2009 were taken as the principal form of data in this study. The data source for this critical raw data was the Zimbabwe's Department of Meteorological Services. Data on forest cover and related ecosystem goods and services were also obtained. A qualitative methodology of inquiry was adopted to allow for the interpretation of events and phenomena such as those identified in determining climate change trends and associated environmental implications in the Makonde Rural District of Zimbabwe.[25] This method included a focus on qualitative interpretation of people's perceptions and meanings attached to social phenomena, attitudes, beliefs and value systems.[25]

As expected of a case study, a mix of data collection instruments was employed in the study: key informant interviews; household questionnaires; documentation and archives' review; and structured field observations. Key informant interviews are among the critical instruments employed in the case study. By means of purposive sampling, a number of institutions were identified based on their special involvement or engagement in the issues of climate change and related biophysical and cultural changes.

The institutions purposively selected in the survey were the Ministries of Environment and Natural Resources (the Department of Meteorological Services and Environmental Management Authority); Local Government; Agriculture and Land Resettlement (Agricultural Research and Extension Services) and Labor and Social Welfare. The key informants provided vital information about the climate change patterns and their biophysical impacts and local level responses to the problem as applicable to the scope of their responsibilities. For the historical profiling of local climate change knowledge and experiences, 10 elders were selected by means of a snowball sampling method and unstructured interviews were conducted with one elder from each of the 10 villages.[26]

The questionnaire survey as a tool was adopted in this particular study because it is the most appropriate and cost-effective method of surveying a large sample population as in this particular case in which 500 households were surveyed. Given the data collection costs and other constraints, a sample of 500 households was viewed as significantly representative of the population under study. A stratified random sampling framework was employed to ensure proportionate representation of all the 10 villages under study, while at the same time giving each village-level household an equal opportunity to be included. From the household survey, we sought to elicit socio-economic characteristics of households and their knowledge and experience concerning local climate change and variability trends together with associated environmental implications.

Field observations were also conducted in and across the 10 villages in the study area to examine selected biophysical indicators and livelihood impacts in the area. Aspects for observation in the case study included: biophysical conditions, land use and artifacts and other indicators of climate and associated environmental change.

A comprehensive literature survey was conducted to review climate change scenarios and associated biophysical implications at various geographical scales ranging from global to local level. Also derived from the theoretical framework was the methodological aspect of the Millennium Ecosystem Assessment (MEA).[27]

The analysis of data in the study involved both quantitative and qualitative techniques. For the quantitative data, Statistical Packages for Social Sciences (SPSS), complimented by Microsoft Excel, was used to determine correlations and cross-cutting issues. In order to understand the biophysical and socio-economic dynamics of the study area, some maps on human settlements and food security situation were generated. A basic geographical information system (GIS) was employed as a means to integrate different components of the prevailing environmental conditions and community well-being. The analysis also involved an evaluation of the natural ecosystem services connected to local livelihood sectors with great emphasis on how climate change impacts on forests and other land resources and in turn how these influence the identified livelihood sectors.

For the qualitative data, detailed descriptions and classifications were constructed to provide insight into the research questions. Typologies were created to analyse qualitative data.[28] The typological technique aided in highlighting the Makonde community's vulnerability setting. The creation of taxonomies highlighted the hierarchical and other relationships among categories and sub-categories of research themes. Visual representations such as conceptual maps, tables and charts were also employed to analyse and organise the data in order to gain insight into the setting.[25]

Finally, the study included the gathering, analysis and interpretation of climatic and other biophysical data with emphasis on the past and current scenarios, with little reference to climate modelling for potential future impacts. This position was taken because of the inherent weaknesses associated with climate modelling for the future 'scenarios'. Smit and Pilifosova[28] argue that assumptions put forward in climate modelling fail to match with behaviours, both natural and human.

Results and discussion

Climate change trends in the Makonde Communal Lands

We examined whether climate change had been experienced in the area and what trends had been observed. Our key findings suggest that climate change in the Makonde Rural District has been significant during the past 30 years. The mean annual temperatures recorded since 1962 show an increasing warming trend. In terms of rainfall pattern, the mean rainfall amounts recorded during the same period illustrate increasing annual rainfall variability, with a falling trend over the recent decades.

Our findings concur with those of Weart[5] – among the most striking environmental challenges affecting the earth is anthropogenic climate change. The climate change trends observed in the Makonde Communal Lands suggest a progressive warming of the temperature conditions, falling of rainfall amounts and increasing variability of rainfall received between and within rainfall seasons. The analysed rainfall and temperature data that were obtained from the Department of Meteorological Services generated a pattern that clearly confirms the changing climate in the district. Figures 2 and 3, respectively, show the temperature and rainfall patterns experienced over a period of 36 years (since 1962) in the Makonde Rural District.

Whilist there is significant interannual variability in mean temperature beween 1962 and 2008 (Figure 2), the trends displayed show a departure (or anomalies) from the general average and the anomalies are mostly positive. The anomalies shown in Figure 2 are larger in more recent years (beginning in the 1980s), suggesting that the rate of increase of mean temperature is increasing over time. This finding is consistent with detected increases in global and regional annual surface temperatures discussed earlier. A trend analysis of temperature in the study area revealed an increase in both annual maximum and minimum temperatures between 1962 and 2008. Further analysis showed that the period of most rapid warming occurred since the early 1980s to date.

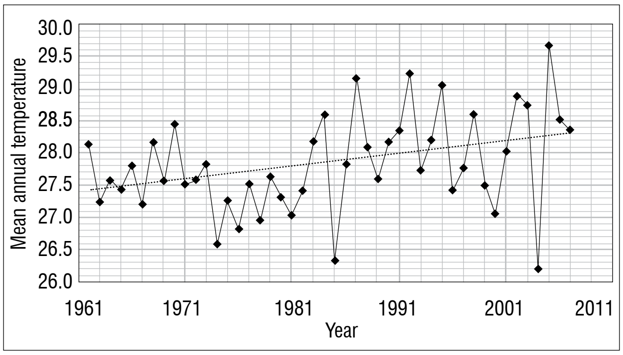


Figure 2: Mean annual temperature (°C) changes in the Makonde Rural District from 1962 to 2008.

In terms of rainfall, the patterns of annual rainfall between 1962 and 2008 shown in Figure 3 are in agreement with already gathered evidence of rainfall trends, which suggest a progressive decrease in annual rainfall over southern Africa. The extreme dips notable in Figure 3 symbolise the increasing interannual rainfall variability associated with increasingly frequent and severe drought spells over the past 20 years. The alternating patterns of below-normal (most frequent) to above-normal rainfall periods reveal the trends of both climate variability and climate change in the Makonde Communal Lands.

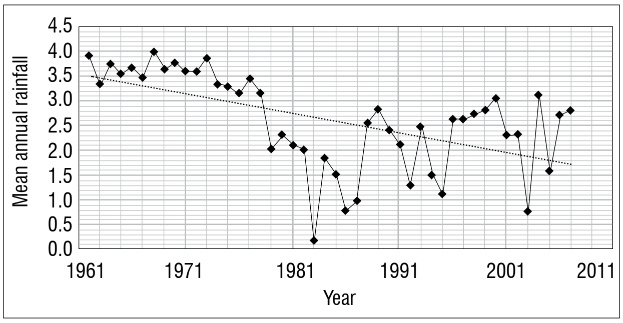
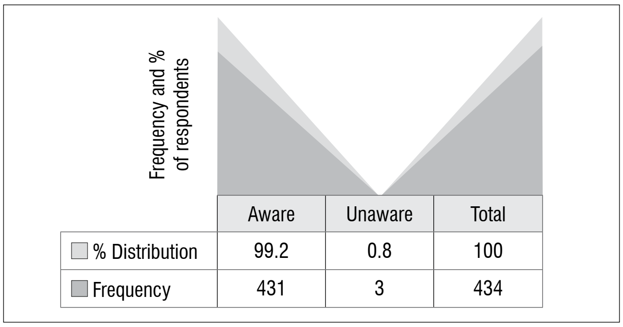


Figure 3: Mean annual rainfall (mm) changes in the Makonde Rural District from 1962 to 2008.

From the perspective of indigenous knowledge systems, community experiences, knowledge and perceptions on climate change in the area were examined in the field. All key informants – from the Environmental Management Authority, Agriculture Research and Extension Services and Makonde Rural District Council to the local leaders and the non-governmental organisations operational in the study area – concurred that there was significant climate change occurring, with progressively warming temperatures and falling rainfall amounts. The most outstanding indicator of climate change among the key informants was the increasing frequency and severity of drought spells occurring in the Makonde Communal Lands. The household questionnaire survey revealed a similar result – awareness of varying and changing climate was almost 100% among households. Community climate change awareness was also investigated (Figure 4). Figure 4 shows that only 0.8% of community members are not aware of the changing climate in the district.

The majority (66.4%) of the study population has experienced significant climate change, with 28.6% suggesting that the climate change is intensifying. Specifically, about 63% of the sampled population suggested that there has been a significant change in the climate. About 80% indicated that they had experienced a progressive fall in mean annual rainfall amounts in the past 30 years. In terms of the prevalence and severity of drought spells over the past years, a total of 49.3% and 44.2% of the community indicated that they either often or very often experienced drought spells over the past 30 years, respectively. As shown in Figure 5, an estimated 76.2% of the sample suggested an increasing severity of the drought spells in the case study.



	Aware	Unaware	Total
▢ % Distribution	99.2	0.8	100
▨ Frequency	431	3	434

Figure 4: Community awareness of local climate change in the Makonde Rural District (*n*=434).

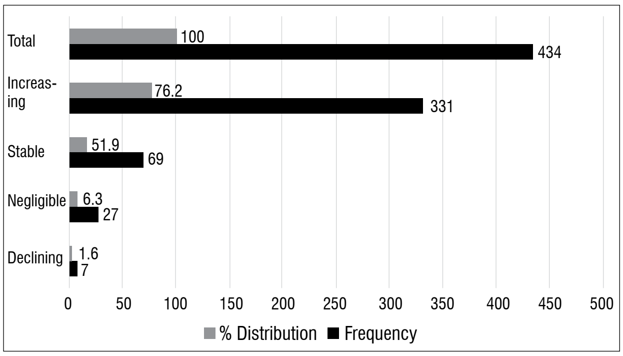


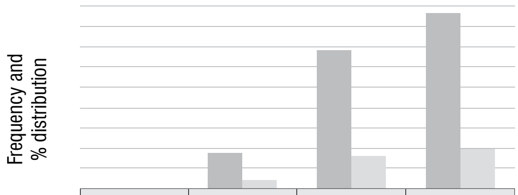
Figure 5: Community perceptions of the severity of drought spells over the years (*n*=434).

Given the observational record of the scientifically availed temperature and rainfall changes for the Makonde Rural District over a period of at least 30 years, examined in parallel with the community experiences and indigenous knowledge systems on climate change, we conclude that there has been significant climate change in the Makonde Communal Lands. In spatial terms, climate change in the study area is more severe in the northeastern part of the study area, covering Villages 17 and 18 in the Kenzamba area (see Figure 1). This part of the study area experienced more severe drought spells because of its extreme northerly position towards the drier agro-ecological Region IV. The soils in this region are more sodic and water resources are generally much scarcer than in the southern and central parts of the study area.

Biophysical impacts of climate change in the Makonde Communal Lands

In light of the MEA's[26] assertion that there is an intricate association between climate and environmental resources on one hand and livelihoods on the other, climate change variables in the Makonde Rural District were also observed to influence local biophysical factors, such as plant and animal growth, water cycles, biodiversity and nutrient cycling. The MEA was adopted in the study to examine the environment through the framework of ecosystem goods and services.

Among the major environmental components affected by the changing climate in the case study is water resources. As much as 78.9% of the study population experienced significant impacts of local climate on local wetlands, springs, rivers and streams (Figure 6). With regard to domestic water supply, 64.3% experienced a decline in availability, 30.6% experienced seasonal availability, and 5.1% suggested that their local water sources had completely dried up. In terms of the state of drinking water, as much as 41% of the community perceived it as poor, 20.7% as deteriorating, and 10.6% apiece perceived it as very poor and critically poor. Focusing on water supply for livestock and irrigation purposes, 40.1% of households sampled suggested that it had become poor and mostly seasonal, whilst 11.3% and 11.1% perceived it as very poor and critically scarce, respectively.

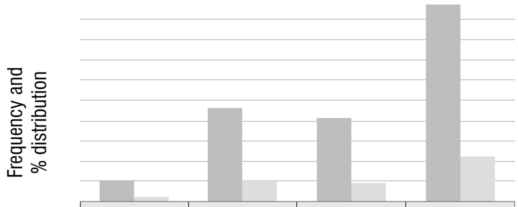


	None	Little	Much	Total
■ Frequency	2	89	343	434
□ % Distribution	0.5	20.6	78.9	100

Figure 6: Community perceptions of the climate impacts on local wetlands (n=434).

It is interesting to note that 17.1% of the local households in the case study found the water supply to be good and reliable for livestock and irrigation purposes. These households were generally from a section of the study area that is agro-ecologically endowed. This community of households is largely situated in the area to the southern tip of the communal area (Hombwe–Mukohwe), bordering the agro-ecological zone II (see Figure 1). Incidentally, in terms of water infrastructure, this community enjoys the privilege of accessing the small earth dam that was built for the St Rupert's Mission Hospital.

The other biophysical resource that is critical in sustaining rural livelihoods is forest resources for food, wood, medicinal and other domestic purposes. Figure 7 illustrates community perceptions of the climate change impacts on wild food and other forest resources. Almost half (47.5%) of the Makonde community has experienced a growing scarcity of forest resources for various domestic purposes. A further 41.8% suggested that the resources were now very scarce because of persistent drought spells undermining ecosystem recovery and productivity. The traditional wild fruits which used to be in abundance were now very scarce and many species that provide fruits, wood and traditional medicines had disappeared from the local agro systems.



	Declining	Now scarce	Now very scarce	Total
■ Frequency	46	206	182	434
□ % Distribution	10.7	47.5	41.8	100

Figure 7: Local community perceptions of the state of forest food resources (n=434).

In terms of wild animals, elders recalled times when there used to be an abundance of small antelopes, bushbucks and other small mammals, birds and fish contributing to a rich food reserve for the community. According to the elders, all that remains is the baboons and monkeys which, because of the growing scarcity of wild food sources, are increasingly encroaching onto crop fields and homesteads, thus further threatening the declining crop yields and domestic small livestock such as goats, chicken and rabbits.

From the perspective of a rural setting of smallholder farmers, climate change has a direct effect on all food security dimensions which include food availability, food accessibility and food systems stability. It also has an impact on human health, livelihood assets and food production. Forest resources in the study area are under threat from climate change and other climate change induced pressures such as excessive wood harvesting, gold panning and veld fires. The changing rainfall and temperature regimes have had an impact on the availability of many forest resources such as food, wood and other livelihood assets in the area. The depletion of wild food resources has not only exacerbated the insecurity of the community, but has also forced baboons, monkeys, wild pigs and other wildlife to turn to crop fields and domestic livestock for food. This wildlife encroachment has intensified food insecurity among smallholder farmers and the conflicts between the community and wild animals. Productive time is often wasted spent guarding the crop against invasion by wildlife.

Among the major assets that support rural livelihoods and sustain household well-being is livestock ownership. Among the ecosystem services that the community traditionally enjoys is pasture for livestock, which is similarly under threat from the climate change–ecosystem degradation interface. About 32% and 19.4% of the community suggested that the pasture was now scarce and very scarce, respectively. The combined climate change impact of increasing water scarcity, forest depletion and declining pasture availability and quality has seriously undermined livestock productivity, variety, health and numbers in the local villages. As much as 49.8% of the surveyed households incurred significant losses in livestock size, particularly during the series of drought periods mentioned earlier. An estimated 49.5% of the sampled households conceded that the health and quality of the remaining livestock had dropped significantly. The combined effect of climate change induced food shortages and increasing risk of loss of livestock from drought and disease has seen many households forced to sell off their livestock at uneconomically low prices.

As earlier noted, even within the relatively small geographical unit of the Makonde Communal Lands, there is some degree of spatial heterogeneity in terms of climate change trends and related environmental manifestations. Whilst the majority of the households suffer the climate change induced scarcities of water, forest resources and pasture, a certain cluster of households in the study area have experienced little to no negative effects on their livelihoods, food security and general well-being. These households are mostly confined to the southeastern tip of the district, where the conditions are more humid, being geographically linked to the agro-ecological Region II, with higher, more reliable rainfall and richer agro-ecosystems.

Conclusion

Based on the observed meteorological records over the past 30 years and local community knowledge systems, the smallholder farming community of Makonde Communal Lands is experiencing significant climate change. An analysis of meteorological data, questionnaire responses and interview transcripts tapping into local community experiences and indigenous knowledge systems led to the conclusion that there has been significant climate change in the study area since 1962. The climate change trends clearly manifest in the form of progressive warming, falling mean annual rainfall amounts and the increasing frequency and severity of drought spells in the Makonde Communal Lands.

As a consequence of climate change occurring in the Makonde Communal Lands, the study has noted three dimensions of vulnerability to the changing climate among which is the physical–environmental dimension which refers to the emergent local climatic conditions and associated biophysical impacts. The biophysical changes experienced in the area are in the form of the deterioration of ecosystem productivity and diversity. From the perspective of the biophysical dimension of vulnerability, the changing climate in the Makonde Communal Lands has had a significant effect on the state, productivity and diversity of local biophysical resources. The survey revealed that local climate change has induced the progressive degradation of local environmental assets such as forests and associated products, wildlife, water, pasture and soils.

Given the multifaceted nature of the climate change phenomenon and the associated impacts on ecosystems, economies and general human well-being, a mix of measures needs to be explored. These measures include the need to bridge climate change information gaps at national and local level; address institutional capacity constraints in climate change and environmental management; and put in place a multi-stakeholder approach to institute, implement and monitor a comprehensive community-based natural resources management system and sustainable intensification of land and animal husbandry practices. Specific areas that need capacity building include: weather focusing, climate change monitoring and early warning systems; climate change education; and appropriate weather information dissemination to farmers in order to strengthen climate resilient livelihoods.

Acknowledgements

We thank the Exxaro Chair in Business and Climate Change at the University of South Africa for partial sponsorship of the fieldwork and the Unisa Bursary Section for payment of 2013 tuition fees for I.S.

Authors' contributions

I.S. performed the data collection and data analysis and wrote the manuscript. G.N. supervised the entire research process, provided guidance on the design of the research project, made conceptual contributions and provided overall quality assurance for the manuscript.

References

1. Davies S. Adaptable livelihoods: Coping with food insecurity in the Malian Sahel. London: Macmillan Press; 2011.
2. Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, et al., editors. Climate change 2007: The physical science basis. Contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge and New York: Cambridge University Press; 2007. p. 1–18.
3. Kulkarmi J, Leary N. Climate change and vulnerability in developing country regions. Draft final report of the AIACC Project: A Global Environmental Facility Enabling Activity in the Climate Change Focal Area, Project no. GFL–2328-2724-4330. Nairobi: UNEP; 2007. Available from: http://www.aiaccproject.org
4. Tol RSJ, Downing TE, Kuikb OJ, Smith JB. Distributional aspects of climate change impacts. Global Environ Chang. 2004;14:60. http://dx.doi.org/10.1016/j.gloenvcha.2004.04.007
5. Weart SR. The discovery of global warming. Cambridge: Harvard University Press; 2004. Available from: http://www.aip.org/history/climate/index.html
6. Leichenko RM, O'Brien KL. The dynamics of rural vulnerability to global change. The case of southern Africa. Mitig Adapt Strat Gl. 2001;7:1–18. http://dx.doi.org/10.1023/A:1015860421954
7. Aguilar Y, Rodriguez E, Tobar J. Current and future sustainability of rural livelihoods: A case study in the community of Isla Mendez, El Salvador [document on the Internet]. c2005 [cited 2014 Aug 04]. Available from: http://www.start.org/Program/advanced_institute3_web/Final%2520Papers/RPaper_MetziAguilar
8. Metz B, Davidson OR, Bosch PR, Dave R, Meyer LA, editors. Climate change 2007: Mitigation of climate change. Contribution of Working Group III to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge and New York: Cambridge University Press; 2007; p. 7–22.
9. Murombedzi JC. Pre-colonial and colonial conservation practices in southern Africa and their legacy today. Southern Rhodesia, Reports of Debates – Legislative Council, 1899–1903. Ordinance No. 6 of 1899 in BSAC Government Gazette, 9/8/1899; 2003.
10. Ranger T. The invention of tradition in colonial Africa. In: Hobsbawm E, Ranger T, editors. The invention of tradition. Cambridge: Cambridge University Press; 1983.
11. Doré D. Transforming traditional institutions for sustainable natural resource management: History, narratives and evidence from Zimbabwe's communal areas. Afr Stud Quart. 2001;5(3), 18 pages. Available from: http://asq.africa.ufl.edu/files/Dore-Vol-5-Issue-3.pdf
12. Eriksen SH, Brown K, Kelly PM. The dynamics of vulnerability: Locating coping strategies in Kenya and Tanzania. Geogr J. 2005;171:287–305. http://dx.doi.org/10.1111/j.1475-4959.2005.00174.x
13. Schneider SH, Semenov S, Patwardhan A, Burton I, Magadza CH, Openheimer M, et al. Assessing key vulnerabilities and the risk from climate change. In: Parry ML, Canziani OF, Palutikof JP, Van der Linden PJ, Hanson CE, editors. Climate Change 2007: Impacts, adaptations and vulnerability. Contribution of Working Group II to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge: Cambridge University Press; 2007. p. 779–810.
14. Christensen JH, Hewitson B, Busuioc A, Chen A, Gao X, Held I, et al. Regional climate projections. In: Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, et al., editors. Climate change 2007: The physical science basis. Contribution of Working Group I to the fourth assessment report of the Intergovernmental Panel on Climate Change. Cambridge and New York: Cambridge University Press; 2007. p. 847–940.
15. Agrawal S, editor. Bridge over troubled waters: Linking climate change and development. Paris: Organisation for Economic Cooperation and Development; 2005.
16. Bryceson DF. The scramble in Africa: Reorienting rural livelihoods. Rev Afr Polit Econ. 2002;31:617–629. http://dx.doi.org/10.1016/s0305-750x(02)00006-2
17. Chishakwe NE. Southern Africa sub regional framework on climate change programs report. Draft working document. SADC–UNEP; 2010 [unpublished report].
18. Smith JB, Schellnhuber H–J, Monirul Qader Mirza M. Vulnerability to climate change and reasons for concern: A synthesis. In: McCarthy JJ, Canziani OF, Leary NA, Dokken DJ, White KS, editors. Climate change 2001: Impacts, adaptation and vulnerability. Contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change Cambridge and New York: Cambridge University Press; 2001. p. 276–322.
19. NCAR. A continent split by climate change: New study projects drought in southern Africa, rain in Sahel [press release]. Boulder, CO: National Center for Atmospheric Research; 2005 May 24.
20. Chenje M, Johnson P. Water management in southern Africa. Maseru, Lesotho: Environment and Land Management Sector, SADC and IUCN; 1996.
21. Mary AL, Majule AE. Impacts of climate change, variability and adaptation strategies on agriculture in semi–arid areas of Tanzania: The case of Manyoni District in Singida Region, Tanzania. Afr J Environ Sci Tech. 2009;3(8):206–218.
22. Yohe G, Tol RSJ. Indicators of social and economic coping capacity – Moving toward a working definition of adaptive capacity. Glob Environ Change. 2002;2:311–341. http://dx.doi.org/10.1016/s0959-3780(01)00026-7
23. Cobretta P. Social research: Theory methods and techniques. London. SAGE; 2003.
24. Lincoln YS, Guba EG. Paradigmatic controversies, contradictions and emerging confluences. In: Denzin NK, Lincoln YS, editors. Handbook of qualitative research. 2nd ed. New York and Thousand Oaks, CA: SAGE; 2000. p. 163–188.
25. Cresswell J. Qualitative inquiry and research design: Choosing among five traditions. New York and Thousand Oaks, CA: SAGE; 1998.
26. Millennium Ecosystem Assessment. Ecosystems and human well-being: Policy responses. Washington DC: Island Press; 2005.
27. Flick U. An introduction to qualitative research. 2nd ed. London: SAGE; 2002.
28. Smit B, Pilifosova O. Adaptation to climate change in the context of sustainable development and equity. In: McCarthy JJ, Canziani OF, Leary NA, Dokken DJ, White KS, editors. Climate change 2001: Impacts, adaptation and vulnerability. Contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change Cambridge and New York: Cambridge University Press; 2001. p. 143–171.

Long-term changes and variability in rainfall and streamflow in Luvuvhu River Catchment, South Africa

AUTHORS:
John O. Odiyo[1]
Rachel Makungo[1]
Tinyiko R. Nkuna[1]

AFFILIATION:
[1]Department of Hydrology and Water Resources, University of Venda, Thohoyandou, South Africa

CORRESPONDENCE TO:
John Odiyo

EMAIL:
john.odiyo@univen.ac.za

POSTAL ADDRESS:
Department of Hydrology and
Water Resources, University
of Venda, Private Bag X5050,
Thohoyandou 0950, South Africa

DATES:
Received: 23 May 2014
Revised: 21 Sep. 2014
Accepted: 22 Oct. 2014

KEYWORDS:
linear regression; Mann–Kendall; trends; climate change; Limpopo Province

HOW TO CITE:
Odiyo JO, Makungo R, Nkuna TR. Long-term changes and variability in rainfall and streamflow in Luvuvhu River Catchment, South Africa. S Afr J Sci. 2015;111(7/8), Art. #2014-0169, 9 pages. http://dx.doi.org/10.17159/sajs.2015/20140169

We investigated long-term changes and variability in daily rainfall and streamflow in the Luvuvhu River Catchment, South Africa. Changes and variability in rainfall and streamflow impact on available water resources and the allocation of these resources. Daily rainfall data for six stations and daily streamflow data for four stations for the period 1920/1921–2005/2006 were grouped into cycles of 5 and 10 years. Daily means and standard deviations were computed for each cycle. Standard deviation was used to define the rainfall and streamflow variability. Linear regression was used to compute trends in 5- and 10-year average rainfall and streamflow and their standard deviations. Paired two-tailed *t*-tests (significance level of 0.05) were carried out to verify the spatial variability of rainfall and streamflow in the study area. Mann–Kendall and linear regression were used to determine trend analyses based on long-term annual rainfall and streamflow data. All but two rainfall stations showed decreasing trends in 5- and 10-year mean rainfall; 10-year mean daily rainfall showed decadal rainfall fluctuations. Contrasting trends were observed in 5- and 10-year mean streamflow, indicating that other factors such as anthropogenic activities and impoundments could be impacting on streamflow. Trend directions identified from Mann–Kendall and linear regression analyses of long-term annual rainfall and streamflow were similar to those identified by linear regression of 5- and 10-year mean daily rainfall. Results of paired two-tailed *t*-tests verified the spatial variability of rainfall and streamflow in the study area. We have shown that the variability of rainfall and streamflow has increased in the Luvuvhu River Catchment over the 86-year study period.

Introduction

Increased pressure on natural water systems and artificial water storage systems as a result of a growing population make southern Africa vulnerable to potential changes in the hydrological cycle as a result of global warming, which could lead to extremely negative impacts on societies within this region.[1] Studies on long-term changes and variability in rainfall and streamflow are therefore of immense interest in South Africa.

According to Ampitiyawatta and Guo[2], precipitation is a good long-term indicator of changes which impact on water resources. Furthermore, changes in precipitation patterns are very important for water resources managers who deal with water resources planning and management. Several studies have been undertaken on rainfall changes in South Africa, including in the Limpopo Province, in which Luvuvhu River Catchment (LRC) is located. Tyson et al.[3] noted discernible but specific regional oscillations of 16–20 and 10–12 years, ubiquity of 3–4-year fluctuations and spatially distinctive occurrences of quasi-biennial oscillations, based on analyses of rainfall data from 157 stations across South Africa for the period between 1880 and 1972. Dyer and Tyson[4] observed a 20-year oscillation in rainfall for the northeastern parts of South Africa over the period 1910–1972. Dyer and Gosnell[5] noted significant long-term oscillations with a mean wavelength of 19.2 years from 18 of the oldest and most reliable rainfall stations within the South African sugar industry. Neither Nicholson[6,7] nor Hulme[8] identified any trends in the mean annual rainfall over southern Africa for the periods 1900–1970, 1931–1960 and 1961–1990, respectively. Lumsden et al.[9] qualitatively analysed potential changes in hydrologically relevant rainfall statistics to determine where convergence existed amongst the different global climate models with respect to changes in rainfall in South Africa. Hydrologically relevant statistics include annual means and variances of the rainfall scenarios, as well as the distribution of daily rainfall amounts. The results of the global climate models evaluated in the study showed that more rainfall was projected for the east of the region while less rainfall was projected along the west coast and the adjacent interior, with the possibility of a slight increase in interannual variability.[9]

In studies on historical trends in precipitation over southern Africa, Kruger[10] reported significant decreases in annual precipitation in northern Limpopo, northeastern Free State, western KwaZulu-Natal and the southeastern regions of the Eastern Cape, and significant increases in precipitation during the wet season in the northern North West Province and an area over the Northern Cape Province, Western Cape Province and Eastern Cape Province. Lynch et al.[11] reported a gradual increase in annual rainfall in the Potchefstroom area from 1925 to 1998. A decrease in median annual rainfall in the Limpopo Province over the latter half of the 20th century was reported by Warburton and Schulze[12]. Dollar and Rowntree[13] did not detect long-term changes in the rainfall pattern over the Bell River Catchment in the Eastern Drakensberg of South Africa, but did find annual, seasonal rainfall cycles with variance peaks every 16–19 years.

There is a limited number of studies on long-term changes and variability in streamflow. Fanta et al.[14] investigated the variability of river flow for 502 river flow gauging stations in nine countries of the southern African region, including South Africa, with a view to document the spatial variability of the river flow regimes. They found evidence of declining run-off in parts of Zambia, Angola, Mozambique and the highveld in South Africa.

Grenfell and Ellery[15] used the coefficient of variation to determine the interannual variations in rainfall and streamflow in the Mfolozi River, South Africa. They found that rainfall and streamflow were highly variable with coefficients of variation for interannual rainfall and streamflow ranging from 22.6% to 36.6% and 61% to 79%, respectively. Research on southern African rainfall trends has focused on annual data series with little information on seasonal or daily data.[13] Thus, long-term daily average changes and variability has not been covered extensively. The same applies to long-term daily streamflow changes and variability. Such investigations have yet to be undertaken in the LRC. Our aim in the current study was to investigate long-term changes and variability in rainfall and streamflow in the LRC and to provide additional information on long-term changes and variability in daily rainfall and streamflow on a local scale.

Study area

The LRC is located in the Luvuvhu/Letaba Water Management Area in the Limpopo Province of South Africa. The LRC is located between longitudes 29°49'46.16''E and 31°23'32.02''E and latitudes 22°17'33.57''S and 23°17'57.31''S (Figure 1). It covers a catchment area of 5941 km^2. The Luvuvhu River flows for about 200 km through a diverse range of landscapes before it joins the Limpopo River near Pafuri in the Kruger National Park. The mean annual rainfall is 608 mm and the mean annual run-off is 520x10^6 m^3. Topography varies from 200 m to 1500 m (Figure 1) and greatly influences rainfall and run-off distribution in the catchment. The highest rainfall occurs in the upper reaches where the Soutpansberg Mountains are located, with little rainfall in the lower reaches around the Kruger National Park.

Land-use activities in the LRC include forestry, agriculture and settlements. Forestry plantations cover the upper reaches of the Luvuvhu and Latonyanda Rivers, declining towards the Albasini Dam. Land cover in the southern highlands of the LRC is dominated by exotic tree plantations of pines and eucalyptus. Land use in the LRC includes commercial forestry (4%), commercial dry land agriculture (10%), commercial irrigation agriculture (3%), range land (50%), conservation areas (30%) and urban areas (3%).[16] According to Griscom et al.[17], notable land-cover changes have occurred in the LRC in the northeastern part of South Africa in the past two decades. These changes are linked to human population growth and may be contributing to observed reductions in winter rivers' base flows and increased events of rivers within the Kruger National Park running dry.

Methodology

Daily rainfall data for six rainfall stations for 1931/1932–2005/2006 were obtained from the South African Weather Service and Lynch's[18] rainfall database. Daily streamflow data for four stations covering the period 1920/1921-2005/2006 were obtained from the Department of Water and Sanitation. Figure 1 shows the distribution of rainfall and streamflow stations within the study area. It was essential to select stations that were spatially distributed throughout the study area. Stations were also selected based on availability of long-term rainfall data (>30 years) with minimal or no gaps. The World Meteorological Organization[19] recommended a period of 30 years or longer as ideal for studies dealing with long-term changes. The selected stations had data from periods ranging from 40–86 years, with gaps in data of less than 5%.

Each rainfall and streamflow data set was divided into 5-year periods (pentads) and 10-year periods (decades). The mean and standard deviation were computed for each pentad and decade for all the stations in order to show the long-term changes and variability, respectively. The standard deviation is one of the most common statistical parameters used to measure overall dispersion (variation) of data[20], and has been widely

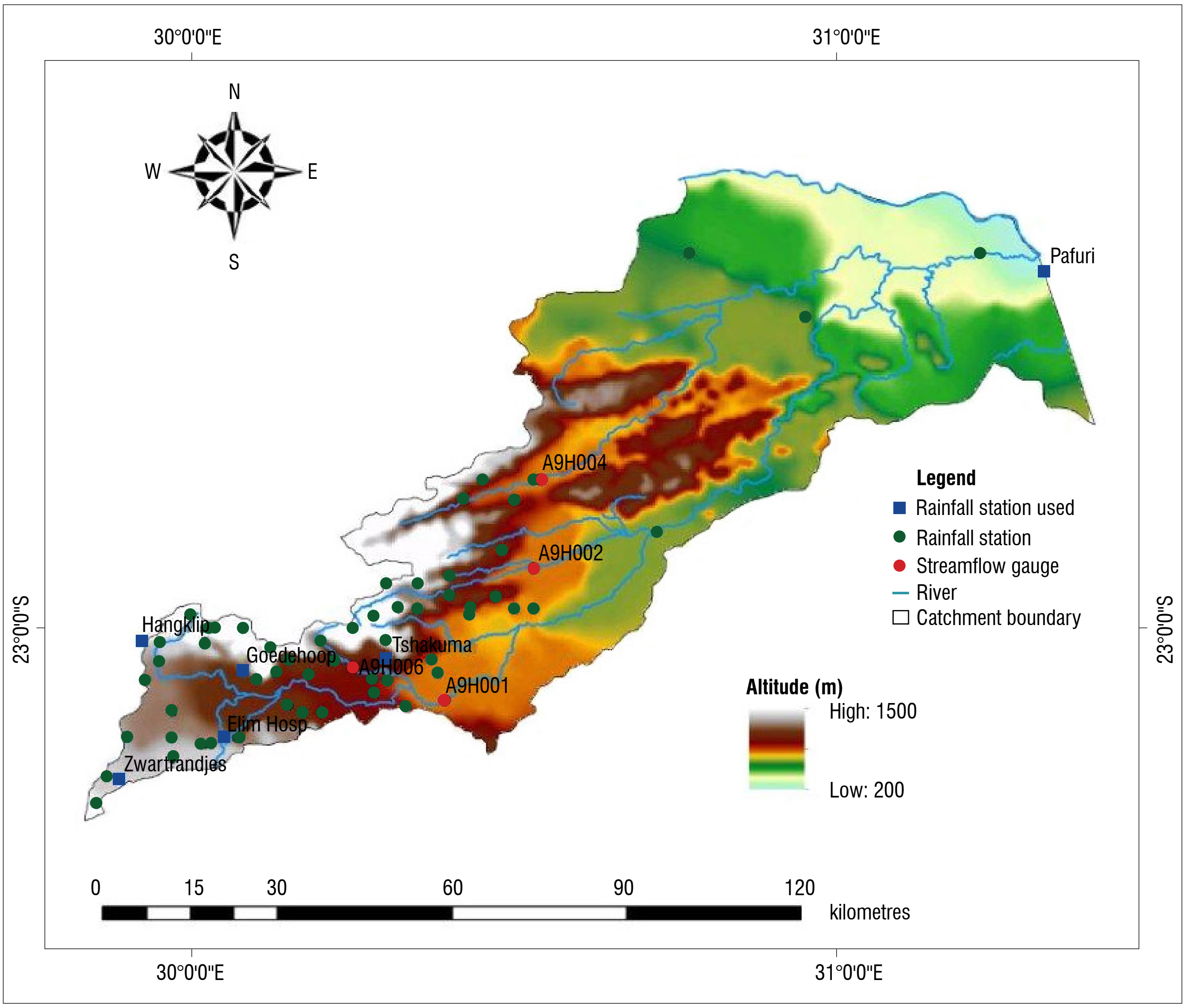


Figure 1: The locations of rainfall and streamflow stations within the study area.

used to study rainfall and/or streamflow variability (examples include the studies of Singh and Mulye[21] and Sanz et al.[22]). Linear regression was used to estimate trends in daily mean rainfall and streamflow over the study period for both the 5- and 10-year cycles. This method fits a regression line to the time series data and the slope indicates whether the trend is strong or not. The null hypothesis is that the slope of the line is zero. Linear regression has been used in a number of trend analysis studies including those of Suppiah and Hennessy[23], Schmidli and Frei[24] and Cheung et al.[25], amongst others. Paired two-tailed *t*-tests (with a significance level of 0.05) were used to verify if there was a significant difference between the means of rainfall and streamflow for any two stations. The samples were paired between sites based on their periods of record. This approach was useful in verifying the spatial variation of rainfall and streamflow in the study area.

Mann–Kendall was used for trend analyses based on long-term annual rainfall and streamflow data. Mann–Kendall is a rank-based non-parametric test. The null hypothesis (H_0) of the test is that there is no trend (the data is independent and randomly ordered) and the alternative hypothesis (H_1) is that there is a trend.[26] The Mann–Kendall test is a statistical test widely used for the analysis of trends in climatological and hydrological time series.[27] It was used together with linear regression to analyse trends in long-term annual rainfall and streamflow data to show comparability of results with the results of linear regression trend analyses based on 5- and 10-year means. This comparison was aimed at showing that the latter method had merit for trend detection, which was essential for the comparison of trends identified with the Mann–Kendall method for which annual data were used. Furthermore, using more than one method for trend analysis improves on the reasonableness of the results.

Results

Rainfall and streamflow trends and variability

The ranges of the 5- and 10-year means and standard deviations for rainfall and streamflow data for all stations are provided in Tables 1 and 2. All the results were organised according to hydrological year (October of one year to September of the following year); for example, years 51–56 refers to the hydrological years in the period 1951/1952–1955/1956. The ranges of the 5- and 10-year rainfall means and standard deviations for each station are comparable. The ranges of the 5- and 10-year means for streamflow are mostly not comparable, although the standard deviation ranges are comparable. The differences in the ranges of 5- and 10-year means for streamflow may be a result of floods which occurred in the pentads 76–81, 91–96 and 96–01 and in the decades 71–81 and 91–01. The major flood events occurred in the hydrological years 1976/1977, 1995/1996 and 1999/2000. Each flood event occurred towards the end of the decade and was not considered in the computation of the pentad mean streamflow preceded by dry cycles. The standard deviation ranges are generally higher than the means for both rainfall and streamflow data, showing high rainfall and streamflow variability in the study area. Fauchereau et al.[28] reported that southern Africa's geographical location, steep topography, contrasted oceanic surroundings and atmospheric dynamics are conducive to great interannual variability in the hydrological cycle.

Figure 2 shows the linear regression results for 5- and 10-year rainfall means for Elim, Goedehoop, Hanglip, Tshakhuma, Pafuri and Zwartrandjes rainfall stations. Goedehoop, Hanglip and Tshakhuma rainfall stations show decreasing trends in 5- and 10-year means. Pafuri and Zwartrandjes rainfall stations show increasing trends in 5- and 10-year means. Elim shows a decreasing trend in 5-year mean and an increasing trend in 10-year mean. Trends in rainfall characteristics are likely to be associated with changes in atmospheric circulation patterns,[23] which could be impacted by site-specific local effects and contribute to different trends. Others have also attributed changes in African rainfall to Indian Ocean processes and atmospheric features such as the intertropical convergence zone and anticyclones.[25] Shongwe et al.[29] projected drying rainfall trends in the southern African region, although decade-to-decade rainfall fluctuations were also noted. The 10-year mean daily rainfall shows decadal rainfall fluctuations which is in agreement to that noted by Shongwe et al.[29] That droughts are becoming more intense and widespread in South Africa,[28] confirms the decreasing trends obtained in this study. Warburton and Schulze[12] also reported a decrease in the median annual rainfall over the later half of the 20th century in the Limpopo Province, while Kruger[10] noted a significant decrease in annual precipitation in the northern Limpopo Province. Our results are mostly in agreement with these results, despite the fact that trends in the current study were based on mean daily rainfall.

Table 1: The 5- and 10-year means and standard deviations for rainfall data over the 86–year study period

Station	Mean		Standard deviation	
	5-year	10-year	5-year	10-year
Goedehoop	1.3–3.8	1.7–3.2	6.5–14.9	1.8–14.8
Hanglip	1.2–2.9	1.4–2.5	5.5–10.1	6.4–9.0
Tshakhuma	1.8–5.1	2.5–4.4	9.7–23.8	10.0–19.8
Zwartrandjes	0.9–1.8	1.1–1.7	4.2–9.8	4.7–9.2
Elim	1.3–2.7	1.3–2.5	5.8–15.5	7.0–13.1
Pafuri	1.0–1.6	0.9–1.7	4.9–8.0	5.2–7.9

Table 2: The 5- and 10-year means and standard deviations for streamflow data over the 86–year study period

Station	Mean		Standard deviation	
	5-year	10-year	5-year	10-year
A9H001	0.5–3.6	1.8–6.6	2.2–30.9	2.5–23.4
A9H002	1.4–9.9	1.0–2.6	0.6–21.9	0.7–18.1
A9H004	0.1–0.8	0.8–4.5	0.7–10.9	1.5–8.5
A9H006	0.7–4.6	0.1–0.5	0.1–1.6	0.2–1.1

The 5- and 10-year mean streamflows for A9H001, A9H002, A9H004 and A9H006 are provided in Figure 3. A9H001 and A9H004 show decreasing trends in 5- and 10-year mean streamflows, which are associated with decreasing trends in rainfall. A9H002 shows increasing trends in 5- and 10-year means while A9H006 shows an increasing trend in the 5-year mean and almost no trend in the 10-year mean (Figure 3). Contrasting trends in 5- and 10-year mean streamflows indicate that other factors such as anthropogenic activities and impoundments could be impacting on the streamflow. Chunzhen[30] noted that detection and attribution of a trend in a hydrological time series are much more difficult because changes in runoff are affected not only by climate factors, but also by non-climate factors, such as increases in water use and water consumption resulting from population growth, economic development, and changes in land use and land cover. These might be the reasons for contrasting trends in streamflow in the study area, as the anthropogenic activities in the area, which include agriculture, afforestation and settlements, occur in the quaternary catchments where each streamflow gauge is located.

Griscom et al.[17] reported considerable land-cover changes associated with human population growth and land-use activities such as agriculture, grazing and fuelwood cutting in the LRC. They reported a 12% increase in bare land between 1978 and 2005, which might have resulted in increased streamflows in parts of the catchment. Sambo[31] reported a 50% decrease in natural vegetation and a 30% increase in agriculture in the LRC between 1980 and 2010. Irrigation return flows and wastewater discharges into the river may also have increased the streamflows in parts of the river. In 2000, return flows of 5 million and 2 million m^3/year from irrigation and urban areas, respectively, were recorded in the LRC. The impact of anthropogenic activities on long-term streamflow changes and trends in the study area requires further investigation. Jewitt et al.[32] assessed the hydrological response of nine land-use scenarios in Mutale River quaternary catchment located within the LRC and showed that an increase in forestry to the maximum possible would result in a 7–9% decrease in streamflow. Odiyo et al.[33] showed increases in streamflows and their frequency of occurrence in the Luvuvhu River as a consequence of the removal of alien vegetation. The removal of alien vegetation could also have contributed to increased streamflow trends in the LRC. The 10-year mean daily streamflow also shows decadal fluctuations similar to those of mean daily rainfall.

Elim, Goedehoop, Tshakhuma and Zwartrandjes show increases in 5- and 10-year rainfall standard deviations (Figure 4). Hanglip station shows no significant changes in rainfall standard deviations while Pafuri station shows a slight increase and decrease in 5- and 10-year rainfall standard deviations, respectively. All streamflow stations show increasing trends of 5- and 10-year standard deviations, except for station A9H004 which shows a slightly decreasing trend for the 10-year standard deviation (Figure 5). Thus the results mostly indicate increased

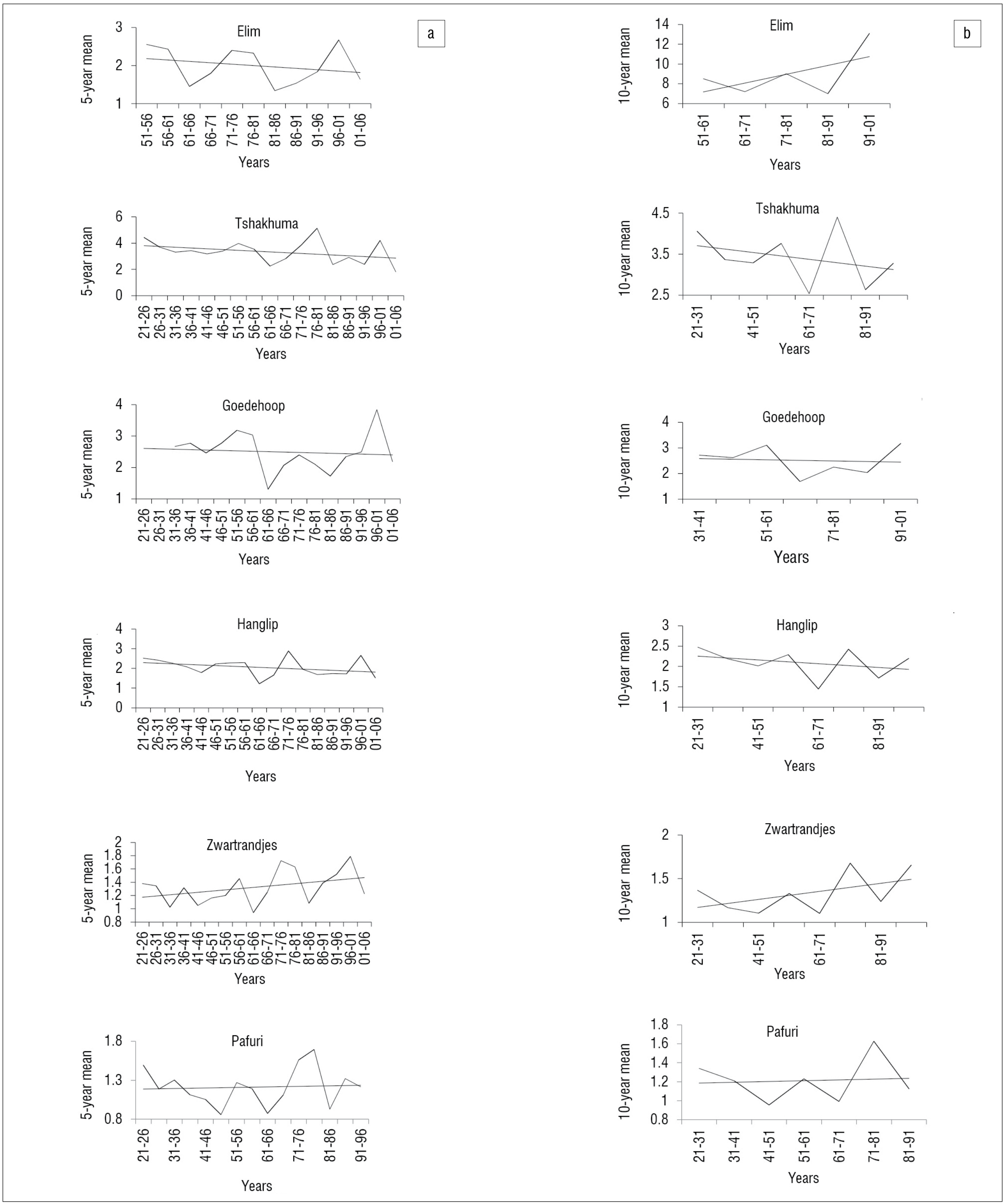


Figure 2: (a) 5- and (b) 10-year mean rainfall (mm) for the available hydrological years for each rainfall station.

variability of rainfall and streamflow. Reason and Rouault[34] reported the connection of ENSO-like decadal variability in South African rainfall. Figures 2–5 show 10-year cyclic increases and decreases in mean and standard deviation for all rainfall and streamflow stations. These findings show that the decadal or pentad rainfall trends and variations influence decadal or pentad streamflow trends and variations, as rainfall is one of the major drivers of run-off generation in a catchment.

Statistical significance of 5- and 10-year means

The differences in the 5- and 10-year rainfall means for Zwartrandjes station vary from not significant ($p>0.05$) to extremely significant ($p<0.001$), with the latter dominating compared with those of all the other stations (Table 3). The difference in 5- and 10-year means for Hanglip and Tshakhuma, and Hanglip and Zwartrandjes stations, is also extremely significant. There is no statistically significant difference in the 5- and 10-year means for Hanglip and Elim stations. The differences in the 5-year means for Goedehoop and Hanglip, and Tshakhuma and Elim are extremely significant ($p<0.001$), while those of Goedehoop and Tshakhuma, and Goedehoop and Elim are very significant ($0.001<p<0.01$). The differences in the 10-year means for Goedehoop and Hanglip and Goedehoop and Tshakhuma are significant ($0.01<p<0.05$) while the difference between the 10-year means of Tshakhuma and Elim stations is very significant ($0.001<p<0.01$). The significant differences in the 5- and 10-year means for the majority of the stations, some of which are highlighted above, verify the highly variable nature of rainfall in the study area, as do the results of the 5- and 10-year standard deviation comparisons.

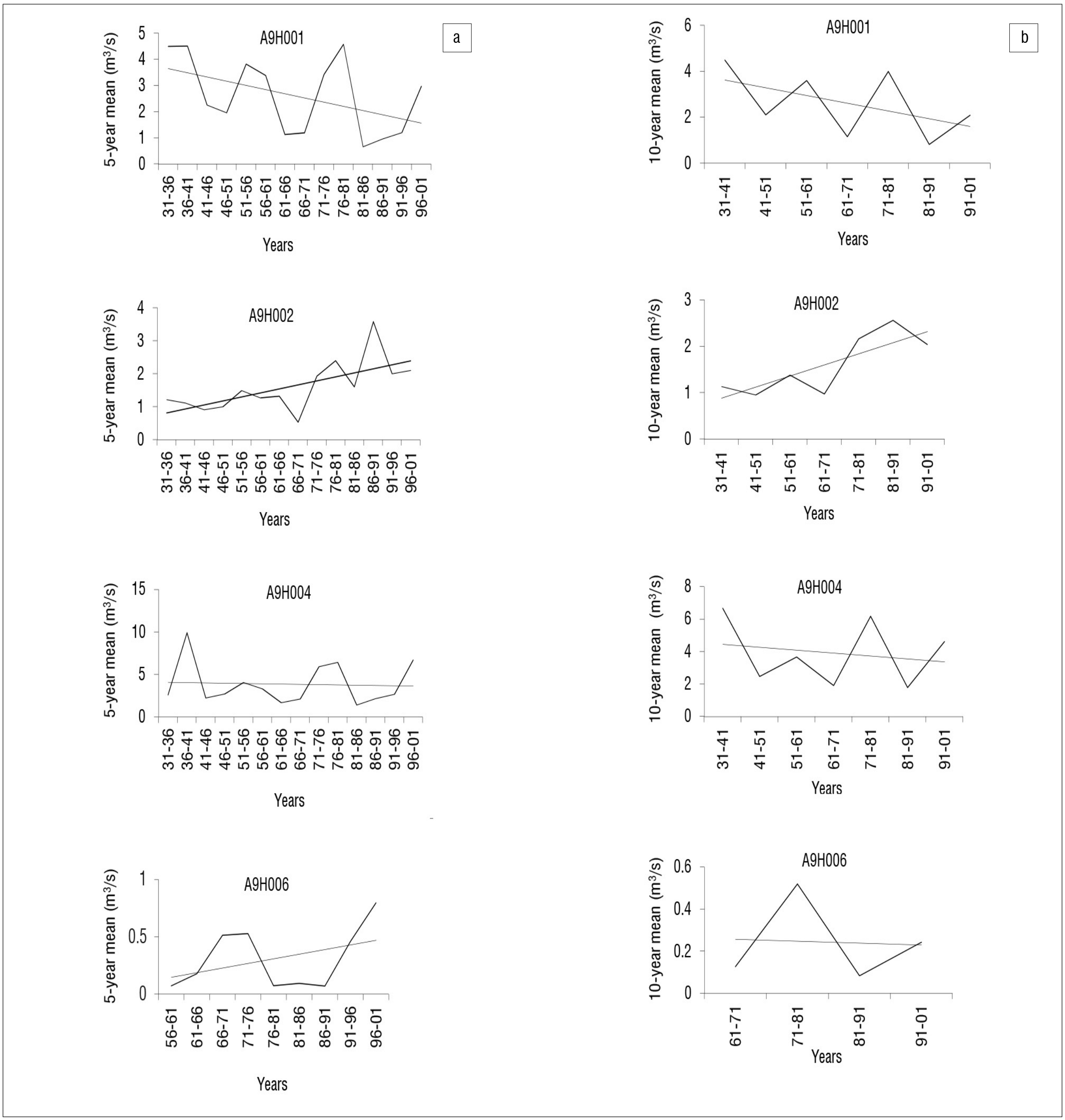


Figure 3: (a) 5- and (b) 10-year mean streamflows (m^3/s) for the available hydrological years for each streamflow station.

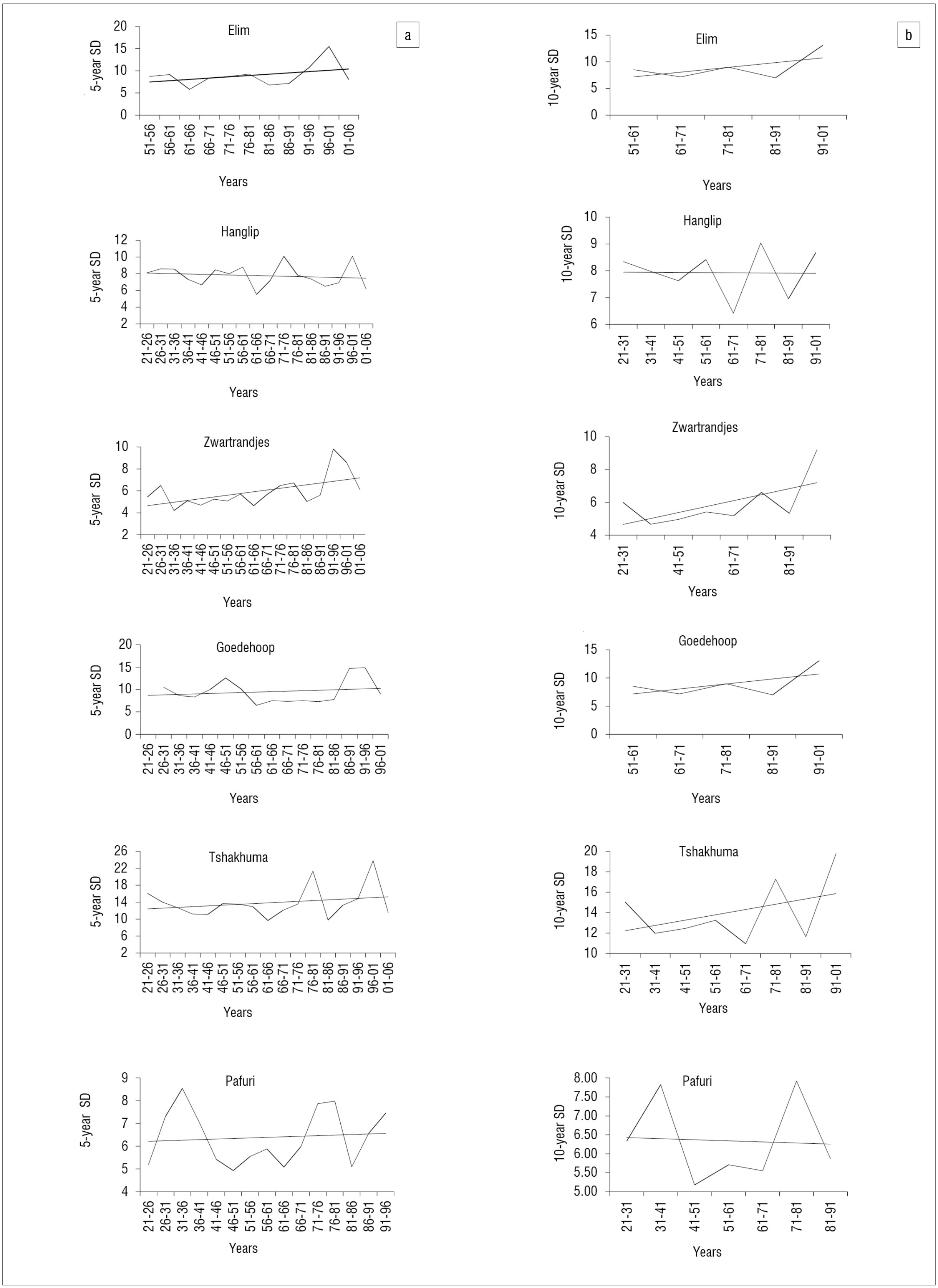


Figure 4: (a) 5- and (b) 10-year rainfall standard deviations (mm) for the available hydrological years for each rainfall station.

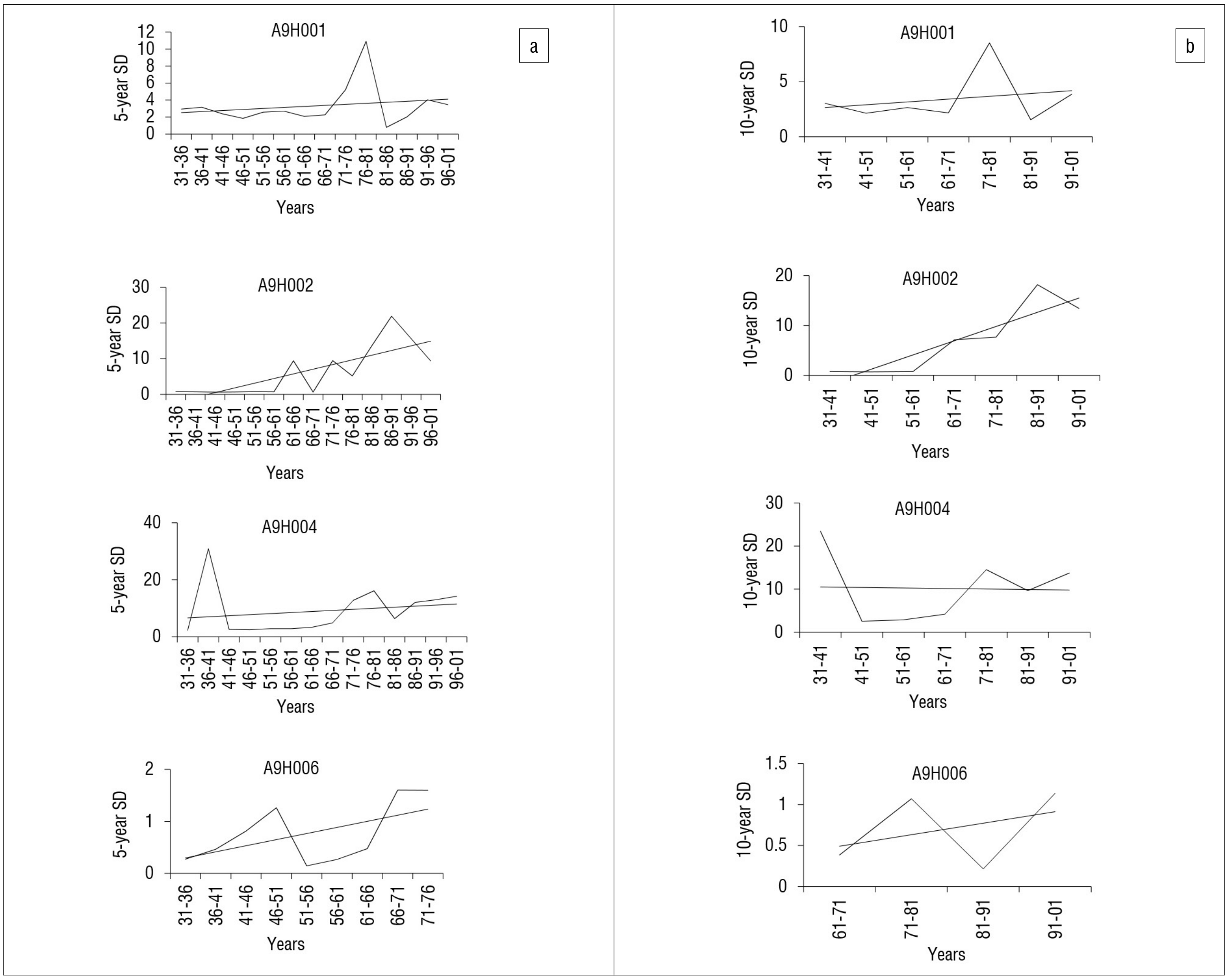


Figure 5: (a) 5- and (b) 10-year streamflow standard deviations (m^3/s) for the available hydrological years for each streamflow station.

Table 3: Results (*p*-values) of the paired two-tailed *t*-test comparing means of rainfall stations

Stations	5-year mean	10-year mean
Goedehoop and Hanglip	0.00039521***	0.016606693*
Goedehoop and Tshakhuma	0.00180114**	0.014944424*
Goedehoop and Zwartrandjes	0.00000074***	0.000873311***
Goedehoop and Elim	0.00725302**	0.088843473^{ns}
Hanglip and Tshakhuma	0.00000026***	0.00001182***
Hanglip and Elim	0.71213733^{ns}	0.726702838^{ns}
Hanglip and Zwartrandjes	0.00000033***	0.000117723***
Zwartrandjes and Elim	0.00017542***	0.009645059**
Tshakhuma and Elim	0.00014043***	0.002477142**
Tshakhuma and Zwartrandjes	0.00000001***	0.00001179***
Pafuri and Zwartrandjes	0.062260798^{ns}	0.024152941*
Pafuri and Tshakhuma	0.000000002***	0.000028993***
Pafuri and Hanglip	0.000000319***	0.000123572***
Pafuri and Goedehoop	0.000002362***	0.002604983***
Pafuri and Elim	0.000272006***	0.022086705

^{ns}*not significant ($p>0.05$);* **significant ($0.01<p<0.05$);* ***very significant ($0.001<p<0.01$);* ****extremely significant ($p<0.001$)*

The differences in the 5- and 10-year means for A9H004 and A9H001 and 5-year mean for A9H002 and A9H001 are significant (Table 4). The differences in the 5-year means for A9H002 and A9H006, A9H004 and A9H006, A9H006 and A9H001, and A9H004 and A9H002 are very significant while the 10-year means for A9H002 and A9H006, A9H004 and A9H006, and A9H004 and A9H002 are significant ($0.01<p<0.05$) (Table 4). There were no statistically significant differences in the 10-year means for A9H002 and A9H001, and A9H006 and A9H001. The results show significant differences in the mean values between different stations in the majority of cases, thus verifying the spatial variability in streamflow. However, the differences in the 5- and 10-year mean streamflows for different stations are not as highly pronounced as those for rainfall, indicating that the spatial variability of rainfall is higher than that of streamflow.

Table 4: Results (*p*-values) of the paired two-tailed *t*-test comparing means of streamflow stations

Stations	5-year mean	10-year mean
A9H002 and A9H006	0.001**	0.015*
A9H002 and A9H001	0.022*	0.094^{ns}
A9H004 and A9H006	0.001**	0.041*
A9H006 and A9H001	0.005**	0.064^{ns}
A9H004 and A9H001	0.021*	0.013*
A9H004 and A9H002	0.005**	0.025*

^{ns}*not significant ($p>0.05$);* **significant ($0.01<p<0.05$);* ***very significant ($0.001<p<0.01$*

Trends for total annual rainfall and streamflow

The Mann–Kendall and linear regression trend analysis results for rainfall are shown in Table 5. The Kendall statistic (S) was used to identify the direction of the trend. A positive value of S indicates an upward trend and a negative value of S indicates a downward trend.[35] Positive and negative signs of the *t*-statistic indicate increasing and decreasing trends, respectively.[36] Goedehoop, Hanglip, Tshakhuma and Elim showed downward (decreasing) rainfall trends according to both Mann–Kendall and linear regression analyses. However, Zwartrandjes and Pafuri stations showed upward (increasing) trends in rainfall. These results are similar to those obtained by linear regression analysis of 5- and 10-year means, indicating that 5- and 10-year means can be used in trend detection. None of the identified trends is statistically significant except for Tshakhuma – the only station that shows a statistically significant trend based on Mann–Kendall analysis. Based on linear regression analysis, three stations – Hanglip, Tshakhuma and Zwartrandjes – show statistically significant trends.

Table 6 shows Mann–Kendall and linear regression results for streamflow stations. A9H001 and A9H002 show decreasing trends, which are statistically significant, while A9H004 and A9H006 show increasing trends that are not statistically significant. Similarity in trend directions identified from both Mann–Kendall and linear regression methods support the use of 5- and 10-year means, as the results are mostly comparable. Thus, the results obtained from the use of 5- and 10-year means for trend detection are realistic, as they are comparable to those of established scientific or statistical methods.

Conclusion

We investigated long-term changes and variability in rainfall and streamflow in the LRC. The 5- and 10-year long-term means and standard deviations, linear regression and Mann–Kendall were used to show the long-term trends and variability. Most of the rainfall stations show decreasing trends in 5- and 10-year mean rainfall; Zwartrandjes and Pafuri stations show increasing trends. Elim station shows an increasing trend in 10-year mean. In agreement with other studies undertaken in the Limpopo Province, 10-year mean daily rainfall shows decadal rainfall fluctuations.

Streamflow stations show contrasting trends in 5- and 10-year mean streamflows, indicating that other factors such as anthropogenic activities and impoundments could be impacting on streamflow. The impact of anthropogenic activities on long-term streamflow changes and variability in the study area requires further investigation. Chunzhen[30] noted that it is important to separate natural climate variability from anthropogenic variability in historical data of hydrometeorological observations for a long-term period. Such a study should be carried out in the current study area. Most rainfall and streamflow stations show increasing trends for 5- and 10-year standard deviations. The results of the study thus, generally, show increased variability of rainfall and streamflow, which increases the variability of the available water resources. The statistically significant differences in the 5- and 10-year means for the majority of the rainfall and streamflow stations verify the highly variable nature of rainfall and streamflow in the study area and hence verify the results of the 5- and 10-year standard deviations. The decadal or pentad rainfall trends and variations influence decadal or pentad streamflow trends and variations as rainfall is one of the major drivers of run-off generation in a catchment. Trend directions identified from long-term annual rainfall and streamflow from Mann–Kendall and linear regression analyses were similar to those identified from linear regression analysis for 5- and 10-year mean daily rainfall, suggesting that the latter approach can be applied for trend analysis.

This simple method based on statistical analysis of available rainfall and streamflow data has clearly demonstrated climate change in the study area. Cheung et al.[25] demonstrated how the use of simple statistical analyses of historical rain gauge data can be used to accurately characterise rainfall. The method we used in the current study is therefore highly recommended for trend and variability detection in rainfall and streamflow in situations in which daily long-term data are available.

Table 5: Mann–Kendall and linear regression trends for rainfall data

Mann–Kendall						Linear regression			
Station	**S**	***p*-value (two-tailed)**	**Trend**	**α**	**Significant**	***t*-statistic**	**Trend**	**α**	**Significant**
Goedehoop	-354	0.18	Decreasing	0.1	No	-0.89	Decreasing	0.1	No
Hanglip	-416	0.115	Decreasing	0.1	No	-1.81	Decreasing	<0.1	Yes
Tshakhuma	-1655	<0.0001	Decreasing	0.1	Yes	-3.12	Decreasing	<0.01	Yes
Zwartrandjes	285	0.281	Increasing	0.1	No	1.82	Increasing	<0.1	Yes
Elim	-215	0.119	Decreasing	0.1	No	-0.28	Decreasing	0.1	No
Pafuri	124	0.609	Increasing	0.1	No	1.06	Increasing	0.1	No

Table 6: Mann–Kendall and linear regression trends for streamflow data

Mann–Kendall						Linear regression			
Station	**S**	***p*-value (two-tailed)**	**Trend**	**α**	**Significant**	***t*-statistic**	**Trend**	**α**	**Significant**
A9H001	-701	0.0004	Decreasing	0.1	Yes	-3.115	Decreasing	<0.01	Yes
A9H002	-445	0.019	Decreasing	0.1	Yes	-1.509	Decreasing	0.1	Yes
A9H004	27	0.895	Increasing	0.1	No	0.829	Increasing	0.1	No
A9H006	4	0.973	Increasing	0.1	No	1.054	Increasing	0.1	No

Acknowledgements

We thank the South African Weather Service and the Department of Water Affairs for providing data used in this study.

Authors' contributions

J.O.O. made conceptual contributions, performed the paired two-tailed *t*-tests and revised the edited manuscript. R.M. calculated the 5- and 10-year standard deviations, plotted the graphs, performed the linear regression analyses based on the annual data and edited the manuscript. T.R.N grouped daily rainfall and streamflow data into 5- and 10-year cycles, calculated the means and performed the linear regressions for 5- and 10-year cycles and the Mann–Kendall analysis.

References

1. Schulze R, Meigh J, Horan M. Present and potential future vulnerability of eastern and southern Africa hydrology and water resources. S Afr J Sci. 2001;67:150–160.

2. Ampitiyawatta AD, Guo S. Precipitation trends in the Kalu Ganga Basin in Sri Lanka. J Agric Sci. 2009;4(1):10–18.

3. Tyson PD, Dyer TGJ, Mametse MN. Secular changes in South African rainfall: 1880 to 1972. Quart J Roy Meteor. 1975;101:817–833. http://dx.doi.org/10.1002/qj.49710143008

4. Dyer TGJ, Tyson PD. Estimating above and below normal rainfall period over South Africa 1972–2000. J Appl Meteorol. 1977;16:145–147. http://dx.doi.org/10.1175/1520-0450(1977)016<0145:EAABNR>2.0.CO;2

5. Dyer TGJ, Gosnell JM. Long term rainfall trends in the South African Sugar Industry. Proceedings of the South African Sugar Technologists' Association. 1978:206–213.

6. Nicholson SE. The nature of rainfall variability in Africa south of the equator. J Climatol. 1986;6:515–530. http://dx.doi.org/10.1002/joc.3370060506

7. Nicholson SE. Long-term changes in African rainfall. Weather. 1989;44:46–56. http://dx.doi.org/10.1002/j.1477-8696.1989.tb06977.x

8. Hulme M. Rainfall changes in Africa: 1931-1960 to 1961-1990. Int J Climatol. 1992;12:685–699. http://dx.doi.org/10.1002/joc.3370120703

9. Lumsden TG, Schulze RE and Hewitson BC. Evaluation of potential changes in hydrologically relevant statistics of rainfall in southern Africa under conditions of climate change. Water SA. 2009;35(5):646–656. http://dx.doi.org/10.4314/wsa.v35i5.49190

10. Kruger AC. Observed trends in daily precipitation indices in South Africa: 1910-2004. Int J Climatol. 2006;26:2275–2285. http://dx.doi.org/10.1002/joc.1368

11. Lynch SD, Zulu JT, King KN, Knoesen DM. The analysis of 74 years of rainfall recorded by the Irwins on two farms south of Potchefstroom. Water SA. 2001;27:459–564. http://dx.doi.org/10.4314/wsa.v27i4.4970

12. Warburton M, Schulze RE. Historical precipitation trends over southern Africa: A hydrology perspective. In: Schulze RE, editor. Climate change and water resources in southern Africa: Studies on scenarios, impacts, vulnerabilities and adaptation. WRC report no. 1430/1/05. Pretoria: Water Research Commission; 2005. p.326–338.

13. Dollar ESJ, Rowntree KM. Hydroclimatic trends, sediment sources, and geomorphic resources in the Bell River Catchment, Eastern Cape Drakensberg, South Africa. S Afr Geogr J. 1995;77(1):21–32. http://dx.doi.org/10.1080/03736245.1995.9713585

14. Fanta B, Zaake BT, Kaachroo RK. A study of variability of annual river flow of the southern African region. Hydrolog Sci J. 2001;46:513–524. http://dx.doi.org/10.1080/02626660109492847

15. Grenfell SE, Ellery WN. Hydrology, sediment transport dynamics and geomorphology of a variable flow river: The Mfolozi River, South Africa. Water SA. 2009;35(3):271–282.

16. Hope RA, Jewitt GPW, Gowing JW. Linking the hydrological cycle and rural livelihoods: A case study in the Luvuvhu catchment, South Africa. Phys Chem Earth. 2004;29:1209–1217. http://dx.doi.org/10.1016/j.pce.2004.09.028

17. Griscom HR, Miller SN, Gyedu-Ababio T, Sivanpillai R. Mapping land cover change of the Luvuvhu catchment, South Africa for environmental modeling. GeoJournal. 2010;75:163–173. http://dx.doi.org/10.1007/s10708-009-9281-x

18. Lynch SD. Development of a raster database of annual, monthly and daily rainfall for southern Africa. WRC report no. 1156/1/04. Pretoria: Water Research Commission; 2004.

19. World Meteorological Organization (WMO). Technical publication no. 335: Compendium of lecture notes in Climatology for class III meteorological personnel. Geneva: WMO; 1976.

20. An H, Eheart JW, Braatz RD. Stability-oriented programs for regulating water withdrawals in riparian regions. Water Resour Res. 2004;40:W12301.

21. Singh N, Mulye SS. On the relations of the rainfall variability and distribution with the mean rainfall over India. Theor Appl Climat. 1991;44(3&4):209–221. http://dx.doi.org/10.1007/BF00868176

22. Sanz DB, Sacristán MM, Rubial GC. The natural variability approach, application to five rivers in the Ebro Basin, Spain. Environ Econ. 2011;2(2):107–121.

23. Suppiah R, Hennessy KJ. Trends in total rainfall, heavy rain events and number of dry days in Australia, 1910–1990. Int J Climatol. 1998;10:1141–1164. http://dx.doi.org/10.1002/(SICI)1097-0088(199808)18:10<1141::AID-JOC286>3.0.CO;2-P

24. Schmidli J, Frei C. Trends of heavy precipitation and wet and dry spells in Switzerland during the 20th century. Int J Climatol. 2005;25:753–771. http://dx.doi.org/10.1002/joc.1179

25. Cheung WH, Senay GB, Singh A. Trends and spatial variation of seasonal and annual rainfall in Ethopia. Int J Climatology. 2008;28:1723–1734. http://dx.doi.org/10.1002/joc.1623

26. Onoz B, Bayazit M. The power of statistical tests for trend detection. Turkish J Eng Env Sci. 2003;27:247–251.

27. Bawden AJ, Linton HC, Burn DH, Prowse TD. A spatiotemporal analysis of hydrological trends and variability in the Athabasca River region, Canada. J Hydrol. 2014;509:333–342. http://dx.doi.org/10.1016/j.jhydrol.2013.11.051

28. Fauchereau N, Trzaska S, Rouault M, Richard Y. Rainfall variability and changes in southern Africa during the 20th century in the global warming context. Nat Hazards. 2003;29:139–154. http://dx.doi.org/10.1023/A:1023630924100

29. Shongwe ME, Van Oldenborgh GJ, Van den Hurk BJJM, De Boer B, Coelho CAS, Van Aalst MK. Projected changes in mean and extreme precipitation in Africa under global warming. Part I: Southern Africa. J Climate. 2009;22(13):3819–3897. http://dx.doi.org/10.1175/2009JCLI2317.1

30. Chunzhen L. Research advances in impacts of natural climate variability and anthropogenic climate change on streamflow. Adv Clim Change Res. 2009;5:47–53.

31. Sambo SP. An investigation of the impacts of land use change on streamflow of the Luvuvhu River Catchment in the Limpopo Province of South Africa [honour's dissertation]. Thohoyandou: University of Venda; 2012.

32. Jewitt GPW, Garratt JA, Calder IR, Fuller L. Water resources planning and modelling tools for the assessment of land use change in the Luvuvhu Catchment, South Africa. Phys Chem Earth. 2004;29:1233–1241. http://dx.doi.org/10.1016/j.pce.2004.09.020

33. Odiyo JO, Makungo R, Maumela D. Impact of alien vegetation clearance on hydrology of Luvuvhu River Catchment. Paper presented at: 11th WaterNet/WARFSA/GWP-SA Symposium on Integrated Water Resources Management for National and Regional Integration. 2010 October 27–29; Victoria Falls, Zimbabwe. p. 551–569.

34. Reason CJC, Rouault M. ENSO-like decadal variability and South African rainfall. Geophys Res Lett. 2002;29(13):16-1–16-4.

35. Motulsky H. Prism 5 Statistics Guide Version 5.0. San Diego, CA: GraphPad Software Inc.; 2007.

36. Nury AH, Koch M, Md Alam JB. Analysis and prediction of time series variations of rainfall in North-Eastern Bangladesh. Brit J Appl Sci Technol. 2014;4(11):1644–1656. http://dx.doi.org/10.9734/BJAST/2014/7722

In-vitro effects of garlic extracts on pathogenic fungi *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba*

AUTHORS:
Chanel K. Daniel[1,2]
Cheryl L. Lennox[1]
Filicity A. Vries[2]

AFFILIATIONS:
[1]Fruit and Postharvest Pathology Research Programme, Department of Plant Pathology, Stellenbosch University, Stellenbosch, South Africa
[2]Agricultural Research Council-Infruitec/Nietvoorbij (The Fruit, Vine and Wine Institute of the Agricultural Research Council), Stellenbosch, South Africa

CORRESPONDENCE TO:
Filicity Vries

EMAIL:
VriesF@arc.agric.za

POSTAL ADDRESS:
Agricultural Research Council-Infruitec/Nietvoorbij, Helshoogte Road, Stellenbosch 7600, South Africa

DATES:
Received: 18 July 2014
Revised: 15 Oct. 2014
Accepted: 05 Nov. 2014

KEYWORDS:
plant extracts; antifungal activity; postharvest pathogens; sulphur compounds; volatiles

HOW TO CITE:
Daniel CK, Lennox CL, Vries FA. In-vitro effects of garlic extracts on pathogenic fungi *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba*. S Afr J Sci. 2015;111(7/8), Art. #2014-0240, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140240

The antifungal activity of garlic extracts applied directly and through volatile release was tested against the growth of postharvest pathogens *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba*. Mycelial growth of *B. cinerea* and *P. expansum* was inhibited by aqueous and ethanol dilutions on garlic extract amended media (direct method) in a dose-response manner. The aqueous dilution was more effective than the ethanol dilution. Both dilutions inhibited mycelial growth of *N. alba* to a similar extent but no trend in data was noted across the concentration range. Calculated EC_{50} values indicated that 13.36% and 8.09% aqueous dilutions could be used to inhibit growth of *B. cinerea* and *P. expansum*, respectively; however, values generated for *N. alba* either bordered on or exceeded the concentration range. The volatile vapour application of garlic was able to inhibit mycelial growth and spore germination of all pathogens at concentrations as low as 20%. Gas chromatography–mass spectrometry analysis showed that 85.95% of compounds present in the garlic sample belonged to a sulphur or sulphur-derived group. Allicin, the active component of garlic, was not found; however, breakdown products of allicin were present in high amounts. Overall, the antifungal activity of garlic extracts for the control of *B. cinerea* and *P. expansum* was confirmed. Further investigations into the antifungal effect of garlic extracts on *N. alba* is required, although garlic volatiles seem to be effective. This report is the first of antifungal activity of garlic extracts against *N. alba* – the causal agent of bull's eye rot, one of the major diseases of apples.

Introduction

Several pathogenic fungi, such as *Botrytis cinerea* Pers., *Penicillium expansum* (Link) Thom. and *Neofabraea alba* (E.J. Guthrie), are major infectious agents of apples, especially in the postharvest stage. Pathogenic fungi are controlled primarily through the use of synthetic fungicides; however, restrictions are being placed on the use of chemicals because of the perceived negative effects that pesticides may have on human health and the environment. Increasing regulations on the use of synthetic fungicides, build-up of chemical residues on the fruit and the emergence of pathogen resistance to the most frequently used fungicides, validate the search for novel biological control strategies.[1]

In recent years, a number of plant extracts, their essential oils and their volatile components have been reported to have strong antifungal activity.[2] In the agricultural sector, plant extracts, essential oils and their components are gaining increasing interest as a result of their volatility, reasonably safe status, eco-friendly and biodegradable properties, and wide consumer acceptance.[3] Some extracts and essential oils of 'medicinal' plants have been found to be effective against fungal and bacterial pathogens.[4] The fungicidal activity of essential oils from citrus, eucalyptus and thymus has been demonstrated. For example, in-vitro studies have shown that the oil of eucalyptus inhibits mycelial growth of important soilborne and postharvest disease pathogens such as *Pythium* spp*.,* *Rhizoctonia solani*[5,6] and *Collectotrichum gloeosporioides*[6]. Nosrati et al.[7] proposed that spearmint essential oil could be used in the control and management of *Fusarium oxysporum* f. sp. *radicis-cucumerinum* which is the causal organism of stem- and crown rot of greenhouse cucumber. Pawar and Thaker[8] found that the essential oils of lemongrass, clove, cinnamon bark, cinnamon leaf, cassia, fennel, basil and evening primrose had an antifungal effect on *Alternaria porri* and *Fusarium oxysporum* f. sp*. cicer*.

Garlic (*Allium sativum* L.) has been used for centuries for culinary purposes and its medicinal properties in traditional and conventional medicine are well documented.[9,10] The wide range of antifungal and antibacterial activities of garlic has been largely attributed to the presence of high concentrations of sulphur-containing compounds.[9,11]

Cavallito and Bailey[12] were responsible for the discovery of an oxygenated sulphur compound called allicin (diallyl thiosulphinate), which they considered to be responsible for the aroma and flavour of garlic. Since then, several researchers have attributed the antimicrobial action of garlic to allicin, which is present as the main active component.[11-14] The formation of allicin is followed by its rapid decomposition into sulphur-derived compounds such as diallyl disulphide, diallyl sulphide, diallyl trisulphide, sulphur dioxide, allyl propyl disulphide and diallyl tetrasulphide.[15,16] This fact has led some researchers to suspect that the antimicrobial activity may be a result of the action of a combination of sulphur and sulphur-related compounds.[14,17]

The antifungal effect of garlic on plant pathogens has been shown by Russel and Mussa[18] for the control of *Fusarium oxysporum* f.sp *phaseoli*. Investigations have also shown inhibitory effects of garlic against *Penicillium digitatum*.[19,20] Whilst many studies have highlighted the antimicrobial action of garlic on pathogens, little research has been done relating to postharvest plant pathogens,[19] and specifically the postharvest pathogens of apples.

In light of the above, in the present study we aimed to evaluate the antifungal efficacy of garlic extracts on the in-vitro mycelial growth and conidial germination of *B. cinerea*, *P. expansum* and *N. alba*. Gas chromatography–mass spectrometry (GC-MS) was employed to attain information on the chemical constituents of the garlic sample, in an effort to highlight the compounds potentially involved in the antimicrobial effect of garlic on pathogens.

Materials and methods

Pathogen isolation

Three pathogens – *B. cinerea* (B62-SUN), isolated from pears, and *P. expansum* (P1110-SUN) and *N. alba* (DOK7-SUN), both isolated from infected apples – were used. All three pathogens were obtained from the fungal collection of Stellenbosch University's Department of Plant Pathology. All the isolates were tested for pathogenicity on apples and pure isolates were prepared. Both *B. cinerea* and *P. expansum* were cultured on potato dextrose agar (PDA, pH 5.6, Merck, Johannesburg, South Africa) at 25 °C for 3 days for mycelial plugs and 7 days for the production of conidia. *Neofabraea alba* was cultured on acidified PDA (pH 3.5, Merck, Johannesburg, South Africa) for 1 month at 25 °C. Conidia were harvested by flooding the medium surface with sterile distilled water together with Tween 80 (0.05% w/v), and gently agitating the plate to dislodge spores. The final inoculum concentration was adjusted to 10^4 conidia/mL for each pathogen.

Preparation of garlic extract

Fresh garlic (*Allium sativum* L.) cloves were purchased from a retail store (Woolworths, Stellenbosch, South Africa). The garlic cloves were peeled and surface sterilised using ethanol (99.9% v/v). The garlic cloves were allowed to air dry before 800 g was weighed out and crushed in a blender. Ethanol (1 L) was added to the crushed garlic and the mixture was then placed into a glass container and incubated overnight at room temperature (20–25 °C). The extract was then filtered through a Büchner funnel using Whatman qualitative filter paper (No. 4). The filtrate was then subjected to a rotary evaporator (at 60–80 °C) to remove the ethanol. The filtrate was evaporated down to a final volume of 150 mL, yielding an extract with a semisolid consistency. This extract was considered to be the 100% concentrate and stored at 4 °C until subsequent use. The 100% extract was diluted down to make up the required concentrations used for efficacy testing.

Effect of garlic extract on mycelial growth

The effect of garlic extract on mycelial growth of *B. cinerea*, *P. expansum* and *N. alba* was determined following the poisoned food technique of Shahi et al.[21], with slight modification. A concentration range (80–2.5% w/v) of garlic extract was prepared by adding the requisite amount of the extract to ethanol or sterile distilled water up to a volume of 2.8 mL, which was then added to 140 mL of PDA medium (pH 5.6), and 20-mL aliquots of the amended PDA were poured into 90-mm Petri plates. Control sets consisted of unamended PDA.

Mycelial discs of 3 mm diameter cut out from the periphery of 3-day-old cultures (*B. cinerea* and *P. expansum*) and 7-day-old culture (*N. alba*) were aseptically transferred, mycelium side down, onto the surface of the agar. Petri plates were incubated at 25 °C for 3 days for *B. cinerea* and *P. expansum* and 7–21 days for *N. alba*. Radial mycelial growth was measured using digital calipers. Percentage of mycelial growth inhibition (MGI) was calculated as follows: MGI (%) = (dc-dt) x 100/dc, where dc is the mycelial growth diameter in control sets and dt is the mycelial growth diameter in treatment sets.

The nature of antifungal activity – fungistatic (temporary inhibition) or fungicidal (permanent inhibition) – of the garlic extract was determined by transferring the inhibited fungal discs from the above-mentioned method onto unamended PDA and observing growth. Three replicates were used for each of the three pathogens and for each concentration tested and the whole experiment was repeated once.

Effect of garlic extract on conidial germination

To determine the effect of garlic extract on conidial germination of *B. cinerea* and *P. expansum,* 100 µL of fungal conidia suspensions (10^4 conidia/mL) were pipetted onto the centre of garlic-amended PDA plates. Inoculated plates were incubated at 25 °C for 3 days. The control plates consisted of the pathogen on unamended PDA. Three replicates were used for each pathogen for each concentration. Plates were evaluated for germination (+) and non-germination (-) of conidia. The experiment was repeated once.

Effect of garlic volatiles on mycelial growth and conidial germination

A phytatray chamber assay was used to determine the effect of the volatile vapour of garlic extracts on all three pathogens in vitro. A glass Petri dish containing 5 mL of garlic extract, diluted with water to concentrations of 0%, 20%, 30% or 40% (wt/v), was fixed to the base of a disposable phytatray (Zibo, Cape Town, South Africa). Sterilised distilled water was used as the control. Four 65-mm PDA Petri plates inoculated with the respective fungi were fixed to the sides of the phytatray. Each chamber contained two plates inoculated with 3-mm mycelial plugs cut from the leading edge of an actively growing culture and placed mycelial side down onto the PDA, as well as two plates inoculated with 100 µL of a 10^4 conidia/mL conidial suspension by means of a spread plate method. The lid was closed and the chamber was then incubated at 20 °C (at 95% RH) and -0.5 °C (at 95% RH) for a total of 3 days for *B. cinerea* and *P. expansum* and 7 days for *N. alba,* before evaluation. Plates incubated at -0.5 °C were further incubated at 20 °C for 3 (*B. cinerea* and *P. expansum*) or 7 (*N. alba*) days. A total of three replicates with five phytatray chambers was used for each concentration. Plates inoculated with mycelial plugs were evaluated by measuring mycelial growth of the fungi using digital calipers and mycelial inhibition was calculated as described previously. Fungal spore plates were evaluated for germination (+) and non-germination (-) and subsequently converted to a percentage for statistical analysis.

GC-MS analysis of garlic extract

Approximately 1 mL of garlic crude extract was transferred to 20-mL solid phase microextraction vials for analysis. The vials were allowed to equilibrate for 2 min in the heating chamber of the CTC autosampler maintained at 30 °C. The volatile compounds were extracted by exposure of a 50/30 m divinylbenzene-carboxen-polydimethylsiloxane coated fibre (Supelco™, Port Edward, South Africa) on the headspace of the samples. Following extraction, desorption of the volatile compounds from the fibre coating was carried out for 10 min in the injection port of the GC–MS operated in splitless mode. The temperature of the injection port was maintained at 240 °C. Separation of the volatile compounds was performed on an Agilent 6890 N (Agilent, Palo Alto, CA, USA) gas chromatograph coupled with an Agilent 5975 MS (Agilent, Palo Alto, CA, USA) mass selective detector. Chromatographic separation was performed on a DB-FFAP (60-m length, 250-µm inner diameter and 0.5-µm film thickness) capillary column from Agilent technologies. Analyses were carried out using helium as a carrier gas with a flow rate of 1.9 mL/min operated in constant flow mode. The injector temperature was maintained at 240 °C.

The oven temperature was as follows: 70 °C for 1 min and then ramped up to 225 °C at 5 °C/min and held for 3 min. The mass selective detector was operated in full-scan mode and the ion source and quadropole were maintained at 230 °C and 150 °C, respectively. The transfer line temperature was maintained at 280 °C and total run time was approximately 46 min. Authentic standards were unavailable so compounds were tentatively identified by comparison with mass spectral libraries (NIST05 and Wiley 275.L). For quantification, the automatically calculated relative abundances were used and are expressed as a percentage. The sample was run twice, each time with three replicates.

Statistical analysis

In all cases, the experimental design was completely randomised. Conidial germination data were binary (present or absent), summed across the Petri dishes and expressed as a percentage. Mycelial growth was measured as a diameter (mm) and converted to percentage inhibition, which was analysed by an appropriate analysis of variance (ANOVA). The treatment means were compared using a Student's *t*-test with least significant difference at 5% ($p=0.05$).[22] A logarithmic growth curve was fitted to the concentration range to calculate the concentration at which 50% inhibition was achieved (EC_{50} values). The percentage inhibition and EC_{50} values were submitted to an appropriate ANOVA to compare treatments. Analysis was performed using SAS version 9.2 statistical software.[23] All three pathogens were analysed separately and no comparison was made between them.

Results

Effect of garlic extract on mycelial growth

A clear dose-response effect was obtained. The aqueous (sterile distilled water) and ethanol diluted extracts showed complete inhibition (100%) of *B. cinerea* at the higher concentrations (80% and 60%), with a fungicidal effect noted at both concentrations (Table 1) for both diluents tested. Aqueous diluted extract at a concentration of 40% showed 92.08% inhibition of *B. cinerea*. At a concentration of 80%, the aqueous and ethanol diluted extracts inhibited *P. expansum* by 96.21% and 99.21%, respectively. Ethanol diluted extracts seemed to be more effective against *N. alba* with 80% extract showing 79.63% inhibition. Overall, comparison between the diluents used indicated that the aqueous diluted extract provides significantly better results than the ethanol diluted extract (Table 1).

The effective concentrations at which 50% pathogen inhibition (EC_{50}) resulted from the use of garlic extracts were calculated. *B. cinerea* could be controlled using 20.59% of an aqueous diluted extract or 13.36% of an ethanol diluted extract. For *P. expansum*, a 19.95% ethanol diluted extract or 8.09% aqueous diluted extract could be used to retard pathogen growth. For *N. alba,* results indicated that 50% control could not be achieved unless the extract was at a concentration of no less than 79.51% (Table 2).

Effect of garlic extracts on conidial germination

Aqueous and ethanol diluted extracts were tested for their antifungal activity against conidial viability of the pathogens *B. cinerea* and *P. expansum* at concentration ranges from 0% to 80%. *B. cinerea* and *P. expansum* germinated when exposed to concentrations of 0–10% aqueous diluted extract (Table 3); however, germination of both pathogens was completely inhibited at the higher concentration ranges of 20–80% aqueous extract. For both pathogens exposed to the ethanol diluted extracts, conidial germination was completely inhibited at all concentrations except the lowest concentration of 2.5% (Table 3).

Effect of garlic volatiles on mycelial growth and conidial germination

The volatile vapours of aqueous diluted extracts were strongly active against mycelial growth of all the fungi, especially at concentrations of 30% and 40%. In all cases it was noted that the percentage mycelial inhibition increased with an increase in garlic concentration (Table 4).

Conidial germination of all fungi tested was almost completely inhibited by volatiles of garlic extracts (Table 4), irrespective of the concentration of the extract.

Mycelial growth for plates incubated at -0.5 °C showed complete inhibition against all pathogens tested, irrespective of the garlic extract concentration (Table 5), in comparison with the control, because of the low incubation temperature.

When the phytatrays from -0.5 °C were incubated further at 20 °C, concentrations of 20–40% were strongly active against mycelial growth of all fungi tested. The combination of garlic volatiles with low temperature (-0.5 °C) resulted in a stronger antifungal activity, especially against *P. expansum* and *N. alba,* than that exhibited for phytatrays that were only incubated at 20 °C, as is suggested by the difference between control and treatment sets.

Table 1: Inhibitory effect of aqueous (sterile distilled water) and ethanol diluted garlic extracts on mycelial growth of *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba*

		Inhibition (%)		
Garlic extracts	**Concentrations (% w/v)**	***B. cinerea***	***P. expansum***	***N. alba***
Aqueous extracts	0	0.00^{g}	0.00^{h}	0.00^{e}
	2.5	11.50^{e}	28.82^{f}	5.95^{de}
	5	12.91^{e}	29.48f	6.03^{de}
	10	36.36^{d}	60.03^{d}	30.08^{bc}
	20	52.05^{c}	70.60^{c}	6.43^{de}
	40	92.08^{a}	83.82^{b}	34.29^{bc}
	60	100.0^{a†}	95.68^{a}	40.15^{b}
	80	100.0^{a†}	96.21^{a}	25.05^{bc}
Ethanol extracts	0	0.00^{g}	0.00^{h}	0.00^{e}
	2.5	0.86^{g}	5.10^{g}	21.53^{cd}
	5	0.44^{g}	11.59^{g}	1.94^{e}
	10	1.62^{g}	10.92^{g}	4.31^{e}
	20	42.17^{c}	41.17^{e}	6.68^{de}
	40	69.20^{b}	72.88^{c}	7.26^{de}
	60	100.0^{a†}	85.40^{b}	39.26^{b}
	80	100.0^{a†}	99.21^{a}	79.63^{a}

Values represent means of measurements made on three independent plates per treatment. In each column, values followed by the same letter do not differ significantly ($p<0.05$*), as determined by a Student's* t*-test.*
[†]Indicates fungicidal effect (permanent inhibition).

Table 2: The effective concentrations of aqueous and ethanol diluted garlic extracts that resulted in 50% inhibition (EC_{50}) of mycelial growth of *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba* in vitro

	EC_{50} value[†]	
Pathogen	**Aqueous extract**	**Ethanol extract**
Botrytis cinerea	13.36^{a}	20.59^{b}
Penicillium expansum	8.09^{a}	19.95^{b}
Neofabraea alba	81.39^{b}	79.51^{a}

[†] EC_{50} values were determined on garlic amended potato dextrose agar growth medium. Values represent means of measurements made on three plates per pathogen. Mean values followed by the same letter(s) represent data that are not significantly different (p<0.05), as determined by a Student's t-test.

Table 3: Inhibitory effects of aqueous and ethanol diluted garlic extracts against conidial germination of *Botrytis cinerea* and *Penicillium expansum*

	Inhibitory effect of garlic extracts			
	Botrytis cinerea		***Penicillium expansum***	
Concentrations (%)	**Aqueous extract**	**Ethanol extract**	**Aqueous extract**	**Ethanol extract**
0	+	+	+	+
2.5	+	+	+	+
5	+	-	+	-
10	+	-	+	-
20	-	-	-	-
40	-	-	-	-
60	-	-	-	-
80	-	-	-	-

Plates were evaluated for germination (+) and non-germination (-) of conidia. The experiment was repeated once.

Table 4: Inhibitory volatile action of aqueous extracts against mycelial growth and conidial germination of *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba* in phytatrays incubated at 20 °C

	Inhibition (%)					
	Mycelial growth			**Conidial germination**		
Concentration (%)	***B. cinerea***	***P. expansum***	***N. alba***	***B. cinerea***	***P. expansum***	***N. alba***
0	0.00^{e}	0.00^{f}	0.00^{d}	0.00^{c}	0.00^{b}	0.00^{b}
20	61.55^{b}	75.41^{d}	100.0^{a}	100.0^{a}	96.67^{a}	100.0^{a}
30	97.14^{a}	89.60^{b}	100.0^{a}	100.0^{a}	100.0^{a}	100.0^{a}
40	99.35^{a}	100.0^{a}	100.0^{a}	100.0^{a}	100.0^{a}	100.0^{a}

Values followed by the same letter, down the column, do not differ significantly (p<0.05) according to a Student's t-test. The pathogens were not compared against each other (i.e. across rows).

Similar results were observed for conidial germination trays incubated at -0.5 °C (data not presented). The data showed that conidial germination of all three fungi was completely inhibited at -0.5 °C by concentrations of 20–40% when plates were further incubated. This outcome is in agreement with the results obtained for mycelial growth.

GC-MS analysis of garlic extract

From the GC-MS analysis of crude garlic extract, 43 volatile compounds were detected. Of this 43, 25 compounds were identified to be sulphur or sulphur-derived compounds, 2 compounds belonged to the alcohol group and one compound was an ester; the remaining compounds were not identified.

Sulphur and sulphur-derived compounds made up 85.95% of the entire sample concentration. The relative abundances (%) of all sulphur compounds identified in this study are presented in Table 6. Allyl methyl sulphide (7.93%), allyl methyl disulphide (7.86%), allyl methyl trisulphide (13.85%), diallyl disulphide (24.10%) and dimethyl trisulphide (11.36%) were present in the highest percentages within the sample. The chromatograph (Figure 1) highlights compound abundance relative to the retention time, in correspondence with samples listed in Table 6. The percentage relative abundances (%) presented in this study were calculated automatically using the peaks obtained from the chromatograph. Similar results were obtained for the second sample run, with the exception of the detection of three additional compounds: 3-vinyl-1,2-dithiacyclohex-4-ene, 3-vinyl-1,2-dithiacyclohex-5-ene and 1-oxa-4,6-diazacyclooctane-5-thione.

Discussion

Garlic extracts had a significant effect on the growth of the pathogens tested in this study. This finding is in agreement with earlier reports on the antifungal properties of garlic. The effect of garlic extracts on postharvest pathogens was determined by direct exposure as well as through volatile action.

Studies have shown that the method by which a plant extract is prepared will influence the type of activity (antifungal or other) it has.[24,25] Different extraction and dilution solvents will affect the extraction of different chemical compounds and the physiological properties within a plant, and therefore different extracts may contain different compounds, or the same compounds in varying quantities.[26,27]

Table 5: Inhibitory volatile action of aqueous diluted garlic extracts against mycelial growth of *Botrytis cinerea*, *Penicillium expansum* and *Neofabraea alba*, incubated at -0.5 °C with further incubation at 20 °C

	Inhibition (%)					
	B. cinerea[†]		***P. expansum***[†]		***N. alba***[‡]	
Concentration (%)	**-0.5 °C**	**20 °C**	**-0.5 °C**	**20 °C**	**-0.5 °C**	**20 °C**
0	0.00^{e}	0.00^{e}	0.00^{e}	0.00^{e}	100.0^{a}	0.00^{c}
20	100.0^{a}	69.41^{b}	100.0^{a}	97.76^{bc}	100.0^{a}	100.0^{a}
30	100.0^{a}	97.62^{a}	100.0^{a}	99.14^{ab}	100.0^{a}	100.0^{a}
40	100.0^{a}	97.12^{a}	100.0^{a}	100.0^{a}	100.0^{a}	100.0^{a}

[†] *Trays incubated at respective temperatures for 3 days.*
[‡] *Trays incubated at respective temperatures for 7 days.*
Values followed by the same letter, down the column, do not differ significantly ($p<0.05$*), as determined by a Student's* t*-test. The pathogens were not compared against each other (i.e. across rows).*

Table 6: Percentage composition of sulphur and sulphur-derived compounds in garlic extract

Retention time (min)	**Name of compound**[†]	**Molecular formula**	**Molecular weight (g/mol)**	**Percentage peak**[‡] **(%)**
3.36	Methanethiol	CH_4S	48.10	0.96
7.20	Dimethyl disulphide	$C_2H_6S_2$	94.18	3.16
9.05	Allyl methyl sulphide	C_4H_8S	88.15	7.93
13.44	Allyl methyl disulphide	$C_4H_8S_2$	120.21	11.23
17.45	Dimethyl trisulphide	$C_2H_6S_3$	126.24	11.36
18.82	Allyl propyl disulphide	$C_6H_{12}S_2$	148.26	0.06
21.14	Diallyl disulphide	$C_6H_{10}S_2$	146.25	24.10
25.63	Allyl methyl trisulphide	$C_4H_8S_3$	152.27	13.85
30.49	3.4-Dihydro-3-vinyl-1,2-dithiin	$C_6H_8S_2$	144.23	4.66
31.87	Diallyl trisulphide	$C_6H_{10}S_3$	178.31	5.07

[†] *Compound identification based on mass spectrum and retention index matching reference samples from the mass spectral libraries (NIST05 and Wiley 275.L).*
[‡] *Percentage reflected is the average of triplicate analysis expressed as the relative percentage of analyte in the garlic sample.*

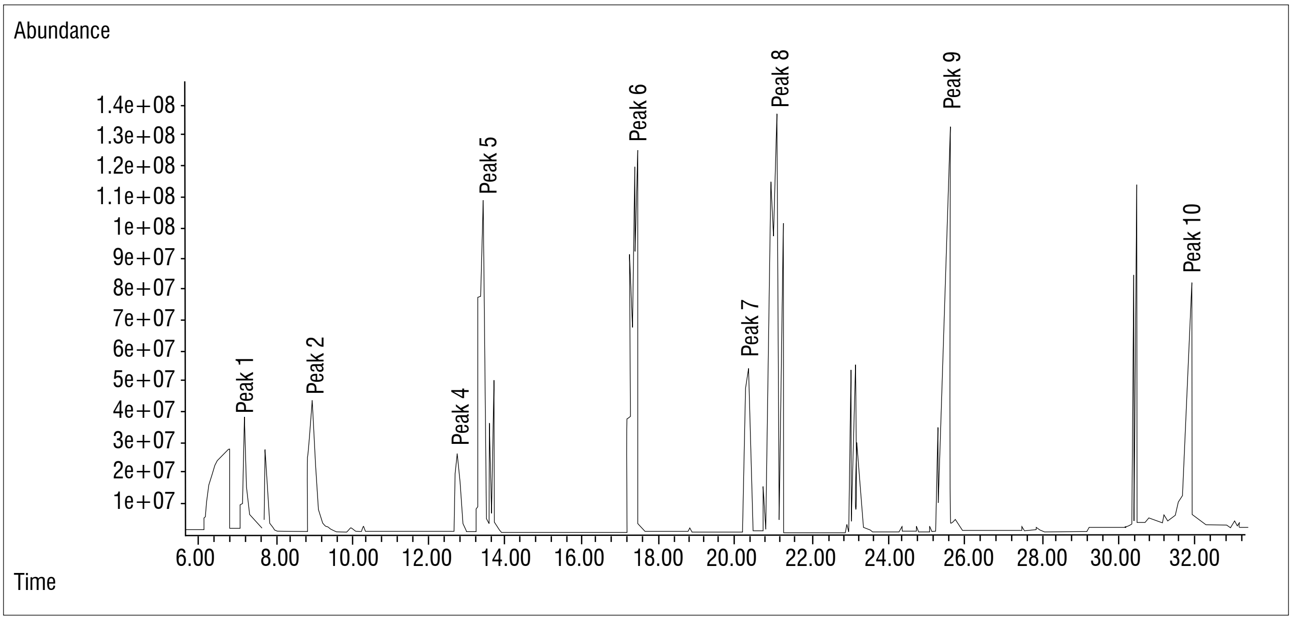


Figure 1: Gas chromatography–mass spectrometry (full scan) chromatogram showing the peaks corresponding to compounds present in the garlic extract in relation to retention time: peak 1, dimethyl disulphide; peak 2, diallyl sulphide; peaks 3 and 4, allyl methyl sulphide; peak 5, allyl methyl disulphide; peak 6, dimethyl trisulphide; peak 7, 1-oxa-4,6-diazacyclooctane-5-thione; peak 8, diallyl disulphide; peak 9, allyl methyl trisulphide; peak 10, diallyl trisulphide.

Aqueous dilutions of the garlic extract showed better activity than the ethanol dilutions of the extract. Previous studies both support[28] and contradict[29] this finding; however, all studies do support the fact that aqueous extract preparations do result in antimicrobial activity. When comparing aqueous preparations of extracts to preparations using other organic solvents tested against *B. cinerea*, a parallel may be drawn between results from the present study and those achieved by Senhaji et al.[27], in that the present study revealed that complete (100%) mycelial inhibition could be achieved by using a 60% aqueous diluted extract (Table 1). Timothy et al.[30] noticed a dose-dependent antifungal activity of leaf extracts of *Cassia alata* when tested against clinical isolates of pathogenic fungi, including a species of *Penicillium*. The same trend was noted in this study, with *B. cinerea* and *P. expansum* each eliciting a decrease in colony diameter with increasing concentration of garlic extract.

For the pathogen *N. alba*, percentage inhibition of mycelial growth was significantly different across the concentration range for both extract dilutions tested. The ethanol dilutions were more effective at reducing mycelial growth of the pathogen; however, results pertaining to the antifungal effect of garlic extracts on *N. alba* were inconclusive.

Plant extracts of *Allium* and *Capsicum* have been shown to completely inhibit spore germination of *B. cinerea*.[31] When garlic extracts were tested in this instance to determine the effect on conidial viability of *B. cinerea* and *P. expansum*, across the chosen concentration range, both pathogens behaved in an identical manner with regard to exposure to the aqueous and ethanol diluted extracts. Exposure to the ethanol extract yielded a greater inhibitory effect on pathogen conidial viability.

In recent years, studies have been carried out concerning the application of essential oils as antimicrobial agents,[32] with the majority of reports focused on the antifungal activity of essential oils and plant extracts exposed directly to fungus. However, few studies concerned the antifungal activity of volatiles of plant essential oils and extracts.[33]

An investigation into the effectiveness of garlic extracts for the control of pathogenic bacteria and fungi has demonstrated that allicin (the putative active ingredient of garlic) supplied via the vapour phase was effective in reducing *Phytophthora infestans* in vitro.[34] In the present in-vitro study, the volatiles released from extracts were effective in limiting mycelial growth and conidial germination of all the pathogens tested. The individual pathogens responded variably to the garlic extracts, with *N. alba* being the most sensitive to the volatile vapours.

At 20 °C, conidial germination and mycelial growth of all three pathogens were effectively inhibited by garlic volatiles across the concentration range, with pathogen inhibition increasing as the concentration of garlic extract volatiles increased (Table 4). Volatile substances released from essential oils derived from *Ocimum sanctum, Prunus persica* and *Zingiber officinale* were reported to have a similar effect on the control of *B. cinerea* on grapes.[35] A recent study showed that the vapours of thyme, peppermint and citronella oils caused a gradual inhibition of the growth of *P. expansum* and other postharvest pathogens.[36]

When the pathogens were incubated at -0.5 °C, the low temperature affected their growth, as indicated by the reduced growth of the control sets of *B. cinerea* and *P. expansum*. However, for both these pathogens, complete inhibition (100%) was noted on all plates exposed to the garlic treatments. The further incubation at 20 °C indicated that mycelial growth was inhibited, while conidial germination remained 100% inhibited (data not presented) for all three pathogens even at 20 °C. This finding suggests a synergistic relationship between the low temperature and the garlic extracts. Under standard conditions, incubation at -0.5 °C is expected to suppress growth of pathogens; however, the garlic extracts enhanced pathogen inhibition and allowed for added control of pathogens, as can be seen when comparing the control sets to the treated sets (Table 5).

No comparison was made between the pathogens in this study but the results indicate that while all three pathogens were sensitive to the garlic extracts, each individual pathogen reacted differently to the extracts. When exposed to garlic volatiles, *N. alba* was most sensitive to the extracts at both temperatures tested, followed by *P. expansum* and *B. cinerea*.

GC-MS has been previously employed to profile chemical compounds in garlic and other plant essential oils. Available literature on garlic composition reveals compounds common across the various studies, but also isolated detection of compounds, suggesting that even though this method is quite sensitive to its purpose, variation in compound detection will occur between sample sets. This variation is probably for a variety of reasons which include the type of analysis carried out, the conditions surrounding the study and also the garlic sample itself – with sample preparation and cultivar type also playing a role in the compounds that would be amplified.[10]

Over 35 different compounds have been identified in garlic,[10] with the sulphur-containing compounds the main focus of research studies conducted on garlic and related species. The full profile analysis in this study rendered a total of 43 compounds highlighted within the garlic sample and further investigation found that the total amount of sulphur-containing compounds made up approximately 85.95% of the sample tested. Allicin (diallyl thiosulphinate) could not be directly detected in this study; however, it has been reported that allicin decomposes to diallyl disulphide, diallyl trisulphide and sulphur dioxide[15] and these compounds, together with other volatiles typically present in crushed garlic, were found in relatively high amounts in this study. Furthermore, exposure time between extract preparation and analysis could be integral to detecting allicin, because of its rapid decay rate.

The major sulphides that have been identified in garlic include diallyl sulphide, allylmethyl, dimethyl- and mono-tohexasulphides together with small amounts of allyl 1-propyl and methyl 1-propyl, and di-, tri- and tetrasulphides,[15] although different studies[9,10,37] including the present study, have reported different amounts of these compounds. Khadri et al.[9] reported that the two major compounds present in a garlic sample tested were methyl allyl trisulphide (34.61%) and diallyl disulphide (31.65%). Both of these compounds were found in the sample tested in this study, but at lower concentrations of 13.85% and 24.10%, respectively. According to the authors, no other reports of allyl methyl trisulphide had been made previously and they concluded that the cultivar used represented a new chemotype typical of eastern Algeria[9]; however, this cannot be the case as the garlic used in the present study was not sourced from that geographical region. The compounds 3-vinyl-1,2-dithiacyclohex-4-ene and 3-vinyl-1,2-dithiacyclohex-5-ene have been reported as the compounds responsible for allinase activity.[38] Another compound detected, which is worth mentioning, is 1-oxa-4,6-diazacyclooctane-5-thione, which was found to be present in ‘rosy garlic’ (*Allium roseum* L.) and was reported in the study to not have been recorded in the literature.[39]

As a result of various compounds highlighted in different studies on components of garlic, Amagase[15] speculated that, while garlic is recognised for the abundance of sulphur compounds present, perhaps compounds other than allicin could contribute to the various antimicrobial activities. The present study supports this hypothesis as allicin was not found in the sample tested; however, other sulphur and sulphur-derived compounds were found in high amounts. The possibility exists that a complex of compounds, rather than one individual compound, is responsible for the antifungal activity noted by garlic samples throughout this study. It is recommended that if individual compounds can be sourced then each individual compound should be subjected to an antimicrobial screening to determine whether or not it makes any contribution to the antimicrobial action of garlic extracts.

In conclusion, we have shown that garlic extracts can have a significant effect on preventing the growth of *B. cinerea* and *P. expansum*. However, growth of *N. alba* was not significantly suppressed by the garlic. The solvent used for dilution concentrations (water or ethanol) had an influence on the antifungal activity of the garlic extract. Aqueous dilutions of the extract had greater antifungal activity than ethanol diluted extracts, possibly because the longer the extracts were exposed to ethanol, the more the antifungal activity was reduced.

When tested in the vapour phase, garlic extracts were able to control growth of *B. cinerea, P. expansum* and *N. alba*. Our findings confirm those of fellow researchers who stated that application in the vapour phase is preferred because of increased volatile activity and the ability to use lower concentrations.[40] In the present study, concentrations used in the volatile experiment were at a lower garlic concentration than the amended media experiments. Furthermore, increased antifungal activity was noted. This finding is significant as it gives a possible lead into using a garlic preparation as a fumigant to control pathogens that may be present in the air and on regular surfaces in a pack house or containers. Also, where garlic extracts are combined with effective storage conditions, this application could be adopted into a closed packaging system.

In conclusion, volatile vapour of garlic extracts showed more potent antifungal activity against conidial germination than against mycelial growth of the test fungi. Volatile vapours of garlic extracts were more effective than the direct method, as efficacy in volatile assays was at concentrations of 20–40% compared with concentrations of 60–80% for the direct method.

Acknowledgements

We thank the Agricultural Research Council (ARC) and the Technology and Human Resources for Industry Programme (THRIP) for financing this study, and research technicians at ARC-Infruitec/Nietvoorbij, Mr E Williams, Ms S Adonis and Ms L Gordon, for technical support.

Authors’ contributions

C.K.D. is a master’s student and performed all the laboratory experiments and wrote the manuscript; C.L. supervised the study; and F.V. initiated the project and supervised the project at the Agricultural Research Council where most of the experiments were performed.

References

1. Dellavalle PD, Cabrera A, Alem D, Larrañaga P, Ferreira F, Rizza MD. Antifungal activity of medicinal plant extracts against phytopathogenic fungus *Alternaria* spp. Chilean J Agric Res. 2011;71:231–239. http://dx.doi.org/10.4067/S0718-58392011000200008

2. Siripornvisal S, Rungprom W, Sawatdikarn S. Antifungal activity of essential oils derived from medicinal plants against grey mould (*Botrytis cinerea*). Asian J Food Agro-Ind. 2009;(special issue):229–223.

3. Tzortzakis NG, Economakis CD. Antifungal activity of lemongrass (*Cymbopogon citrates* L.) essential oil against key postharvest pathogens. Inn Food Sci Emerg Technol. 2007;8:253–258. http://dx.doi.org/10.1016/j.ifset.2007.01.002

4. Amini M, Safaie N, Salmani MJ, Shams-Bakhsh M. Antifungal activity of three medicinal plant essential oils against some phytopathogenic fungi. Trakia J Sci. 2012;10:1–8.

5. Katooli N, Maghsodlo R, Razavi SE. Evaluation of *Eucalyptus* essential oil against some plant pathogenic fungi. J Plant Breed Crop Sci. 2011;3:41–43.

6. Huy JS, Ahn SY, Koh YJ. Antimicrobial properties of cold-tolerant *Eucalyptus* species against phytopathogenic fungi and food-borne bacterial pathogens. Plant Path J. 2000;16:286–289.

7. Nosrati S, Esmaeilzadeh-Hosseini SA, Sarpeleh A, Soflaei-Shahrbabak M, Soflaei-Shahrbabak Y. Antifungal activity of (*Mentha spicata L.*) essential oil on *Fusarium oxysporum* f. sp. *radicis-cucumerinum* the causal agent of stem and crown rot of greenhouse cucumber in Yazd, Iran. ICEAE. 2011;15:52–56.

8. Pawar C, Thaker VS. Evaluation of the anti *Fusarium oxysporum* f. sp. *cicer* and *Alternaria porri* effect of some essential oils. World J Micro Biotech. 2007;23:1099–1106. http://dx.doi.org/10.1007/s11274-006-9339-6

9. Khadri S, Boutefnouchet N, Dekhil M. Antibacterial activity evaluation of *Allium sativum* essential oil compared to different *Pseudomonas aeruginosa* strains in Eastern Algeria. St. Cerc. St. CICBIA. 2010;11:421–428.

10. Clemente JG, Williams JD, Cross M, Chambers CC. Analysis of garlic cultivars using head space solid phase microextraction/gas chromatography/mass spectroscopy. Open Food Sci J. 2011;6:1–4. http://dx.doi.org/10.2174/1874256401206010001

11. Ankri S, Mirelman D. Antimicrobial properties of allicin from garlic. Microbes Infect. 1999;2:125–129. http://dx.doi.org/10.1016/S1286-4579(99)80003-3

12. Cavallito C, Bailey JH. Allicin, the antibacterial principle of *Allium sativum*. Isolation, physical properties and antibacterial action. J Am Chem Soc. 1944;66:1950–1952. http://dx.doi.org/10.1021/ja01239a048

13. Bocchini P, Andalo C, Pozzi R, Galletti GC, Antonelli A. Determination of diallyl thiosulfinate (allicin) in garlic (*Allium sativum* L.) by high-performance liquid chromatography with a post-column photochemical reactor. Anal Chim Acta. 2001;441:37–43. http://dx.doi.org/10.1016/S0003-2670(01)01104-7

14. Josling P. Preventing the common cold with a garlic supplement: A double-blind, placebo controlled survey. Adv Nat Ther. 2001;18:4. http://dx.doi.org/10.1007/bf02850113

15. Amagase H. Significance of garlic and its constituents in cancer and cardiovascular disease: Clarifying the real bioactive constituents of garlic. J Nutr. 2006;2:716–725.

16. Verma SK, Jain V, Verma D. Garlic – 'The spice of life': Composition, cooking chemistry and preparations. J Herbal Med Toxic. 2008;2:21–28.

17. Harris JC, Cottrell S, Lloyd D. Antimicrobial properties of *Allium sativum* (garlic). Appl Microbiol Biotechnol. 2001;57:282–286. http://dx.doi.org/10.1007/s002530100722

18. Russel PE, Mussa AEA. The use of garlic (*Allium sativum*) extracts to control foot rot of *Phaseolus vulgaris* caused by *Fusarium solani* f.sp. *phaseoli*. Ann Appl Biol. 1977;86(Abstr.):369–372.

19. Obagwu J, Korsten L. Control of citrus green and blue moulds with garlic extracts. Euro J Plant Pathol. 2003;109:221–225. http://dx.doi.org/10.1023/A:1022839921289

20. Kanan GJ, Al-Najar RA. *In vitro* antifungal activities of various plant crude extracts and fractions against citrus postharvest disease agent *Penicillium digitatum*. Jordan J Biol Sci. 2008;1:89–99.

21. Shahi SK, Patra M, Shukla AC, Dikshit A. Use of essential oil as botanical-pesticide against post harvest spoilage in *Malus pumilo* fruits. BioControl. 2003;48:223–232. http://dx.doi.org/10.1023/A:1022662130614

22. Ott RL. An introduction to statistical methods and data analysis. Belmont, CA: Duxbury Press; 1993.

23. SAS Institute. SAS version 9.2 64 bit. Cary, NC: SAS Institute Inc.; 2012.

24. Arora SD, Kaur GJ. Antibacterial activity of some Indian medicinal plants. J Nat Med. 2007;61:313–317. http://dx.doi.org/10.1007/s11418-007-0137-8

25. Raghavendra MP, Satish S, Raveesha KA. Alkaloid extracts of *Prosopis juliflora* (Sw.) DC. (Mimosaceae) against *Alternaria alternata*. J Biopest. 2009;2:56–59.

26. Mendonca-Filho RR. Bioactive phytocompounds: New approaches in the phytosciences. In: Ahmad I, Aqil F, Owais M, editors. Modern phytomedicine: Turning medicinal plants into drugs. Weinheim: Wiley-VCH Verlag GmbH & Co.; 2006. p. 1–24. http://dx.doi.org/10.1002/9783527609987.ch1

27. Senhaji B, Ben Hmamou D, Salghi R. *Asteriscus imbricatus* extract: Antifungal activity and anticorrosion inhibition. Int J Electrochem Sci. 2013;8:6033–6046.

28. Gull I, Saeed M, Sahukat H, Aslam SM, Samra ZQ, Athar AM. Inhibitory effect of *Allium sativum* and *Zingiber officinale* extracts on clinically important drug resistant pathogenic bacteria. Ann Clin Microbiol Antimicrob. 2012;11:8. http://dx.doi.org/10.1186/1476-0711-11-8

29. Saravanan P, Ramya V, Sridhar H, Balamurugan V, Umamaheswari S. Antibacterial activity of *Allium sativum* L. on pathogenic bacterial strains. Glob Vet. 2010;4:519–522.

30. Timothy SY, Wazis CH, Adati RG, Maspalma ID. Antifungal activity of aqueous and ethanolic leaf extracts of *Cassia alata* Linn. J Appl Pharm Sci. 2012;2:182–185. http://dx.doi.org/10.7324/japs.2012.2728

31. Wilson CL, Solar JM, El Ghaouth A, Wisniewski ME. Rapid evaluation of plant extracts and essential oils for antifungal activity against *Botrytis cinerea.* Plant Dis. 1997;81:204–210. http://dx.doi.org/10.1094/PDIS.1997.81.2.204

32. Barratta MT, Dorman HJD, Deans SG, Figueiredo C, Barroso JG, Ruberto G. Antimicrobial and antioxidant properties of some commercial essential oils. Flavour Fragr J. 1998;13:235–244. http://dx.doi.org/10.1002/(SICI)1099-1026(1998070)13:4<235::AID-FFJ733>3.0.CO;2-T

33. Chee HY, Lee MH. Antifungal activity of clove essential oil and its volatile vapour against dermatophytic fungi. Mycobiology. 2007;35:241–243. http://dx.doi.org/10.4489/MYCO.2007.35.4.241

34. Curtis H, Noll U, Stormann J, Slusarenko AJ. Broad-spectrum activity of the volatile phytoantocipin allicin in extracts of garlic (*Allium sativum* L.) against plant pathogenic bacteria, fungi and oomycetes. Physiol Mol Plant Pathol. 2004;65:79–89. http://dx.doi.org/10.1016/j.pmpp.2004.11.006

35. Tripathi P, Dubey NK, Shukla AK. Use of some essential oils as postharvest botanical fungicides in the management of grey mould of grapes caused by *Botrytis cinerea*. World J Microbiol Biotechnol. 2008;24:39–46. http://dx.doi.org/10.1007/s11274-007-9435-2

36. Sellamuthu PS, Sivakumar D, Soundy P. Antifungal activity and chemical composition of thyme, peppermint and citronella oils in vapour phase against avocado and peach postharvest pathogens. J Food Safe. 2013;33:86–93. http://dx.doi.org/10.1111/jfs.12026

37. Borrego S, Valdes O, Vivar I, Guiamet P, Battistoni P, Gomez de Saravia S, et al. Essential oils of plants as biopesticides against microorganisms isolated from Cuban and Argentine documentary heritage. ISRN Microbiology. 2012:1–7.

38. Chen Z, Zhang H, Liu B, Yang G, Aboul-Enein HY, Wang W, et al. Determination of herbicide residues in garlic by GC-MS. Chromatographia. 2007;66:887–891. http://dx.doi.org/10.1365/s10337-007-0425-1

39. Zouari S, Ketata M, Boudhrioua N, Ammar E. *Allium roseum* L. volatile compounds profile and antioxidant activity for chemotype discrimination – Case study for the wild plant of Sfax (Tunisia). Indust Crops Prod. 2013;41:172–178. http://dx.doi.org/10.1016/j.indcrop.2012.04.020

40. Laird K, Phillips C. Vapour phase: A potential future use for essential oils as antimicrobials? Lett Appl Microbiol. 2011;54:169–174. http://dx.doi.org/10.1111/j.1472-765X.2011.03190.x

Directed genetic modification of African horse sickness virus by reverse genetics

AUTHORS:
Elaine Vermaak[1]
Duncan J. Paterson[2]
Andele Conradie[1]
Jacques Theron[1]

AFFILIATIONS:
[1]Department of Microbiology and Plant Pathology, University of Pretoria, Pretoria, South Africa
[2]Sir William Dunn School of Pathology, University of Oxford, Oxford, United Kingdom

CORRESPONDENCE TO:
Jacques Theron

EMAIL:
jacques.theron@up.ac.za

POSTAL ADDRESS:
Department of Microbiology and Plant Pathology, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

DATES:
Received: 28 Sep. 2014
Revised: 06 Nov. 2014
Accepted: 08 Nov. 2014

KEYWORDS:
orbivirus; core transcripts; genome modification; vaccine

HOW TO CITE:
Vermaak E, Paterson DJ, Conradie A, Theron J. Directed genetic modification of African horse sickness virus by reverse genetics. S Afr J Sci. 2015;111(7/8), Art. #2014-0331, 8 pages. http://dx.doi.org/10.17159/sajs.2015/20140331

African horse sickness virus (AHSV), a member of the *Orbivirus* genus in the family Reoviridae, is an arthropod-transmitted pathogen that causes a devastating disease in horses with a mortality rate greater than 90%. Fundamental research on AHSV and the development of safe, efficacious vaccines could benefit greatly from an uncomplicated genetic modification method to generate recombinant AHSV. We demonstrate that infectious AHSV can be recovered by transfection of permissive mammalian cells with transcripts derived in vitro from purified AHSV core particles. These findings were expanded to establish a genetic modification system for AHSV that is based on transfection of the cells with a mixture of purified core transcripts and a synthetic T7 transcript. This approach was applied successfully to recover a directed cross-serotype reassortant AHSV and to introduce a marker sequence into the viral genome. The ability to manipulate the AHSV genome and engineer specific mutants will increase understanding of AHSV replication and pathogenicity, as well as provide a tool for generating designer vaccine strains.

Introduction

African horse sickness, of which African horse sickness virus (AHSV) is the aetiological agent, is a non-contagious disease of equids that is transmitted by *Culicoides* biting midges. African horse sickness is the most devastating of all equine diseases, with a mortality rate that can exceed 90% in susceptible horse populations.[1] Although endemic to sub-Saharan Africa, AHSV sporadically escapes from its geographical area and outbreaks of the disease have occurred in North Africa, the Middle East, the Arabian Peninsula and Mediterranean countries.[2,3] As a consequence of its severity, economic impact and ability to spread rapidly from endemic regions, African horse sickness is listed by the World Organization for Animal Health as a notifiable equine disease.

AHSV is a member of the *Orbivirus* genus in the Reoviridae family and has a genome composed of 10 segments of linear double-stranded RNA (dsRNA), designated from segment 1 (S1) to S10 in decreasing order of size.[4] The viral genome encodes seven structural proteins and at least three non-structural proteins (NS1 to NS3). Structural proteins VP1, VP3, VP4, VP6 and VP7 form the virus core particle, which is surrounded by an outer capsid layer composed of VP2 and VP5.[5] Shortly after cell entry by the virus, the outer capsid proteins are removed and the transcriptionally active core particle is released into the host cell cytoplasm. Within the core particle, each of the dsRNA genome segments is repeatedly transcribed by the core-associated enzymes VP1 (RNA-dependent RNA polymerase), VP4 (capping enzyme) and VP6 (helicase), resulting in extrusion of newly synthesised, capped, viral single-stranded RNA (ssRNA). The extruded transcripts, in turn, function as templates for the synthesis of viral proteins and also act as templates for the synthesis of genomic dsRNA following their encapsidation in progeny viral cores.[6,7]

Although a combination of mutagenesis and re-expression of AHSV proteins in heterologous hosts has allowed progress to be made regarding the structure–function relationships of selected AHSV proteins,[8-13] definitive roles for many of these proteins in the context of a replicating virus remain unresolved. Indeed, a major obstacle to studies aimed at understanding AHSV biology has been the inability to genetically manipulate the segmented dsRNA viral genome. However, recent technological advances for members of the Reoviridae family have demonstrated that it is possible to recover genetically engineered recombinant viruses through a reverse genetics approach. Despite variations in their molecular design and methodology, these reverse genetics systems all share a common feature, which is the availability of cloned complementary DNA (cDNA) copies of the viral genomes that can be genetically modified and manipulated to generate live viruses containing precisely engineered changes in their genomes.[14-17] Recently, a plasmid-based reverse genetics system was reported for AHSV.[18] This system requires the construction of recombinant plasmids in which cDNA copies of the 10 viral genome segments are each cloned under the control of a T7 RNA polymerase promoter in order to synthesise synthetic T7 viral transcripts. Transfection of permissive cells with these in–vitro synthesised synthetic T7 transcripts resulted in the recovery of infectious AHSV, indicating that the synthetic T7 transcripts can serve as functional substitutes of the authentic viral transcripts. A variation of this approach requires the construction of additional expression helper plasmids to synthesise the AHSV inner core proteins and two non-structural proteins in permissive cells, followed by transfection of these cells with all 10 synthetic T7 transcripts.[18] It can be envisaged that a reverse genetics approach which is not dependent on extensive cloning procedures may represent an attractive alternative to these plasmid-based reverse genetics systems.

Our objective in this study was to establish an efficient, broadly applicable method to generate recombinant AHSV. Here we show that transfection of permissive mammalian cells with a mixture of purified AHSV core-derived transcripts and an in-vitro synthesised T7 transcript derived from a polymerase chain reaction (PCR) amplicon leads to the production of recombinant AHSV. The performance of this system was validated by isolation of both a viable cross-serotype single-gene reassortant virus and a mutant virus containing a defined mutation in the replicating viral genome. The implications of targeted alterations of the AHSV genome are highly significant to research focused on basic studies of AHSV biology and the development of recombinant vaccine strains.

Materials and Methods

Cells and viruses

BSR cells (a clone of baby hamster kidney-21 cells) were cultured at 37 °C and 5% CO$_2$ in Eagle's minimum essential medium (EMEM) supplemented with Earle's balanced salt solution, 2 mM L-glutamine, 1% (v/v) non-essential amino acids, 5% (v/v) foetal bovine serum (FBS) and antibiotics (100 U/mL penicillin, 100 μg/mL streptomycin, 25 μg/mL amphotericin B) (HyClone Laboratories, Logan, UT, USA).

AHSV serotypes 3 (AHSV-3) and 4 (AHSV-4) were used for cell infections. Cell monolayers were rinsed twice with incomplete EMEM (lacking FBS and antibiotics) and then infected with the virus at the desired multiplicity of infection (MOI). Virus adsorptions were performed at 37 °C for 1 h, followed by incubation in complete EMEM.

Isolation of AHSV core particles

BSR cell monolayers were infected with AHSV-3 or AHSV-4 at a MOI of 0.08 plaque-forming units (pfu) per cell and harvested at 72 h post-infection. Core particles were purified using a modification of the method described by Mertens et al.[19] The virus-infected cells were lysed by incubation on ice for 30 min in chilled lysis buffer (100 mM Tris-HCl [pH 8.8], 10 mM ethylenediaminetetraacetic acid, 50 mM NaCl, 0.5% [v/v] NP-40), and the cell debris and nuclei were removed by centrifugation at 1000 *g* for 10 min at 4 °C. To remove the outer capsid proteins of virions, α-chymotrypsin (Sigma-Aldrich, St. Louis, MO, USA) was added to the cytoplasmic extract to a final concentration of 60 μg/mL, followed by incubation at 37 °C for 1 h. *N*-lauroylsarcosine (Merck, Darmstadt, Germany) was then added to 0.2% (w/v) final concentration, and incubation was continued at 25 °C for 1 h. The sample was subjected to high-speed ultracentrifugation (141 000 *g* for 2 h at 20 °C) through a 1-mL cushion of 40% (w/v) sucrose, prepared in 600 mM MgCl$_2$ and 20 mM Tris-HCl (pH 8.0). The pelleted core particles were suspended in 20 mM Tris-HCl (pH 8.0).

Characterisation of AHSV core particles

Proteins from purified cores were resolved by 10% sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) and visualised by staining of the gel with Coomassie brilliant blue (Merck, Darmstadt, Germany). Aliquots of the purified core particles were also adsorbed onto Formvar/carbon-coated 300-mesh copper grids for 90 s, rinsed with distilled water and then negatively stained with 2% (w/v) uranyl acetate. The grids were examined under a JEOL 2100F transmission electron microscope at 200 kV.

Synthesis and purification of AHSV core-derived transcripts

Core transcripts were synthesised in vitro at 30 °C for 6 h by incubating the AHSV cores (75 μg/mL) in AHSV core transcription buffer (100 mM Tris-HCl [pH 8.0], 4 mM ATP, 2 mM GTP, 2 mM CTP, 2 mM UTP, 500 μM S-adenosylmethionine, 6 mM DTT, 9 mM MgCl2, 1 U/μL of RiboLock RNase Inhibitor [Promega, Madison, WI, USA]). The sample was centrifuged at 20 000 *g* for 30 min at 4 °C and the supernatant was subjected to a second, identical centrifugation step to ensure removal of the core particles. The core particles were suspended in 20 mM Tris-HCl (pH 8.0) for re-use in subsequent in-vitro transcription reactions. The core transcripts were precipitated from the supernatant by addition of 8 M LiCl to a final concentration of 2 M, followed by incubation at 4 °C for 16 h. Following centrifugation as above, the pelleted core transcripts were suspended in 100 μL of sterile diethylpyrocarbonate (DEPC)-treated water and purified with the Nucleospin RNA Clean-up kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's protocol. Purified core transcripts were mixed with an equal volume of denaturing RNA loading buffer (Thermo Fisher Scientific Inc., Waltham, MA, USA) and analysed by electrophoresis on a 1% (w/v) agarose gel.

Preparation of templates for synthesis of synthetic T7 transcripts

All molecular biology and cloning procedures were performed using standard methodology.[20] For reassortant virus studies, recombinant plasmid pJET/blunt-S10, which contains a cloned cDNA copy of the full-length S10 genome segment of AHSV-4, was used as a template in a PCR with the 5' end primer NS3T7F (5'-CCGGTAATACGACTCACTATAGTTAAAATTATCCC-3'; T7 RNA polymerase promoter underlined) and the 3' end-specific primer NS3R (5'-GTAAGTTGTTATCCCACTCCCTAGAAAACG-3'). The resulting amplicon was purified from an agarose gel with the Zymoclean Gel DNA Recovery kit (Zymo Research Corporation, Irvine, CA, USA) and used as the template in T7 transcription reactions.

A mutant version of the AHSV-4 S10 genome segment, containing an introduced *Eco*RI restriction enzyme recognition sequence at nucleotide position 270, was constructed by a megaprimer PCR-based mutagenesis strategy in which three primers and two PCRs were used.[21] In the first round of PCR, pJET/blunt-S10 was used as the template together with primer NS3T7F and an internal mutagenic primer (NS3mR: 5'-ACGTATGGGAATTCGTTCTGGATCAC-3'; *Eco*RI sequence underlined). The purified amplicon was then used as a 'megaprimer' along with primer NS3R in the second round of PCR to generate the full-length S10 genome segment containing the newly introduced sequence. The amplicon was purified from an agarose gel and then blunt-end cloned into the *Sma*I site of pUC19 to generate pUC-S10E. The mutation was subsequently verified by nucleotide sequencing using the BigDye Terminator Cycle Sequencing Ready Reaction kit v.3.1 (Perkin Elmer, Foster City, CA, USA), followed by resolution on a Model 377 DNA sequencer (Perkin Elmer, Foster City, CA, USA). The AHSV-4 mutant S10 genome segment was PCR amplified from pUC-S10E with primers NS3T7F and NS3R, and used as the template in T7 transcription reactions. Transcription of the respective amplified products with T7 RNA polymerase is expected to yield synthetic T7 transcripts with authentic virus terminal sequences.

In-vitro T7 transcription reactions

T7 transcripts, containing a 5' cap analogue, were synthesised from the above purified amplicons with the MessageMAX T7 ARCA-capped Message Transcription kit (Epicentre, Madison, WI, USA) using the manufacturer's protocol. In these reactions, a ratio of 4:1 of anti-reverse cap analogue to rGTP was used. Following removal of the DNA template with RNase-free DNase I by incubation at 37 °C for 15 min, the T7 transcripts were purified with the Nucleospin RNA Clean-up kit (Macherey-Nagel, Düren, Germany).

Cell transfections and recovery of infectious virus

For recovery of AHSV from core transcripts, BSR cell monolayers at 80% confluence in 24-well tissue culture plates were transfected twice (16 h apart) with 800 ng of the core transcripts using Lipofectamine LTX reagent (Life Technologies, Carlsbad, CA, USA) according to the manufacturer's instructions. The same protocol was used for recovery of directed reassortant AHSV, except that 400 ng each of AHSV-3 core transcripts and the AHSV-4 S10 genome segment T7 transcript were mixed prior to transfection of BSR cells. For recovery of a mutant AHSV-4 virus, BSR cells were likewise transfected with a mixture consisting of 400 ng each of AHSV-4 core transcripts and the AHSV-4 mutant S10 genome segment T7 transcript. At 72 h post-second transfection, the cells were lysed and used for infection of confluent BSR cell monolayers in 6-well tissue culture plates. Virus adsorption was performed at 37 °C for 1 h, after which the lysate was replaced with a 1-mL overlay consisting of complete EMEM and 0.5% (w/v) agarose. At 72 h post-infection, plaques were visualised by staining the cell monolayers with 0.1% (w/v) MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide). Fresh BSR cell monolayers were infected with virus eluted from single plaques and the cells were incubated for 72 h to allow amplification of the virus.

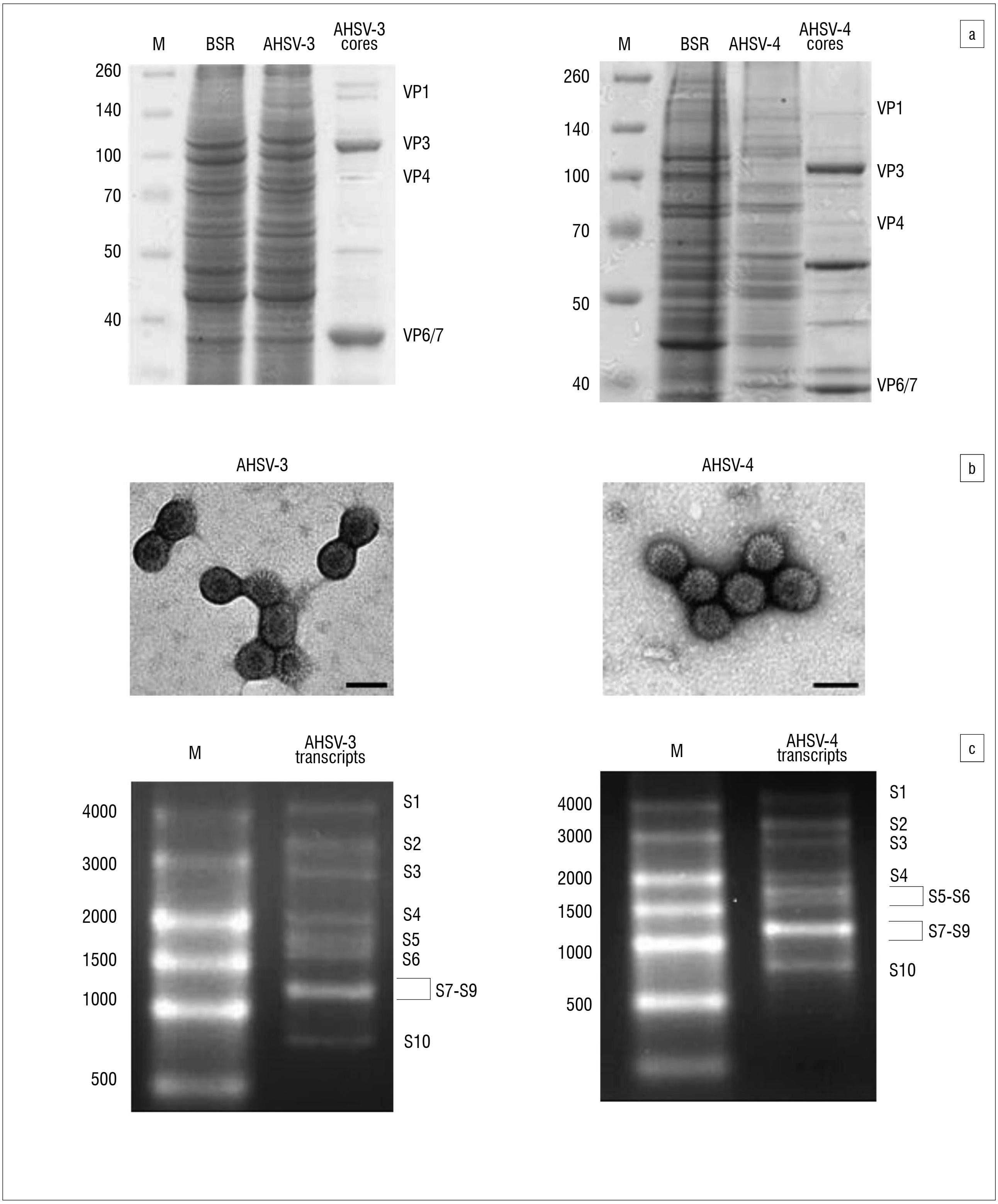


Figure 1: Characterisation of purified African horse sickness virus (AHSV) core particles and synthesis of core-derived transcripts. (a) SDS-polyacrylamide gel electrophoresis of purified AHSV-3 and AHSV-4 core particles. Uninfected (BSR) and virus-infected cells were included for comparative purposes. The sizes of reference protein markers (M, in kDa) are indicated on the left of the figure, and the AHSV core proteins on the right. (b) Transmission electron micrographs of negatively stained purified core particles (scale bar = 60 nm). (c) Agarose gels showing transcripts synthesised in vitro from purified core particles. The sizes of ssRNA markers (in nucleotides) are indicated on the left of the figure, and the core transcripts are indicated on the right.

RNase A and Proteinase K treatment of AHSV core transcripts and cores

BSR cell monolayers were transfected, as described above for the recovery of AHSV from core transcripts, with 800 ng of AHSV-4 core transcripts or 100 ng of AHSV-4 cores that had been pre-treated at 37 °C for 1 h with 100 μg/mL Proteinase K or 50 μg/mL RNase A.

Immunoblot analysis

Standard immunoblotting techniques[20] were used for detection of NS2 and VP7 protein expression at 12-h intervals over a time course of 72 h in BSR cells transfected with AHSV-4 core transcripts. Guinea pig polyclonal antiserum raised against NS2[8] or VP7[22] were used as primary antibodies and Protein A conjugated to horseradish peroxidase as the secondary antibody (Sigma-Aldrich, St. Louis, MO, USA).

Virus growth curves

Confluent BSR cell monolayers in 6-well tissue culture plates were infected with transfection-derived AHSV-3 or AHSV-4 at a MOI of 0.1 pfu/cell. At different time points post-infection, the virus titres were determined by serial dilution and plaque assays on BSR cells as described above.

Extraction of dsRNA

Total RNA was extracted from virus-infected cells with the Nucleospin RNA II kit (Macherey-Nagel, Düren, Germany) according to the manufacturer's protocol. The ssRNA was removed by precipitation with 2 M LiCl and centrifugation at 17 000 *g* for 30 min at 4 °C. The dsRNA was subsequently precipitated from the collected supernatant with 0.1 volume of NaOAc and 2 volumes of absolute ethanol. Following centrifugation as above, the pelleted dsRNA was washed twice with 70% ethanol and then suspended in DEPC-treated water. The purified dsRNA samples were analysed by agarose gel electrophoresis or used for cDNA synthesis.

Screening of transfection-derived viruses for reassortant and mutant AHSV

Purified dsRNA was used as a template to synthesise cDNA copies of the respective AHSV genome segments using the RevertAid H Minus First Strand cDNA Synthesis kit (Thermo Fisher Scientific, Waltham, MA, USA) according to the manufacturer's protocol. To identify AHSV reassortants, PCR amplification was subsequently performed using S10 genome segment-specific primer pairs for either AHSV-3 or AHSV-4. The primer pairs used for AHSV-3 and AHSV-4 were: A3S10-186F (5'-GATACTTAACCAAGCCATGTC-3') and A3S10-700R (5'-GTTTGATCCACCCAACACTG-3'), and A4S10-239F (5'-CTATGGCGGAAGCATTGC-3') and A4S10-742R (5'-GTATGTTGTTATCCCACTCC-3'), respectively. The reaction mixtures were analysed by agarose gel electrophoresis. To identify mutant AHSV-4, the PCR amplification was performed with primers NS3F (5'-GTTAAAATTATCCCTTGTCATGAATCTAGCTAC-3') and NS3R. The amplicons were subjected to *Eco*RI restriction endonuclease digestion, followed by analysis of the digestion products on an agarose gel. As a final confirmation, the nucleotide sequences of the amplified S10 genome segments from putative reassortant and mutant viruses were also determined.

Results

Purification of AHSV cores and in-vitro synthesis of core transcripts

AHSV-3 and AHSV-4 core particles were purified from virus-infected cells and the protein composition of the core particles was verified by SDS-PAGE. The core proteins VP1, VP3 and VP4 were readily visualised in the stained SDS-polyacrylamide gels, whereas the core proteins VP6 and VP7 co-migrated as a single band as a result of their similar molecular masses (38.4 kDa and 37.9 kDa, respectively) (Figure 1a). To furthermore confirm the absence of both of the outer capsid proteins VP2 and VP5, the AHSV core particles were examined by negative-staining transmission electron microscopy. Large quantities of core particles with a diameter of 65 nm were observed (Figure 1b). These particles displayed the characteristic capsomeres of orbivirus particles and were smaller in size to intact virion particles (80 nm in diameter).[5,23] To confirm that the purified AHSV-3 and AHSV-4 core particles were transcriptionally active, the cores were used for in-vitro synthesis of AHSV ssRNA. Agarose gel electrophoretic analysis of the purified core transcripts indicated that all 10 viral transcripts were synthesised, with no obvious evidence of premature termination or degradation of the transcripts (Figure 1c).

AHSV protein expression in transfected BSR cells

To determine whether the in-vitro synthesised AHSV core transcripts could serve as templates for viral protein synthesis in vivo, BSR cells were transfected with purified AHSV-4 core transcripts and harvested at different time intervals after infection. As markers through which the expression of viral proteins could be assessed over time, the non-structural protein NS2 and the structural protein VP7 were used. As shown in Figure 2, immunoblot analysis of cell lysates prepared from the transfected BSR cells confirmed that the respective proteins were synthesised from 60 h post-transfection onwards. These results thus indicate that transfection could effectively deliver in-vitro synthesised AHSV-4 core transcripts to the cytoplasm of BSR cells where they serve as templates for virus protein synthesis.

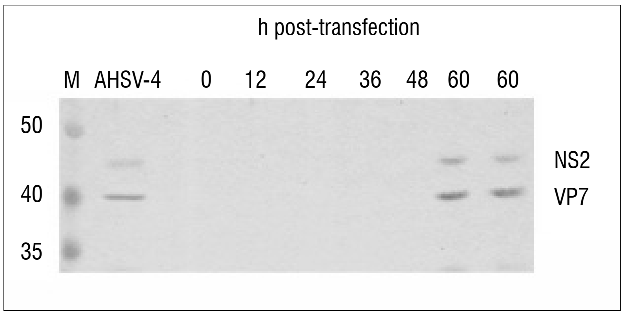


Figure 2: Viral protein expression in transfected BSR cells. Expression of the NS2 and VP7 proteins in BSR cells transfected with purified AHSV-4 core transcripts was determined at 12-h intervals by immunoblotting using appropriate antibodies. BSR cells infected with AHSV-4 served as a positive control. The sizes of reference protein markers (M, in kDa) are indicated on the left of the figure.

Recovery of infectious AHSV from transfected BSR cells

To determine whether infectious AHSV could be recovered from in-vitro synthesised core transcripts, plaque assays were performed of lysates prepared from BSR cells transfected with purified AHSV-3 or AHSV-4 core transcripts. Clear, well-defined plaques were observed, suggesting the presence of replicating virus (Figure 3a). To confirm virus recovery, viruses from individual plaques were amplified, after which dsRNA was extracted and analysed by agarose gel electrophoresis. The electrophoresis pattern of dsRNA segments extracted from transfection-derived AHSV-3 (Figure 3b) and AHSV-4 (Figure 3c) were identical to that of the corresponding parental AHSV serotype derived from cell infection. These results not only confirmed that AHSV-3 and AHSV-4 were recovered successfully from their corresponding core transcripts, but also that the core transcripts had been replicated to produce new genomic dsRNA. Moreover, the parental and transfection-derived AHSV-3 and AHSV-4 displayed similar growth curves in BSR cells, indicating that the recovered viruses had no gross replication or growth defects (Figure 3d).

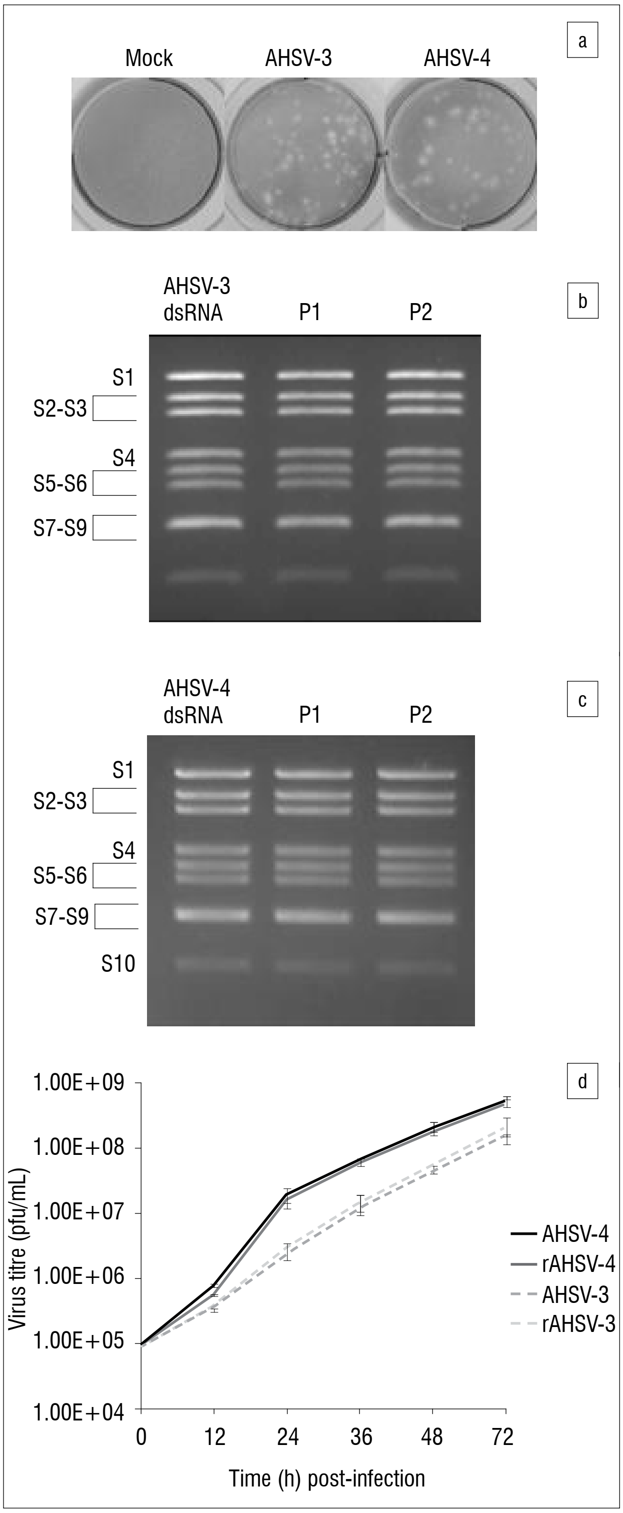


Figure 3: Recovery of African horse sickness virus (AHSV) following transfection of purified core transcripts into BSR cells. (a) Plaque assays performed on lysates of BSR cells transfected with purified AHSV-3 or AHSV-4 core transcripts. Mock-transfected BSR cells were included as a control. (b and c) Agarose gels showing the migration patterns of AHSV-3 and AHSV-4 dsRNA compared to the dsRNA extracted from transfection-derived viruses (indicated as P1 and P2 in each figure). (d) Growth curves of the transfection-derived AHSV-3 (rAHSV-3) and AHSV-4 (rAHSV-4) in BSR cells. For comparative purposes, the BSR cells were also infected with the parental AHSV-3 and AHSV-4 serotypes. The titre of the respective viruses at different times post-infection, as indicated in the figure, was determined by serial dilution and plaque assays on BSR cells. Each data point represents the mean±standard deviation from three independent experiments.

Confirmation that AHSV core transcripts are infectious

To characterise the infectious material present in the AHSV core transcript preparations, the sensitivity of the in-vitro synthesised AHSV-4 core transcripts to RNase A and Proteinase K was compared with that of AHSV-4 cores. Following transfection of BSR cells and plaque assays, the results indicated that the infectivity of the purified core transcripts was eliminated by the RNase A treatment, whereas the cores retained their infectivity. In contrast, Proteinase K treatment of core particles completely inhibited their ability to cause infection, whereas the infectivity of purified core transcripts was unaffected (Figure 4). These results confirmed that core transcripts are the sole source of infectivity in the purified core transcript preparations.

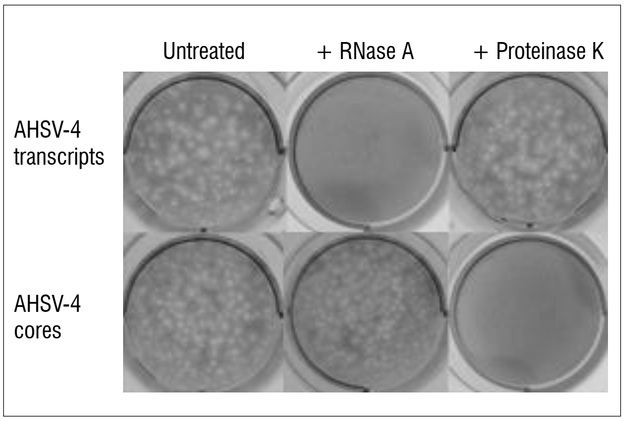


Figure 4: Confirmation that cores are not the source of infectivity in purified African horse sickness virus (AHSV) transcript preparations. Plaque assays were performed on lysates prepared from BSR cells that had been transfected with core transcripts and cores of AHSV-4, following treatment with RNase A or Proteinase K. Untreated samples were included as controls.

Generation of a targeted cross-serotype reassortant virus

After the infectivity of the purified AHSV core transcripts was established and the recovery of virus following their transfection into BSR cells was shown, we attempted to modify the AHSV genome by replacing the S10 genome segment of AHSV-3 with an in-vitro synthesised synthetic T7 transcript of the AHSV-4 S10 genome segment. For this purpose, purified AHSV-3 core transcripts were mixed with the AHSV-4 S10 genome segment T7 transcript (Figure 5a) and transfected into BSR cells. To screen for reassortant viruses, viruses from randomly selected plaques were amplified, dsRNA was extracted from the virus-infected cells and cDNA was synthesised. The origin of the S10 genome segment was then determined by PCR using primers specific for the S10 genome segment of AHSV-3 and AHSV-4. The amplicons were analysed by agarose gel electrophoresis and the identity of the amplicons obtained by PCR amplification (using the AHSV-4 S10 genome segment-specific primers) was confirmed by nucleotide sequencing. The results, presented in Figure 6, indicate that a cross-serotype reassortant AHSV could be recovered successfully using this approach.

Generation of a mutant AHSV-4

To determine if a specific mutation could be introduced into the AHSV genome using the above approach, we attempted to incorporate a unique *Eco*RI restriction enzyme site into the S10 genome segment of AHSV-4. To recover this mutant virus, purified AHSV-4 core transcripts and an AHSV-4 mutant S10 genome segment T7 transcript (Figure 5b) were transfected into BSR cells. Viruses resulting from the transfection were amplified and the dsRNA extracted from virus-infected BSR cells was used for cDNA synthesis. To screen for viruses containing the introduced mutation, the S10 genome segment was PCR amplified, digested with *Eco*RI and the reaction mixtures analysed by agarose gel electrophoresis.

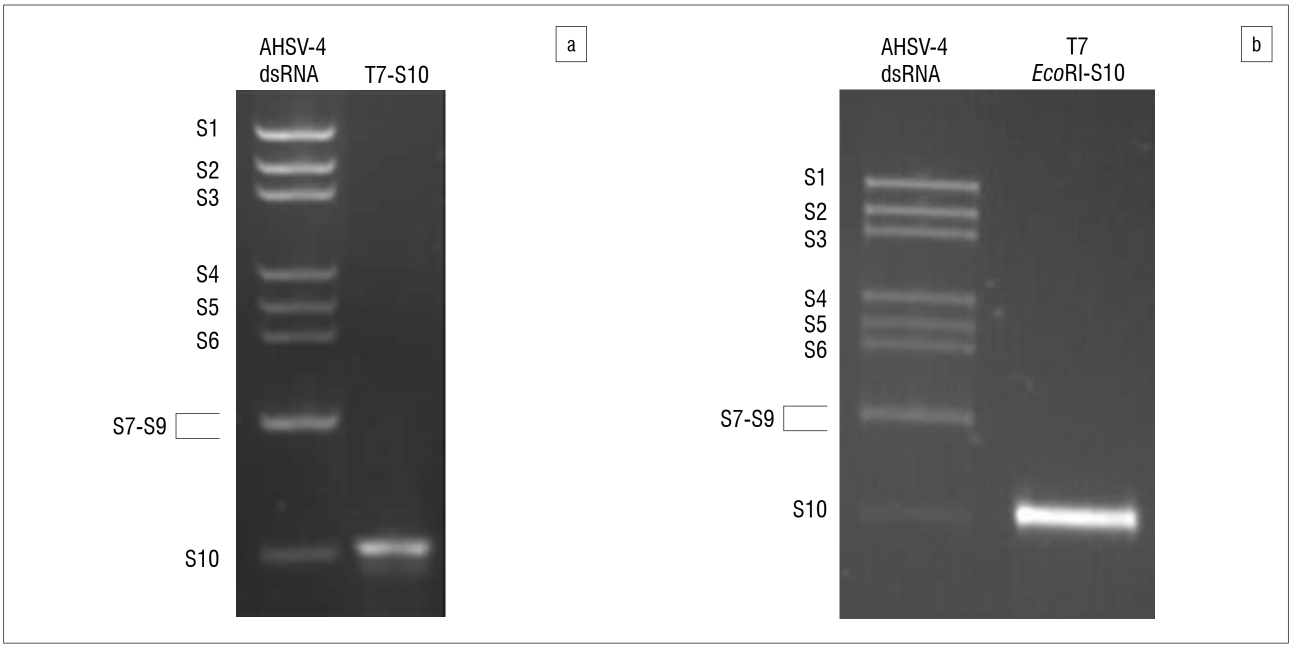


Figure 5: Agarose gels showing the in-vitro synthesised T7 transcripts of (a) African horse sickness virus (AHSV)-4 S10 genome segment (T7-S10) and (b) AHSV-4 mutant S10 genome segment (T7-*Eco*RI-S10) that were used in cell transfections to recover directed reassortant and mutant AHSV, respectively.

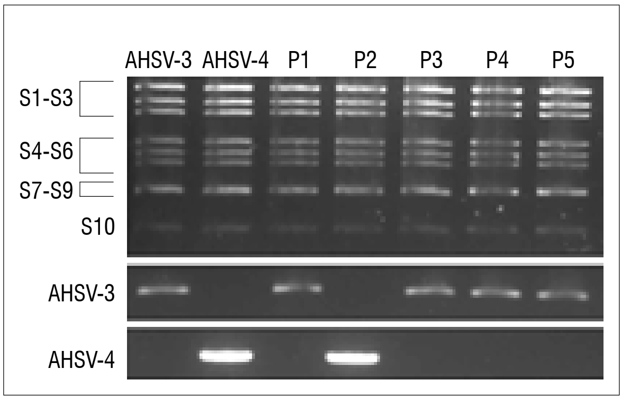


Figure 6: Recovery of cross-serotype reassortant African horse sickness virus (AHSV) after transfection of BSR cells with a mixture of AHSV-3 core transcripts and an AHSV-4 S10 genome segment T7 transcript. Agarose gels show viral dsRNA extracted from individual recovered viruses (indicated as P1 through P5 in top panel) and used for subsequent cDNA synthesis. The origin of the S10 genome segment in the recovered viruses was determined by PCR using primers specific for the S10 genome segment of AHSV-3 (middle panel) or AHSV-4 (bottom panel). BSR cells infected with AHSV-3 or AHSV-4 served as positive controls.

The results indicated that, in contrast to the non-mutated S10 genome segment which was not cleaved by the restriction enzyme, the S10 genome segments of mutant AHSV-4 viruses were cleaved and yielded two DNA fragments of the expected sizes (270 and 493 base pairs) (Figure 7). The nucleotide sequences of the purified amplicons were also determined, the results of which confirmed the presence of the unique *Eco*RI restriction enzyme site.

Discussion

The greatest obstacle preventing translation of basic research to AHSV biology has been the inability to manipulate the 10-segmented dsRNA viral genome through reverse genetics. Moreover, AHSV is a significant equine pathogen, and an unencumbered reverse genetics system would allow the development of novel recombinant vaccine strains. Thus, towards establishing a simple, broadly applicable genetic modification system for AHSV, we first examined whether transcripts synthesised in vitro from purified AHSV core particles have the same coding potential as those produced from transcribing core particles in virus-infected cells. As is the case for viral transcripts produced during natural infection, the in-vitro synthesised core transcripts were shown to be infectious when transfected into the cytoplasm of permissive BSR cells, resulting in the recovery of viable viruses that were identical to the parental AHSV serotypes in terms of growth characteristics in the mammalian cells. These findings were subsequently extended to demonstrate the genetic manipulation of the AHSV genome. Targeted reassortant and mutant viruses were generated successfully by making use of a combination of core-derived transcripts and a synthetic T7 viral transcript, which was derived from a PCR amplicon consisting of either a wild-type or a mutant cDNA copy of the S10 genome segment of AHSV-4 under the control of an upstream T7 promoter sequence. The approach used in this study is especially suited to investigations focusing on a single viral genome segment and may be applied to any one of the 10 AHSV genome segments. Although not investigated in this study, the approach may furthermore be suitable to generate reassortant AHSV in which more than one genome segment has been exchanged.

Although a reverse genetics system has been described recently for AHSV that makes it possible to generate strains with a single recombinant gene, this system relies on cloning cDNA copies of the entire viral genome and a modified version requires the construction of additional expression helper plasmids.[18] Compared to these systems, the approach used in this study is relatively easy to perform. Our approach obviates the need to construct a full set of 10 cDNA clones as only a single PCR product or cDNA clone needs to be generated, and the approach is sufficiently robust to enable recovery of targeted reassortant and mutant AHSV. Moreover, the time and labour involved in the isolation and purification of core particles can be reduced through ‘recycling’ of the core particles. In this study, we found that the same sample of core particles could be re-used up to four times in in-vitro transcription reactions without any significant loss in the yield and quality of synthesised transcripts. It should, however, be noted that screening is required to identify recombinant AHSV. Reassortant viruses were recovered at a 20–50% frequency, and mutant viruses at a frequency of between 70% and 90%. The differences in virus recovery may reflect variations in the degree of capping of the T7 transcript preparations.[24] In addition to being poorly translated, the 5' triphosphate moiety of uncapped transcripts is known to be a pathogen-associated molecular pattern recognised by RIG-I, which may lead to the induction of antiviral responses and thus also negatively influence the recovery of the virus.[25,26] The lower efficiency associated with the recovery of reassortant AHSV compared to mutant AHSV might be because of serotype-specific differences in genome segment sorting and packaging signals, but requires further investigation. Nevertheless, the efficiencies are sufficiently high so that mass screening of the transfection-derived viruses can be easily performed depending on the targeted gene and available tools, for example, by making use of discriminating monoclonal antibodies and/or discriminating PCR assays.

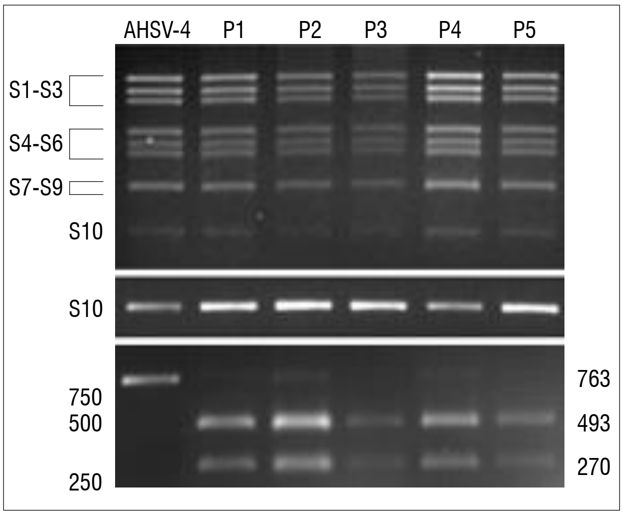


Figure 7: Recovery of mutant African horse sickness virus (AHSV)-4 after transfection of BSR cells with a mixture of AHSV-4 core transcripts and an AHSV-4 mutant S10 genome segment T7 transcript containing an *Eco*RI marker mutation. Agarose gels showing viral dsRNA extracted from individual recovered viruses (indicated as P1 through P5 in top panel), cDNA of the viral S10 genome segments (middle panel) and digestion of the cDNA products with *EcoRI* (bottom panel). BSR cells infected with AHSV-4 served as a control. The sizes of DNA markers (in base pairs) are indicated on the left of the figure, and the sizes of the digestion products are indicated on the right.

The ability to generate directed reassortant and mutant AHSV may have important implications for vaccine development. No effective treatment exists for African horse sickness and control of the disease relies on preventative vaccination. The current commercially available vaccine comprises a cocktail of live attenuated strains of only seven of the nine different AHSV serotypes, and the shortcomings of this vaccine have motivated efforts to develop alternative safe vaccines.[27,28] As opposed to random attenuation by serial passage of the virus, a reverse genetics approach may enable the rational design of attenuated vaccine strains via directed mutagenesis and/or by engineering of disabled infectious single-cycle (DISC) vaccine strains. DISC strains lack an essential gene product, and consequently are capable of replicating only once in infected cells, but they are nevertheless capable of eliciting both cellular and humoral immune responses.[29-31] Unlike the current commercial AHSV vaccine, DISC vaccines would make it possible to differentiate between animals that have been vaccinated and animals that have been infected with replicating virus (DIVA principle). In addition, the potential capability of replacing the antigenically important outer capsid proteins VP2 and VP5[32,33] from one AHSV serotype with those of another can be used to generate 'serotyped' vaccines. This capability could represent a means to rapidly produce vaccines against the different AHSV serotypes and create the option to custom engineer vaccines in order to protect against emerging epidemic strains.

In conclusion, transfection of cultured mammalian cells with a combination of viral core transcripts and a synthetic T7 transcript represents a feasible genetic modification system for AHSV. As such, it may provide a means by which safe designer live attenuated AHSV vaccine strains could be generated and also represents a potentially powerful tool for studies aimed at understanding AHSV biology, including aspects pertaining to replication, transmission, pathogenesis and immunity.

Acknowledgements

This research was funded by the University of Pretoria's Institutional Research Theme programme. The National Research Foundation (South Africa) provided E.M., D.J.P. and A.C. with bursaries. We are thankful to Andrea Crous for providing the recombinant pJET/blunt-S10 plasmid construct, Flip Wege for supplying viruses and cell cultures, and Prof. Henk Huismans and Dr Christiaan A. Potgieter for insightful discussions.

Authors' contributions

The initiation, conception and planning of the study involved E.V., D.J.P. and J.T. The experiments were performed by E.V., D.J.P. and A.C. All authors contributed to the data analysis and preparation of the manuscript.

References

1. Stassen L, Vermaak E, Theron J. African horse sickness, an equine disease of emerging global significance. In: Paz-Silva A, Sol Arias Vázquez M, Sánchez-Andrade Fernández R, editors. Horses: Breeding, health disorders and effects on performance and behaviour. New York: Nova Science Publishers; 2014. p. 145–170.

2. Coetzer JAW, Guthrie AJ. African horse sickness. In: Coetzer JAW, Tustin RC, editors. Infectious diseases of livestock. 2nd ed. Cape Town: Oxford University Press; 2004. p. 1231–1246.

3. MacLachlan NJ, Guthrie A. Re-emergence of bluetongue, African horse sickness, and other *Orbivirus* diseases. Vet Res. 2010;41:35–47. http://dx.doi.org/10.1051/vetres/2010007

4. Bremer CW, Huismans H, Van Dijk AA. Characterization and cloning of the African horsesickness virus genome. J Gen Virol. 1990;71:793–799. http://dx.doi.org/10.1099/0022-1317-71-4-793

5. Manole V, Laurinmäki P, Van Wyngaardt W, Potgieter CA, Wright IM, Venter GJ, et al. Structural insight into African horsesickness virus infection. J Virol. 2012;86:7858–7866. http://dx.doi.org/10.1128/JVI.00517-12

6. Roy P. Orbiviruses. *In:* Knipe DM, Howley PM. Fields virology. Philadelphia, PA: Lippincott Williams and Wilkins; 2001. p. 1679-–1728.

7. Mertens PP, Diprose J. The bluetongue virus core: A nano-scale transcription machine. Virus Res. 2004;101:29–43. http://dx.doi.org/10.1016/j.virusres.2003.12.004

8. Uitenweerde JM, Theron J, Stoltz MA, Huismans H. The multimeric nonstructural NS2 proteins of bluetongue virus, African horsesickness virus, and epizootic hemorrhagic disease virus differ in their single-stranded RNA-binding ability. Virology. 1995;209:624–632. http://dx.doi.org/10.1006/viro.1995.1294

9. Maree FF, Huismans H. Characterization of tubular structures composed of nonstructural protein NS1 of African horsesickness virus expressed in insect cells. J Gen Virol. 1997;78: 1077–1082.

10. Van Niekerk M, Smit CC, Fick WC, Van Staden V, Huismans H. Membrane association of African horsesickness virus nonstructural protein NS3 determines its cytotoxicity. Virology. 2001;279:499–508. http://dx.doi.org/10.1006/viro.2000.0709

11. De Waal PJ, Huismans H. Characterization of the nucleic acid binding activity of inner core protein VP6 of African horse sickness virus. Arch Virol. 2005;150:2037–2050. http://dx.doi.org/10.1007/s00705-005-0547-4

12. Stassen L, Huismans H, Theron J. Membrane permeabilization of the African horse sickness virus VP5 protein is mediated by two N-terminal amphipathic α-helices. Arch Virol. 2011;156:711–715. http://dx.doi.org/10.1007/s00705-010-0897-4

13. Bekker S, Huismans H, Van Staden V. Factors that affect the intracellular localization and trafficking of African horse sickness virus core protein, VP7. Virology. 2014;456–457:279–291. http://dx.doi.org/10.1016/j.virol.2014.03.030

14. Boyce M, Celma CCP, Roy P. Development of reverse genetics systems for bluetongue virus: Recovery of infectious virus from synthetic RNA transcripts. J Virol. 2008;82:8339–8348. http://dx.doi.org/10.1128/JVI.00808-08

15. Van Gennip RG, Van de Water SG, Potgieter CA, Wright IM, Veldman D, Van Rijn PA. Rescue of recent virulent and avirulent field strains of bluetongue virus by reverse genetics. PLoS ONE. 2012;7(2):e30540. http://dx.doi.org/10.1371/journal.pone.0030540

16. Trask SD, Boehme KW, Dermody TS, Patton JT. Comparative analysis of Reoviridae reverse genetics methods. Methods. 2013;59:199–206. http://dx.doi.org/10.1016/j.ymeth.2012.05.012

17. Komoto S, Taniguchi K. Genetic engineering of rotaviruses by reverse genetics. Microbiol Immunol. 2013;57:479–486. http://dx.doi.org/10.1111/1348-0421.12071

18. Kaname Y, Celma CC, Kanai Y, Roy P. Recovery of African horse sickness virus from synthetic RNA. J Gen Virol. 2013;94:2259–2265. http://dx.doi.org/10.1099/vir.0.055905-0

19. Mertens PP, Burroughs JN, Anderson J. Purification and properties of virus particles, infectious subviral particles, and cores of bluetongue virus serotypes 1 and 4. Virology. 1987;157:375–386. http://dx.doi.org/10.1016/0042-6822(87)90280-7

20. Sambrook J, Russell DW. Molecular cloning: A laboratory manual. New York: Cold Spring Harbor Laboratory Press; 2001.

21. Landt O, Grunert H-P, Hahn U. A general method for rapid site-directed mutagenesis using the polymerase chain reaction. Gene. 1990;96:125–128. http://dx.doi.org/10.1016/0378-1119(90)90351-Q

22. Rutkowska DA, Meyer QC, Maree F, Vosloo W, Fick W, Huismans H. The use of soluble African horse sickness viral protein 7 as an antigen delivery and presentation system. Virus Res. 2011;156:35–48. http://dx.doi.org/10.1016/j.virusres.2010.12.015

23. Breese SS, Ozawa Y, Dardiri AH. Electron microscopic characterization of African horse-sickness virus. J Am Vet Med Assoc.1969;155:391–400.

24. Jemielity J, Fowler T, Zuberek J, Stepinski J, Lewdorowicz M, Niedzwiecka A, et al. Novel 'anti-reverse' cap analogs with superior translational properties. RNA. 2003;9:1108–1122. http://dx.doi.org/10.1261/rna.5430403

25. Hornung V, Ellegast J, Kim S, Brzozka K, Jung A, Kato H, et al. 5'-Triphosphate RNA is the ligand for RIG-I. Science. 2006;314:994–997. http://dx.doi.org/10.1126/science.1132505

26. Cui S, Eisenacher K, Kirchhofer A, Brzozka K, Lammens A, Lammens K, et al. The C-terminal regulatory domain is the RNA 5'-triphosphate sensor of RIG-I. Mol Cell. 2008;29:169–179. http://dx.doi.org/10.1016/j.molcel.2007.10.032

27. Mellor PS, Hamblin C. African horse sickness. Vet Res. 2004;35:445–466. http://dx.doi.org/10.1051/vetres:2004021

28. MacLachlan NJ, Balasuriya UB, Davis NL, Collier M, Johnston RE, Ferraro GL, et al. Experiences with new generation vaccines against equine viral arteritis, West Nile disease and African horse sickness. Vaccine. 2007;25:5577–5582. http://dx.doi.org/10.1016/j.vaccine.2006.12.058

29. Zevenhoven-Dobbe JC, Greve S, Van Tol H, Spaan WJM, Snijder EJ. Rescue of disabled infectious single-cycle (DISC) equine arteritis virus by using complementing cell lines that express minor structural glycoproteins. J Gen Virol. 2004;85:3709–3714. http://dx.doi.org/10.1099/vir.0.80443-0

30. Dudek T, Knipe DM. Replication-defective viruses as vaccines and vaccine vectors. Virology. 2006;344:230–239. http://dx.doi.org/10.1016/j.virol.2005.09.020

31. Celma CC, Boyce M, Van Rijn PA, Eschbaumer M, Wernike K, Hoffmann B, et al. Rapid generation of replication-deficient monovalent and multivalent vaccines for bluetongue virus: Protection against virulent virus challenge in cattle and sheep. J Virol. 2013;87:9856–9864. http://dx.doi.org/10.1128/JVI.01514-13

32. Martínez-Torrecuadrada JL, Diaz-Laviada M, Roy P, Sanchez C, Vela C, Sánchez-Vizcaíno JM, et al. Full protection against African horsesickness (AHS) in horses induced by baculovirus-derived AHS virus serotype 4 VP2, VP5 and VP7. J Gen Virol. 1996;77:1211–1221. http://dx.doi.org/10.1099/0022-1317-77-6-1211

33. Alberca B, Bachanek-Bankowska K, Cabana M, Calvo-Pinilla E, Viaplana E, Frost L, et al. Vaccination of horses with a recombinant modified vaccinia Ankara virus (MVA) expressing African horse sickness (AHS) virus major capsid protein VP2 provides complete clinical protection against challenge. Vaccine. 2014;32:3670–3674. http://dx.doi.org/10.1016/j.vaccine.2014.04.036

Relationships between student throughput variables and properties

AUTHOR:
Lucas C.A. Stoop[1]

AFFILIATION:
[1]Division of Institutional Planning, University of Johannesburg, Johannesburg, South Africa

CORRESPONDENCE TO:
Lucas Stoop

EMAIL:
Lcastoop@gmail.com

POSTAL ADDRESS:
PO Box 8397, Edenglen 1613,
South Africa

DATES:
Received: 29 Sep. 2014
Revised: 04 Nov. 2014
Accepted: 16 Nov. 2014

KEYWORDS:
throughput efficiency; student success rates; analytical process model; graduates; dropouts

HOW TO CITE:
Stoop LCA. Relationships between student throughput variables and properties. S Afr J Sci. 2015;111(7/8), Art. #2014-0336, 5 pages. http://dx.doi.org/10.17159/sajs.2015/20140336

Many different models have been designed to describe the plethora of factors that influence student throughput and success and how these factors affect throughput system variables and properties. System variables include headcounts (H) and successful credits (S) of throughput systems; some examples of system properties are the percentage of the new student intake graduating annually, and the average number of years to graduate or to drop out of a degree. However, no past study has defined the analytical relationships between these variables and properties from a process perspective – which was the purpose of this study. Three simple analytical equations were derived for 4-year degrees, and then geometrically interpreted. The behaviour of a simplified throughput system can be described by the position of a point in the admissible region of the H-S plane, with each point relating to a specific set of system properties. The successful credits ratio (S/H) is shown to be the ideal process efficiency ratio for throughput systems. The results were also extended to degrees of shorter duration. The behaviour of real throughput systems is broadly found to be similar to the behaviour of simplified throughput systems. In this study, only the mathematical foundations for the general relationships between throughput properties and throughput variables for a degree were established. The way in which this mathematical basis finds application in practice is illustrated for a few selected cases only, because of the specific focus of this paper.

Introduction

Understanding the complexities of the throughput patterns of students enrolled for degree studies at universities has always been a major challenge. Numerous models are reported in the literature to describe the plethora of factors that influence student throughput and success and how these factors affect throughput system variables and properties. A review of the current literature in this regard can be found in recent studies.[1,2] From a process perspective, however, no study could be found that defines the analytical relationships between the most important variables and properties of a student throughput system; this absence is the reason for the current study. The throughput variables included in this study were the headcount number of students, the total number of successful module credits earned by students and the full-time equivalent (FTE) value for students enrolled for the degree. The annual intake of new students was also a special throughput variable considered here. Throughput properties considered were the percentage of new entrants graduating annually as well as the average time to graduate or to drop out of a degree. The word 'degree' in this paper is used in a generic sense and also refers to a diploma or certificate. In this paper, a simplified cohort survival model for a 4-year degree will be used to establish analytical formulae for the key system variables as functions of the system properties and the annual intake of new students. A geometrical interpretation of the analytical equations will also be given to show exactly how all the system variables and system properties are connected to one another simultaneously. The results developed for a 4-year degree will then be extended to degrees of shorter duration.This paper only establishes the mathematical basis for the general relationships between throughput properties and throughput variables for a degree. Although the application of the results to real throughput systems will be discussed, the way in which this mathematical basis finds application in practice will be illustrated for a few selected cases only. The theory as presented is not a model of South African higher education, but can be used to assess throughput process efficiency within the local higher education system as well as in other systems. The many interventions that can be undertaken by universities to improve the efficiency of the throughput process are equally important but are excluded from the scope of this paper.

Simplified cohort survival model calculations

The background calculations in this paper are structured in terms of a standard cohort survival model for a 4-year degree with a maximum time for completion assumed to be 8 years. The 8 years cut-off is necessary to ensure that the analytical expressions that will be derived are of a simple form. Treating the number of cut-off years as a variable increases the number of parameters in the model without necessarily contributing to a better understanding of throughput systems. In a simplified student throughput system, the cohort size remains constant, and the student graduation and dropout patterns for cohorts repeat year after year. These characteristics give rise to stationary throughput patterns that have also been referred to as equilibrium throughput patterns.[3] The standard cohort survival model used, together with the simplifying assumptions made in this paper, created the simplest model possible for deducing analytical expressions for the various throughput variables in terms of throughput properties. Cohort survival models in general produce numbers as outcomes with very little chance of discovering the analytical relationships that exist between the said entities. The details of the cohort calculations documented in this paragraph as new research results are important in order to understand the full impact of this work, but may be skipped by readers not necessarily interested in such detail. A summary is provided in the next paragraph.

In the simplified cohort survival model, the student throughput history for a 4-year degree will be reflected by the simultaneous presence of a set of eight cohorts with cohort 1 being the youngest cohort in the year under observation. Each cohort has the same throughput profile characterised by the percentage of the cohort graduating or dropping out from the system at the end of the year. Hence, if G_i is the percentage of the annual intake N of new students graduating and D_i the percentage of the annual intake of new students dropping out from cohort i at the end of the year, then the following equations will apply:

$(G_4 + \ldots\ldots + G_8)N = GN$ and $(D_1 + \ldots\ldots + D_8)N = (1\text{-}G)N,$ Equation 1

where GN is the number of graduates and G is the percentage of the annual intake of new students graduating at the end of the year. The very restrictive nature of Equation 1 already reflects the assumptions underlying the simplified student throughput model. More general expressions can be provided but will only increase the complexity of the analytical expressions to be derived without necessarily contributing to a better understanding of student throughput systems. Furthermore, the average number of years J taken by students to graduate at the end of the year is given by:

$J = (4G_4 + 5G_5 + 6G_6 + 7G_7 + 8G_8)/G,$ Equation 2

and the average number of years K studied by students dropping out at the end of the year is derived similarly as:

$K = (1D_1 + 2D_2 + \ldots 7D_7 + 8D_8)/(1\text{-}G).$ Equation 3

The headcount H_G of students in the throughput system who will eventually graduate can be derived through the cohort approach as:

$H_G = (4G_4 + 5G_5 + 6G_6 + 7G_7 + 8G_8)N = GJN.$ Equation 4

A similar expression for the headcount H_D of students who will eventually drop out of the throughput system can be derived through the cohort approach as:

$H_D = (1D_1 + 2D_2 + \ldots 7D_7 + 8D_8)N = (1\text{-}G)KN.$ Equation 5

The total headcount H for the throughput system is therefore defined in terms of the three independent system properties G, J and K as:

$H = [GJ + (1\text{-}G)K]N = [K + G(J\text{-}K)]N.$ Equation 6

This formula resembles a similar formula that was derived in a different way by Breneman for a production function for PhDs, as reported in Hopkins[4].

To calculate the total successful module credits earned annually by students who will eventually graduate (a maximum of 1 credit per student per year), the following is considered: a student who graduates in 4 years would earn 4/4 successful module credits per year; a student who graduates in 5 years would earn an average of 4/5 successful module credits per year; etc. The total number of successful module credits S_G earned annually by students who will eventually graduate is then given by:

$S_G = (4/4)(4N)G_4 + (4/5)(5N)G_5 + \ldots (4/8)(8N)G_8 = 4GN.$ Equation 7

In the case of students eventually dropping out of the throughput system, the successful module credits S_D in total earned annually by these students is:

$S_D = C(1\text{-}G)KN = 0.25(1\text{-}G)KN,$ Equation 8

where C is the average number of successful credits earned by each of these students per year. For simplicity, C is assumed to be equal to 0.25 to ensure that in the case of K=8, in which all students that eventually drop out stay on for 8 years, the students will each have accumulated only 2 successful credits on average. This average is presumably the result of some dropouts earning up to 4 successful credits and others earning very few credits. More realistic assumptions about the functional form of C have been tried, but these assumptions only increased the analytical complexity of Equation 8 without changing the broad conclusions reached in this paper. The total successful module credits (S) earned annually by all students is now defined in terms of the two independent system properties G and K, as:

$S = [4G + 0.25(1\text{-}G)K]N.$ Equation 9

The relationship between S and G is a linear relationship which shows a relatively direct response between S and G, especially for low values of K.

Throughput system variables and properties in summary

The number of graduates produced by the simplified throughput system at the end of the year under observation would be equal to GN – where N is the annual intake of new students and G is the percentage of the annual intake of new students graduating (ranging between 0% and 100%). The average number of years taken by students who have graduated at the end of the year under observation is denoted by J. Similarly, the average number of years studied by students dropping out at the end of the year under observation is denoted by K. The value of J ranges between 4 and 8, and the value of K ranges between 1 and 8. The three quantities G, J and K are independent of one another. They also describe the main characteristics of the student throughput process and are therefore referred to as throughput system properties. The characteristics of a throughput system can therefore be described in terms of all the possible combinations of G, J and K. System variables are used to describe bulk system quantities that only change if the properties of the throughput system change. These system variables are the headcount H of the throughput system, the successful module credits S annually earned by students, and the FTE value for students enrolled for the degree V. It has already been shown in Equations 6 and 9 that the system variables H and S depend on the system properties G, J and K and on the annual intake of new students N. Student headcount H, which in the South African higher education system refers to an unduplicated count of students irrespective of the academic course load of the student, depends on G, J and K as follows:

$H = [GJ + (1\text{-}G)K]N = [K + G(J\text{-}K)]N.$ Equation 6

The perfect throughput system in which all students graduate in minimum time, has a size equal to H=4N when G=100% and J=4. The influence of K diminishes as G approaches 100%. The total successful module credits S earned annually by students enrolled for a 4-year degree (a maximum credit of 1 per student per year) depends on G and K as follows:

$S = [4G + 0.25(1\text{-}G)K]N.$ Equation 9

The number of successful module credits S earned annually by students for the perfect throughput system is equal to S=4N when G=100%. Again, the influence of K diminishes as G approaches 100%. Apart from the successful credits (S) earned annually by students, the credits assigned to modules not successfully passed by students in the same year will be referred to here as failed credits (F). Furthermore, some students often take fewer modules than required by a full academic load with module credits therefore adding up to less than 1. In this paper, the balance of module credits not attempted by students in a particular year is referred to as unutilised credits (U). Using the fact that each of the H headcount students can at most generate 1 credit per year, it clearly follows that:

$S + F + U = H$ or $F + U = H - S.$ Equation 10

The FTE value for students enrolled for the degree V as a throughput variable, which in the South African higher education system depends on the successful credits as well as the failed credits for students enrolled for the degree, is defined by:

$$V = S + F = LS + (1 - L)H, \qquad \text{Equation 11}$$

where the fourth system property L=U/(U+F) defines the balance of credits between U and F. This fundamental relationship between V, S and H implies that V will range between S, when L=1 with no failed credits (F) present, and H, when L=0 with no unutilised credits (U) present. In a perfect throughput system, H=4N=S with the FTE value V also being equal to 4N. In the South African higher education system, the biggest part of the funding of universities, as well as the provision of building facilities, mainly depends on V. The FTE value is also a direct measure of the actual academic load on students in a particular year and is therefore also generally used by universities as a basis for the provision of lecturing staff.

Relationship between H and S in the throughput system

The system variables H and S defined above are not independent but are each dependent on the same set of G, J and K values for the student throughput configuration under consideration. This connection is mathematically defined by combining Equations 6 and 9 under the assumption that K be treated as a parameter. This definition establishes specific relationships between H (as well as S) and the two system properties J and G. These relationships can be made visible through graphs in the J-G plane, which unfortunately produces rather complex patterns of H and S. This complexity can be avoided by analysing these relationships in the H-S plane, as shown in Figure 1. Such an analysis reveals that only certain combinations of H and S can be realised, namely those included in the so-called admissible region of the H-S plane. The admissible region is a triangle, WXZ, bounded by the line WX representing J=4, by the line WZ representing J=8, as well as by the horizontal lines of G=0% and G=100%. A second vertical axis on the right has been added to show the corresponding values of G. The parameter K has been set equal to 4 by way of example with the latter value corresponding to a specific system property. The value of K also defines the position of the pivot W of lines of constant J and their intersection with the line G=0%. In the admissible region WXZ, pairs of admissible H and S values, such as the throughput configuration Y with H=5N and S=3N, would always be connected to lines of constant J values (in this case J=5.5) and lines of constant G values (in this case G=67%). The perfect throughput system is located at X now defined by H=4N=S and is produced by the intersection of the lines G=100% and J=4. Only within the admissible region defined by a given value of the parameter K, will each combination of G and J correspond to a unique combination of H and S, and vice versa.

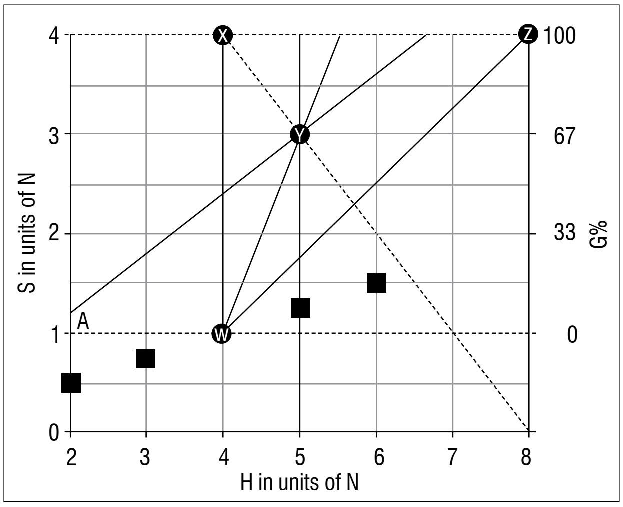


Figure 1: Admissible region for H (headcounts) and S (successful credits) in the H-S plane for K=4 and L=0.5, where K is the average number of years studied by students dropping out at the end of the year and L is the balance of failed and unutilised credits; N is the annual intake of new students and G is the percentage of the annual intake of new students graduating.

Of particular importance in this analysis is the successful credits ratio for the simplified throughput system defined in this paper by the ratio S/H. This ratio which can be regarded as the ideal process efficiency ratio compares the output of the successful credits S produced during the year under consideration to the input of the total credits H available for that year; both S and H depend on the same set of system properties G, J and K. The successful credits ratio is considered to be ideal in the sense that it will be shown to apply to simplified as well as real throughput systems. The successful credits ratio as a single number produces a simultaneous account of the combined efficiency status of the throughput system characterised by a specific set of system properties G, J and K. The successful credits ratio for the throughput system Y is given by S/H=60%. As the efficiency of the throughput system increases, the lines of constant S/H will move towards X. The successful credits ratio S/H will be equal to 100% for the perfect throughput system when G=100% and J=4. This scenario is true for all degrees irrespective of the duration of the degree.

Furthermore, lines of constant FTE values are defined by V/N=constant. If L=0, the lines V/N=constant are vertical lines, and if L=1, these lines are horizontal; all of these lines are also parallel to one another. The line V=4N (with L=0.5) is also shown in Figure 1 as a dotted line passing through X. The fact that Y lies on the line V=4N implies that the V value of the throughput system Y is 4N. Above the line V=4N, the FTE values V are larger than 4N, and below the line, the values are smaller than 4N. With much of the attention presumably focused on the migration of the system Y along the line V=4N, it would appear that a value of K equal to 4 would conveniently be required to restrict migrations of Y diverging too far to the left. This situation implies that K as a system property would have to be managed in such a way as to remain at a value equal to the minimum time of completion of the degree. The geometrical interpretation of Equations 6, 9 and 11 is also shown in Figure 1, again illustrating the interdependence of throughput variables and properties. As the throughput system Y migrates within the triangular admissible domain, Y carries along with it the lines of constant H/N, S/N, V/N, G, J and L, simultaneously showing the relationships between these quantities. The value of K defining the pivot W of the triangle, can be read from the H axis with various possible positions of the pivot as shown by the square markers. The value of J can also be read from the H axis at the top, and the value of G can be read on the right-hand axis. Constant successful credits ratio lines S/H=constant all pass through the fictitious origin of the H-S plane. In summary, the behaviour of a simplified throughput system can be described by the position of a point in the admissible region of the H-S plane, each point relating to a specific set of system properties. The admissible region is bounded by lines of maximum and minimum values of J and G with K preferably managed to be equal to the minimum time for completion of the degree. The migration of the throughput system in the H-S plane means changing the system properties and allowing enough time for the system to establish equilibrium in its throughput patterns. The successful credits ratio for the system is equal to S/H and its FTE value is equal to (S+H)/2 for L=0.5. For a given value of N, the perfect throughput system is found at H=4N=S with S/H=100%.

Throughput systems for 1- to 4-year degrees

Here the simplified model for 4-year degrees is extended to degrees of shorter duration. In the case of the 4-year degree, it was assumed that the minimum time of 4 years to be taken by students to graduate should be limited to a maximum of 8 years. It is suggested that the same rule be applied for degrees of shorter duration, the reason being the simplification of the analytical expressions to be derived. In the case of degrees of shorter duration, the relevant simplified cohort survival model can be used to derive simple expressions for H and S, similar to Equations 6 and 9. The assumption of the maximum time for completion of the degree being double the minimum time for completion, leads to strikingly similar analytical expressions, each being a function of the minimum time for completion of the degree. Therefore, with M the minimum time for completion of a degree in general, J will then range between M and 2M, whereas K would range between 1 and 2M. The headcount H of the student throughput system is given by H=[K+G(J-K)]N, whereas the successful module credits S earned annually by students would then be given by S=[MG+0.25(1-G)K]N.

Application of the findings to a few selected topics

The main purpose of this paper is to establish the mathematical basis for the general relationships between throughput properties and throughput variables for a degree. However, a few simple applications of the equations derived in this paper will now be given for degrees offered in the South African higher education system. More complex applications to issues such as the conversion of 3-year degrees into 4-year degrees for students who are expected to find it difficult to complete the 3-year degree in minimum time,[5] will be discussed in a follow-up paper.

Clarity on process efficiency measurements

In the South African higher education system, the success rate S/(S+F) that can be written as S/(S+F)=(S/H)/[1-L(1-S/H)] is used as a measure of efficiency instead of the successful credits ratio (S/H) as defined in this paper. The success rate will clearly measure the process efficiency of degrees differently because of its dependence on L. Degrees with L values close to 1 and a success rate therefore approximately equal to 1, will seem to be more efficient than degrees with L values close to 0 and a success rate therefore equal to approximately S/H. This difference in efficiency measurement becomes very pronounced in the case of degrees with relatively low successful credits ratios, such as S/H=50%. The use of the success rate should therefore be discontinued in favour of the use of the successful credits ratio, S/H.

Another ratio that is often used in the South African higher education system as a measure for process efficiency is the FTE to headcount ratio (V/H). With V equal to LS+(1-L)H, it is clear that this ratio would likewise produce different measurements of process efficiency as a result of its dependence on L; its use should therefore also be discontinued.

The graduate output to size ratio (GN/H) is known in the South African higher education system as the graduation rate, and its application to real throughput systems is widely contested because of fluctuations in both the values of G and the intake N of new students.[6] Despite this disadvantage in the case of real systems, it may still be useful in simplified systems as an absolute output to size ratio, focusing on the process efficiency of the teaching learning process irrespective of the duration of the degree. However, a 4-year degree with GN/H being equal to 0.25 at best, would then from a process perspective appear to be less efficient than a 3-year degree with its ratio GN/H being equal to 0.33 at best. This outcome would certainly restrict the use of this ratio as a process efficiency measurement for degrees in general. The general perception that GN/H is an unrelated or even a more comprehensive measurement of process efficiency is also incorrect. The fact is that the two measurements (S/H) and (GN/H) are related to one another through Equation 9 with (GN/H)=(S/H)/4=0.25 for the perfect throughput system. It is noted that lines of GN/H=constant all pass through the fictitious origin of the H-G plane; in particular, the line GN/H=0.133 passes through Y in Figure 1.

Inequities in the funding framework for South African universities

The biggest part of the funding framework for South African universities is based on the sum total of the FTE values V=LS+(1-L)H for each university degree. Degrees with L values close to 1 will have FTE values very close to S and therefore a funding base that will be largely performance driven through the output variable S. However, degrees with L values close to 0 will have FTE values very close to H and therefore a funding base that will be largely input driven through the input variable H. This difference is considerable, especially in the case of degrees in which the successful credits ratio is S/H=50%. In such a case, the L=0 degree will receive double the funding received by the L=1 degree. The question is whether two degrees with exactly the same successful credits ratio S/H should be funded at such disparate levels especially if no specific reason can be identified to justify the existence of different values of L? It therefore seems reasonable that L should be assumed to be 0.5 for all degrees and that the quantity LS+(1-L)H=(S+H)/2 be used as a more appropriate basis for the funding of degrees offered by South African universities. This approach would result in all degrees being funded on the basis of their average of S and H, which, however, does not affect the actual L value for the degree, although one may eventually find a tendency amongst faculties to manage the L values of their degrees towards L=0.5. Consideration could perhaps also be given to change the definition of the FTE value to be based on the average of S and H. It is noted that, whereas H can be regarded as the nominal size of the degree relating to an unduplicated count of student names on a list, the quantity S can be regarded as the credit earning size of the degree. The redefined FTE values (S+H)/2 could then be regarded as the effective size of the degree. Again, determining the full implications for the funding of South African higher education should, because of the complexity of the topic, require much more research and should rather be pursued outside of the scope of this introductory paper.

Enrolment management

The size of the South African higher education system is currently regulated by government-approved enrolment plans for universities with quotas for both H and V set for each university for each year. A minor relaxation of these constraints has recently been proposed,[7] which will not address the difficulties highlighted below. The enrolment plans also call for meaningful annual improvement of the success rates and graduation rates for each university, thus signalling that higher successful credits ratios (S/H) need to be achieved for each university. These enrolment plans have three unintended consequences. Firstly, by imposing the FTE values constraint on throughput systems, universities would have to manage the H part of this constraint at the beginning of each academic year according to Equation 11 but would only know the value of the S part at the end of the year as a consequence of successful teaching outcomes. Such a constraint is very difficult to manage in practice. Secondly, when written in the form V/H=[1-L(1-S/H)], Equation 11 states that a student throughput system in which both V and H are constant, or even directly proportional to one another, can only produce the same but not higher successful credits ratios (S/H) as required by the enrolment plans. This dilemma can apparently be resolved by only retaining the constraint on V, which is the more important constraint relating to the funding of the system. However, this again leaves the university with a constraint which is very difficult to manage in practice. Thirdly, imposing headcount quotas on the throughput system has led to enrolment management practices at South African universities which amount to registering returning students first and then using the new student intake (N) to make up for the shortfall. This unfortunate way of setting the size of the new student intake (N) introduces awkward fluctuations into the student throughput patterns during subsequent years, which in turn undermines proper planning with regard to the provision of facilities and lecturing staff. These three unintended consequences of enrolment planning within the South African higher education system, also point to the need for further research to be undertaken to resolve these difficulties.

Real student throughput systems

A better understanding of real throughput systems should follow directly from a study of simplified systems. Simplified systems are not theoretical constructs but in fact special cases of real throughput systems. The behaviour of a simplified throughput system can be described by the position of a point in the admissible region of the H-S plane, with each point relating to a specific set of system properties. The successful credits ratio of the throughput system is equal to S/H with the redefined FTE values equal to (S+H)/2. The same behaviour, however, is observed from the data for real throughput systems. This is explained by the fact that H can be calculated independently of Equation 6 by simply adding together the actual unduplicated number of students enrolled for a degree. The same calculation can be performed for S independent of Equation 9. Hence, for all real throughput systems the successful credits ratio would be defined by S/H with the redefined FTE value equal to (S+H)/2. Only in the case of a constant annual intake of new students N, would it be possible to define the position of the perfect throughput system H=4N=S. It is also noted that these calculations may now be performed at the university level because both H and S can be added together for a group of degrees.

Conclusions

In this paper, the mathematical foundation for the general relationships between throughput properties and throughput variables for a degree has been established using a simplified or equilibrium cohort survival approach. The simplified model assumes a constant annual intake of new students and that throughput system properties such as the graduation and dropout patterns for each cohort also remain the same. Throughput properties include the percentage G of the annual intake of new students graduating annually as well as the average number of years J to graduate and the average number of years K to drop out of a degree. The balance L between the unutilised and failed module credits within the throughput system is the fourth system property required. Three analytical formulae have been derived for important system variables, such as the headcounts H and total successful module credits S of the throughput system, both of which depend on the system properties G, J and K as well as on the annual intake of new students N. The FTE value V has been expressed in terms of S, H and L.

In this paper, it has been demonstrated that the system variables H and S are not independent but are each dependent on the same set of G, J and K values for the simplified student throughput configuration under consideration. Furthermore, only certain H and S values can simultaneously be realised – namely those in the admissible region of the H-S plane. The admissible region has a triangular shape bounded by edges corresponding to the maximum and minimum values of J and G. The shape of the triangle is determined by the value of K, which should preferably be managed to be equal to the minimum time for completion of the degree. Furthermore, within this triangle, the relationships between throughput variables and throughput properties become visible through suitable geometrical constructions. In essence, the behaviour of a simplified throughput system can be described by the position of a point in the admissible region of the H-S plane, with each point relating to a specific set of system properties. The migration of the throughput system in the H-S plane means changing the system properties and allowing enough time for the system to establish equilibrium in its throughput patterns. The successful credits ratio S/H produces a simultaneous account of the combined efficiency status of the three system properties G, J and K. More importantly, the FTE number of students is given by (S+H)/2 for L=0.5, and for a given value of N, the perfect throughput system for a 4-year degree is located at H=4N=S with S/N=100%. This paper provides indications on how the simplified model for a 4-year student throughput system can be changed to apply to degrees of different duration.

References

1. Brinkman PT, McIntyre C. Methods and techniques of enrolment forecasting. In: Layzell DT, editor. New directions for institutional research, no. 93. San Francisco, CA: Jossey-Bass; 1997. p. 67–80.

2. Prinsloo P. Modelling throughput at Unisa: The key to successful implementation of ODL. Pretoria: University of South Africa; 2009. Available from: http://hdl.handle.net/10500/6035

3. Oliver RM, Hopkins DSP, Armacost RL. An equilibrium flow model of a university campus. Oper Res. 1972;2:249–264. http://dx.doi.org/10.1287/opre.20.2.249

4. Hopkins DSP. The higher education production function: Theoretical foundations and empirical findings. In: Hoenack SA, Collins EL, editors. The economics of American universities. Albany, NY: State University of New York Press; 1990. p. 21.

5. Council on Higher Education. A proposal for undergraduate curriculum reform in South Africa: The case for a flexible curriculum structure. Pretoria: Council on Higher Education; 2013. Available from: http://www.che.ac.za/sites/default/files/publications/Full_Report.pdf

6. Watson P. Measuring postgraduate cohort throughput: A case study. S Afr J High Educ. 2008;22:725–739.

7. Department of Higher Education and Training. Report of the Ministerial Committee for the review of the funding of universities. Pretoria: Department of Higher Education and Training; 2013. Available from: http://www.dhet.gov.za/SiteAssets/Latest%20News/Report%20of%20the%20Ministerial%20Committee%20for%20the%20Review%20of%20the%20Funding%20of%20Universities.pdf

Allele frequencies of *AVPR1A* and *MAOA* in the Afrikaner population

AUTHORS:
J. Christoff Erasmus[1]
Anton Klingenberg[1]
Jaco M. Greeff[1]

AFFILIATION
[1]Department of Genetics, University of Pretoria, Pretoria, South Africa

CORRESPONDENCE TO:
Jaco Greeff

EMAIL:
jaco.greeff@up.ac.za

POSTAL ADDRESS:
Department of Genetics, University of Pretoria, Private Bag X20, Hatfield 0028, South Africa

DATES:
Received: 20 Feb. 2015
Revised: 23 Apr. 2015
Accepted: 01 June 2015

KEYWORDS:
admixture; founder effect; monoamine oxidase; arginine vasopressin receptor 1a; non-paternity

HOW TO CITE:
Erasmus JC, Klingenberg A, Greeff JM. Allele frequencies of *AVPR1A* and *MAOA* in the Afrikaner population. S Afr J Sci. 2015;111(7/8), Art. #2015-0074, 6 pages. http://dx.doi.org/10.17159/sajs.2015/20150074

The Afrikaner population was founded mainly by European immigrants that arrived in South Africa from 1652. However, female slaves from Asia and Africa and local KhoeSan women may have contributed as much as 7% to this population's genes. We quantified variation at two tandem repeats to see if this historical founder effect and/or admixture could be detected. The two loci were chosen because they are in the promoters of genes of neurotransmitters that are known to be correlated with social behaviour. Specifically, arginine vasopressin receptor 1A's (*AVPR1A*) RS3 locus has been shown to correlate with age of sexual onset and happiness in monogamous relationships while the tandem repeat in the promoter of the monoamine oxidase A (*MAOA*) gene correlates with reactive aggression. The Afrikaner population contained more *AVPR1A* RS3 alleles than other Caucasoid populations, potentially reflecting a history of admixture. Even though Afrikaners have one of the lowest recorded non-paternity rates in the world, the population did not differ at *AVPR1A* RS3 locus form other European populations, suggesting a non-genetic explanation, presumably religion, for the low non-paternity rate. By comparing population allele-frequency spectra it was found that different studies have confused *AVPR1A* RS3 alleles and we make some suggestions to rectify these mistakes in future studies. While *MAOA* allele frequencies differed between racial groups, the Afrikaner population showed no evidence of admixture. In fact, Afrikaners had more 4-repeat alleles than other populations of European origin, not fewer. The 4-repeat allele may have been selected for during colonisation.

Introduction

The Afrikaner population of South Africa derives from about the same proportion of German, French and Dutch immigrants that came to the Cape from 1652 to 1806.[1] Because the number of immigrants was finite, the Afrikaner population is considered a textbook example of a founder effect[2], with many genetic diseases in overabundance compared with European populations[3-5]. However, as many as 5000 European men settled at the Cape between 1657 and 1866.[6] As most of the immigrants were men, they occasionally married non-European women[6,7] who were either slaves from Africa and India (including Indonesia and East Asia) or local Khoe and San (KhoeSan) women[6-9]. This practice is reflected genetically by the presence of non-European alleles in the Afrikaner population.[10]

The Afrikaner population fought several local wars against the KhoeSan, Xhosa, Zulus and British. It could be argued that these aggressive encounters may have been frequent and severe enough to have left traces of selection on the population. One gene that may have played an important role in this regard is monoamine oxidase A (*MAOA*) which breaks down serotonin and dopamine and which has been linked to increased reactive aggression.[11-14]

Reactive aggression in humans may be affected by the alleles they carry at the variable number of tandem repeats (VNTR) in the promoter region of the *MAOA* gene, which is located on the X chromosome at Xp11.23.[11-15] These VNTRs occur in 2, 3, 3.5, 4 or 5 repeats of 30 base pairs (bp).[13,16] These repeats can be classified as either high or low activity alleles, with the 2, 3 and 5 repeats constituting the low activity (*MAOA*-L) alleles and the 3.5 and 4 repeats the high activity (*MAOA*-H) alleles.[12] The *MAOA*-H allele has a 2–10 times more effective transcription than the *MAOA*-L alleles.[13,16,17] Others[11,12,18] have determined that carriers of *MAOA*-L react more aggressively in provocational circumstances than their *MAOA*-H counterparts. The *MAOA*-H carriers better tolerated maltreatment and were also less likely to develop antisocial traits.[18]

Recent studies revealed a very low non-paternity rate of less than 1% in the Afrikaner population[19,20] (see Greeff and Erasmus[21] for an exception to the rule). Strong religious convictions, as was the case for Afrikaners[22], have been suggested as an important determinant of marital fidelity[23]. However, the low rate of non-paternity in Afrikaners may have a genetic component: two studies point to the potential importance of arginine vasopressin receptor 1a (*AVPR1A*) in this context. Prichard et al.[24] have shown that age of first sexual encounter is correlated to repeat length of alleles of *AVPR1A*. Similarly, Walum et al.[25] have shown that certain alleles of *AVPR1A* seem to predispose their carriers to a less fulfilling monogamous life. Given Afrikaners' low non-paternity rates, it is of interest to quantify *AVPR1A* for this population.

AVPR1A is located at 12q14-15 and there are three polymorphic repeat regions in its 5' flanking region.[26] One of these, a complex$(CT)_4$-TT-$(CT)_8$-$(GT)_{24}$ repeat known as RS3, is 3625 bp upstream from transcription initiation[26] and has been linked to human social behaviour in a number of studies. Given the early lead from voles in which longer microsatellite length results in higher levels of transcription[27], functional magnetic resonance imaging showed that carriers of longer repeats had significantly stronger activation of their amygdala upon an emotional test[28]. Similarly, longer *AVPR1A* RS3 alleles were found to be significantly more transcribed in post-mortem examination of the hippocampal area of humans.[29] From the behavioural side, male individuals with two long alleles are significantly more likely to have sexual intercourse before the age of 15 than male individuals with a short/long genotype.[24] Individuals with longer alleles are also more likely to be altruistic in the dictator game.[29] Other studies have linked specific alleles with altruism in pre-schoolers[30], happiness in monogamous relationships[25], social behaviour and autism[31-34], musicality[35-37], creative dance[38] and eating attitudes[39].

Both of these genes, *AVPR1A* and *MAOA*, could have unusual frequencies in the Afrikaner population because of the founder effect and/or admixture and may have affected the population's average behaviour. The aim of this study was to characterise the frequencies of *AVPR1A*'s RS3 microsatellite and the *MAOA* VNTR alleles in the Afrikaner population and to compare them to other populations. However, comparing allele frequencies across populations highlighted a problem with standardised allele calls at the RS3 locus.

Materials and methods

Sample collection

Ethical clearance was obtained from the Ethics Committee of the Faculty of Natural and Agricultural Sciences, University of Pretoria (no. EC11912-065). A total of 200 male volunteers from the Afrikaner population were confirmed not to be fourth-degree relatives through self-supplied ancestries and the majority were clustered into 23 groups that are very distantly related by paternal ancestry (at least 14 degrees). Note that this is no more than can be expected of random individuals.[9] These men considered themselves Afrikaners and have typical Afrikaner surnames. All volunteers completed an informed consent form and signed an agreement which stated that they understood that their DNA was going to be used for analysis and that they donated it willingly. The study adhered to the principles of the Declaration of Helsinki. Saliva samples were collected from participants using the Oragene-DNA self-collection kit supplied by DNA Genotek (Kanata, Ontario, Canada) and genomic DNA was isolated according to the manufacturer's instructions.

Genotyping

For both loci, the polymerase chain reaction (PCR) set-up consisted of 50 ng DNA, 1X AmpliTaq® 360 Buffer (Applied Biosystems, Foster City, CA, USA), 20 μM dNTPs, 250 μM $MgCl_2$, 0.4 μM forward and reverse primers, 2% AmpliTaq® 360 GC Enhancer and 1.25 units AmpliTaq® 360, in a final reaction volume of 10 μL, and the PCR reactions were run in the 2720 Thermal Cycler (Applied Biosystems, Foster City, CA, USA). The basic PCR cycle was repeated 35 times and included an initial denaturation step of 5 min at 95 °C and a final elongation step of 5 min at 72 °C.

MAOA

The genomic DNA of the promoter region in the *MAOA* gene was PCR amplified for subjects with the primer MAOaPT1 5'-ACAGCCTGACCGTGGAGAAG-3' and MAOaPB1 5'-GAACGGACGCTCCATTCGGA-3'.[16] The cycling conditions were as follows: 95 °C denaturation step for 1 min followed by primer annealing at 62 °C for 1 min and elongation at 72 °C for 1 min. The PCR products were separated on 3% agarose gels, with a 20-bp ladder (Promega, Madison, WI, USA). Six bands were excised from the gel and purified with the High Pure PCR purification kit (Roche Diagnostics, Germany) and were successfully sequenced with BigDye (Applied Biosystems, Foster City, CA, USA). Cycle sequencing products were purified with ethanol precipitation and ran on a 3130xl genetic analyser (Applied Biosystems, Foster City, CA, USA). These sequences were aligned to the reference sequence from the National Centre for Biotechnology Information (GenBank: M89636.1) with BioEdit version 7.2.2[40] and three unique sequences were deposited in the European Nucleotide Archive (accession numbers LN813020 – LN813022). Based on size differences, cases with 3, 4 and 5 repeats were selected to serve as size markers for the identification of the remaining samples on the gel.

AVPR1A

The RS3 microsatellite of *AVPR1A* was amplified with a labelled forward primer 5'-6-FAM-TCCTGTAGAGATGTAAGTGC-3' and the reverse 5'-GTTTCTTTCTGGAAGAGACTTAGATGG-3'.[32] Cycling conditions were as follows: denaturation at 94 °C for 1 min, primer annealing at 54.7 °C for 45 s and elongation at 72 °C for 1 min. Amplicons were run on an ABI 3500 XL genetic analyser (Applied Biosystems, Foster City, USA). The final allele lengths were scored with GeneMapper software version 4.1.1 (Applied Biosystems, Foster City, USA). Individuals homozygous for alleles 332, 334, 336, 338, 340, 342 and 346 were sequenced with the PCR primers to determine the actual number of CT and GT repeats and the sequences were deposited in the European Nucleotide Archive (accession numbers LN812321 to LN812340).

Comparable allele frequencies

Like other microsatellites, *AVPR1A* RS3 allele calling can easily vary among studies because of different Taq, dye, polymers, size standards and machines, which prevents comparisons among studies. To complicate matters further, different studies have used primers that result in different sized amplicons. Working from the sequence published by Thibonnier et al.[26] (AF208541), primer sets can result in an amplicon of 260 bp[41], 324 bp[25,28,32,37] (and the present study), 316 bp[29-31,33,38,39] or 317 bp[34], while others do not report primers used[36]. Fortunately, the allele spectrum for Caucasoids has a very characteristic profile that can be used to slide the alleles along so that they align well (compare Figure 1a and 1b). It can be seen that typically there are two alleles that are much more frequent than the others, and that there is another frequent allele that is 5 repeat units larger than the biggest common allele (Figure 1). We used this allele profile to make data sets comparable.

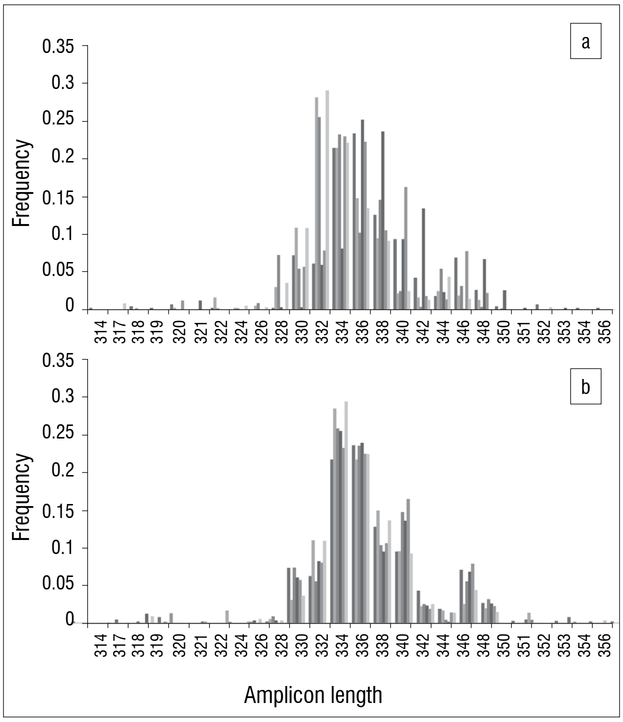


Figure 1: The population frequencies for *AVPR1A* alleles when (a) expected amplicon differences caused by primer differences are taken into account and (b) allele sizes are adjusted so that allele-frequency spectra of populations correspond best with each other. The data are drawn from Table 2.

Statistical analysis

For both loci, pairwise F_{ST}s were calculated between the populations using Arlequin version 3.5.1.2.[42] For *AVPR1A* we also compared the Afrikaner's frequency of allele 334 to that of the other populations by first testing if the allele's frequency varied significantly over all five population samples (prop.test as implemented in R[43]). Then the two European populations[32,37] and the two Israeli samples[30,33] were combined and compared to each other and the Afrikaner population in a pairwise proportion test as implemented in R (pairwise.prop.test).[43]

Results

MAOA

Sequencing confirmed that we were amplifying the correct DNA and we used these confirmed allele sizes as standards for electrophoresis. The frequencies of the observed alleles in the Afrikaner population and other populations are summarised in Table 1. The 3 and 4 repeat alleles were most frequent and varied across populations (Table 1; Figure 2). Pairwise F_{ST} values split the populations into two groups (Supplementary table 1 online). One group consisted of those of European descent for whom the 4-repeat allele was more common (Figure 2, clear symbols), and the other group consisted of those of African and Asian descent for whom the 3-repeat allele was more frequent (Figure 2, filled circles). Within each group, F_{ST} values were generally smaller than 0.01, and F_{ST} values between groups were mostly greater than 0.1 and significantly different (all $p<0.001$). A sample from Italy was significantly different from the Afrikaner population but not significantly different from the African American sample. Interestingly, two of the admixed populations – Hispanics and Afrikaners – did not fall in between European and non-European populations but had a higher frequency of the 4-repeat allele.

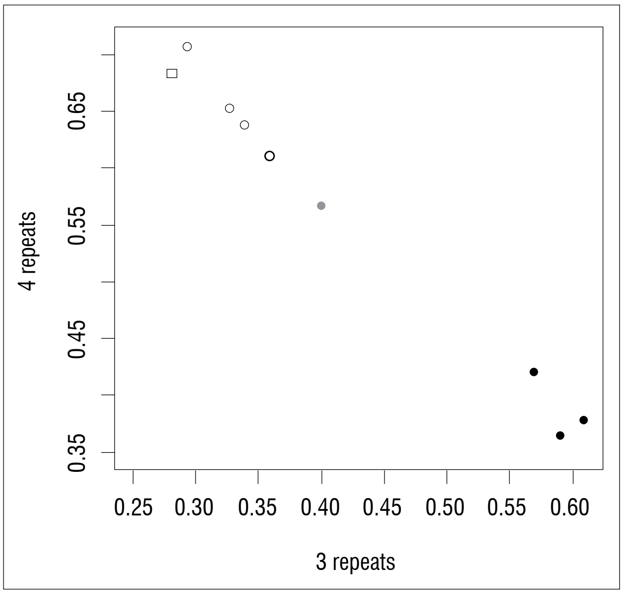


Figure 2: The frequencies of the two common *MAOA* alleles for the 10 populations included in Table 1. Open symbols indicate the six populations of a mostly European background that are not significantly different from one another and the filled symbols indicate the Asian and African populations. The grey circle represents the Italian population which differs significantly from the Afrikaner population (indicated by the square symbol), but not from the African American population.

AVPR1A

Alleles that contained a combined number of 32 CT and GT repeats ran to a length of 334 on our machine (Table 2). We identified 20 alleles in the Afrikaner population whereas the other studies identified 15 or 16 alleles (Table 2). Afrikaners had a higher expected heterozygosity of 0.86 compared to values ranging from 0.83 to 0.84 (Supplementary table 2). For population comparisons it was impossible to align the allele frequencies of one Asian study[34] with those of the Caucasoid populations because the allele frequency spectrum did not have the characteristic hallmarks of the locus in Caucasoid populations (Figure 1). As a result, no African or Asian populations could be compared to the Caucasoid populations. The frequencies of the adjusted allele sizes in other studies and ours are given in Table 2. With the allele sizes used as published and corrected for primer differences, pairwise F_{ST}s were as high as 0.07 and all populations differed significantly from one another (Supplementary table 3). With the corrected sizes as given in Table 2, the F_{ST} values were all lower than 0.006 and mostly considerably lower (Supplementary table 4). The two Israeli samples[30,33] were not significantly different from one another ($p=0.775$) but were significantly different from the other populations (all $p<0.045$), which in turn were not significantly different from one another (all $p>0.11$). The frequency of allele 334 differed among the population samples ($\chi^2 = 16.723$, df = 4, $p=0.0022$) as follows: Afrikaners = 0.21; Israeli 1[30] = 0.28; Israeli 2[33] = 0.29; British[37] = 0.25; and American, mainly Caucasoid[32] = 0.25. The pairwise test suggests that the Israeli sample had a significantly higher frequency of allele 334 than the Afrikaner ($p=0.006$) and European populations ($p=0.010$), but that the latter two did not differ significantly from each other ($p=0.143$).

Table 1: The observed number of the *MAOA* alleles in the Afrikaner population and in nine other populations

	Alleles					
Population group	**2**	**3**	**3.5**	**4**	**5**	***N***
Hispanic/Latino[a]	0	27	0	65	0	92
Afrikaner	0	55	0	134	7	196
White/non-Hispanic[a]	0	529	8	1056	26	1629
New Zealand, European origin[b]	3	658	9	1238	32	1940
German, European origin[c]	0	47	1	80	3	131
German, European origin[d]	3	140	3	238	6	390
Italian, European origin[c]	3	72	0	102	3	180
Chinese[e]	1	122	0	90	1	214
Asian/Pacific Islander[a]	0	50	1	31	0	82
African American[a]	0	52	2	32	2	88

Sources: [a]Sabol et al.[15]; [b]Caspi et al.[18]; [c]Deckert et al.[44]; [d]Kuepper et al.[13]; [e]Lu et al.[45]
Alleles 2–5 refer to the number of times the repeat element is repeated and N *is the total sample size. Populations are arranged in decreasing frequency of the allele with 4 repeats.*

Discussion

The two loci we considered provided two very different depictions. *MAOA* did not reveal any traces of admixture in the Afrikaner population as its allele frequency was displaced away from the African (as gauged from African American frequencies) and Asian populations rather than towards them (Figure 2). On the other hand, the *AVPR1A* showed an increased number of alleles in the Afrikaner population compared to other European populations, which could indicate the influence of admixture with older African and KhoeSan populations.[6] Neither locus suggested a strong deviation from European frequencies caused by a founder effect.

The unexpected high frequency of the 4-repeat allele of *MAOA* in the Afrikaner population (Figure 2) requires an explanation. We need to take into account that the founder effect was more severe for female individuals in the population[9]; despite an influx of male individuals, there was no such influx of female individuals.[6,9] In addition, because male individuals contribute only a single X chromosome, X chromosomes may have experienced a more severe bottleneck than other autosomal chromosomes. If we also consider that all non-European genetic contribution to this population was female derived,[6,7,9] it seems questionable that the frequency could be skewed away from African and Asian frequencies. It also is interesting that one of the other populations with an admixed heritage, the Latinos from America, also had a higher frequency of the 4-repeat allele (Figure 2).

Table 2: Allele frequencies at the *AVPR1A* RS3 locus for the Afrikaner and seven other populations

x=y[a]	Allele	Afrikaners	Israeli[b]	Israeli[c]	Israeli[d]	European origin			
						American[e]	British[f]	Swedish[g]	American[h]
	Correction[i]	0	+9	+9	+9	+2	-1	0	+2
	314	1							
	317			7					
	318	2							
	319	1		62			13		
	320	3				1		21	1.3
	322	1	1						
	324		6			1			
	326		1	39		1	4		
	328	1	2	26	0.6	4	4		0.2
	330	27	11	249	4.3	32	61	92	6.1
30	332	23	39	746	13.1	24	83	128	9.3
32	334	80	101	2000	24.4	112	257	**371**	22.2
33	336	87	**77**	1528	**19.2**	102	241	359	21.3
34	338	47	53	929	13.6	45	96	170	12.0
35	340	35	34	630	9.8	64	137	263	10.2
36	342	16	8	173	**2.4**	11	24	30	1.7
	344	7	6	95	2.4	2	2	23	2.8
38	346	26	9	304	6.1	24	69	126	8.3
	348	10	7	101		14	27	37	3.5
	349						3		
	350	2	5		2				0.2
	351						3		
	352	3			1				0.2
	354	1		26					
	356	1							0.2
	Total	374	360	6915		440	1024	1620	

Alleles in bold indicate alleles that were significantly linked to behavioural traits, whereas underlined alleles showed non-significant linkage with traits.

[a] *The combined number of CT and GT repeats confirmed by sequencing for the Afrikaner population. Thibonnier et al.'s[26] (CT)$_8$(GT)$_{24}$ corresponds with allele 334.*

[b] *Avinum et al.[30] The authors kindly provided us with the allele frequencies of the parent population. Allele 336 was significantly linked to reduced altruistic behaviour.*

[c] *Yirmiya et al.[33] frequencies from control samples. While RS3 locus did not show a significant link with autism, it did so in the proband when combined in haplotypes with other SSR loci of* AVPR1A.

[d] *Bachner-Melman et al.[39] relative frequencies based on 280 families. Not used for population frequency estimates.*

[e] *Kim et al.[32] 82% of the sample were Caucasoid, the remainder were African American, Asian American and Hispanic. Samples were from parents of autistic children. Allele 342 was marginally over-transmitted.*

[f] *Morley et al.[37] British choral singers and controls combined as RS3 had no significant effect.*

[g] *Walum et al.[25] This is not a random sample as twins were genotyped and so should not be used for population genetic comparisons. However, the allele spectrum should show similar hallmarks. Allele 334 was linked to reduced happiness in monogamous relationships.*

[h] *Meyer-Lindenberg et al.[28] relative frequencies based on 258 healthy people with an European ancestry. Allele 336 had the highest activation of amygdala, and overall, longer alleles had significantly stronger activation than shorter ones.*

[i] *Number of base pairs added to reported allele sizes to match allele frequency spectra best (compare to Figure 1).*

This finding may suggest that this allele may integrate more easily into a heterogeneous genetic background or that this small deviation simply stems from a small founder effect. Another explanation may be selection on the social phenotype of this allele. If *MAOA*-H (4-repeat allele) carriers are less likely to react aggressively in provocational circumstances[12,13] and by extension be less likely to be berserkers in war scenarios, cope better with maltreatment[18], and be less likely to develop antisocial behaviour[11], then the allele may have been selected for in this founding population. It is, however, important to note that this allele's frequency is not significantly higher in the Afrikaner population and we should caution against over-interpreting this result as such.

For *AVPR1A* RS3, it is firstly important to make sure that alleles compared among studies are indeed the same. Because of the complex nature of *AVPR1A* RS, the various repeat alleles that run to the same length may differ in their proportions of GT and CT repeats.[41] However, for a number of studies in which microsatellites in the promoter regions affect expression of the allele it seems to be the length rather than the content that is important.[41] In addition to this complication, it is easy to systematically call alleles a number of base pairs shorter or longer when their identity is inferred from the rates of amplicon migration in different genetic analysers and/or when labelled with different dyes etc.; for this reason, it is important to standardise allele size in some way. We followed two approaches that were both effective. Comparable lengths can be obtained either by sequencing amplicons to confirm their length or by comparing allele-frequency spectra between populations. In this light, it is important to note that this study only considered Caucasoid populations, and the same characteristic peaks may not be observed in other populations. In fact, it was impossible to align the Asian study[34] with the Caucasoid ones.

Confusion over which alleles are which is not trivial. The Israeli group has correctly linked their allele 327 to allele 334 from the Meyer-Lindenberg et al.[28] and Kim et al.[32] studies, but Walum et al.'s[25] 334 is in fact one base pair repeat longer than those in these other studies. As several researchers are comparing the effects of this locus on many behavioural patterns, it is important to have a gold standard to avoid confusion. We have compiled a ladder that can be used for such standardisation that is available on request. The confusion is not limited to comparisons of specific alleles; binning of alleles into groups of short, medium and long alleles can also lead to confusion. For instance, Prichard et al.[24] classified short alleles as those that have 12–19 repeats, medium alleles as those with 20–21 repeats and long alleles as those with 22–29 repeats; however, these numbers of repeats are substantially lower than those observed in this and other[26] studies (Table 2).

While the Afrikaner population had a lower frequency of the 334 allele, the frequency was not significantly lower and it would be premature to link their low levels of non-paternity with their low frequency of this allele. It is more likely that the Afrikaners' strong religious convictions[24] could explain their low non-paternity rate[19,20]. The fact that the Israeli samples had a significantly higher frequency of allele 334 suggests that an investigation into marital happiness may be interesting for this population.

Acknowledgements

We thank the participants for providing samples, and two reviewers and the associate editor for their thought-provoking comments. This work was supported by a National Research Foundation (NRF) incentive grant 77256 to J.M.G. This work is based on research supported in part by a number of grants from the NRF (UID: 78566 (NRF RISP grant for the ABI3500)). The grantholders acknowledge that the opinions, findings and conclusions or recommendations expressed in any publication generated by NRF-supported research are those of the authors, and that the NRF accepts no liability whatsoever in this regard.

Authors' contributions

A.K. performed the work under the mentorship of J.C.E. and J.M.G. J.C.E. and J.M.G. performed the analyses. A.K. wrote the initial draft of the manuscript, J.M.G. made major editorial adjustments and A.K., C.J.E. and J.M.G. made further editorial adjustments; all authors approved the final version.

References

1. De Villiers CC, Pama C. Geslagsregisters van die ou Kaapse families,1 A-K [Genealogies of old Cape families]. Cape Town: AA Balkema; 1966. Afrikaans.

2. Ridley M. Evolution. Oxford: Blackwell Science Ltd; 2004.

3. Botha MC, Beighton PB. Inherited disorders in the Afrikaner population of southern Africa. Part I. Historical and demographic background, cardiovascular, neurological, metabolic and intestinal conditions. S Afr Med J. 1983;64:609–612.

4. Botha MC, Beighton PB. Inherited disorders in the Afrikaner population of southern Africa. Part II. Skeletal, dermal and haematological conditions; the Afrikaners of Gamkaskloof; demographic considerations. S Afr Med J.1983;64:664–667.

5. Nurse GT, Weiner JS, Jenkins T. The peoples of southern Africa and their affinities. Oxford: Oxford University Press; 1985.

6. Heese JA. Die herkoms van die Afrikaner 1657–1867 [Ancestry of the Afrikaner 1657–1867]. Cape Town: AA Balkema; 1971. Afrikaans.

7. Heese HF. Groep sonder grense [Group without borders]. Pretoria: Protea Boekhuis; 2005. Afrikaans.

8. De Bruyn GFC. Die samestelling van die Afrikanervolk [The composition of the Afrikaner population]. Tydskr Geesteswet. 1976;15:39–42. Afrikaans.

9. Greeff JM. Deconstructing Jaco: Genetic heritage of an Afrikaner. Ann Hum Genet. 2007;71:674–688. http://dx.doi.org/10.1111/j.1469-1809.2007.00363.x

10. Botha MC, Pritchard J. Blood group gene frequencies – An indication of the genetic constitution of population samples in Cape Town. S Afr Med J. 1972;46:S1–S27.

11. Eisenberger NI, Way BM, Taylor SE, Welch WT, Lieberman MD. Understanding genetic risk for aggression: Clues from the brain's response to social exclusion. Biol Psychiatry. 2007;61:1100–1108. http://dx.doi.org/10.1016/j.biopsych.2006.08.007

12. Gallardo-Pujol D, Andres-Pueyo A, Maydeu-Olivares A. *MAOA* genotype, social exclusion and aggression: An experimental test of a gene–environment interaction. Genes Brain Behav. 2013;12:140–145. http://dx.doi.org/10.1111/j.1601-183X.2012.00868.x

13. Kuepper Y, Grant P, Wielpuetz C, Hennig J. *MAOA*-uVNTR genotype predicts interindividual differences in experimental aggressiveness as a function of the degree of provocation. Behav Brain Res. 2013;247:73–78. http://dx.doi.org/10.1016/j.bbr.2013.03.002

14. Lea R, Chambers G. Monoamine oxidase, addiction, and the "warrior" gene hypothesis. N Z Med J. 2007;120:u2441.

15. Lan NC, Heinzmann C, Gal A, Klisak I, Orth U, Lai E, et al. Human monoamine oxidase A and B genes map to Xp 11.23 and are deleted in a patient with Norrie disease. Genomics. 1989;4:552–559. http://dx.doi.org/10.1016/0888-7543(89)90279-6

16. Sabol SZ, Hu S, Hamer D. A functional polymorphism in the monoamine oxidase A gene promoter. Hum Genet. 1998;103:273–279. http://dx.doi.org/10.1007/s004390050816

17. Reti IM, Xu JZ, Yanofski J, Mckibben J, Uhart M, Cheng Y-J, et al. Monoamine oxidase A regulates antisocial personality in whites with no history of physical abuse. Compr Psychiatry. 2011;52:188–194. http://dx.doi.org/10.1016/j.comppsych.2010.05.005

18. Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, et al. Role of genotype in the cycle of violence in maltreated children. Science. 2002;297:851–854. http://dx.doi.org/10.1126/science.1072290

19. Greeff JM, Greeff FA, Greeff AS, Rinken L, Welgemoed DJ, Harris Y. Low nonpaternity rate in an old Afrikaner family. Evol Hum Behav. 2012;33:268–273. http://dx.doi.org/10.1016/j.evolhumbehav.2011.10.004

20. Greeff JM, Erasmus JC. Three hundred years of low non-paternity in a human population. Heredity. Forthcoming 2015. http://dx.doi.org/10.1038/hdy.2015.36

21. Greeff JM, Erasmus JC. Appel Botha Cornelitz: The abc of a three hundred year old divorce case. Forensic Sci Int Genet. 2013;7:550–554. http://dx.doi.org/10.1016/j.fsigen.2013.06.008

22. Giliomee H. Die Afrikaners: 'n Biografie [The Afrikaners: Biography of a people]. Cape Town: Tafelberg Uitgewers; 2004. Afrikaans.

23. Strassmann BI, Kurapati NT, Hug BF, Burke EE, Gillespie BW, Karafet TM, et al. Religion as a means to assure paternity. Proc Natl Acad Sci USA. 2012;109:9781–9785. http://dx.doi.org/10.1073/pnas.1110442109

24. Prichard ZM, Mackinnon AJ, Jorm AF, Easteal S. *AVPR1A* and OXTR polymorphisms are associated with sexual and reproductive behavioral phenotypes in humans. Hum Mutat. 2007;28:1150. http://dx.doi.org/10.1002/humu.9510

25. Walum H, Westberg L, Henningsson S, Neiderhiser JM, Reiss D, Igl W, et al. Genetic variation in the vasopressin receptor 1a gene (*AVPR1A*) associates with pair-bonding behavior in humans. Proc Natl Acad Sci USA. 2008;105:14153–14156. http://dx.doi.org/10.1073/pnas.0803081105

26. Thibonnier M, Graves MK, Wagner MS, Chatelain N, Soubrier F, Corvol P, et al. Study of V1-vascular vasopressin receptor gene microsatellite polymorphisms in human essential hypertension. J Mol Cell Cardiol. 200;32:557–564. http://dx.doi.org/10.1006/jmcc.2000.1108

27. Hammock EAD, Young LJ. Variation in the vasopressin V1a receptor promoter and expression: Implications for inter- and intraspecific variation in social behaviour. Eur J Neurosci. 2002;16:399–402. http://dx.doi.org/10.1046/j.1460-9568.2002.02083.x

28. Meyer-Lindenberg A, Kolachana B, Gold B, Olsh A, Nicodermus KK, Mattay V, et al. Genetic variants in *AVPR1A* linked to autism predict amygdala activation and personality traits in healthy humans. Mol Psychiatry. 2009;14:968-975. http://dx.doi.org/10.1038/mp.2008.54

29. Knafo A, Israel S, Darvasi A, Bachner-Melman R, Uzefovsky F, Cohen L, et al. Individual differences in allocation of funds in the dictator game associated with length of the arginine vasopressin 1a receptor RS3 promoter region and correlation between RS3 length and hippocampal mRNA. Genes Brain Behav. 2008;7:266–275. http://dx.doi.org/10.1111/j.1601-183X.2007.00341.x

30. Avinum R, Israel S, Shalev I, Gritsenko I, Bornstein G, Ebstein RP, et al. *AVPR1A* variant associated with preschoolers' lower altruism behaviour. PLoS One. 2011;6:e25274. http://dx.doi.org/10.1371/journal.pone.0025274

31. Bachner-Melman R, Zohar AH, Bacon-Shnoor N, Elizur Y, Nemanov L, Gritsenko I, et al. Link between vasopressin receptor *AVPR1A* promoter region microsatellites and measures of social behaviour in humans. J Indiv Diff. 2005;26:2–10. http://dx.doi.org/10.1027/1614-0001.26.1.2

32. Kim SJ, Young LJ, Gonen D, Veenstra-vanderWeele J, Courchesne R, Lord C, et al. Transmission disequilibrium testing of arginine vasopressin receptor 1a (*AVPR1A*) polymorphisms in autism. Mol Psychiatry. 2002;7:503–507. http://dx.doi.org/10.1038/sj.mp.4001125

33. Yirmiya N, Rosenberg C, Levi S, Salomon S, Shulman C, Nemanov L, et al. Association between the arginine vasopressin 1a receptor (*AVPR1a*) gene and autism in a family-based study: Mediation by socialization skills. Mol Psychiatry. 2006;11:488–494. http://dx.doi.org/10.1038/sj.mp.4001812

34. Yang SY, Cho S-C, Yoo HJ, Cho IH, Park M, Yoe J, et al. Family-based association study of microsatellites in the 5' flanking region of *AVPR1A* with autism spectrum disorder in the Korean population. Psychiatry Res. 2010;178:199–201. http://dx.doi.org/10.1016/j.psychres.2009.11.007

35. Granot RY, Frankel Y, Gritsenko V, Lerer E, Gritsenko I, Bachner-Melman R, et al. Provisional evidence that the arginine vasopressin 1a receptor gene is associated with musical memory. Evol Hum Behav. 2007;28:313–318. http://dx.doi.org/10.1016/j.evolhumbehav.2007.05.003

36. Ukkola-Vuoti L, Oikkonen J, Onkamo P, Karma K, Raijas P, Järvelä I. Association of the arginine vasopressin receptor 1A (*AVPR1A*) haplotypes with listening to music. J Hum Genet. 2011;56:324–329. http://dx.doi.org/10.1038/jhg.2011.13

37. Morley AP, Narayanan M, Mines R, Molokhia A, Baxter S, Craig G, et al. *AVPR1a* and *SLC6A4* polymorphisms in choral singers and non-musicians: A gene association study. PloS One. 2012;7:e31763. http://dx.doi.org/10.1371/journal.pone.0031763

38. Bachner-Melman R, Dina C, Zohar AH, Constantini N, Lerer E, Hoch S, et al. AVPR1a and SLC6A4 gene polymorphisms are associated with creative dance performance. PloS Genet. 2005;1:e42. http://dx.doi.org/10.1371/journal.pgen.0010042

39. Bachner-Melman R, Zohar AH, Elizur Y, Nemanov L, Gritsenko I, Konis D, et al. Association between a vasopressin receptor *AVPR1a* promoter region microsatellite and eating behaviour measured by a self-report questionnaire (Eating Attitudes Test) in a family-based study of a nonclinical population. Int J Eat Dis. 2004;36:451–460. http://dx.doi.org/10.1002/eat.20049

40. Hall TA. Bioedit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 1999;41:95–98.

41. Pritchard Z, Easteal S. Characterization of simple sequence repeat variants linked to candidate genes for behavioural phenotypes. Hum Mut. 2006;27:120–126. http://dx.doi.org/10.1002/humu.9394

42. Excoffier L, Laval G, Schneider S. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinform Online. 2005;1:47–50.

43. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2013.

44. Deckert J, Catalano M, Syagailo YV, Bosi M, Okladnova O, Di Bella D, et al. Excess of high activity monoamine oxidase A gene promoter alleles in female patients with panic disorder. Hum Mol Genet. 1999;8:621–624. http://dx.doi.org/10.1093/hmg/8.4.621

45. Lu R-B, Lee J-F, Ko H-C, Lin W-W, Chen K, Shih JC. No association of the MAOA gene with alcoholism among Han Chinese males in Taiwan. Prog Neuro-Psychopharmacol Biol Psychiatry. 2002;26:457–461. http://dx.doi.org/10.1016/S0278-5846(01)00288-3

Note: This article is supplemented with online only material.

Fibre optic cables navigating the world (credit: Greatstock/Corbis). In an article on page 67, Pillay and colleagues discuss the use of polarisation-encoded quantum key distribution in optical fibre networks to encode digital information.

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