

Biodiversity informatics: Building a lifeboat for high functionality data to decision pipeline

Edited by

Cang Hui, Nick Isaac, Quentin Groom,
Vernon Visser and Sandra MacFadyen

Published in

Frontiers in Ecology and Evolution
Frontiers in Environmental Science



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-4578-2
DOI 10.3389/978-2-8325-4578-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Biodiversity informatics: Building a lifeboat for high functionality data to decision pipeline

Topic editors

Cang Hui — Stellenbosch University, South Africa

Nick Isaac — NERC Centre for Ecology & Hydrology, United Kingdom

Quentin Groom — Botanic Garden Meise, Belgium

Vernon Visser — University of Cape Town, South Africa

Sandra MacFadyen — Stellenbosch University, South Africa

Citation

Hui, C., Isaac, N., Groom, Q., Visser, V., MacFadyen, S., eds. (2024). *Biodiversity informatics: Building a lifeboat for high functionality data to decision pipeline*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4578-2

Table of contents

- 05 **Editorial: Biodiversity informatics: building a lifeboat for high functionality data to decision pipeline**
Cang Hui, Sandra MacFadyen, Vernon Visser, Quentin Groom and Nick J. B. Isaac
- 08 **Drivers of compositional turnover of narrow-ranged versus widespread naturalised woody plants in South Africa**
Isabella W. de Beer, Cang Hui, Christophe Botella and David M. Richardson
- 20 **BIRDIE: A data pipeline to inform wetland and waterbird conservation at multiple scales**
Francisco Cervantes, Res Altwegg, Francis Strobbe, Andrew Skowno, Vernon Visser, Michael Brooks, Yvan Stojanov, Douglas M. Harebottle and Nancy Job
- 33 **Taking state of biodiversity reporting into the information age – A South African perspective**
Carol Jean Poole, Andrew Luke Skowno, Jock C. Currie, Kerry Jennifer Sink, Brenda Daly and Lize von Staden
- 39 **Collating biodiversity occurrence data for conservation**
Dian Spear, Nicola J. van Wilgen, Anthony G. Rebelo and Judith M. Botha
- 45 **How the Freshwater Biodiversity Information System (FBIS) is supporting national freshwater fish conservation decisions in South Africa**
Mohammed Kajee, Dominic A. W. Henry, Helen F. Dallas, Charles L. Griffiths, Josephine Pegg, Dewidine Van der Colff, Dean Impson, Albert Chakona, Domitilla C. Raimondo, Nancy M. Job, Bruce R. Paxton, Martine S. Jordaan, Roger Bills, Francois Roux, Tsungai A. Zengeya, Andre Hoffman, Nick Rivers-Moore and Jeremy M. Shelton
- 56 **Iterative mapping of marine ecosystems for spatial status assessment, prioritization, and decision support**
Kerry J. Sink, Luther A. Adams, Mari-Lise Franken, Linda R. Harris, Jock Currie, Natasha Karenyi, Anisha Dayaram, Sean Porter, Stephen Kirkman, Maya Pfaff, Lara van Niekerk, Lara J. Atkinson, Anthony Bernard, Mariel Bessinger, Hayley Cawthra, Willem de Wet, Loyiso Dunga, Zoleka Filander, Andrew Green, David Herbert, Stephen Holness, Stephen Lamberth, Tamsyn Livingstone, Melanie Lück-Vogel, Fiona Mackay, Mapula Makwela, Ryan Palmer, Wilhem Van Zyl and Andrew Skowno
- 73 **South Africa's initiative toward an integrated biodiversity data portal**
Brenda Daly and Fhatani Ranwashe

- 80 **A conceptual approach to developing biodiversity informatics as a field of science in South Africa**
Fatima Parker-Allie, Mark J. Gibbons and Douglas M. Harebottle
- 95 **Reproducible WiSDM: a workflow for reproducible invasive alien species risk maps under climate change scenarios using standardized open data**
Amy J. S. Davis, Quentin Groom, Tim Adriaens, Sonia Vanderhoeven, Rozemien De Troch, Damiano Oldoni, Peter Desmet, Lien Reyserhove, Luc Lens and Diederik Strubbe



OPEN ACCESS

EDITED AND REVIEWED BY
Alexander Kokhanovsky,
German Research Centre for Geosciences,
Germany

*CORRESPONDENCE
Sandra MacFadyen
✉ sandra@biogis.co.za

RECEIVED 16 February 2024
ACCEPTED 19 February 2024
PUBLISHED 28 February 2024

CITATION
Hui C, MacFadyen S, Visser V, Groom Q
and Isaac NJB (2024) Editorial: Biodiversity
informatics: building a lifeboat for high
functionality data to decision pipeline.
Front. Ecol. Evol. 12:1386917.
doi: 10.3389/fevo.2024.1386917

COPYRIGHT
© 2024 Hui, MacFadyen, Visser, Groom and
Isaac. This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use,
distribution or reproduction in other forums
is permitted, provided the original author(s)
and the copyright owner(s) are credited and
that the original publication in this journal is
cited, in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Editorial: Biodiversity informatics: building a lifeboat for high functionality data to decision pipeline

Cang Hui^{1,2,3}, Sandra MacFadyen^{1,2*}, Vernon Visser^{4,2},
Quentin Groom⁵ and Nick J. B. Isaac⁶

¹Mathematical Biosciences Lab, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa, ²National Institute for Theoretical and Computational Sciences (NITheCS), Stellenbosch, South Africa, ³Biodiversity Informatics Unit, African Institute for Mathematical Sciences, Muizenberg, South Africa, ⁴Centre for Statistics in Ecology, The Environment and Conservation, University of Cape Town, Cape Town, South Africa, ⁵Meise Botanic Garden, Meise, Belgium, ⁶UK Centre for Ecology and Hydrology, Wallingford, United Kingdom

KEYWORDS

ecological modelling, data management, citizen science, FAIR (findable accessible interoperable and reusable) principles, climate change, invasive species

Editorial on the Research Topic

Biodiversity informatics: building a lifeboat for high functionality data to decision pipeline

Biodiversity informatics is a multidisciplinary field that focuses on the use of computer technology to manage, explore, analyse, and interpret biological data (Soberon and Peterson, 2005; Guralnick and Hill, 2009). This field is designed to meet the worldwide requirements for biodiversity monitoring, addressing both the necessity for and the challenges associated with sharing data. Some of the key focus areas within biodiversity informatics include: 1) Taxonomy and genomics: the classification and naming of organisms, as well as the construction of phylogenetic trees to show evolutionary relationships. 2) Ecological modelling: the use of statistical and computational methods to understand the distribution and abundance of species across different habitats. 3) Data management and sharing: the development of databases and online platforms to store and share biodiversity data with fellow researchers and stakeholders. 4) Citizen science: engaging the public in collecting and sharing biodiversity data, using tools such as smartphone apps and online portals.

Data availability serves as the cornerstone of biodiversity science, fuelling research and discovery to inform evidence-based decisions for biodiversity conservation. However, resource limitations (financial, technological skills and infrastructure) severely hamper the establishment of inclusive information and communication technology (ICT) solutions and research and development of related approaches and tools. While the world is firmly cemented in a data age, we are drowning in data but thirsty for information and the synthesis of knowledge into understanding. The volume, diversity and speed at which new environmental and ecological data, in particular, are being generated is growing exponentially as biodiversity continues to decline worldwide (MacFadyen et al., 2022). Those able to successfully generate, collect, store and curate, share, analyse and

communicate the existence of clarified synthesised biodiversity data, will become central players in informing and influencing the debate around global biodiversity change.

The Biodiversity Informatics Symposium was held at the Stellenbosch Institute for Advanced Study (STIAS) from 28-30 November 2022. The symposium brought together 68 researchers, managers, and practitioners from across South Africa and abroad, with expertise in a wide range of fields including conservation, ecology, information science, information technology, mathematics and statistics. During the symposium, we discussed challenges, highlighted opportunities, and encouraged innovative solutions for biodiversity informatics in South Africa and beyond. Discussions were centred around six keynote addresses, 34 topical presentations and three facilitated panel discussions over three days. Emerging from this symposium, nine pivotal articles are featured here: [Davis et al.](#); [Cervantes et al.](#); [Daly and Ranwashe](#); [Kajee et al.](#); [Sink et al.](#); [Poole et al.](#); [Parker-Allie et al.](#); [de Beer et al.](#); and [Spear et al.](#). These articles focus on enhancing data management through standardisation, embracing citizen science, and innovative tools, while also highlighting the critical roles of macroecological insights, the impact of alien species, and robust data infrastructure in conservation.

Emergent questions and solutions within the field of biodiversity informatics in South Africa include challenges like reluctance in data sharing, the vast amount of undigitized data, and the need for improved data management practices. Solutions proposed include changing perceptions about data sharing, employing Optical Character Recognition (OCR) for faster digitization, and advocating for standardised data formats and interoperability. These efforts aim to enhance the accessibility and utility of biodiversity data, fostering collaborative and informed conservation strategies. More specifically, biodiversity informatics in South Africa presents both challenges and promising solutions. One of the primary questions in this field revolves around data management and integration. How can data from various sources, including government agencies, research institutions, and citizen science initiatives, be effectively harmonised and centralised? Solutions involve the development of robust databases and data-sharing protocols, fostering collaboration among stakeholders, and utilising advanced technologies like machine learning for data analysis ([Parker-Allie et al.](#)).

Other challenges stem from elements contributing to the global biodiversity crisis, including: i) Invasive species, which pose a significant threat to biodiversity. How can biodiversity informatics aid in tracking and managing these invasions effectively? Solutions include the development of early warning systems using remote sensing and geographic information systems (GIS) to monitor changes in vegetation and habitat. Additionally, promoting public awareness and community involvement in invasive species management is essential. ii) Climate change is another key concern for biodiversity conservation. How can climate data be integrated with biodiversity information to predict ecological changes? Proposed solutions encompass interdisciplinary research, leveraging climate models, and working with policymakers to implement conservation measures that account for climate impacts. iii) Habitat fragmentation and urbanisation

threaten South Africa's unique biodiversity. How can we address the challenge of fragmented habitats? Solutions include prioritising conservation in land-use planning, establishing wildlife corridors, and incentivising private landowners to participate in conservation efforts. iv) Engaging local communities in biodiversity informatics efforts is crucial. How can indigenous knowledge be incorporated, and communities be empowered to become stewards of their natural resources? Solutions involve community-based monitoring programs, culturally sensitive conservation initiatives, and integrating traditional ecological knowledge into biodiversity databases. iv) Capacity building and education are vital for the sustainability of biodiversity informatics in South Africa. How can we train a skilled workforce and promote data literacy? Solutions encompass the establishment of training programs, workshops, and educational initiatives at various levels, from school curricula to professional development opportunities. Implementing innovative solutions in these areas is essential for preserving the nation's rich biodiversity and contributing to global conservation efforts.

Looking forward, we anticipate significant growth in the field of biodiversity informatics both regionally and globally in the coming decades. This growth will be supported by well-established and documented data pipelines and analysis protocols that facilitate the following key developments: (1) Development of standards for data exchange and interoperability; (2) Improvement of data quality and completeness; (3) Enhancement of data integration and synthesis; (4) Development of tools for data analysis and modelling; (5) Improvement of access to data and information; (6) Development of methods for assessing and predicting biodiversity change; (7) Building capacity in biodiversity informatics; (8) Addressing ethical and legal issues related to data sharing and use. These developments are expected to empower researchers, policymakers, and conservationists with better tools and insights for effectively managing South Africa's rich biodiversity and addressing environmental challenges.

Author contributions

CH: Writing – original draft, Writing – review & editing. SM: Writing – original draft, Writing – review & editing. VV: Writing – review & editing. QG: Writing – review & editing. NI: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The symposium was funded by the National Institute for Theoretical and Computational Sciences (NITheCS), research programme Advancing Biodiversity Informatics & Ecological Modelling. CH and SM acknowledge support from the NRF (grant 89967); QG, CH, and SM acknowledge support from the EU Horizon project (Biodiversity Building Blocks for policy; 101059592); NI and CH also acknowledge support from the UK NERC project (GLocal Insect Threat-Response Synthesis; NE/V007548/1).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organisations, funders, or publishers.

References

- Guralnick, R., and Hill, A. (2009). Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. *Bioinformatics* 25, 421–428. doi: 10.1093/bioinformatics/btn659
- MacFadyen, S., Allsopp, N., Altwegg, R., Archibald, S., Botha, J., Bradshaw, K., et al. (2022). Drowning in data, thirsty for information and starved for understanding: A biodiversity information hub for cooperative environmental monitoring in South Africa. *Biol. Conserv.* 274, 109736. doi: 10.1016/j.biocon.2022.109736
- Soberon, J., and Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Inf.* 2, 1–10. doi: 10.17161/bi.v2i0.4



OPEN ACCESS

EDITED BY

Rubén G. Mateo,
Autonomous University of Madrid, Spain

REVIEWED BY

Res Altwegg,
University of Cape Town,
South Africa
Irene Martin-Fores,
University of Adelaide,
Australia

*CORRESPONDENCE

Isabella W. de Beer
✉ isabella.debeer8@gmail.com

[†]These authors have contributed equally to this work

SPECIALTY SECTION

This article was submitted to
Biogeography and Macroecology,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 23 November 2022

ACCEPTED 30 January 2023

PUBLISHED 21 February 2023

CITATION

de Beer IW, Hui C, Botella C and
Richardson DM (2023) Drivers of compositional
turnover of narrow-ranged versus widespread
naturalised woody plants in South Africa.
Front. Ecol. Evol. 11:1106197.
doi: 10.3389/fevo.2023.1106197

COPYRIGHT

© 2023 de Beer, Hui, Botella and Richardson.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Drivers of compositional turnover of narrow-ranged versus widespread naturalised woody plants in South Africa

Isabella W. de Beer^{1*}, Cang Hui^{2,3,4†}, Christophe Botella^{1,2†} and David M. Richardson^{1,5}

¹Centre for Invasion Biology, Department of Botany and Zoology, Stellenbosch University, Stellenbosch, South Africa, ²Centre for Invasion Biology, Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa, ³Biodiversity Informatics Unit, African Institute for Mathematical Sciences, Cape Town, South Africa, ⁴National Institute for Theoretical and Computational Sciences (NITheCS), Stellenbosch University, Stellenbosch, South Africa, ⁵Department of Invasion Ecology, Institute of Botany, Czech Academy of Sciences, Průhonice, Czechia

Introduction: Alien trees and shrubs have become increasingly common invaders globally and have caused major negative impacts to ecosystems and society. Non-native woody plant species make up the majority of legislated invasive alien taxa in South Africa and contribute substantially to recorded negative impacts. It is of management interest to elucidate the macroecological processes that mediate the assembly of alien taxa, as this is expected to be associated with anthropogenic factors (e.g., human activity, introduction events, pathways of propagule dispersal mediated by humans) and bioclimatic factors (such as diurnal temperature range and precipitation gradients). These analyses require large species-occurrence datasets with comprehensive sampling across broad environmental conditions. Efforts of citizen scientists produce large numbers of occurrence records in a consistent manner which may be utilised for scientific investigations.

Methods: Research Grade occurrence data on naturalised plants of South Africa were extracted from the citizen scientist platform iNaturalist. Sampling bias was mitigated using statistical modelling of background points estimated from a Target Group of species which identifies well sampled communities. The drivers of assembly for alien plants at different range sizes were identified using multi-site generalised dissimilarity modelling (MS-GDM) of zeta diversity. The predicted compositional similarity between all cells was computed based on the subset of identified well sampled communities and using generalised dissimilarity modelling (GDM). From this, alien bioregions were identified using a k-means cluster analysis.

Results and Discussion: Bioclimatic factors significantly influenced community turnover in inland areas with large diurnal temperature ranges, and in areas with high precipitation. Communities separated by large geographical distances had significantly different compositions, indicating little contribution of long-range propagule movement by humans, and the presence of localised introduction hubs within the country which harbour unique species compositions. Analyses also showed a significant contribution of road density to turnover, which may be moderated by the habitat service provided by road verges. The same is true for natural dispersal via rivers in arid areas. The distribution of naturalised tree and shrub species is geographically clustered and forms six alien bioregions that are distinct from the South African biomes defined by native species distribution analysis.

KEYWORDS

invasion biology, generalized dissimilarity model, zeta diversity, iNaturalist, species occurrence data, biogeography

1. Introduction

The consequences of introductions of non-native species have been investigated globally and remain a major concern. Invasions of non-native trees and shrubs have increased rapidly in many parts of the world in recent decades, and these growth forms now feature prominently on many lists of the most problematic invasive species worldwide (Richardson and Rejmánek, 2011; Richardson et al., 2014). In South Africa, woody tree and shrub species are overrepresented in the invasive alien flora (Richardson et al., 2020). In reviewing the impacts of alien plant species in South Africa, van Wilgen et al. (2022) showed that invasive trees and shrubs contribute the majority of recorded impacts. The cause of impacts in invaded ecosystems is mainly through competitive exclusion of native species and by changes to ecosystem functioning. Formal impact assessments of species in the genus *Acacia* have shown that these species consistently have massive and major impact potential across many habitat types in South Africa (Jansen and Kumschick, 2022). The management of these impacts can be aided by understanding large-scale macroecological processes which govern the community assembly of taxa. Such insights can inform regional strategies to reduce the impacts of invasions.

Little attention has been given to the assemblage-level drivers of alien taxa. This is akin to studying assembly processes of native vegetation units, or biomes, which have been shaped over evolutionary time scales. Biomes can be defined using multiple criteria, but because of the evolutionary time scale, regional climatic conditions or climate niches are often used (Rutherford et al., 2006). In contrast, assemblages of alien taxa have had relatively little residence time to form stable communities (Hui et al., 2013). Their distributions may thus be largely constrained by processes like introduction history and diverse dispersal pathways which act during the initial spread of alien populations (Donaldson et al., 2014). Of the many natural and human-mediated dispersal pathways, river systems have been found to contribute substantially to turnover at the local and regional scales (Merritt et al., 2010). Evidently, dependent on river flow dynamics and the physical properties of propagules, hydrochorous seeds often accumulate seedbanks in riparian zones (Gurnell et al., 2008). In contrast, human-mediated dispersal, such as along road verges (Christen and Matlack, 2009) can facilitate long-distance spread of alien species that possess specific reproductive strategies (Skultety and Matthews, 2017). For instance, cultivation of ornamental and horticultural species by commercial nurseries has increasingly contributed to plant invasions (van Kleunen et al., 2018).

It may be informative for the management of invasions, therefore, to investigate whether the assembly processes of alien species are influenced by the relatively short time scale of introduction in new ranges, and their close association with anthropogenic factors (e.g., Lenzner et al., 2022). We would expect a close association between anthropogenic factors (such as the presence of roads which are correlated with human activity), introduction history and alien plant distribution. In studying the drivers of community composition of alien assemblages it is thus appropriate to include both climatic conditions which place fundamental constraints on plant survival and reproduction, and factors associated with human mediation and facilitation of invasions.

Alien biomes in South Africa have been studied using a phytogeographic classification which clusters communities as regions with distinct species composition. In South Africa, the geography of alien phytochoria, or regions inhabited by unique communities of alien

plants, was first investigated by Hugo et al. (2012); they found eight clusters of communities of naturalised and invasive plant taxa closely associated with biomes and prevailing climatic conditions, but with limited influence of human activities like irrigation and agriculture. Similar results were reported by Rouget et al. (2015) using a classification tree. The five clusters found by Rouget et al. (2015) were best explained by climatic niches and overlapped with biomes, although biomes are also shaped by climatic conditions. Little evidence was found for the role of introduction histories, disturbance, and possible emerging competition between native and alien plants, which is unexpected given the close association between anthropogenic factors and invasions. Richardson et al. (2020) followed similar methods to divide alien plant taxa into four distinct regions, although the drivers of alien composition were only informed by literature and expert knowledge. Questions remain regarding the number of distinct alien biomes and drivers of alien compositional turnover in South Africa.

Another option for determining alien biomes is to apply Generalised Dissimilarity Modelling (GDM; Ferrier et al., 2007) using the relationship between species turnover (i.e., beta diversity, between two sites) and change in environmental drivers. Mapped species turnover can further determine areas likely to contain unique species composition and unveil spatial regionalisation of biodiversity (Basel et al., 2021; Mokany et al., 2022). To our knowledge, this method has not been applied previously in a study of non-native species assemblages over large spatial scales. Estimated contributions of different drivers to species turnover can be tentatively projected to under-sampled areas for compositional similarity mapping of the entire region. However, identified drivers from GDM of pairwise beta diversity are biased towards explaining compositional turnover by predominantly narrow-ranged species present in few sites. To alleviate this bias, a newer method, multi-site generalised dissimilarity modelling (MS-GDM) of Zeta diversity (ζ), was designed to explicitly account for the different contributions to compositional turnover of narrow-ranged versus widespread species (Hui and McGeoch, 2014; McGeoch et al., 2019). Zeta diversity is the average number of species shared by a given number (order) of sites and has diverse applications when investigating compositional similarity of communities (e.g., of stream invertebrates between different streams, Simons et al., 2019; the parasites associated with small mammal hosts, Krasnov et al., 2020; or of native versus invasive ants across oceanic islands, Latombe et al., 2019). What makes the concept of zeta diversity particularly useful in the context of biological invasions is that it allows us to use MS-GDM and extract the drivers responsible for different stages of the introduction-naturalisation-invasion continuum (hereafter the invasion continuum; Blackburn et al., 2011). Indeed, according to Richardson et al. (2000) the naturalised (self-sustaining populations in the invaded range) and invasive (naturalised species which have spread significantly from their point of introduction and are reproductively successful at distant locations) stages of the continuum are generally associated with different range sizes (see also Richardson and Pyšek, 2012). Considering the zeta diversity metric in this context, lower-order zeta diversity thus represents compositional turnover from predominantly newly established and thus narrow-ranged alien species, with identified drivers thus primarily associated with the establishment stage. In contrast, higher-order zeta diversity reflects compositional turnover of more widespread alien species, and identified drivers are thus associated with the invasion stage (Hui and McGeoch, 2014; Latombe et al., 2017). Previous work using zeta diversity in understanding drivers of community turnover has described unique drivers at play for native and alien plant species with

different range sizes in the Czech Republic (Latombe et al., 2017). This method has not been applied to non-native taxa in South Africa.

Most literature on the biogeography and macroecology of non-native plants in South Africa is based on analyses of professionally curated data sources. Occurrence databases like the South African Plant Invaders Atlas (SAPIA) have been widely used in studies of alien taxa for which species identifications were confirmed by members of the scientific community (Henderson and Wilson, 2017). Occurrence records in SAPIA are largely from roadside surveys (Rouget et al., 2015). Standardised data such as local abundance, percentage cover and occurrence of alien introductions are exclusively mapped during focussed expert studies (Watt et al., 2009; Kalwij et al., 2014; Cheek and Semple, 2016). However, these methods are limited by time and resources and may be too data poor to address macroecological questions. A review by Richardson et al. (2020) on the biogeography of terrestrial plant invasions highlights the need for accurate and extensive data on the distributions of alien plants in order to manage present invasions and assess their possible future impacts. The increasing popularity of crowdsourced data platforms may contribute to fill this data gap where large datasets are needed to guide biodiversity conservation (MacFadyen et al., 2022).

Crowdsourced presence-only data are useful to the scientific community since they continuously produce massive amounts of information in a consistent format and a sustainable manner. Such data have been used in investigations of the distribution of invasive species, like the African carder bee in Australia (Dart et al., 2022). Although bias and low data quality are potentially problematic, methods are available to reduce the influence of these issues (Bird et al., 2014). When using such data, it is important to account for users' sampling behaviour and spatial sampling bias in data collection, as reviewed by Dickinson et al. (2010), which can result in filtrating uninformative data. Spatial sampling bias, namely the distortion of species distributions due to spatial heterogeneity in sampling effort of the observers, can be approximated by the spatial concentration of records from a group of species generally sampled together by the observers (called a Target-Group, TG; Phillips et al., 2009). This approximation is justified if the cumulated distribution of TG species does not vary greatly over space (Botella et al., 2020) and if observers have similar sampling preferences towards TG species (although not necessarily true as observer experience can influence species reporting rate; Johnston et al., 2018). Accounting for sampling heterogeneity, this study aimed at capitalising on crowdsourced plant records generated through iNaturalist (iNaturalist, 2022), a popular global crowdsourcing platform. Observers can upload geolocated records of any species, and the community assists in verifying species identifications. Although this data source is appealing, not all studies will benefit from using iNaturalist data over those collected and collated by experts. This is because some taxonomic groups receive less attention on citizen science platforms, and some cannot easily be recorded with standard smart-phone cameras (Hochmair et al., 2020). Nonetheless, there are indications of the valuable contribution that this dataset offers to the scientific community, especially in countries or regions where there are many active users and where the subject species are conspicuous (Mesaglio and Callaghan, 2021), which is the case for naturalised trees and shrubs in South Africa.

Our aim is to understand the drivers of community assembly of woody naturalised trees and shrubs in South Africa, and to investigate whether these drivers differ between species at different stages along the invasion continuum, particularly species at an early stage of establishment with localised spread versus those that have attained

larger geographical ranges. We hypothesise that the drivers of narrow-ranged and widespread species will differ. Specifically, narrow-ranged species are hypothesised to be driven by anthropogenic factors since these species are still confined to areas close to sites of introduction, whereas invasions of widespread species should be driven more by environmental conditions since these species would have been filtered by the environment in the expanded range. For better area-based management, we also aim to describe spatial clusters of alien tree and shrub species across South Africa and summarise properties of these groupings, making use of data curated by citizens on iNaturalist. Although this project aims to investigate many factors, we can do so because of the versatile method of multi-site generalised dissimilarity modelling with zeta diversity.

2. Methods

2.1. Focal species

Focal species of this investigation included tree and shrub species naturalised in South Africa. These species were identified using the definitions of trees and shrubs proposed by Richardson and Rejmánek (2011) and additional expert advice whilst cross-referencing the classifications of alien trees and shrubs made by Henderson (2020) (Supplementary Table S1).

In this study, the species composition of a single quarter-degree grid cell, an area of approximately 25 × 25 km, is considered a 'community'. Occurrence data were extracted from the iNaturalist website on 27 July 2022 with a query specifying the iNaturalist project "Naturalized Plants of South Africa"¹. This extraction contained all Research Grade (RG) observations made by iNaturalist observers within South Africa's national borders for naturalised plant species listed in Richardson et al. (2020) based largely on the list compiled for South Africa's National Status Report (Zengeya and Wilson, 2020) with additions as detailed on the iNaturalist project page. Plant observations labelled RG are expected to be wild-growing individuals where species identification have been verified and where two-thirds of identifications agree. However, even with this stringency measure in place there is still not absolute confidence in the data quality as identifications may be mislabelled, especially in cases where subspecies are identified, and where cultivated specimens are incorrectly recorded as growing wild. For research on invasions, the separation of wild-growing, self-sustaining populations from planted individuals is crucial for the understanding of spread and progression of the invasion. Consequently, substantial effort was made to manually filter the iNaturalist records to ensure correct identifications and labelling. As a further measure to ensure data quality, we only used records identified to the species level; if a record was identified at a finer level (e.g., subspecies), we retained only the species name.

This occurrence dataset was converted to a presence/absence matrix based on the grid of quarter-degree cells widely used for biodiversity atlases in South Africa, using the *letsR* package v4.0 (Vilela and Villalobos, 2015). All analyses were conducted in R version 4.1.2 (R Core Team, 2021). We initially identified a total of 246 naturalised tree and shrub species (all used to build the Target-Group, explained below), but finally kept for the analysis only the 190 species that were

¹ <https://www.inaturalist.org/projects/naturalized-plants-of-south-africa>

present in the 68 selected cells (see section 2.2). These species are hereafter called the focal species. Focal species occurred in 801 quarter-degree grid cells of South Africa (42% of all 1900 grid cells across the country) (Figure 1A) based on 40,491 records uploaded between the period of 2011–2022 by 3,153 observer, with *Acacia mearnsii* having the largest coverage of 265 grid cells (13.9%). The number of FS occurrence records showed near identical trends in concentration of richness (Figure 1B), indicating that high species richness may be skewed by sampling effort in those cells, and not necessarily representative of true richness. When visualising the relationship between record count and richness, we see that cells around the Cape Peninsula in the southwest are consistently highly sampled with many records per species, and many regions are poorly sampled in the country (Figure 2). The distribution of record counts per cell was highly left-skewed, ranging from 1 to 6,334, with an average of 51 and a median of 5 observations per cell.

The least visited cells will likely have many false absences, under-representing their true species richness, with a likely detection bias towards the most conspicuous species. Following the rationale presented in Botella et al. (2020) and implemented in Botella et al. (2022), we designed a procedure to tackle this problem and detection variability across observers. Briefly, we first computed an approximation of sampling effort per cell which is positively correlated with the detection probability of all focal species. We then minimised the effects of sampling bias by retaining only cells with minimal numbers of potentially undetected species estimated from the sampling effort approximation per cell.

Botella et al. (2020) showed that the spatial variations of sampling effort may be well approximated by the total number of records of a Target-Group (TG) of species under certain conditions. Firstly, the TG species must be generally reported together with the focal species by the observers, so that cells having more TG occurrences can be assumed to have been better sampled for the focal species, thereby increasing the chances of detecting the ones that are present. Since our

focal species are trees and shrubs, we selected all terrestrial plant species in the generic sense as our candidate TG; hence we extracted all the Research Grade (RG) iNaturalist observations of terrestrial plants from the Global Biodiversity Information Facility extracted through (GBIF.org, 2022). Secondly, to reduce observer bias, we kept only records from observers who reported more records than species, thereby favouring observers who have repeatedly sampled some species and thus are assumed to sample species irrespective of whether they recorded these species previously. Thirdly, the TG species must be selected so that their cumulated abundances across space is roughly constant (Botella et al., 2020), otherwise species-rich areas would have more TG record counts even under constant sampling effort (this is likely to be the case here given the spatial variability of richness across South Africa; see, e.g., Supplementary Figure S1). To approximate this spatially homogeneous situation, we used a heuristic method for TG selection proposed by Botella et al. (2022). The procedure sequentially adds species to the TG, firstly maximising the spatial coverage of TG species to increase the breadth of environments sampled, and secondly maximising the evenness of richness (Shannon entropy) of TG species across cells to reduce the unequal distribution of richness of the TG. It follows that the species with the highest number of presences across grid cells would be selected first, and that each subsequent addition of species to the TG should increase the coverage of the TG and then the evenness of species richness across cells. From the GBIF dataset, a subset of 1,788 species was included in the TG, covering all 1,548 grid cells of the GBIF dataset with the Shannon entropy improved from 6.214 to 6.355.

The number of TG occurrences per grid cell provides an estimate of the sampling effort. We used this sampling effort estimate and one extra parameter, called K , to determine, for each cell and each undetected focal species in that cell, whether the focal species absence was certain or not. The value of K (integer between 1 and 100) determines the amount of sampling effort required to detect each species in any cell. Specifically, we computed one detection threshold per focal species as

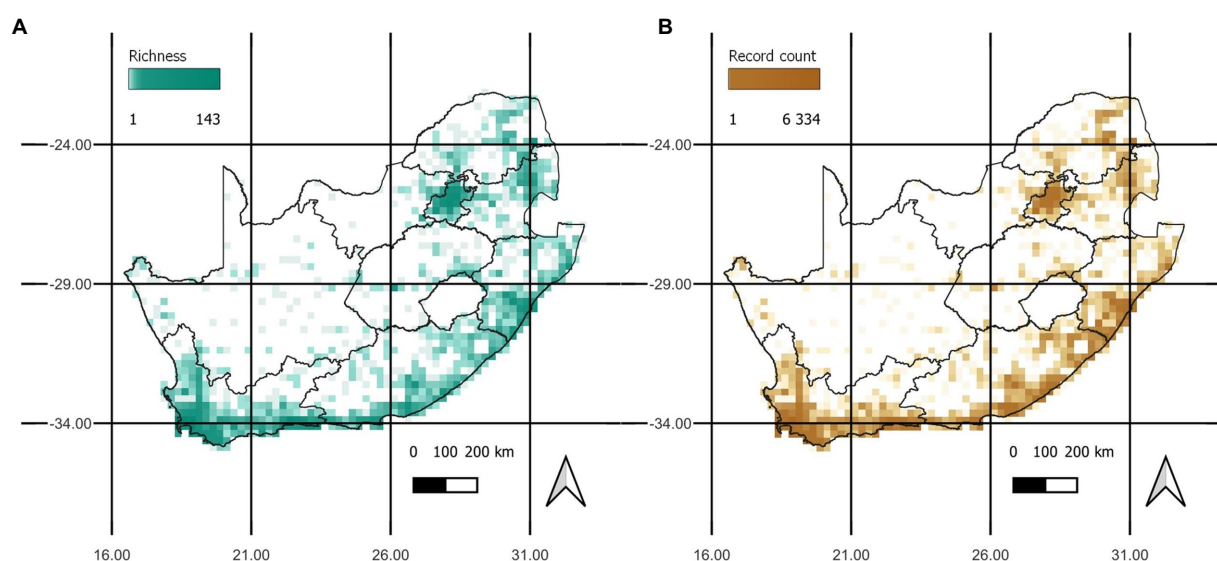
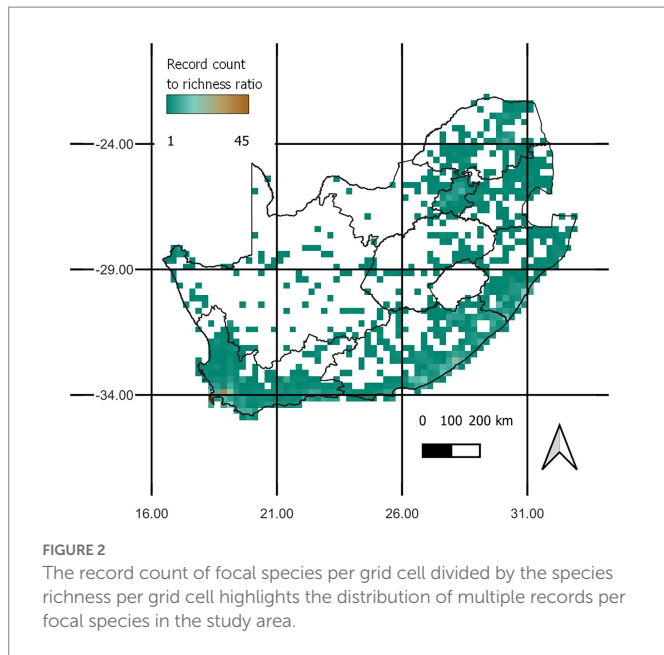


FIGURE 1

(A) The distribution of focal species richness, presented as the number of species per grid cell. (B) The distribution of record counts per grid cell, which relates to sampling effort for focal species. The similar patterns in the two plots highlight that the richness per cell may be biased by the sampling effort in those cells.



the K th percentile of the sampling effort values across cells where this species was detected. Finally, a cell was selected if the number of focal species whose absence is uncertain divided by the number of detected focal species in the cell was less than a threshold called *ratio*. After investigating the influence of changing the two parameters, K , and *ratio*, with differing levels of stringency, the values of $K = 10$, and *ratio* = 1 were decided and 68 grid cells selected in South Africa. These parameter values were selected to retain enough cells for subsequent analyses, whilst still limiting the false-absence errors. These selected cells were all distributed in highly sampled areas, i.e., around urban centres (Supplementary Figure S2).

2.2. Drivers of turnover

Plant distributions are shaped by environmental factors such as local bioclimates defined by precipitation and temperature variables (Mod et al., 2016). We included bioclimatic variables to test the hypothesis that prevailing environmental conditions drive the turnover of naturalised trees and shrubs. Given the limited number of selected cells and to avoid over-parameterization, the 19 bioclimatic variables at 30-s resolution from the WorldClim database (Fick and Hijmans, 2017) were reduced *via* a principal component analysis. Fine-grained raster datasets were cropped to South Africa and then resampled to quarter-degree resolution using the bilinear method in R statistical software. This dimension reduction makes the interpretation of contribution to turnover less explicit since each principal component axis now captures the variation of the linear combination of variables. The contribution of each axis to bioclimatic variation is indicated by the corresponding eigenvalues, whilst the contribution of each bioclimatic variable in an axis is the variable's weight in the linear combination. The first and third axes explained 35.8 and 14.8% of bioclimatic variation, respectively. The second axis was dropped due to its strong correlation with other included predictors (particularly, river density). According to bioclimatic variable weights of these two selected principal component axes (Supplementary Table S2), the first axis reflects mainly the effect of diurnal temperature range, and the third axis precipitation.

Road density is a measure of human activity in the Human Footprint Index (HFI) (Venter et al., 2016). To investigate the role of pathways created by all roads and trails in naturalised plant community assembly, the road density per quarter-degree cell was computed based on the complete vector shapefile of roads in South Africa (Humanitarian OpenStreetMap Team, 2022) in QGIS and included in the analyses.

River systems can affect community turnover *via* mediating propagule dispersal and storage. The river density per quarter-degree cell was included as a predictor to capture the contribution of rivers and streams to turnover. We computed river density using the line density interpolation function in QGIS software version 3.18.0-Zurich (QGIS Development Team, 2022) based on a complete vector shapefile of South African rivers and streams. All projections were according to standard WGS84 projection.

The iNaturalist interface allows its users to label species observations as cultivated. We can expect that species labelled as cultivated were planted in the area and that such occurrences are the result of current or historical horticultural practices in the area. To test the hypothesis that local horticultural activities are a driver of turnover of focal species, a planting effort metric was computed by dividing the number of cultivated iNaturalist records of all terrestrial plants, by the sum of focal species and cultivated observations per cell in order to control for sampling bias. Observations of planted specimens were extracted from the iNaturalist website on 28 July 2022 with a query specifying only cultivated plant records within South Africa.

Variable pairs were checked for collinearity with the accepted level indicated by Pearson's correlation coefficient below 0.7. Variance Inflation Factor (VIF) measures how much of the variance of a coefficient in the regression model is attributed to multicollinearity within the model, where a VIF below 10 is acceptable. This was tested using the *car* package v3.1-0 in R (Fox and Weisberg, 2019). The variable with the highest VIF above 10 was removed and the VIFs of all other variables recalculated and assessed, and this procedure was repeated till all the remaining variables had their VIFs below 10. Key variables for hypothesis testing were retained whilst checking collinearity and VIF (Araújo et al., 2019). Consequently, only six variables reflecting various mechanisms of turnover were included in the final analyses, including site-based factors (two bioclimatic axes, PC1 and PC3), dispersal pathways (road density, river density and geographical distance between quarter-degree cells) and human disturbance activity (road density and planting effort).

2.3. Statistical analyses

2.3.1. Zeta diversity

We aim to understand the processes mediating the assembly of species of alien trees and shrubs. The broadscale assembly processes can be investigated by analysing the behaviour of the zeta diversity metric with its order, called the zeta-decline. This can describe whether compositional turnover is governed by deterministic or stochastic processes. If deterministic processes such as niche differentiation shape the turnover in species composition, the curve of zeta-decline would necessarily be explained by a power-law regression function. If stochastic processes such as introduction history and pathway-based factors shape the turnover in species composition, the curve of zeta-decline would necessarily be described by a negative exponential regression function (Hui and McGeoch, 2014). The significance of these regressions was tested using the zeta decline function in the *zetadiv* package v1.2.1 in R

(Latombe et al., 2022). We acknowledge that this binary interpretation of zeta decline can be oversimplified (Deane et al., 2023) and thus included further regression analyses with predictors.

2.3.2. Species richness

The first order of zeta ζ_1 describes the average number of species in a single cell which is the species richness. The drivers of change in species richness across sites were investigated using a generalised additive model (GAM) with restricted maximum likelihood (REML) estimation. This method is similar to GDM models except that it considers the non-linear relationship between explanatory variables and species richness, as opposed to turnover of species. This was tested to understand the determinants of higher levels of biodiversity within sites.

2.3.3. Multi-site generalized dissimilarity modelling

We aim to understand the drivers of compositional turnover by relating variations of zeta diversity to environmental gradients. In generalised dissimilarity modelling (GDM), species turnover between sites are regressed with changes in selected predictor variables. Using zeta diversity as a metric of multi-site compositional turnover, the GDM model is adapted to Multi-Site Generalised Dissimilarity Modelling (MS-GDM). In this model, ζ_2 is equivalent to the GDM, quantifying pairwise site dissimilarity (beta diversity) driven largely by the gain and loss of narrow-ranged species between sites, whereas higher orders of zeta are equivalent to compositional turnover amongst more than two sites, hence driven largely by more widespread species. In this study, we selected the I-spline regression in the MS-GDM which captures the local and nonlinear response of compositional turnover to predictor gradients (Latombe et al., 2017). That is, a change in the value of a predictor variable, like temperature, does not affect the compositional turnover equally along the range of values. This is important since an equal amount of change at different ranges of a predictor can have very different ecologically meaningful effects. For example, increasing 1°C in cold regions has a different impact on compositional turnover when compared to increasing 1°C in warm regions. The I-splines were specified with three knots, thus allowing us to disentangle contribution to turnover throughout the low, medium, and high ranges of a predictor which corresponds to the three knots. The range of distances influencing turnover in this study is interpreted as change within a bioregion (short distance), change between adjacent bioregions (medium distance) and change between distant bioregions (long distance), where the longest distance would be between the south-western (around Cape Town) and the north-eastern parts of South Africa (around Mbombela). The MS-GDM for Simpson-equivalent zeta diversity was constructed for zeta orders of two and five to describe, respectively, the drivers of turnover of narrow-range (order two) and widespread (order five) naturalised species.

2.3.4. Mapping alien biomes

To predict species composition of alien trees and shrubs across South Africa, even in unsampled and under-sampled areas, we first trained a GDM with species compositions in the 68 selected cells and associated environments (whilst including geographical distance between cells). From this trained GDM model, the predicted compositional dissimilarity between site pairs was mapped using the predict function in the *gdm* package v1.5.0–3 in R (Fitzpatrick et al., 2022) by multidimensional scaling (MDS) and plotted as a Red-Green-Blue (RGB) plot of the principal component scores of sites. Alien biomes (bioregion clusters) of predicted compositional

dissimilarity were identified using the K-means algorithm, with the optimal number of clusters investigated using cluster visualisation methods (Supplementary Figures S3, S4). Choosing the optimal number of clusters is not definitive and we could equally well have chosen between five and seven clusters. Informed by the number of native biomes in South Africa, we chose to present six clusters to relate the communities of these alien species to native vegetation communities. The characteristic species of each cluster were identified as the 10 species with the highest count of presence per cluster (Supplementary Table S3).

3. Results

3.1. Zeta diversity

Communities of naturalised trees and shrubs appeared to be structured by deterministic rather than stochastic processes. Indeed, the shape of the zeta decline (Figure 3) for the selected 68 cells conforms best to a power-law regression function ($\zeta_n = e^{1.891} n^{-2.251}$ with AIC = −18.21) compared to an exponential decline regression function ($\zeta_n = e^{1.220 - 1.332n}$ with AIC = −9.35) for the first 20 orders of zeta. The zeta diversity declines to below 1 after order eight, which means that very few species overlap with more than eight grid cells. The average number of shared species between two sites is 14.45 ± 12.90 (mean \pm standard deviation), amongst three sites 6.81 ± 6.07 , amongst four sites 3.87 ± 3.70 , amongst five sites 2.49 ± 2.58 .

3.2. Species richness

The average number of species found in a single site of the 68 selected cells was 41.50 ± 30.52 (mean \pm standard deviation) (Figure 3). Species richness of selected cells was best explained by the planting effort and secondly by the diurnal temperature range (Table 1) (53.3%

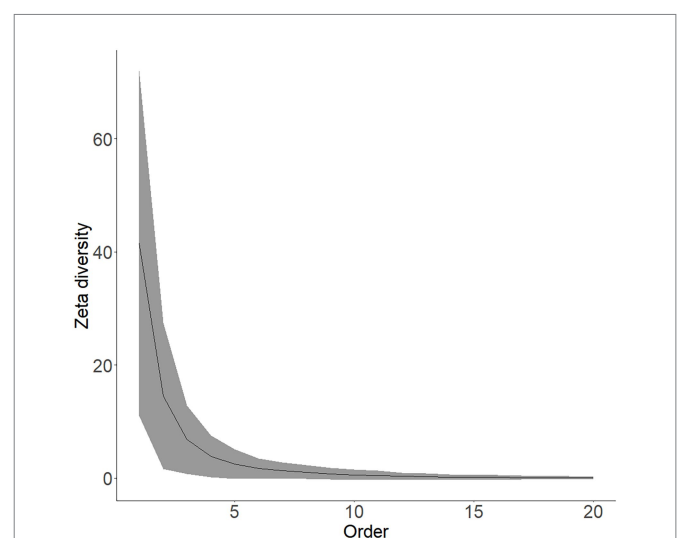


FIGURE 3

The relationship between zeta diversity and the order of zeta is shown as the zeta diversity decline. The zeta diversity decline was fitted to an exponential and a power-law regression function to test whether the assembly is driven largely by stochastic or deterministic factors. The solid line indicates mean zeta diversity and is bounded by the values of standard deviation.

TABLE 1 Results of GAM testing contribution of predictors to change in species richness of selected cells.

	<i>F</i>	<i>p</i> -value
Diurnal temperature range	6.764	0.012*
Precipitation	1.475	0.270
Road density	3.242	0.053
River density	2.318	0.108
Planting effort	14.369	<0.001***

Significance levels are indicated with * <0.05 , ** <0.01 , *** <0.001 .

TABLE 2 Results of multi-site generalised dissimilarity modelling of narrow-ranged species (ζ_2).

Predictor	Range	Estimate	<i>t</i> value	Pr(> <i>t</i>)
Diurnal temperature range	Low	0.000	0.000	1.000
	Medium	−0.079	−2.301	0.022*
	High	−0.115	−2.622	0.009**
Precipitation	Low	0.000	0.000	1.000
	Medium	0.000	0.000	1.000
	High	−0.102	−2.888	0.004**
Road density	Low	−0.096	−3.579	≤0.001***
	Medium	0.000	0.000	1.000
	High	0.000	0.000	1.000
River density	Low	−0.075	−3.080	0.002**
	Medium	−0.058	−1.858	0.063
	High	0.000	0.000	1.000
Distance	Low	−0.073	−2.173	0.030*
	Medium	−0.071	−1.660	0.097
	High	−0.341	−5.687	≤0.001***
Planting effort	Low	0.000	0.000	1.000
	Medium	0.000	0.000	1.000
	High	0.000	0.000	1.000

Significance levels are indicated with * ≤ 0.05 , ** ≤ 0.01 , *** ≤ 0.001 . Zeta similarity is the response variable, which is opposite to compositional dissimilarity or turnover; as such, estimates of regression coefficients in this constrained regression are all non-positive.

deviance explained by the GAM). There was a marginally non-significant effect of road density ($p = 0.053$) and no significant effect of precipitation and river density (Table 1).

3.3. Multi-site generalized dissimilarity modelling

Compositional similarity of narrow-ranged species at ζ_2 decayed mainly with increasing distance between sites (Table 2; note the significantly negative estimates), as indicated by the high magnitude of the distance I-spline (i.e., its contribution to compositional turnover) relative to those of other drivers (Figure 4A), and secondly diurnal temperature range and river density. It is important to note that the compositional similarity of zeta diversity is the opposite of the compositional dissimilarity (i.e., compositional turnover). As the compositional zeta similarity were fitted with three knots, we could understand contribution to turnover

across the low, medium, and high range of each predictor. Changes in diurnal temperature range of areas with larger difference between temperature extremes, and changes in precipitation in wet areas, not in arid zones, affected compositional turnover of narrow ranged species (Table 2). Changes in the road density at lower values of the predictor which corresponds to rural areas with few roads, and changes in river density lower values of the predictor which corresponds to areas with little drainage, were found to significantly affect compositional turnover of narrow-ranged species. Additionally, river density was marginally significant at medium ranges ($p < 0.1$). Compositional turnover of narrow-ranged species was sensitive to distance change across high (roughly more than 1,000 km) and low (roughly less than 500 km) ranges; note the significantly negative estimates in Table 2 and the notable steepness over these ranges for the distance I-spline in Figure 4A. Notice that it is marginally significant for medium ranges ($p < 0.1$). Planting effort had no significant impact on compositional turnover of narrow ranged species. The model performed well, with 24.1% deviance explained. When not accounting for the distance between sites, the variance explained decreases by half to 12.6% deviance explained and negligible changes in the contribution to compositional turnover of predictors, with no additional significant predictors.

Community turnover of increasingly widespread species at ζ_5 was driven by very similar drivers to those for narrow-ranged species (Figure 4B for compositional turnover, Table 3 for compositional similarity) although for these species changes at medium distance ranges significantly affected turnover, compared to low distance ranges for narrow-ranged species. Road density was the second most important predictor for widespread species, followed by river density, which was again marginally significant in the low range. Lower road density in more rural areas was a very strong influence of compositional turnover, even stronger than distance for this range of the predictor (Figure 4B). Again, planting effort had no significant contribution to turnover. The model had a slightly weaker performance, which is expected as the order of zeta increases, with 19.9% deviance explained.

3.4. Predicted compositional similarity

The GDM model explained reasonably well the pairwise compositional dissimilarity between the selected cells with 34.1% of the deviance explained by the predictors. The visualisation of the predicted compositional similarity *via* GDM shows areas of similar community composition with similar colours in the Red-Green-Blue space (Figure 5A). Results of the K-means cluster analysis shows the geographical distributions of the six clusters (Figure 5B). Many species are present in multiple clusters, indicating that there are many widespread species which homogenises alien communities across South Africa (see Table S3). These clusters are compared to the native biomes of South Africa (Figure 5C). The planting effort metric was not considered to predict dissimilarity across all pairs of cells as it cannot be computed for the many cells without focal species and planted species since this results in dividing by zero (Figure 5D). Including planting effort would have limited our capacity to predict across South Africa since planting effort could only be computed for 732 cells (Figure 5D). We consider it justified to remove planting effort from GDM modelling since it consistently made no significant contribution to turnover.

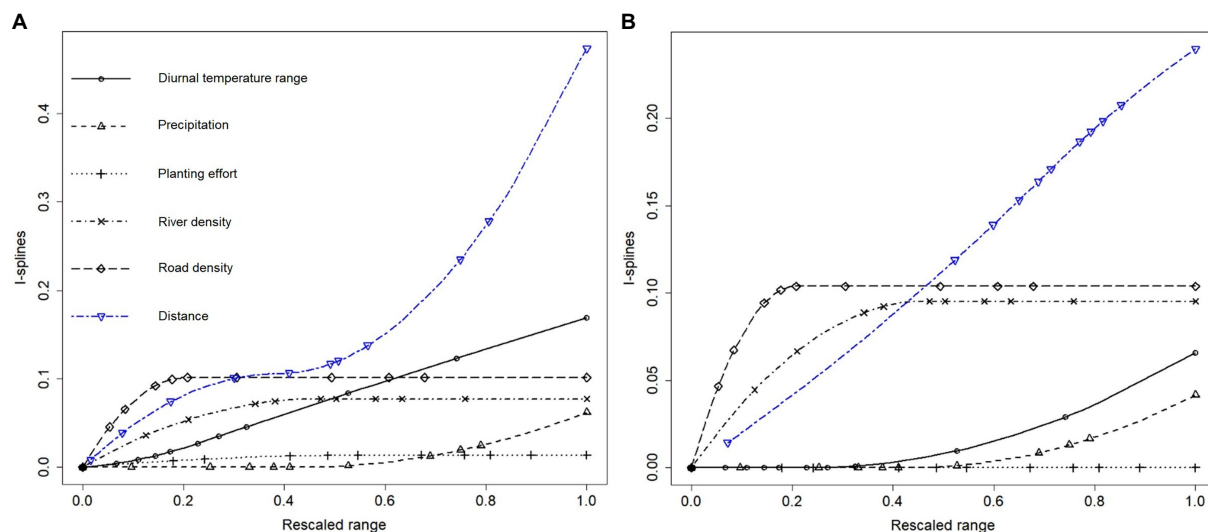


FIGURE 4

(A) I-spline regression of narrow-ranged species (zeta order two) and (B) widespread species (zeta order five) naturalised trees and shrubs, for which the magnitude of an I-spline is interpreted as the total contribution of the predictor to compositional turnover (i.e., the opposite to compositional zeta similarity). Following the convention of dissimilarity modelling, the vertical axis of I-splines represents compositional dissimilarity or turnover, which is opposite to zeta similarity. Notably, distance is consistently a strong predictor at all ranges, but with low road density being more important than distance for turnover of widespread species. The steepness of the response curve at a particular value for a predictor represents the contribution of the predictor to compositional turnover at this value. For instance, river density contributes more to compositional turnover in low-drainage areas (a steeper response curve for river density < 0.3) than in high-drainage areas (a flatter response curve for river density > 0.5). Each variable is rescaled to a value between minimum and maximum, with markers separating 10th percentiles. Distance is rescaled between zero and maximum and thus the minimum distance between sites (the leftmost marker) is above zero.

4. Discussion

The drivers of compositional turnover of narrow-ranged and widespread naturalised trees and shrubs in South Africa are largely similar (Table 2 and Table 3) and composition showed a strong response to geographic distance in both cases. This study also revealed for the first time the importance of pathway-based factors in shaping the compositional turnover of naturalised woody plant communities in South Africa. This highlights the central role of introduction history and dispersal constraints in shaping the distribution of these species, along with the documented role of bioclimatic factors (Hugo et al., 2012; Rouget et al., 2015). However, narrow-ranged and widespread species compositions differ in their sensitivity to the range of distance. The difference in their response to distance between sites merits further investigation. When species are not limited by distance, and with enough residence time, they are expected to be distributed across all biotically and abiotically suitable environments. In reality, the dispersal capabilities of organisms are limited, and the physical distance between sites poses a limit to movement. Physical barriers such as mountain ranges, rivers, and regions of unsuitable habitat through which they cannot traverse further modify the distribution of species (i.e., the connectivity between sites; Vasudev et al., 2015). However, this distance predictor may also serve as a surrogate for environmental variables which are autocorrelated with distance. This is known as the Moran effect, where populations geographically close to each other, and which experience more similar environmental variables in space and time, have similar population dynamics (Hansen et al., 2020). Thus, it may be that a change in distance is synonymous with changing some environmental factors not considered in this study because of the limited number of well-sampled sites available. The relationship between

TABLE 3 Results of multi-site generalised dissimilarity modelling of widespread species (ζ_s).

Predictor	Range	Estimate	t value	Pr(> t)
Diurnal temperature range	Low	0.000	0.000	1.000
	Medium	−0.039	−1.292	0.196
	High	−0.075	−2.791	0.005**
Precipitation	Low	−0.008	−0.542	0.588
	Medium	0.000	0.000	1.000
	High	−0.058	−2.810	0.005**
Road density	Low	−0.116	−7.047	≤0.001***
	Medium	0.000	0.000	1.000
	High	0.000	0.000	1.000
River density	Low	−0.093	−5.639	≤0.001***
	Medium	0.000	0.000	1.000
	High	0.000	0.000	1.000
Distance	Low	−0.079	−1.767	0.078
	Medium	−0.107	−4.066	≤0.001***
	High	−0.013	−0.442	0.659
Planting effort	Low	0.000	0.000	1.000
	Medium	0.000	0.000	1.000
	High	0.000	0.000	1.000

Significance levels are indicated with *≤0.05, **≤0.01, ***≤0.001. Zeta similarity is the response variable, which is opposite to compositional dissimilarity or turnover; as such, estimates of regression coefficients in this constrained regression are all non-positive.

distance and community turnover is likely further modulated by the long-distance or haphazard movement and introduction of propagules

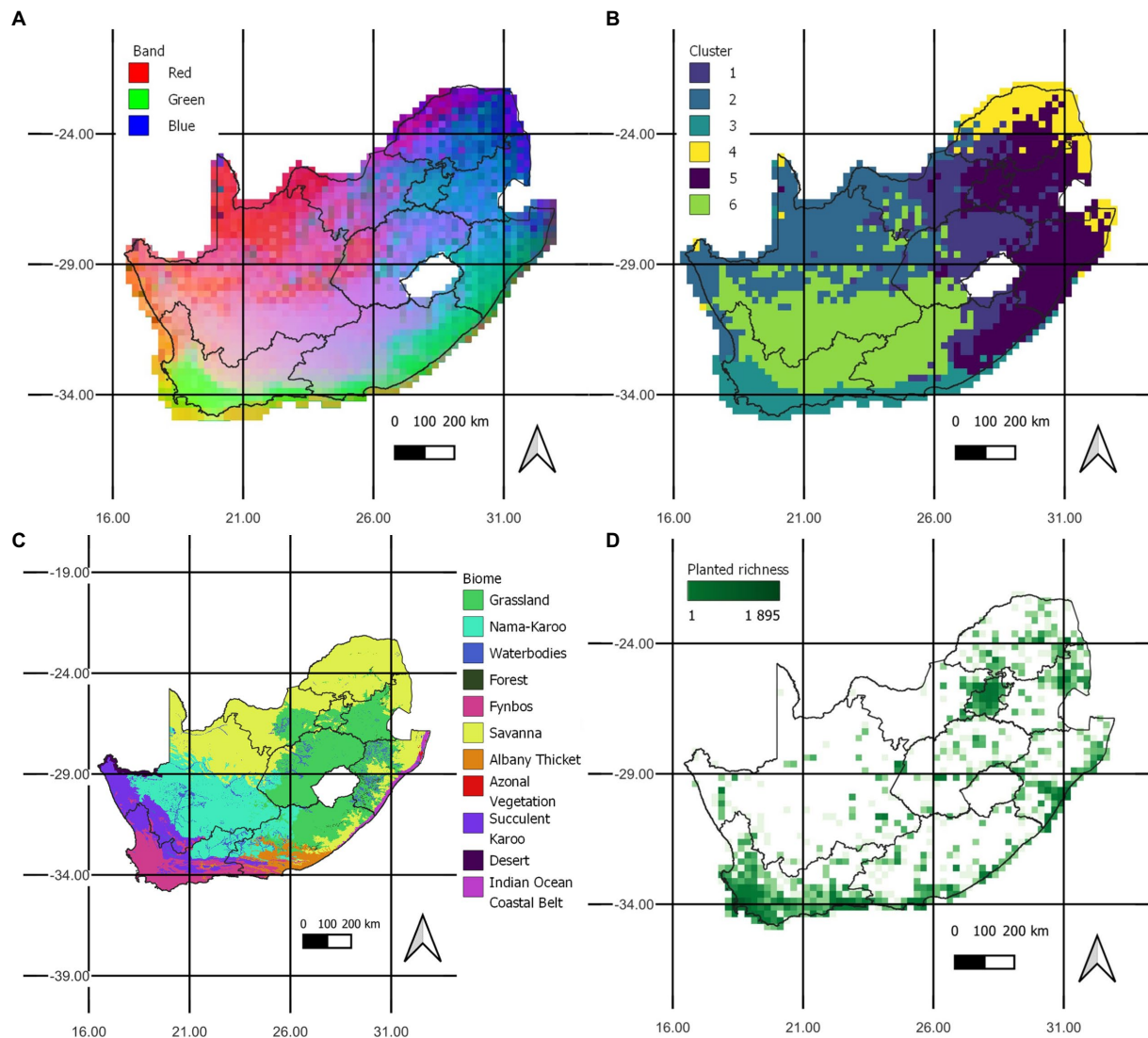


FIGURE 5

(A) The predicted compositional similarity of alien biomes of trees and shrubs in the RGB space and (B) the predicted geographies of communities of trees and shrubs as results from a cluster analysis of k-means when specifying $k=6$ compared to (C) the native biomes of South Africa (South African National Biodiversity Institute (SANBI), 2012) and (D) the coverage of record counts of planting effort.

and organisms through diverse human activities (Auffret et al., 2014). Thus, geographical distance in this study may be representative of not just a limit to the natural dispersal abilities of species, but also the effect of untested autocorrelated variables and the influence of humans which are known to modulate alien species movements.

There is high turnover within bioregions for narrow-ranged species, for which the communities also differ between distant bioregions. There is little evidence for dispersal over short distances since they remain unhomogenised even over very short ranges; this may reflect the influence of competitive exclusion between species. Propagules rarely disperse *via* natural long-distance dispersal and rather a large proportion of dispersal events occur at the local scale. Thus, the distinction between distant communities more likely reflects the influence of human movement or introduction of propagules to these sites (Nathan et al., 2008). These narrow-ranged species thus have a patchy distribution across South Africa, which may reflect multiple introduction hubs, which may also differ in the species they harbour. This may be due to a

difference in the history of ornamental plantings in different parts of the country, different agricultural crops and ornamental species, and different forestry species, a view that is supported by the findings of Thuiller et al. (2006). This indicates a close association of human use of alien plants and their distribution in South Africa. This interpretation requires further research into the regional scale, current and historical trends in the use of alien trees and shrubs by humans.

In contrast, widespread species showed high turnover across medium distances, indicating little homogenisation at these ranges and thus low rates of natural dispersal by species between adjacent bioregions, where only some species successively spread far from introduction hubs. The high density of invasions in urban centres may reflect communities outside the radius of major introduction hubs which differ from those within the introduction hub. This could thus indicate a concentration of widespread species at sites of introduction, with high levels of homogenisation within these sites, and high turnover of species at the outskirts, possibly from a handful of particularly

successful invaders having spread to this distance. These species have low turnover at local scales which indicates homogenisation of communities within regions, possibly reflecting frequent within-region dispersal. This is likely the effect of a few very widely distributed species having spread substantially across the country, or those which have been introduced to multiple distant locations. If this distance reflects the Moran effect, we can understand that these very widespread species are generalist invaders that occur widely across many habitats and climatic conditions. This complex effect of distance on turnover could be further elucidated by studies of network topology and connectivity between sites (e.g., Banks et al. (2015) that may reflect contributions to propagule movement from multiple human-mediated vectors.

Passive dispersal of propagules *via* roads likely only acts over short distances. Parendes and Jones (2000) indicate that dispersal pathways such as rivers and roads may provide a more complex contribution to turnover than simply through dispersal of propagules. Rauschert et al. (2017) found evidence for dispersal of seeds across a maximum distance of a few hundred metres on rural roads, and this *via* the action of local road maintenance. However, there is strong evidence that these pathways contribute to turnover through more complex mechanisms as they have been shown to simultaneously provide suitable habitat for species, act as a store of propagules, and aid dispersal *via* physical propagule movement. Kalwij et al. (2008) found little role for dispersal along the road verges, but rather that the nature of the road explained turnover. This was divided into rural gravel roads and tar roads, with high and low road traffic. Areas with higher road density had higher species richness and the distance of these roads from urban centres did not predict species richness. In a survey along South African roads, Milton and Dean (1998) found high alien species richness on road verges across all biomes. Higher species richness was found to be closely associated with cultivated areas, rather than rangelands. Thus, the high turnover of species in regions of low road density found in this study may be attributed to short-distance dispersal of some species for which seeds are able to disperse along the road verges. The effect of turnover at relatively rural sites is more likely associated with the habitat features provided by the road verges, and possibly untested land-use practices at these sites.

An interesting observation is that species in areas with high road densities, such as urban centres, are more homogenised. This means that urban invaders are interspersed with little turnover, possibly reflecting suites of invaders widespread throughout urban areas. These areas likely contain the highest species richness, the planting effort per grid cell, which is associated with urban areas and human cultivation, was a strong contributor to richness. This indicates that urban areas with higher planting effort facilitate higher species richness of wild-growing trees and shrubs, possibly as a consequence of the nurturing of parent populations (Donaldson et al., 2014). Thus, although the planting effort does not increase spatial turnover between sites of wild growing naturalised trees and shrubs, it still influences the presence of these species. This may be as a possible consequence of the urban landscape with high road densities, rather than a direct consequence of cultivation, although these factors are tightly linked. In contrast, bioclimatic factors are likely sorting the communities of wild-growing species.

Rivers likely contribute in a complex way to turnover, *via* both propagule dispersal and storage. For example, Foxcroft et al. (2007) showed the possibility of spread of riparian invaders along river corridors. On the other hand, Gurnell et al. (2008) found that a large proportion of propagules deposited in the riverbed of a UK river were of species present much higher up along the river. The inundation cycle

of the river allows these propagules to be moved out of the riparian zone to colonise new sites. However, this depends on the physical nature of propagules which determines their deposition and germination. The alternative would be that the river systems act as a barrier to dispersal which has been identified in old river systems in Amazonia (Dambros et al., 2020). In this case, a higher river density would equal higher compositional turnover because of limitations to dispersal between sites across rivers. However, this is not the case, and rather we see that the river systems increase dispersal by homogenising compositions between sites in areas with high river density. In this study, river corridors likely act to disperse propagules of both narrow ranged and widespread species over significant distances. This is also true in more arid regions where the river may act as refugia for these species, thus harbouring and homogenising communities rather than limiting dispersal.

We see high turnover in areas which experience high temperature variation and precipitation, for both narrow-ranged and widespread species. High-altitude inland areas which experience greater temperature extremes have high compositional turnover, compared to moderate coastal plains. Wetter areas such as in the eastern half of the country, along with the southern coast and near the Cape Peninsula, have high species turnover, with little turnover in more arid regions. There is thus little evidence of unique environmental drivers of species turnover of comparatively less progressed alien taxa, versus more progressed widespread taxa. The finding that narrow-ranged and widespread species turnover are influenced by the same factors sheds light on the determinants of invasion by these species. This suggests that range expansion of invaders is ongoing and not characterised by any particular predictors tested in this study. However, this study was limited to few predictors and future studies should incorporate biotic factors to fully investigate the determinants of range expansion of these species.

The geographical distribution of naturalised tree and shrub communities has little resemblance to native biomes. The strong effect of introduction hubs and road density as a driver of turnover indicates that there is an association with areas of urbanisation or human activity and use of alien species. This is supported by the finding that species richness is higher in areas of high planting effort which are around urban areas. The influence of diurnal temperature ranges is seen to explain clusters two and five along the coast in areas that experience moderate temperatures and relatively high precipitation. Further inland, clusters are separated between the east and west which corresponds to the influence of precipitation gradients. This seems to be further modulated by different road densities in the sparsely populated Northern Cape province in the north-west which is represented roughly by clusters three and six, and the densely populated province of Gauteng in the north-east, which is mainly contained in cluster five.

5. Conclusion

Similar to the results of Hugo et al. (2012) and Rouget et al. (2015), our study provided evidence for the role of bioclimatic factors in shaping alien biomes. New evidence emerged of the role of human introduction and movement of species, and both natural and human-made pathways in aiding dispersal and providing habitat. However, the limitations of the presence-only dataset meant that not all possible drivers could be tested. This study did not include biotic factors in the analysis, nor land-use

practices. Further work, including these factors, is needed to further clarify the dynamics of species assemblages of naturalised trees and shrubs in South Africa. Turnover of narrow-ranged and widespread species are maintained by similar drivers and there is little evidence for unique drivers which determine range expansion of naturalised trees and shrubs. The geographical clustering of species indicates that mainly roads, rivers and to some extent distance explains within-bioregion compositional turnover; these factors are related to dispersal pathways. Between bioregions the effect of bioclimatic factors and introduction hubs with distinct species drive compositional turnover. Further research could apply this to the entire suite of alien taxa present in South Africa to see whether results of this larger grouping would mimic more closely the results of Hugo et al. (2012) and Rouget et al. (2015) for the geographical distribution of alien plant biomes. Overall, this investigation has showed evidence for the role of human activity in the assembly processes of naturalised trees and shrubs in South Africa, with little evidence found for environmental drivers which may assist the range expansion of alien taxa. Further studies are needed to compare the results of analysis of data from this platform to results from professionally curated sources to further our knowledge on the usability of this dataset.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

References

- Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early, R., et al. (2019). Standards for distribution models in biodiversity assessments. *Sci. Adv.* 5, eaat4858–eaat4874. doi: 10.1126/sciadv.aat4858
- Auffret, A. G., Berg, J., and Cousins, S. A. O. (2014). The geography of human-mediated dispersal. *Divers. Distrib.* 20, 1450–1456. doi: 10.1111/ddi.12251
- Banks, N. C., Paine, D. R., Bayliss, K. L., and Hodda, M. (2015). The role of global trade and transport network topology in the human-mediated dispersal of alien species. *Ecol. Lett.* 18, 188–199. doi: 10.1111/ele.12397
- Basel, A. M., Simaika, J. P., Samways, M. J., Midgley, G. F., MacFadyen, S., and Hui, C. (2021). Assemblage reorganization of south African dragonflies due to climate change. *Divers. Distrib.* 27, 2542–2558. doi: 10.1111/ddi.13422
- Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., et al. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biol. Conserv.* 173, 144–154. doi: 10.1016/j.biocon.2013.07.037
- Blackburn, T. M., Pyšek, P., Bacher, S., Carlton, J. T., Duncan, R. P., Jarošík, V., et al. (2011). A proposed unified framework for biological invasions. *Trends Ecol. Evol.* 26, 333–339. doi: 10.1016/j.tree.2011.03.023
- Botella, C., Bonnet, P., Hui, C., Joly, A., and Richardson, D. M. (2022). Dynamic species distribution modeling reveals the pivotal role of human-mediated long-distance dispersal in plant invasion. *Biol.* 11:1293. doi: 10.3390/biology11091293
- Botella, C., Joly, A., Monestiez, P., Bonnet, P., and Munoz, F. (2020). Bias in presence-only niche models related to sampling effort and species niches: lessons for background point selection. *PLoS One* 15:e0232078. doi: 10.1371/journal.pone.0232078
- Cheek, M. D., and Semple, J. C. (2016). First official record of naturalised populations of *Solidago altissima* L. var. *pluricephala* M.C. Johnst. (Asteraceae: Astereae) in Africa. *S. Afr. J. Bot.* 105, 333–336. doi: 10.1016/j.sajb.2016.05.001
- Christen, D. C., and Matlack, G. R. (2009). The habitat and conduit functions of roads in the spread of three invasive plant species. *Biol. Invasions* 11, 453–465. doi: 10.1007/s10530-008-9262-x
- Dambros, C., Zuquim, G., Moullet, G. M., Costa, F. R. C., Tuomisto, H., Ribas, C. C., et al. (2020). The role of environmental filtering, geographic distance and dispersal barriers in shaping the turnover of plant and animal species in Amazonia. *Biodivers. Conserv.* 29, 3609–3634. doi: 10.1007/s10531-020-02040-3
- Dart, K., Latty, T., and Greenville, A. (2022). Citizen science reveals current distribution, predicted habitat suitability and resource requirements of the introduced African carder bee *Pseudanthidium (Immanthidium) repetitum* in Australia. *Biol. Invasions* 24, 1827–1838. doi: 10.1007/s10530-022-02753-2
- Deane, D., Hui, C., and McGeoch, M. A. (2023). Two dominant forms of multisite similarity decline – their origins and interpretation. *Ecol. Evol.* in press
- Dickinson, J. L., Zuckerberg, B., and Bonter, D. N. (2010). Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Syst.* 41, 149–172. doi: 10.1146/annurev-ecolsys-102209-144636
- Donaldson, J. E., Hui, C., Richardson, D. M., Robertson, M. P., Webber, B. L., and Wilson, J. R. U. (2014). Invasion trajectory of alien trees: the role of introduction pathway and planting history. *Glob. Chang. Biol.* 20, 1527–1537. doi: 10.1111/gcb.12486
- Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Divers. Distrib.* 13, 252–264. doi: 10.1111/j.1472-4642.2007.00341.x
- Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: new 1km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi: 10.1002/joc.5086
- Fitzpatrick, M., Mokany, K., Manion, G., Nieto-Lugilde, D., and Ferrier, S. (2022). gdm: generalized dissimilarity modeling. Available at: <https://CRAN.R-project.org/package=gdm> (Accessed October 24, 2022).

Funding

We thank the DSI-NRF Centre of Excellence for support. DMR acknowledges support from Mobility 2020 project no. CZ.02.2.69/0.0/0.0/18_053/0017850 (Ministry of Education, Youth, and Sports of the Czech Republic) and long-term research development project RVO 67985939 (Czech Academy of Sciences). CH and CB acknowledge support from the NRF (grant 89967). We are grateful to Guillaume Latombe for useful discussions on the methodology. We thank the Department of Botany and Zoology at Stellenbosch University for their support during the research process.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2023.1106197/full#supplementary-material>

- Fox, J., and Weisberg, S. (2019). An {R} companion to applied regression. Available at: <https://socialsciences.mcmaster.ca/jfox/Books/Companion/> (Accessed October 18, 2022).
- Foxcroft, L. C., Rouget, M., and Richardson, D. M. (2007). Risk assessment of riparian plant invasions into protected areas. *Conserv. Biol.* 21, 412–421. doi: 10.1111/j.1523-1739.2007.00673.x
- GBIF.org (2022). Gbif Occurrence Download. doi: 10.15468/dl.vrsyxr
- Gurnell, A., Thompson, K., Goodson, J., and Moggridge, H. (2008). Propagule deposition along river margins: linking hydrology and ecology. *J. Ecol.* 96, 553–565. doi: 10.1111/j.1365-2745.2008.01358.x
- Hansen, B. B., Grøtan, V., Herfindal, I., and Lee, A. M. (2020). The Moran effect revisited: spatial population synchrony under global warming. *Ecography* 43, 1591–1602. doi: 10.1111/ecog.04962
- Henderson, L. (2020). *Invasive Alien Plants in South Africa*. Pretoria: Agricultural Research Council.
- Henderson, L., and Wilson, J. R. U. (2017). Changes in the composition and distribution of alien plants in South Africa: an update from the southern African plant invaders atlas. *Bothalia Afr. Biodivers. Conserv.* 47, 1–26. doi: 10.4102/abc.v47i2.2172
- Hochmair, H. H., Scheffrahn, R. H., Basille, M., and Boone, M. (2020). Evaluating the data quality of iNaturalist termite records. *PLoS One* 15:e0226534. doi: 10.1371/journal.pone.0226534
- Hugo, S., van Rensburg, B. J., van Wyk, A. E., and Steenkamp, Y. (2012). Alien phytoecogeographic regions of southern Africa: numerical classification, possible drivers, and regional threats. *PLoS One* 7:e36269. doi: 10.1371/journal.pone.0036269
- Hui, C., and McGeoch, M. A. (2014). Zeta diversity as a concept and metric that unifies incidence-based biodiversity patterns. *Am. Nat.* 184, 684–694. doi: 10.1086/678125
- Hui, C., Richardson, D. M., Pyšek, P., Le Roux, J. J., Kučera, T., and Jarošík, V. (2013). Increasing functional modularity with residence time in the co-distribution of native and introduced vascular plants. *Nat. Commun.* 4:2454. doi: 10.1038/ncomms3454
- Humanitarian OpenStreetMap Team (2022). HOTOSM South Africa Roads (OpenStreetMap Export). Available at: <https://data.humdata.org/> (Accessed March 18, 2022).
- iNaturalist (2022). iNaturalist. Available at: <https://www.inaturalist.org> (Accessed March 18, 2022).
- Jansen, C., and Kumschick, S. (2022). A global impact assessment of acacia species introduced to South Africa. *Biol. Invasions* 24, 175–187. doi: 10.1007/s10530-021-02642-0
- Johnston, A., Fink, D., Hochachka, W. M., and Kelling, S. (2018). Estimates of observer expertise improve species distributions from citizen science data. *Methods Ecol. Evol.* 9, 88–97. doi: 10.1111/2041-210X.12838
- Kalwij, J. M., Milton, S. J., and McGeoch, M. A. (2008). Road verges as invasion corridors? A spatial hierarchical test in an arid ecosystem. *Landsc. Ecol.* 23, 439–451. doi: 10.1007/s10980-008-9201-3
- Kalwij, J. M., Steyn, C., and le Roux, P. C. (2014). Repeated monitoring as an effective early detection means: first records of naturalised *Solidago gigantea* Aiton (Asteraceae) in southern Africa. *S. Afr. J. Bot.* 93, 204–206. doi: 10.1016/j.sajb.2014.04.013
- Krasnov, B. R., Shenbrot, G. I., van der Mescht, L., and Khokhlova, I. S. (2020). Drivers of compositional turnover are related to species' commonness in flea assemblages from four biogeographic realms: zeta diversity and multi-site generalised dissimilarity modelling. *Int. J. Parasitol.* 50, 331–344. doi: 10.1016/j.ijpara.2020.03.001
- Latombe, G., Hui, C., and McGeoch, M. A. (2017). Multi-site generalised dissimilarity modelling: using zeta diversity to differentiate drivers of turnover in rare and widespread species. *Methods Ecol. Evol.* 8, 431–442. doi: 10.1111/2041-210X.12756
- Latombe, G., McGeoch, M., Nipperess, D., and Hui, C. (2022). Zetadiv: Functions to compute compositional turnover using zeta diversity. R package version 1.2.1. Available at: <https://CRAN.R-project.org/package=zetadiv>.
- Latombe, G., Roura-Pascual, N., and Hui, C. (2019). Similar compositional turnover but distinct insular environmental and geographical drivers of native and exotic ants in two oceans. *J. Biogeogr.* 46, 2299–2310. doi: 10.1111/jbi.13671
- Lenzner, B., Latombe, G., Schertler, A., Seebens, H., Yang, Q., Winter, M., et al. (2022). Naturalized alien floras still carry the legacy of European colonialism. *Nat. Ecol. Evol.* 6, 1723–1732. doi: 10.1038/s41559-022-01865-1
- MacFadyen, S., Allsopp, N., Altwegg, R., Archibald, S., Botha, J., Bradshaw, K., et al. (2022). Drowning in data, thirsty for information and starved for understanding: a biodiversity information hub for cooperative environmental monitoring in South Africa. *Biol. Conserv.* 274:109736. doi: 10.1016/j.biocon.2022.109736
- McGeoch, M. A., Latombe, G., Andrew, N. R., Nakagawa, S., Nipperess, D. A., Roigé, M., et al. (2019). Measuring continuous compositional change using decline and decay in zeta diversity. *Ecology* 100:e02832. doi: 10.1002/ecy.2832
- Merritt, D. M., Nilsson, C., and Jansson, R. (2010). Consequences of propagule dispersal and river fragmentation for riparian plant community diversity and turnover. *Ecol. Monogr.* 80, 609–626. doi: 10.1890/009-1533.1
- Mesaglio, T., and Callaghan, C. T. (2021). An overview of the history, current contributions and future outlook of iNaturalist in Australia. *Wildl. Res.* 48, 289–303. doi: 10.1071/WR20154
- Milton, S. J., and Dean, W. R. J. (1998). Alien plant assemblages near roads in arid and semi-arid South Africa. *Divers. Distrib.* 4, 175–187. doi: 10.1046/j.1472-4642.1998.00024.x
- Mod, H. K., Scherrer, D., Luoto, M., and Guisan, A. (2016). What we use is not what we know: environmental predictors in plant distribution models. *J. Veg. Sci.* 27, 1308–1322. doi: 10.1111/jvs.12444
- Mokany, K., Ware, C., Woolley, S. N. C., Ferrier, S., and Fitzpatrick, M. C. (2022). A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment. *Glob. Ecol. Biogeogr.* 31, 802–821. doi: 10.1111/geb.13459
- Nathan, R., Schurr, F. M., Spiegel, O., Steinitz, O., Trakhtenbrot, A., and Tsoar, A. (2008). Mechanisms of long-distance seed dispersal. *Trends Ecol. Evol.* 23, 638–647. doi: 10.1016/j.tree.2008.08.003
- Parendes, L. A., and Jones, J. A. (2000). Role of light availability and dispersal in exotic plant invasion along roads and streams in the H.J. Andrews experimental Forest. *Oreg. Conserv. Biol.* 14, 64–75. doi: 10.1046/j.1523-1739.2000.99089.x
- Phillips, S. J., Dudik, M., Dudik, D., Elith, J., Graham, C. H., Lehmann, A., et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197. doi: 10.1890/07-2153.1
- QGIS Development Team (2022). QGIS geographic information system. Available at: <http://qgis.osgeo.org>.
- R Core Team (2021). R: A language and environment for statistical computing. Available at: <https://www.R-project.org/>.
- Rauschert, E. S. J., Mortensen, D. A., and Bloser, S. M. (2017). Human-mediated dispersal via rural road maintenance can move invasive propagules. *Biol. Invasions* 19, 2047–2058. doi: 10.1007/s10530-017-1416-2
- Richardson, D. M., Foxcroft, L. C., Latombe, G., Le Maitre, D. C., Rouget, M., and Wilson, J. R. (2020). “The biogeography of south African terrestrial plant invasions,” in *Biological Invasions in South Africa*, eds B. W. van Wilgen, J. Measey, D. M. Richardson, J. R. Wilson and T. A. Zengeya. Cham: Springer.
- Richardson, D. M., Hui, C., Nuñez, M. A., and Pauchard, A. (2014). Tree invasions: patterns, processes, challenges and opportunities. *Biol. Invasions* 16, 473–481. doi: 10.1007/s10530-013-0606-9
- Richardson, D. M., and Pyšek, P. (2012). Naturalization of introduced plants: ecological drivers of biogeographical patterns. *New Phytol.* 196, 383–396. doi: 10.1111/j.1469-8137.2012.04292.x
- Richardson, D. M., Pyšek, P., Rejmánek, M., Barbour, M. G., Panetta, F. D., and West, C. J. (2000). Naturalization and invasion of alien plants: concepts and definitions. *Divers. Distrib.* 6, 93–107. doi: 10.1046/j.1472-4642.2000.00083.x
- Richardson, D. M., and Rejmánek, M. (2011). Trees and shrubs as invasive alien species - a global review. *Divers. Distrib.* 17, 788–809. doi: 10.1111/j.1472-4642.2011.00782.x
- Rouget, M., Hui, C., Renteria, J., Richardson, D. M., and Wilson, J. R. U. (2015). Plant invasions as a biogeographical assay: vegetation biomes constrain the distribution of invasive alien species assemblages. *S. Afr. J. Bot.* 101, 24–31. doi: 10.1016/j.sajb.2015.04.009
- Rutherford, M. C., Mucina, L., and Powrie, L. W. (2006). “Biomes and bioregions of South Africa” in *Vegetation of South Africa, Lesotho and Swaziland*. eds M. C. Rutherford and L. Mucina (Pretoria: South African National Biodiversity Institute), 32–51.
- Simons, A. L., Mazor, R., Stein, E. D., and Nuzhdin, S. (2019). Using alpha, beta, and zeta diversity in describing the health of stream-based benthic macroinvertebrate communities. *Ecol. Appl.* 29:e01896. doi: 10.1002/eap.1896
- Skultety, D., and Matthews, J. W. (2017). Urbanization and roads drive non-native plant invasion in the Chicago metropolitan region. *Biol. Invasions* 19, 2553–2566. doi: 10.1007/s10530-017-1464-7
- South African National Biodiversity Institute (SANBI) (2012). *Vegetation Map of South Africa, Lesotho and Swaziland*. Pretoria: SANBI. [vector geospatial dataset]. Available at: <http://bgis.sanbi.org/SpatialDataset/Detail/18> (Accessed April 26, 2022).
- Thuiller, W., Richardson, D. M., Rouget, M., Procheş, Ş., and Wilson, J. R. U. (2006). Interactions between environment, species traits, and human uses describe patterns of plant invasions. *Ecology* 87, 1755–1769. doi: 10.1890/0012-9658(2006)87[1755:LBEST]2.0.CO;2
- van Kleunen, M., Essl, F., Pergl, J., Brundu, G., Carboni, M., Dullinger, S., et al. (2018). The changing role of ornamental horticulture in alien plant invasions. *Biol. Rev.* 93, 1421–1437. doi: 10.1111/brv.12402
- van Wilgen, B. W., Zengeya, T. A., and Richardson, D. M. (2022). A review of the impacts of biological invasions in South Africa. *Biol. Invasions* 24, 27–50. doi: 10.1007/s10530-021-02623-3
- Vasudev, D., Fletcher, R. J., Goswami, V. R., and Krishnadas, M. (2015). From dispersal constraints to landscape connectivity: lessons from species distribution modelling. *Ecography* 38, 967–978. doi: 10.1111/ecog.01306
- Venter, O., Sanderson, E. W., Magrath, A., Allan, J. R., Beher, J., Jones, K. R., et al. (2016). Global terrestrial human footprint maps for 1993 and 2009. *Sci. Data* 3:160067. doi: 10.1038/sdata.2016.67
- Vilela, B., and Villalobos, F. (2015). letsR: a new R package for data handling and analysis in macroecology. *Methods Ecol. Evol.* 6, 1229–1234. doi: 10.1111/2041-210X.12401
- Watt, M. S., Kriticos, D. J., and Manning, L. K. (2009). The current and future potential distribution of *Melaleuca quinquenervia*. *Weed Res.* 49, 381–390. doi: 10.1111/j.1365-3180.2009.00704.x
- Zengeya, T. A., and Wilson, J. R. (eds.). (2020). *The Status of Biological Invasions and their Management in South Africa in 2019*. Stellenbosch: South African National Biodiversity Institute, Kirstenbosch and DSI-NRF Centre of Excellence for Invasion Biology.



OPEN ACCESS

EDITED BY

Sawaid Abbas,
University of the Punjab, Pakistan

REVIEWED BY

Rakesh Bhutiani,
Gurukul Kangri University, India
Michael Lorenzo Casazza,
United States Geological Survey (USGS),
United States
Kate Brandis,
University of New South Wales, Australia
Muhammad Usman,
University of the Punjab, Pakistan

*CORRESPONDENCE

Francisco Cervantes
✉ f.cervantesperalta@gmail.com

SPECIALTY SECTION

This article was submitted to
Environmental Informatics and Remote
Sensing,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 24 December 2022

ACCEPTED 21 February 2023

PUBLISHED 10 March 2023

CITATION

Cervantes F, Altwegg R, Strobbe F, Skowno A,
Visser V, Brooks M, Stojanov Y, Harebottle DM
and Job N (2023) BIRDIE: A data pipeline to
inform wetland and waterbird conservation at
multiple scales. *Front. Ecol. Evol.* 11:1131120.
doi: 10.3389/fevo.2023.1131120

COPYRIGHT

© 2023 Cervantes, Altwegg, Strobbe, Skowno,
Visser, Brooks, Stojanov, Harebottle and Job.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

BIRDIE: A data pipeline to inform wetland and waterbird conservation at multiple scales

Francisco Cervantes^{1,2*}, Res Altwegg¹, Francis Strobbe³,
Andrew Skowno², Vernon Visser¹, Michael Brooks⁴,
Yvan Stojanov³, Douglas M. Harebottle⁵ and Nancy Job²

¹Department of Statistical Sciences, Centre for Statistics in Ecology, The Environment and Conservation, University of Cape Town, Cape Town, South Africa, ²South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa, ³Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, Brussels, Belgium, ⁴Department of Biological Sciences, FitzPatrick Institute of African Ornithology, University of Cape Town, Cape Town, South Africa, ⁵School of Natural and Applied Sciences, Risk and Vulnerability Science Centre, Sol Plaatje University, Kimberley, South Africa

Introduction: Efforts to collect ecological data have intensified over the last decade. This is especially true for freshwater habitats, which are among the most impacted by human activity and yet lagging behind in terms of data availability. Now, to support conservation programmes and management decisions, these data need to be analyzed and interpreted; a process that can be complex and time consuming. The South African Biodiversity Data Pipeline for Wetlands and Waterbirds (BIRDIE) aims to help fast and efficient information uptake, bridging the gap between raw ecological datasets and the information final users need.

Methods: BIRDIE is a full data pipeline that takes up raw data, and estimates indicators related to waterbird populations, while keeping track of their associated uncertainty. At present, we focus on the assessment of species abundance and distribution in South Africa using two citizen-science bird monitoring datasets, namely: the African Bird Atlas Project and the Coordinated Waterbird Counts. These data are analyzed with occupancy and state-space models, respectively. In addition, a suite of environmental layers help contextualize waterbird population indicators, and link these to the ecological condition of the supporting wetlands. Both data and estimated indicators are accessible to end users through an online portal and web services.

Results and discussion: We have designed a modular system that includes tasks, such as: data cleaning, statistical analysis, diagnostics, and computation of indicators. Envisioned users of BIRDIE include government officials, conservation managers, researchers and the general public, all of whom have been engaged throughout the project. Acknowledging that conservation programmes run at multiple spatial and temporal scales, we have developed a granular framework in which indicators are estimated at small scales, and then these are aggregated to compute similar indicators at broader scales. Thus, the online portal is designed to provide spatial and temporal visualization of the indicators using maps, time series and pre-compiled reports for species, sites and conservation programmes. In the future, we aim to expand the geographical coverage of the pipeline to other African countries, and develop more indicators specific to the ecological structure and function of wetlands.

KEYWORDS

biodiversity informatics, citizen science, data pipeline, waterbirds, wetlands, species distribution, species abundance, diversity

1. Introduction

Freshwater ecosystems are among the most productive, biodiverse, and efficient at capturing and storing carbon (Convention on Wetlands, 2021). Unfortunately, they are also among the most impacted by human activity (Skowno et al., 2019; Convention on Wetlands, 2021), and climate change will likely exacerbate the pressure on freshwater resources. This is particularly true for the African continent, home to some of the largest wetlands, which not only host a wealth of freshwater species, but are also key in supporting human communities (Stephenson et al., 2020). Such critical issues have fueled unprecedented efforts to collect and mobilize freshwater biodiversity data (Wetzel et al., 2015; Dallas et al., 2021).

While we must strive to keep monitoring programmes that deliver data funded and alive, it is clear that data on their own are not enough (MacFadyen et al., 2022). If we are to take effective action to stop ecosystem degradation, it is important that data are analyzed to extract indicators that are meaningful for decision- and policy-making (Harebottle and Underhill, 2016; Stephenson et al., 2017b; Jetz et al., 2019). Furthermore, with continuous data collection, we need to implement workflows that update indicators and support decisions in a timely fashion (Yenni et al., 2019; MacFadyen et al., 2022). Automated data pipelines allow us to keep datasets updated and free of errors (Yenni et al., 2019), make model-based forecasts, and evaluate previous forecasts in light of new data (White et al., 2019). These modern and automated data workflows require multidisciplinary skills in ecology, statistics, data science, and software development, but their end products should ideally be free, accessible and easy to interpret (Stephenson et al., 2017b). It would also be desirable that they integrate multiple datasets and environmental layers to produce a holistic understanding of biodiversity structure and function (MacFadyen et al., 2022).

South Africa is leading the African continent in terms of biodiversity data availability (Barnard et al., 2017), with successful citizen-science programmes such as the Southern African Bird Atlas Project (Brooks et al., 2022), and biodiversity data platforms, such as the Biodiversity Advisor [South African National Biodiversity Institute (SANBI), 2023] or the Freshwater Biodiversity Information System (FBIS, Dallas et al., 2021). In contrast, dashboards and tools that facilitate the timely uptake of information and unlock the utility of current data are still limited. There is also an imbalance in data availability across taxonomic groups and habitats. Regular monitoring of the status, distribution, and condition of wetlands ecosystems is urgently required to understand environmental pressures on wetland habitats, but challenges associated with limited human and budget capacity hamper the collection of the necessary data. Conversely, available waterbird species data are rich in detail and coverage, and could provide a stronger basis for both adaptive management and reporting at priority wetland sites.

Here, we describe a data pipeline that implements a workflow of wetland- and waterbird-related biodiversity data, the South African Biodiversity Data Pipeline for Wetlands and Waterbirds (BIRDIE). At present, most of BIRDIE's functionality focuses on computing indicators related to waterbird distribution and

abundance, which are considered the minimum set of variables necessary to study changes in species populations (Pereira et al., 2013; Jetz et al., 2019). BIRDIE utilizes two long-term citizen-science programmes that have collected waterbird data in South Africa for more than two decades, and are still active: the Southern African Bird Atlas Project (SABAP; Brooks et al., 2022) and the Coordinated Waterbird Counts (CWAC; FIAO, 2022). Apart from waterbird data, BIRDIE uses and serves ancillary environmental data for contextualizing the aforementioned waterbird population variables, and also for describing the state of the wetlands that support them. In a next phase, we plan to expand the functionality of the pipeline to provide indicators of wetland ecosystem structure and function.

BIRDIE is embedded into the South African National Biodiversity Institute (SANBI) biodiversity informatics infrastructure and it was conceived as a tool to inform environmental strategies, identify priorities for the protection and sustainable use of biodiversity, and to guide land-use management. Because such policy-linked objectives require updated and timely information, the pipeline was designed to run periodically (yearly in principle), and automatically (but supervised). Currently, BIRDIE provides indicators for South Africa only, but in the future we expect to expand its coverage to other African countries. In what follows we describe BIRDIE's data pipeline workflow from data acquisition to display of final outputs (Figure 1), as well as the technologies we have used and the general modeling frameworks adopted.

2. Framework and target users

The main objective of BIRDIE is to provide information to support authorities that need to report on the state of wetlands or waterbird populations at multiple levels: (1) as required by national and international programmes and agreements, (2) provincial authorities, site managers and other stakeholders who need to make a range of decisions specific to certain wetlands, and (3) the general public could make use of BIRDIE's freely available outputs for a variety of reasons, including recreation and local conservation initiatives.

Indicators on the state of biodiversity have been adopted by a range of multilateral environmental agreements including the United Nations Convention on Biological Diversity (CBD, 2022) and Sustainable Development Goals (SDGs; United Nations, 2022). New indicators are under development and established processes, such as the International Union for the Conservation of Nature (IUCN, 2022) species red-listing efforts, are receiving renewed attention (Han et al., 2017). With these indicators come various global and national initiatives and targets for reducing rates of biodiversity loss (Mace et al., 2018). Essential Biodiversity Variables (EBVs) have been conceptualized and developed to help standardize and improve interoperability of biodiversity data and monitoring (Pereira et al., 2013). Within this framework, BIRDIE gives support to both national and international programs contributing information about the state of waterbird populations in South Africa, with a view to expand to the Southern Africa

region. We focus primarily on species population EBVs, with the assessment of waterbird abundance, distribution and diversity, and changes of these over time (Kissling et al., 2018; Jetz et al., 2019).

At an international scale, the BIRDIE team has engaged in conversation with two strategic partners from the project outset: the Ramsar Convention Secretariat and the Technical Committee of the Agreement on the Conservation of African-Eurasian Migratory Waterbirds. South Africa is signatory to the Ramsar Convention (Convention on Wetlands, 2021), hosts 28 Wetlands of International Importance, and needs to produce reports on the state of these sites every 3 years. National reports must also be compiled for the Agreement on the Conservation of African-Eurasian Migratory Waterbirds (AEWA; United Nations Environmental Programme, 2022), an international agreement, framed under the Convention on Migratory Species, and focused on protecting migratory waterbirds and their habitats. The Ramsar Convention and AWEA both require information on changes in overall abundance and distribution of waterbirds, with AWEA focusing on migratory species. Both conventions also report on indicators such as change in wetland extent and condition. Engagement with the South African national government bodies for both of these conventions ensures the reporting component of the BIRDIE project responds directly to their needs.

At the national level, South Africa produces a National Biodiversity Assessment every 4 years, which constitutes the main reporting tool of the state of biodiversity in the country, and informs policy and conservation strategies (Skowno et al., 2019). At the same time, there are regular efforts to address the conservation status of South African species within the IUCN Red-List framework. Changes in abundance and distribution of species are key in these assessments to track and report on population trends, and shifts in species ranges and community diversity. BIRDIE is embedded within SANBI, which is the organization mandated to report on the state of biodiversity in South Africa. As such, the outputs produced by the pipeline have a direct connection to needs specified for National Biodiversity Assessments, the Freshwater Biodiversity Programme and other national decision processes regarding freshwater ecosystems and species.

Keeping these main reporting channels in mind, BIRDIE also intends to support local management actions and basic research. Site-scale wetland monitoring is severely limited in South Africa, lagging far behind monitoring of other aquatic ecosystems such as rivers and estuaries. Managers ideally need to report on the state of the wetland (e.g., wetland condition, flux in surface water extent) as well as the species that the wetland supports, including species of special concern. Local waterbird and wetland information can facilitate the development of site-specific management actions and management plans, and support permitting decisions. At the same time, linking the local manager inputs and feedback into the data pipeline closes the gap between large-scale assessments and local data collection. In this sense, throughout the development of the pipeline, we have engaged with stakeholders at a pilot site, the Barberspan Nature Reserve. These conversations were enormously insightful to understand the variety of questions that may arise when working at a local level. One key take-away message from these engagements was that we should favor a flexible online portal,

where users can customize their queries, over a rich but fixed set of outputs.

Finally, we hope that the data pipeline will also allow citizen scientists to more actively interact with the data they have collected, and to see it taken up into the statistical analyses and data visualizations. The general public could also benefit from a flexible wetland and waterbird portal, with the right information to aid their interpretation.

3. Input data

In South Africa, we have a number of long-running citizen science projects that help monitor waterbird populations throughout the country. At its core, BIRDIE leverages two bird-related datasets: the Coordinated Waterbird Counts (CWAC, FIAO, 2022) and the second phase of the South African Bird Atlas Project (SABAP2, Brooks et al., 2022), which is part of the larger African Bird Atlas Project (ABAP). These datasets have well-established citizen scientist support and offer information about: (1) bird abundance, with waterbird counts taken twice a year at 731 water bodies across Southern Africa (mostly South Africa) since 1992, and (2) species occurrence, with visits to a grid of pentads ($5' \times 5'$ grid cells) initiated in 2007 and covering several African countries.

The Coordinated Waterbird Counts project provides regular counts of all waterbirds at just over 700 sites throughout South Africa. Counts are predominantly conducted by field observers from a set of observation points defined for each site, and that are visited twice a year; although in some sites other types of counts, such as count by boat, are also used (FIAO, 2022). The project was launched in 1992 and since then, it has accumulated a long time series for many sites. However, not all sites have been monitored since the start of the project, some regions are better represented than others, and not all sites have been monitored continuously (Figure 2). Waterbird species have diverse habitat requirements and life histories; some use the same sites year-round, whereas others are migratory or undergo local movements. To capture this diversity, CWAC counts are carried out twice per year: once in mid-summer and once in mid-winter. Although counts incorporate errors due to imperfect waterbird detection by observers, with appropriate statistical analyses, they can reveal long-term temporal trends and seasonal fluctuations in waterbird populations.

ABAP offers occurrence, rather than abundance data. In ABAP, volunteers collect checklists of all birds observed over a grid of pentads ($5' \times 5'$ minute grid) covering different African countries (Brooks et al., 2022). We are currently restricting our analysis to South Africa, and therefore we are using the SABAP2 component of ABAP (Figure 3). However, we plan to expand BIRDIE's functionality to cover other countries contributing data to ABAP, such as Kenya or Nigeria. Under the SABAP2 protocol, which started in 2007, observers need to spend at least 2 h of intensive birding at a pentad and are asked to visit as many habitats within it as possible. They can add new species for up to 5 days. SABAP2 currently has ca. 17 million records, and >2 million records are added per year. The structured sampling protocol, together with the spatial and temporal extent of SABAP2 allow us to examine how bird

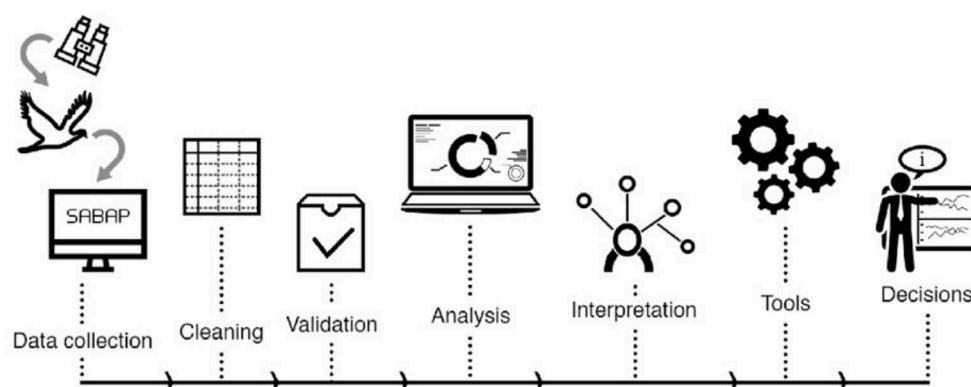


FIGURE 1

Basic workflow of the BIRDIE pipeline covering all steps from data collection, to analysis and presentation of digested, decision-ready indicators. Note that this is not a detailed sequence of all steps data go through, but rather a simplified view of the main processes.

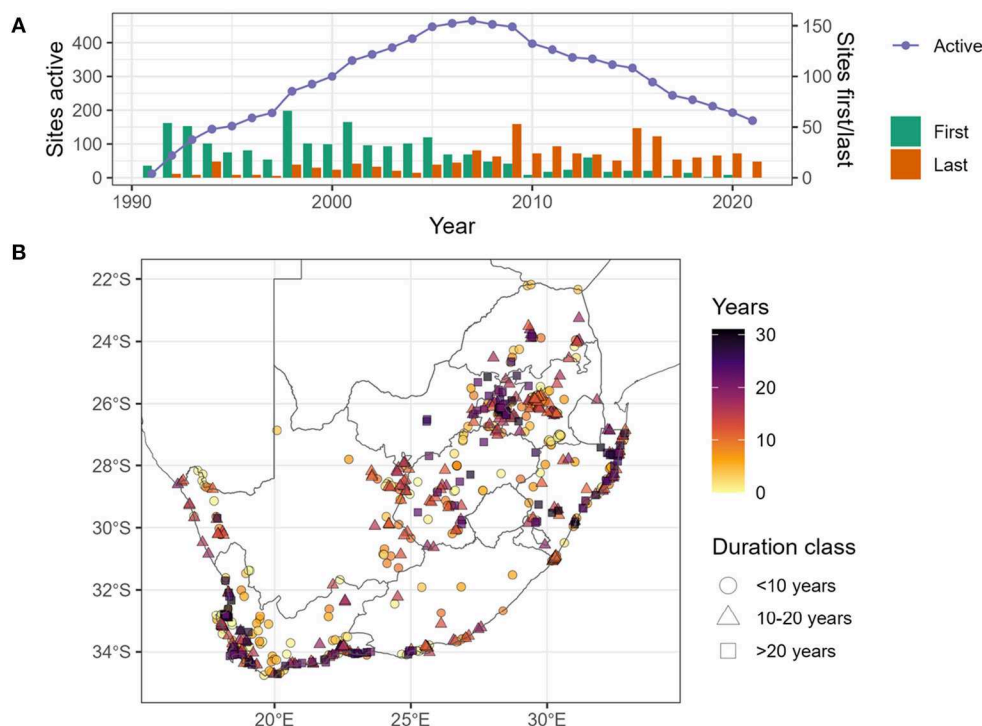


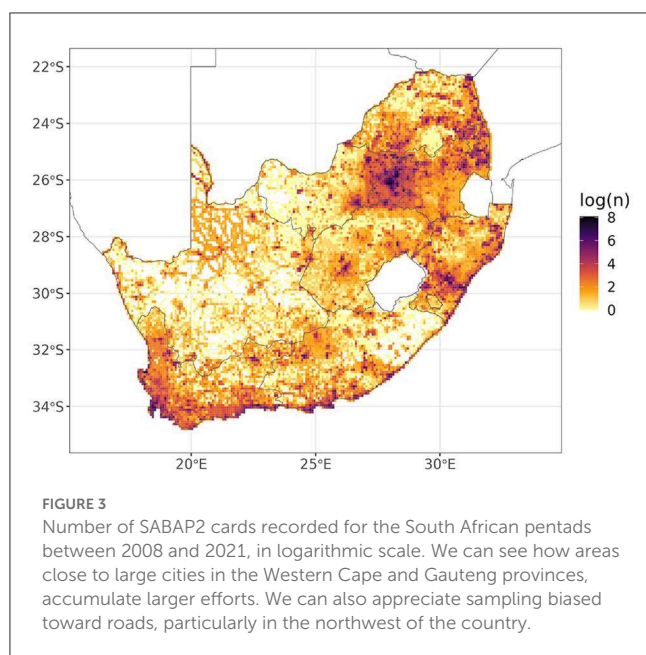
FIGURE 2

The graph (A) shows, the number of CWAC sites active (purple dots and line), number of sites firstly counted each year (green bars) and number of sites last counted each year (orange bars), between 1991 and 2021. Note that some of the sites that were last counted before 2021, might be counted again in the future, so orange bars do not represent sites removed from the programme. In map (B), we show the spatial distribution of CWAC sites in South Africa. The color gradient represent the duration of the period the site was counted for. To aid visualization, we show different shapes for different duration categories.

distributions are changing over time, although statistical modeling is required to account for imperfect detection and spatial sampling biases (Figure 3).

There are a variety of other data sources that BIRDIE uses for adding environmental information into its analytical workflows. Most of these data sources are conveniently accessed through

Google Earth Engine, such as TerraClimate (Abatzoglou et al., 2018), the JRC surface water dataset (Pekel et al., 2016), MODIS Vegetation Indices (Didan, 2015), and Digital Elevation Models (DEM, Yamazaki et al., 2017). Other data not yet available on Google Earth Engine, such as the National Wetland Map (van Deventer et al., 2020) are managed independently.



4. Indicators and statistical methods

Capturing good quality raw data is a fundamental first step to monitor the state of biodiversity. However, raw data reflect not only the biological signal of interest but also the sampling process, which is typically spatially biased and subject to imperfect detection (Yoccoz et al., 2001). Therefore, some level of statistical analysis is required to estimate the state of the system of interest, and separate it from observational artifacts introduced by the observation process used for capturing the data (Yoccoz et al., 2001; Gimenez et al., 2014; King, 2014). The BIRDIE pipeline broadly uses two types of models: (1) occupancy models (MacKenzie et al., 2002; Altwegg and Nichols, 2019) to estimate the probability of a species being present at the different SABAP2 pentads, and (2) state-space models (Buckland et al., 2004; Newman et al., 2014) to estimate the number of individuals at the sites monitored by the CWAC programme. Contrary to raw observations (counts and detection/non-detection of a species), model-based estimates (abundance and occupancy probabilities) allow us to quantify uncertainty.

The variety of end-user needs requires a pipeline that provides waterbird population indicators at multiple spatial and temporal scales. Therefore, in addition to estimating basic occupancy and abundance at small scales (i.e., individual site/pentad), the BIRDIE pipeline produces other high-level indicators obtained by aggregation (Table 1). The idea is to follow a process whereby raw data are used to inform models that estimate indicators at the smallest temporal and spatial scales possible, and then to aggregate these estimates at larger scales, as required. For example, species abundance can be estimated for a set of regularly monitored wetlands in South Africa, and these site-specific estimates can then be combined to calculate an abundance index for all sites as a group. We can follow this procedure to estimate abundance and occupancy probabilities at national, regional and local levels, as well

as for specific groups of wetlands (e.g., designated Ramsar sites, estuaries, or artificial sites).

The main indicators computed by the BIRDIE pipeline for waterbird species are:

- **Abundance:** estimated for CWAC sites in two seasons per year. For each species, only those wetlands with at least a 10-year coverage between 1997 and 2021 are analyzed statistically.
- **Occurrence:** estimated for ABAP pentads on an annual basis.
- **Diversity:** the simplest and most easily understood metric is species richness. Species richness can be calculated based on the occupancy analysis, by summing occupancy probabilities of all species potentially present in each pentad, to estimate the expected number of species present.
- **Important records:** sightings of rarities, invasive species. Although this information does not require any statistical processing, it does make particular records more visible.

In addition to estimates of static indicators, the pipeline also estimates their associated dynamics, such as: changes in abundance, occupancy probabilities and diversity. The temporal reference for these dynamics can also vary ranging from a single season to multiple years (typically ca. 5 years, for short-term changes, and ca. 15 years for long-term changes).

It is important that uncertainty is correctly propagated when aggregating, and also when estimating dynamic indicators. We work in a Bayesian framework and use the posterior distribution of occupancy probabilities and species abundance to define indicators at the various scales. Working with full posterior distributions allows us to conveniently keep track of the uncertainty in the estimates used as building blocks for other derived indicators.

4.1. Delineating species distributions

Occupancy models are fitted to detection/non-detection data from SABAP2 to delineate the distribution of waterbird species and its dynamics over time. Within the SABAP2 framework, observers visit pentads and make a list of the bird species detected during the visit. We assume that observers identify species correctly and only list species observed (the rigorous vetting process of SABAP2 data justifies this assumption), but non-detections may be caused by either species not being present in the pentad or by observers not being able to detect them, when present. Therefore, occupancy models describe two processes simultaneously: (i) the underlying occupancy of the sites (pentads), and (ii) the observation process whereby species present might or might not be observed.

More precisely, we define z_{jt} to be the true occupancy of site j in year t , which can be 1 (if species present) or 0 (if species absent) and has distribution:

$$z_{jt} | \psi_{jt} \sim \text{Bernoulli}(\psi_{jt})$$

where ψ_{jt} is the occupancy probability at site j and year t . The logit transformation of ψ_{jt} can be modeled as a linear combination

TABLE 1 Main indicators produced by the BIRDIE pipeline for waterbird species.

Indicator	Input	Model	Spatial scale	Temporal scale
Abundance	CWAC	SSM	CWAC site	2 seasons/year
Diversity	ABAP	Occupancy	Pentad	Annual
Extent of occurrence	Occurrence	Aggregated	National	Annual
Area of occupancy	Occurrence	Aggregated	National	Annual
Population size	Abundance	Aggregated	National	2 seasons/year
Pop. proportion on site	Abundance	Aggregated	CWAC site/national	2 seasons/year
Waterbird Conservation Value	Abundance	Aggregated	CWAC site/national	2 seasons/year
Number of sites	Abu./occu.	Aggregated	National	Annual

For each indicator, we show: “Input”, which can be databases (CWAC, Coordinated Waterbird Counts; ABAP, African Bird Atlas Project), or other indicators; “Model” used to estimate the indicator (SSM, state-space model; Occupancy, occupancy model) or whether it was computed by aggregating lower-level indicators; the smallest “Spatial scale” of assessment; and the smallest “Temporal scale” of assessment. Indicators are: “Abundance”, number or individuals; “Diversity”, number of species; “Extent of occurrence”, area of minimum complex polygon enclosing sites with species presence; “Area of occupancy”, area of sites (SABAP2 pentads) with species presence; “Population size”, number of individuals in South Africa; “Population proportion on site”, percentage of the population present on each site; “Waterbird Conservation Value”, index based on [Harebottle and Underhill \(2016\)](#); “Number of sites”, number of CWAC sites where the species is present. Annual changes in all of these indicators are also computed, and other indicators will be added over time as needed.

of covariates and smooth functions of covariates, such that:

$$\text{logit}(\psi_{jt}) = \mathbf{x}_{jt}^T \boldsymbol{\beta} + \sum_{k=1}^K f_k(u_{jk})$$

where $f_k(u_{jk})$ is a smooth function of the covariate u_k , which is defined as

$$f_k(u_{jk}) = \sum_{l=1}^L B_l(u_{jkl}) \gamma_{jkl}$$

where the smooth function f is represented by a set of L basis functions B_l evaluated at the value of the covariates associated with site j at year t ([Wood, 2006](#)).

We can then write the likelihood of observation y_{ij} as:

$$y_{ij}|z_{jt}, p_{ij} \sim \text{Bernoulli}(z_{jt} p_{ij})$$

The probability of detection of a species that is present in site j on visit i is denoted by p_{ij} . Following the same logic as for the probability of occupancy, the logit transformation of p is modeled as a linear combination of covariates and smooth functions:

$$\text{logit}(p_{ij}) = \mathbf{w}_{ij}^T \boldsymbol{\alpha} + \sum_{h=1}^H f_h(v_{ih}),$$

Spatial, spatio-temporal, and unstructured random effects can be specified for either occupancy or detection probabilities to account for variation across sites, observers and visits, not accounted for by the covariates incorporated in the models.

Each checklist is treated as an independent survey, but occupancy is assessed on a yearly basis. This means that if a species is detected in any one survey it is considered present that year. Therefore, missing a species because it left the site is considered part of the observation process and not the occupancy process. Migratory birds, for example, are considered present at a site even if they are only there for part of the year.

We are fitting single-season occupancy models without spatial random effects to most species. However, all models incorporate

random effects to account for pentad- and observer-specific detection probabilities. If model diagnostics indicate poor model fit (see Section 4.3 below), we assess models individually to understand the reasons, and if necessary we add spatial random effects for occupancy probabilities with an exponential decay function. Currently, we fit the models in R ([R Core Team, 2022](#)), in a Bayesian framework using the package `spOccupancy` ([Doser et al., 2022](#)), and running three MCMC chains for 20,000 iterations, with a thinning interval of 20. We use non-informative priors for all parameters when no information from other years is available, but we incorporate information obtained from other model fits if available, by centering the priors on the closest model's posterior means. However, it is important noticing that modeling details may differ among species and may be updated in future versions of BIRDIE.

4.2. Estimating abundance and population trends

State-space models ([Buckland et al., 2004](#); [Newman et al., 2014](#)) are used to describe and understand dynamic systems that may not be perfectly observed. Within this framework, we consider waterbird abundance to be a process that evolves over time, and which we observe during visits to CWAC sites. However, counts conducted by observers are distorted by imperfect detection that translates into counting errors. By counting repeatedly over time, and assuming that abundance evolves smoothly over time compared to observation error, we can disentangle these two processes.

We consider that the observed counts (y_i) at sampling occasion i (generally there were two sampling occasions per year, one in mid-summer and one in mid-winter), at any given site, arise from a Poisson (λ_i) distribution

$$y_i \sim \text{Poisson}(\lambda_i)$$

And we model the log of the intensity λ_i as:

$$\log(\lambda_i) \sim N(\mu_i, \sigma^2)$$

where μ_i is the mean abundance of waterbirds present at a site on sampling occasion i and σ^2 is the corresponding variance of the observers counting error, both in the log scale. Therefore, counts depend both on the number of waterbirds present on site, and on errors in the counts of these birds.

To model changes in waterbird abundance between the two-seasons of year t , we define s_t to be the summer abundance and w_t the winter abundance. Note that there might be multiple counts in a single year and season, but the underlying true abundance is considered to stay constant in any given year and season (for clarity, note also that while sampling occasions were indexed by i , years are indexed by t). Thus, the expected (log) abundance for any given count can be written as

$$\mu_i = s_t \text{summer} + w_t \text{winter}$$

where “summer” is an indicator variable that takes on the value 1 in summer and 0 in winter, and “winter” is the opposite. We then define abundance dynamics as:

$$\begin{aligned} s_t &= s_{t-1} + \beta_t \\ w_t &= s_t + \xi_t \end{aligned}$$

where β_t corresponds to the change in summer abundance from year $t - 1$ to year t , and ξ_t is the difference between summer and winter abundance, both in the log scale. If exponentiated, these parameters can be interpreted as the rate of change in the population and the winter-to-summer ratio of the population, respectively.

We impose relatively smooth changes in abundance by defining autocorrelation in β_t and ξ_t terms over time. In addition, we define relationships between the rate of change in the population β_t and environmental covariates. These relationships facilitate the estimation of abundance for those years in which counts are missing, and it is particularly useful to contain uncertainty in long periods with missing data between counts. Thus, we set

$$\begin{aligned} \beta_t &= \phi\beta_{t-1} + \eta_{t-1} + \zeta_{t-1} \\ \xi_t &= \xi_{t-1} + \epsilon_{t-1} \end{aligned}$$

where ϕ lies between zero and one, and it defines an autoregressive term on β_{t-1} ; η_t captures the effect of covariates in the expected change in abundance, and can be expanded to $\gamma^T U$, where U is a matrix of covariate values and γ a vector of coefficients; ζ_t and ϵ_t are random variables that represent change in abundance change, and change in winter to summer ratio, respectively.

We mentioned at the beginning that this model applies to each monitored site. However, we have multiple sites, and counts are often missing for some seasons or even full years. To facilitate the estimation of abundance with missing data, we borrow information from sites with counts, by defining a hierarchical structure such that:

$$\begin{aligned} \zeta_{ij} &\sim N(0, \sigma_{\zeta_t}^2) \\ \epsilon_{ij} &\sim N(0, \sigma_{\epsilon_t}^2) \end{aligned}$$

Therefore, random changes at any site and year come from a common distribution of changes across all sites for that year. We thus ensure that variation is contained within similar values in most sites. These distributions are normal with variances $\sigma_{\zeta_t}^2$ and $\sigma_{\epsilon_t}^2$ for changes in abundance and winter to summer ratio, respectively.

We fit these models in R (R Core Team, 2022) with the additional functionality provided by JAGS (Plummer, 2003) using the package jagsUI (Kellner, 2021). We work on a Bayesian framework, using non-informative priors, and running three chains for 10,000 iterations each. Similar to the occupancy models, these are the details of the models we are working with at the time of writing, and they are intended to give an idea of the type of model we are using. The modular nature of BIRDIE allows us to update these models when necessary and the updated modeling details will be published on the BIRDIE website.

4.3. Data and model diagnostics

The pipeline needs to run for a multitude of species, with different ecological requirements and geographical distributions. Therefore, finding a model that suits all species is challenging. Not only may a model not be a good fit for a particular species, but the algorithms used for fitting the model may fail to converge due to characteristics of the data.

In a first control stage, we have defined the minimum requirements that the data should meet to enter the model-fitting process. Species that have been observed in five or less pentads in a year are considered to not have enough data to inform an occupancy model. Similarly, we chose only those CWAC sites where the species of interest has been counted at least ten times between 1993 and 2021, to fit state-space models. Otherwise, data tend to be too sparse to assess trends in abundance reliably. These thresholds are based on our own experiences working with these data, and they are considered to be the minimum requirements for models to converge successfully. However, meeting these requirements does not guarantee model convergence or a good fit. To keep track of potential issues arising during model fitting, and to improve the algorithms of the pipeline, each time the pipeline runs it generates several reports that are later examined.

To decide whether any occupancy or state-space model converges, we calculate the Gelman-Rubin (Rhat) diagnostic (Gelman et al., 2014) for each estimated parameter. These diagnostics are then tabulated and stored for future revision. Any Rhat value above 1.1 or below 0.9 is considered to represent lack of convergence. Distinctive characteristics of the models with convergence issues are explored and addressed on a case by case basis, after the pipeline has finished running.

In addition to convergence, we assess goodness of fit using posterior predictive checks (Gelman et al., 2014; Doser et al., 2022). This procedure compares some quantity of interest calculated using pseudo-data simulated from the model posterior distribution, with that same quantity calculated from the observed data. In a well-fitting model we would expect real and synthetic data to produce similar values. For occupancy models, we produce simulated detection/non-detection data for each site, species and year and compute the expected number of detections out of as

many visits as there were in the data. We compare the results of the simulations with the observed number of detections recorded in the data using Chi-square tests (Doser et al., 2022). For state-space models we follow a similar procedure, but instead of simulating detection/non-detection data for 1 year, we simulate count data for summer and winter, and aggregate these in a single annual count. Results from the goodness of fit Chi-square tests are also tabulated and stored for revision. Significant deviations detected with these tests are addressed for each case individually.

Due to the computational burden of the pipeline, it is not possible to run multiple models for each species, site and year, to perform model selection. Therefore, model selection is performed on a sample of species, selected to have representation of common and scarce taxa, but that are otherwise selected arbitrarily. Our general approach has been to include a rich set of variables that we believe can explain the main environmental gradients within our geographical range, without paying too much attention to multi-collinearity and overfitting. We are therefore cautious about making causal inference or predictions outside of the range of the data, and so should be other users.

5. Systems and technology

In this section, we describe the technology that underpins the flow of data along the pipeline until it is transformed into indicators that are presented to the BIRDIE user. BIRDIE's data, code and outputs are stored and run on three main systems (Figure 4): the Africa Bird Data servers, and the two BIRDIE servers (servers A and B).

The Africa Bird Data servers are hosted at the FitzPatrick Institute for African Ornithology, University of Cape Town, and contain the CWAC and ABAP databases. They also serve these data through an Application Programming Interface (API).

BIRDIE's server A is the access point of the final user to the information generated by the pipeline. This information is stored in a data mart, which in essence, is a MySQL database (version 8.0.27), a widely used, open source, relational database management system. Its main objective is to store the outputs of BIRDIE's data analyses and provide easy and flexible access to the final user. At the same time, the structure of the database ensures that inputs and outputs conform to a given standard, and creates security back-ups for the stored data. The main mechanism BIRDIE uses to present data to the user is through OpenAPI web services (OpenAPI Specification, Version 3.1.0), which was designed to provide a standard interface for documenting and exposing APIs. The public web services offered by the OpenAPI give users the flexibility to access and download data from the database without being constrained by the specific functionality of a web application. This technology facilitates the integration of BIRDIE's outputs into other workflows. However, for the user that is interested in readily accessing the information through a dashboard, we have deployed a web application, written in HTML5, CSS, and the most common and popular JavaScript libraries, including OpenLayers (<https://openlayers.org/>) and Plotly (<https://plotly.com/>). Among other elements (see Section 6), the web application features a map viewer, based on mviewer (<https://mviewer.netlify.app/en/>), a free and

open-source cartographic application, that has an easy-to-use and intuitive interface.

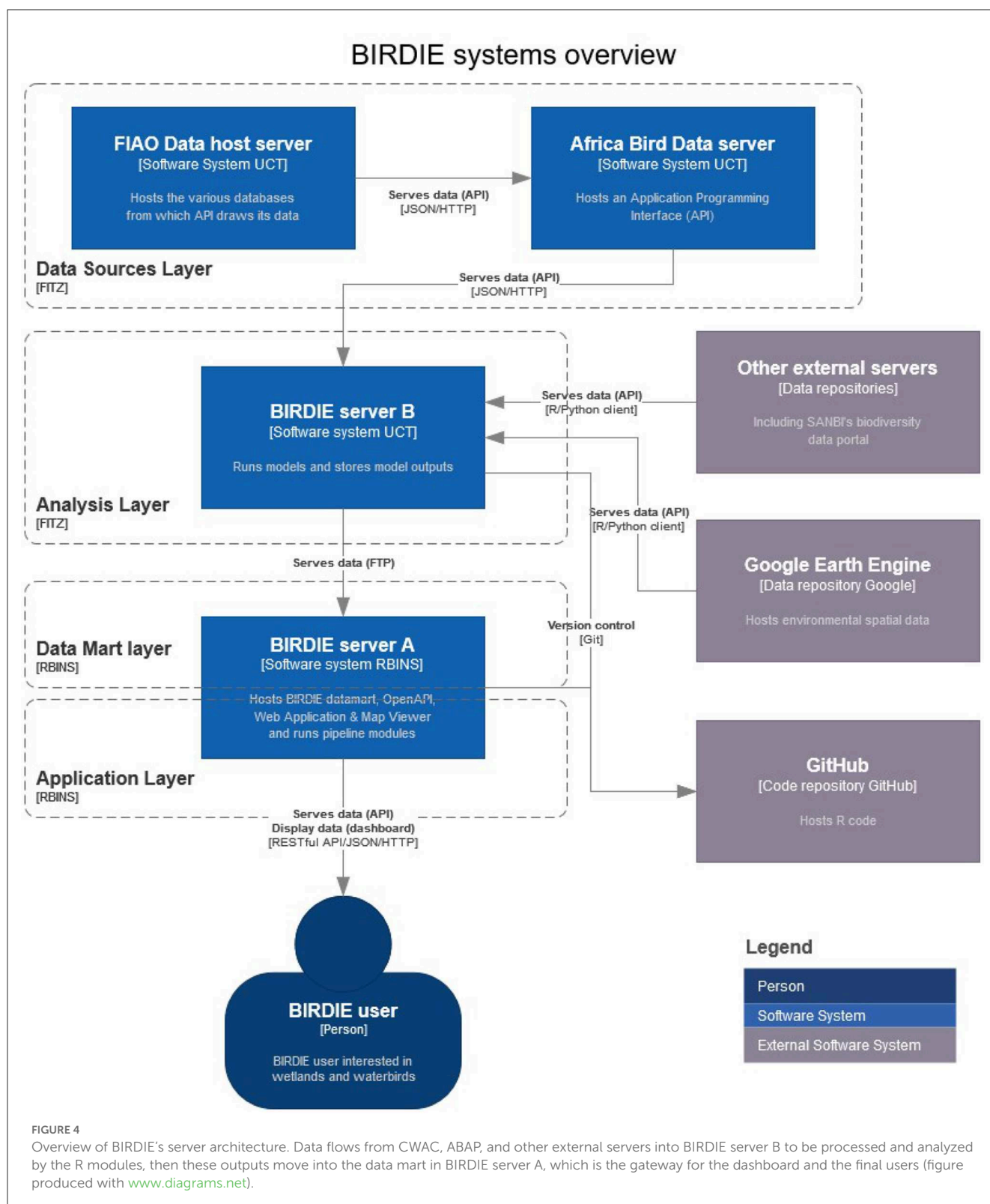
If we thought of server A as being the face of the pipeline, server B would be the brain. All the functionality in this server revolves around statistical modeling. This server connects with the Africa Bird Data servers to obtain CWAC and ABAP data, and with other external systems, such as Google Earth Engine or SANBI servers to obtain environmental information. It then runs the main analytical modules of the pipeline, where occupancy and state-space models are fitted. At the time of writing, the analytical workflows were supported by an Intel Xeon Dual 8 core, with 64 GB RAM and an 8 TB hard drive. The model outputs are made available to server A, where they are incorporated into the data mart, used to compute derived high-level indicators by aggregation (see Section 4), and prepared to be presented to the final users.

In terms of code structure, the BIRDIE data pipeline consists of several fundamental building blocks or modules. The first module, which we call the data source layer (Figure 4) hosts and curates the raw data. The second module, the analysis layer, analyses the data and estimates the fundamental quantities of interest, like abundance and occurrence of each species at each wetland or pentad. The third module consists of the data mart where the outputs of the analyses are stored and indicators are aggregated or disaggregated to multiple scales. The final module serves the information to the user *via* APIs, web services and a web application. The modular structure of BIRDIE enables us to maintain and update individual parts independently. For example, we could replace the current statistical routines with more efficient ones without changing the other parts of the pipeline. Or we could add new indicators to the data mart layer without needing to change the statistical routines that produce the underlying components.

6. Web application

To cater for different user needs, BIRDIE's web application offers four main menus that provide access to the pipeline outputs in different ways (see Figure 5):

1. An exploration map. Through this menu the user can explore the different indicators BIRDIE computes on a map. This spatial framework can be configured to display information layers, such as occupancy probabilities for ABAP pentads or waterbird abundance at CWAC sites. Users can also zoom in and out to find the scale that best fits their needs. In addition to this, there are also environmental layers that can be overlaid to provide context and generate hypotheses on what might be driving the observed indicators.
2. Site and species summaries, are detailed reports elaborated for users focused on some sites or species in particular, rather than in general exploration. At the moment, site summaries are only available for those sites that have sufficient CWAC data to be included in BIRDIE's data analysis step. These reports contain a description of the site/species, links to other resources of interest (e.g., to criteria motivating declaration of Ramsar site or IUCN conservation status) and summaries prepared from BIRDIE's indicators. These reports can be exported as a document, and BIRDIE's data used for generating the reports can be accessed



through the data mart and downloaded in common formats such as .json or .csv.

3. Reporting tools. We mentioned in Section 2 that BIRDIE was developed to support reports for national and international

conservation programmes. In this menu, users interested in elaborating, or accessing the information underpinning these reports, will find this information conveniently packed in programme-specific summaries. Similar to site and species



summaries, reports for conservation programmes can also be printed, and the data used to compute statistics and create plots can be downloaded.

4. Web services. Through this menu users can access BIRDIE's API and retrieve its outputs in the most flexible way. It is through BIRDIE's API that all maps and plots in the web application are produced. By accessing this functionality directly, users can download the data themselves and incorporate them into their own workflows.

7. Discussion

Data on biodiversity and related environmental drivers are collected at increasingly faster rates. Although these data can be accessed to support decisions at various levels, it can be difficult for decision makers to extract relevant information in a timely fashion (Stephenson et al., 2017b; MacFadyen et al., 2022). Apart from data availability and accessibility, obstacles for using biodiversity data in decision-making include (Stephenson et al., 2017a; MacFadyen et al., 2022): lack of analysis and interpretation, lack of technical accessibility with excessive use of jargon, and timely use of data. Here, we introduce BIRDIE, the South African Biodiversity Data Pipeline for Wetlands and Waterbirds; a data pipeline that aims to provide the information needed for making evidence-based decisions on wetlands and waterbirds in southern Africa. Target users of BIRDIE include government and public entities that need to report on the status of wetlands and waterbirds, as well as site managers, and the general public (e.g., birdwatchers).

BIRDIE is the first African full biodiversity data pipeline (from raw data to indicators) that we are aware of at the time of writing. Although biodiversity data portals are proliferating (Saran et al., 2022), examples of fully operational workflows for computing and displaying biodiversity indicators are still scarce (but see Brlik et al., 2021; Boyd et al., 2022). Compared to other richer countries, long-term datasets from biodiversity monitoring programmes are still scarce in many African countries (Proença et al., 2017; Stephenson et al., 2017a). In South Africa we are lucky to have two good bird monitoring programmes that provide data on waterbirds. However, even these well established programmes can be hampered by lack of funds and qualified personnel in remote locations, as we can see by the decreasing coverage of the CWAC project in the last decade (Figure 2). Critical data on the location, structure and dynamics of freshwater ecosystems are still scarce and highly local. Thus, BIRDIE relies heavily on citizen science projects such as ABAP and CWAC, which poses clear challenges in terms of uneven efforts and imperfect detection, but also adds the advantage of having the support of a large community of observers that provides a continuous and steady flow of data. These data inputs allow us to run the pipeline periodically to keep the indicators updated and timely. Although we would like to update our indicators more often, at the time of writing we only update once per year due to the computational requirements of the pipeline, and certain characteristics of the data (e.g., CWAC counts are conducted only twice a year).

All data used by BIRDIE are freely available, so one of the main contributions of BIRDIE is to facilitate information uptake by statistically analyzing these data and filtering out observational

artifacts introduced during data collection. Uneven sampling efforts, imperfect detection and missing data are all examples of how data collection methods can affect data (Yoccoz et al., 2001), and if undealt with, mislead decision making. Furthermore, statistical models also provide measures of uncertainty in their estimates, which must be clearly communicated to the stakeholders (Kissling et al., 2018). With all their benefits, these statistical analyses require technical knowledge and are time-consuming. Therefore, having their outputs pre-computed and readily available could dramatically increase the impact of the data. In this context, one of our main challenges was running models automatically and periodically for multiple species, which requires pre-defining and using similar models for all species. Therefore we faced a trade-off between having accurate models for individual species and having a pipeline that works reasonably well for all species in general. Users should keep in mind this compromise, and think of BIRDIE's outputs as useful approximations rather than accurate estimates. We recommend designing bespoke models for those species for which accuracy is required. Similarly, rare species are likely to appear too sparsely in datasets designed for monitoring common species for models to work well (Bellingham et al., 2020). For these species, we should design monitoring protocols and models that are tailored for them. Setting up feedback channels whereby users can suggest model improvements (e.g., relevant covariates) for certain species is a possible avenue for development in BIRDIE. However, in this first phase, the idea is to create a baseline pipeline in which the model structure is similar to all sites and species.

In addition, model structure was not designed for making causal inference and therefore confounders could mislead the user to believe that certain variables are driving emerging patterns, when there is only a correlation (Stewart et al., 2022). To avoid misinterpretation by the casual user, we favored displaying environmental layers that can overlay with model state estimations, rather than presenting marginal covariate effects estimated by the model. In future versions of BIRDIE, we might consider presenting this type of information in specific sections with extensive explanations on how to interpret it. The current version of BIRDIE has a portal that presents indicators that are easily accessed, visualized and interpreted, avoiding unnecessary jargon. At the same time, and for the interested user, we have allocated some space for clearly explaining the analytical routines used in all the analyses in dedicated sections. In BIRDIE, we followed the Findable, Accessible, Interoperable, Reusable (FAIR) principles (Wilkinson et al., 2016), making all processes reproducible and transparent. All the code used by the pipeline is public, freely available (<https://github.com/AfricaBirdData>) and based on open-source software.

In BIRDIE we envision several avenues for further development. Integration of multiple EBVs into a common assessment has important advantages for understanding drivers of change and designing conservation interventions (Bellingham et al., 2020). In the next phase, we intend to develop more profound links between waterbird population indicators and wetlands. Waterbirds are often regarded as good indicators of wetland biodiversity and condition. However, this assumption is rarely proven empirically, and it is apparent that it needs careful consideration on a case by case basis (Amat and Green, 2010). With advances in the accessibility to biodiversity data, we are

now in a better position to investigate whether these claims hold, and if so, under which conditions. Data portals such as GBIF.org and in South Africa, the Freshwater Biodiversity Information System (FBIS), and SANBI's biodiversity data portal, could help us understand how waterbird occurrence, abundance and diversity relates to the general ecological condition of the hosting wetlands. However, we are aware that the integration of opportunistic data with different sampling schemes and scales poses additional challenges that we will need to carefully address (Kissling et al., 2018; Boyd et al., 2022).

We will also extend BIRDIE's functionality to cover other African countries with similar available data, such as Kenya and Nigeria that also use the ABAP protocol. There is also a wealth of information that BIRDIE has not yet used, such as eBird or iNaturalist, that could improve the outputs of the pipeline. While integrating data sources with different sampling designs, coverages and biases is not trivial, the modular design of BIRDIE allows us to update the modeling step as new statistical methods are being developed. Data integration is a very active topic in the field of statistical ecology (Isaac et al., 2020). Approaches to combining data range from pooling multiple data sources together disregarding their different assumptions and biases, to much more accurate integrated models in which characteristics of each data source are explicitly accounted for (Fletcher et al., 2019). Although at the expense of increased model complexity, with the application of newly-developed statistical methods for data integration, we can now explore how different species interrelate, and inform more effective and efficient conservation actions.

We wish BIRDIE can contribute to closing the existing gap between data providers and decision makers, facilitating effective conservation action. We also hope it will provide a feedback channel to SABAP, CWAC, SANBI's Freshwater Biodiversity Programme and other data providers. Not only serving as a platform to analyse the data collected, but also to investigate coverage deficiencies and potential new priorities. Finally, we would like to see that BIRDIE exposes the importance of existing monitoring programmes, and that it helps prioritize new data-driven initiatives to understand and protect freshwater biodiversity.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found here: <https://sabap2.birdmap.africa/>, <https://cwac.birdmap.africa/>, <https://github.com/AfricaBirdData>.

References

- Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., and Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* 5, 170191. doi: 10.1038/sdata.2017.191
- Altwegg, R., and Nichols, J. D. (2019). Occupancy models for citizen-science data. *Methods Ecol. Evol.* 10, 8–21. doi: 10.1111/2041-210X.13090
- Amat, J. A., and Green, A. J. (2010). "Waterbirds as bioindicators of environmental conditions," in *Conservation Monitoring in Freshwater Habitats: A Practical Guide and Case Studies*, eds C. Hurford, M. Schneider, and I. Cowx (Dordrecht: Springer), 45–52.
- Barnard, P., Altwegg, R., Ebrahim, I., and Underhill, L. G. (2017). Early warning systems for biodiversity in southern Africa – How much can citizen science mitigate imperfect data? *Biol. Conserv.* 208, 183–188. doi: 10.1016/j.biocon.2016.09.011
- Bellingham, P. J., Richardson, S. J., Gormley, A. M., Allen, R. B., Cook, A., Crisp, P. N., et al. (2020). Implementing integrated measurements of essential biodiversity variables at a national scale. *Ecol. Solut. Evid.* 1, e12025. doi: 10.1002/2688-8319.12025
- Boyd, R., August, T., Cooke, R., Logie, M., Mancini, F., Powney, G., et al. (2022). An operational workflow for producing periodic estimates of species occupancy at large scales. *EcoEvoRxiv [Preprint]*. doi: 10.32942/osf.io/2v7jp

Author contributions

NJ was the project director. FC, RA, and VV developed the analyses of the pipeline. NJ, AS, and DH worked on reporting and indicators. FS and YS designed and implemented the data mart, APIs, web services and web application. MB managed the citizen science database. FC led the writing of the manuscript with contribution from all authors. All authors contributed to conception and design of the project. All authors contributed to the article and approved the submitted version.

Funding

This project was funded by the JRS Biodiversity Foundation, Grant Number 60908.

Acknowledgments

We are really grateful to other members of the BIRDIE team without whom this project would not be viable: Sediqa Khatieb, Monica Klass, and Carol Poole. We are also grateful for the support of the JRS Biodiversity Foundation, and to the many interested users that have engaged and shared useful insights with us. Finally, we would like to recognize the tremendous contribution of all the citizen scientists that devote their time and effort to collect the valuable data that we use.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Brlik, V., Silarova, E., Skoropilova, J., Alonso, H., Anton, M., Aunins, A., et al. (2021). Long-term and large-scale multispecies dataset tracking population changes of common European breeding birds. *Sci. Data* 8, 21. doi: 10.1038/s41597-021-00804-2
- Brooks, M., Rose, S., Altwegg, R., Lee, A. T., Nel, H., Ottosson, U., et al. (2022). The African bird atlas project: a description of the project and birdmap data-collection protocol. *Ostrich* 93, 223–232. doi: 10.2989/00306525.2022.2125097
- Buckland, S. T., Newman, K. B., Thomas, L., and Koesters, N. B. (2004). State-space models for the dynamics of wild animal populations. *Ecol. Model.* 171, 157–175. doi: 10.1016/j.ecolmodel.2003.08.002
- CBD (2022). *Convention on Biological Diversity*. Available online at: <https://www.cbd.int/> (accessed December 22, 2022).
- Convention on Wetlands (2021). *Global Wetland Outlook: Special Edition 2021*. Secretariat of the Convention on Wetlands, Gland.
- Dallas, H., Shelton, J., Sutton, T., Tri Cuptura, D., Kajee, M., and Job, N. (2021). The Freshwater Biodiversity Information System (FBIS) –mobilising data for evaluating long-term change in South African rivers. *Afr. J. Aquat. Sci.* 47, 291–306. doi: 10.2989/16085914.2021.1982672
- Didan, K. (2015). *MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006 [Data set]*. NASA EOSDIS Land Processes DAAC. Available online at: <https://doi.org/10.5067/MODIS/MOD13A2.006> (accessed February 28, 2023).
- Doser, J. W., Finley, A. O., Kery, M., and Zipkin, E. F. (2022). spOccupancy: an R package for single-species, multi-species, and integrated spatial occupancy models. *Methods Ecol. Evol.* 13, 1670–1678. doi: 10.1111/2041-210X.13897
- FIAO (2022). *CWAC: Coordinated Waterbird Counts*. Available online at: <https://cwac.birdmap.africa/index.php> (accessed December 21, 2022).
- Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology* 100, e02710. doi: 10.1002/ecy.2710
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press; Taylor and Francis Group.
- Gimenez, O., Buckland, S. T., Morgan, B. J. T., Bez, N., Bertrand, S., Choquet, R., et al. (2014). Statistical ecology comes of age. *Biol. Lett.* 10, 20140698. doi: 10.1098/rsbl.2014.0698
- Han, X., Josse, C., Young, B. E., Smyth, R. L., Hamilton, H. H., and Bowles-Newark, N. (2017). Monitoring national conservation progress with indicators derived from global and national datasets. *Biol. Conserv.* 213, 325–334. doi: 10.1016/j.biocon.2016.08.023
- Harebottle, D. M., and Underhill, L. G. (2016). Assessing the value of wetlands to waterbirds: Exploring a population-based index at flyway and regional levels. *Ostrich* 87, 7–21. doi: 10.2989/00306525.2015.1104396
- Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., et al. (2020). Data integration for large-scale models of species distributions. *Trend. Ecol. Evolut.* 35, 56–67. doi: 10.1016/j.tree.2019.08.006
- IUCN (2022). *International Union for the Conservation of Nature*. Available online at: <https://www.iucn.org/content/home-page> (accessed December 22, 2022).
- Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., et al. (2019). Essential biodiversity variables for mapping and monitoring species populations. *Nat. Ecol. Evol.* 3, 539–551. doi: 10.1038/s41559-019-0826-1
- Kellner, K. (2021). *jagsUI: A Wrapper Around “rjags” to Streamline “JAGS” Analyses*. R package version 1.5.2. Available online at: <https://CRAN.R-project.org/package=jagsUI>
- King, R. (2014). Statistical ecology. *Annu. Rev. Stat. Appl.* 1, 401–426. doi: 10.1146/annurev-statistics-022513-115633
- Kissling, W. D., Ahumada, J. A., Bowser, A., Fernandez, M., Fernandez, N., Garcia, E. A., et al. (2018). Building essential biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biol. Rev.* 93, 600–625. doi: 10.1111/brev.12359
- Mace, G. M., Barrett, M., Burgess, N. D., Cornell, S. E., Freeman, R., Grooten, M., et al. (2018). Aiming higher to bend the curve of biodiversity loss. *Nat. Sustain.* 1, 448–451. doi: 10.1038/s41893-018-0130-0
- MacFadyen, S., Allsopp, N., Altwegg, R., Archibald, S., Botha, J., Bradshaw, K., et al. (2022). Drowning in data, thirsty for information and starved for understanding: a biodiversity information hub for cooperative environmental monitoring in South Africa. *Biol. Conserv.* 274, 109736. doi: 10.1016/j.biocon.2022.109736
- MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 2248–2255. doi: 10.1890/0012-9658(2002)083%5B2248:ESORWD%5D2.0.CO;2
- Newman, K. B., Buckland, S. T., Morgan, B. J. T., King, R., Borchers, D. L., Cole, D. J., et al. (2014). *Modelling Population Dynamics: Model Formulation, Fitting and Assessment Using State-Space Methods*. New York, NY: Springer.
- Pekel, J.-F., Cottam, A., Gorelick, N., and Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 418–422. doi: 10.1038/nature20584
- Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., et al. (2013). Essential biodiversity variables. *Science* 339, 277–278. doi: 10.1126/science.1229931
- Plummer, M. (2003). “JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling,” in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vienna).
- Proença, V., Martin, L. J., Pereira, H. M., Fernandez, M., McRae, L., Belnap, J., et al. (2017). Global biodiversity monitoring: From data sources to essential biodiversity variables. *Biol. Conserv.* 213, 256–263. doi: 10.1016/j.biocon.2016.07.014
- R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available online at: <https://www.R-project.org/>
- Saran, S., Chaudhary, S. K., Singh, P., Tiwari, A., and Kumar, V. (2022). A comprehensive review on biodiversity information portals. *Biodivers. Conserv.* 31, 1445–1468. doi: 10.1007/s10531-022-02420-x
- Skowno, A., Poole, C. J., Raimondo, D. C., Sink, K. J., Van Deventer, H., Van Niekerk, L., et al. (2019). *National Biodiversity Assessment 2018: The Status of South Africa's Ecosystems and Biodiversity: Synthesis Report*. South African National Biodiversity Institute, Department of Environment, Forestry and Fisheries, Pretoria.
- South African National Biodiversity Institute (SANBI) (2023). *Biodiversity Advisor*. Available online at: <http://biodiversityadvisor.sanbi.org/> (accessed December 21, 2022).
- Stephenson, P. J., Brooks, T. M., Butchart, S. H., Fegraus, E., Geller, G. N., Hoft, R., et al. (2017b). Priorities for big biodiversity data. *Front. Ecol. Environ.* 15, 124–125. doi: 10.1002/fee.1473
- Stephenson, P. J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagona, M., Hoft, R., et al. (2017a). Unblocking the flow of biodiversity data for decision-making in Africa. *Biol. Conserv.* 213, 335–340. doi: 10.1016/j.biocon.2016.09.003
- Stephenson, P. J., Ntiamoa-Baidu, Y., and Simaika, J. P. (2020). The use of traditional and modern tools for monitoring wetlands biodiversity in africa: challenges and opportunities. *Front. Environ. Sci.* 8, 61. doi: 10.3389/fevns.2020.00061
- United Nations (2022). *Sustainable Development Goals*. Available online at: <https://sdgs.un.org/> (accessed December 22, 2022).
- United Nations Environmental Programme (2022). *AEWA: Agreement on the Conservation of African-Eurasian Migratory Waterbirds*. Available online at: <https://www.unep-awea.org/> (accessed December 21, 2022).
- van Deventer, H., van Niekerk, L., Adams, J., Dinala, M. K., Gangat, R., Lamberth, S. J., et al. (2020). National Wetland Map 5: An improved spatial extent and representation of inland aquatic and estuarine ecosystems in South Africa. *Water SA* 46, 66–79. doi: 10.17159/wsa/2020.v46.i1.7887
- Wetzel, F. T., Saarenmaa, H., Regan, E., Martin, C. S., Mergen, P., Smirnova, L., et al. (2015). The roles and contributions of Biodiversity Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case study. *Biodiversity* 16, 137–149. doi: 10.1080/14888386.2015.1075902
- White, E. P., Yenni, G. M., Taylor, S. D., Christensen, E. M., Bledsoe, E. K., Simonis, J. L., et al. (2019). Developing an automated iterative near-term forecasting system for an ecological study. *Methods Ecol. Evol.* 10, 332–344. doi: 10.1111/2041-210X.13104
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 18. doi: 10.1038/sdata.2016.18
- Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC Text in Statistical Science. New York, NY: Chapman and Hall/CRC. doi: 10.1111/j.1467-985X.2006.00455_15.x
- Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., et al. (2017). A high-accuracy map of global terrain elevations: accurate global terrain elevation map. *Geophys. Res. Lett.* 44, 5844–5853. doi: 10.1002/2017GL072874
- Yenni, G. M., Christensen, E. M., Bledsoe, E. K., Supp, S. R., Diaz, R. M., White, E. P., et al. (2019). Developing a modern data workflow for regularly updated data. *PLoS Biol.* 17, e3000125. doi: 10.1371/journal.pbio.3000125
- Yoccoz, N. G., Nichols, J. D., and Boulmier, T. (2001). Monitoring of biological diversity in space and time. *Trends Ecol. Evol.* 16, 446–453. doi: 10.1016/S0169-5347(01)02205-4



OPEN ACCESS

EDITED BY

Quentin Groom,
Botanic Garden Meise,
Belgium

REVIEWED BY

Richelle Li Tanner,
Chapman University,
United States

*CORRESPONDENCE

Carol Jean Poole
✉ c.poole@sanbi.org.za

[†]These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION

This article was submitted to
Environmental Informatics and Remote
Sensing,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 25 November 2022

ACCEPTED 27 February 2023

PUBLISHED 15 March 2023

CITATION

Poole CJ, Skowno AL, Currie JC, Sink KJ,
Daly B and von Staden L (2023) Taking state of
biodiversity reporting into the information age
– A South African perspective.
Front. Ecol. Evol. 11:1107956.
doi: 10.3389/fevo.2023.1107956

COPYRIGHT

© 2023 Poole, Skowno, Currie, Sink, Daly and
von Staden. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Taking state of biodiversity reporting into the information age – A South African perspective

Carol Jean Poole^{1*†}, Andrew Luke Skowno^{1,2†}, Jock C. Currie¹,
Kerry Jennifer Sink^{1,3}, Brenda Daly¹ and Lize von Staden^{1,4}

¹Kirstenbosch Research Centre, South African National Biodiversity Institute, Cape Town, South Africa,

²Plant Conservation Unit, Department of Biological Science, University of Cape Town, Cape Town,

South Africa, ³Institute for Coastal and Marine Research, Nelson Mandela University, Gqeberha, South

Africa, ⁴Department of Botany, Nelson Mandela University, Gqeberha, South Africa

South Africa's National Biodiversity Assessment (NBA) is the primary tool for monitoring and reporting on the state of biodiversity, with a focus on spatial information and key indicators. The NBA distills information that informs policies and strategies, meets national and international reporting requirements, and helps prioritize limited resources for managing and conserving biodiversity. The three previous versions of the NBA (2004, 2011 and 2018) are in the form of detailed thematic technical reports and a synthesis report, served on a simple, static web page. Selected spatial products from the report are available *via* a dedicated web platform (<http://nba.sanbi.org.za/>). While all methods and data are clearly described in the technical reports, most of the underlying analyses are inaccessible, lacking reproducibility and transparency. This makes iterative updates to indicators or metrics challenging and inefficient, complicates version control, and exacerbates the risk of capacity, knowledge and data loss during staff turnover. To move the assessment process into the information age we aim to develop well documented and reproducible workflows, and to serve the indicators and their accompanying synthesis on an interactive web platform that facilitates uptake. Achieving these aims will deliver efficiency, greater transparency and trust in future NBA products and will strengthen communication and engagement with the content by the many different users of those products. While these visions will not be realized overnight, the skills and systems required to achieve them can be adaptively built towards an improved NBA that better serves the needs of our society.

KEYWORDS

national biodiversity assessment, South Africa, data science, state of biodiversity, convention on biological diversity

1. Introduction

Biodiversity monitoring and reporting at national and global scales plays an important role in meeting the goals of the Rio conventions (Convention on Biological Diversity, United Nations Convention to Combat Desertification, United Nations Framework Convention on Climate Change) and other multilateral environmental agreements (e.g., Sustainable Development Goals, the Ramsar Convention on Wetlands, Convention on the Conservation of Migratory Species of Wild Animals). As signatories to these agreements, parties need to report regularly against a series of indicators that draw on a wide range of biodiversity and environmental observations

(including pressures and drivers). This data-to-knowledge pipeline is undergoing rapid change in the information age, with an explosion of available data and the evolution of new tools for analysis and information delivery (Wilkinson et al., 2016; MacFadyen et al., 2022). Policy, planning and decision-making bodies with a mandate over biodiversity conservation and sustainable use are set to benefit from these changes; if the supporting agencies can adapt their processes and avoid “drowning in data.” This is particularly important—and challenging—in the parts of the world where high biodiversity coincides with pressing social and employment imperatives that require economic development.

The National Biodiversity Assessment (NBA) of South Africa is an iterative body of work that collates and summarizes biodiversity information for both national and global reporting requirements, and informs local to national policies that influence, or are influenced by, biodiversity considerations (Reyers and McGeoch, 2007; Skowno et al., 2019). Many of the components of the NBA are used in systematic conservation planning, which has a clear statutory influence on land and sea use decision making and strategic planning in South Africa (Reyers et al., 2007; Botts et al., 2020; Skowno and Monyeki, 2021).

The NBA is led by the South African National Biodiversity Institute (SANBI) as a core part of their mandate [in terms of the National Environmental Management: Biodiversity Act (Act 10 of 2004)], to monitor and report on the status of the Republic’s biodiversity. SANBI does not work alone; the NBA 2018 was a collaborative effort from more than 470 individuals from approximately 90 institutions. This co-production of knowledge both improves the credibility of the science and promotes the collective ownership and application of the products by the biodiversity science and management communities.

The NBA presents findings on the state of biodiversity (i.e., reports on metrics and indicators), but also includes messaging that aims to explain the implications of the findings and what can be done in response. The goals of the NBA are to (i) inform policy and decision making without being prescriptive, (ii) support planning and prioritization for conservation action, (iii) present indicators for national and international reporting, (iv) report on key issues for educational and fund-raising purposes, and (v) provide a platform for collaboration and capacity building across the biodiversity sector.

At the heart of the most recent NBA lies a series of documents (a Synthesis Report and eight technical reports) with associated appendices and spatial datasets. The Synthesis Report is available as a hardcopy book (Skowno et al., 2019), but all other outputs are digital products served on the NBA website.¹ None of the web content is dynamic or interactive; it is purely a repository of reports and files that can be downloaded for offline use.

In this perspective, we consider the current structure and workflows of the NBA and its delivery, and how they can be improved for greater efficiency, transparency and impact in a world of escalating data availability. By highlighting systems that succeed in effectively delivering robust data to decision makers, and considering NBA user needs, we describe a vision, of improved workflows and an effective, interactive web delivery.

2. Current context

The NBA has been undertaken three times in the last two decades (Driver et al., 2005, 2012; Skowno et al., 2019). Each iteration has seen an increase and broadening in the scope, content and contributor base. All three iterations essentially followed the same approach of collating the best available information on biodiversity, undertaking analyzes, and writing up a series of reports, with key datasets posted to an online spatial data repository on SANBI’s Biodiversity Advisor web platform.² All reports and layers are static, so information and messages contained in them age between version releases, regardless of whether updated information becomes available for certain components. The majority of analyzes that constitute the NBA (e.g., threat status and protection level assessments of taxonomic groups and ecosystem types) were conducted manually using spreadsheets and GIS platforms, generally without prescribed or explicitly documented data and analytical workflows or version control. Staff turnover and methodological advances between releases mean most analyzes have to be conducted from scratch, making the process inefficient and difficult to reproduce (Figure 1).

Global efforts to operationalize the collection of Essential Biodiversity Variables and establish global biodiversity observation networks (Pereira et al., 2013; Han et al., 2017; Turak et al., 2017), combined with parallel initiatives to promote improved data management, stewardship and uptake (Wilkinson et al., 2016; MacFadyen et al., 2022), make it clear that the past NBA workflows are inadequate and will greatly benefit from the incorporation of tools and platforms of the information age.

3. Future plans

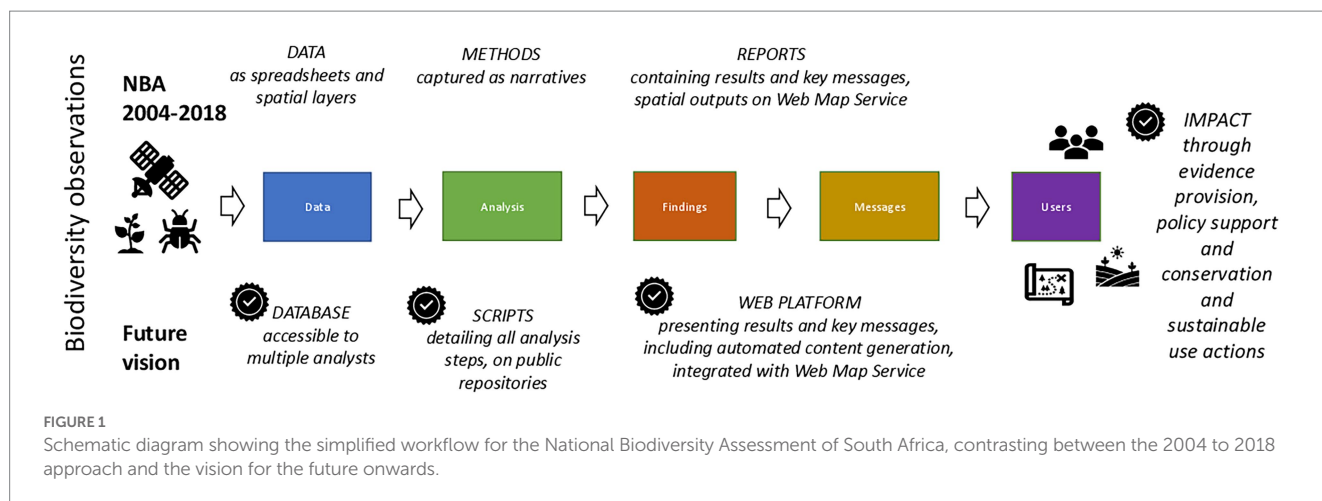
The vision for moving the NBA into the information age is of a ‘living’, interactive, online platform, with clear supporting workflows that can:

- Deliver suitable content for the full range of outputs of the NBA (data, indicators and messages).
- Efficiently accommodate updates to metrics and indicators as new data or methods become available.
- Facilitate easy access to programming scripts and source data, enhancing the reproducibility and transparency of the NBA.

Moving from flat data file-based approaches to relational databases is an important step in making consistent datasets available across a broader user base and to ensure that web-based systems such as SANBI’s Biodiversity Advisor can access information. Using centralized databases containing expert validated data also improves preservation and simplifies version control and integration of other products and services. Data providers often lack the capacity or resources to develop Application Programming Interfaces (APIs) and indexing directly from an institutional database is not typically supported. Building capacity for database design and maintenance is critical, when members of the team have been accustomed to working

¹ <http://nba.sanbi.org.za/>

² <http://biodiversityadvisor.sanbi.org>



with their own, diverse file-based systems. Examples of effective data science solutions applied in environmental or ecological monitoring and assessments, such as the United States Long Term Ecological Research (LTER) Program (Michener et al., 2011; Kaplan et al., 2021) and Ocean Health Index project (Lowndes et al., 2017), speak to the importance of data organization and wrangling, versioning, and the documentation (metadata) aspects of data management. An additional consideration in the context of the NBA is that data are often spatial in nature and comprise raster (grid-based data and imagery) and complex vector data, which require appropriate database types and structures. Bastin et al. (2017) explain some lessons learned from the Digital Observatory for Protected Areas, which include: tracking 'change-only' updates of key spatial datasets, recognizing the value of using different software tools suited to different steps in the workflow, and tips on how to overcome challenges such as legal restrictions of sharing certain datasets.

The challenge in maintaining and updating a centralized database is the interoperability of various data types and formats (e.g., Csv, MS Access, and shape files) from many and varied source datasets. For such integration to be successful, data partners need to agree on a fixed file schema and data standards that enable interoperability [e.g., Atlas of Living Australia and Global Biodiversity Information Facility use the Darwin Core schema for species-related data; the Spatial Data Infrastructure Act (Act 54 of 2003) outlines standards for spatial data infrastructure implementation in South Africa]. Migration tools, such as FME Workbench or Node-Red, can be used to facilitate data integration from disparate flat data files and existing databases. In this way, source data are maintained, and project leads can continue using their preferred systems. Centralizing all NBA datasets will ensure the integration of the necessary data to monitor biodiversity change, increasing accessibility and improving the quality and efficiency of workflows.

Once the data have been queried from the database(s), the next steps in research computing tend to be data preparation, analyzes and presentation. To promote transparency and reproducibility, these should be implemented *via* clearly documented programming scripts, preferably with widely used and open-source data science languages, such as Python and R. Besides their strength of enabling replication of results, such scripted workflows greatly enhance efficiency when iterative adjustments or updates are needed over time—i.e., 'better science in less time' (Lowndes et al., 2017). They also lend themselves to effective version control and collaboration, as the scripted 'recipe'

and input files can be organized within a project structure that is easily shared, within a collaborative team and online once it is finalized (Wilkinson et al., 2016). Developing these data science skills within the NBA team is critical to making inroads to the 'smarter' and more transparent NBA vision. Lowndes et al. (2017) illustrate how the application of data science tools improved the quality, efficiency, reproducibility and accessibility of iterative research outputs from the Ocean Health Index project. In line with this reasoning, many scientific journals are increasingly emphasizing open science standards, with a requirement for authors to have reproducible workflows and their data in accessible online repositories.

The use of databases combined with scripts that largely automate the analyzes, lays the foundation for building 'live' web platforms that deliver information to users interactively and can be updated as new results, methods or data become available. Bastille et al. (2020) gives a technical overview of their workflows for integrated ecosystem assessments, including ideas for creating reproducible data visualizations for various programming languages so that time spent customizing visualizations is reduced. Also noted is how the custom web coding (e.g., in JavaScript, CSS, and HTML) is no longer a barrier, because entire websites can be generated in the scientific programming languages such as R and Python.

Delivering a complete data pipeline for the NBA, from databases through to web platform, will require substantial development of staff capacity, supporting infrastructure, and shifts in thinking and practice, all of which should not be underestimated. Fortunately, these aspirations can be developed and implemented in steps that improve the workflows over time, as demonstrated by Lowndes et al. (2017).

A key feature of the envisaged new NBA format is that the work is broken down into smaller 'modules', each with leads and contributing authors assigned. Each module would typically aim to be published as a peer-reviewed journal article, a GitHub repository containing the code required to replicate the analyzes, and a link to an online data repository with the input data. From these, summary text and figures optimized for delivery to a web platform will be created. The efforts of all NBA authors and contributors, many of whom are not SANBI employees, should be acknowledged and the ability to cite each module will re-enforce trust between stakeholders. It is also crucial that the NBA still meets the needs of its numerous and diverse users, including those that have become accustomed to the current report format. The option to download and print certain summary text and figures must therefore be explored.

Defining a clear plan, managing expectations and communicating clearly about the changes will be essential for updating the format and delivery of the NBA. Since the release of the last NBA in October 2019, SANBI has held several discussions with key users and authors through internal SANBI workshops, presentations at various fora regarding the proposed change in form, and a survey on the discovery and use of NBA 2018. The survey was distributed *via* mailing lists and was completed between

August and November 2022. It received 153 responses from a cross-section of the biodiversity sector in South Africa. See [Supplementary Information](#) and Box 2.

4. Promoting understanding and action through clear messaging

Effective state of biodiversity reporting hinges on efforts to distill and communicate a wide array of findings, spanning multiple levels of biodiversity, realms, pressures, states and responses. The NBA process in South Africa has demonstrated a process of iteratively improving messaging strategies and practice, to promote understanding across the user base ([Maze et al., 2016](#)). For example, the NBA summarizes benefits of biodiversity, with vignettes covering subjects such as pollination services, the traditional medicine economy, biodiversity-related jobs, food security, and spiritual and cultural uses of biodiversity ([SANBI, 2019](#)). The latest NBA includes 19 key messages each comprising a summary of findings, what they mean (the benefit) and what action can be taken. All three elements of the key message (i.e., the 'finding', the 'so what' and the 'call to action') are vital, as they promote understanding and inspire action ([UNEP-WCMC and SANBI, 2022](#)). For example, the finding that 30% of estuaries are impacted by freshwater flow reduction should explain the multiple benefits and requirements of sediment and freshwater flow reaching the coast (i.e., a complex interaction of fish nursery function, beach and dune stability, coastal water quality and

Box 2 NBA use survey

Key findings of the NBA 2018 use/uptake survey (see [Supplementary material](#)) indicate that users need both web-based content and access to detailed digital reports, while hard copy books are not widely sought. Users discovered the NBA products primarily through internet searches, or used a known SANBI information portal such as Biodiversity Advisor (<http://biodiversityadvisor.sanbi.org>) or the NBA's short URL (<http://nba.sanbi.org.za/>), though email distribution lists were noted as important by some respondents. A substantial portion of users still rely on the PDF reports to access NBA information.

Most users wanted access to the maps and spatial data that accompany the NBA reports. Key messages (narratives) and high-level statistics were also sought-after items ([Figure 2](#)). The detailed technical reports were used by the specialist audiences—over 50% of respondents stated they use the terrestrial technical report 'frequently' or 'sometimes', while more specialized reports (e.g., those for the sub-Antarctic or genetic diversity) were used 'rarely' or 'never' by over 70% of respondents. Terrestrial and species datasets are in high demand, followed by freshwater, estuarine, marine and coastal datasets.

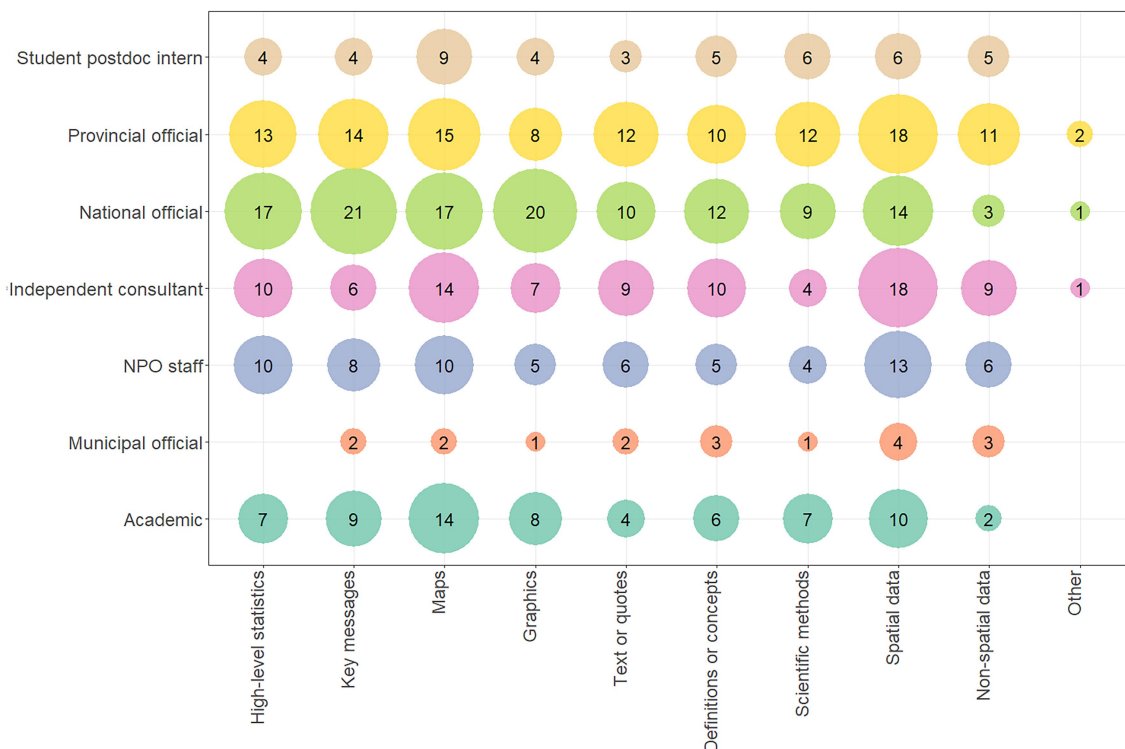


FIGURE 2

Results of the questions 'What type of information did you use from the NBA 2018?' and 'What is your role?'. Respondents could choose multiple responses for the first question. Academics, consultants, non-profit organizations (NPO) and provincial officials used mostly the spatial data and maps, while national officials were the main users of key messages. The high-level statistics were used broadly across all user groups.

other issues), and be followed by recommended actions for freshwater flow strategies and management.

Key messages are often the first thing presented to users, while the underlying detail of indicators and trends are provided as supporting material. NBA 2018's 'Facts, Findings and Key Messages' booklet was a vital product to ensure that the findings were succinctly articulated and acted as a 'summary for policy makers'—a recognized method of ensuring policy makers engage with the scientific findings (IPBES, 2018). This aspect of the NBA needs to be retained and its delivery enhanced. A web platform offers many advantages over the static documents of previous assessments. Through clever design, the most important 'headline' information or succinct messages can be summarized on header pages, with links to the explanations and technical details. In this way, users can access the relevant level of detail they require, from highly summarized messages to fully referenced scientific findings for those wanting to access technical and scientific details.

5. Four key requirements for biodiversity reporting in the information age

5.1. Data science capacity development

Achieving the vision of reproducible NBA workflows requires the development of institutional data science skills. As such skills have not been a priority at SANBI in the past, they need to be built through structured training programs, ongoing mentorship arrangements with key partner institutions and an emphasis of data science skills in the selection of new staff. Traditionally, SANBI staff working on the NBA analyzes have been ecologists or GIS specialists, so it is important to promote the vision that the 'modern analyst' requires some data science skills.

5.2. Enhanced information architecture

SANBI is in the process of redeveloping its Biodiversity Advisor platform, an upgrade that will integrate geospatial, species and ecosystem data, literature and other data made available by SANBI projects such as the NBA and many data partners (Daly and Ranwashe, n.d.). Funding and governance constraints, and the complex nature of the information SANBI serves, necessitate a phased approach to this redevelopment. A modern, web-based NBA requires that these efforts are fast tracked and remains an institutional priority.

5.3. Promoting biodiversity monitoring

A key message in NBA 2018 spoke to South Africa's need for investment in existing and future biodiversity monitoring programs. Without the continuation of monitoring programs and flow of fresh biodiversity observations, the NBA's trend analyzes and iterative computation of key indices would not be possible. Platforms to promote and support focused biodiversity monitoring are essential,

requiring dedicated resources and sustainable funding models within and among relevant institutions.

5.4. Partnerships

SANBI operates within a network of partnerships, acknowledging that it is impossible to achieve its mandate or fulfill its vision and mission without the support of those partnerships. A policy and clear mechanisms are in place to operate in this 'network of partners' model. There is an ongoing need to maintain, strengthen and widen this network and SANBI welcomes discussions with parties who could assist with implementation of the NBA vision outlined here.

6. Conclusion

The NBA is a valuable instrument for communicating the state of South Africa's biodiversity, but there are opportunities to leverage tools of the information age for improved science and more effective product and message delivery. Key improvements include better managed and more accessible data, transparent and reproducible, scripted workflows, effective version control and a user-friendly delivery of findings and messages on a regularly updated 'living' platform. Such changes are going to be necessary to ease and strengthen the uptake of key biodiversity messages and priority actions in a society that lives among an increasingly crowded information flow, supporting improved decision making on the ground and in the water. SANBI welcomes offers of support or partnerships to achieve this vision of taking the NBA into the information age.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Author contributions

CP conceptualized and wrote initial drafts. AS conceptualized, revised and edited, developed figures and analyzes. JC restructured manuscript and revised sections. KS, BD, and LS revised and edited manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors recognize that the 'NBA team' consists of a large number of people, all of whom have been contributing to the discussions about the next NBA's scope and form since NBA 2018 was released. Our sincerest thanks to this entire 'NBA team' that cannot be named here individually. We thank Jeffrey Manuel from SANBI who provided guidance on the design of the article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2023.1107956/full#supplementary-material>

References

- Bastille, K., Hardison, S., deWitt, L., Brown, J., Samhour, J., et al. (2020). Improving the IEA approach using principles of open data science. *Coast. Manag.* 49, 72–89. doi: 10.1080/08920753.2021.1846155
- Bastin, L., Mandrici, A., Battistella, L., and Dubois, G. (2017). Processing conservation indicators with open source tools: lessons learned from the digital Observatory for Protected Areas. *Free Open Source Softw. Geospatial Conf. Proc.* 17:14. doi: 10.7275/R5XK8CQ5
- Botts, E., Skowno, A., Driver, A., Holness, S., Maze, K., Smith, T., et al. (2020). More than just a (red) list: over a decade of using South Africa's threatened ecosystems in policy and practice. *Biol. Conserv.* 246:e108559:108559. doi: 10.1016/j.biocon.2020.108559
- Daly, B., and Ranwashe, F. (n.d.). South Africa's initiative towards an integrated biodiversity data portal. *Front. Ecol. Evol.*
- Driver, A., Maze, K., Lombard, A. T., Nel, J., Rouget, M., Turpie, J. K., et al. (2005). *South African National Spatial Biodiversity Assessment 2004: Priorities for biodiversity conservation in South Africa-Strelitzia 17*. Pretoria: South African National Biodiversity Institute.
- Driver, A., Sink, K. J., Nel, J. N., Holness, S., Van Niekerk, L., Daniels, F., et al. (2012). *National Biodiversity Assessment 2011: An Assessment of South Africa's Biodiversity and Ecosystems*. Synthesis Report.
- Han, X., Josse, C., Young, B. E., Smyth, R. L., Hamilton, H. H., and Bowles-Newark, N. (2017). Monitoring national conservation progress with indicators derived from global and national datasets. *Biol. Conserv.* 213, 325–334. doi: 10.1016/j.biocon.2016.08.023
- IPBES. (2018). *IPBES Guide on the Production of Assessments*. Bonn, Germany: Secretariat of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services Available at: <https://www.ipbes.net/deliverables/2a-assessment-integration>.
- Kaplan, N. E., Baker, K. S., and Karasti, H. (2021). Long live the data! Embedded data management at a long-term ecological research site. *Ecosphere* 12:3493. doi: 10.1002/ecs2.3493
- Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., et al. (2017). Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.* 1. doi: 10.1038/s41559-017-0160
- MacFadyen, S., Allsopp, N., Altwegg, R., Archibald, S., Botha, J., Bradshaw, K., et al. (2022). Drowning in data, thirsty for information and starved for understanding: a biodiversity information hub for cooperative environmental monitoring in South Africa. *Biol. Conserv.* 274:109736. doi: 10.1016/j.biocon.2022.109736
- Maze, K., Barnet, M., Bots, E. A., Stephens, A., Freedman, M., and Guenther, L. (2016). Making the case for biodiversity in South Africa: re-framing biodiversity communications. *Bothalia* 46, 1–8. doi: 10.4102/abc.v46i1.2039
- Michener, W. K., Porter, J., Servilla, M., and Vanderbilt, K. (2011). Long term ecological research and information management. *Ecol. Inform.* 6, 13–24. doi: 10.1016/j.ecoinf.2010.11.005
- Pereira, H. M., Ferrier, S., Walters, M., Geller, G. N., Jongman, R. H. G., Scholes, R. J., et al. (2013). Essential biodiversity variables. *Science* 339, 277–278. doi: 10.1126/science.1229931
- Reyers, B., and McGeoch, M. A. (2007). A biodiversity monitoring framework for South Africa: Progress and directions. *S. Afr. J. Sci.* 103, 295–300.
- Reyers, B., Rouget, M., Jonas, Z., Cowling, R. M., Driver, A., Maze, K., et al. (2007). Developing products for conservation decision-making: lessons from a spatial biodiversity assessment for South Africa. *Divers. Distrib.* 13, 608–619. doi: 10.1111/j.1472-4642.2007.00379.x
- SANBI. (2019). *National Biodiversity Assessment 2018 Supplementary Material: Compendium of Benefits of Biodiversity*. Pretoria, South Africa: South African National Biodiversity Institute. Available at: <http://hdl.handle.net/20.500.12143/8052>.
- Skowno, A. L., and Monyeki, M. S. (2021). South Africa's red list of terrestrial ecosystems (RLEs). *Land* 10, 1–14. doi: 10.3390/land10101048
- Skowno, A. L., Poole, C. J., Raimondo, D. C., Sink, K. J., Van Deventer, H., Van Niekerk, L., et al. (2019). *National Biodiversity Assessment 2018: The Status of South Africa's Ecosystems and Biodiversity-Synthesis Report*. Pretoria, South Africa: South African National Biodiversity Institute Available at: <http://opus.sanbi.org/handle/20.500.12143/6362>.
- Turak, E., Regan, E., and Costello, M. J. (2017). Measuring and reporting biodiversity change. *Biol. Conserv.* 213, 249–251. doi: 10.1016/j.biocon.2017.03.013
- UNEP-WCMC and SANBI. (2022). *Mainstreaming Biodiversity Priorities: A Practical Guide on how to Integrate Spatial Biodiversity Products into National Policy, Planning and Decision-making*. Pretoria, South Africa: South African National Biodiversity Institute Available at: <http://opus.sanbi.org/jspui/handle/20.500.12143/8735>.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). Comment: the FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18



OPEN ACCESS

EDITED BY

Vernon Visser,
University of Cape Town,
South Africa

REVIEWED BY

Cossi Jean Ganglo,
University of Abomey-Calavi,
Benin
Maarten Trekels,
Botanic Garden Meise,
Belgium

*CORRESPONDENCE

Dian Spear
✉ dian.spear@sanparks.org

SPECIALTY SECTION

This article was submitted to
Conservation and Restoration Ecology,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 05 September 2022

ACCEPTED 28 February 2023

PUBLISHED 16 March 2023

CITATION

Spear D, van Wilgen NJ, Rebelo AG and
Botha JM (2023) Collating biodiversity
occurrence data for conservation.
Front. Ecol. Evol. 11:1037282.
doi: 10.3389/fevo.2023.1037282

COPYRIGHT

© 2023 Spear, van Wilgen, Rebelo and Botha.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Collating biodiversity occurrence data for conservation

Dian Spear^{1*}, Nicola J. van Wilgen^{1,2}, Anthony G. Rebelo^{3,4} and
Judith M. Botha⁵

¹Cape Research Centre, Scientific Services, South African National Parks, Cape Town, South Africa,

²Department of Botany and Zoology, Centre for Invasion Biology, Stellenbosch University, Stellenbosch, South Africa, ³Threatened Species Unit, South African National Biodiversity Institute, Cape Town, South Africa, ⁴Pearson Chair of Botany, University of Cape Town, Cape Town, South Africa, ⁵Savanna Research Unit, Scientific Services, South African National Parks, Skukuza, South Africa

Plant and animal checklists, with conservation status information, are fundamental for conservation management. Historical field data, more recent data of digital origin and data-sharing platforms provide useful sources for collating species locality data. However, different biodiversity datasets have different formats and inconsistent naming systems. Additionally, most digital data sources do not provide an easy option for download by protected area. Further, data-entry-ready software is not readily available for conservation organization staff with limited technical skills to collate these heterogeneous data and create distribution maps and checklists for protected areas. The insights presented here are the outcome of conceptualizing a biodiversity information system for South African National Parks. We recognize that a fundamental requirement for achieving better standardization, sharing and use of biodiversity data for conservation is capacity building, internet connectivity, national institutional data management support and collaboration. We focus on some of the issues that need to be considered for capacity building, data standardization and data support. We outline the need for using taxonomic backbones and standardizing biodiversity data and the utility of data from the Global Biodiversity Information Facility and other available sources in this process. Additionally, we make recommendations for the fields needed in relational databases for collating species data that can be used to inform conservation decisions and outline steps that can be taken to enable easier collation of biodiversity data, using South Africa as a case study.

KEYWORDS

biodiversity data, GBIF, iNaturalist, taxonomic backbones, species checklists, conservation, data management

1. Introduction

1.1. The need for collated, standardized species data

Protecting biodiversity requires knowing what plants and animals occur in and around protected areas. As such, the Convention on Biological Diversity (CBD) requires biodiversity monitoring and maintenance of biodiversity information (United Nations, 1992). Despite the inherent value of readily available biodiversity data, biodiversity data management is often overlooked in conservation organizations. Relevant biodiversity data is often inconsistent, incomplete, inaccessible and unusable without relevant metadata (Stephenson et al., 2017). This is not for a lack of available systems, protocols and best practices (see Wilkinson et al., 2016; Hackett et al., 2019). Biodiversity data standards have been produced, such as Darwin Core,

which enables comparable data sharing through standardizing data fields and requires certain ancillary data (Wieczorek et al., 2012). Additionally, the Global Biodiversity Information Facility (GBIF) provides a platform for sharing and accessing biodiversity data shared by others (Gaiji et al., 2013). Yet, conservation organizations fall short in their data management. The insights presented here are the outcome of conceptualizing a biodiversity information system for South African National Parks and provide considerations for building capacity to enhance conservation data management more generally, with the [Supplementary materials](#) detailing particular processes and tools that can be used for collating biodiversity occurrence data.

1.2. Challenges of making biodiversity data accessible

Biodiversity status and trend data are needed for making conservation decisions (Jimenez-Valverde et al., 2010). However, accessing and sharing biodiversity data is a key challenge of implementing the CBD (Chandra and Idrisova, 2011), and even though data collected using public funds should be publically available (Costello, 2009; Chavan and Penev, 2011; Thessen and Patterson, 2011), it often is not. In Africa, capacity and skills to collect, process and curate data is limited in many institutions (Stephenson et al., 2017), with data sharing not being a priority. This is not surprising: biodiversity scientists globally resist sharing biodiversity data (Mandeville et al., 2021), partly because of the time required to make their data sharable (Enke et al., 2012). Despite the multitude of conservation benefits of publishing biodiversity data (Tulloch et al., 2018), historically, there has not been a data-sharing culture among biodiversity researchers (Huang et al., 2012; Costello et al., 2013), with many researchers being reluctant to share data before publishing (Huang et al., 2012) and not knowing how to share their data (Enke et al., 2012). Funders and journals now require data to be made available (see Costello, 2009), and this approach could be extended to institutions including data sharing as criteria for assessing job performance. However, shared data also needs to be standardized and usable (see Costello et al., 2013).

Disparate data are collected by different people (Alves et al., 2018) using different approaches (Berkley et al., 2001; Heidorn, 2008) to answer different questions, and limited human resources are dedicated to curating conservation data (Heidorn, 2008). In South Africa, one limiting area is data cleaning (Coetzer and Hamer, 2019), which is the correction or removal of inaccurate data and standardization of formatting to enable data to be more useful. However, some data management support and capacity building is being provided by GBIF nodes (Parker-Allie et al., 2021). In conservation organizations, there is often limited post-field processing and availability of expertise to guide this, and even where expertise exists, staff turnover and insufficient hand-overs can lead to substantial data loss (see Wiser et al., 2001; Sato et al., 2019). Therefore, although conservation organizations collect large quantities of biodiversity data, e.g., through rangers, these data need to be digitized, standardized according to protocols, checked for consistency through quality control procedures and collated so that they can contribute to decision making in conservation organizations (see [Supplementary material 1](#)). A further complication is obtaining accurate locality data for sensitive species. Locality records for sensitive species are necessarily obscured on data

sharing platforms to protect these species from poaching, and obtaining this locality data can be challenging due to its conditional use and the limited data processing capacity of data holders and collators, such as the South African National Biodiversity Institute (SANBI).

1.3. The wealth of biodiversity data sources

Systematic long-term monitoring programmes are fundamental for assessing population trends (see Kamp et al., 2016), but they are resource intensive. Notably, there are unstructured sources of biodiversity data that can be integrated into monitoring programmes (Kühl et al., 2020; Stephenson and Stengel, 2020), including herbarium and museum specimen records and citizen science data (see [Supplementary material 2](#)). Most of these data are already collated by the GBIF, which enables access to data stored outside its country of origin, which is useful, as a substantial amount of biodiversity collections and data from the global south is in the global north (Tydecks et al., 2018), and research published in journals may be inaccessible to staff of conservation organizations (see Veríssimo et al., 2020). Additionally, there are increasing opportunities for volunteers to curate, identify and categorize data to assist with data analysis (see [Supplementary material 2](#)). However, uploading image and video files to online platforms requires having sufficient bandwidth, and many conservation organizations have slow internet.

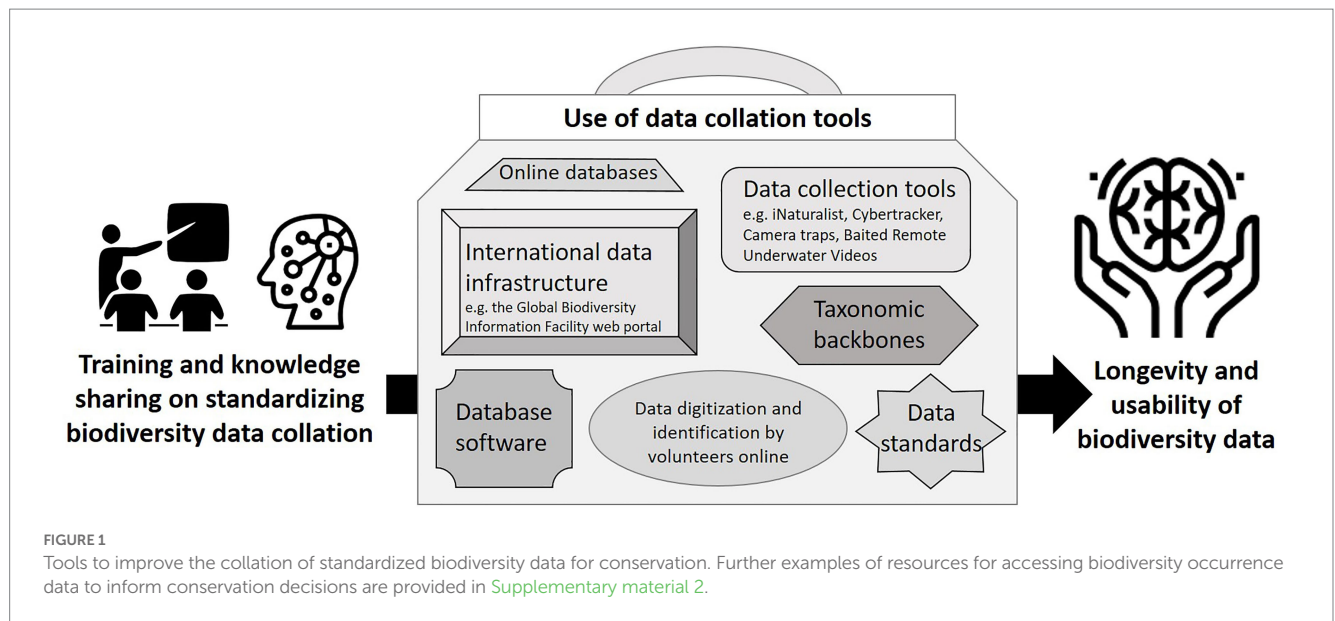
1.4. Tools for managing biodiversity data

Although global biodiversity data systems are advancing (Farley et al., 2018) and many tools are available (see Gadelha et al., 2021; [Figure 1](#); [Supplementary material 2](#)), awareness of and capacity to use these resources is limited, and there are limited data-entry-ready software options available to collate species data across the multiple available data sources, match names to accepted taxonomies and develop species checklists for protected areas. Database software, e.g., BRAHMS and Specify, has been developed for specimen data, but it does not have the flexibility to incorporate other data types, such as iNaturalist and CyberTracker (see Ansell and Koenig, 2011) observation data, which can be collected and curated much faster than traditional specimen-based data (see Kays et al., 2020).

Making species data available to inform protected area management requires standardizing existing biodiversity data (see [Supplementary material 1](#)), incorporating data from global sources, e.g., GBIF, applying a consistent approach to naming species using taxonomic backbones, using a relational database with relevant fields and formats ([Supplementary material 3](#)), implementing data management systems and best practice quality control (Michener et al., 2011; Veiga et al., 2017; Ball-Damerow et al., 2019), and making these data available on a user-friendly platform.

1.5. Taxonomic backbones

Taxonomic backbones are essential foundations of biodiversity information systems (Thomson et al., 2021) and require regular updates. They are exhaustive taxon- and area-specific checklists of



names that include unique identifiers for each name, taxonomic information (family, order, phylum, kingdom etc.), taxonomic status (accepted, synonym, inclusive, misapplied), and the unique identifier and accepted name for each synonym. All species names in a database or checklist should be checked against an authoritative taxonomic backbone (Costello and Wieczorek, 2014). There are many online taxonomic backbones that enable checking for accepted names (Grenié et al., 2022). The Catalogue of Life (CoL; Hobern et al., 2021), which is the primary source of names for the GBIF, incorporates many taxonomic databases. Keeping taxonomies up to date is challenging because of continuous changes in nomenclature requiring historical lists to be updated. These updates are complex, as species can change names, swap names with another species, be combined with another species, or a name could apply to several species that were split since an identification was made (Godfray, 2002). Additionally, the subspecies or variety, which may be of conservation interest, is often not specified or ‘aff.’ (similar potentially new species) or ‘cf.’ (uncertain identification) is associated with a listed name.

In South Africa, SANBI is mandated to maintain national species checklists and has made efforts toward compiling checklists of accepted species names in the country. The most comprehensive of these lists is the annually-updated South African National Plant Checklist (SANPC; SANBI, 2022a), which is part of the taxonomic backbone of the Botanical Database of Southern Africa (BODATSA). Updates to this checklist are guided by a policy that requires that only published name changes are included in the checklist, and updates to the checklist have to be checked by taxonomic experts and approved by a committee, which includes three SANBI taxonomists and six external taxonomists (Victor et al., 2013). The BODATSA is maintained in BRAHMS software, and the published version of the checklist includes the necessary information required to form the basis of a taxonomic backbone for South African plants in a new species database.

One challenge with using the SANPC, and likely many other national checklists, is that there is no accessible species matching tool. This is problematic for the staff of conservation organizations, who often have outdated species lists without authorities and lack the

technical skills to automate name matching. Matching to online lists, such as World Flora Online, is relatively easy with the use of available fuzzy matching tools, e.g., the Taxonomic Name Resolution Service (Boyle et al., 2013) and an R package called WorldFlora (Kindt, 2020). Additionally, GBIF names are easily matched using GBIF’s species lookup tool¹, which would be even more useful if the accepted names of synonyms could be included as part of downloads for matched species names. More manually-intensive methods are available in the absence of a matching tool, including the use of functions in MS Excel (see [Supplementary material 4](#)) and R. Another challenge with the SANPC is that it does not include all alien species that are found in the country: it only includes those species that are considered naturalized, and some nationally-regulated alien species are missing from the list. This necessitates having a standardized way of capturing the scientific names of non-naturalized alien species found on protected area species lists. While GBIF or the Global Register of Introduced and Invasive Species (see [Pagad et al., 2018](#)), which is available on the GBIF platform, could be used, ideally, the SANPC should be updated. It would be useful for SANPC names to be used across organizations in South Africa, and not just herbaria, to standardize national name usage annually and inform more accurate species assessments and conservation prioritization. Using a national checklist can help keep names relevant and useable, and users can submit relevant published updates, changes and errors.

In contrast to the SANPC, there is no comprehensive animal checklist that has been produced by SANBI. The animal names used by SANBI are also overseen by taxonomic experts and a committee. These taxonomists consult specific databases for different groups, e.g., the Amphibian Species of the World database hosted by the American Museum of Natural History for amphibians (Frost, 2021) and The Reptile Database for reptiles (Uetz et al., 2022). There is limited capacity to develop the animal list, particularly given the large variety of species groups, which have been focused on in

¹ <https://www.gbif.org/tools/species-lookup>

isolation, and to date, only some vertebrate and freshwater invertebrate names have been included on SANBI's national animal checklist (see [SANBI, 2022b](#)). In the absence of a comprehensive national list of animals, one interim solution is to use an easily available source of animal names, such as iNaturalist or GBIF, which incorporates the CoL.

1.6. The use of GBIF and iNaturalist to inform conservation

Extensive data sources are easily accessed through the online GBIF data platform ([Gaiji et al., 2013](#)). However, locality data specific to protected areas are not easily downloaded (see [Supplementary material 5](#) for how to access species locality data for a protected area using Geographical Information System software). An option to download locality data by protected area administrative boundaries would be a useful addition to the GBIF data platform, as conservation organization staff often do not know how to access biodiversity data for protected areas. While GBIF includes functionality to filter data by IUCN Red List threat status, it would also be useful to be able to filter by occurrence (indigenous or alien) status per country.

iNaturalist data are available directly from the iNaturalist website and research grade data, which are observations for which two-thirds or more than two iNaturalist identifications concur, are available from the GBIF. iNaturalist provides a wealth of data, which can be used to inform conservation ([Dobson et al., 2020](#)), particularly when iNaturalist is used in bioblitzes and by specialist groups.

iNaturalist data are often organized into projects and places on the iNaturalist website, making it easy to access and curate locality data for protected areas. However, similar to the case for museum and herbarium specimen data, a key caveat is that not all identifications are accurate. The accuracy of iNaturalist records can be assessed by considering who made the identification, as all iNaturalist users are not equal. Specialists can be asked to verify identifications that are not made by taxon experts. Experts can be identified by looking at the 'top identifiers' tab and asking one of the reliable 'top identifiers' to verify the species identification. Unlike Cybertracker or Cmore² data, the images that accompany iNaturalist data make identifications easier to verify. iNaturalist users should also be encouraged to submit diagnostic pictures, such as flowers, seeds and male and female specimens. Further, data users should be aware that because of obscured locality data records may appear at locations where species do not occur. iNaturalist also has a powerful curation tool to assist in rapid and efficient identification of large volumes of records.

2. Discussion

Although species, and lower taxonomic rank, data are a vital aspect of conservation management, these data are not easily available and regularly updated for many protected area networks. Given the parameters that need to be considered for collating, storing and sharing species and locality data for protected areas there are some

minimum requirements that need to be considered for plant and animal locality databases, which should form part of organizational data management plans (see [Donnelly et al., 2010](#); [Strasser et al., 2011](#); [Michener and Jones, 2012](#); [Hampton et al., 2013](#)). Additionally, capacity building, such as the training provided by GBIF's biodiversity data mobilization course³ and training on using OpenRefine and Wikibase software, is vital but under resourced.

2.1. Recommendations for collating biodiversity occurrence data

Species databases should include the wide range of heterogeneous species occurrence data ([Kühl et al., 2020](#), [Stephenson and Stengel, 2020](#); see [Supplementary material 2](#)), and a relational database should be used with some compulsory fields, such as unique taxon numbers (see [Anderson et al., 2020](#)), to allow updates to taxon names and conservation statuses in the taxon table and link to the rest of the database. Although more sophisticated software is available, such a database can be set up in MS Excel with a taxon table in one spreadsheet and occurrence records in another spreadsheet and the use of functions or Power Pivot to link the data between the tables. Additionally, a protocol is required for updating databases regularly to incorporate new data from data sources that are constantly being updated, e.g., searching GBIF for particular time periods and accession dates.

There are several standardized fields that are required to develop a database that will be useful for informing conservation (see [Supplementary material 3](#)). Consistent, standardized and accepted naming systems are needed, and while these should ideally be driven by taxonomists, available taxonomic backbones, such as the GBIF backbone taxonomy, provide a work around where resources are limited. Incorrect taxonomic identifications and inaccurate coordinates are well-known issues. Ideally, detail is needed about the accuracy of the locality information ([GBIF, 2010](#); [Faith et al., 2013](#)) and the source and reliability of the identification to determine the validity of the identification ([Anderson, 2012](#); [Costello and Wicczorek, 2014](#)). For data generated in a conservation organization this would include noting who made the observations and identifications of species. Occurrence status (endemic, indigenous, extralimital, alien) for the protected area and conservation classifications are also needed as these are relevant to conservation management.

2.2. Enabling easier collation of biodiversity occurrence data

The accessibility of biodiversity data for informing conservation in South Africa could be improved through enhanced institutional data management support, inter-organization collaboration and capacity building to enable the use of standardized electronic data capture and data sharing protocols, templates and tools and the use of standardized names for all taxa for species reporting, listing and conservation status assessment.

² <https://www.csir.co.za/cmored>

³ <https://docs.gbif.org/course-data-mobilization/en/>

Increased data collation and sharing by researchers, conservation staff and specimen collectors is possible through the use of iNaturalist, which provides a useful platform for uniform data sharing and access. Additionally, prerequisites and available support for researchers to upload biodiversity occurrence data to GBIF would improve data availability and reduce the data management burden on conservation authorities. An agreed set of backbones for taxa names, e.g., GBIF, that can be used by organizations nationally and national species lookup tools for looking up scientific names would improve consistent name usage nationally. For example, it would be useful if the SANPC could include all alien plants in South Africa to assist with managing and reporting on alien species in a standardized way at the national level. Name matching to the SANPC could also be made easier through an Application Programming Interface being made available for linking the SANPC to existing species matching tools.

Having the functionality to download species occurrence data from global and national platforms using protected area boundaries and the incorporation of GBIF data into local and national biodiversity information systems would improve the accessibility of data to the staff of conservation organizations. The inclusion of a term, such as `protectedAreaName`, in Darwin Core, through engagement with the TDWG (Biodiversity Information Standards), would also be useful. Further, it would be useful for conservation organizations to have easy ways to securely access occurrence data for sensitive species from SANBI to enable effective monitoring of these species, which is currently constrained by limited access to data as a consequence of human resource constraints.

To conclude, biodiversity data needs an overhaul, with a focus on data sharing, to improve data availability and standardization for biodiversity data to become more useful for informing conservation decisions. Incentives, institutional support and capacity building are needed to enhance the sharing of biodiversity occurrence data to data-sharing platforms, such as GBIF, and enable conservation organizations to access this data.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

References

- Alves, C., Castro, J. A., Ribeiro, C., Honrado, J. P., and Lomba, Á. (2018). "Research data management in the field of ecology: an overview," in *International Conference on Dublin Core and Metadata Applications*. 87–94.
- Anderson, R. P. (2012). Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Ann. N. Y. Acad. Sci.* 1260, 66–80. doi: 10.1111/j.1749-6632.2011.06440.x
- Anderson, R. P., Araújo, M. B., Guisan, A., Lobo, J. M., Martínez-Meyer, E., Peterson, A. T., et al. (2020). Optimizing biodiversity informatics to improve information flow, data quality, and utility for science and society. *Front. Biogeogr.* 12:e47839. doi: 10.21425/F5FBG47839
- Ansell, S., and Koenig, J. (2011). CyberTracker: an integral management tool used by rangers in the Djelk indigenous protected area, Central Arnhem Land, Australia. *Ecol. Manag. Restor.* 12, 13–25. doi: 10.1111/j.1442-8903.2011.00575.x
- Ball-Damerow, J. E., Brenskelle, L., Barve, N., Soltis, P. S., Sierwald, P., Bieler, R., et al. (2019). Research applications of primary biodiversity databases in the digital age. *PLoS One* 14:e0215794. doi: 10.1371/journal.pone.0215794
- Berkley, C., Jones, M., Bojilova, J., and Higgins, D. (2001). "Metacat: a schema-independent XML database system," in *Proceedings thirteenth international conference on scientific and statistical database management. SSDBM 2001*. IEEE. 171–179.
- Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzerin, D., Rees, T., et al. (2013). The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinf.* 14, 1–15. doi: 10.1186/1471-2105-14-16
- Chandra, A., and Idrisova, A. (2011). Convention on biological diversity: a review of national challenges and opportunities for implementation. *Biodivers. Conserv.* 20, 3295–3316. doi: 10.1007/s10531-011-0141-x
- Chavan, V., and Penev, L. (2011). The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinf.* 12, 1–12. doi: 10.1186/1471-2105-12-S15-S2
- Coetzer, W., and Hamer, M. (2019). Managing south African biodiversity research data: meeting the challenges of rapidly developing information technology. *S. Afr. J. Sci.* 115, 1–5. doi: 10.17159/sajs.2019/5482
- Costello, M. J. (2009). Motivating online publication of data. *Bioscience* 59, 418–427. doi: 10.1525/bio.2009.59.5.9

Author contributions

DS conceptualized, wrote and revised the manuscript. NvW, AR, and JB conceptualized and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

DS is funded by the JRS Biodiversity Foundation, grant 60916.

Acknowledgments

Our colleagues at SANBI are thanked for explaining their biodiversity data management systems. Three reviewers are thanked for their constructive comments, which improved the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2023.1037282/full#supplementary-material>

- Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q., and Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends Ecol. Evol.* 28, 454–461. doi: 10.1016/j.tree.2013.05.002
- Costello, M. J., and Wiczorek, J. (2014). Best practice for biodiversity data management and publication. *Biol. Conserv.* 173, 68–73. doi: 10.1016/j.biocon.2013.10.018
- Dobson, A. D., Milner-Gulland, E. J., Aebischer, N. J., Beale, C. M., Brozovic, R., Coals, P., et al. (2020). Making messy data work for conservation. *One Earth* 2, 455–465. doi: 10.1016/j.oneear.2020.04.012
- Donnelly, M., Jones, S., and Pattenden-Fail, J. W. (2010). DMP online: the digital curation Centre's web-based tool for creating, maintaining and exporting data management plans. *Int. J. Digit. Curation* 5, 187–193. doi: 10.2218/ijdc.v5i1.152
- Enke, N., Thessen, A., Bach, K., Bendix, J., Seeger, B., and Gemeinholzer, B. (2012). The user's view on biodiversity data sharing—investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecol. Inform.* 11, 25–33. doi: 10.1016/j.ecoinf.2012.03.004
- Faith, D., Collen, B., Ariño, A., Koleff, P. K. P., Guinotte, J., Kerr, J., et al. (2013). Bridging the biodiversity data gaps: recommendations to meet users' data needs. *Biodivers. Inform.* 8, 41–58. doi: 10.17161/bi.v8i2.4126
- Farley, S. S., Dawson, A., Goring, S. J., and Williams, J. W. (2018). Situating ecology as a big-data science: current advances, challenges, and solutions. *Bioscience* 68, 563–576. doi: 10.1093/biosci/biy068
- Frost, D. R. (2021). *Amphibian species of the world: An online reference. Version 6.1 (Date of access)*. (New York, USA: American Museum of Natural History). Available at: <https://amphibiansoftheworld.amnh.org/index.php>
- Gadelha, L. M. Jr., de Siracusa, P. C., Dalcin, E. C., da Silva, L. A. E., Augusto, D. A., Krempser, E., et al. (2021). A survey of biodiversity informatics: concepts, practices, and challenges. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 11:e1394. doi: 10.1002/widm.1394
- Gaiji, S., Chavan, V., Ariño, A. H., Otegui, J., Hobern, D., Sood, R., et al. (2013). Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. *Biodivers. Inform.* 8, 94–172. doi: 10.17161/bi.v8i2.4124
- GBIF. (2010). *GBIF position paper on future directions and recommendations for enhancing fitness-for-use across the GBIF network, version 1.0*. A. W. Hill, J. Otegui, A. H. Ariño and R. P. Guralnick. Copenhagen: Global Biodiversity Information Facility, 25. Available at: <http://www.gbif.org>
- Godfray, H. C. J. (2002). Challenges for taxonomy. The discipline will have to reinvent itself if it is to survive and flourish. *Nature* 417, 17–19. doi: 10.1038/417017a
- Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M. L., Sagouis, A., and Winter, M. (2022). Harmonizing taxon names in biodiversity data: a review of tools, databases and best practices. *Methods Ecol. Evol.* 14, 12–25. doi: 10.1111/2041-210X.13802
- Hackett, R. A., Belitz, M. W., Gilbert, E. E., and Monfils, A. K. (2019). A data management workflow of biodiversity data from the field to data users. *Appl. Plant Sci.* 7:e11310. doi: 10.1002/aps.11310
- Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., et al. (2013). Big data and the future of ecology. *Front. Ecol. Environ.* 11, 156–162. doi: 10.1890/120103
- Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Libr. Trends* 57, 280–299. doi: 10.1353/lib.0.0036
- Hobern, D., Barik, S. K., Christidis, L., T. Garnett, S., Kirk, P., Orrell, T. M., et al. (2021). Towards a global list of accepted species VI: the catalogue of life checklist. *Org. Divers. Evol.* 21, 677–690. doi: 10.1007/s13127-021-00516-w
- Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., et al. (2012). Willing or unwilling to share primary biodiversity data: results and implications of an international survey. *Conserv. Lett.* 5, 399–406. doi: 10.1111/j.1755-263X.2012.00259.x
- Jimenez-Valverde, A., Lira-Noriega, A., Peterson, A. T., and Soberon, J. (2010). Marshalling existing biodiversity data to evaluate biodiversity status and trends in planning exercises. *Ecol. Res.* 25, 947–957. doi: 10.1007/s11284-010-0753-8
- Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T., and Donald, P. F. (2016). Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Divers. Distrib.* 22, 1024–1035. doi: 10.1111/ddi.12463
- Kays, R., McShea, W. J., and Wikelski, M. (2020). Born-digital biodiversity data: millions and billions. *Divers. Distrib.* 26, 644–648. doi: 10.1111/ddi.12993
- Kindt, R. (2020). WorldFlora: an R package for exact and fuzzy matching of plant names against the world Flora online taxonomic backbone data. *Appl. Plant Sci.* 8:e11388. doi: 10.1002/aps.11388
- Kühl, H. S., Bowler, D. E., Bösch, L., Bruelheide, H., Dauber, J., Eichenberg, D., et al. (2020). Effective biodiversity monitoring needs a culture of integration. *One Earth* 3, 462–474. doi: 10.1016/j.oneear.2020.09.010
- Mandeville, C. P., Koch, W., Nilsen, E. B., and Finstad, A. G. (2021). Open data practices among users of primary biodiversity data. *Bioscience* 71, 1128–1147. doi: 10.1093/biosci/biab072
- Michener, W. K., and Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends Ecol. Evol.* 27, 85–93. doi: 10.1016/j.tree.2011.11.016
- Michener, W. K., Porter, J., Servilla, M., and Vanderbilt, K. (2011). Long term ecological research and information management. *Ecol. Inform.* 6, 13–24. doi: 10.1016/j.ecoinf.2010.11.005
- Pagad, S., Genovesi, P., Carnevali, L., Schigel, D., and McGeoch, M. A. (2018). Introducing the global register of introduced and invasive species. *Sci. Data* 5, 1–12. doi: 10.1038/sdata.2017.202
- Parker-Allie, F., Pando, F., Telenius, A., Ganglo, J. C., Vélez, D., Gibbons, M. J., et al. (2021). Towards a post-graduate level curriculum for biodiversity informatics. Perspectives from the global biodiversity information facility (GBIF) community. *Biodivers. Data J.* 9:e68010. doi: 10.3897/BDJ.9.e68010SANBI
- SANBI. (2022a). *Official yearly release of South African National Plant Checklist for 2022*. South African National Biodiversity Institute. Available at: <https://hdl.handle.net/20.500.12143/6880> (Accessed June 7, 2022).
- SANBI. (2022b). *South African animal checklist for reptiles, birds, frogs, freshwater fish, dobsonflies, caddisflies, Mollusca, long-tongue flies, stoneflies, decapods, amphipods and mayflies*. South African National Biodiversity Institute. Available at: <https://hdl.handle.net/20.500.12143/8511> (Accessed June 7, 2022).
- Sato, C., Westgate, M. J., Barton, P., Foster, C., O'Loughlin, L. S., Pierson, J. C., et al. (2019). The use and utility of surrogates in biodiversity monitoring programmes. *J. Appl. Ecol.* 56, 1304–1310. doi: 10.1111/1365-2664.13366
- Stephenson, P. J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagana, M., Höft, R., et al. (2017). Unblocking the flow of biodiversity data for decision-making in Africa. *Biol. Conserv.* 213, 335–340. doi: 10.1016/j.biocon.2016.09.003
- Stephenson, P. J., and Stengel, C. (2020). An inventory of biodiversity data sources for conservation monitoring. *PLoS One* 15:e0242923. doi: 10.1371/journal.pone.0242923
- Strasser, C., Cook, R., Michener, W., Budden, A., and Koskela, R. (2011). “Promoting data stewardship through best practices,” in *Proceedings of the environmental information management conference*. 126–131.
- Thessen, A. E., and Patterson, D. J. (2011). Data issues in the life sciences. *ZooKeys* 150, 15–51. doi: 10.3897/zookeys.150.1766
- Thomson, S. A., Thiele, K., Conix, S., Christidis, L., Costello, M. J., Hobern, D., et al. (2021). Towards a global list of accepted species II. Consequences of inadequate taxonomic list governance. *Org. Divers. Evol.* 21, 623–630. doi: 10.1007/s13127-021-00518-8
- Tulloch, A. I., Auerbach, N., Avery-Gomm, S., Bayraktarov, E., Butt, N., Dickman, C. R., et al. (2018). A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nat. Ecol. Evol.* 2, 1209–1217. doi: 10.1038/s41559-018-0608-1
- Tydecks, L., Jeschke, J. M., Wolf, M., Singer, G., and Tockner, K. (2018). Spatial and topical imbalances in biodiversity research. *PLoS One* 13:e0199327. doi: 10.1371/journal.pone.0199327
- Uetz, P., Freed, P., Aguilar, R., and Hošek, J. (eds.) (2022). The reptile database. Available at: <http://www.reptile-database.org>
- United Nations (1992). *Convention on biological diversity*. United Nations, New York.
- Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., et al. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLoS One* 12:e0178731. doi: 10.1371/journal.pone.0178731
- Verissimo, D., Pienkowski, T., Arias, M., Cugnière, L., Doughty, H., Hazenbosch, M., et al. (2020). Ethical publishing in biodiversity conservation science. *Conserv. Soc.* 18, 220–225. doi: 10.4103/cs.cs_19_56
- Victor, J., Klopper, R. R., Winter, P. J. D., and Hamer, M. (2013). *South African National Plant Checklist Policy*. Pretoria: South African National Biodiversity Institute. Available at: <http://hdl.handle.net/20.500.12143/6880> (Accessed March, 2013).
- Wiczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., et al. (2012). Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* 7:e29715. doi: 10.1371/journal.pone.0029715
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. data* 3, 160018–160019. doi: 10.1038/sdata.2016.18
- Wiser, S. K., Bellingham, P. J., and Burrows, L. E. (2001). Managing biodiversity information: development of New Zealand's National Vegetation Survey databank. *New Zeal. J. Ecol.* 25, 1–17. Available at: <http://www.jstor.org/stable/24055293>



OPEN ACCESS

EDITED BY

Cang Hui,
Stellenbosch University, South Africa

REVIEWED BY

Jeyaraj Antony Johnson,
Wildlife Institute of India, India
Rui Manuel Vitor Cortes,
University of Trás-os-Montes and Alto
Douro, Portugal

*CORRESPONDENCE

Mohammed Kajee,
✉ kxjmoh007@myuct.ac.za

SPECIALTY SECTION

This article was submitted to
Conservation and Restoration Ecology,
a section of the journal
Frontiers in Environmental Science

RECEIVED 12 December 2022

ACCEPTED 22 February 2023

PUBLISHED 22 March 2023

CITATION

Kajee M, Henry DAW, Dallas HF,
Griffiths CL, Pegg J, Van der Colff D,
Impson D, Chakona A, Raimondo DC,
Job NM, Paxton BR, Jordaan MS, Bills R,
Roux F, Zengeya TA, Hoffman A,
Rivers-Moore N and Shelton JM (2023),
How the Freshwater Biodiversity
Information System (FBIS) is supporting
national freshwater fish conservation
decisions in South Africa.
Front. Environ. Sci. 11:1122223.
doi: 10.3389/fenvs.2023.1122223

COPYRIGHT

© 2023 Kajee, Henry, Dallas, Griffiths,
Pegg, Van der Colff, Impson, Chakona,
Raimondo, Job, Paxton, Jordaan, Bills,
Roux, Zengeya, Hoffman, Rivers-Moore
and Shelton. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

How the Freshwater Biodiversity Information System (FBIS) is supporting national freshwater fish conservation decisions in South Africa

Mohammed Kajee^{1,2,3*}, Dominic A. W. Henry^{4,5}, Helen F. Dallas^{2,6},
Charles L. Griffiths¹, Josephine Pegg^{3,7}, Dewidine Van der Colff⁸,
Dean Impson², Albert Chakona⁹, Domitilla C. Raimondo⁸,
Nancy M. Job⁸, Bruce R. Paxton², Martine S. Jordaan^{9,10,11},
Roger Bills⁹, Francois Roux¹², Tsungai A. Zengeya^{8,13},
Andre Hoffman¹², Nick Rivers-Moore² and Jeremy M. Shelton²

¹Department of Biological Sciences, University of Cape Town, Cape Town, South Africa, ²Freshwater Research Centre, Cape Town, South Africa, ³DSI/NRF Research Chair in Inland Fisheries and Freshwater Ecology, South African Institute for Aquatic Biodiversity (SAIAB), Makhanda, South Africa, ⁴Endangered Wildlife Trust, Johannesburg, South Africa, ⁵Centre for Statistics in Ecology, Environment and Conservation, University of Cape Town, Cape Town, South Africa, ⁶Faculty of Science, Formerly Port Elizabeth, Nelson Mandela University, Gqeberha, South Africa, ⁷Department of Ichthyology and Fisheries Science, Rhodes University, Makhanda, South Africa, ⁸Kirstenbosch Research Centre, South African National Biodiversity Institute, Cape Town, South Africa, ⁹NRF-South African Institute for Aquatic Biodiversity (NRF-SAIAB), Makhanda, South Africa, ¹⁰CapeNature Biodiversity Capabilities Unit, Stellenbosch, South Africa, ¹¹Center of Excellence for Invasion Biology, CapeNature Biodiversity Capabilities Unit, Stellenbosch, South Africa, ¹²Scientific Services, Mpumalanga Tourism and Parks Agency, Mbombela, South Africa, ¹³Centre for Invasion Biology, Department of Zoology and Entomology, University of Pretoria, Pretoria, South Africa

In South Africa, anthropogenic pressures such as water over-abstraction, invasive species impacts, land-use change, pollution, and climate change have caused widespread deterioration of the health of river ecosystems. This comes at great cost to both people and biodiversity, with freshwater fishes ranked as the country's most threatened species group. Effective conservation and management of South Africa's freshwater ecosystems requires access to reliable and comprehensive biodiversity data. Despite the existence of a wealth of freshwater biodiversity data, access to these data has been limited. The Freshwater Biodiversity Information System (FBIS) was built to address this knowledge gap by developing an intuitive, accessible and reliable platform for freshwater biodiversity data in South Africa. The FBIS hosts high quality, high accuracy biodiversity data that are freely available to a wide range of stakeholders, including researchers, conservation practitioners and policymakers. We describe how the system is being used to provide freshwater fish data to a national conservation decision-support tool—The Department of Forestry, Fisheries, and the Environment (DFFE) National Environmental Screening Tool (NEST). The NEST uses empirical and modelled biodiversity data to guide Environmental Impact Assessment Practitioners in conducting environmental assessments of proposed developments. Occurrence records for 34 threatened freshwater fishes occurring in South Africa were extracted from the FBIS and verified by taxon specialists, resulting in 6 660 records being used to generate modelled and empirical national

distribution (or sensitivity) layers. This represents the first inclusion of freshwater biodiversity data in the NEST, and future iterations of the tool will incorporate additional freshwater taxa. This case study demonstrates how the FBIS fills a pivotal role in the data-to-decision pipeline through supporting data-driven conservation and management decisions at a national level.

KEYWORDS

freshwater fish, South Africa, FBIS, screening tool, conservation, SDM (species distribution model), threatened species, decision-making

Introduction

Freshwater ecosystems are among the most biodiverse habitats on earth, covering less than 1% of the planet's total surface area, but accounting for nearly 25% of all vertebrates and more than 50% of all fishes (Hughes et al., 2021; Fricke et al., 2022). However, despite providing essential ecosystem services, protection of freshwater habitats and their associated biodiversity remain a low priority for policymakers when developing protected areas and legislation (Hughes et al., 2021). Recent global studies (Abell et al., 2007; Adams et al., 2015; Hughes et al., 2021; Williams-Subiza and Epele, 2021) have highlighted significant gaps in protected area networks for freshwater systems, and freshwater biodiversity is declining twice as fast as in marine and terrestrial ecosystems (Grooten and Almond, 2018), with nearly a third of freshwater fishes now threatened with extinction (WWF, 2020; IUCN Red List, 2020; Hughes et al., 2021).

In South Africa, freshwater fishes are recognised as the country's most threatened species group (Skowno et al., 2019). Of the 106 formally described native fish species, 27 are threatened with extinction (classified as Vulnerable, Endangered or Critically Endangered according to the IUCN), with at least eight reported to be in decline over the last decade (Chakona et al., 2022). Ongoing taxonomic revisions indicate that several endemic taxa are genetically distinct and have narrower distribution ranges than previously thought (Chakona et al., 2022). As such, some threatened endemic 'species suites' are likely to be split taxonomically, resulting in individual species being more vulnerable to extinction and raising the total number of threatened taxa in the country (Chakona et al., 2022). The key anthropogenic pressures impacting freshwater fishes in South Africa are land-use change (O'Brien et al., 2019; Chakona et al., 2020a), pollution (Wepener et al., 2011; Horak et al., 2021), excessive abstraction of water (Dallas and Rivers-Moore, 2014; Cerrilla et al., 2022; Evans et al., 2022), spread of invasive species (Ellender and Weyl, 2014; Weyl et al., 2020; Zengeya et al., 2020; Cerrilla et al., 2022) and climate change (Dallas and Rivers-Moore, 2014; Ziervogel et al., 2014). Consequently, there is an urgent need to develop current, data-driven conservation plans and policies to guide effective management and protection of freshwater habitats in South Africa, to safeguard their unique biodiversity and to sustain their essential ecosystem services.

South Africa's protected area network (Republic of South Africa, 2004a; Republic of South Africa, 2004b; Chakona et al., 2022), covers less than 10% of South Africa's total land area (Skelton et al., 1995; Russell, 2011; Skowno et al., 2019). Whilst protected areas generally provide some level of protection for freshwater fishes, few protect entire catchments (Skelton et al., 1995; Acreman et al., 2019; Jordaan

et al., 2020), which is problematic given the linear nature of river ecosystems (Jordaan et al., 2020). Of the country's rivers, only 18% are regarded as Well Protected and 12% as Moderately Protected, with the remainder of the being classified as Not or Poorly Protected (Department of Environmental Affairs, 2016). Even rivers considered as Well or Moderately Protected are often not in a pristine or healthy condition. For example, Kleynhans (2000) and Nel et al. (2009) found that almost half of the river systems falling within protected areas in South Africa are already degraded upstream as a result of human activities (Kleynhans, 2000; Nel et al., 2009). Russell (2011) found that of the 19 national parks managed by South African National Parks, only 13 included habitat for freshwater fishes, with protection often being an unintended by-product of the targeted protection of threatened terrestrial plants and mammals. Moreover, protection of freshwater fishes within protected areas is compromised by human-linked impacts such as climate change, invasive species and habitat degradation impacts further upstream (Impson et al., 2002; Abell et al., 2007), with 84% freshwater fish regarded as under-protected (Jordaan et al., 2020). As such, South Africa's current network of protected area network does not adequately protect freshwater fishes (Nel et al., 2004; Abell et al., 2007; Jordaan et al., 2020).

Having recognised the limitations of the current protected areas network, expansion of these areas in South Africa is supported through the country's commitment to the Convention on Biological Diversity (CBD) in The National Protected Area Expansion Strategy (NPAES, 2016; Republic of South Africa, 2010; Department of Environmental Affairs, 2016). In conjunction with the NPAES, the Government of South Africa also published amendments (24 (5) (a) and 24 (5) (h)) to the National Environmental Management Act, 1998 (Act 107 of 1998; Republic of South Africa, 1998), stipulating that future developments will need to be guided by an objective Environmental Impact Assessments (EIA) process that takes into consideration presence or absence of threatened taxa (South African National Biodiversity Institute (SANBI), 2021). As such, the Department of Forestry, Fisheries, and the Environment (DFFE) developed a National Environmental Screening Tool (screening.environment.gov.za; SANBI, 2021)—hereafter, the 'NEST'. The NEST is a national web-enabled application that allows applicants seeking environmental authorisation for development to screen their proposed development site for any environmental sensitivities, for example, the presence of threatened species (DFFE, 2021). The NEST uses empirical and modelled biodiversity data (packaged as 'sensitivity layers') to guide EIAs of proposed developments—a process that has in the past been criticized for not being sufficiently transparent or robust (i.e., data-

driven). A brief description of the NEST and an explanation of the species distribution layers are outlined below.

The NEST consists of theme-specific spatial datasets, which have been assigned various sensitivity levels (SANBI, 2021), allowing for pre-screening of the proposed development footprint. The NEST assesses the likelihood that a proposed development will have a negative impact on either the environment or any threatened species that may occur in the same area. One of the main components of the NEST is the Plant and Animal Sensitivity Layers, which uses a four-tiered sensitivity rating system to identify and classify habitat where threatened species occur (SANBI, 2021). These layers are briefly described below (SANBI, 2021).

- **Very high layer**—Habitat for highly range-restricted threatened taxa that has an extent of occurrences of less than or equal to 10 km². For each taxon, critical habitat is manually mapped at a fine scale by taxon experts. Combined data for all taxa are combined into a single spatial layer: Very high sensitivity layer.
- **High layer**—The current distribution of threatened taxa are included in the high sensitivity level by developing spatial polygons around known recent occurrence records (defined by the NEST as records collected since the year 2002 with reasonably high spatial accuracy). Combined data for all taxa are combined into a single spatial layer: High sensitivity layer.
- **Medium layer**—Species distribution models (SDMs), which use species occurrence records combined with multiple environmental variables to quantify and predict areas of suitable habitat, were used to model suitable habitat areas where threatened species are expected to occur. Combined data for all taxa are combined into a single spatial layer: Medium sensitivity layer.
- **Low sensitivity layer**—Areas where no threatened taxa are currently known or expected to occur.

Whilst the NEST and associated Species Assessment Guidelines (SANBI, 2021) do currently include aquatic habitat sensitivity layers, they do not currently include any freshwater species-specific layers (SANBI, 2021). Developing sensitivity layers for freshwater taxa requires access to reliable and comprehensive freshwater biodiversity data. However, despite a wealth of current and historic biodiversity data existing for South Africa's freshwater ecosystems, access to these data has been limited by the lack of a dedicated and resourced freshwater information system (Dallas et al., 2022).

Several databases have been developed in South Africa in the past for collating and preserving freshwater biodiversity data at both national and provincial levels (Dallas et al., 2022). However, these databases are generally difficult to access, use different data formats and standards, and are not always maintained due to limited resources and funds (Dallas et al., 2022)—a problem also experienced in both terrestrial and marine ecosystems. Some South African organisations (such as the South African Institute for Aquatic Biodiversity and the Albany Museum) publish their freshwater biodiversity data to the Global Biodiversity Information Facility (GBIF), but the data currently available on GBIF under-represent what is actually available for South African freshwater fish, and require substantial time and effort to

clean, format and analyse. Whilst these platforms all contribute valuable data, none of them adequately meet the needs for national freshwater conservation and decision-making in a South Africa (Dallas et al., 2022). However, such information is critical for informing national and international biodiversity assessments, measuring the impact of anthropogenic activities (such as climate change), and enhancing our ability to make informed policy, management, and conservation decisions at both provincial and national levels (Dallas et al., 2022). As such, access to comprehensive and reliable freshwater biodiversity data is imperative and will help to safeguard critical freshwater biodiversity from anthropogenic threats, and will allow freshwater resources in South Africa to be sustainably used and managed.

The Freshwater Biodiversity Information System (FBIS; freshwaterbiodiversity.org), an open-access, online platform for freshwater biodiversity data in South Africa launched in 2017, was developed by the Freshwater Research Centre (FRC) in partnership with SANBI and Kartoza to bridge this data gap by improving access to comprehensive and reliable freshwater biodiversity data. The FBIS is a powerful, open-access system for hosting, standardising, analysing and serving freshwater biodiversity data for South Africa. As such, the FBIS functions as a repository for South African freshwater data, and has been populated with data from key available sources including data mobilised through manual literature searches, and data pulled in *via* links with existing online platforms such as GBIF (Dallas et al., 2022). The FBIS represents the first comprehensive, accessible, national-level resource for freshwater biodiversity data records in South Africa, and thus provides an opportunity for national-level freshwater biodiversity data to be utilised by researchers, policymakers, and conservation practitioners in real time (Dallas et al., 2022).

Here we describe how the FBIS was recently used to provide freshwater fish data for informing a national-level conservation decision-support tool (the NEST), and we evaluate the role of the FBIS in the data-to-decision pipeline. Specifically, we examine how data extracted from the FBIS were used to develop national freshwater fish sensitivity layers—core components of the NEST—thereby allowing for the first inclusion of freshwater species spatial coverage in the tool, and the potential for improved freshwater biodiversity conservation at a national scale. This case-study demonstrates the importance of collecting and collating comprehensive, high quality biodiversity data sets, and being able to synthesize these data and make them accessible for analysis and uptake and use in conservation planning and decision making at a national scale. We also highlight the importance of expert consultation and multi-disciplinary stakeholder engagement during various stages of this process.

Methods

Data source

The development of the NEST sensitivity layers for freshwater fish was a collaborative effort that included

TABLE 1 List of threatened (Vulnerable, Endangered and Critically Endangered) South African freshwater fish taxa included in the NEST analysis. Conservation status and the number of occurrence records sourced from the Freshwater Biodiversity Information System (FBIS) pre and post 2002 are shown. Taxa #1–27 are formally-describes species, while taxa #28–34 represent recognised genetically-distinct lineages. Threat categories are: VU = Vulnerable; EN = Endangered; CR = Critically Endangered.

#	Scientific name	Common name	IUCN Conservation Status	SANBI Conservation Status	Number of Records Pre 2002	Number of Records Post 2002
1	<i>Austroglanis barnardi</i>	Barnard's rock catfish	EN	EN	66	34
2	<i>Chetia brevis</i>	Orange-fringed largemouth	EN	EN	17	27
3	<i>Chiloglanis bifurcus</i>	Incomati suckermouth	CR	CR	81	45
4	<i>Chiloglanis emarginatus</i>	Pongolo suckermouth	VU	VU	26	27
5	<i>Ctenopoma multispine</i>	Many-spined climbing perch	LC	VU	21	3
6	<i>Enteromius gurneyi</i>	Redtail barb	VU	VU	73	14
7	<i>Enteromius treurensis</i>	Treur River barb	CR	CR	20	10
8	<i>Labeo rubromaculatus</i>	Tugela labeo	VU	VU	29	16
9	<i>Labeo seeberi</i>	Clanwilliam sandfish	EN	EN	42	230
10	<i>Marcusenius caudisquamatus</i>	Natal bulldog	EN	EN	1	8
11	<i>Oreochromis mossambicus</i>	Mozambique tilapia	VU	VU	1,581	1,292
12	<i>Pseudobarbus afer</i>	Eastern Cape redbfin	EN	EN	93	95
13	<i>Pseudobarbus asper</i>	Smallscale redbfin	VU	VU	294	54
14	<i>Pseudobarbus burchelli</i>	Barrydale redbfin	CR	CR	18	32
15	<i>Pseudobarbus burgi</i>	Berg River redbfin	EN	EN	84	109
16	<i>Pseudobarbus capensis</i>	Berg-Breede River whitefish	EN	EN	45	24
17	<i>Pseudobarbus erubescens</i>	Twee River redbfin	CR	CR	122	20
18	<i>Pseudobarbus phlegethon</i>	Fiery redbfin	EN	EN	118	72
19	<i>Pseudobarbus quathlambae</i>	Maloti minnow	EN	EN	94	59
20	<i>Pseudobarbus senticeps</i>	Krom River redbfin	CR	CR	31	4
21	<i>Pseudobarbus skeltoni</i>	Giant redbfin	EN	EN	2	36
22	<i>Pseudobarbus swartzi</i>	Gamtoos redbfin	VU	VU	130	35
23	<i>Pseudobarbus trevelyani</i>	Border barb	EN	EN	23	10
24	<i>Pseudobarbus verlorenei</i>	Verlorenvlei redbfin	EN	EN	29	24
25	<i>Sandelia bainsii</i>	Eastern Cape rocky	EN	EN	83	71
26	<i>Serranochromis meridianus</i>	Lowveld largemouth	EN	EN	75	10
27	<i>Silhouettea sibayi</i>	Sibayi Goby	EN	EN	10	2
28	<i>Kneria</i> sp. nov. south africa		EN	EN	29	19
29	<i>Marcusenius</i> sp. nov. kosi		EN	EN	0	4

(Continued on following page)

TABLE 1 (Continued) List of threatened (Vulnerable, Endangered and Critically Endangered) South African freshwater fish taxa included in the NEST analysis. Conservation status and the number of occurrence records sourced from the Freshwater Biodiversity Information System (FBIS) pre and post 2002 are shown. Taxa #1–27 are formally-describes species, while taxa #28–34 represent recognised genetically-distinct lineages. Threat categories are: VU = Vulnerable; EN = Endangered; CR = Critically Endangered.

#	Scientific name	Common name	IUCN Conservation Status	SANBI Conservation Status	Number of Records Pre 2002	Number of Records Post 2002
30	<i>Pseudobarbus</i> sp. nov. breede		N/A	VU	216	324
31	<i>Pseudobarbus</i> sp. nov. doring		CR	CR	26	19
32	<i>Pseudobarbus</i> sp. nov. heuningnes		EN	EN	7	48
33	<i>Pseudobarbus</i> sp. nov. keiskamma		EN	EN	38	9
34	<i>Pseudobarbus</i> sp. nov. keurbooms		N/A	EN	19	16

contributions from several organisations: FRC, SANBI, NRF-South African Institute for Aquatic Biodiversity (NRF-SAIAB), CapeNature, Endangered Wildlife Trust (EWT), DFFE, Mpumalanga Tourism and Parks Agency (MTPA), Free State Department: Economic, Small Business Development, Tourism and Environmental Affairs, and University of Cape Town (UCT), as well as multiple other individuals and organisations *via* SANBI's National Freshwater Fish Observation Group (see Supplementary S3 and S4). Occurrence records for 34 threatened (Vulnerable, Endangered and Critically Endangered classified as per the IUCN) freshwater fishes occurring in South Africa (Table 1) were downloaded on 14 September 2022 from the FBIS online database (FBIS, 2022). These 34 fish taxa included 27 formally described species and seven lineages, yet to be formally described (Chakona et al., 2022). Formally described species have been assessed both globally (on the IUCN Red List) and nationally (Red List of South African Species; SANBI, 2022), whereas undescribed lineages have only been assessed nationally (SANBI, 2022). Only primary, secondary, and catadromous threatened freshwater fishes were included, while marine peripheral/sporadic and primary marine fishes were excluded.

Data cleaning and validation

A series of online workshops were held with freshwater fish taxon specialists from across South Africa during 2021 to verify the quality and comprehensiveness of the data downloaded from the FBIS (see Supplementary Table S1). Relevant experts were identified for each threatened fish taxon, based on prior involvement in species Red-Listing assessments, and consulted. Occurrence records for each fish taxon from the FBIS were thoroughly scrutinized, erroneous records were removed, and missing data sets identified and added. Resultant occurrence data sets were used to generate updated distribution maps for each taxon, and these were sent to all specialists for approval prior to further analysis.

Data analysis

All data cleaning, processing and analyses were conducted using R Software (R Core Team, 2020; Version 3.5.0). Data visualizations and final maps were produced using R Software (R Core Team, 2020; Version 3.5.0) and ArcGIS Pro (Esri Inc, 2022). Cleaned and validated occurrence records were used to produce different sensitivity layers, for inclusion into the NEST. The protocols for developing the different sensitivity layers followed those described by DFFE (2021). A brief description of how each sensitivity layer was developed is outlined below. Final data and spatial layers were presented to taxon specialists for review before being combined and submitted for inclusion in the NEST.

Very high sensitivity layers

The 'very high' sensitivity category only applied to freshwater fishes that were assessed as Critically Endangered. Given that freshwater fish occur in linear river systems, the criteria used to develop the Very high sensitivity layer were adapted as follows: *All historic (pre-2002) and current (post-2002) occurrence records of freshwater fish taxa that are categorised as Critically Endangered*. As such, all known, valid, historic occurrence records were used to build the very high sensitivity layer. Occurrences were intersected with the FEPA Sub-Quaternary Catchment layer (Nel, 2011; SANBI, 2011) to create catchment-specific occurrence polygons, indicating the presence of a Critically Endangered freshwater fish in that catchment.

High sensitivity layers

The high sensitivity category only applied to freshwater fishes that were assessed as Vulnerable or Endangered. The 'high' sensitivity layer is comprised of all valid, post-2002 occurrence records (SANBI, 2021). Occurrence records for each freshwater fish (i.e., the assumed, current distribution of the species) were plotted in R (R Core Team, 2020; Version 3.5.0). Occurrence data

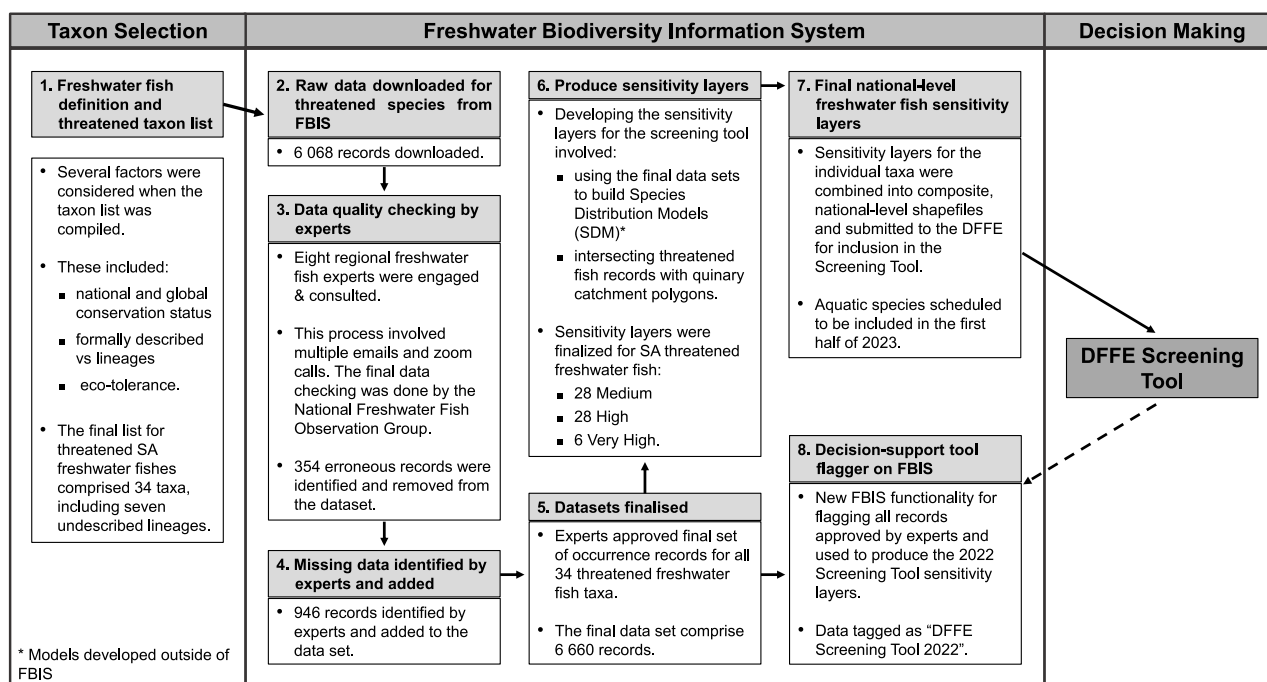


FIGURE 1

Schematic illustrating the data flow from the Freshwater Biodiversity Information System (FBIS) into the Department of Forestry, Fisheries and Environment (DFFE) Screening Tool and for developing the necessary functionality to improve the ease of future such projects.

were then intersected with the River Freshwater Ecosystem Priority Areas (FEPA) Sub-Quaternary Catchment layer (Nel, 2011; SANBI, 2011) to create catchment-specific occurrence polygons, indicating the recent presence of a Vulnerable or Endangered freshwater fish in that catchment.

Medium sensitivity layers

Taxa that qualify for inclusion in the 'medium' sensitivity layer were categorised as either Vulnerable or Endangered. For these taxa, species distribution models (SDMs) were used to generate predictive geographic ranges. All valid occurrence records for each freshwater fish were used to independently develop a unique, accurate SDM for each taxon using a Bayesian additive regression trees (BART) algorithm via functions from the *embarcadero* R package (Carlson, 2020). A comprehensive suite of environmental and hydrological variables was used to generate these SDMs (see Supplementary Table S2). Modelled distributions were converted to a binary output (presence/absence) for each species using a threshold that maximizes the true skill statistic from the species SDM.

Results

Initially, 6 068 records were downloaded from the FBIS (Figure 1; FBIS, 2022). Data cleaning, guided by expert consultation, resulted in the deletion of 354 erroneous records

(Figure 1). An additional 946 records provided by the experts originating from private, unpublished datasets owned by relevant experts (Figure 1) were uploaded to the FBIS. The final cleaned dataset was also tagged in FBIS as "DFFE Screening Tool 2022", which will ensure that expert input and feedback will not be lost and will help streamline this process in the future.

In total, 6 660 occurrence records for the 34 threatened (Vulnerable = 8; Endangered = 20; Critically Endangered = 6) freshwater fishes were used to develop the NEST layers (Table 1; Figure 1). Of the 34 threatened freshwater fishes included in the NEST, 14 had fewer than 50 records in South Africa (Table 1). Three threatened taxa, *Marcusenius caudisquamatus* ($n = 9$), *Marcusenius* sp. nov. kosi ($n = 4$) and *Silhouettea sibayi* ($n = 12$) were found to have less than 20 records. Only two taxa, *Oreochromis mossambicus* ($n = 2 873$) and *Pseudobarbus* sp. nov. breede ($n = 540$), had more than 500 records (Table 1).

Modelled taxon distribution maps (contributing to the medium sensitivity layer) and current taxon distribution maps (high sensitivity layer) were produced for all taxa classified as Vulnerable and Endangered, except for *S. sibayi*, which had no records post-2002. For Critically Endangered taxa, occurrence data were used to develop a single distribution map for each of the seven taxa (making up the very high sensitivity layer). Example outputs are presented for two fish species, namely, *Labeo seeberi* (Endangered; Figure 2) and *Pseudobarbus erubescens* (Critically Endangered; Figure 2), illustrating finalised occurrence data (Figures 2A, E), SDMs derived from these data (Figures 2B, F), and how these were used to develop the medium (Figure 2C), high (Figure 2D) and very high (Figure 2G) sensitivity layers for the DFFE Screening Tool.

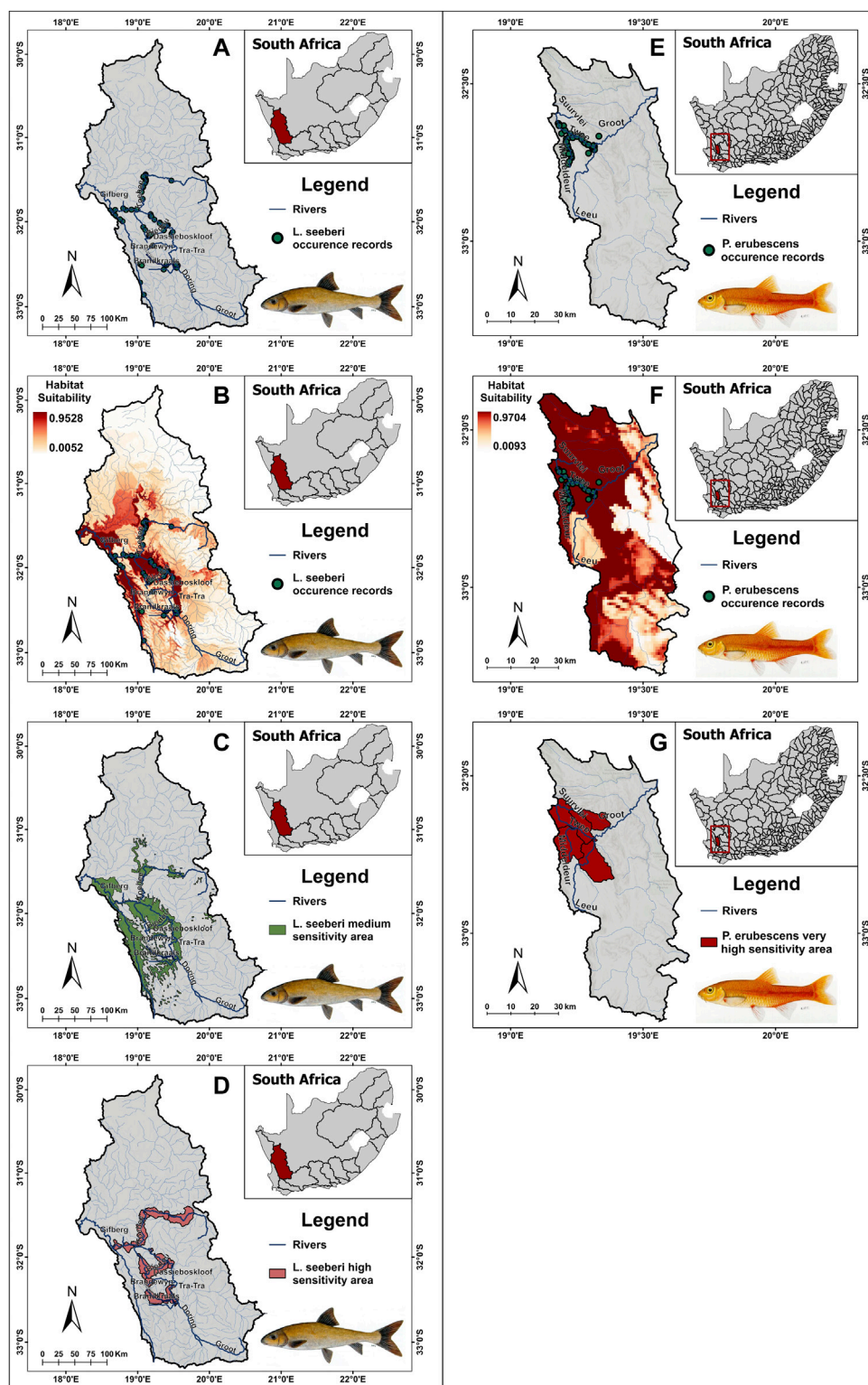
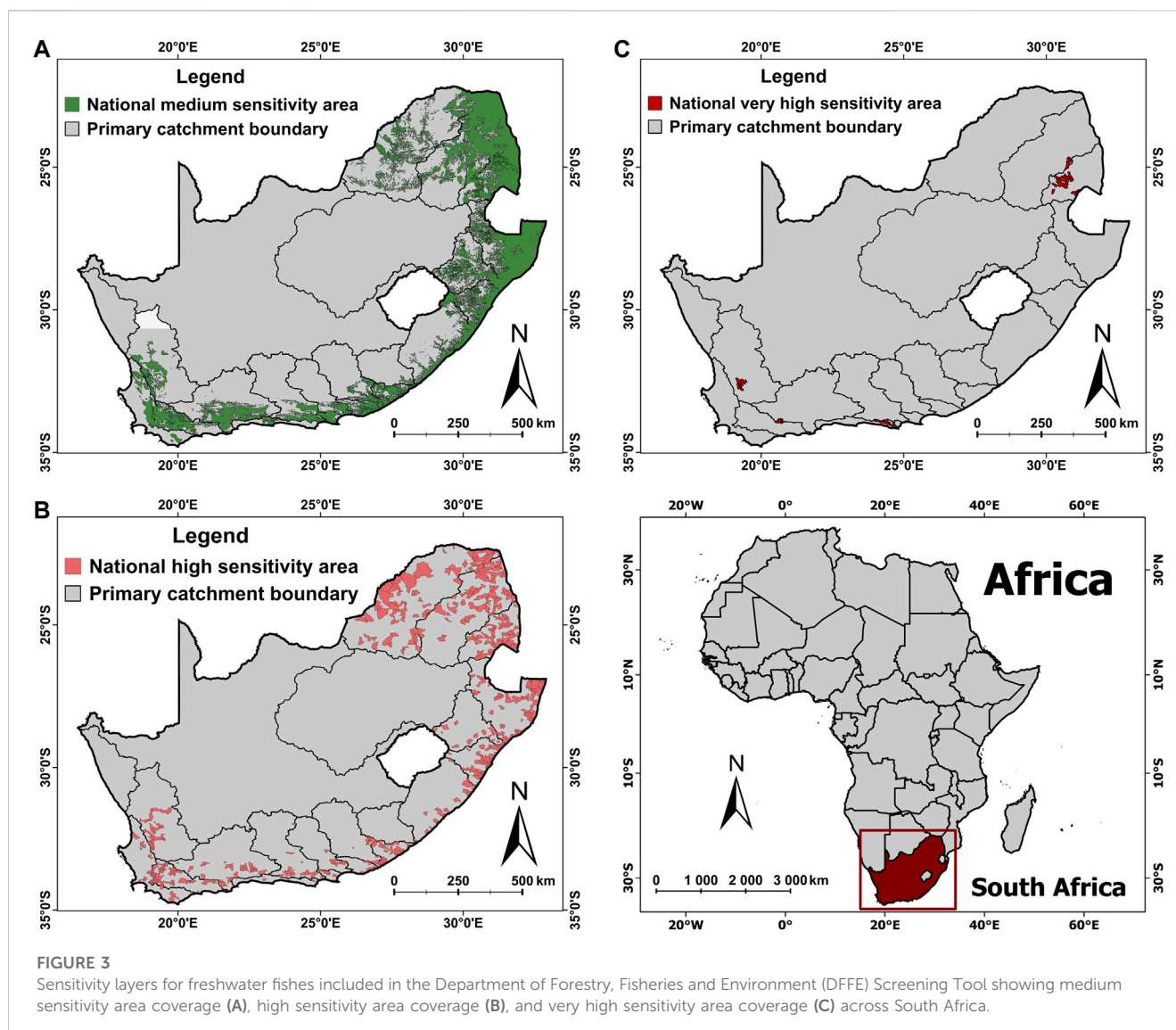


FIGURE 2

Labeo seeberi occurrence records (A), species distribution model (B), medium sensitivity area (C) and high sensitivity area (D) located in the Olifants-Doring primary catchment area of South Africa. *Pseudobarbus erubescens* occurrence records (E), species distribution model (F), and very high sensitivity area (G) located in tertiary catchment E21.



The final composite (combined across all taxa) national sensitivity layers (medium, high, and very high) for all threatened freshwater fishes are presented in [Figure 3](#). The medium sensitivity layer for freshwater fish spans 49 385 km of river and covers a total catchment area of 251 264 km², whilst the high sensitivity layer spans 15 162 km of river and covers a catchment area of 117 412 km², with the very high sensitivity layer providing much-needed protection for Critically Endangered freshwater taxa spanning 1,024 km of river and covering a total catchment area of 5 992 km².

Discussion

We present the first inclusion of freshwater species coverage in a key national conservation decision-support tool—the DFFE EIA NEST—and describe how data from the FBIS were used as the basis for threatened fish sensitivity layers (core elements of the tool) through a multi-stakeholder, collaborative approach (see

Supplementary S3 and S4). Sensitivity layers were successfully developed for 34 threatened freshwater fishes occurring in South Africa and combined to produce national-level sensitivity layers for inclusion into the tool. Given that the majority of South Africa's rivers are either poorly protected or not protected ([Department of Environmental Affairs, 2016](#)), the updated coverage of freshwater fishes in the NEST should provide much-needed protection for the country's threatened freshwater fishes by preventing or minimising further destruction of critical freshwater habitats due to new developments.

Importantly, all new applications for development under the EIA regulations will be compelled to make use of the tool before authorisations are granted. This will identify and protect sensitive catchment areas that in the past would have been overlooked due to a lack of access to freshwater biodiversity data. This will no doubt support future conservation efforts, especially in locations where threatened taxa occur outside of formal, protected areas ([Impson et al., 2002](#); [Nel et al., 2004](#); [Abell et al., 2007](#); [Russell, 2011](#); [Jordaan et al., 2020](#)). In addition to the conservation and policy benefits, the

distribution maps produced for the medium, high and very high sensitivity layers can also help guide expansion of protected areas and stewardship initiatives and can be used in future national biodiversity assessments and IUCN Red List assessments.

The FBIS supported the development of freshwater fish sensitivity layers through providing quick and easy access to reliable and comprehensive freshwater fish occurrence records for South Africa. The relatively large number of records provided by the FBIS is especially noteworthy given that most distribution modelling and mapping of threatened taxa are heavily hampered by limited access to occurrence records (Stockwell and Peterson, 2002; Wisz et al., 2008). A substantial percentage of these data originating from published literature and private data repositories would not have been available had it not been for a platform like the FBIS. Whilst databases like GBIF and FishBase do provide opportunities for the scientific community to share and access freshwater fish data, there are not any concerted efforts to drive the data collation process (especially within freshwater systems in South Africa). In this regard, the FBIS has a proven advantage over global biodiversity databases, through leveraging personal relationships and thorough ongoing stakeholder engagement to ensure that both the quantity and quality of records uploaded to the FBIS are maximised. Furthermore, the inclusion of various taxon specialists, government organisations and conservation authorities (see Supplementary S1) via multiple online workshops ensured that the final occurrence dataset was cleaned efficiently and included only verified, high-certainty records. This culture of cooperation and collaboration was one of the cornerstones of success in developing the freshwater fish sensitivity layers for the NEST and should be adopted by all future bioinformatics projects that seek to impact conservation management and planning on a national scale.

Understandably, we did encounter some limitations whilst developing the NEST sensitivity layers. Firstly, there were 14 taxa with fewer than 50 records in South Africa, with three taxa having less than 20 records each (Table 1). Data scarcity increased the difficulty of producing accurate SDMs (Stockwell and Peterson, 2002; Wisz et al., 2008) potentially reducing the accuracy of the sensitivity layers for these taxa. This highlights the urgent need to increase threatened species monitoring efforts at both provincial and national levels (Chakona et al., 2022). We recommend, based on their low numbers of recent records, that the following taxa be considered as a very high priority for baseline data collection and monitoring: *Chetia brevis*, *Ctenopoma multispine*, *Enteromius treurenensis*, *Kneria* sp. nov. south africa, *Labeo rubromaculatus*, *Marcusenius caudisquamatus*, *Marcusenius* sp. nov. kosi, *Pseudobarbus senticeps*, *Pseudobarbus skeltoni*, *Pseudobarbus* sp. nov. doring, *Pseudobarbus* sp. nov. keiskamma, *Pseudobarbus* sp. nov. keurbooms, *Amatolacypis trevelyani*, and *Silhouettea sibayi*. In this regard, a standardised sampling protocol for freshwater fishes is currently under development in South Africa.

Secondly, the NEST sensitivity layers are only restricted to the modelled and current distribution of threatened taxa. Several studies have highlighted the importance of ensuring that catchment-level impacts (Skelton et al., 1995; Kleynhans, 2000; Nel et al., 2004; Abell et al., 2007), specifically upstream-impacts of anthropogenic activities, also be considered when developing protected areas for freshwater systems. The current iteration of the NEST does not fully account for this, providing protection for threatened taxa at the sub-

quaternary catchment level only. Future iterations of the tool should consider ways to provide upstream protection at a larger primary catchment scale. Lastly, the overall success in terms of providing on-the-ground protection to South Africa's threatened freshwater fishes will be dependent on local government authorities correctly and competently interpreting the outputs of the NEST and making responsible land-use change decisions by preventing developments where the risk to threatened taxa is deemed to be too high.

The freshwater fish component of the NEST provides a critical first step towards adequately incorporating threatened freshwater taxa into the EIA and development application process. However, there are additional steps that could improve upon and update the tool, thereby supporting more effective freshwater conservation in the future. Firstly, there is an urgent need to update and resolve the taxonomy of South Africa's freshwater fishes, as there are a number of distinct genetic lineages that await formal description (Chakona et al., 2015, 2020b; Martin and Chakona, 2019; Bronaugh et al., 2020; Kambikambi et al., 2021; Mazungula and Chakona, 2021; Ramoejane et al., 2021). Once the taxonomy of South Africa's freshwater fishes has been revised, the FBIS and NEST can be updated accordingly. Secondly, the NEST process described here can now be replicated for other species groups such as freshwater invertebrates. Although anurans are already included in the NEST (SANBI, 2021), there is also the potential to use the FBIS and the methodology developed here to improve upon these sensitivity layers. Lastly, distribution mapping used for threatened species should be repeated for non-native freshwater species (specifically invasives) occurring in South Africa—the primary threat to native freshwater fishes in the country (Weyl et al., 2020; Zengeya and Wilson, 2020; Chakona et al., 2022). Worryingly, the distributions of many of the country's non-native fish species are not well known, and a lack of spatial data is holding back reliable system-wide invasive species assessments (De Moor, 1996; Ellender and Weyl, 2014; Zengeya and Wilson, 2020). Although mapping non-native species distributions falls beyond the scope of the NEST, this could provide insights into non-native species impacts on native taxa, and assist with identifying key areas for alien species management interventions (Weyl et al., 2020). Actioning these next steps will amplify the impact that the NEST may have on preventing further population declines in South Africa's freshwater fish fauna.

This case study demonstrates how the FBIS has been used successfully to provide spatial data on threatened freshwater fishes to inform a national-level conservation decision-support tool in South Africa. Key to the success of this project was investing substantial time and effort into manually identifying, collating and cataloguing historic biodiversity data into a standardised, digital format. This generated comprehensive, high-quality biodiversity data sets that, through the FBIS, were then made accessible for analysis and uptake into national conservation planning and decision-making. Collaboration, networking and stakeholder engagement from the outset encouraged data-sharing and facilitated inter-disciplinary skill-sharing (e.g., modelling, mapping and data management skills)—two critical elements of the projects' success - and we recommend this approach to similar bioinformatics efforts elsewhere, particularly in developing countries.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repository and accession number(s) can be found in the article/Supplementary Material. Alternatively, the dataset can be accessed via the Freshwater Biodiversity Information System (freshwaterbiodiversity.org).

Author contributions

MK is the primary author and was involved in all aspects of publishing this manuscript. JS, CG, and HD contributed to the conception and design of the project as well as assisting with writing the first draft of the manuscript. DH contributed to conception and design of the work as well performing statistical analysis and assistance with writing the first draft of the manuscript. JP, DV, DI, AC, DR, NJ, BP, MJ, RB, FR, TZ, AH, and NR-M all assisted with data cleaning and provided conceptual design and advice throughout the project. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

Funding for the development of the Freshwater Biodiversity Information System (FBIS) was provided by the JRS Biodiversity Foundation (Grants 60606 and 60919) and the South African National Biodiversity Institute (SANBI). This work is based on the research supported in part by the National Research Foundation (NRF) of South Africa (Grant Number: MND2000621534710–UID: 133692) and the NRF-SAIAB DSI/NRF Research Chair in Inland Fisheries and Freshwater Ecology (UID 110507). Student funding, in the form of an MSc bursary, was

also provided by the Freshwater Biodiversity Unit of the South African National Biodiversity Institute (SANBI).

Acknowledgments

We would like to acknowledge the contributions of the FRC staff and interns, specifically Aneri Swanepoel and Toni Olsen, who provided invaluable assistance with data collection and cleaning. We would also like to thank Cecilia Cerrilla for providing moral support during drafting and editing of this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fenvs.2023.1122223/full#supplementary-material>

References

- Abell, R., Allan, J. D., and Lehner, B. (2007). Unlocking the potential of protected areas for freshwaters. *Biol. Conserv.* 134, 48–63. doi:10.1016/j.biocon.2006.08.017
- Acreman, M., Hughes, K. A., Arthington, A. H., Tickner, D., and Dueñas, M. A. (2019). Protected areas and freshwater biodiversity: A novel systematic review distills eight lessons for effective conservation. *Conserv. Lett.* 13, 1–14. doi:10.1111/conl.12684
- Adams, V. M., Setterfield, S. A., Douglas, M. M., Kennard, M. J., and Ferdinands, K. (2015). Measuring benefits of protected area management: Trends across realms and research gaps for freshwater systems. *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140274. doi:10.1098/rstb.2014.0274
- WWF (2020). in *Living planet report 2020 - bending the curve of biodiversity loss*. Editors R. E. A. Almond, M. Grooten, and T. Petersen Gland (Switzerland).
- Bronaugh, W. M., Swartz, E. R., and Sidlauskas, B. L. (2020). Between an ocean and a high place: Coastal drainage isolation generates endemic cryptic species in the cape kurper *Sandelia capensis* (Anabantiformes: Anabantidae), Cape region, South Africa. *J. Fish. Biol.* 96 (5), 1087–1099. doi:10.1111/jfb.14182
- Carlson, C. J. (2020). Embarcadero: Species distribution modelling with Bayesian additive regression trees in *r*. *Methods Ecol. Evol.* 11 (7), 850–858. doi:10.1111/2041-210X.13389
- Cerrilla, C., Afrika, J., Impson, D., Kotze, N., Paxton, B. R., Reed, C., et al. (2022). Rapid population decline in one of the last recruiting populations of the endangered Clanwilliam sandfish (*Labeo seeberi*): The roles of climate change and non-native fish. *Aquatic Conservation Mar. Freshw. Ecosyst.* 32 (5), 781–796. doi:10.1002/aqc.3785
- Chakona, A., Gouws, G., Kadye, W. T., Mpopetsi, P. P., and Skelton, P. H. (2020b). Probing hidden diversity to enhance conservation of the endangered narrow-range endemic Eastern Cape rocky, *Sandelia bainesii* (Castelnau 1861). *Koedoe Afr. Prot. Area Conserv. Sci.* 62, 1–6. doi:10.4102/koedoe.v62i1.1627
- Chakona, A., Gouws, G., Kadye, W. T., Jordaan, M. S., and Swartz, E. R. (2020a). Reconstruction of the historical distribution ranges of imperilled stream fishes from a global endemic hotspot based on molecular data: Implications for conservation of threatened taxa. *Aquat. Conserv. Mar. Freshw. Ecosyst.* 30, 144–158. doi:10.1002/aqc.3251
- Chakona, A., Jordaan, M. S., Raimondo, D. C., Bills, R. I., Skelton, P. H., and van Der Colff, D. (2022). Diversity, distribution and extinction risk of native freshwater fishes of South Africa. *J. Fish. Biol.* 100, 1044–1061. doi:10.1111/jfb.15011
- Chakona, A., Malherbe, W. S., Gouws, G., and Swartz, E. R. (2015). Deep genetic divergence between geographically isolated populations of the goldie barb (*Barbus pallidus*) in SouthSouth Africa: Potential taxonomic and conservation implications. *Afr. Zool.* 50, 5–10. doi:10.1080/15627020.2015.1021164
- Dallas, H., and Rivers-Moore, N. (2014). Ecological consequences of global climate change for freshwater ecosystems in South Africa. *S. Afr. J. Sci.* 110, 48–58. Discovery Service for Rhodes University Library. doi:10.1590/sajs.2014/20130274
- Dallas, H., Shelton, J., Sutton, T., Tri Cuptura, D., Kajee, M., and Job, N. (2022). The Freshwater Biodiversity Information System (FBIS)—mobilising data for evaluating long-term change in South African rivers. *Afr. J. Aquat. Sci.* 47, 291–306. doi:10.2989/16085914.2021.1982672

- De Moor, I. J. (1996). Case studies of the invasion by four alien fish species (*Cyprinus carpio*, *Micropterus salmoides*, *Oreochromis macrochir* and *O. mossambicus*) of freshwater ecosystems in southern Africa. *Trans. R. Soc. S. Afr.* 51, 233–255. doi:10.1080/00359199609520609
- Department of Environmental Affairs (2016). National protected area expansion Strategy for South Africa 2016. Available at: http://bgis.sanbi.org/protectedareas/Nationa_Protected_Area_Expansion_Strategy.pdf.
- Department of Forestry Fisheries and the Environment (2021). National web based environmental screening tool. Available at: <https://screening.environment.gov.za/screeningtool/#/pages/welcome> (Accessed August 22, 2022).
- Ellender, B., and Weyl, O. (2014). A review of current knowledge, risk and ecological impacts associated with non-native freshwater fish introductions in South Africa. *Aquat. Invasions* 9, 117–132. doi:10.3391/ai.2014.9.2.01
- Esri Inc (2022). ArcGIS Pro (version 3.0). Available at: <https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview>.
- Evans, W., Downs, C. T., Burnett, M. J., and O'Brien, G. C. (2022). Assessing fish community response to water quality and habitat stressors in KwaZulu-Natal, South Africa. *Afr. J. Aquat. Sci.* 47, 47–65. doi:10.2989/16085914.2021.1952158
- FBIS (2022). Freshwater biodiversity information system (FBIS). FBIS version 3. Available at: <https://freshwaterbiodiversity.org/> (Accessed September 14, 2021).
- Fricke, R., Eschmeyer, W. N., and Van der Laan, R. (2022). Eschmeyer's catalog of fishes: Genera, species, references. San Francisco. Available at: <http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp>.
- Grooten, M., and Almond, R. E. A. (2018). *Living planet report-2018: Aiming higher*. Gland, Switzerland: WWF international.
- Horak, I., Horn, S., and Pieters, R. (2021). Agrochemicals in freshwater systems and their potential as endocrine disrupting chemicals: A South African context. *Environ. Pollut.* 268, 115718. doi:10.1016/j.envpol.2020.115718
- Hughes, K., Harrison, L., Darwall, W., Lee, R., Muruvu, D., Revenga, C., et al. (2021). The world's forgotten fishes. Gland, Switzerland. Available at: https://c402277.ssl.cf1.rackcdn.com/publications/1460/files/original/wwfintl_freshwater_fishes_report.pdf?1617110723.
- Impson, N. D., Bills, I. R., and Cambray, J. A. (2002). "A conservation plan for the unique and highly threatened freshwater fishes of the Cape Floral Kingdom," in *Conserv. Freshw. Fishes options futur* (Hoboken, New Jersey: Oxford Blackwell Sci), 432–440.
- IUCN Red List (2020). International Union for the Conservation of Nature (IUCN) Red List of Threatened Species. Available at: <https://www.iucnredlist.org/> (Accessed December 1, 2020).
- Jordaan, M. S., Chakona, A., and Colff, D. V. D. (2020). Protected areas and endemic freshwater fishes of the cape fold ecoregion: Missing the boat for fish conservation? *Front. Environ. Sci.* 8, 1–13. doi:10.3389/fenvs.2020.502042
- Kambikambi, M. J., Kadye, W. T., and Chakona, A. (2021). Allopatric differentiation in the *Enteromius anoplus* complex in South Africa, with the revalidation of *Enteromius cernuus* and *Enteromius oraniensis*, and description of a new species, *Enteromius mandelai* (Teleostei: Cyprinidae). *J. Fish. Biol.* 99, 931–954. doi:10.1111/jfb.14780
- Kleynhans, C. J. (2000). *Desktop estimates of the ecological importance and sensitivity categories (EISC), default ecological management classes (DEMC), present ecological status categories (PESC), present attainable ecological management classes (present AEMC), and best attainab.* Pretoria: Quaternary Catchments In South Africa.
- Martin, M. B., and Chakona, A. (2019). Designation of a neotype for *Enteromius pallidus* (smith, 1841), an endemic cyprinid minnow from the cape fold ecoregion, South Africa. *Zookeys* 848, 103–118. doi:10.3897/zookeys.848.32211
- Mazungula, D. N., and Chakona, A. (2021). An integrative taxonomic review of the Natal mountain catfish, *Amphilius natalensis* Boulenger 1917 (Siluriformes, Amphiliidae), with description of four new species. *J. Fish. Biol.* 99, 219–239. doi:10.1111/jfb.14714
- Nel, J. L., Reyers, B., Roux, D. J., and Cowling, R. M. (2009). Expanding protected areas beyond their terrestrial comfort zone: Identifying spatial options for river conservation. *Biol. Conserv.* 142, 1605–1616. doi:10.1016/j.biocon.2009.02.031
- Nel, J. L. (2011). Technical report for the national freshwater ecosystem priority areas project: Report to the water research commission. Pretoria. Available at: https://www.karooroever.org.za/images/NFEPA_Technical_Report.pdf.
- Nel, J., Maree, G., Roux, D., Moolman, J., Kleynhans, N., Silberbauer, M., et al. (2004). *South African national spatial biodiversity assessment 2004: Technical report*. Volume 2. Stellenbosch: River component.
- O'Brien, G. C., Ross, M., Hanzen, C., Dlamini, V., Petersen, R., Diedericks, G. J., et al. (2019). River connectivity and fish migration considerations in the management of multiple stressors in South Africa. *Mar. Freshw. Res.* 70, 1254–1264. doi:10.1017/MF19183
- R Core Team (2020). *R core Team*. 2020. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
- Ramoejane, M., Weyl, O. L. F., Swartz, E. R., and Gouws, G. (2021). Identifying multiple geographically restricted phylogeographic lineages of moggel (*Cyprinidae: Labeo umbratus*) in South Africa. *Afr. J. Aquat. Sci.* 46, 225–235. doi:10.2989/16085914.2020.1818546
- Republic of South Africa (1998). *National environmental management Act 107 of 1998*. doi:10.4324/9781315664262-18
- Republic of South Africa (2004a). National environmental management: Biodiversity Act (10/2004): Threatened or protected species regulations. Available at: <https://www.gov.za/documents/national-environmental-management-biodiversity-act-0>.
- Republic of South Africa (2004b). National environmental management: Protected areas Act, No. 57 of 2003. Available at: http://www.nsw.gov.au/sites/default/files/Government_Gazette_2_December.pdf#page=15.
- Republic of South Africa (2010). *National protected area expansion Strategy 2008 protected areas*.
- Russell, I. A. (2011). Conservation status and distribution of freshwater fishes in South African national parks. *Afr. Zool.* 46, 117–132. doi:10.1080/15627020.2011.11407485
- South African National Biodiversity Institute (2021). *Species environmental assessment guideline. Guidelines for the implementation of the terrestrial fauna and terrestrial flora species protocols for environmental impact assessments in South Africa*. Pretoria, South Africa: South African National Biodiversity Institute.
- Skelton, P. H., Cambray, J. A., Lombard, A., and Benn, G. A. (1995). Patterns of distribution and conservation status of freshwater fishes in South Africa. *South Afr. J. Zool.* 30, 71–81. doi:10.1080/02541858.1995.11448375
- Skowno, A. L., Poole, C. J., Raimondo, D. C., Sink, K. J., Van Deventer, H., van Niekerk, L., et al. (2019). *National Biodiversity Assessment 2018: The status of South Africa's ecosystems and biodiversity*. Pretoria. doi:10.1017/CBO9781107415324.004
- South African National Biodiversity Institute (2011). *GIS metadata: Detailed report*. Stellenbosch: ESRI.
- South African National Biodiversity Institute (2022). Red list of South African species. Available at: www.speciesstatus.sanbi.org (Accessed September 14, 2022).
- Stockwell, D. R. B., and Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecol. Modell.* 148, 1–13. doi:10.1016/S0304-3800(01)00388-X
- Wepener, V., van Dyk, C., Bervoets, L., O'Brien, G., Covaci, A., and Cloete, Y. (2011). An assessment of the influence of multiple stressors on the Vaal River, South Africa. *Phys. Chem. Earth* 36, 949–962. doi:10.1016/j.pce.2011.07.075
- Weyl, O. L. F., Ellender, B. R., Wassermann, R. J., Truter, M., Dalu, T., Zengeya, T. A., et al. (2020). "Alien freshwater fauna in South Africa," in *Biological invasions in South Africa*. Editors B. W. van Wilgen, J. Measey, D. M. Richardson, J. R. Wilson, and T. A. Zengeya (Switzerland: Springer), 153–183. doi:10.1007/978-3-030-32394-3
- Williams-Subiza, E. A., and Epele, L. B. (2021). Drivers of biodiversity loss in freshwater environments: A bibliometric analysis of the recent literature. *Aquat. Conserv. Mar. Freshw. Ecosyst.* 31, 2469–2480. doi:10.1002/aqc.3627
- Wisn, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., and Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Divers. Distrib.* 14, 763–773. doi:10.1111/j.1472-4642.2008.00482.x
- Zengeya, T. A., and Wilson, J. (2020). "The status of biological invasions and their management in South Africa in 2019," in *Stellenbosch: South African National Biodiversity Institute, Kirstenbosch and DSI-NRF Centre of Excellence for Invasion Biology*. Editors T. A. Zengeya and J. R. Wilson doi:10.5281/zenodo.3947613
- Zengeya, T. A., Kumschick, S., Weyl, O. L. F., and van Wilgen, B. W. (2020). An evaluation of the impacts of alien species on biodiversity in South Africa using different assessment methods. *Biol. invasions S. Afr.* 14, 487–512. doi:10.1007/978-3-030-32394-3_17
- Ziervogel, G., New, M., Archer van Garderen, E., Midgley, G., Taylor, A., Hamann, R., et al. (2014). Climate change impacts and adaptation in South Africa. *Wiley Interdiscip. Rev. Clim. Chang.* 5, 605–620. doi:10.1002/wcc.295



OPEN ACCESS

EDITED BY

Nick Isaac,
UK Centre for Ecology and Hydrology,
United Kingdom

REVIEWED BY

Amy Whitehead,
National Institute of Water and Atmospheric
Research (NIWA),
New Zealand
Jean-Noel Druon,
European Commission, Joint Research Centre,
Belgium

*CORRESPONDENCE

Kerry J. Sink
✉ k.sink@sanbi.org.za

SPECIALTY SECTION

This article was submitted to
Environmental Informatics and Remote
Sensing,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 25 November 2022

ACCEPTED 28 February 2023

PUBLISHED 31 March 2023

CITATION

Sink KJ, Adams LA, Franken M-L, Harris LR,
Currie J, Karenzi N, Dayaram A, Porter S,
Kirkman S, Pfaff M, van Niekerk L, Atkinson LJ,
Bernard A, Bessinger M, Cawthra H, de Wet W,
Dunga L, Filander Z, Green A, Herbert D,
Holness S, Lamberth S, Livingstone T,
Lück-Vogel M, Mackay F, Makwela M, Palmer R,
Van Zyl W and Skowno A (2023) Iterative
mapping of marine ecosystems for spatial
status assessment, prioritization, and decision
support.
Front. Ecol. Evol. 11:1108118.
doi: 10.3389/fevo.2023.1108118

COPYRIGHT

© 2023 Sink, Adams, Franken, Harris, Currie,
Karenzi, Dayaram, Porter, Kirkman, Pfaff, van
Niekerk, Atkinson, Bernard, Bessinger, Cawthra,
de Wet, Dunga, Filander, Green, Herbert,
Holness, Lamberth, Livingstone, Lück-Vogel,
Mackay, Makwela, Palmer, Van Zyl and Skowno.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Iterative mapping of marine ecosystems for spatial status assessment, prioritization, and decision support

Kerry J. Sink^{1,2*}, Luther A. Adams¹, Mari-Lise Franken^{1,3},
Linda R. Harris², Jock Currie¹, Natasha Karenzi^{3,4},
Anisha Dayaram^{1,5}, Sean Porter⁶, Stephen Kirkman^{2,7},
Maya Pfaff^{3,7}, Lara van Niekerk^{2,8}, Lara J. Atkinson^{3,9},
Anthony Bernard¹⁰, Mariel Bessinger¹, Hayley Cawthra¹¹,
Willem de Wet¹², Loyiso Dunga^{1,3}, Zoleka Filander⁷,
Andrew Green¹³, David Herbert¹⁴, Stephen Holness²,
Stephen Lamberth^{2,15}, Tamsyn Livingstone¹⁶,
Melanie Lück-Vogel^{8,17}, Fiona Mackay⁶, Mapula Makwela¹⁸,
Ryan Palmer¹⁰, Wilhem Van Zyl¹¹ and Andrew Skowno^{1,3}

¹South African National Biodiversity Institute, Cape Town, South Africa, ²Institute for Coastal and Marine Research, Nelson Mandela University, Gqeberha, South Africa, ³Biological Sciences, University of Cape Town, Cape Town, South Africa, ⁴Centre for Statistics in Ecology, Environment and Conservation, University of Cape Town, Cape Town, South Africa, ⁵Restoration and Conservation Biology Research Group, School of Animal, Plant and Environmental Sciences, University of the Witwatersrand, Johannesburg, South Africa, ⁶Oceanographic Research Institute, Durban, Kwa-Zulu Natal, South Africa, ⁷Department of Forestry, Fisheries and the Environment, Branch Oceans and Coasts, Cape Town, South Africa, ⁸Coastal Systems Research Group, Council for Scientific and Industrial Research, Stellenbosch, South Africa, ⁹Egagasinini Node, South African Environmental Observation Network, Cape Town, South Africa, ¹⁰South African Institute for Aquatic Biodiversity, Makhanda, South Africa, ¹¹Geophysics and Remote Sensing Unit, Council for Geoscience, Cape Town, South Africa, ¹²Marine Research Institute and Department of Geological Sciences, University of Cape Town, Cape Town, South Africa, ¹³Geological Sciences, University of KwaZulu-Natal, Durban, South Africa, ¹⁴Department of Natural Sciences, National Museum of Wales, Cardiff, United Kingdom, ¹⁵Department of Forestry, Fisheries and the Environment, Branch Fisheries Research and Development, Cape Town, South Africa, ¹⁶Ezemvelo KwaZulu-Natal Wildlife, Pietermaritzburg, South Africa, ¹⁷Department for Geography and Environmental Studies, Stellenbosch University, Stellenbosch, South Africa, ¹⁸Biodiversity and Conservation Biology, University of the Western Cape, Bellville, South Africa

South Africa has taken an iterative approach to marine ecosystem mapping over 18 years that has provided a valuable foundation for ecosystem assessment, planning and decision-making, supporting improved ecosystem-based management and protection. Iterative progress has been made in overcoming challenges faced by developing countries, especially in the inaccessible marine realm. Our aim is to report on the approach to produce and improve a national marine ecosystem map to guide other countries facing similar challenges, and to illustrate the impact of even the simplest ecosystem map. South Africa has produced four map versions, from a rudimentary map of 34 biozones informed by bathymetry data, to the latest version comprising 163 ecosystem types informed by 83 environmental and biodiversity datasets that aligns with the IUCN Global Ecosystem Typology. Data were unlocked through academic and industry collaborations; multi-disciplinary, multi-realm and multi-generational networks of practitioners; and targeted research to address key gaps. To advance toward a more transparent, reproducible and data-driven approach, limitations, barriers and opportunities for improvement were identified. Challenges included limited human and data infrastructure capacity to collate, curate and assimilate

many data sources, covering a variety of ecosystem components, methods and scales. Five key lessons that are of relevance for others working to advance ecosystem classification and mapping, were distilled. These include (1) the benefits of iterative improvement; (2) the value of fostering relationships among a co-ordinated network of practitioners including early-career researchers; (3) strategically prioritizing and leveraging resources to build and curate key foundational biodiversity datasets and understand drivers of biodiversity pattern; (4) the need for developing, transferring and applying capacity and tools that enhance data quality, analytical workflows and outputs; and (5) the application of new technology and emerging statistical tools to improve the classification and prediction of biodiversity pattern. South Africa's map of marine ecosystem types has been successfully applied in spatial biodiversity assessment, prioritization to support protected area expansion and marine spatial planning. These successes demonstrate the value of a co-ordinated network of practitioners who continually build an evidence base and iteratively improve ecosystem mapping while simultaneously growing ecological knowledge and informing changing priorities and policy.

KEYWORDS

ecosystem mapping, evidence-based biodiversity management, marine ecosystem map, ecosystem types, benthic ecosystems, pelagic ecosystems

1. Introduction

The classification and mapping of ecosystem types is foundational to their assessment and effective management (Borja et al., 2010; Galparsoro et al., 2012; Keith et al., 2022). Developing ecosystem classification frameworks and typologies can also provide an opportunity to 'marshal' biodiversity data and knowledge to inform efficient and appropriate action (Keith et al., 2020). As such, one of the key purposes of classifications is to simplify the complexity of biodiversity data into synthesized spatial information that is fundamental for reporting and management (Nikolopoulou et al., 2021). The main objectives for classifying ecosystems include providing: (i) surrogates to represent biodiversity patterns; (ii) ecologically relevant units to support management and planning; and (iii) ecosystem units that can support ecosystem service and accounting work (Dayaram et al., 2021). Applications of ecosystem classification and mapping outputs include those related to assessment, monitoring, spatial biodiversity planning, allocating environmental flows and natural capital accounting (Bland et al., 2017; Bogaart et al., 2019; Botts et al., 2020).

Globally, there have been multiple efforts to classify and map marine regions, habitats and ecosystem types (Grega and Bodtke, 2007; Longhurst, 2007; Howell, 2010; Last et al., 2010; Grega et al., 2012; Hu et al., 2021; Nikolopoulou et al., 2021). Approaches include those with a focus on biogeography (Spalding et al., 2007) or bio-physical habitats (McArthur et al., 2010; Sayre et al., 2017; Nikolopoulou et al., 2021), with more capacitated countries possessing substantial datasets increasingly applying statistical methods to support classification, mapping and dissemination (Costello, 2009; Dove et al., 2018; Gerovasileiou et al., 2019). In the context of this paper, ecosystems are considered as functionally connected complexes of living organisms and their non-living environment (often referred to as habitat), and it is recognized that

ecosystems can be classified at multiple scales. It is also important to acknowledge the limitations of mapping continuous ecological patterns as distinct polygons that may be less discrete in reality (Keith et al., 2022). Despite the challenges and limitations, national maps of marine ecosystem types or marine ecosystem maps (MEMs) help scientists and managers to organize biodiversity information into spatially discernable units that can support spatial assessment and planning (Dove et al., 2018).

In 2020, the IUCN released a Global Ecosystem Typology (IUCN GET) "to support global, regional and national efforts to assess and manage risks to ecosystems" (Keith et al., 2020, 2022). The IUCN GET provides a globally consistent classification of ecosystems underpinned by ecosystem functioning that spans all realms and the entire biosphere. The typology is hierarchical and developed by top-down classification of biomes within each realm with functional ecosystem groups, but can also accommodate local-scale classifications, including those developed by bottom-up approaches that dovetail with the ecosystem functional groups (Keith et al., 2020). It includes indicative maps of the global distribution at the third hierarchical level (Ecosystem Functional Groups), and there are ongoing efforts to map selected ecosystem types at a biogeographic ecotype level (the fourth hierarchical level). For national applications, more detailed maps at hierarchical levels 4 (biogeographic ecotypes), 5 (global ecosystem types) or 6 (local ecosystem types) are required. The classification and mapping of ecosystems at a national or greater scale is challenging, because of the diversity of ecosystems involved, lack of data availability and coverage, limitations in understanding the drivers of ecosystem pattern, and discrepancies in knowledge among different types of ecosystems (Keith et al., 2022). These challenges are exacerbated in developing countries where there are often less financial and human resources to support strategic biodiversity data collection, analyses and translation of technical outputs into decision-support tools.

Marine ecosystem classification and mapping lags far behind the equivalent efforts on land. The collection of marine biodiversity and environmental data is particularly difficult, because the sea is less accessible (and most of it opaque to satellites) compared to the terrestrial and fresh-water realms (Roberson et al., 2017; Smit et al., 2022; Bell et al., 2022a). Ecosystem mapping is more complex in the marine realm because the ocean environment is three-dimensional, fluid with few prominent boundaries that limit connectivity, and can be highly variable and dynamic (Howell, 2010; Roberson et al., 2017; Sayre et al., 2017; Sink et al., 2019). The costs associated with deep-sea research increase exponentially with increasing depth, resulting in a rapid decrease in data at greater depths (Bell et al., 2022a,b). Furthermore, the technical skills needed to collect, process and disseminate relevant marine geological, oceanographic and biodiversity data are often limited or concentrated within isolated research groups. Different disciplines of marine science have often operated in silos, focused in different areas or conducted research at different scales, which has limited multi-disciplinary collaborative utility. Moreover, oceanographers or marine geoscientists rarely rely on biodiversity data, but ecosystem mappers rely on oceanographic, bathymetric and geological data that complement biodiversity datasets. These different types of data products rarely align in space and time.

South Africa presents a unique opportunity to report on the process and progress in advancing marine ecosystem mapping and is a relevant case study for three reasons. Firstly, we have an established record in harnessing biodiversity data to support systematic biodiversity assessment and conservation planning (e.g., Balmford et al., 2002; SANBI and UNEP-WCMC, 2016; Botts et al., 2019; Skowno et al., 2019; Holness et al., 2022), often leading to implementation success (Botts et al., 2019; Sink et al., 2019; Harris et al., 2022a; von Staden et al., 2022). Secondly, South Africa has three ocean systems: the Indian, Atlantic and Southern Oceans, with a high associated diversity of ecosystems and species (Gibbons, 1999; Costello et al., 2010; Griffiths et al., 2010) that exemplify the challenges in classifying and mapping many different ecosystem types. Thirdly, as a developing country, we can provide relevant lessons for other countries requiring evidence to support sustainable development of their ocean economies, but due to resource and capacity constraints, are faced with different challenges and opportunities to those of more developed nations (Smit et al., 2022). In addition, South Africa's national MEM has been adjusted for improved alignment with global typologies where appropriate, and comparisons of classifications at multiple scales provide opportunities for joint learning and adjustments of classifications and typologies.

We aim to report on South Africa's approach to produce and improve the national MEM as a foundation for assessment, planning and ecosystem-based management of the marine realm. The objectives are: (1) to document the process, results and application of the map; (2) to report key limitations, barriers and challenges; and (3) to identify enabling factors and opportunities for more transparent, data-driven and repeatable improvements. In doing so, we contribute an approach to collate, collect, share and integrate multi-disciplinary data to iteratively improve maps of marine ecosystem types in a developing-country context. Because this paper shares lessons from iterative development over the past two decades and also reports on future plans, it departs from the standard paper format and presents methods, results and reflections on several components before distilling key lessons. In so doing, we demonstrate a collaborative

approach to iteratively improve ecosystem maps to feed into a national system that synthesizes biodiversity data and knowledge to support biodiversity assessment, spatial planning, decision-making and protection.

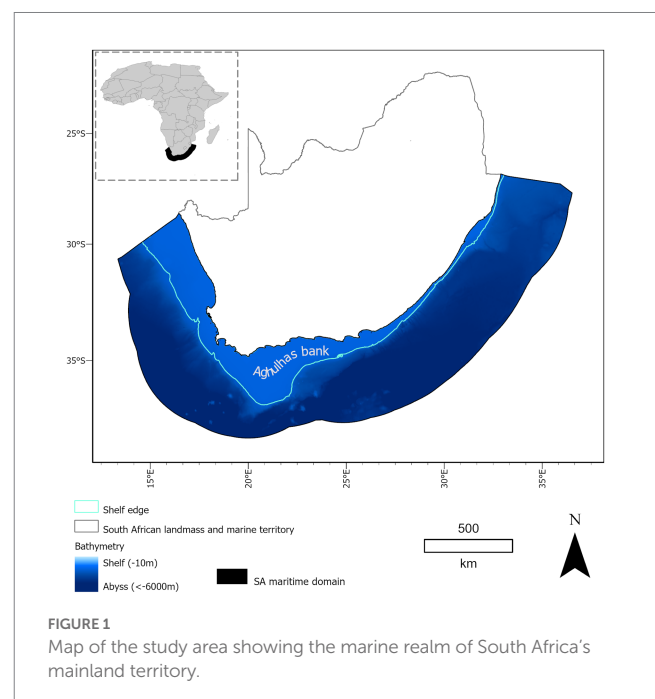
2. South Africa's iterative approach

2.1. Study area

The study area is South Africa's mainland maritime domain (i.e., territorial seas plus exclusive economic zone), which extends from the dune base (Harris et al., 2019) to 200 NM seaward (Figure 1). The width of the continental shelf varies between 1.3 and 260 km, and the depth of the shelf break ranges between 80 and 600 m (de Wet and Compton, 2021). The western continental margin is wide and deep and terraced in places (de Wet and Compton, 2021), and tapers in a southern-easterly direction. The southern continental margin includes the wide Agulhas Bank resembling the outline of the South African shoreline, and abruptly tapers to the eastern continental margin, which lies parallel to the coastline. Slope and abyssal depth zones account for more than 60% of the mainland maritime domain. This paper does not cover the classification developed for the Prince Edward Island territory (Whitehead et al., 2019) or the estuarine realm (van Niekerk et al., 2020), and the coast as a cross-realm zone is also considered elsewhere (Harris et al., 2019).

2.2. Evolving approach

South Africa's approach to coordinate, guide and advance ecosystem classification and mapping has developed, evolved and iteratively improved over time. A key driver of this process has been the series of three National Biodiversity Assessments (NBAs) that



South Africa has undertaken over the past two decades (Driver et al., 2005, 2012; Skowno et al., 2019), with recent changes in the ecosystem map undertaken to facilitate alignment with the IUCN GET (Keith et al., 2020, 2022). The first National Biodiversity Assessment (2003–2005) brought scientists together to discuss marine bioregions and spatial biodiversity data for assessment and planning for the first time (Lombard et al., 2005). Ecosystems were assessed at a scale of 34 broad biozones (broad categories reflecting main biogeographic and depth zones) across the seascape, and although a finer-scale national map of ecosystem types was not yet developed, digitization of existing hard-copy maps of relevant features such as sediment and submarine canyons was initiated. The biozone map (MEM 2004) can be considered as a precursor to an ecosystem map proper, but nevertheless was used to determine the two headline indicators in South Africa's NBAs: ecosystem threat status and protection level (Skowno et al., 2019).

Further advances in data collation and mapping were made (2006–2011) through a project working to identify and support implementation of an ecologically representative offshore marine protected area network (Sink, 2016; Roberson et al., 2017; Kirkman et al., 2021). The MEM 2011 (Sink et al., 2012) was produced through collaborations from 25 organizations facilitated by a series of workshops that were part of the assessment process (2009–2011). At these workshops, initial habitat classifications initiated in the 2004 assessment and global efforts in this context were discussed, the pelagic ecosystem classification developed from cluster analysis of sea-surface data (now published as Roberson et al., 2017) was reviewed, and reef classifications were considered. The most recent NBA (Sink et al., 2019) produced a revised map of marine ecosystem types, MEM 2018, based on additional collated data, and minor changes were made for MEM 2022 to improve alignment with the IUCN GET (Keith et al., 2020). There have thus been four versions of the MEM, with MEM 2022 being most advanced and described in more detail in this paper (Figure 2), each discussed in more detail in Sections 2.4–2.6.

2.3. Network of practitioners to advance ecosystem classification and mapping and input data

The team and governance arrangements to support ecosystem classification and mapping have also evolved over time, with ongoing efforts to improve representation, capacity and coordination. Initial efforts were informal with South Africa's first Marine Ecosystem Committee, formalized in 2015, drawing from experience in the terrestrial realm (Dayaram et al., 2021) and the collaborations that advanced the map of marine ecosystem types for the NBA 2011. A clear terms of reference was developed (Supplementary material) and potential members were approached or nominated from key organizations and departments. The Marine Ecosystem Committee was tasked with: (1) facilitating collaboration between institutions and individuals involved in regional and national marine ecosystem classification, mapping and description; (2) agreeing on the purpose, approach, data sources and methods for classifying and mapping marine ecosystems in South Africa; (3) supporting compatibility and alignment (spatially and conceptually), with the national vegetation map, the national

estuary map and more recently, the emerging IUCN GET; and (4) assisting with decisions pertaining to the curatorship, update and changes in the national marine ecosystem classification system and map, including the review of protocols and procedures. Due to the substantial volume of work involved in advancing this task, a less formal broader network was then developed in 2018 with multiple task teams to advance more specific, focal areas (Figure 3). The inclusion of an emerging researcher task team helped develop long-term capacity, tackle specific research questions and diversify perspectives and approaches including the application of more modern methods in ecosystem classification and mapping (Supplementary material).

South Africa's current ecosystem committees and the associated network that advance marine classification and mapping are shown in Figure 3. A National Ecosystem Committee guides the body of work across realms and ensures consistent principles and alignment in classifications and maps. Realm-specific committees lead and coordinate the work of each realm (terrestrial, river, wetland, estuarine and marine). Each committee has broad institutional and ecosystem representation and a chair. These chairs also serve on the National Ecosystem Committee, which is also where, *inter alia*, issues pertaining to the cross-realm coastal zone are discussed and resolved. Sharing of experience across realms has supported consistency in conceptual approaches, learning across realms, and standardization in governance.

The work of the Marine Ecosystem Committee is supported by a less formal marine ecosystem network that comprises nine task teams (Figure 3), including seven thematic task teams, a cross-cutting team, and a dedicated emerging researchers task team to ensure that the skills and relationships are cultivated in the next generation of scientists who will advance this work. This task team consists of an inclusive group of early career scientists and postgraduate students who volunteer or are invited to collaborate and develop skills in the ecosystem classification and mapping process. The emerging researchers also participate in one or more of the thematic task teams and are invited to all Marine Ecosystem Committee meetings. Task teams have an elected lead and the leads make up the Marine Ecosystem Committee together with additional members to ensure institutional representation. The cross-cutting task team supports alignment between task teams within the network and the classification and mapping of transitional ecosystem types (Keith et al., 2020), coastal ecosystem types, and other ecosystems like bays that are challenging to incorporate in the classification scheme. The leads of the terrestrial (based largely on national vegetation types) and estuarine committees form part of the marine ecosystem network. All these committees and task teams are governed by specific terms of reference (Supplementary material). Task teams work informally and advance work through email, research collaborations and dedicated meetings. They are independently convened but the leads report to the Marine Ecosystem Committee that meets annually. Annual meetings of the network and some task teams were supported by government funding (less than \$500 per year) with actual field expenses and analyses funded by organizations supporting members of the network and project funding. The COVID-19 pandemic caused all meetings to move online and this eliminated travel and meeting expenses. Proposed amendments to the ecosystem classification map are presented for deliberation and decision by the Marine Ecosystem Committee. Where consensus is not reached within a realm-specific


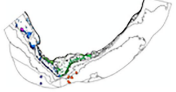
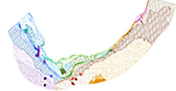
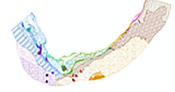



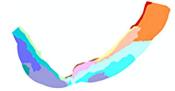
	SA MEM 2004	SA MEM 2011	SA MEM 2018	SA MEM 2022
Informed by	Expert input, Literature & 1 Dataset (Bathymetry)	Expert input, Literature & 12 Datasets	Expert input, Literature & 76 Datasets	Expert input, Literature & 83 Datasets (Table S1)
Key elements of mapped pattern	Depth; Biogeography	Depth & topography; Substratum; Broad ecosystem groups; Wave exposure, grain size, geology or beach morphodynamics; Oceanographic variables; Biogeography	Ecoregion; Bathymetry, Substratum & seabed features; Wave exposure & beach morphodynamics, Biotopes, Vulnerable Marine Ecosystems, River influence & fluvial fans, Oceanographic features	Ecoregion; Bathymetry, Substratum & seabed features; Wave exposure & beach morphodynamics, Biotopes, Vulnerable Marine Ecosystems, River influence & fluvial fans; Oceanographic variables
New Elements		Shores mapped as buffered lines and beaches classified into morphodynamic types; Inclusion of sediment types & seabed features; First application of data	Shores mapped as polygons at <1:3000; Inclusion of kelp forests, fluvial fans, stromatolites & bays; Finer depth strata on shelf & slope; Introduction of mosaic ecosystems on dynamic shelves; Inclusion of visual survey and remote sensing data; Benthic and pelagic components combined	Alignment with IUCN GET including hierarchical changes; Introduction of ecosystem functional groups; Separation of benthic and pelagic ecosystem types beyond the shelf edge
No. of Ecosystem Types for Assessments	34 Biozones	136 Marine and Coastal Habitat Types	150 Marine Ecosystem Types	163 Marine Ecosystem Types
Benthic				
Pelagic				

FIGURE 2

Schematic illustrating the evolution of South Africa's national marine ecosystem map (SA MEM) used to assess ecosystem threat status and protection level between 2004 and 2022 in South Africa. This evolved from a literature- and expert-based approach to a data-informed but expert-driven method drawing from 83 multi-disciplinary datasets with recent changes to align with the IUCN Global Ecosystem Typology (IUCN GET). Note that the key elements of mapped patterns are listed hierarchically as they were applied in the classification.

committee, the National Ecosystem Committee undertakes final decision making.

2.4. Conceptual/classification framework

South Africa's maps of ecosystem types aim to simplify biodiversity into spatially distinct units that represent areas of more cohesive biodiversity pattern. They are discernable by the main factors influencing their composition, structure and function (Dayaram et al., 2021). South Africa follows a hierarchical approach in ecosystem classification. The marine classification hierarchy and the incorporation of ecosystem types has evolved over time (see overview above). The precursor to the ecosystem types map nested tidal, topographic and depth strata within 10

bioregions (Lombard et al., 2005), but the 2011 scheme nested broad ecosystem groups (similar to the IUCN level 3 Ecosystem Functional Groups; EFGs) within depth zones, which were sub-classified using biogeography in an approach more similar to the marine typology described by Keith et al. (2020). In 2018, the Marine Ecosystem Committee took a decision to nest ecosystem types within ecoregions (Sink et al., 2019). In 2022, the classification and hierarchy was reorganized to align with the emerging IUCN Global Ecosystem Typology 2.0 (Keith et al., 2020, 2022). South Africa considers the IUCN GET marine biomes as bathomes because of their focus on depth zones rather than biomes, which should consider broader elements, such as ecosystem function, biogeographic patterns and evolutionary history (Mucina, 2019). The marine ecosystem committee discussed implications before amendments were implemented and reviewed the IUCN

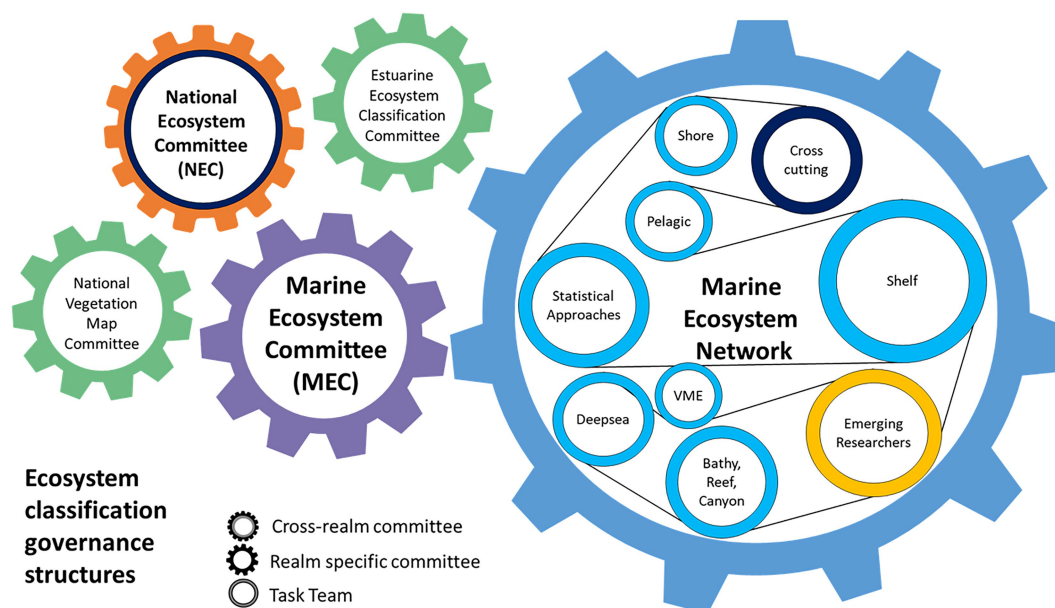


FIGURE 3

Ecosystem classification governance structures pertinent to the marine realm in South Africa. Marine ecosystem classification and mapping is led by a Marine Ecosystem Committee guided by a cross-realm National Ecosystem Committee (orange) but supported by other coastal committees (green) and the marine ecosystem network that includes nine task teams, seven of which are thematic (blue), one cross-cutting (black), and one dedicated to emerging researchers (yellow). VME refers to Vulnerable Marine Ecosystems.

framework in 2020, proposing some adjustments. Meetings were held between committee members and members of the IUCN team to explore alignment and discuss challenges.

In 2022, South Africa adopted the IUCN's hierarchical classification system with minor modifications to create an updated hierarchy for South Africa's marine ecosystems (Figure 4). The IUCN GET recognizes 6 levels, 5 of which were used in South Africa. Within the marine realm (level 1), the 4 marine biomes (level 2) specified by Keith et al. (2020) were adopted (but considered as bathomes): Shore, Shelf, Deepsea Benthic and Pelagic. To match the IUCN typology, a level of EFGs was developed with each type assigned into a relevant GET EFGs, with the exception of bays which are temporarily considered as a distinct additional EFG (level 3). The EFGs are a central concept in the IUCN GET, which classifies ecosystem types on the basis of function rather than through a biogeographical or biophysical lens (Keith et al., 2022).

We did not make use of some functional groups within the marine IUCN GET because these types are not known to occur in the study area (e.g., muddy shores and hadal ecosystems >6,000 m), fall into other realms (e.g., coastal shrubland and grassland, which are beyond the extent of the marine ecosystem map), there is insufficient data to map them (shellfish beds, reefs and chemosynthetically based ecosystems), or the EFG was too broad and contained multiple EFGs (e.g., upwelling shelves). We also did not apply artificial shorelines, which are mapped as the historical natural ecosystem type, so that coastal development could be mapped as a pressure and that portion of the ecosystem extent reported as lost in a threat assessment.

The fourth level in our classification is made up of 58 biogeographic ecotypes nested within the EFGs (consistent with the IUCN typology). We did not attempt to define global ecosystem types (IUCN GET level 5). As such, the lowest and 5th level in our typology

is made up of the national ecosystem types, which relate to 'local ecosystem types' in the 6th level of the IUCN GET.

The most contentious areas in South Africa's work has been in nesting different components related to biogeography and depth, coupling or separating benthic and pelagic elements, and the challenges in classifying and mapping more connected or transitional ecosystem types. Depth and biogeography are universal elements in most marine and seabed classification schemes (Howell, 2010; Keith et al., 2020; Swanborn et al., 2021), with many separating biogeographic elements before depth. This was regularly debated and alternated between years in South Africa, but like the IUCN GET, the committee considered depth and topography (bathome) to be a more important driver of differences in biodiversity pattern than other assembly drivers, such as substratum, nutrients, oceanographic variables, disturbance regimes, and biotic interactions (Howell, 2010; Keith et al., 2022).

Bays also elicited substantial discussion. Although there is recognition of the distinct processes that operate in bay systems, there is less evidence that these translate into notable differences in community composition, structure and function. There were requests to distinguish different ecosystem types within bays (such as sandy shelves, kelp forests, and reefs) from those outside bays, by scientists and managers working at different scales. Islands and lagoons also posed challenges with South Africa's only marine lagoon transferred to the estuarine realm in 2018, in recognition of the groundwater flow that influences ecosystem functioning (Whitfield, 2005).

The IUCN classification does not cater for mixed shores or the mosaic shelves that are recognized in the South African classification. Discussions with the IUCN GET team indicated that the absence of a mixed shore Ecosystem Functional Group was an oversight rather than a deliberate omission in the typology. We are

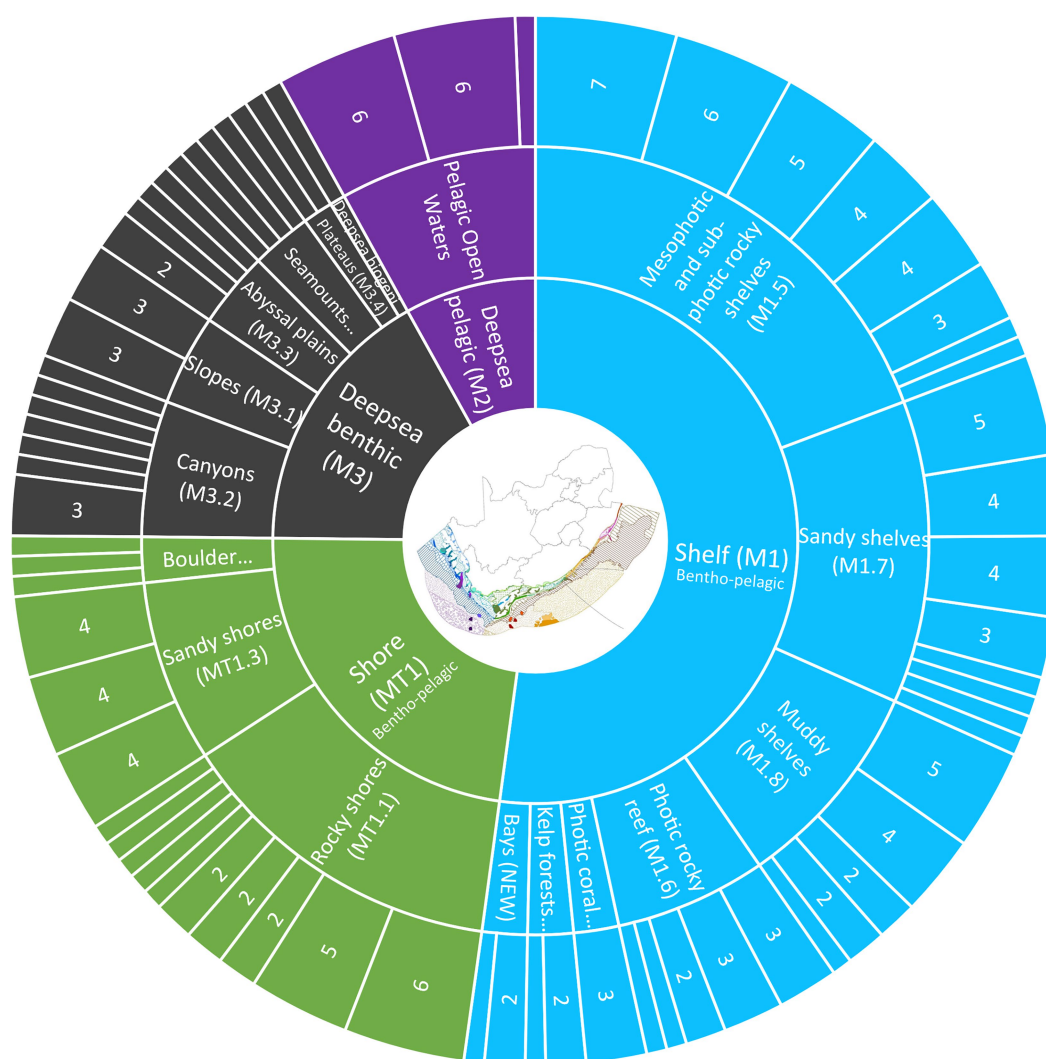


FIGURE 4

Overview of South Africa's latest marine ecosystem classification with four bathomes (inner ring), 17 Ecosystem Functional Groups (central ring) nested within the four bathomes 58 biogeographic ecotypes (segments in outer ring) and 163 national marine ecosystem types (not shown as a distinct ring). The number and relative width of each segment in the outer ring reflects the number of national marine ecosystem types nested within each of the biogeographic ecotypes. Where there is no number, there is only one ecosystem type associated with that biogeographic ecotype. Note that the shore and shelf do not separate into benthic and pelagic components but the deepsea includes overlapping benthic and pelagic layers (see Table 1).

currently exploring whether it is more appropriate to split mixed shores into sandy, rocky or boulder and cobble types, based on the dominant shore type, and to align with the existing groups, or to propose addition of a mixed shore Ecosystem Functional Group to the IUCN GET. This decision not only requires a better understanding of the biodiversity pattern in space and time within these systems but also improved ecological information on ecosystem functioning, species interactions and connectivity across habitats. Emerging research results have shown that many parts of the shelf, especially in the high current eastern margin, are more dynamic than anticipated. Rocky areas experiencing regular sand inundation host assemblages that are distinct from sandy and muddy areas where rocky substrate is absent, although the mosaic systems host a subset of species that characterize higher profile and more permanently rocky habitats (Porter et al., 2017; Supplementary material).

South Africa's deep shelf, extending to 600 m in parts of the western margin (Sink et al., 2019) also posed challenges in aligning with international classification schemes such as the IUCN GET, which does not accommodate subphotic reefs on the shelf. South Africa combined mesophotic and subphotic reefs and assigned the IUCN GET code for photo-limited marine animal forests (M1.5 Keith et al., 2020), but recognize the need for further work in collaboration with the IUCN GET team.

Accommodating the three-dimensional and dynamic nature of the ocean in ecosystem typologies and maps is particularly challenging (Porter et al., 2017). South Africa separated benthic and pelagic elements in 2011 but re-combined these in 2018 due to challenges linked to apportioning impacts in coupled benthic-pelagic ecosystems in the 2018 assessment (Sink et al., 2019) and because overlapping ecosystem types raised concerns from managers and practitioners working to advance ecosystem accounting. One of South Africa's key

TABLE 1 Description of vertical stratification in South Africa's 2022 marine ecosystem map.

	Ecosystem layers	Description	Number of types
Shore and shelf	Benthic-Pelagic	Benthic and pelagic components not distinguished on the shore or shelf but these ecosystem types are considered to have tightly coupled benthic and pelagic elements. Ecosystem types extend from the sea surface to the seabed.	128
Beyond the shelf	Deepsea pelagic	Ecosystem type extends from the sea surface to 500 m above the seabed, noting the limitations indicated in Roberson et al. (2017).	13
	Deepsea benthic	Ecosystem type extends from the seabed with its associated biodiversity to 500 m above the seabed.	22
			Total 163

There is a single combined benthic-pelagic layer on the shore and shelf and separate but overlapping benthic and pelagic layers in the deepsea.

fisheries, the demersal hake trawl fishery targets benthic-pelagic hake *Merluccius capensis* and *M. paradoxus* which feed in the water column but rests on the seabed (Pillar and Barange, 1995; Huse et al., 1998), and trawling (benthic and mid-water) affects both benthic and pelagic ecosystem components. Managers were particularly concerned about the complexity that would be caused by introducing four overlapping pelagic classes (epipelagic, mesopelagic, bathypelagic, and abyssopelagic, as per the IUCN GET) in addition to a benthic component, and scientists indicated that there was insufficient data to reliably classify the vertical component in South Africa at this stage. However, in 2022, benthic and pelagic elements were separated beyond the shelf edge to align with the IUCN GET (Table 1).

This was feasible because South Africa had undertaken a pelagic bioregionalization in 2010 to support the development of a representative Marine Protected Area network (Roberson et al., 2017), and assessed ecosystem threat status for 13 pelagic ecosystem types in 2011 (Sink et al., 2012). The 500 m depth boundary used to delineate the area beyond the shelf edge where separate benthic and pelagic layers were introduced, was a pragmatic decision noting that the depth of the shelf break varies mostly between 100 and 500 m in South Africa (de Wet and Compton, 2021). Note that three of the pelagic ecosystem types (Aa1, Ab1, and Ab2) defined by Roberson et al. (2017) were excluded as pelagic ecosystem types in the 2022 map, because these are confined to the shelf where all ecosystem types are considered benthic-pelagic. A better understanding of vertical stratification, benthic-pelagic coupling in different bathomes and ecosystem functional groups and the impact of pressures across recognized vertical components, is needed. More international collaboration in resolving pelagic ecosystem pattern, process and functioning could benefit global and national ecosystem classification and mapping efforts.

2.5. Key datasets

The MEM 2004 classification was based only on literature and expert input, but initiated mapping of some features for which data were readily accessible (see Lombard et al., 2005 for details). Intertidal habitats were mapped by digitizing the shoreline from 1:50000 topo-cadastral maps, split into 12 habitats based on several expert-based maps (e.g., wave exposure) and/or hard-copy maps (e.g., Jackson and Lipschitz, 1984). Offshore sediment and seabed features (seamounts and submarine canyons) were also mapped, drawing from published geological and sediment maps (Birch et al., 1986; Dingle, 1986; Dingle et al., 1987), untrawled grounds (data provided by the fisheries department), and unpublished canyon maps from academics.

The MEM 2011 drew from these and other data sources, including additional substrate, geology and oceanographic datasets (Sink et al., 2012). Because beaches were represented in 2004 as 'sand', a dedicated effort went into improving representation of beach morphodynamic types (Harris et al., 2011). The scale and accuracy of the shoreline was also improved by mapping a consistent midshore line on SPOT5 satellite imagery (Harris et al., 2011) that corrected errors introduced in 2004 from mapping the symbology for different shore types off the 1:50000 topo-cadastral maps. Substrate (e.g., rocky, sandy, muddy, gravel, mixed), wave exposure (sheltered, exposed, or very exposed), grain size, and biogeography were used to classify 58 coastal and inshore habitat types (Sink et al., 2012). Depth and slope, substrate, geology (e.g., sandy, muddy, gravel, reef, hard grounds, canyons, and ferro-manganese deposits), and biogeographic data were used to map 62 offshore benthic habitats (Sink et al., 2012). Sea surface temperature, primary productivity and chlorophyll-*a* content, depth, turbidity, frequency of eddies, and distribution of temperature and chlorophyll-*a* fronts were used to map 16 offshore pelagic ecosystem types (Roberson et al., 2017).

Between 2013 and 2018, experts identified and prioritized key datasets (Supplementary Table S1) to represent established marine biodiversity surrogates drawing from relevant international literature (e.g., Last et al., 2010; Briggs and Bowen, 2012; Spalding et al., 2012; Douglass et al., 2014; Sutton et al., 2017; Sink et al., 2019). New datasets, including contemporary and historical data, were discovered or made accessible through collaborations with academics, government and industry partners. A total of 83 datasets were used, including: shore maps ($n=1$), remote sensing data ($n=2$), bathymetric data ($n=7$), geoscience data ($n=35$), sediment data ($n=13$), historical data that could inform seabed types ($n=2$), visual survey datasets ($n=8$), biotope data that classified biological assemblages ($n=7$), a dataset for Vulnerable Marine Ecosystems [VMEs as defined by FAO (2009)] ($n=1$) and oceanographic data ($n=8$; Supplementary Table S1). Of these, 67 datasets were of abiotic variables, 16 of biotic variables, and one of mapped ecosystem types (shores). These data sets include different spatial and temporal resolutions that influence the quality and granularity of ecosystem mapping efforts. Where finer resolution abiotic data or biodiversity data were available these took precedence over coarser grained or environmental data. This meant that the better sampled shore and shelf and some finer resolution or granular ecosystem types (e.g., reefs and kelp forests) were mapped at a higher resolution than deep sea or broad scale ecosystem types (e.g., abyssal plains and pelagic habitats; Supplementary Figure S1).

Shores were delineated using expert mapping and remote sensing in a separate process for direct use in MEM 2018. Importantly, there

was fine-scale delineation (<1:3000) of the shores to their actual extent (dunefoot to the back of the surf zone) and the marine, estuarine and terrestrial ecosystem maps were seamlessly aligned (see [Harris et al., 2019](#) for a detailed description of the methods). For the rest of the marine realm, collation of additional historical and contemporary data provided information for parts of the country that previously lacked data, facilitated improved seabed (e.g., through the compilation of a national sediment layer), and shore mapping (e.g., inclusion of stromatolites) and the first inclusion of biodiversity data from biotope classifications ([Supplementary Table S1](#)). The use of more modern methods such as remote sensing and visual surveys ([Supplementary Table S1](#)) facilitated the inclusion of kelp forests and supported improved mapping of shelf habitats including reefs and reef mosaics ([Supplementary Table S1](#)).

There were marked increases in the number and diversity of datasets and spatial layers that informed the iterative MEM versions, enabling more sophisticated mapping, and inclusion of greater detail with more confidence. The growing number of informative data layers used increases the complexity of data management and curation. The diverse array of environmental (bathymetry, geology, oceanography) and biodiversity (biotope, VME, remote sensing) data have been sourced from broad network of institutions and research partners ([Supplementary Table S1](#)), without a focus on data or metadata standards. The data and their resultant spatial layers have been managed as a collection of files on analyst's computers, without dedicated data management, metadata curation and version control. To attain the vision of well documented and reproducible workflows ([Poole et al., 2023](#)), SANBI, including its marine program, needs to strengthen its data management capacity. Improved data management will include the use of spatially enabled databases, centralized, documented and version-controlled data repositories that can feed scripted, reproducible workflows. As with most improvements in the MEM, these capabilities will be developed and improved upon iteratively over time.

2.6. Producing the National ecosystem maps

The identification and delineation of biozones for MEM 2004 was informed by expert workshops where discussions focused on biogeographic patterns and depth zones (supratidal, intertidal, shallow and deep photic, sub photic, upper and lower slope, and abyss). For MEM 2011, the first national map of coastal and benthic ecosystem types and a separate national map of pelagic ecosystem types ([Roberson et al., 2017](#)) were produced. For the former, the midshore line [Harris et al. \(2011\)](#) was buffered by 500 m landward and seaward to create polygons representing the shores. The intertidal habitats from 2004 and beach morphodynamic types from [Harris et al. \(2011\)](#) were combined by coding wave exposure from the former to the rocky and mixed shores mapped by [Harris et al. \(2011\)](#), and then all shores were split by bioregions to give the shore ecosystem types. Offshore, a top-down approach was used to delineate refined depth zones and digitized geological maps facilitate mapping of seabed types using GIS.

By 2018, a range of new data, both biotic and abiotic, had been collated to support an improved MEM ([Supplementary Table S1](#)). As mentioned above, shores were included directly from [Harris et al. \(2019\)](#). For the rest of the MEM 2018, prioritization and exclusion rules were used in cases where the data layers overlapped spatially, as

described for the MEM 2011 above. Typically, the most recent data (in the case of conflicting sediment type), data with greater confidence (such as visual surveys) or finer resolution mapping (such as kelp forest mapped by remote sensing, or multi-beam data interpreted by geoscientists) were prioritized in assigning seabed types. The distribution of biotopes ([Supplementary Table S1](#)) was useful in determining whether non-contiguous polygons with similar abiotic characteristics hosted different epifaunal, infaunal or fish biodiversity. For example, two areas of similar depth and sediment might be separated on the basis of infauna or epifauna assemblage data even though the existing abiotic data does not reflect any differences in environmental conditions. Polygons representing ecosystem types were thus mapped in ArcGIS Pro 10.4 using an expert-driven but data-informed approach, with a focus on identifying more uniform areas separated by discontinuities in multiple datasets. Mosaic ecosystems were defined and mapped as areas with high spatial and temporal variability in the multiple data sources used. Further, a decision was made to combine the benthic and pelagic maps from 2011 into a single MEM in 2018. The mapping process revealed aspects of the classification that needed further consideration (e.g., how to position certain ecosystem types that spanned bathomes in the hierarchy, such as kelp). This required collaboration and feedback between teams and different experts including refinements to achieve alignment and edge-mapping ([Dayaram et al., 2021](#)). This was particularly important in the cross-realm coastal zone when the maps of ecosystem types from the terrestrial, estuarine, and marine realms were aligned (e.g., extending estuarine shores all the way to the back of the surf zone; see [Harris et al., 2019](#) for details).

The MEM 2022 is the latest version of South Africa's evolving map of marine ecosystem types ([Figure 5](#)). It consisted of a relatively minor update, with a focus on alignment with the IUCN GET ([Keith et al., 2020](#)), as discussed above. The benthic-pelagic ecosystems of the MEM 2018 were separated into benthic and pelagic ecosystem type layers in the deepsea (beyond approximately 500 m depth), reintroducing the pelagic components defined by [Roberson et al. \(2017\)](#). South Africa's 163 marine ecosystem types are now nested within 58 biogeographic ecotypes and 17 ecosystem functional groups ([Supplementary Table S2](#)), demonstrating alignment with the IUCN GET.

3. Value and application of an iteratively improving map

The challenges and size of the task of producing a national MEM, especially if it has not been done before, may paralyze efforts to get started. Collating the available and accessible knowledge into a first 'best available' version that could be iteratively improved upon, rather than focusing on perfection, was key to South Africa's marine classification and mapping journey. Producing a first, even if rudimentary, version is better than none. South Africa's national MEM has not only supported the assessment of marine biodiversity at a national scale ([Lombard et al., 2005](#); [Sink et al., 2012, 2019](#)) but has also played a key role in spatial planning and prioritization ([Botts et al., 2020](#); [Kirkman et al., 2021](#); [Harris et al., 2022a,b](#); [Sink et al., 2023](#)), and is being used to support national ecosystem accounts ([Figure 5](#)).

The ecosystem map is a foundational layer for the assessment of ecosystem threat status and protection level ([SANBI and UNEP-WCMC, 2016](#); [Keith et al., 2022](#)) and even in its simplest form, provided

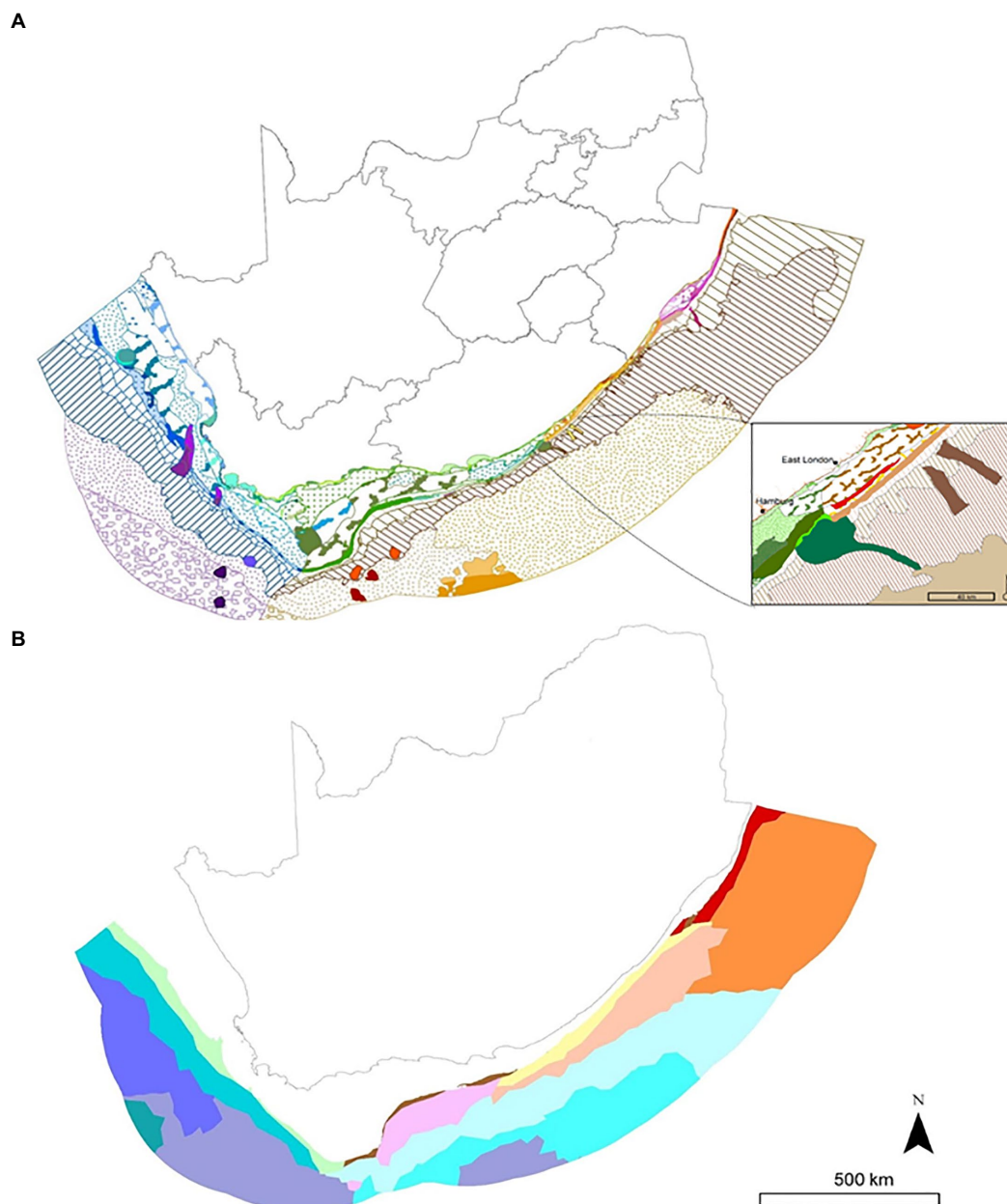
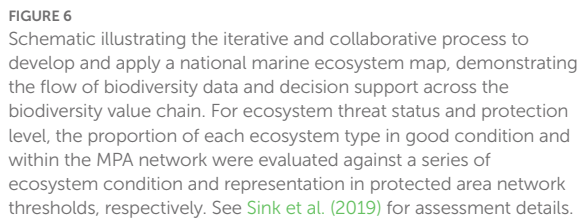


FIGURE 5

Map of the 163 marine ecosystem types in South Africa, including (A) 128 combined benthic-pelagic (shelf) and 22 deep-sea benthic (deep-sea) types and (B) 13 deep-sea pelagic types. The insert on A shows the transition between the Agulhas (green colors) and Natal ecoregions (warm colors). The key is available in Sink et al. (2019) and the map can be explored at this <https://aniday.maps.arcgis.com/apps/webappviewer/index.html?id=d123de04da384e52b53baad8e6ca749c>.

an effective foundation to report on these headline indicators. One of the headline messages from South Africa's first biodiversity assessment was that offshore ecosystems are the most poorly protected among realms (Driver et al., 2005). This message leveraged funding, research, stakeholder engagement and systematic conservation planning that culminated in the proclamation of 20 new Marine Protected Areas in 2019 representing a tenfold increase in area and a dramatic improvement (from 54 to 87%) in ecosystem representation in South Africa's marine protected area network (Kirkman et al., 2021; Sink et al., 2023). South Africa has recently undertaken a review of its Ecologically or Biologically Significant Marine Areas (EBSAs), revising

their delineation and descriptions, assessing their status, and providing management recommendations for each EBSA (Harris et al., 2022a). The MEM was a key dataset for the revised EBSA delineations, especially for assessing the EBSA criteria: biological diversity, and importance for threatened species and habitats, and in the EBSA status assessment. The MEM was also a key component of South Africa's National Coastal and Marine Spatial Biodiversity Plan (Harris et al., 2022b) that in turn underpins the Biodiversity Sector Plan (DFFE, 2022), which is being developed as the biodiversity sector's input to the national marine spatial planning process (Figure 6). Ecosystem characteristics, vulnerabilities, threat status and protection levels are



4. Limitations, challenges, enabling factors, and future opportunities

MEM 2018 that were identified as part of the NBA 2018 process, because recognition and discussion of limitations is an explicit element used to identify research gaps and priorities for improvement in order to focus funders' attention on nationally agreed priorities. The main impediments and most significant enabling factors for the classification and mapping process were identified in task team and committee meetings and reviewed in a dedicated workshop where opportunities for future improvements were jointly distilled.

Current limitations and areas for improvement of the marine ecosystem classification and map in South Africa relate broadly to the need for additional data coverage and quality, particularly in deep water (Supplementary Figure S1), with a recognized need to better document uncertainty in the resulting map (Jansen et al., 2022). Data management and provision platforms such as the Copernicus Marine Environment Monitoring Service (CMEMS) and the European Marine Observation and Data Network (EmodNet) provide leading examples of digital platforms for data acquisition, sharing and production, with potential access to global datasets. Similar regional platforms for the Western Indian Ocean include the Ocean Data and Information Network for Africa (ODINAFRICA)¹ and the African Marine Atlas,² which could also provide data for future improvement of national maps. Identified gaps include improved mapping of Vulnerable Marine Ecosystems, reefs, rhodolith beds (Adams et al., 2020), rocky, mixed and shingle shores, and better understanding of benthic-pelagic coupling and pelagic biodiversity pattern. A limitation for species distribution modeling is the lack of a national bathymetry layer, which could be addressed by collating data from researchers, global databases, the navy and industry to produce one publicly available bathymetry map with the best available data at a national scale. Similarly, nationally collated datasets for oceanographic variables are of high priority, but are currently limited. Development of platforms such as the National Oceans and Coastal Information Management System (OCIMS) may be able to play this role at a national level in future. Linking in and collaborating with existing data platforms and networks, such as ODINAFRICA, the Ocean Info Hub and Indian Ocean Global Ocean Observing System (IO-GOOS), may improve access to data and streamlining data management workflows. An improved understanding of the role of freshwater flow in shelf environments and mapping of fluvial fans is also required. Those ecosystem types that posed challenges in aligning with international classification schemes such as the IUCN GET require further international collaboration to resolve. South Africa also needs to advance conceptual models of its ecosystem functional groups, drawing from, and providing feedback to the IUCN GET team.

South Africa could draw additional lessons from international efforts to collate and share relevant marine ecosystem data and indicators in support of marine management. This includes other international ecosystem classification programs such as the European Nature Information System (EUNIS; [Davies et al., 2004](#)), the marine habitat classification for Britain and Ireland ([Connor et al., 2004](#); [Parry et al., 2015](#)), the Coastal and Marine Classification Standard (CMECS) ([United States Federal Geographic Data Committee, 2012](#)) and the Australian hierarchical seabed classification framework ([Last et al., 2010](#)). European platforms that also support the assessment and

1 <http://odinafrica.org/>

2 www.africanmarineatlas.org

monitoring of environmental status include CMEMS and EModNet, which not only collate, curate and provide co-ordinated ocean observations but also ready-to-use data products and indicators of ocean health for application in ocean state reports and marine environmental decision making. Relevant improvements that South Africa must consider include the provision of robust, well curated open access datasets; advances in data architecture; application of new technologies to improve observations; modeling to address data gaps; development of forecasting approaches; improved workflows to integrate diverse data sources in a more transparent and reproducible framework; innovations in indicator development and steps to improve user uptake (Le Traon et al., 2019; von Schuckmann et al., 2022). Key challenges and barriers to data collation, curation and integration are shown in Table 2 and include capacity, data and analytical challenges. Many scientists were concerned that data were inadequate, outdated or needed further time to be improved or completed before being incorporated into ecosystem mapping. Enabling factors and opportunities for improvement of data, analyses, maps and capacity are also outlined in Table 2. Finally, the reliance on an expert-driven approach leads to potential bias and a lack of transparency in classification, which points to a need for more transparent, repeatable data-driven methods to improve classification and mapping at the national scale.

5. Lessons

South Africa's first lesson from the efforts to classify and map marine ecosystem types is the value of an iterative approach to develop and improve classification and mapping. Dayaram et al. (2019) describe how iterative versions of South Africa's vegetation map have improved the mapping of terrestrial ecosystems and how this translated into improved foundation for biodiversity planning and management (Botts et al., 2020). The inclusion of a simple map of marine biodiversity pattern in 2004 was useful in ensuring that aquatic ecosystems were not omitted from assessments, and also highlighted discrepancies relative to the more advanced terrestrial effort. Increased participation by scientists was catalyzed by omissions or mismatches in the data used in marine ecosystem mapping, and also through evidence of the map being applied in decision support. There was an understanding that the process was iterative and that there would be opportunities to influence the map with contributed data, participation and knowledge, facilitated by scheduled updates to allow scientists to plan their research for future inclusion of their new data.

The second lesson that emerged was the value of a co-ordinated network of practitioners including early-career researchers to support this work, which is a lesson shared in other realms (Botts et al., 2019). Benefits include greater transparency and inclusiveness, sharing of experience and knowledge between multiple fields and institutes thus supporting learning and capacity development, and helping to leverage desirable data. This serves to enhance South Africa's marine ecosystem network (Figure 3), which spans numerous disciplines including geology, geoscience, oceanography, biodiversity and ecology and draws from experience from different habitats and functional ecosystem groups. Inclusion of a dedicated emerging researcher task team within the marine ecosystem network and the participation of these young scientists in the Marine Ecosystem Committee meetings

provides them with opportunities for learning, developing leadership skills and gaining insight into the science-policy interface, while also fostering institutional memories that are important for inter-generational continuity. The network and task teams and the emerging researcher group in particular have allowed for a greater diversity of researchers to participate in South Africa's marine ecosystem work, which is important in transitioning to more inclusive and equitable efforts in marine science and capacity development (Amon et al., 2022; Harden-Davies et al., 2022). The established governance structures with multiple teams and committees provide opportunities for discussions that facilitate agreed priorities, highlight key gaps and knowledge shortfalls, and identify potential expertise, datasets, approaches or research opportunities to address gaps. Our experience accords with the expectations shared by Keith et al. (2022) in the context of the IUCN GET, where highlighting areas of limited knowledge promote research to fill significant gaps. In our experience, a formally appointed champion (individual or research team within an appropriately mandated institute) that can drive this process and integrate results as they become available is an important element in coordination, fostering research relationships and facilitating progress.

The third lesson relates to the importance of building and curating key foundational biodiversity datasets and research to understand drivers of biodiversity patterns. The majority of *in situ* marine biological or ecological data collection that is undertaken is focused on relatively narrow geographic extents and optimized to address specific hypotheses, rather than to map biodiversity. The development of national-scale maps of ecosystem types, however, requires expansive foundational data layers that inform biodiversity and environmental patterns. To produce these layers, substantial effort is often required to collate, 'clean' and prepare datasets from many sources into a single, standardized format. In developed countries this is typically a function of national data centers, but such infrastructure and services are often lacking in developing countries. The importance of curation, documentation, expansion and iterative improvement of these datasets is a key requirement to build improved biodiversity knowledge and maps over time (SANBI and UNEP-WCMC, 2016) and helps identify gaps that can guide future research priorities.

The benefit of identifying gaps and research priorities is that it can help funders focus their funding calls to strategically address key gaps. In South Africa, the Foundational Biodiversity Information Programme (FBIP)³ has played an enabling role in funding research to collect, collate and prepare foundational biodiversity data used in South Africa's Marine Ecosystem classification and map. Similarly, the Department of Science and Innovation/National Research Foundation funded African Coelacanth Ecosystem Programme (ACEP) provides the research community with competitive access to funding and marine research platforms and infrastructure (such as remote imagery and mapping platforms) provided through the South African Institute for Aquatic Biodiversity. The ACEP Open Call draws research priorities from among other strategy documents, the National Biodiversity Assessment which communicates research priorities identified by the Marine Ecosystem Committee, to ensure that the research supported is relevant and addresses gaps in ecosystem understanding. Research supported by ACEP has contributed to ecosystem classification and mapping

³ <https://fbip.co.za/>

TABLE 2 Challenges, enabling factors, and further opportunities for improvement, as identified by authors during reflections from more than a decade of efforts to improve ecosystem maps.

	Challenges and constraints	Enabling factors	Opportunities for improvement
General capacity	<ul style="list-style-type: none"> Human resources capacity, including coordination, data management and analytical skills 	<ul style="list-style-type: none"> Commitment to iterative improvement 	<ul style="list-style-type: none"> Support strategic investment in training and job positions where capacity is lacking
	<ul style="list-style-type: none"> Financial resources, including difficulties in funding data collation, curation, preparation and analyses 	<ul style="list-style-type: none"> Network of practitioners and cross-disciplinary dialog 	<ul style="list-style-type: none"> Strengthen international and national collaboration
	<ul style="list-style-type: none"> Institutional memory limited to few individuals 	<ul style="list-style-type: none"> Academic and industry partnerships 	<ul style="list-style-type: none"> Focus on well documented, reproducible and iterative workflows
		<ul style="list-style-type: none"> Student projects and intern support 	<ul style="list-style-type: none"> Stagger outputs with ecosystem maps and assessment components completed in different years
		<ul style="list-style-type: none"> Receptive funders including research priorities in funding calls 	
		<ul style="list-style-type: none"> Emerging researcher task team 	
Data acquisition	<ul style="list-style-type: none"> Lack of data, caused by survey gaps and undiscoverable, inaccessible, dispersed data 	<ul style="list-style-type: none"> A focus on building key foundational datasets 	<ul style="list-style-type: none"> Commit to improved data protocols, standards and version control aligned with international best practice and empower partners to apply these
	<ul style="list-style-type: none"> Existing data frequently not fit for purpose 	<ul style="list-style-type: none"> Investment in targeted research to address spatial survey gaps 	<ul style="list-style-type: none"> Thoroughly document data and transparent data pipelines
	<ul style="list-style-type: none"> Poor data quality and inter-operability challenges due to lacking application of data and metadata standards; 	<ul style="list-style-type: none"> Commitment to iterative data improvement 	<ul style="list-style-type: none"> Prioritize data management training and positions in research teams
	<ul style="list-style-type: none"> Lacking institutional data storage and curation expertise 	<ul style="list-style-type: none"> Network of practitioners and champions to provide data 	<ul style="list-style-type: none"> Improve, contribute and draw from national data repositories
	<ul style="list-style-type: none"> Lacking centralized coordination in data collation, and curation 	<ul style="list-style-type: none"> Industry and academic partnerships to leverage inaccessible data 	<ul style="list-style-type: none"> Use open access data and open-source software and non-proprietary storage solutions to ensure long-term, equitable access.
		<ul style="list-style-type: none"> Identification and wide communication of priority gaps and opportunities 	<ul style="list-style-type: none"> Improve data acquisition and processing (especially genetic and imagery data) by adopting innovative new technologies
		<ul style="list-style-type: none"> Developing research infrastructure platforms and improved data protocols 	<ul style="list-style-type: none"> Work with research agencies to produce data fit for purpose and support long-term monitoring surveys
Management and analyses	<ul style="list-style-type: none"> Lack of technical skills, methods and personnel to effectively integrate patchy data for classification and mapping 	<ul style="list-style-type: none"> Network of practitioners with cross-realm experience in ecosystem assessment and spatial planning 	<ul style="list-style-type: none"> Commit to collaborate with statistical ecologists locally and internationally to explore and apply emerging statistical approaches for ecosystem classification and mapping
	<ul style="list-style-type: none"> Short timeframes for co-ordinated national ecosystem assessments 	<ul style="list-style-type: none"> Commitment to iterative analytical improvement 	<ul style="list-style-type: none"> Employ scripted and well-documented workflows and version control to enhance reproducibility and transparency
	<ul style="list-style-type: none"> Few case studies to learn from, especially in a developing country context 	<ul style="list-style-type: none"> Strategic student projects 	<ul style="list-style-type: none"> Access high-performance computing centers
	<ul style="list-style-type: none"> Computational limitations 	<ul style="list-style-type: none"> Establishment of statistical task team to guide and improve analyses for classification and mapping 	

These were grouped into three categories relative to general capacity, data acquisition and management and analyses noting these are related and cross-cutting.

through the provision of data layers and dedicated research outputs in the form of several student theses and peer-reviewed publications.

The fourth lesson is the importance of developing capacity and transferring skills and tools that enhance data quality, analytical workflows and reproducible outputs. Developing transparent and repeatable workflows requires meticulous organization and documentation of datasets, data preparation, analyses and all decisions along the way (Wilson et al., 2017). Accomplishing this for a wide range of input data from different fields and many different institutions, with a small, constrained team, is a daunting undertaking. Fortunately, data science has adopted and developed many highly effective tools, most of them open source and freely available, that enable improved documentation, management and sharing of data and analyses. Developing such data science and data management skills within the team is critical to working 'smarter' and building repeatable workflows. This philosophy and its benefits to iterative programs of work are well demonstrated by Lowndes et al. (2017) in their annual Ocean Health Index assessments. The code, systems and approaches developed by better-capacitated international teams can be shared and adopted to greatly benefit resource-limited teams in developing countries.

Lastly, the fifth lesson is the importance of innovation and adopting modern solutions and emerging statistical tools to advance ecosystem classification and mapping. Rapid advances in the areas of genomics (van Oppen and Coleman, 2022), visual survey methodologies (Mallet and Pelletier, 2014) and machine learning (Beyan and Browman, 2020) have unlocked big data streams for inclusion in the mapping process. Improved environmental and biological data access and extent worldwide has led to increasing focus on data-driven approaches to marine ecosystem classification and mapping (Howell, 2010; Shumchenia and King, 2010; Hill et al., 2020; Woolley et al., 2020). These approaches are robust, transparent, repeatable and may include measures of uncertainty (Jansen et al., 2022), making use of the most complete available data for classifying and predicting biodiversity patterns. The existence of earlier versions of ecosystem maps may necessitate some level of continuity between successive iterations, which can be achieved through expert-based decisions or decision trees. This is feasible due to recent growth in the field of statistical ecology in Africa (Minoarivelo et al., 2021). Three aspects to consider when pursuing data-driven statistical approaches, include: (i) the types of data that are available; (ii) surrogacy (Mellin et al., 2011; Flannery and Przeslawski, 2015), prioritization or integration of datasets (Zipkin et al., 2021) at a relevant scale; and (iii) the statistical method that best captures or predicts biodiversity patterns at this scale (e.g., Verfaillie et al., 2009; Murillo et al., 2018; Hill et al., 2020; McQuaid et al., 2020). Due to the scale mismatch between localized project-based biodiversity research and the standardized, national-scale biodiversity datasets required for ecosystem mapping (Sink et al., 2019), there is typically a need for prioritization or integration of multiple datasets covering different regions and assemblages, or consideration of surrogacy in some cases. Though statistical data integration methods exist, these require parameters or species in common (Zipkin et al., 2021), which is not always the case for datasets collected using different methods such as eDNA, trawls, dredges, grabs, and visual surveys (Lange et al., 2014; Flannery and Przeslawski, 2015). Emerging statistical tools for marine ecosystem classification rely on multivariate biodiversity data, either environmental or biological (or both), and include a technique of clustering the data and predicting the clusters across space (e.g.,

Verfaillie et al., 2009; Howell, 2010; Ovaskainen et al., 2017; Murillo et al., 2018; Hill et al., 2020). Currently, there is no accepted standard methodology and multiple emerging methods can be explored to classify and predict biodiversity patterns from growing datasets.

6. Conclusion

A co-ordinated and inclusive process enabled the collation and improvement of multi-disciplinary data to advance ecosystem mapping in a developing country. The national MEM has supported spatial biodiversity assessment, a tenfold increase in MPA estate, the identification of biodiversity priority areas to inform marine spatial planning, and has been applied in biodiversity management. Reflecting on progress, limitations and failures in South Africa reveals important lessons for others working to synthesize multiple datasets for robust decision support and improved biodiversity management. Key lessons include the value of (1) a clear commitment to iterative improvement even if the ideal frameworks or data are lacking; (2) developing relationships and skills among a network of current and emerging practitioners; (3) strategically prioritizing and leveraging resources to build and curate key foundational biodiversity datasets; (4) enhancing data quality and analytical workflows; and (5) the application of new technology and emerging statistical tools to improve the classification and prediction of biodiversity patterns. South Africa's marine ecosystem mapping process provides insights that are relevant for developing countries but are also applicable in global efforts to unlock, share and use multi-disciplinary data for better biodiversity decisions.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <http://bgis.sanbi.org/SpatialDataset/Detail/2681>.

Author contributions

KS, NK, LH, AS, JC, LJA, and LAA contributed to the conception and planned content of this manuscript. KS, LAA, M-LF, JC, LH, SP, NK, AD, AS, ML-V, LJA, LD, SK, and RP wrote sections of the manuscript. AD, LAA, M-LF, MP, and KS produced the figures. All authors contributed data and analyses to the 2018 and 2022 versions of the ecosystem map and reviewed the draft manuscript. Most authors contributed to the reflections and lessons emanating from this manuscript.

Funding

The development and application of national maps of ecosystem types is part of the mandate of the South African National Biodiversity Institute. This manuscript was feasible due to funding received from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 862428 (MISSION ATLANTIC). KS and NK acknowledge the ACEP Deep Connections, Agulhas Bank

Connections and Deep Forests projects funded through the National Research Foundation (NRF Grants 129216, 129213, and 110765). LJA acknowledges support from the SeaMap project funded through the NRF Grant number 138572. LH and SH were supported by the Benguela Current Marine Spatial Management and Governance (MARISMA) Project. The MARISMA Project is funded by the German Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) through its International Climate Initiative (ICI), with in-kind contributions by the Benguela Current Commission (BCC) and its contracting parties (South Africa, Namibia and Angola). It is implemented by GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH; German Development Cooperation) in partnership with the BCC and its contracting parties. LH, LVN, SH, KS, MP, LD, and SK are also supported by CoastWise, a project of the MeerWissen initiative, which is funded by the German Federal Ministry for Economic Cooperation and Development, and implemented by GIZ.

Acknowledgments

We acknowledge all participants in South Africa's three national biodiversity assessments and the broader communities of practice that support spatial biodiversity assessment and planning in South Africa. South Africa's first marine assessment and first map of marine biozones was led by Amanda Lombard. We thank Kristal Maze and Mandy Driver for vision and leadership in supporting the science to policy value chain for biodiversity in South Africa. We acknowledge Megan van der Bank for her work in co-ordination and support of the marine ecosystem network and Prideel Majiedt for her technical support and

policy insights. We thank Lynne Shannon for sharing perspectives on benthic-pelagic interactions. We acknowledge Robyn Adams for her contributions to Figure 5 in the form of the schematic component depicting the ecosystem assessment. The organizations and departments of all authors are acknowledged for their support as is the National Research Foundation and the African Coelacanth Ecosystem Program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2023.1108118/full#supplementary-material>

References

- Adams, L. A., Maneveldt, G. W., Green, A., Karenyi, N., Parker, D., Samaai, T., et al. (2020). Rhodolith bed discovered off the south African coast. *Diversity* 12:125. doi: 10.3390/d12040125
- Amon, D. J., Filander, Z., Harris, L., and Harden-Davies, H. (2022). Safe working environments are key to improving inclusion in open-ocean, deep-ocean, and high-seas science. *Mar. Policy* 137:104947. doi: 10.1016/j.marpol.2021.104947
- Andrews, J., Scarcella, G., and Pierre, J. (2022). *South African Hake Fishery Surveillance Report*. South Africa: Marine Stewardship Council Fisheries Assessments.
- Balmford, A., Bruner, A., Cooper, P., Costanza, R., Farber, S., Green, R. E., et al. (2002). Economic reasons for conserving wild nature. *Science* 297, 950–953. doi: 10.1126/science.1073947
- Bell, K. L. C., Chow, J. S., Hope, A., Quinzin, M. C., Cantner, K. A., Amon, D. J., et al. (2022b). Low-cost, deep-sea imaging and analysis tools for deep-sea exploration: a collaborative design study. *Front. Mar. Sci.* 9:873700. doi: 10.3389/fmars.2022.873700
- Bell, K. L. C., Quinzin, M. C., Sarti, O., Cañete, T., Smith, A., Baldwin, H., et al. (2022a). "Global Summary," in 2022 *Global Deep-Sea Capacity Assessment*. eds. K. L. C. Bell, M. C. Quinzin, S. Poulton, A. Hope and D. Amon (Saunderstown, USA: Ocean Discovery League).
- Beyan, C., and Browman, H. I. (2020). Setting the stage for the machine intelligence era in marine science. *ICES J. Mar. Sci.* 77, 1267–1273. doi: 10.1093/icesjms/fsaa084
- Birch, G. F., Rogers, J., and Bremner, J. M. (1986). "Texture and composition of sediments of the continental margin of the republics of South Africa, Transkei and Ciskei," in *Marine Geoscience Series* 3, ed. Plessis, A. Du Pretoria: Department of Mineral and Energy Affairs, Government Printer, 1–4.
- Bland, L., Keith, D., Miller, R., Murray, N., and Rodríguez, J. (2017). *Guidelines for the Application of IUCN Red List of Ecosystems Categories and Criteria, Version 1.1*. Gland: IUCN.
- Bogaart, P., Chan, J. Y., Horlings, H., Keith, D., Larson, T., Sayre, R., et al. (2019). Discussion paper 1.1: An ecosystem type classification for the SEEA EEA. Paper submitted to the SEEA EEA Technical Committee as input to the revision of the technical recommendations in support of the System on Environmental-Economic Accounting.
- Borja, Á., Elliott, M., Carstensen, J., Heiskanen, A.-S., and van de Bund, W. (2010). Marine management—towards an integrated implementation of the European marine strategy framework and the water framework directives. *Mar. Pollut. Bull.* 60, 2175–2186. doi: 10.1016/j.marpolbul.2010.09.026
- Botts, E. A., Pence, G., Holness, S., Sink, K., Skowno, A., Driver, A., et al. (2019). Practical actions for applied systematic conservation planning. *Conserv. Biol.* 33, 1235–1246. doi: 10.1111/cobi.13321
- Botts, E. A., Skowno, A., Driver, A., Holness, S., Maze, K., Smith, T., et al. (2020). More than just a (red) list: over a decade of using South Africa's threatened ecosystems in policy and practice. *Biol. Conserv.* 246:108559. doi: 10.1016/j.biocon.2020.108559
- Briggs, J. C., and Bowen, B. W. (2012). A realignment of marine biogeographic provinces with particular reference to fish distributions. *J. Biogeogr.* 39, 12–30. doi: 10.1111/j.1365-2699.2011.02613.x
- Connor, D. W., Allen, J. H., Golding, N., Howell, K. L., Lieberknecht, L. M., Northen, K. O., et al. (2004). *The marine habitat classification for Britain and Ireland*. Version 04.05 Infralittoral Rock Section.
- Costello, M. J. (2009). Distinguishing marine habitat classification concepts for ecological data management. *Mar. Ecol. Prog. Ser.* 397, 253–268. doi: 10.3354/meps08317
- Costello, M. J., Coll, M., Danovaro, R., Halpin, P., Ojaveer, H., and Miloslavich, P. (2010). A census of marine biodiversity knowledge, resources, and future challenges. *PLoS One* 5:e12110. doi: 10.1371/journal.pone.0012110
- Davies, C. E., Moss, D., and Hill, M. O. (2004). *EUNIS habitat classification revised 2004*. Report to: European environment agency-European topic Centre on nature protection and biodiversity, 127–143.
- Dayaram, A., Harris, L. R., Grobler, B. A., Van der Merwe, S., Rebelo, A. G., Ward Powrie, L., et al. (2019). Vegetation map of South Africa, Lesotho and Swaziland 2018: a description of changes since 2006. *Bothalia African Biodiver. Conserv.* 49, 1–11. doi: 10.4102/abc.v49i1.2452
- Dayaram, A., Skowno, A. L., Driver, A., Sink, K., Van Deventer, H., Smith-Adao, L., et al. (2021). *South African National Ecosystem Classification System Handbook: 1st Edn*. Pretoria: South African National Biodiversity Institute.
- de Wet, W. M., and Compton, J. S. (2021). Bathymetry of the south African continental shelf. *Geo-Mar. Lett.* 41, 1–19. doi: 10.1007/s00367-021-00701-y

- DFFE (2022). *Biodiversity Management Plans*, Department of Forestry, Fisheries and the Environment, Pretoria.
- Dingle, R. V. (1986). *Revised bathymetric map of the cape canyon*. Technical Report, Joint Geological Survey, University of Cape Town Marine Geoscience Unit 16, 20–25.
- Dingle, R., Birch, G., Bremner, J., de Decker, R., du Plessis, A., Engelbrecht, J., et al. (1987). Deep-sea sedimentary environments around southern Africa south-East Atlantic and south-west Indian oceans. *Ann. South African Museum* 98, 1–27.
- Douglass, L. L., Turner, J., Grantham, H. S., Kaiser, S., Constable, A., Nicoll, R., et al. (2014). A hierarchical classification of benthic biodiversity and assessment of protected areas in the Southern Ocean. *PLoS One* 9:e100551. doi: 10.1371/journal.pone.0100551
- Dove, D., Acoba, T., and DesRochers, A. (2018). *Seafloor substrate characterization from shallow reefs to the abyss: Spatially-continuous seafloor mapping using multispectral satellite imagery, and multibeam bathymetry and backscatter data within the Pacific Remote Islands marine National Monument and the Main Hawaiian islands*. PIFSC Internal Report IR-18-08. Issued 18 October 2018.
- Driver, A., Maze, K., Rouget, M., Lombard, A., Nel, J., Turpie, J., et al. (2005). *National Spatial Biodiversity Assessment 2004: Priorities for Biodiversity Conservation in South Africa*. Pretoria: South African National Biodiversity Institute.
- Driver, A., Sink, K. J., Nel, J. N., Holness, S., Van Niekerk, L., Daniels, F., et al. (2012). *National Biodiversity Assessment 2011: An Assessment of South Africa's Biodiversity and Ecosystems, Synthesis Report*. Pretoria: South African National Biodiversity Institute and Department of Environmental Affairs.
- FAO (2009). *International Guidelines for the Management of Deep-Sea Fisheries in the High Seas*. Rome: Food and Agriculture Organization of the United Nations.
- Flannery, E., and Przeslawski, R. (2015). Comparison of sampling methods to assess benthic marine biodiversity. Are spatial and ecological relationships consistent among sampling gear? *Geosci. Australia*. 007–71. doi: 10.11636/Record.2015.007
- Galparsoro, I., Connor, D. W., Borja, Á., Aish, A., Amorim, P., Bajjouk, T., et al. (2012). Using EUNIS habitat classification for benthic mapping in European seas: present concerns and future needs. *Mar. Pollut. Bull.* 64, 2630–2638. doi: 10.1016/j.marpolbul.2012.10.010
- Gerovasileiou, V., Smith, C. J., Sevastou, K., Papadopoulou, N., Dailianis, T., Bekkby, T., et al. (2019). Habitat mapping in the European seas—is it fit for purpose in the marine restoration agenda? *Mar. Policy* 106:103521. doi: 10.1016/j.marpol.2019.103521
- Gibbons, M. (1999). The taxonomic richness of South Africa's marine fauna: a crisis at hand. *S. Afr. J. Sci.* 95, 8–12.
- Gregg, E. J., Ahrens, A. L., and Ian Perry, R. (2012). Reconciling classifications of ecologically and biologically significant areas in the world's oceans. *Mar. Policy* 36, 716–726. doi: 10.1016/j.marpol.2011.10.009
- Gregg, E. J., and Bodtker, K. M. (2007). Adaptive classification of marine ecosystems: identifying biologically meaningful regions in the marine environment. *Deep Sea Res. I Oceanogr. Res. Pap.* 54, 385–402. doi: 10.1016/j.dsr.2006.11.004
- Griffiths, C. L., Robinson, T. B., Lange, L., and Mead, A. (2010). Marine biodiversity in South Africa: an evaluation of current states of knowledge. *PLoS One* 5:e12008. doi: 10.1371/journal.pone.0012008
- Harden-Davies, H., Amon, D. J., Vierros, M., Bax, N. J., Hanich, Q., Hills, J. M., et al. (2022). Capacity development in the ocean decade and beyond: key questions about meanings, motivations, pathways, and measurements. *Earth System Governance* 12:100138. doi: 10.1016/j.esg.2022.100138
- Harris, L. R., Bessinger, M., Dayaram, A., Holness, S., Kirkman, S., Livingstone, T.-C., et al. (2019). Advancing land-sea integration for ecologically meaningful coastal conservation and management. *Biol. Conserv.* 237, 81–89. doi: 10.1016/j.biocon.2019.06.020
- Harris, L. R., Holness, S. D., Finke, G., Amunyele, M., Braby, R., Coelho, N., et al. (2022a). Practical marine spatial management of ecologically or biologically significant marine areas: emerging lessons from evidence-based planning and implementation in a developing-world context. *Front. Mar. Sci.* 9:678. doi: 10.3389/fmars.2022.831678
- Harris, L. R., Holness, S. D., Kirkman, S. P., Sink, K. J., Majiedt, P., and Driver, A. (2022b). A robust, systematic approach for developing the biodiversity sector's input for multi-sector marine spatial planning. *Ocean Coastal Manag.* 230:106368. doi: 10.1016/j.ocecoaman.2022.106368
- Harris, L., Nel, R., and Schoeman, D. (2011). Mapping beach morphodynamics remotely: a novel application tested on south African sandy shores. *Estuar. Coast. Shelf Sci.* 92, 78–89. doi: 10.1016/j.eess.2010.12.013
- Hein, L., Bagstad, K. J., Obst, C., Edens, B., Schenau, S., Castillo, G., et al. (2022). Progress in natural capital accounting for ecosystems. *Science* 367, 514–515. doi: 10.1126/science.aaz890
- Hill, N., Woolley, S. N. C., Foster, S., Dunstan, P. K., McKinlay, J., Ovaskainen, O., et al. (2020). Determining marine bioregions: a comparison of quantitative approaches. *Methods Ecol. Evol.* 11, 1258–1272. doi: 10.1111/2041-210X.13447
- Holness, S. D., Harris, L. R., Chalmers, R., De Vos, D., Goodall, V., Truter, H., et al. (2022). Using systematic conservation planning to align priority areas for biodiversity and nature-based activities in marine spatial planning: a real-world application in contested marine space. *Biol. Conserv.* 271:109574. doi: 10.1016/j.biocon.2022.109574
- Howell, K. L. (2010). A benthic classification system to aid in the implementation of marine protected area networks in the deep/high seas of the NE Atlantic. *Biol. Conserv.* 143, 1041–1056. doi: 10.1016/j.biocon.2010.02.001
- Hu, W., Zhou, Q., Chen, B., Yang, S., Xiao, J., Du, J., et al. (2021). Progress in marine habitat mapping: concept, methods, and applications. *Biodivers. Sci.* 29, 531–544. doi: 10.17520/biods.2020176
- Huse, I., Hamukuaya, H., Boyer, D. C., Malan, P. E., and Strømme, T. (1998). The diurnal vertical dynamics of cape hake and their potential prey. *S. Afr. J. Mar. Sci.* 19, 365–376. doi: 10.2989/025776198784126746
- Jackson, L. F., and Lipschitz, S. (1984). *Coastal Sensitivity Atlas of Southern Africa*. Goodwood, Cape Town: National Printing Press.
- Jansen, J., Woolley, S. N. C., Dunstan, P. K., Foster, S. D., Hill, N. A., Haward, M., et al. (2022). Stop ignoring map uncertainty in biodiversity science and conservation policy. *Nat. Ecol. Evol.* 6, 828–829. doi: 10.1038/s41559-022-01778-z
- Keith, D. A., Ferrer-Paris, J. R., Nicholson, E., Bishop, M. J., Polidoro, B. A., Ramirez-Llodra, E., et al. (2022). A function-based typology for Earth's ecosystems. *Nature* 610, 513–518. doi: 10.1038/s41586-022-05318-4
- Keith, D. A., Ferrer-Paris, J. R., Nicholson, E., and Kingsford, R. T. (2020). *IUCN Global Ecosystem Typology 2.0: Descriptive Profiles for Biomes and Ecosystem Functional Groups*. Gland, Switzerland: IUCN, International Union for Conservation of Nature.
- Kirkman, S., Mann, B., Sink, K., Adams, R., Livingstone, T., Mann-Lang, J., et al. (2021). Evaluating the evidence for ecological effectiveness of South Africa's marine protected areas. *Afr. J. Mar. Sci.* 43, 389–412. doi: 10.2989/1814232X.2021.1962975
- Lange, E., Petersen, S., Rüpke, L., Söding, E., and Wallmann, K. (2014). Marine resources—opportunities and risks. *World Ocean Review* 3:165.
- Last, P. R., Lyne, V. D., Williams, A., Davies, C. R., Butler, A. J., and Yearsley, G. K. (2010). A hierarchical framework for classifying seabed biodiversity with application to planning and managing Australia's marine biological resources. *Biol. Conserv.* 143, 1675–1686. doi: 10.1016/j.biocon.2010.04.008
- Le Traon, P. Y., Reppucci, A., Alvarez Fanjul, E., Aouf, L., Behrens, A., Belmonte, M., et al. (2019). From observation to information and users: the Copernicus marine service perspective. *Front. Mar. Sci.* 6:234. doi: 10.3389/fmars.2019.00234
- Lombard, A., Strauss, T., Harris, J., Sink, K., Attwood, C., and Hutchings, L. (2005). *National Spatial Biodiversity Assessment 2004: Technical Report. Volume 4: Marine Component*. Pretoria: South African National Biodiversity Institute.
- Longhurst, A. R. (2007). “Toward an ecological geography of the sea,” in *Ecological Geography of the Sea* (San Diego: Academic Press), 1–17.
- Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., et al. (2017). Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.* 1:0160. doi: 10.1038/s41559-017-0160
- Mallet, D., and Pelletier, D. (2014). Underwater video techniques for observing coastal marine biodiversity: a review of sixty years of publications (1952–2012). *Fish. Res.* 154, 44–62. doi: 10.1016/j.fishres.2014.01.019
- McArthur, M. A., Brooke, B. P., Przeslawski, R., Ryan, D. A., Lucieer, V. L., Nichol, S., et al. (2010). On the use of abiotic surrogates to describe marine benthic biodiversity. *Estuar. Coast. Shelf Sci.* 88, 21–32. doi: 10.1016/j.eess.2010.03.003
- McQuaid, K. A., Attrill, M. J., Clark, M. R., Cobley, A., Glover, A. G., Smith, C. R., et al. (2020). Using habitat classification to assess Representativity of a protected area network in a large, data-poor area targeted for Deep-Sea mining. *Front. Mar. Sci.* 7:8860. doi: 10.3389/fmars.2020.558860
- Mellin, C., Delean, S., Caley, J., Edgar, G., Meekan, M., Pitcher, R., et al. (2011). Effectiveness of biological surrogates for predicting patterns of marine biodiversity: a global meta-analysis. *PLoS One* 6:e20141. doi: 10.1371/journal.pone.0020141
- Minoarivelo, H. O., Peralta, F. C., and Kuiper, T. (2021). Statistical Ecology can Unlock the Power of Biodiversity Data in Africa. The Conversation Africa. Available at: <https://theconversation.com/statistical-ecology-can-unlock-the-power-of-biodiversity-data-in-africa-171513> (Accessed February 23, 2022).
- Mucina, L. (2019). Biome: evolution of a crucial ecological and biogeographical concept. *New Phytol.* 222, 97–114. doi: 10.1111/nph.15609
- Murillo, F. J., Kenchington, E., Tompkins, G., Beazley, L., Baker, E., Knudby, A., et al. (2018). Sponge assemblages and predicted archetypes in the eastern Canadian Arctic. *Mar. Ecol. Prog. Ser.* 597, 115–135. doi: 10.3354/meps12589
- Nikolopoulou, S., Berov, D., Klayn, S., Dimitrov, L. I., Velkovsky, K., Chatzinikolaou, E., et al. (2021). Benthic habitat mapping of Plazh Gradina – Zlatna ribka (Black Sea) and Karpathos and Saria Islands (Mediterranean Sea). *Biodivers. Data J.* 9:e71972. doi: 10.3897/BDJ.9.e71972
- Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., et al. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecol. Lett.* 20, 561–576. doi: 10.1111/ele.12757
- Parry, M., Howell, K. L., Narayanaswamy, B., Bett, B., Jones, D. O. B., Hughes, D., et al. (2015). *A Deep-Sea Section for the Marine Habitat Classification of Britain and Ireland*, Peterborough: JNCC.
- Pillar, S. C., and Barange, M. (1995). Diel feeding periodicity, daily ration and vertical migration of juvenile cape hake off the west coast of South Africa. *J. Fish Biol.* 47, 753–768. doi: 10.1111/j.1095-8649.1995.tb06000.x
- Poole, C. J., Skowno, A. L., Currie, J. C., Sink, K. J., Daly, B., and von Staden, L. (2023). Taking state of biodiversity reporting into the information age – a south African perspective. *Front. Ecol. Evol.* In press

- Porter, S. N., Branch, G. M., and Sink, K. J. (2017). Changes in shallow-reef community composition along environmental gradients on the east African coast. *Mar. Biol.* 164:101. doi: 10.1007/s00227-017-3130-0
- Roberson, L. A., Lagabriele, E., Lombard, A. T., Sink, K., Livingstone, T., Grantham, H., et al. (2017). Pelagic bioregionalisation using open-access data for better planning of marine protected area networks. *Ocean Coastal Manag.* 148, 214–230. doi: 10.1016/j.ocecoaman.2017.08.017
- SANBI and UNEP-WCMC (2016). *Mapping Biodiversity Priorities: A Practical, Science-Based Approach to National Biodiversity Assessment and Prioritisation to Inform Strategy and Action Planning*. Cambridge, UK: UNEP-WCMC.
- Sayre, R. G., Wright, D. J., Breyer, S. P., Butler, K. A., Van Graafeiland, K., Costello, M. J., et al. (2017). A three-dimensional mapping of the ocean based on environmental data. *Oceanography* 30, 90–103. doi: 10.5670/oceanog.2017.116
- Shumchenia, E. J., and King, J. W. (2010). Comparison of methods for integrating biological and physical data for marine habitat mapping and classification. *Cont. Shelf Res.* 30, 1717–1729. doi: 10.1016/j.csr.2010.07.007
- Sink, K. J. (2016). *Operation Phakisa African Coelacanth Ecosystem Programme (ACEP) Deep Secrets Cruise: The Outer Shelf and Slope Ecosystems of South Africa–RV Algoa voyage 230*. Pretoria: South African National Biodiversity Institute.
- Sink, K. J., Harris, L. R., Skowno, A. L., Livingstone, T., Franken, M., Porter, S., et al. (2019). “Chapter 3: marine ecosystem classification and mapping,” in *South African National Biodiversity Assessment 2018 Technical Report Volume 4: Marine Realm*, eds K. J. Sink, Bank M. G. van der, P. A. Majiedt, L. R. Harris, L. J. Atkinson and S. P. Kirkman et al. Pretoria: South African National Biodiversity Institute.
- Sink, K. J., Holness, S., Harris, L., Majiedt, P., Atkinson, L., Robinson, T., et al. (2012). *National Biodiversity Assessment 2011: Technical Report, Volume 4: Marine and Coastal Component*. Pretoria: South African National Biodiversity Institute, 328.
- Sink, K. J., Lombard, A. T., Attwood, C. G., Livingstone, T., Grantham, H., and Holness, S. D. (2023). Integrated systematic planning and adaptive stakeholder process support tenfold increase in South Africa's marine protected area estate. *Conservation Letters*. In review.
- Skowno, A. L., Poole, C. J., Raimondo, D. C., Sink, K. J., van Deventer, H., Van Niekerk, L., et al. (2019). *National Biodiversity Assessment 2018: The Status of South Africa's Ecosystems and Biodiversity, Synthesis Report*. Pretoria: South African National Biodiversity Institute, an entity of the Department of Environment, Forestry and Fisheries.
- Smit, K. P., Van Niekerk, L., Harris, L. R., McQuatters-Gollop, A., Shannon, L. J., and Sink, K. J. (2022). A roadmap to advance marine and coastal monitoring, biodiversity assessment, and international reporting: a developing nation perspective. *Front. Mar. Sci.* 9:6373. doi: 10.3389/fmars.2022.886373
- Spalding, M. D., Agostini, V. N., Rice, J., and Grant, S. M. (2012). Pelagic provinces of the world: a biogeographic classification of the world's surface pelagic waters. *Ocean Coastal Manag.* 60, 19–30. doi: 10.1016/j.ocecoaman.2011.12.016
- Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., Ferdaña, Z. A., Finlayson, M., et al. (2007). Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. *Bioscience* 57, 573–583. doi: 10.1641/b570707
- Sutton, T. T., Clark, M. R., Dunn, D. C., Halpin, P. N., Rogers, A. D., Guinotte, J., et al. (2017). A global biogeographic classification of the mesopelagic zone. *Deep Sea Res. I Oceanogr. Res. Pap.* 126, 85–102. doi: 10.1016/j.dsr.2017.05.006
- Swanborn, D. J. B., Huvenne, V. A. I., Pittman, S. J., and Woodall, L. C. (2021). Bringing seascape ecology to the deep seabed: a review and framework for its application. *Limnol. Oceanogr.* 67, 66–88. doi: 10.1002/lno.11976
- United States Federal Geographic Data Committee (2012). *Coastal and Marine Ecological Classification Standard (CMECS)*. Silver Spring, MD: NOAA Repository.
- van Niekerk, L., Adams, J., James, N., Lamberth, S., MacKay, C., Turpie, J., et al. (2020). An estuary ecosystem classification that encompasses biogeography and a high diversity of types in support of protection and management. *Afr. J. Aquat. Sci.* 45, 199–216. doi: 10.2989/16085914.2019.1685934
- van Oppen, M. J. H., and Coleman, M. A. (2022). Advancing the protection of marine life through genomics. *PLoS Biol.* 20:e3001801. doi: 10.1371/journal.pbio.3001801
- Verfaillie, E., Degraer, S., Schelfaut, K., Willems, W., and Van Lancker, V. (2009). A protocol for classifying ecologically relevant marine zones, a statistical approach. *Estuar. Coast. Shelf Sci.* 83, 175–185. doi: 10.1016/j.ecss.2009.03.003
- von Schuckmann, K. P., Le Traon, N., Smith, A., Pascual, S., Djavidnia, P., Brasseur, M., et al. (2022). Copernicus Ocean state report, issue 6. *J. Operat. Oceanograph.* 15, 1–220. doi: 10.1080/1755876X.2022.2095169
- von Staden, L., Lötter, M. C., Holness, S., and Lombard, A. T. (2022). An evaluation of the effectiveness of critical biodiversity areas, identified through a systematic conservation planning process, to reduce biodiversity loss outside protected areas in South Africa. *Land Use Policy* 115:106044. doi: 10.1016/j.landusepol.2022.106044
- Whitehead, T. O., Von der Meden, C. E., Skowno, A. L., Sink, K. J., Van der Merwe, S., Adams, R., et al. (2019). *South African National Biodiversity Assessment 2018 Technical Report Volume 6: Sub-Antarctic Territory*. Pretoria: South African National Biodiversity Institute.
- Whitfield, A. (2005). Langebaan – a new type of estuary? *Afr. J. Aquat. Sci.* 30, 207–209. doi: 10.2989/16085910509503859
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Comput. Biol.* 13:e1005510. doi: 10.1371/journal.pcbi.1005510
- Woolley, S. N. C., Foster, S. D., Bax, N. J., Currie, J. C., Dunn, D. C., Hansen, C., et al. (2020). Bioregions in marine environments: combining biological and environmental data for management and scientific understanding. *Bioscience* 70, 48–59. doi: 10.1093/biosci/biz133
- Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., et al. (2021). Addressing data integration challenges to link ecological processes across scales. *Front. Ecol. Environ.* 19, 30–38. doi: 10.1002/fee.2290



OPEN ACCESS

EDITED BY

Sandra MacFadyen,
Stellenbosch University,
South Africa

REVIEWED BY

Maria M. Romeiras,
University of Lisbon,
Portugal
Glenn R. Moncrieff,
South African Environmental Observation
Network (SAEON),
South Africa

*CORRESPONDENCE

Brenda Daly
✉ B.Daly@sanbi.org.za

SPECIALTY SECTION

This article was submitted to
Environmental Informatics and Remote
Sensing,
a section of the journal
Frontiers in Ecology and Evolution

RECEIVED 15 December 2022

ACCEPTED 29 March 2023

PUBLISHED 26 April 2023

CITATION

Daly B and Ranwashe F (2023) South Africa's
initiative toward an integrated biodiversity data
portal.
Front. Ecol. Evol. 11:1124928.
doi: 10.3389/fevo.2023.1124928

COPYRIGHT

© 2023 Daly and Ranwashe. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

South Africa's initiative toward an integrated biodiversity data portal

Brenda Daly* and Fhatani Ranwashe

Biodiversity Information and Planning Directorate, Division of Biodiversity Information and Policy
Advice, Kirstenbosch Research Centre, South African National Biodiversity Institute, Cape Town, South
Africa

Researchers and policymakers have called on the South African National Biodiversity Institute (SANBI), in its role as the statutory biodiversity organisation of South Africa, to develop a coordinated and integrated biodiversity informatics hub. While biodiversity information is increasingly available from several providers, there is no platform through which to access comprehensive biodiversity information from a single source. In response, SANBI is redeveloping the Biodiversity Advisor platform, which will integrate geospatial, species and ecosystem data, literature and other data made available by a wide variety of data partners. To do so it has adopted a Service Orientated Architecture, whereby existing, independent biodiversity datasets are integrated. Consolidating such an extensive and varied set of databases, however, introduces some significant operational challenges. Solutions had to be found to address limited infrastructure, the complexity of the system, the lack of taxonomic identifiers, as well as the need for access and attribution. Solutions had to be pragmatic, given limited financial resources and limited capacity for information technology. The emerging outcome is a system that will easily allow users to access most biodiversity data within South Africa from a single, recognised platform.

KEYWORDS

infrastructure, services, system, attribution, biodiversity data integration, information resources

1. Introduction

Currently, there is a global impetus toward an interconnected network to link other sources of biodiversity and environmental data and in this way provide interdisciplinary information (Hardisty et al., 2022). Hardisty et al. (2022) use specimens as a digital anchor connecting other discipline-specific data. A recent example is modelling, understanding, and preventing potential pandemics, following COVID-19 (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services [IPBES], 2020). Conservation efforts such as modelling invasive species, food and water security, and restoration are among the major forces behind data-driven prioritisation in many countries and organisations. New opportunities have become possible with the availability of big datasets and the advances in artificial intelligence technologies and their use in different fields of study (e.g., image and text recognition, systematic conservation planning) is an upcoming innovation (Silvestro et al., 2022).

With the increase in global and local online platforms that offer biodiversity data, such as iNaturalist, Global Biodiversity Information Facility (GBIF), Plants of Southern Africa and Biodiversity Geographic Information System (BGIS), access to biodiversity data has become easier (MacFadyen et al., 2022). However, it is still difficult to obtain all this biodiversity information from just one source. The South African National Biodiversity Institute (SANBI)

was maintaining disparate information systems that required a broad range of skills to support and management costs were escalating (Daly et al., 2013). It was at this point that SANBI started working to recreate the Biodiversity Advisor (URL)¹, an interoperable biodiversity data portal, that will provide comprehensive biodiversity information to a wide range of users. Users will have access to geospatial data, plant and animal species distribution data, ecosystem-level data, literature, images, and metadata. The newly developed system promotes a shift from tactical information systems, which deliver products and services for individual projects, to a strategic system that builds capacity within organisations and networks.

The overarching goal of the new system is to integrate available biodiversity data by unifying information resources across SANBI and its data partners, to improve quality and use, and thereby transform data into knowledge. Joining these data infrastructures will give researchers a collective overview that better facilitates answering research questions and will provide policymakers with the necessary information to make more informed decisions. This paper describes how biodiversity information sources, systems, and services in South Africa are being integrated into a national information system, as part of a project called the National Biodiversity Information System (NBIS). The range of different data platforms that are being brought together presents significant operational challenges that have required expedient and resource-efficient solutions.

SANBI recognised the success of similar international initiatives and to avoid reinventing existing solutions, a comparison of eight national research infrastructures was completed during the scoping phase of the NBIS project. These included the Atlas of Living Australia (ALA), SiB Colombia, National Biodiversity Data Centre (Biodiversity Ireland), National Biodiversity Network Atlas (NBN Atlas), LifeWatch Marine Virtual Research Environment, Conabio, Zoo Universe and Catchments. Most of the systems investigated were bespoke solutions. The ALA, NBN Atlas and Biodiversity Ireland showed the greatest fit (20%) with SANBI's requirements in developing its biodiversity informatics infrastructure. The criteria that made these systems more similar to what was required were how adaptable these systems would be to the unique existing source repositories at SANBI, the data types made available online and the requirements identified. The ALA architectural model consisted of numerous modular tools and software suites (e.g., Sensitive Data Service, Image Service, BioLink, etc.) linked together via a micro-services architecture (Chapman et al., 2016). Several of these modules were later made available to other organisations to use as open-source software reusable modules. The existing international systems would therefore be used as exemplars for the South African system, with the necessary deviations to account for the unique local context.

Biodiversity information management is not just about creating new methods or tools, it is about the coordination of stakeholders (e.g., data partners, communities of practise, etc.), standards, digitisation processes, integration, processing and using data effectively to support decisions. Biodiversity data and its processed products such as the Red List of Ecosystems and Species, routinely inform spatial planning, environmental authorisation, and protected area expansion through established channels (Botts et al., 2019, 2020).

When compared to other countries, South Africa is ahead of many others in the global context because it covers the whole spectrum.

This perspective article is targeted toward institutions that are starting on the journey of developing biodiversity informatics infrastructure. It highlights aspects of NBIS technical design that are particularly challenging and solutions that have proven successful.

2. The biodiversity information architecture

2.1. Strategy and building blocks

The South African National Biodiversity Institute (SANBI) is a statutory organisation established under the (National Environmental Management: Biodiversity Act, No.10 of 2004, 2004). South Africa is one of the few countries in the world to have a statutory entity with a dedicated biodiversity focus. In fulfilment of its mandate, SANBI leads and coordinates research, monitors and reports on the state of biodiversity in South Africa, gives planning and policy advice, engages in ecosystem restoration, and has a variety of managed collections of preserved and living specimens, seed banks, biological samples (BioBank), literature and library records. SANBI has responded to identified needs over time and developed a range of systems, tools, and policies, however, the value of these resources has been undermined as they are not integrated.

Despite its extensive data and information holdings, SANBI is not the only biodiversity organisation in the country that collects and serves biodiversity data. SANBI recognises that it does not have the capacity to achieve its mandate single-handedly and has adopted a Network of Partners Model where partners, through formal agreements, can contribute toward delivering on the SANBI mandate. Partnerships are not established with individual consultants or organisations working purely for profit and there is a set of criteria that each institution must meet. There are legal and non-legal mechanisms for implementation, for example, data sharing agreements, collaboration agreements, secondments, etc. (South African National Biodiversity Institute [SANBI], 2017). For example, the South African Environmental Observation Network (SAEON) collects long-term environmental observation data in South Africa (such as weather, soil moisture and temperature, etc.), so SAEON is a data partner.

Due to the complexity of the source repositories and the significant investments in developing large biological information resources, a more streamlined technical and operational model was needed to integrate all information resources. The challenge was to combine the existing information environment despite limited financial resources and in-house information technology expertise within SANBI. Consequently, decisions made during the development of the Biodiversity Advisor sought pragmatic but innovative ways to achieve more in a resource-constrained setting.

The NBIS project, therefore, began with the existing set of established information resources that were largely independent. It made no sense to go to the significant effort of migrating data into the available open-source ALA software suites when these functional components already existed in the existing infrastructure. Instead, to accomplish the data synthesis required, a Service Oriented Architecture (SOA) was implemented, which is a style of software design that integrates distributed, separately deployed and maintained

¹ <https://biodiversityadvisor.sanbi.org>

software components that may be controlled by various owners (Reference Architecture Foundation for Service Oriented Architecture Version 1.0, 2012). The basic tenet of SOA is that it is independent of vendors, products, and technologies.

The benefits of following an SOA architecture are the ability to assemble services (functionality and data) that leverage existing investments. Another benefit is that, although software and application upgrades are required to ensure compliance, there has been minimal impact within the source datasets and information resources landscape, which has meant that users continue to use the existing software and applications. Independent data storage has meant that each application (authoring layer) or service is independently changeable and deployable and can use a different technology stack. The web application, however, will need to be modified to accommodate this change.

2.2. User needs analysis

A survey, as part of a thesis project, was completed to understand who the user community is and what their needs are. These findings are currently being built into the Biodiversity Advisor and will help inform the development of products and services through a clear understanding of user needs. Using the initial needs analysis, the following user-level functionality was highlighted (Daly, 2020):

- The ability to aggregate information from other relevant fields of study (social, political, and economic) for more informed decision-making.
- Presentation of useful (solve a problem or decision) case studies.
- Tailored information views (consider the viewpoint of the information seeker).
- Include intuitive navigation as users are often unfamiliar with the content of the website.
- An advisory section on emerging science and policy topics.
- Focus on the information most in demand (distribution, ecological and threatened species data).
- Provide sources of environmental change information.
- Crowd-source data deficient species.
- Increase accessibility to peer-reviewed research outputs.

2.3. The service orientated architecture model

The basic structural elements of an SOA model are: (1) the underlying source datasets accompanied by their independent authoring layers, (2) an index and services layer that catalogues the information in the source datasets and acts as a bridge between them, (3) the front-end website and app that users will interact with, and (4) a search engine option that offers the ability to navigate the information (Figure 1). The authoring layer is the ready-made, often commercial application that supports business activities.

The source datasets that will be integrated within the Biodiversity Advisor system hosted by SANBI include:

- BODATSA – Botanical Database of Southern Africa, official plant names and descriptions (taxonomic backbone), specimen, living and seed collection, medicinal plant data, national vegetation database, and invasive species data.
- ZODATSA – Zoological Database of Southern Africa, official animal names and descriptions.
- Institutional repository – document repository to store all SANBI historical collections, library services and publications.
- Invasive Species Management System – tracks invasive species locations, abundance, and control efforts.
- Ecosystem database – ecosystem type, description, threat status, protection level assessment, distribution, and extent.
- BGIS – Biodiversity Geographic Information System (BGIS), a stakeholder website hosting products of various biodiversity plans (conservation plans) and other related initiatives.
- Metadata portal.
- IPT – Integrated Publishing Toolkit, publish biodiversity datasets from data partners.
- SEIS – SANBI Enterprise Image System, specimen and other digital images.

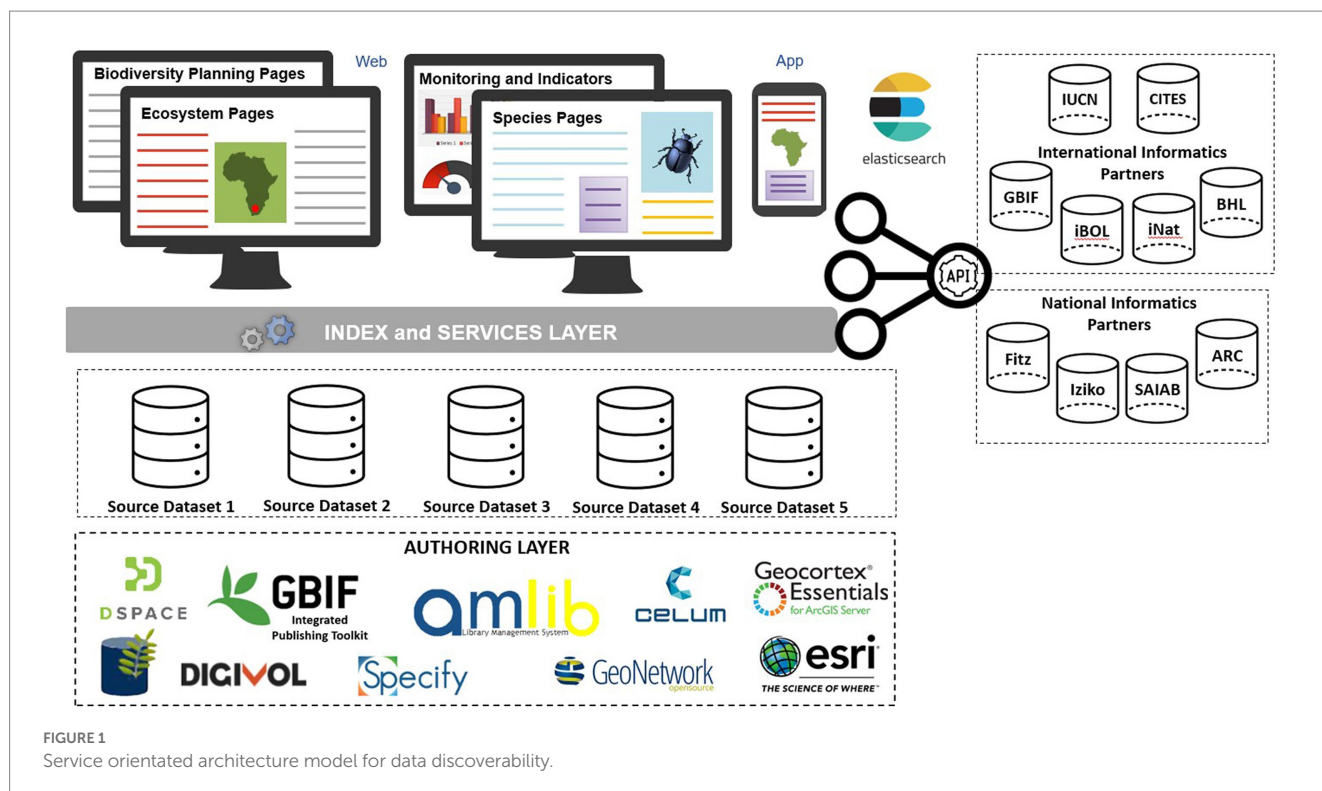
National Informatics Partners datasets:

- Fitz – FitzPatrick Institute of African Ornithology, hosts several biological resources.
- Iziko – Iziko South African Museum, museum specimens.
- SAIAB – South African Institute for Aquatic Biodiversity, fish specimens.
- ARC – Agricultural Research Council, conduct research in the agricultural sector.

International Informatics Partners datasets:

- GBIF – Global Biodiversity Information Facility, a global aggregator of species occurrence records.
- iBOL/Genbank – International Barcode of Life and Genbank, annotated collection of available DNA sequences.
- iNat – iNaturalist, Citizen Scientists can capture and upload sightings.
- BHL – Biodiversity Heritage Library, biodiversity literature.
- IUCN – International Union for Conservation of Nature, develops and promotes international standards for evaluating the conservation status of plant and animal species.
- CITES – Convention on International Trade in Endangered Species of Wild Fauna and Flora, a global agreement that ensures international trade does not threaten species' survival in the wild.

The transition to the SOA centred on the creation of an indexing system, which is a highly ordered set of lists of frequently searched data, coupled with the “ElasticSearch” search engine. The index and search engine are what allow calls to be made to the respective systems for data. In instances where data partners do not have application programming interfaces (APIs), data will be moved into the “index and services layer” with an extract-transform-load (ETL) process. An ETL is where data is extracted, transformed, and loaded into an output data container.



3. Operational challenges and solutions

3.1. Lack of infrastructure and tools when integrating data

The greatest challenge at the start of the project was the limited infrastructure concerning the resources assigned at the server level so the server specifications, poor network bandwidth, and separating servers to reduce response times by spreading the computational load. This task's resource considerations were underestimated, meaning it took up to 4 weeks to index the various source datasets. These technological challenges were overcome by procuring additional infrastructure and scaling to meet demands. Going forward, once the source datasets are indexed, incremental indexing can be used when changes are made to the source datasets. This will bypass the need for a complete resource-intensive reindex. Resulting updates will be run separately and published to the live portal once complete.

An API is a data interchange tool that allows applications to communicate and is most often developed and deployed by the vendor. An API means access to the data without having to understand all the system detail such as the database schema, functionality, etc. In some cases, web-based APIs were developed for the data sources, however, many software applications do not have a stable API, which meant the data was indexed directly from the backend database. The disadvantage of this solution is if any major changes are made to the software database it means changes need to be made to the platform. Therefore, it is essential to consider how future versions or changes in application products will fit with the current architecture and account for the time and resources to maintain the system. APIs can often also

be a constraint in a project as they only unlock certain data depending on the intended use case.

Many national informatics partners do not have the necessary resources to manage and provide data. System maintenance is also resource intensive and the sustainability of these projects is a risk. The lack of suitable mechanisms and infrastructure is often a barrier to data partners publishing their data. To overcome this challenge SANBI has offered support with data management and set up an Integrated Publishing Toolkit (IPT) instances used to publish biodiversity datasets (Robertson et al., 2014). This is working toward recommendations made by Costello et al. (2014) on strategies for the sustainability of datasets being the integration of datasets into a collaborative information system such as the IPT, within an institute with a suitable mandate.

3.2. Complexity of the system

With different data platforms being brought together various data categories are integrated under one architecture and the challenge is integrating these heterogeneous data. Source repositories have data categories ranging from geographic or spatial data (BGIS), key biodiversity areas data, occurrence records (BODATSA, SANBI IPT), ecosystem data (Ecosystem Descriptions Database), checklists (accepted and synonym names; protologue citations; type information; classification), distribution and residency status (BODATSA, ZODATSA), specimen data (BODATSA), descriptive information (BODATSA, ZODATSA), literature (SANBI Institutional Repository, SANBI library catalogue), metadata (SANBI Metadata), genetics (BioBank), images (SANBI Enterprise Image System (SEIS)), and threatened species and ecosystem data (Red List Assessment Systems), National Biodiversity Assessment (NBA) data to taxonomic descriptions (BODATSA and ZODATSA) and indicator data

(Figure 2). Interlinking information systems in interoperable ways in a consistent manner is challenging and much time needs to be allocated to tailoring data structures and query processing.

Another challenge faced by organisations is related to people or skills, the first implication of this skills shortage is that organisations often need service providers to fill this resource gap. It is vital with outsourcing to select the right service provider that understands the core business and time to be able to develop specifications for the project correctly.

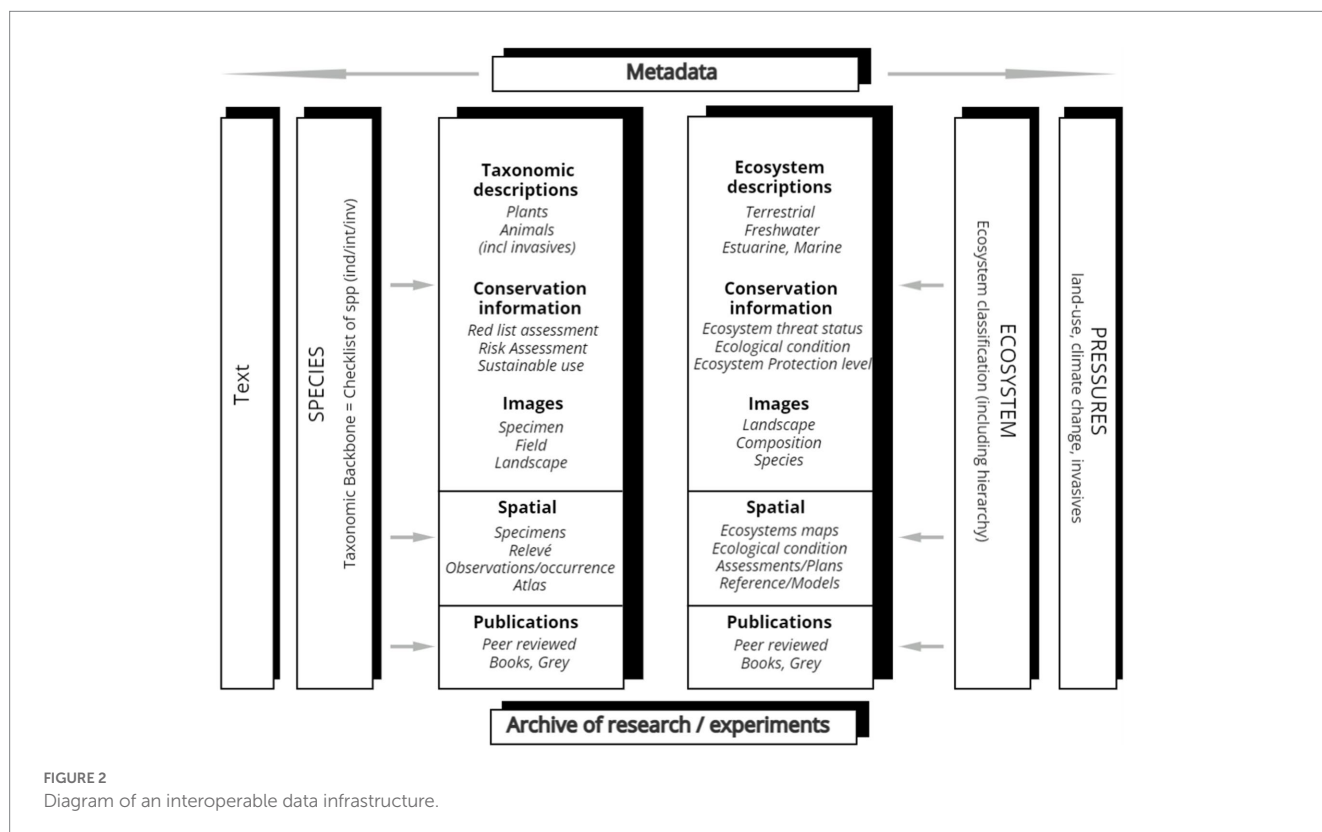
3.3. Taxonomic service backbone

SANBI is mandated to maintain and provide an up-to-date South African National Plant and Animal Checklist with accurate taxonomic information. This is achieved by publishing a consolidated national checklist of plants and animal species in South Africa yearly, with updates happening throughout the year as taxonomic changes are made available in the literature. Updating the checklist involves monitoring published literature. South Africa has over 67,000 described species of animals (Skowno et al., 2019), 21,467 species of plants (Klopper and Winter, 2022), and 1,422 alien plant and animal species (van Wilgen et al., 2020). As with any biodiversity data integration process, species names are often used as the common identifier, however, the limitations are that they are not unique or stable (Page, 2008). An enabling feature is assigning persistent identifiers (long-lasting references consisting of letters or numbers to a digital resource often machine generated) that will support linkages between data sources and allow for global compatibility. Historically, due to a lack of persistent identifiers, users have used primary keys

from the authoring layer which has resulted in obsolete numbers being migrated back into the system, as systems and software are updated. The lack of a common identifier has created a barrier to making data publicly available, accentuating shortcomings in available data, generating, or integrating any other types of data and ultimately the conservation and management of the species (Ely et al., 2017). Therefore, linking Globally Unique IDentifiers (GUIDs) in an authoritative taxonomic resource when integrating biodiversity data and using these GUIDs to link other source datasets is imperative (Guralnick et al., 2015).

3.4. Anonymous usage

SANBI's mandate is clear that as a public organisation, the biodiversity information it provides must be openly accessible (South African National Biodiversity Institute [SANBI], 2010a). The management of data is covered by the Biodiversity Information Policy Framework (South African National Biodiversity Institute [SANBI], 2010a), the Intellectual Property (IP) policy (South African National Biodiversity Institute [SANBI], 2010b) and the Protection of Sensitive Taxa policy (South African National Biodiversity Institute [SANBI], 2010c). Therefore, it is essential that the Biodiversity Advisor is available for anonymous usage and that data is free to download. However, to better understand users, manage the system and determine if the information being provided is having an impact, it is also necessary to monitor use. Several types of activity, such as the download of spatial data, must take the user through a confirmation process that makes it clear who owns the data and the terms of use. In addition, sensitive data, such as the location of species that are



vulnerable to collecting or over-exploitation, requires a process of access request and approval. Some data owned by partners and made available as part of data sharing agreements includes embargoes, redactions, or restrictions. Data sharing agreements within the framework of attributions ensure the data is suitably shared.

In response, a secure authentication mechanism with no restriction to register is being implemented. The authentication mechanism helps to manage the data attribution and can be used to manage access to some projects and functionality (such as authorising access to documents or download links). It also helps to analyse usage patterns and capture business intelligence data when required. A role-based security module was developed, known as the Biodiversity Passport, where functionality and datasets are available only to users who have been authenticated and authorised, such as allowing authorised officials of government conservation agencies to access content for conservation management purposes.

Access to literature is often as essential as raw data, however, copyrights and paywalls often stifle necessary access to information in the conservation of species and ecosystems. There is a push for open access when linking literature associated with biodiversity data. The solution here is to provide Uniform Resource Identifiers (URI) to the physical resource or a Uniform Resource Locator (URL) to the payment gateway to ensure the researcher's work is recognised. SANBI's Institutional Repository allows users to request a copy from the SANBI author as an alternative to buying the paper.

3.5. Ensuring attribution

Data citation and attribution seem to be a consistent struggle in building biodiversity informatics infrastructure (Reichman et al., 2011; Patterson et al., 2014). Attribution is defined as assigning appropriate credit for an organisation or individual's contribution (Haak, 2014; Franz and Sterner, 2018) perspective is that it holds authors accountable for data accuracy and potential criticism. Ensuring a consistent user experience across multiple channels and still preserving attribution is an underrated challenge. The issue then confounds when to increase data accessibility, users can reuse subsets of data by downloading .csv files linked to the data behind any user interface. In this case, any data record downloaded needs to include the contributor's name.

Metadata is considered a form of attribution. However, the challenge here again is ensuring the metadata is accurate and up to date as often the metadata-related changes are not documented. Metadata describes the origin and tracks dataset changes (Biodiversity Conservation Information System, 2000). It is essential to ensure that every record includes the author's name on the website and within any downloaded data. Another solution is an acknowledgements page or listing contributors or editors of data. A solution used by Figshare is to cite datasets using formatted references (Haak, 2014).

4. Conclusion

In the past, the numerous disaggregated and disparate systems, tools, and policies made it difficult to leverage biodiversity information to support research, policy development and decision-making. The Biodiversity Advisor is a service-orientated data management system, built largely from contributing systems. These data platforms are being

brought together to ensure the information resource is more adequate for national planning and management. By addressing limited informatics infrastructure, obtaining necessary human resources and skills, and establishing a system of unique identifiers, the complexity of the system can be overcome. Developing authentication systems and mechanisms for assigning credit can navigate the balance required between accessibility and attribution. The reimagined Biodiversity Advisor is thus a milestone in establishing a fully integrated data information system for South Africa.

The Biodiversity Advisor² is scheduled to launch in 2023. At the time of launch, the BODATSA (plant) and ZODATSA (animal) will be fully indexed, providing comprehensive species pages, including specimen collection records, iNaturalist observational records and occurrence records from various data partners. The SANBI library catalogue and institutional repository will also be indexed providing numerous literature resources. Systematic integration of the remainder of the systems and data will follow as datasets become available.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BD took the lead in writing the manuscript. FR compiled the presentation abstract and gave the presentation. All authors contributed to the article and approved the submitted version.

Acknowledgments

Thank you to Emily Botts for her helpful comments on this manuscript. Thank you to SANBI and Data Partners for their participation in the NBIS project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

² <https://biodiversityadvisor.sanbi.org>

References

- Biodiversity Conservation Information System (2000). Framework for information sharing: Principles. Handbook Series. 1–8.
- Botts, E. A., Pence, G., Holness, S., Sink, K., Skowno, A., Driver, A., et al. (2019). Practical actions for applied systematic conservation planning. *Conserv. Biol.* 33, 1235–1246. doi: 10.1111/cobi.13321
- Botts, E. A., Skowno, A., Driver, A., Holness, S., Maze, K., Smith, T., et al. (2020). More than just a (red) list: Over a decade of using South Africa's threatened ecosystems in policy and practice. *Biol. Conserv.* 246:108559. doi: 10.1016/j.biocon.2020.108559
- Chapman, A. D., Berzin, L., Smith, R., and Tann, J. (2016). Atlas of living Australia infrastructure implementation. Atlas of living Australia (Issue October). Available at: <https://www.ala.org.au/wp-content/uploads/2017/01/ALA-Infrastructure-Implementation-overview-October-2016-final.pdf>
- Costello, M. J., Appeltans, W., Bailly, N., Berendsohn, W. G., de Jong, Y., Edwards, M., et al. (2014). Strategies for the sustainability of online open-access biodiversity databases. *Biol. Conserv.* 173, 155–165. doi: 10.1016/j.biocon.2013.07.042
- Daly, B. (2020). Building biodiversity data infrastructure for science and decision-making: Information needs and information-seeking patterns in South Africa. Faculty of Humanities, Department of Knowledge and Information Stewardship. Available at: <http://hdl.handle.net/11427/32635>
- Daly, B., Hathorn, P., Roberts, R., and Willoughby, S. (Eds.) (2013). Proceedings of an Information Architecture Workshop. South African National Biodiversity Institute.
- Ely, C. V., De Loreto Bordinon, S. A., Trevisan, R., and Boldrini, I. I. (2017). Implications of poor taxonomy in conservation. *J. Nat. Conserv.* 36, 10–13. doi: 10.1016/j.jnc.2017.01.003
- Franz, N. M., and Sterner, B. W. (2018). To increase trust, change the social design behind aggregated biodiversity data. *Database* 2018:bax100. doi: 10.1093/database/bax100
- Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., et al. (2015). Community next steps for making globally unique identifiers work for biocollections data. *ZooKeys* 494, 133–154. doi: 10.3897/zookeys.494.9352
- Haak, L. L. (2014). Persistent identifiers can improve provenance and attribution and encourage sharing of research results. *Inf. Serv. Use* 34, 93–96. doi: 10.3233/ISU-140736
- Hardisty, A. R., Ellwood, E. R., Nelson, G., Zimkus, B., Buschbom, J., Addink, W., et al. (2022). Digital extended specimens: Enabling an extensible network of biodiversity data records as integrated digital objects on the internet. *Bioscience* 72, 978–987. doi: 10.1093/biosci/biac060
- Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services [IPBES] (2020). “Workshop Report on Biodiversity and Pandemics of the Intergovernmental Platform on Biodiversity and Ecosystem Services” in . eds. P. Daszak, J. Amuasi, C. G. Das Neves, D. Hayman, T. Kuiken and B. Roehet al. (Bonn, Germany: IPBES Secretariat)
- Klopper, R.R., and Winter, P.J.D. (2022). *South African National Plant Checklist statistics, version 1.2022*. South African National Biodiversity Institute, Pretoria.
- MacFadyen, S., Allsopp, N., Altwegg, R., Archibald, S., Botha, J., Bradshaw, K., et al. (2022). Drowning in data, thirsty for information and starved for understanding: A biodiversity information hub for cooperative environmental monitoring in South Africa. *Biol. Conserv.* 274:109736. doi: 10.1016/j.biocon.2022.109736
- National Environmental Management: Biodiversity Act, No.10 of 2004 (2004). *Government Gazette*. 467(700). 7 June. Government notice no. 26436. Cape Town: Government Printer.
- Page, R. D. M. (2008). Biodiversity informatics: The challenge of linking data and the role of shared identifiers. *Brief. Bioinform.* 9, 345–354. doi: 10.1093/bib/bbn022
- Patterson, D. J., Egloff, W., Agosti, D., Eades, D., Franz, N., Hagedorn, G., et al. (2014). Scientific names of organisms: Attribution, rights, and licensing. *BMC. Res. Notes* 7:79. doi: 10.1186/1756-0500-7-79
- Reference Architecture Foundation for Service Oriented Architecture Version 1.0 (2012). OASIS committee specification 01. Available at: <http://docs.oasis-open.org/soa-rm/soa-ra/v1.0/cs01/soa-ra-v1.0-cs01.html>
- Reichman, O. J., Jones, M. B., and Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. *Science* 331, 703–705. doi: 10.1126/science.1197962
- Robertson, T., Döring, M., Guralnick, R. P., Bloom, D., Wiczorek, J. R., Braak, K., et al. (2014). The GBIF integrated publishing toolkit: Facilitating the efficient publishing of biodiversity data on the internet. *PLoS One* 9:e102623. doi: 10.1371/journal.pone.0102623
- Silvestro, D., Gorla, S., Sterner, T., and Antonelli, A. (2022). Improving biodiversity protection through artificial intelligence. *Nat. Sustain.* 5, 415–424. doi: 10.1038/s41893-022-00851-6
- Skowno, A.L., Poole, C.J., Raimondo, D.C., Sink, K.J., Van Deventer, H., Van Niekerk, L., et al. (2019). National Biodiversity Assessment 2018: The status of South Africa's ecosystems and biodiversity. Synthesis Report. South African National Biodiversity Institute, an entity of the Department of Environment, Forestry and Fisheries, Pretoria. pp. 1–214.
- South African National Biodiversity Institute [SANBI] (2010a). Biodiversity information policy framework principles and guidelines. Available at: <http://hdl.handle.net/20.500.12143/7450>
- South African National Biodiversity Institute [SANBI] (2010b). Biodiversity information policy framework–intellectual property rights policy. Available at: <http://hdl.handle.net/20.500.12143/7449>
- South African National Biodiversity Institute [SANBI] (2010c). Biodiversity information policy framework-digital access to sensitive taxon data. Available at: <http://hdl.handle.net/20.500.12143/7087>
- South African National Biodiversity Institute [SANBI] (2017). *Network of partners policy*. Pretoria: South African National Biodiversity Institute.
- van Wilgen, B. W., Measey, J., Richardson, D. M., Wilson, J. R., and Zengeya, T. A. (2020). “Biological invasions in South Africa: An overview” in *Biological invasions in South Africa. Invading nature-springer series in invasion ecology*. eds. B. Wilgen, J. Measey, D. Richardson, J. Wilson and T. Zengeya (Cham: Springer).



OPEN ACCESS

EDITED BY

Sandra MacFadyen,
Stellenbosch University, South Africa

REVIEWED BY

Judith Botha,
South African National Parks, South Africa
Hayley Clements,
Stellenbosch University, South Africa

*CORRESPONDENCE

Fatima Parker-Allie

✉ F.Parker@sanbi.org.za

RECEIVED 24 November 2022

ACCEPTED 14 June 2023

PUBLISHED 05 July 2023

CITATION

Parker-Allie F, Gibbons MJ and
Harebottle DM (2023) A conceptual
approach to developing biodiversity
informatics as a field of science
in South Africa.

Front. Ecol. Evol. 11:1107212.

doi: 10.3389/fevo.2023.1107212

COPYRIGHT

© 2023 Parker-Allie, Gibbons and
Harebottle. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

A conceptual approach to developing biodiversity informatics as a field of science in South Africa

Fatima Parker-Allie^{1,2*}, Mark J. Gibbons²
and Douglas M. Harebottle³

¹South African National Biodiversity Institute, Foundational Biodiversity Science, Cape Town, South Africa, ²Biodiversity and Conservation Biology Department, Faculty of Natural Science, University of Western Cape, Cape Town, South Africa, ³Risk and Vulnerability Science Centre, School of Natural and Applied Sciences, Sol Plaatje University, Kimberley, South Africa

In South Africa, as in other parts of the world, *Biodiversity Informatics* (BDI) has been identified as a young field of science that lies at the nexus of several disciplines, including informatics, biology and mathematics/statistics. Being such a new and dynamic field, there are challenges in the recruitment, training and retention of personnel that can support *inter alia* the mobilisation, management, coordination, and utilisation of biodiversity information for key conservation and biodiversity outcomes. The lack of human capital also place at risk the implementation of (e.g.) the *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (IPBES), and hinders attainment of the Convention on Biological Diversity post-2020 framework targets. There is a clear demand for broad efforts to build human capital in the field. Using our experiences in South Africa, we provide a framework for establishing BDI as a field of science in developing countries and look at the potential building blocks towards this broad objective, including the need and requirements for the establishment of a Centre for BDI. We explore this concept against a backdrop of the South African government's 2019 White Paper on *Science, Technology and Innovation*, and the associated Decadal Plan, both developed under the auspices of the *Department of Science and Innovation*. We also reflect on efforts in the broader landscape to look at the establishment of BDI curricula.

KEYWORDS

biodiversity informatics, framework, science–policy interface, science technology & innovation (STI) policy, universities & higher education institutions, capacity development, training & development, elearning

Introduction

The South African science landscape has evolved dramatically since the dawn of democracy through government's commitment to transform the inward-looking and embattled sector into a system that is innovative, flexible and responsive to the needs of our society (DST, 2007). South Africa's prospects for improved competitiveness and

economic growth rely, to a great degree, on science and technology. Its ten-year innovation plan (DST, 2007) is built on the foundation of the *National System of Innovation*, and recognizes that the country needs to take further steps to becoming a knowledge-based economy in order to meet its developmental objectives (Manzini, 2012; DSI, 2021): the *National System of Innovation* is an organising framework for policies and institutions supporting the knowledge economy. In this review, South Africa's role in the Global Biodiversity Information Facility (GBIF) and the potential role that *Biodiversity Informatics* (BDI) can play in meeting a stronger *Science, Technology and Innovation* agenda (DST, 2019; DSI, 2021) is explored (see Box 1).

The field of BDI deals with the interrelated challenges of collection, collation, integration, analysis, prediction, and dissemination of data and information related to the diversity of life on Earth (Hobern et al., 2012; Hardisty et al., 2013; Walters & Scholes, 2016). It is a field that has been massively enhanced following developments in information technology (Berners-Lee, 1999), which have led to exciting new opportunities for data consolidation and interchange (Chapman, 2005). Today, BDI is regarded as the application of informatics techniques to biodiversity data, with much of biodiversity informatics resting on physical objects that ground the digital information i.e., specimens of organisms (Parr and Thessen, 2018).

In the last two decades there has been an unprecedented increase in the acquisition of actual biodiversity data and data types, driven by mobile-cellular applications, sensors, mass digitization and next generation sequencing. More recently, public participation in research ("community science") has led to an increase in the mobilisation of biodiversity data. One such initiative is the United States National Science Foundation's iDigBio (Integrated Digitized Biocollections) which has mobilized more than 120 million specimens held in national institutions, and similar efforts are continuing in parallel across the world (Nelson and Ellis, 2018). New observation-based records collected by community science platforms have proliferated (eBird has >1 billion records whilst iNaturalist has >58 million (Auer et al., 2022; iNaturalist Contributors, 2023), outpacing museum specimen digitization by orders of magnitude (Chandler et al., 2017; Troudet et al., 2018).

The infrastructure (knowledge management systems) linked to BDI-based research, such as GBIF or the *Ocean Biodiversity Information System*, enable users to navigate and put to use vast quantities of biodiversity information. This can be used to advance scientific research in areas such as (e.g.) agriculture and conservation (GBIF Secretariat, 2022b), while species distribution modelling (Anderson et al., 2016) allows for the management of alien species (Faulkner et al., 2014) and understanding the possible impact of climate change on biodiversity (Burrows et al., 2019) and human health (Peterson, 2009).

The accessibility of data serves the economic and quality-of-life interests of society, and provides a basis from which our knowledge of the natural world can grow rapidly in a manner that avoids duplication of effort and expenditure (OECD, 1999; Parker-Allie et al., 2021). While opportunities abound for this new and dynamic field to impact many areas of science linked to human well-being, there is a critical need for increased capacity enhancement precisely because it is an emerging field – especially so in South Africa and other developing countries (Schalk, 1998; Sarkar, 2009; Parker-Allie et al., 2021).

The global context for biodiversity informatics

Global changes, including those with a socio-economic, geopolitical, scientific, technological, or environmental basis, have profound implications for the *National System of Innovation* in South Africa. Inter- and transdisciplinary knowledge is increasingly important, as research becomes progressively more data-driven (OECD, 2013a; Visalli et al., 2020), with greater access to existing information being facilitated by an open science approach (UNESCO, 2021). The success of South Africa's response to the Fourth Industrial Revolution will depend on how well we exploit the pivotal role of information and communication technology (ICT) and harness the potential of big data (DSI, 2022). With the increase in data volumes, velocity and types, data have become a core asset that can create a significant competitive advantage and drive innovation, sustainable growth and development (OECD, 2013a).

Box 1. Growing the knowledge economy through enhanced STI efforts

Data can be described as key elementary units of new knowledge, with data-driven initiatives like the South African National Biodiversity Institute – Global Biodiversity Information Facility (SANBI-GBIF) strengthening South Africa's role for enhanced activities in *Science, Technology and Innovation*. Knowledge provides the basic capital for innovation, through generation, accumulation and exploitation (OECD, 2013b). Economic growth is driven by innovation, and the key driver for innovation is "high-end" human capital. In South Africa, there is a need to significantly strengthen both the production of human capital and the institutional environment for knowledge generation and this can best be done with the collaborative assistance of international partners (DST, 2007). This especially as a growing percentage of the wealth in the world's largest economies is created by knowledge-based industries that rely heavily on human capital and technological innovation (Hadad, 2017; Department of Science and Technology, 2019).

In the "Scientific Impact of Nations", King (2004) made a correlation between the economic wealth of 31 nations and their "citation intensity", where citation intensity was used as a proxy for investment in science and technology. In King's (2004) study, South Africa was clustered with Brazil, South Korea, Russia, China and Poland at the lower end of the spectrum. This demonstrates that if South Africa wishes to increase its economic growth, it needs to prioritize an increased investment in research capacity and development output (DEA, 2016; DST, 2019; DSI, 2021). One of the methodologies that can be used to measure readiness for the knowledge economy is the Global Knowledge Index. This index builds on the Knowledge Economy Index (World Bank, 2012), and measures the knowledge performance of countries based on a country's general "enabling environment", and six other components. The latter include the levels of (1) pre-university education, (2) technical and vocational education and training, (3) higher education, (4) research, development and innovation, (5) ICT, and (6) the economy.

Initiatives such as GBIF fully support these philosophies as they enable vast amounts of data to be published in an open access manner through a knowledge management platform. South Africa's membership to GBIF encourages efforts to grow capacities in BDI. This intergovernmental mega-science, data-driven initiative provides a solid foundation for South Africa to implement its capacity development efforts in BDI, building as it does on the fact that training and capacity development are integral to GBIF's Implementation Programme (GBIF, 2022a).

At the global level the Global Biodiversity Informatics Outlook (Hobern et al., 2012) provides a framework for BDI, as it aims to harness the immense power of information technology and an open data culture to gather unprecedented evidence about biodiversity and so inform better decision-making. It proposes action in four key areas i.e. data, culture, evidence and understanding (Hobern et al., 2012). This framework, can be applied to the national biodiversity landscape, to help focus BDI effort and investment.

National mandates and initiatives to support biodiversity data mobilization and growing biodiversity informatics science in South Africa

The South African National Biodiversity Institute (SANBI), as a knowledge management institute, has a mandate to “collect, generate, process, coordinate and disseminate information about biodiversity and sustainable use of indigenous biological resources and maintain databases” in line with the National Environmental Management: Biodiversity Act, No. 10 of 2004. It supports data sharing and harmonising the sharing of biodiversity data through efforts such as knowledge-brokering (Godfrey et al., 2010), which it effects between its network of partners.

In support of this mandate, a Memorandum of Understanding (MoU) was signed between national government and GBIF in 2004. The establishment of the South African GBIF Node, initially called the South African Biodiversity Information Facility, represented a commitment by national government to the sharing and publishing of biodiversity data and to support open science and open access philosophies (The African Open Science Platform, 2018; UNESCO, 2021).

Growing human capital in biodiversity informatics

BDI is a new and rapidly evolving field of science and as such there are enormous challenges in the recruitment and retention of experienced personnel in biodiversity information management: newly recruited staff often arrive in the workplace without the adequate/appropriate combination of skills required (GreenMatterZa, 2009). The *Biodiversity Human Capital Development Strategy*, for the biodiversity sector (produced through stakeholder engagement by SANBI and the Lewis Foundation, which is one of the largest private funders of

conservation activities in South Africa) has identified several skills, as being of “absolute scarcity”. These include, not exclusively, database developers and managers, modellers, curators of biodiversity collections, Geographic Information Systems specialists and technicians and statistical ecologists. ICT specialists and technicians with biodiversity skills, such as systems analysts, web and multimedia developers, applications programmers and database designers and administrators (SANBI and Lewis Foundation, 2010; Rosenberg, 2012), are also thin-on-the-ground. These skills are core to BDI science, and their scarcity has been recognised by the South African Department of Science and Innovation, which has committed to supporting the development of this area of work.

In this paper, we review the efforts that SANBI and its strategic partners have taken to develop human capital in BDI and we provide an account of the processes, philosophies, and approaches that we have taken as a country. South Africa has leveraged its role in big data initiatives like GBIF and drawn from its participation in science-policy platforms, like the *Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services* (IPBES) to advance this field of work. The opportunity for South African institutions to look at synergies between BDI tools and techniques, aligned to IPBES efforts around knowledge and data catalysis is also possible, thereby strengthening the ability for BDI to support science-policy endeavours. We believe that the model followed here holds much promise as a template for other countries to also develop a BDI agenda. In the South African context, some important policy drivers are in place, including national mandates, while several interventions were also needed, to drive a stronger human capital agenda for BDI and to grow the field of science. These are elaborated below.

Strategic high-level interventions to support efforts in BDI

In building capacity in BDI, several strategic high-level interventions have been implemented, with some clear outputs (Table 1). In 2007, the first Biodiversity Information Management Workshop was held, thereby initiating the development of a community of practice in biodiversity information management (Figure 1, Table 1). The workshop allowed participants to understand the needs of the community in relation to biodiversity data and how it is being managed across organisations. It also provided an opportunity to discuss community approaches around data sharing and publishing, and how ultimately these data could be analysed and used. Consultations with the community continued at the newly established annual Biodiversity Information Management Forum (BIMF) and later at the Joint Biodiversity Information Management and Foundational Biodiversity Information Programme (BIM-FBIP) Forum. These activities served to provide a national platform to discuss highlights, opportunities and challenges with regards to biodiversity data, data standards, biodiversity information systems, data publishing and data use.

In 2010, a Data Handover Event was held by the national Node to celebrate the publication of millions of primary biodiversity data

TABLE 1 Timeline of SANBI led activities and approximate related costs supporting the Capacity Development in Biodiversity Information Management/BDI (2007–current).

Year	Activity	Outcome	Approximate Costing (ZAR)
2007	The start of national engagements through the Biodiversity Information Management Forum (BIMF), with the aim to harmonise biodiversity information sharing	Capacity Enhancement in BDI and the development of a community of practice in biodiversity information management.	R100,000
2008–2011	Through annual BIMF meetings the idea of building BIM capacity was further supported and strengthened	In 2010 as an outcome of the BIMF discussions SANBI was elected to drive the development of a Centre for BDI, in an endeavour to build capacity in the field	R400,000
2009	A skills profile for Biodiversity Information Management was developed by SANBI		
2010	DST recommended for the development of a Center for Biodiversity Information Management, at the South African Biodiversity Information Facility instead of the acronym (SABIF) data handover event.		R200,000
2010	SANBI's Human Capital Development Strategy Report for the Biodiversity Sector was developed	Biodiversity Information Management skills critical to scarce skills in South Africa	
2011	Training coordinator for BIM Directorate funded by the South African Biodiversity Information Facility (3 years)	<ul style="list-style-type: none"> • A Learning Network Strategy developed for BIM • Training events were coordinated 	R1,100,000
2012	A Priority Skills Report [GreenMatter, 2012] was developed	Biodiversity Information Management components was listed as part of the absolute scarce skills areas, within the top 21 priority skills for South Africa's Biodiversity sector	
2012	SANBI signed an MoU with University of Western Cape (UWC) on the 31 st of March 2012, which outlined intended areas of cooperation in teaching, research, technical and policy-related biodiversity issues. Also, to look towards the establishment of a postgraduate research hub.	<ul style="list-style-type: none"> • High level support from the Deputy Minister of Science and Technology, the Vice Chancellor of UWC and Chief Executive Officer of SANBI • Two academics were in place to support the initiative. One as champion and other costed through a Memorandum of Agreement (MoA) 	R1,000,000 SANBI R1,000,000 UWC (co-funding)
2012	SANBI secured funding for two post-doctoral students to work on BDI curricula and development of a research strategy and developing Biodiversity Information content and tools.		R1,400,000
2014	Appointment of first post-doctoral fellow	Focus Area: South African BDI Research Strategy	R100,000 (Running Costs)
2015	Appointment of second post-doctoral fellow	Focus Area: BDI content and tools	R100,000 (Running Costs)
2015	SANBI-GBIF led two sessions at the GBIF Nodes meeting entitled: "Towards a Curriculum for BDI"	<ul style="list-style-type: none"> • SANBI identified to drive the process for a GBIF endorsed Global Curricula for BDI • A Taxonomic Data Working Group (TDWG) – GBIF Interest Group was established 	R60,000
2016	Capacity development Session at Joint BIM-FBIP Forum	• Community engagement and additional interest and support for the roll-out of a Centre identified	R50,000
2020	Appointment of SANBI-GBIF MSc Student (Registered at University of KwaZulu Natal)	• Project title: The phylogeographic diversity and connectivity of intertidal sponges as well as determination of whether substrate type influences species settlement along the East of South Africa	R288,000
2020	Appointment of SANBI-GBIF MSc Student (Registered at Rhodes University)	• Project title: Freshwater Amphipoda in South Africa: diversity, distribution and taxonomic review	R288,000
2020	Appointment of third Post-doc	Research focus: BDI Research and Curriculum Development (Biodiversity Data Science)	R300,000
2021	Services rendered by training experts through both direct funding and in-kind contributions of individuals in terms of people time	• Capacity of stakeholders developed in BDI content areas	R250,000
2022	Funding available for 2 Postdoc appointments		R600,000
			R7,236,000

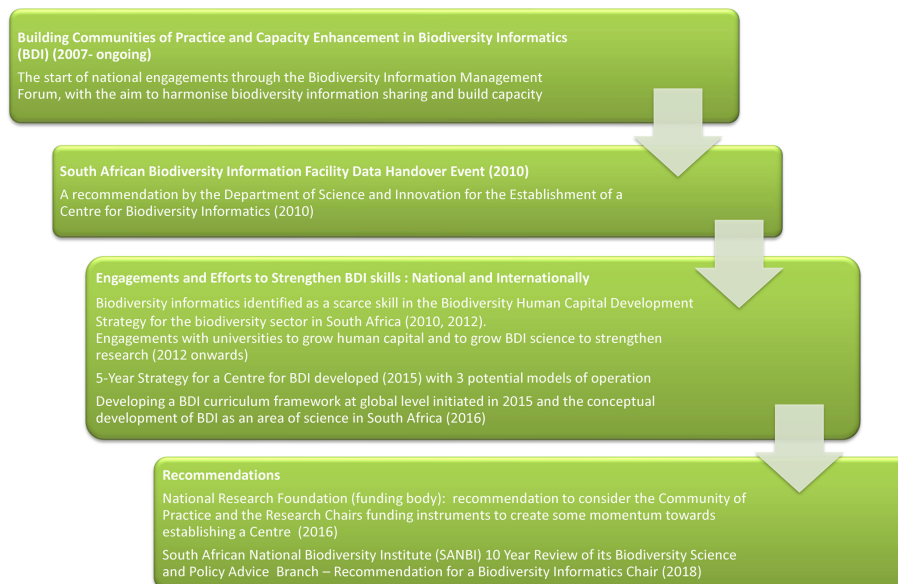


FIGURE 1
Processes, activities and outcomes put in place to support the advancement of BDI.

records to GBIF. Although this event recognized the rewards of investment in BDI by the Department of Science and Innovation (2005–2012), it also resulted in a recommendation being made by national government for the establishment of a Centre for BDI. To achieve this and to grow the field of BDI science, a holistic approach to capacity development was proposed as part of the SANBI-GBIF programme of work. This approach focused on (1) ensuring that relevant, high quality biodiversity information would be available for use by decision makers and managers, and (2) a coordinated network of partners with the commitment and capacity to digitise, share and use biodiversity information would develop. Following engagements with programmes such as GreenMatter^{za}, which is a national initiative focused on the promotion of human capital and skills development in the biodiversity sector (GreenMatterZa, 2009; Rosenberg, 2012; The Lewis Foundation and SANBI, 2021), it became clear that strategic interventions with other parties would be critical for success. The other parties have included academia, the biodiversity community and national government, as well as the National Research Foundation, which is the primary funding body for research in the country. Moreover, this holistic approach included a series of objectives and activities that are detailed in the discussion below (Figure 2), and a Five (5) Year Strategy for the establishment of the Centre was developed in 2015.

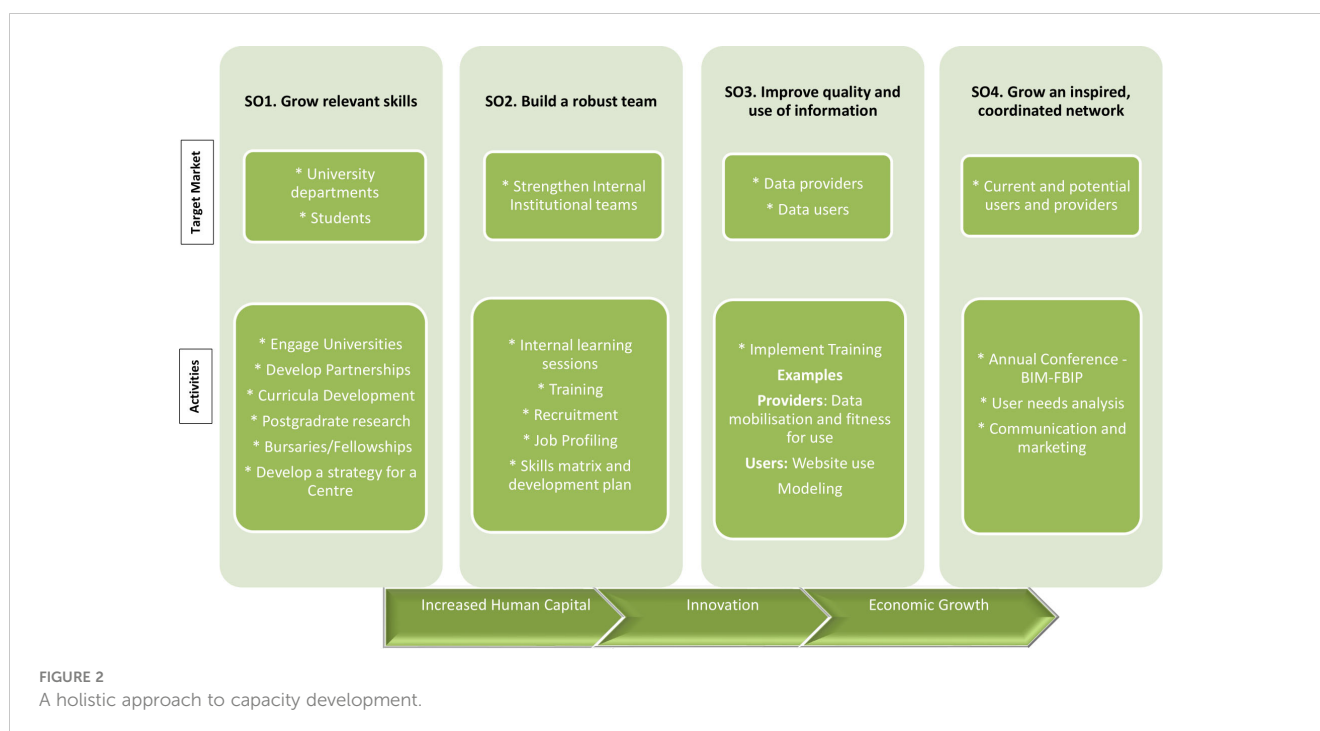
In May 2016, a facilitated session was conducted with the biodiversity science community, at the Joint BIM-FBIP Forum (Box 2). One of the outcomes of that meeting was a series of subsequent engagements with the National Research Foundation to discuss the requirements for a Centre for BDI (detailed later in this study). In line with this, and following an external branch review of SANBI in 2017 (Njobe et al., 2017), it was recommended that the development of a critical mass to enhance Research Leadership was needed. In response to this, a recommendation from SANBI was

made in 2018, to explore the opportunity to develop a *Research Chair* in BDI.

Discussion

The need for increasing human capacity in biodiversity information management was identified by the scientific community in South Africa back in 2007 (Willoughby, 2008; GreenMatterZa, 2009; Coetzer et al., 2012), and it continues to be expressed at the annual joint BIM-FBIP Forum of stakeholders (FBIP, 2020). Efforts to support the need for enhanced capacity has been addressed in several ways, through the efforts of SANBI and the SANBI-GBIF Node over time. A holistic approach to capacity development has been developed (Figure 2), especially since efforts has predominantly focused on short-term, work-based training. The composition of participants has mainly been BDI practitioners, biodiversity scientists and personnel from research and government departments nationally, needing specific skills in their work or study area.

The holistic approach taken addresses activities across four strategic objectives (Figure 2): (1) growing relevant skills, (2) building a robust team, (3) improving the quality and use of information, and (4) growing an inspired and coordinated network. Such an approach was aimed at strengthening the national BDI human capital so that a larger pool of professionals would be available to support this growing field of work. It would additionally improve the quality, use and dissemination of biodiversity data and information, and it addresses several of the opportunities and obstacles identified by the Biodiversity Information Management community in the facilitated capacity building session held in 2016 (Supplementary Table 1). These are further elaborated below.



Strategic objective 1. Grow relevant skills in biodiversity informatics

To deliver on this objective, engagements with institutions of Higher Learning are essential (Parker-Allie et al., 2021). Universities provide the existing, enabling physical infrastructure (buildings, lecture halls, laboratories, equipment), the student population and the opportunity to implement a BDI curriculum as a part of undergraduate and postgraduate degree programmes. Capacities can also be developed through research project activities, which often form a key part of postgraduate degrees.

Discussions with universities to support capacity development in BDI, through the establishment of Centres focussed on Biodiversity Information Management activity, have also been conducted (Harebottle et al., 2016). Such centres have the potential to rapidly grow the field of BDI and will have a marked impact for *Science, Technology and Innovation* through the generation of increased MSc and PhD outputs. These centres will also lead to the production of research-led publications, and have innovation and economic outcomes (Department of Science and Technology, 2007; Department of Science and Technology, 2019). As noted earlier, a Five (5) Year Strategy for a Centre for BDI has been developed (Harebottle et al., 2016), and this addresses some of the challenges and opportunities outlined previously (Box 2). Table 1 shows a detailed breakdown of activities and investments over time.

The Five (5) Year Strategy provides detailed guidelines with regards to potential models for a Centre, as well as the requirements and outputs. Three models for a Centre are proposed, with various hosting and co-hosting options by academic institutions, research entities (e.g., SANBI) and partnership links with research organizations and other

institutions (Figure 3). The Five (5) Year Strategy (Harebottle et al., 2016) also provides a framework for the development of content and curricula suitable for a BDI honours degree (a postgraduate specialisation following an undergraduate degree) or an extended elective module. Strategic partnerships have already been established (via MoUs and/or collaboration agreements) with the University of Western Cape, the University of Sol Plaatje and University of Cape Town to support capacity development endeavours (Figure 4). These efforts have resulted in the appointment of postdoctoral students, the provision of fellowships and the development of research projects focused on the use of open data, through various data pipelines. BDI course content has also been developed to support undergraduate and postgraduate modules at the University of the Western Cape and Sol Plaatje University (Sol Plaatje University, 2020). Additional efforts to strengthen engagements with academia will continue, especially to support efforts to grow research outcomes and to grow capacities at various levels like MSc and PhD.

Strategic objective 2. Grow a robust internal team

Having a robust internal team that will support capacity development for BDI is essential. A job skills profile was conducted by SANBI in 2012 to understand the skills, scope of jobs, and roles and responsibilities that would be required to adequately support national needs. While this exercise was partly initiated in response to the recruitment of personnel into the organisation and/or sector, it also allowed SANBI and partners to improve the sharing and harmonisation of data, as well as the analysis and publishing of said data. At a more global level, this type

Box 2. National stakeholder engagements and efforts towards the development of a Centre for BDI

The Joint BIM-FBIP Forum in 2016 was attended by over 50 national participants (SANBI, 2016; Supplementary Table 2). Here, SANBI-GBIF led a facilitated session focusing on a holistic approach to capacity development in BDI. This focused on both academic engagements with universities and training programmes/events for professional skills development, with the long-term vision for exploring the requirements, potential opportunities, contributions, and obstacles for establishment of a Centre for BDI. The key objectives of this session were twofold, 1. to determine the issues that delegates thought needed to be unlocked to realise a fully operational Centre for BDI, 2. To determine the key opportunities to move the Five (5) Year Strategy for a Centre towards implementation.

The issues/concerns and opportunities identified by the participants were classified into nine focus areas (Supplementary Table 1). These included: 1 funding, 2 relevant skills required, 3 the need and opportunity, 4 interest and communication, 5 content, 6 partnerships and collaborations, 7 science-policy interface, 8 regional collaboration and 9 institutional buy-in, career relevance and private sector opportunities.

Funding was identified as one of the biggest constraining factors with investments required for infrastructure, sustainability and scholarships, as well as for research into emerging areas and for catalysing affiliations with other institutions. While a key challenge, funding was also recognised as a potential opportunity and it was advised that a fund-raising strategy should be developed. It was also suggested that the opportunity should be leveraged, where institutions were already providing funding for research and postgraduate studies.

The lack of persons with the necessary skills to develop such a Centre, to train students and develop training materials that could also be used to “train-the-trainers” was flagged as critical. In the context of developing potential new multidisciplinary BDI modules, the global curriculum for BDI was discussed, and it was agreed that opportunities for knowledge transfer, international collaboration, exchange programmes and experiential learning should be addressed.

Building collaborations and partnerships between committed partners was considered key to success, and it was suggested that SANBI’s network of partners should be leveraged to effect this, especially with tertiary institutions. This would provide, at the very least, experiential learning opportunities. In line with this, it was suggested that co-hosting options should also be considered, to develop a network of collaborators with university partners. Other collaborations identified included partnerships with the science-policy interface Directorate of the national Department of Forestry, Fisheries and the Environment, to mainstream biodiversity data into decision-making. Additionally, exploring avenues or prospects with the private sector was also suggested.

It was recommended that career pathing processes by institutions should be implemented at a strategic level. This includes accessibility to clear development pathways for technical versus academic positions. It was also indicated that the institutions themselves need to recognise and buy-in to the need for information managers, enabling clear opportunities for graduates. Hence, there needs to be a link between supply and demand. Additionally, exploring avenues or prospects with the private sector was also suggested.

of exploration was also conducted for the bioinformatics field of science (Welch et al., 2014).

Strategic objective 3. Improve the quality and use of information

Given that the world is increasingly becoming more data-driven (IEAG, 2014; The Economist, 2016), it is critical that we improve the quality and use of data in the science of BDI. The target markets for training activities can be identified as data providers/publishers and data users. To best support these communities, and various research, data analysis, publishing and use options (Asase and Peterson, 2016; Peterson et al., 2018; Freeman and Peterson, 2019), training has been identified as critical. This will enable the mobilization and publishing of high quality data which is fit-for-use (Hill et al., 2010; Chapman et al., 2020) and comprehensive in terms of taxonomy, geographic and temporal scope.

In this regard, a number of targeted capacity development opportunities have been offered by SANBI and the SANBI-GBIF Node since 2008 (Supplementary Table 3). The topics covered have included biodiversity data standards, species distribution modelling using R, improving fitness-for-use of biodiversity data, data management and cleaning, biodiversity georeferencing etc. (Figure 5). More generally, training opportunities provided by GBIF have also been taken advantage of, when available. These have included those associated with the *Biodiversity Information for Development* Initiative e.g., those hosted and supported by the JRS funded *Africa Biodiversity Challenge* initiative and SANBI-GBIF (GBIF, 2017), (GBIF, 2018). These courses include the Biodiversity Data Mobilisation (GBIF, 2018) and the Data Use for Decision Making. The content for the latter course has been adapted, from the original course (GBIF, 2022b); GBIF Secretariat, 2022a), and evolves

as GBIF enhances the curriculum following these training events. A list of training events can be found in Supplementary Table 3.

Other SANBI implemented initiatives such as the Foundational Biodiversity Information Programme (FBIP) have provided funding opportunities for partner institutions to conduct biodiversity related training workshops (Supplementary Table 4). This investment into the training and capacity building efforts supports the need raised at the facilitated session of 2016 (Supplementary Table 1).

SANBI-GBIF has also put in place an eLearning Platform (SANBI-GBIF, 2022), which can be accessed through the SANBI-GBIF website. This platform aims to be a repository of BDI course contents offered by the Node. Agendas, academic course content, lecture presentations, videos, scripts for relevant informatics software programmes, are made accessible for re-use and for subsequent training. This will enable stakeholders to pick up modules that are relevant for specific training needs. Following a registration process, stakeholders can access SANBI-GBIF training content and materials for topics such as “Fitness for Use of Biodiversity Data”, “Species Distribution Modelling”, and “Data Management and Cleaning supporting Science, Policy and Sustainable Development”. The scope of the training will increase as more training events and courses are rolled out.

At the global scale, training courses, resources and materials, especially those related to the GBIF nodes have been summarised elsewhere (Parker-Allie et al., 2021). But there are others, such as training in Data Carpentries and BDI (Peterson and Ingenloff, 2015) that also support the development of fundamental data skills needed to conduct research, and so provide researchers with high-quality, domain-specific training covering the full lifecycle of data-driven research (management, analysis and use).

A “train-the-trainers” mechanism to grow local expertise has also been employed since 2012 and is ongoing. The two areas of

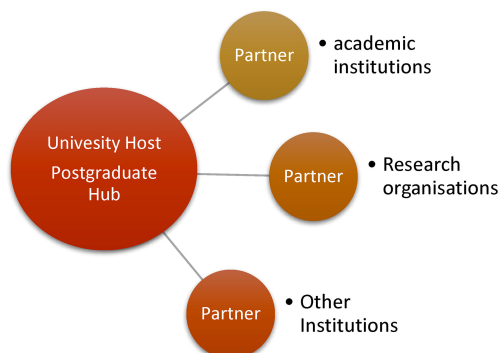
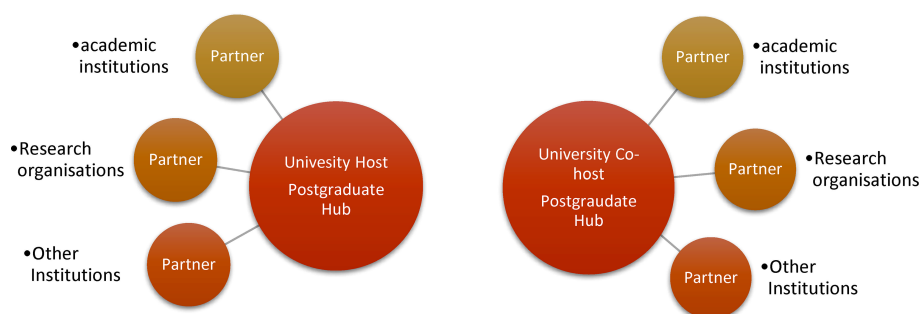
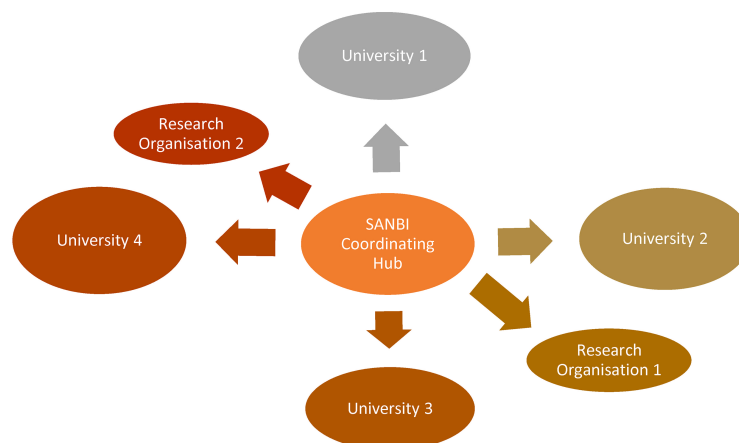
Model 1. Centre for BDI Hosted by a Single Academic Institution**Model 2.** Centre for BDI Hosted by an Academic Institution with a Co-host from another Academic Partner**Model 3.** Decentralised Centre for BDI hosted by a Research Institute like SANBI, with identified Partnerships with academic institutions and other research organisations

FIGURE 3

Three proposed models for the Centre for BDI, with various hosting and co-hosting options by academic and research organizations.

focus have been in biodiversity data geo-referencing and biodiversity data management. In 2012, an initial workshop was supported by the VertNet international team of experts, who also supported a follow-up geo-referencing workshop. Training was provided to more than 30 participants, and a small core team of local trainers have conducted subsequent training. These local trainers already had skills in geo-referencing and were then also able to act as trainers and support more training events (FBIP, 2021). The same mechanism was also employed for

the establishment of a core team training in the field of Data Management.

Strategic objective 4. Grow an inspired and coordinated network of partners

The fourth strategic objective has been to grow an inspired and coordinated network of partners, and the annual BIM-FBIP Forum

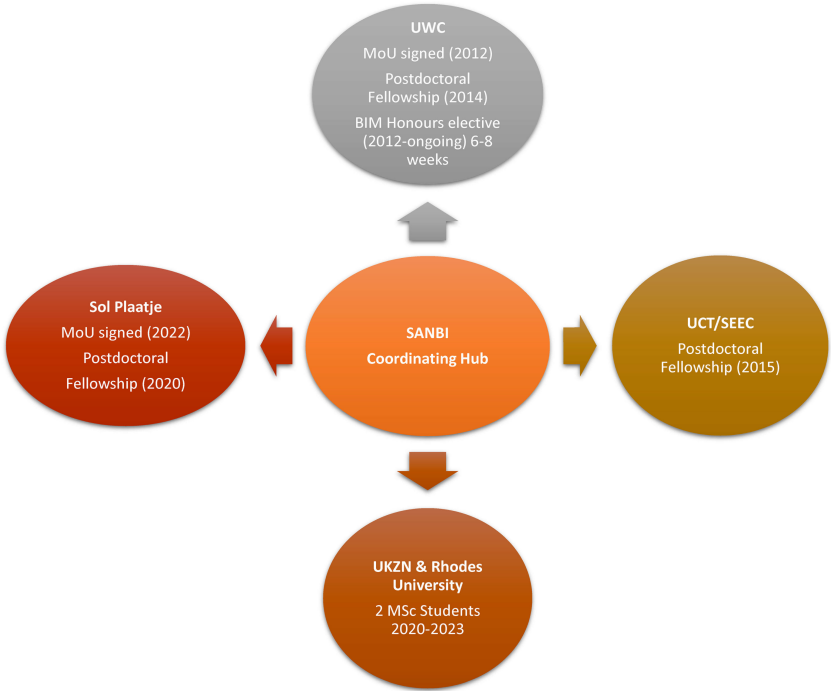


FIGURE 4
Strategic partnerships and progress in developing BDI as a science.

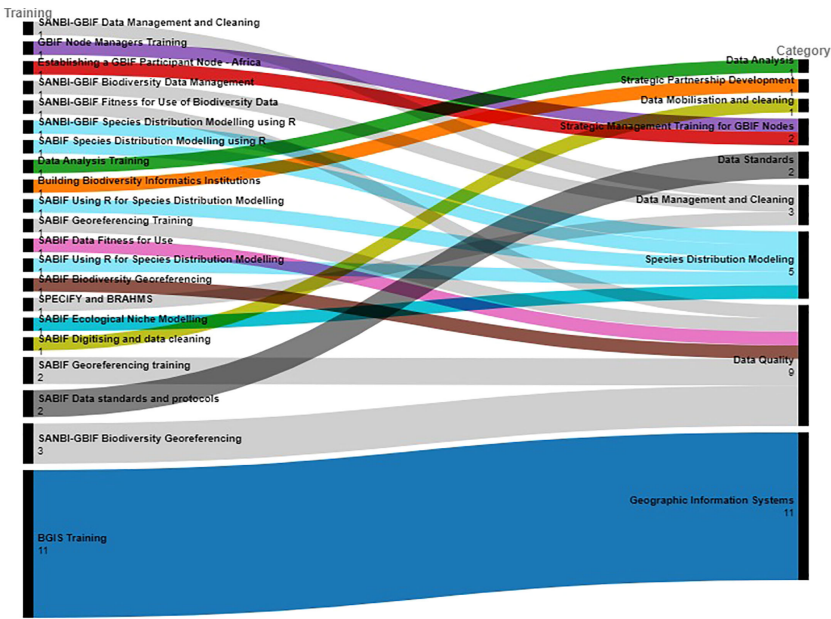


FIGURE 5
Training workshops/short courses offered or supported by SANBI (Biodiversity Information Management Directorate) and SABIF/SANBI-GBIF Programmes over time. The numbers represent the number of training events offered over time. The training events were categorised into data content areas on the left.

provides a platform for engagement to this end (see above). To further support the coordination of BDI, training events and stakeholder workshops are often held alongside the Forum (SANBI, 2016; SANBI, 2019; FBIP, 2020).

Having a national platform to engage stakeholders and share experiences is important to ensure ongoing engagement between partners and the growth of BDI programmes and endeavours. More recently, these Forums have also been instrumental in connecting the biodiversity science and information management community to the *Science, Technology and Innovation* agenda (Box 3). In 2019, the theme for the BIM-FBIP Forum was “Biodiversity Open Data Supporting Open Science, Technology and Innovation”. Here the aim was to look at how to grow efforts and galvanise the thinking of the community, in-line with the objectives of the White Paper in *Science, Technology and Innovation*. It aimed to prepare and ready the community for effective engagements in the DSI Decadal Plan when this came into play. It was identified that there are several initiatives funded by the DSI i.e., SANBI-GBIF, the *Foundational Biodiversity Information Programme* and the *Natural Science Collections Facility*, as well as many activities taking place in line with this. Thus, in part of the Forum it was identified to look at how the concepts presented in the *Science, Technology and Innovation* White Paper could be taken forward strategically, within institutions, across institutions and more broadly i.e., regionally and globally.

Global opportunities that support our biodiversity informatics work in South Africa

For many countries, especially developing countries, a case needs to be made for the investment of national resources towards participating in global initiatives. The value and subsequent impact need to be demonstrated. It is therefore imperative that participation at the global level provides value and makes an impact at the local level and vice versa, through the sharing of relevant information with stakeholders and appropriate work planning through implementation plans, to effect change. To

demonstrate greatest value, it is imperative that countries are also able to inform the global agenda, and to fully contribute to and participate in global initiatives.

The GBIF Graduate Researchers Award provides an opportunity to raise awareness and increase the visibility of BDI, and the value of data mobilisation. The award fosters innovative research and discovery in BDI by graduate students, whose studies rely on GBIF mediated data, in countries participating in the GBIF network. To support national efforts to develop and grow BDI efforts and capacity, SANBI-GBIF has developed a national process to support the selection of candidates and has established a SANBI-GBIF Young Researchers Award Advisory Panel, to provide support and foster additional champions for this scarce skills area of science.

A partnership project entitled “BioDATA Advanced – Accelerating biodiversity research through DNA barcodes, collection and observation data”, was approved for funding and is being led by the Museum of Natural History at the University of Oslo. This initiative is a collaboration between national GBIF Nodes from Norway, South Africa and the Altai State University (Russia), and is being funded by the Norwegian Agency for International Cooperation and Improvement of the Quality of Higher Education (DIKU). The project will offer young researchers academic mobility, as well as professional training in the study of biodiversity using modern methodologies in *inter alia* processing, publishing and using open data (University of Oslo, 2022). Eight courses in advanced biodiversity data skills are planned, and six student MSc and PhD internships will be provided. The courses are designed to create a network of exchange between professionals and students through targeted internships, in the participating countries and around the world.

Mainstreaming our data to support the science–policy interface

One of the opportunities highlighted by the South African community (Supplementary Table 1, Box 2) as contributing towards enhanced BDI endeavours is ensuring good partnerships

Box 3. National STI policy drivers towards change for biodiversity informatics

Two key policy drivers including the *Science, Technology and Innovation* White Paper (DST, 2019) and the *Science, Technology and Innovation* Decadal Plan provide key opportunities and an enabling framework for taking BDI capacity building efforts forward. The White Paper sets a vision for *Science, Technology and Innovation* to accelerate more inclusive and sustainable socio-economic development and improving the quality of life of its citizens. Here some key focus areas or goals provide the ideal mechanisms to support the development of BDI. This includes supporting a more digital society, targeted strategic internationalization, increasing funding across the *National System of Innovation*, expanding the research enterprise (e.g. increasing *Centre's of Excellence*, and *Research Chairs*) and transforming the human capacity for *Science, Technology and Innovation* (DSI, 2021).

The *Science, Technology and Innovation* White Paper, is based on an extensive review of the *National System of Innovation*, which is described as a system of interconnected institutions to create, store and transfer the knowledge, skills and artefacts that define new knowledge (DSI, 2021). Data and information are the building blocks of new knowledge (Chaim, 2013). With the huge volumes of data being mobilised through various new technological advances, it is imperative that we have the computational abilities and capacities in place to use, analyse and mine this data. This especially, towards ensuring outcomes related to societal grand challenges (DSI, 2021). Initiatives such as the FBIP are just one data intensive programme, providing investment in the generation of new data and knowledge, which ultimately informs policy aspects such as global change and the bio-economy (Foundational Biodiversity Information Programme, 2020). One such impactful outcome includes the work conducted through the SeaKeys consortium project, with data mobilised through this project contributing to the expansion of the South African coastal area under marine protection, from 0.4% to 5% (Save our Seas Foundation, 2019; Sink et al., 2019; Parker-Allie, 2021).

The *Science, Technology and Innovation* White Paper also sets out a long term policy direction for government to ensure a growing role for the sector in South Africa, and thereby supporting the objectives of the National Development Plan for 2030 (National Planning Commission, 2015). This policy aims to help South Africa benefit from global developments in rapid technological advancements and respond to the four societal grand challenges, i.e., climate change, future-proofing education and skills, re-industrialising the modern economy and future of society, as well as two *Science, Technology and Innovation* priorities (health innovation and energy innovation).

and alignment of biodiversity data and information with the Science–Policy Interface. South Africa, through SANBI-GBIF has supported the efforts of the IPBES, as a member of the Task Force on Knowledge and Data. As part of its workplan, the Task-Force has developed a Data Management Policy (IPBES, 2020b) to provide overarching guidance on the management of data and knowledge in current assessments and IPBES products. The policy is grounded in the principles of open science (UNESCO, 2021), accessibility, and building knowledge through partnerships. Efforts have also included the development of downloadable curriculum and webinars, to support IPBES authors, in the development of IPBES assessment chapters or other IPBES knowledge products (IPBES, 2020a). These modules cover topics ranging from data management policy, reports, active research data, tools, and examples (Seebens et al., 2022). Efforts relating to the management of active research data were also supported by SANBI-GBIF through the development of content and video presentation (Parker-Allie and IPBES task force on knowledge and data, 2020).

An additional objective of the Task Force on Knowledge and Data is the development of templates and guidelines for IPBES authors on knowledge gaps identification and knowledge catalysis, as identified in the IPBES 2030 Rolling Workplan (IPBES, 2019). Gaps will be used to engage and dialogue with research funders and programmers, to catalyse investment in priority research and data mobilization. In June 2022 as part of the Sustainability, Research, and Innovation Congress, which is a joint initiative of Future Earth and the Belmont Forum, dialogue workshops on the identification of knowledge gaps in the IPBES Global Assessment of Biodiversity and Ecosystem Services was conducted. These efforts are all relevant for mobilising data and knowledge at the national and regional level. Nationally, this also provides an opportunity to leverage funds from existing biodiversity programmes and prioritise these gaps identified through the IPBES Assessments.

At the national level, contributions to the data–science–policy interface are also ensured through ongoing engagement with the IPBES Focal Point, for inclusion into the IPBES Plenary Meetings. IPBES is also a standing item on the BIM-FBIP Forum Agenda, which is in-line with the recommendations from the Joint BIM-FBIP Forum in 2016, indicating that partnerships, synergy and alignment be sought with DFFE, and that data should be mainstreamed to support the science–policy interface and be used for effective decision-making. This engagement with IPBES also supports downstream activities with the scientific community to catalyse data mobilisation activities at national level to support data and knowledge gaps identified in the IPBES assessments, and ensuring alignment with the data and knowledge needs identified in the National Biodiversity Assessment (Skowno et al., 2019). Data gaps exist with the estuarine realm being identified as the most threatened realm in South Africa and freshwater fish identified as the most threatened species group in the country, and freshwater invertebrates identified as a challenge. The marine realm was also identified as lacking adequate taxonomic knowledge, limited occurrence records and a lack of abundance and long-term population trend data, insufficient knowledge of species life histories and ecology, limiting marine threat assessments (Sink et al., 2019; Skowno et al., 2019).

A roadmap for advancing biodiversity informatics and the way ahead

Partnerships and collaboration

This article has set out to provide an overview of the processes that have been undertaken to drive BDI efforts in South Africa, and therefore provides the baseline for a roadmap for the advancement of biodiversity informatics capacity development in the country. In line with some of the opportunities highlighted from the 2016 stakeholder workshop, partnerships with universities have been established and areas of cooperation have been identified. As has previously been stressed, research/academic expertise is the backbone of skills development and high-level partnerships need to be reinforced with additional capacity that can support teaching and research outputs, whilst leveraging internal staff capacities at the universities.

In 2012 an MoU was signed with the University of Western Cape and in March 2022, SANBI and the Sol Plaatje University signed an MoU for cooperation to strengthen efforts to grow the field of BDI science. Future efforts must focus on the development of courses and curriculum content to grow capacity in BDI and data science, to grow human capital through provision of bursary opportunities, internships, postdoctoral fellowships and the development of collaborative research projects.

Experiential learning opportunities through partnerships with Iziko Museums and the City of Cape Town Biodiversity Branch will be explored from 2023, using the Groen Sebenza internship initiative as a model. The latter is a job creation programme that aims to provide a bridge for graduate students leaving university into the work environment, and is funded by national Treasury (SANBI, 2013). With Groen Sebenza it is hoped to support research activities and to develop critical skills in BDI and collections' management in the process. Additional opportunities to expand experiential learning efforts will be through the BioData Advanced initiative, as mentioned above.

Establishment of the Centre and funding opportunities

To look at how a Centre for BDI could be established, a workshop was held in August 2016 which was catalysed by GreenMatter^{za}, and included SANBI and the Centres of Excellence team at the *National Research Foundation*. It was identified that a good instrument to move this work forward would be the *South African Research Chairs Initiative* that is funded by the national *Department of Science and Innovation* and the *National Research Foundation*. This initiative is aimed at strengthening the research and innovation capacity of public universities in strategic institutional niches of excellence. This appointed *Research Chair* would be catalytic in supporting student activity and would create momentum in this research area. The Terms of Reference for a *Research Chair* have been developed, and the process will be taken forward as the

opportunity for funding arises. Engagements between SANBI and the *Department of Science and Innovation* have otherwise been ongoing since 2022 to further identify mechanisms and instruments to move capacity building efforts forward.

Funding for bursaries and studentships

Some funding prospects exist by tapping into bursary opportunities available through the universities, SANBI (Joan Wrench Scholarship Fund) and the National Research Foundation. That said, we need to scale up our efforts in this regard by identifying and engaging with relevant players in the biodiversity science community, and by identifying suitable projects that will support BDI initiatives.

Curriculum and content development

To develop relevant content and materials, SANBI-GBIF has been developing modular re-usable course content and curriculum in component areas of BDI, as highlighted above. *The University of Western Cape* has successfully implemented a 6–8 week BSc Honours module focused on Biodiversity Information Management which has been available since 2012 as part of the Biodiversity and Conservation Biology Honours degree (Parker-Allie et al., 2021). Going forward, SANBI-GBIF will be developing a course in collaboration with GBIF-Spain that will investigate data mining approaches for impactful data use cases and stories (GBIF, 2022c). Other courses that are planned will be developed within the framework of the BioData Advanced initiative with a focus on the mobilisation of molecular, observation and natural history collection data (University of Oslo, 2021; SANBI-GBIF, 2021).

Challenges towards development of a Centre for BDI

While much has been achieved in pursuit of creating a Centre for BDI in South Africa, we have experienced a number of challenges, and we reflect on those here, because they may prove to be valuable to other countries that wish to embark on a similar journey. Our initial efforts were very ambitious, and they were aimed at ensuring high level support and buy-in from both government departments and institutions, and universities. While this was, and remains, of utmost importance, a three-pronged approach should perhaps have been taken. This would additionally have included engagement with a greater critical mass of research expertise to help drive efforts that would support capacity development, as well as a clear science plan. The latter ensuring a more targeted approach towards impactful research outcomes for both science and policy.

Our successes in growing BDI capacity/science since 2012 have largely been achieved through the existing series of programmes and initiatives of SANBI and the managed network. However, to be

truly successful in the creation of a Centre for BDI, several additional supportive components should perhaps have been in place. This includes: (1) A governance structure and identified leadership at various levels that could provide oversight and advisory support i.e., a Board and/or Advisory/Steering Committee including Working Committees depending on the need/s. (2) A clear funding strategy, with an approved budget and a plan for financial sustainability. (3) A clear science strategy and implementation plan for research and Human Capital Development, building in collaborators to lead research and teaching components. (4) A marketing and awareness raising plan to promote the ambitions and opportunities provided by the Centre. (5) A monitoring and evaluation plan that would ensure the Centre meets key performance areas, metrics, and targets. This would support administrative and business plan reporting related to outcomes/outputs to government.

Conclusion

This review article has highlighted our efforts towards the development of BDI as a field of science in South Africa. Although significant achievements have been made in enhancing capacity in BDI, we are not yet at the stage where a Centre has been established.

The study has provided some key lessons that can be implemented by other countries in pursuit of the same goal. For us, the following points are key, including: (1) The annual Forum has been invaluable in developing of a community of practice in BDI, as it provides all interested parties with an opportunity to openly discuss issues, to unblock challenges and to catalyse new projects and endeavours, through strategic engagement with other stakeholder groups. (2) A holistic approach to capacity development has enabled activities across several key strategic areas and target markets. (3) The implementation of platforms such as the SANBI-GBIF website and the eLearning platform, has provided local context, and access to the data published by the South African community, as well as access to the BDI course content implemented by the Node, respectively. (4) High level buy-in and support from national government. (5) Ongoing training events provide capacity development opportunities to learn a range of skills (6) A strong pool of research and teaching expertise in this field of science is lacking, thus there is a clear need for high level skills that could perhaps be enabled through funding for a *Research Chair*. (7) Strong institutional buy-in by host institution/s and associated sustainable resources is needed. (8) A critical mass of students and researchers involved in BDI is needed. (9) Internationalization efforts will be required to grow research and teaching capacities.

For South Africa, it is imperative that the value of the *Science, Technology and Innovation* policy agenda and the alignment with the DSI decadal plan be leveraged to ensure implementation with activities, outcomes and impacts related to the societal grand challenges i.e. futureproofing education and skills, human resources for STI, climate change, expanded and transformed research systems,

expanded and transformed strategic internationalisation, to establish a fully operational Centre and to grow biodiversity informatics as a field of science in South Africa.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

FP-A conceptualized this research and was the primary investigator and author. FP-A leads this research area of work for SANBI in collaboration with University Partners. MG provided critical thinking and strategic input into this research endeavour as a key university partner, and contributed towards the research paper. DH was also a key university partner in this work and was instrumental in the strategic thinking with regards to the Centre. All authors contributed to the article and approved the submitted version.

Acknowledgments

FP-A would like to thank the Department of Science and Innovation for ongoing support to the SANBI-GBIF Node. FP-A also wishes to acknowledge GBIF, and specifically for support

through the GBIF Graduate Researchers Award. We also wish to thank all contributors of the BIM-FBIP Forums who have contributed to the capacity development sessions in the Biodiversity Information Management arena over time. The Authors also wish to thank the reviewers for their valuable inputs and contributions to the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2023.1107212/full#supplementary-material>

References

- Anderson, R. P., Araújo, M. B., Guisan, A., Lobo, J. M., Martinez-Meyer, E., Peterson, A. T., et al. (2016). *Final report of the task group on GBIF data fitness for use in distribution modelling: are species occurrence data in global online repositories fit for modeling species distributions? the case of the global biodiversity information facility (GBIF)* (Copenhagen: Global Biodiversity Information Facility). doi: 10.13140/RG.2.2.27191.93608
- Asase, A., and Peterson, A. T. (2016). Completeness of digital accessible knowledge of the plants of Ghana. *Biodiversity Inf.* 11, 1–11. doi: 10.17161/bi.v11i0.5860
- Auer, T., Barker, S., Borgmann, K., Charnoky, M., Childs, D., Curtis, J., et al. (2022). *EOD – eBird observation dataset* (New York, United States of America: Cornell Lab of Ornithology). doi: 10.15468/aomfnb
- Berners-Lee, T. (1999). "The original design and ultimate destiny of the world wide web by its inventor," in *Weaving the web*. New York, United States of America.
- Burrows, M. T., Bates, A. E., Costello, M. J., Edwards, M., Edgar, G. J., Fox, C. J., et al. (2019). Ocean community warming responses explained by thermal affinities and temperature gradients. *Nat. Climate Change* 9, 959–963. doi: 10.1038/s41558-019-0631-5
- Chaim, Z. (2013). Conceptual approaches for defining data, information, and knowledge. *J. Am. Soc. Inf. Sci. Technol.* 58 (4), 479–493. doi: 10.1002/asi
- Chandler, M., See, L., Copas, K., Bonde, A. M. Z., López, B. C., Danielsen, F., et al. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biol. Conserv.* 213, 280–294. doi: 10.1016/j.biocon.2016.09.004
- Chapman, A. (2005). *Uses of primary species- occurrence data* (Copenhagen: Global Biodiversity Information Facility).
- Chapman, A., Belbin, L., Zermoglio, P., Wicczorek, J., Morris, P., Nicholls, M., et al. (2020). Developing standards for improved data quality and for selecting fit for use biodiversity data. *Biodiversity Inf. Sci. Standards* 4, 1–47. doi: 10.3897/biss.4.50889
- Coetzer, W., Gon, O., Hamer, M., and Parker-Allie, F. (2012). A new era for specimen databases and biodiversity information management in south Africa. *Biodiversity Inf.* 8, 1–11. doi: 10.3897/zokeys.209.3083
- DEA (2016). *National biodiversity economy strategy (NBES)* (Pretoria: Department of Environmental Affairs). Available at: <https://www.environment.gov.za/sites/default/files/reports/nationalbiodiversityeconomystrategy.pdf>.
- DSI (2021). *Science, technology and innovation decadal plan* (Pretoria: Department of Science and Innovation). doi: 10.31826/9781463236984-toc
- DSI (2022). *A National Big Data Strategy For Research , Development And Innovation* (Pretoria: Department of Science and Innovation). Available at: https://www.csir.co.za/sites/default/files/Documents/BDpublicationFinal22021003_0.pdf.
- DST (2007). *Ten-year innovation plan for science and technology* (Pretoria: Department of Science and Technology). Available at: <https://www.dst.gov.za/images/pdfs/>.
- DST (2019). *White paper on science , technology and innovation* (Pretoria: Department of Science and Technology). Available at: <https://www.dst.gov.za/index.php/legal-statutory/white-papers/2775-white-paper-on-science-technology-and-innovation>.
- Faulkner, K. T., Robertson, M. P., Rouget, M., and Wilson, J. R. U. (2014). A simple, rapid methodology for developing invasive species watch lists. *Biol. Conserv.* 179, 25–32. doi: 10.1016/j.biocon.2014.08.014
- FBIP (2020) *Annual forum*. Available at: <https://fbip.co.za/annual-forum/> (Accessed 4 April 2022).
- FBIP (2021) *Training workshops*. Available at: <https://fbip.co.za/training-workshops/> (Accessed 3 April 2022).
- Foundational Biodiversity Information Programme (2020). *Foundational biodiversity information programme (FBIP) framework document and funding guide* (Pretoria: SANBI). Available at: <https://fbip.co.za/wp-content/uploads/2020/02/FBIP-Framework-2020.pdf>.
- Freeman, B., and Peterson, A. T. (2019). Completeness of digital accessible knowledge of the birds of western Africa: priorities for survey. *Condor* 121, 1–10. doi: 10.1093/condor/duz035

- GBIF (2017) *Biodiversity data mobilization workshop for Sub-Saharan Africa*. Available at: <https://www.gbif.org/event/11AZg9RKSQekwuGQ6kKGyq/biodiversity-data-mobilization-workshop-for-sub-saharan-africa> (Accessed 22 May 2022).
- GBIF (2018) *Data use for decision making workshop*. Available at: <https://www.gbif.org/event/2BqrOvYZPKo8wcuqYeA6Sm/data-use-for-decision-making-workshop> (Accessed 22 May 2022).
- GBIF (2022a). *GBIF work programme 2023: annual update to implementation plan 2023–2027* (Copenhagen: Global Biodiversity Information Facility). Available at: <https://docs.gbif.org/2023-work-programme/en/>.
- GBIF (2022b) *Biodiversity data use*. Available at: <https://docs.gbif.org/course-data-use/en/> (Accessed 23 May 2022).
- GBIF (2022c) *Cross-continental partnership to investigate data mining approaches for impactful data use cases and stories*. Available at: <https://www.gbif.org/project/CESP2022-005/cross-continental-partnership-to-investigate-data-mining-approaches-for-impactful-data-use-cases-and-stories> (Accessed 18 November 2022).
- GBIF Secretariat (2022a). *Biodiversity data use* (Copenhagen, Denmark: Global Biodiversity Information Facility). Available at: <https://docs.gbif.org/course-data-use/en/biodiversity-data-use.en.pdf>.
- GBIF Secretariat (2022b). *GBIF science review 2021* (Copenhagen, Denmark: GBIF). doi: 10.35035/w3p0-8729.
- Godfrey, L., Funk, N., and Mbizvo, C. (2010). Bridging the science-policy interface: a new era for south African research and the role of knowledge brokering. *South Afr. J. Sci.* 106 (5–6), 1–8. doi: 10.4102/sajs.v106i5/6.247
- GreenMatterZa (2009) *Priority biodiversity skills. biodiversity information Management : human capital development issues and challenges*. Available at: <http://www.greenmatterza.com/uploads/9/1/5/3/91536866/biodiversity-information-management-human-capital-development-issues-and-challenges.pdf>.
- Hadad, S. (2017). Knowledge Economy : characteristics and dimensions. *Management Dynamics in the Knowledge Economy* 5 (2), 203–225. doi: 10.25019/MDKE/5.2.03
- Hardisty, A., Roberts, D. The Biodiversity Informatics Community (2013). A decadal view of biodiversity informatics: challenges and priorities. *BMC Ecol.* 13 (1), 16. doi: 10.1186/1472-6785-13-16
- Harebottle, D., Parker-Allie, F., and Knight, R. (2016). *5 year strategic plan. development of a centre for biodiversity information management, 2015-2019* (Cape Town: SANBI).
- Hill, A. W., Otegui, J., Ariño, A. H., and Guralnick, R. P. (2010). Copenhagen: global biodiversity information facility, 25 pp. Available at: <http://www.gbif.org>.
- Hobern, D., Apostolico, A., Arnaud, E., Bello, J. C., Canhos, D., Dubois, G., et al. (2012). *Global biodiversity informatics outlook: delivering biodiversity knowledge in the information age* (Copenhagen: Global Biodiversity Information Facility). doi: 10.15468/6jxa-yb44
- IEAG (2014). *A world that counts. mobilising the data revolution for sustainable development* (New York, NY: United Nations). Available at: <https://www.undatarevolution.org/wp-content/uploads/2014/12/A-World-That-Counts2.pdf>.
- iNaturalist Contributors (2023). *iNaturalist research-grade observations*. (San Francisco, United States of America: California Academy of Sciences). doi: 10.15468/ab3s5x
- IPBES (2019) *Strengthening the knowledge foundations*. Available at: <https://ipbes.net/o3-strengthening-knowledge-foundations> (Accessed 30 April 2022).
- IPBES (2020a) *Data management tutorials*. Available at: <https://ipbes.net/dmp/tutorials> (Accessed 28 April 2022).
- IPBES (2020b). *IPBES data management policy ver. 1.0. task force on knowledge and data. version 1*. Eds. R. M. Krug, B. Omare and A. Niamir (Bonn, Germany: IPBES secretariat). doi: 10.5281/zenodo.3551079
- King, D. (2004). The scientific impact of nations: what different countries get for their research spending. *Nature* 430, 311–316. doi: 10.1038/430311a
- Manzini, S. T. (2012). The national system of innovation concept: an ontological review and critique. *South Afr. J. Sci.* 108 (9–10), 1–7. doi: 10.4102/sajs.v108i9/10.1038
- National Planning Commission (2015) *National development plan 2030. our future - make it work, the Philippine journal of nursing*. Available at: <https://www.gov.za/issues/national-development-plan-2030>.
- Nelson, G., and Ellis, S. (2018). The history and impact of digitization and digital data mobilization on biodiversity research. *Philos. Trans. R. Soc.* 374 (1763), 2–10. doi: 10.1098/rstb.2017.0391
- Njobe, K., Pyoos, M., Scholes, B., and Barker, N. (2017). *Biodiversity science and policy advice external review on behalf of SANBI* (Pretoria: SANBI). Available at: <https://www.sanbi.org/wp-content/uploads/2018/12/Review-Report.pdf>.
- OECD (1999) *Final report of the OECD megascience working group on biological informatics*. Available at: <http://www.oecd.org/dataoecd/24/32/2105199.pdf>.
- OECD (2013a). Exploring data-driven innovation as a new source of growth: mapping the policy issues raised by big data. *OECD Digital Economy Papers* 222, 1–43. doi: 10.1787/5k47zw3fcp43-en
- OECD (2013b) *New sources of Growth : knowledge-based capital - key analyses and policy conclusions- synthesis report*. Available at: <https://www.oecd.org/sti/inno/newsourcesofgrowthknowledge-basedcapital.htm>.
- Parker-Allie, F. (2021) *Data mobilization efforts supporting science and policy outcomes*. GBIF.org. Available at: <https://www.gbif.org/article/1kxGi42MT4qflz2jXfc65K/data-mobilization-efforts-supporting-science-and-policy-outcomes> (Accessed 6 February 2022).
- Parker-Allie, F., Pando, F., Telenius, A., Ganglo, J. C., Vélez, D., Gibbons, M. J., et al. (2021). Towards a post-graduate level curriculum for biodiversity informatics. perspectives from the global biodiversity information facility (GBIF) community. *Biodiversity Data J.* 9, 1–25. doi: 10.3897/BDJ.9.e68010
- Parker-Allie, F. and IPBES task force on knowledge and data (2020). IPBES Data Management Tutorials - Session 4.1: General introduction to the management of active research data (1.0). *Zenodo* 1–11. doi: 10.5281/zenodo.4018634
- Parr, C. S., and Thessen, A. (2018). “Biodiversity informatics,” in *Ecological informatics: data management and knowledge discovery*, 3rd ed. Eds. F. Recknagel and W. K. Michener (Cham, Switzerland: Springer), 375–399. doi: 10.1007/978-3-319-59928-1
- Peterson, A. T. (2009). Shifting suitability for malaria vectors across Africa with warming climates. *BMC Infect. Dis.* 9, 1–6. doi: 10.1186/1471-2334-9-59
- Peterson, A. T., Asase, A., Canhos, D., de Souza, S., and Wiczeorek, J. (2018). Data leakage and loss in biodiversity informatics. *Biodiversity Data J.* 6, e26826. doi: 10.3897/bdj.6.e26826
- Peterson, A. T., and Ingenloff, K. (2015). Biodiversity informatics training curriculum, version 1.2. *Biodiversity Inf.* 11, 65–74. doi: 10.17161/bi.v11i0.5008
- Rosenberg, E. (2012). Priority skills for biodiversity. *GreenMatterZa*, 1–13.
- SANBI (2013) *Groen sebenza programme – growing our future*. Available at: <https://www.sanbi.org/community-initiatives/groen-sebenza/> (Accessed 19 March 2023).
- SANBI (2016). *Joint biodiversity information management & foundational biodiversity information programme forum programme*. Eds. F. Parker-Allie and L. Pauw (Pretoria: SANBI). Available at: <https://fbip.co.za/annual-forum/>.
- SANBI (2019). “Joint biodiversity information management & foundational biodiversity information programme forum 2019,” in *Biodiversity open data supporting open science, technology and innovation*. Eds. F. Parker-Allie and L. Pauw (Pretoria: SANBI). Available at: https://fbip.co.za/wp-content/uploads/2019/11/Final-2019_08_12-BIMF-FBIP-programme.pdf.
- SANBI and Lewis Foundation (2010). *A human capital development strategy for the biodiversity sector 2010 - 2030* (Pretoria: SANBI). Available at: <https://www.sanbi.org/wp-content/uploads/2021/03/Biodiversity-hcds-august-2010.pdf>.
- SANBI-GBIF (2021) *BioDATA advanced*. Available at: <https://www.sanbi-gbif.org/post/2021/biota/> (Accessed 18 November 2022).
- SANBI-GBIF (2022) *SANBI-GBIF training and eLearning*. Available at: <https://www.sanbi-gbif.org/e-learning> (Accessed 3 April 2022).
- Sarkar, I. N. (2009). Biodiversity informatics: the emergence of a field. *BMC Bioinf.* 10 (SUPPL.14), 1–2. doi: 10.1186/1471-2105-10-S14-S1
- Save our Seas Foundation (2019) *South Africa announces 20 new marine protected areas*. Available at: <https://saveourseas.com/south-africa-announces-20-new-marine-protected-areas/> (Accessed 6 January 2022).
- Schalk, P. H. (1998). Management of marine natural resources through by biodiversity informatics. *Mar. Policy* 22 (3), 269–280. doi: 10.1016/S0308-597X(98)00013-X
- Seebens, H., Drucker, D. P., Hirsch, T., Nelson, H. P., Parker-allie, F., and Thau, D. (2022). *IPBES data management tutorials chapter 4 : data management of active research data*. Eds. N. Krug, R. M. Aboki Omare and B. D. Bonn (Germany: IPBES Task Force on Knowledge and Data). doi: 10.5281/zenodo.4014796
- Sink, K. J., Van Der Bank, M. G., Majiedt, P. A., Harris, L. R., Atkinson, L. J., Kirkman, S. P., et al. (2019). “Marine realm,” in *South African national biodiversity assessment 2018 technical report*, vol. 4. (Pretoria, South Africa: South African National Biodiversity Institut), 1–597.
- Skowno, A. L., et al. (2019) *National biodiversity assessment 2018: the status of south africa's ecosystems and biodiversity*. Available at: <http://opus.sanbi.org/handle/20.500.12143/6362>.
- Sol Plaatje University (2020) *First post-doctoral fellow at SPU*. Available at: <https://www.spu.ac.za/index.php/first-postdoctoral-fellow-at-spu/> (Accessed 10 November 2022).
- The African Open Science Platform (2018) *The future of science and science for the future*. Available at: https://www.nrf.ac.za/sites/default/files/documents/AOSP_Strategy_Final_HR.pdf.
- The Economist (2016) *The data deluge: five years on*. Available at: <https://perspectives.eiu.com/technology-innovation/data-deluge-five-years>.
- The Lewis Foundation and SANBI (2021). *Biodiversity human capital development strategy review - synthesis report march 2021*. (Pretoria: SANBI). 143–159. doi: 10.1016/b978-0-08-021985-1.50014-x
- Troudet, J., Vignes-Lebbe, R., Grandcolas, P., and Legendre, F. (2018). The increasing disconnection of primary biodiversity data from specimens: how does it happen and how to handle it? *Systematic Biol.* 67 (6), 1110–1119. doi: 10.1093/sysbio/syy044

- UNESCO (2021). *UNESCO Recommendation on open science* (France: UNESCO). Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en>.
- University of Oslo (2021) *BioDATA advanced– accelerating biodiversity research through DNA barcodes, collection and observation data*. Available at: <https://www.nhm.uio.no/english/research/projects/biodata-advanced/>.
- University of Oslo (2022) *BioDATA advanced– accelerating biodiversity research through DNA barcodes, collection and observation data*. Available at: <https://www.nhm.uio.no/english/research/projects/biodata-advanced/> (Accessed 30 March 2023).
- Visalli, M. E., Best, B. D., Cabral, R. B., Cheung, W. W. L., Clark, N. A., Garilao, C., et al. (2020). Data-driven approach for highlighting priority areas for protection in marine areas beyond national jurisdiction. *Mar. Policy* 122, 103927 doi: 10.1016/j.marpol.2020.103927
- M. Walters and R. J. Scholes eds. (2016). *Handbook on biodiversity observation networks*. Pretoria: Springer Open, 326pp. doi: 10.1007/978-3-319-27288-7
- Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., et al. (2014). Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput. Biol.* 10 (3), e1003496. doi: 10.1371/journal.pcbi.1003496
- Willoughby, S. (2008). *Biodiversity information management forum report* (Cape Town: SANBI). Available at: http://biodiversityadvisor.sanbi.org/wp-content/uploads/2012/08/BIMF_2008_Report-final.pdf.
- World Bank (2012) *Measuring knowledge in the world 's economies - knowledge assessment methodology and knowledge economy index*. Available at: https://web.worldbank.org/archive/website01030/WEB/IMAGES/KAM_V4.PDF.



OPEN ACCESS

EDITED BY

Stelios Katsanevakis,
University of the Aegean, Greece

REVIEWED BY

Christophe Botella,
Institut National de Recherche en
Informatique et en Automatique
(INRIA), France
Bharat Babu Shrestha,
Tribhuvan University, Nepal

*CORRESPONDENCE

Amy J. S. Davis

✉ amy.davis@uni-konstanz.de

†PRESENT ADDRESS

Rozemien De Troch,
Belgian Climate Centre, Brussels, Belgium

RECEIVED 20 January 2023

ACCEPTED 17 January 2024

PUBLISHED 09 February 2024

CITATION

Davis AJS, Groom Q, Adriaens T,
Vanderhoeven S, De Troch R, Oldoni D,
Desmet P, Reyserhove L, Lens L and
Strubbe D (2024) Reproducible WiSDM: a
workflow for reproducible invasive alien
species risk maps under climate change
scenarios using standardized open data.
Front. Ecol. Evol. 12:1148895.
doi: 10.3389/fevo.2024.1148895

COPYRIGHT

© 2024 Davis, Groom, Adriaens,
Vanderhoeven, De Troch, Oldoni, Desmet,
Reyserhove, Lens and Strubbe. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Reproducible WiSDM: a workflow for reproducible invasive alien species risk maps under climate change scenarios using standardized open data

Amy J. S. Davis^{1,2*}, Quentin Groom³, Tim Adriaens³,
Sonia Vanderhoeven⁴, Rozemien De Troch^{5†}, Damiano Oldoni³,
Peter Desmet³, Lien Reyserhove⁶, Luc Lens³
and Diederik Strubbe¹

¹Terrestrial Ecology Unit TERE, Department of Biology, Ghent University, Ghent, Belgium, ²Ecology, Department of Biology, University of Konstanz, Konstanz, Germany, ³Research Institute for Nature and Forest (INBO), Brussels, Belgium, ⁴Belgian Biodiversity Platform, Département du Milieu Naturel et Agricole, Service Public de Wallonie, Gembloux, Belgium, ⁵Royal Meteorological Institute of Belgium, Brussels, Belgium, ⁶Meise Botanic Garden, Meise, Belgium

Introduction: Species distribution models (SDMs) are often used to produce risk maps to guide conservation management and decision-making with regard to invasive alien species (IAS). However, gathering and harmonizing the required species occurrence and other spatial data, as well as identifying and coding a robust modeling framework for reproducible SDMs, requires expertise in both ecological data science and statistics.

Methods: We developed WiSDM, a semi-automated workflow to democratize the creation of open, reproducible, transparent, invasive alien species risk maps. To facilitate the production of IAS risk maps using WiSDM, we harmonized and openly published climate and land cover data to a 1 km² resolution with coverage for Europe. Our workflow mitigates spatial sampling bias, identifies highly correlated predictors, creates ensemble models to predict risk, and quantifies spatial autocorrelation. In addition, we present a novel application for assessing the transferability of the model by quantifying and visualizing the confidence of its predictions. All modeling steps, parameters, evaluation statistics, and other outputs are also automatically generated and are saved in a R markdown notebook file.

Results: Our workflow requires minimal input from the user to generate reproducible maps at 1 km² resolution for standard Intergovernmental Panel on Climate Change (IPCC) greenhouse gas emission representative concentration pathway (RCP) scenarios. The confidence associated with the predicted risk for each 1km² pixel is also mapped, enabling the intuitive visualization and understanding of how the confidence of the model varies across space and RCP scenarios.

Discussion: Our workflow can readily be applied by end users with a basic knowledge of R, does not require expertise in species distribution modeling, and only requires an understanding of the ecological theory underlying species distributions. The risk maps generated by our repeatable workflow can be used to support IAS risk assessment and surveillance.

KEYWORDS

uncertainty in SDMs, conformal prediction, spatial sampling bias, ecological models, confidence assessment, invasive alien species

1 Introduction

Climate change and biological invasions represent two of the largest threats to biodiversity in the Anthropocene (Mazor et al., 2018; Urban, 2015). As a result of climate change, it is expected that a wide range of species will migrate to follow their shifting climatic niche and introduced species will find novel areas suitable for their establishment (Bellard et al., 2018). Some of these introduced species are likely to have negative impacts on native biodiversity and human well-being (Simberloff et al., 2013). Assessing the risk of invasion by alien species is a crucial step for proactive management, including identifying species for preventive actions such as legal bans on trade, transport, and possession, targeting early detection efforts both at entry points and in susceptible ecosystems, as well as risk management decisions to remove established populations or limit their further spread (Srivastava et al., 2019). Regardless of the specific protocol used, risk assessment is defined as the standardized evaluation of entry, exposure, and consequence of the introduction of an alien species (Vanderhoeven et al., 2017; González-Moreno et al., 2019). An evaluation of the risks of introduction, establishment, spread, and impact are the four main components of alien species risk assessments (Roy et al., 2017).

Species distribution models (SDMs) are the main tool for forecasting the risk of establishment of an alien species in a spatially explicit way (Guisan and Thuiller, 2005; Jeschke and Strayer, 2008). Correlative SDMs delineate the realized niche of the organism based on species-environmental relationships obtained from georeferenced species occurrence data (i.e., species presence located at specific geographic coordinates) and spatial environmental predictors. This way, SDMs predict the probability of species presence in unsampled areas. Additionally, SDMs also predict environmentally suitable areas where the species is currently absent, but can potentially be established in the future, depending on dispersal success. SDMs can be used to guide spatial decision-making, but recent critiques have highlighted how uncertainties in species distribution modeling practice have hindered their widespread uptake in decision-making workflows (Muscatello et al., 2021; Lee-Yaw et al., 2022; Nguyen and Leung, 2022; Liu et al., 2020). These issues include the impact of methodological choices on model outcomes including accuracy, ease of

interpretation, and predicted distribution (Wenger et al., 2013; Sofaer et al., 2019; Brun et al., 2020). For example, algorithm choice is a major source of variability in model forecasts (Elith et al., 2006; Hallgren et al., 2019). Also, the choice of predictors, parameter settings, and spatial grain are all sources of variability that affect model predictions (Peterson et al., 2018; Fourcade, 2021; Chauvier et al., 2022). In addition to the uncertainty in model predictions stemming from the numerous choices to be made during model development, the failure to record and share these decisions prevents reproducibility (Feng et al., 2019a). Governmental and non-governmental nature conservation agencies often use SDMs to guide management and decision-making regarding invasive species but need transparent and reproducible workflows for acceptance by stakeholders and policy-makers (Schwartz et al., 2018; Ferraz et al., 2021; Baker et al., 2021).

There is an active debate on how to improve the reliability and transferability of invasive species distribution models, and new conceptual and methodological approaches are regularly published (e.g. Barbet-Massin et al., 2018; Bellard et al., 2018; Chapman et al., 2019; Hao et al., 2019; Sillero et al., 2023). However, as far as we are aware, most of these proposed innovations are not geared toward automated reproducibility (Kass et al., 2018; Feng et al., 2019a; Mostert et al., 2023).

To address this, we developed the WiSDM workflow to generate reproducible risk maps for potentially invasive alien species under scenarios of climate change at a high spatial resolution (1 km²). Our workflow semi-automatically: 1) identifies highly correlated predictors; 2) mitigates the impact of sampling bias; 3) generates IAS risk maps for standard RCP scenarios using an ensemble of multiple machine learning algorithms; 4) quantifies spatial autocorrelation in the residuals to assess the impact of clustering of species occurrences; and 5) generates confidence maps for each IAS risk map. This species distribution modeling workflow is part of the Tracking Invasive Alien Species (TrIAS) project, a broader data-to-decision pipeline guiding alien species detection and management (Vanderhoeven et al., 2017). TrIAS encompasses the development and publication of alien species checklists (Reyserhove et al., 2020), identification of emerging species, and risk assessment. WiSDM is written in R

markdown and can be exported as an HTML or notebook file instantly recording all methodological decisions, parameter choices, and outputs, thereby facilitating reproducibility and transparency for risk assessments.

2 Methods

2.1 Overview of WiSDM

Our workflow (Figure 1) uses a hierarchical approach, whereby models are first created at a global scale and then integrated into the European-level models to characterize invasive species' realized niches as extensively as available occurrence data allow. This is achieved by using the model forecast derived from the global model as a probability surface to guide the selection of pseudoabsences for the European model(s). Our SDMs at both the global and European levels use an ensemble of machine learning (ML) algorithms: random forests (RF), gradient boosted machine (GBM), generalized linear model (GLM), and multivariate adaptive regression splines (MARS). These algorithms were chosen because they use distinct approaches: bagging, boosting, linear- and piecewise regression (Table 1). The resulting predictions from each model are stacked together using a GLM as a meta-model to combine the predictions in a weighted combination that optimizes model accuracy (Van der Laan et al., 2007). We used a GLM-based meta-model, instead of a simple averaging of the invasion risk predictions produced by the different modeling algorithms, so that the more accurate models are given a higher weight in the final model while minimizing the risk of overfitting (Hao et al., 2019). The meta-model refers to any statistical or ML model used to combine the information gained from each model's prediction in an ensemble, producing the final model for baseline conditions. Individual country-level maps are a subset of the European model.

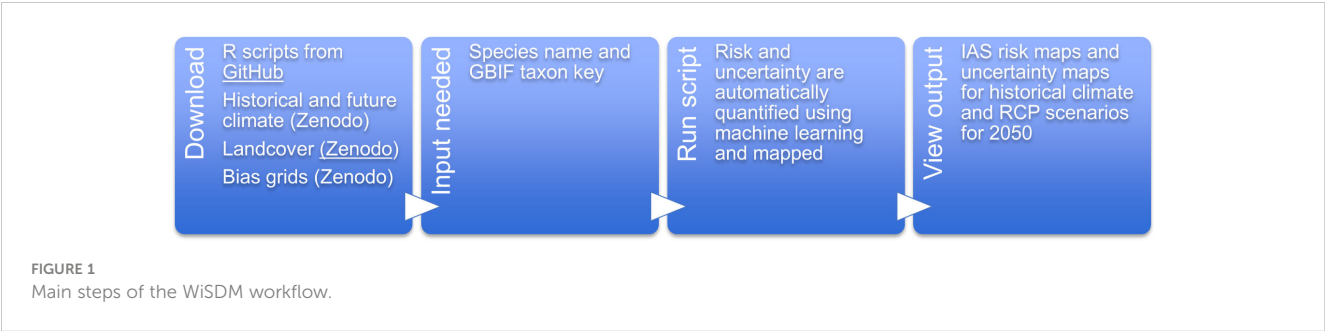
The code necessary to run the workflow is available on GitHub (<https://github.com/trias-project/risk-modelling-and-mapping>).

The global models are climate-only (Pearson and Dawson, 2003), and use high-resolution climate data layers (30 arc second, ~ 1 km) which are available from CHELSA (Karger et al., 2017). The European model uses climate data layers developed specifically for Europe as part of the TrIAS project (De Troch et al., 2020). The TrIAS climate data have been bias-corrected to be compatible with the CHELSA data layers. In order to use our workflow, we have made available the

TABLE 1 Classification algorithms used in the WiSDM workflow.

Algorithm	Type	Technique	Reference
Random Forests (RF)	Supervised	Bagging	Breiman, 2001
Gradient Boosted Machine (GBM)	Supervised	Boosting	Friedman, 2001
Logistic regression (GLM)	Supervised	Regression	Cox, 1958
Multivariate adaptive regression splines (MARS)	Supervised	Piece wise regression	Friedman, 1991

environmental and climate data layers developed for TrIAS via Zenodo (De Troch et al., 2020). The climate layers summarize 30-year climate data (1976-2005), and for three emission scenarios of future climate (the representative concentration pathways (RCP 2.6, RCP 4.5, RCP 8.5) as defined by the IPCC with coverage for Europe. They are based on an ensemble of regional climate models from the EURO-CORDEX archive (Kotlarski et al., 2014), that have been statistically downscaled from a 12.5 × 12.5 km to a 1 km² spatial resolution. WiSDM includes predictors characterizing land use/land cover for Europe derived from the CORINE (Coordination of Information on the Environment) landcover product, anthropogenic pressure from the global terrestrial human footprint dataset (Venter et al., 2016), the distance to the nearest freshwater body, and climate variables based on historical (1976-2005) and future (2040-2070) scenarios. These data have been aligned with the 1 km² EEA Reference Grid (European Environment Agency, 2011). The outputs of WiSDM include 1) a risk map for Europe produced by global ensemble model based on historical climate conditions; 2) risk map(s) produced by European ensemble model based on historical climate conditions; 3) assessment of the predictive performance of all models; 4) country-level risk maps based on European ensemble model for historical climate conditions and under RCP scenarios; 5) country level maps that visualize differences in current vs projected risk under each of the three RCP scenarios; 6) country-level confidence maps; 7) table of variable importance; 8) response curves for all predictors used in the European model; and 9) HTML file from R markdown document saving all code including, decisions, parameters, thresholds, and model outputs (GBIF Secretariat, 2022b). Currently, WiSDM is suitable for modelling plants, mammals, reptiles, amphibians, and birds in Europe, but it can be adapted for other regions. A list of all the predictors used in the workflow and links to download via Zenodo is available (see Data Availability statement). We provide a list of known ecologically relevant predictors for each taxonomic group as Supplementary Information (Supplementary Table 1).



2.2 Occurrence data preparation

The WiSDM workflow utilizes GBIF as it is the largest collector of occurrence data in the world (Waller et al., 2021) with over 2 billion species occurrence records (GBIF, 2023) and follows FAIR data principles (Wilkinson et al., 2016). GBIF has taxonomic and geographic data gaps, notably for insects and Asia, respectively which have been the focus of data mobilization efforts (GBIF Secretariat, 2022a). We recommend that users check for the availability of additional occurrence datasets from regional and national environmental agencies if they are not already present on GBIF. In Belgium, data from the relevant agencies (e.g. the Institute for Nature and Forest Research (INBO), Waarnemingen.be, Florabank) are already contributed to GBIF and regularly updated.

Species names are matched with GBIF taxon keys to download only those occurrences with accepted or synonymous names, minimizing taxonomic uncertainty (GBIF Secretariat, 2022b). All species occurrences that have geographic coordinates and are within the time frame specified were downloaded. The default end dates of 1971 and 2010 were chosen to maximize the number of available observations while staying with ± 5 years of the end dates used for the climate data to minimize a temporal mismatch between the two datasets (Davis et al., 2017). Data with spatial uncertainty greater than 1 km, and duplicate occurrences in the same grid cell are removed. Occurrences that correspond to geographic centroids, biodiversity institutions, and invalid coordinates are flagged and removed using the Coordinate Cleaner package (Zizka et al., 2019). If most of these occurrences are outside of Europe, and there are fewer than ~ 80 –100 occurrences in Europe, we recommend running only the global model until more occurrence data become available. Although it is possible to obtain accurate SDMs with low numbers of occurrences as few as five (Pearson et al., 2007; van Proosdij et al., 2016), we recommend a minimum of 30 and restricting the number of predictors used to the number of occurrences divided by 10 to reduce the risk of overfitting.

2.2.1 Mitigating spatial bias in occurrence data

To achieve large geographic coverage, species occurrence databases that are composed of aggregated species data collections such as those provided by GBIF are often used. A drawback to using these databases is their potential for geographic sampling bias (Beck et al., 2013). Uneven sampling or search effort can mislead conclusions about the extent and drivers of species distributions (Gotelli and Colwell, 2001; Lobo, 2008). Sampling bias in our workflow is mitigated by using taxonomic occurrence grids to exclude areas of low sampling effort from the background when randomly placing pseudoabsences (Phillips et al., 2009). The occurrence grids have a 1-degree spatial resolution in the WGS84 coordinate system (EPSG:4326). Each 1-degree grid cell contains the number of records present in GBIF corresponding to a specific taxonomic group: plants, mammals, reptiles, amphibians, and birds. These are also available for download via Zenodo (Davis et al., 2023).

2.3 Global model

The global climate SDM is constructed using all available species occurrence data, employing CHELSA high-resolution climate data layers to delineate the complete range of suitable climate conditions for each species. The number of pseudoabsences equal to the number of species occurrences (Barbet-Massin et al., 2012) are randomly located in the same ecoregions (Olson et al., 2001) inhabited by occurrences, but not in areas of low sampling effort as indicated by the taxonomic occurrence grid (see below). Ecoregions are hypothesized to delineate the area considered theoretically accessible to the organism (Barve et al., 2011; Guisan et al., 2014). To avoid inflating model performance metrics, pseudoabsences are sampled within relevant ecoregions rather than over a large, unrealistic area (Lobo et al., 2008). These pseudoabsences are then combined with the occurrences to form a presence-pseudoabsence spatial point dataset. We use an equal number of pseudoabsences and presences in both the global and European models because large numbers of pseudoabsences relative to presences bias the model towards predicting absences (maximizing specificity). Reported gains in model performance and accuracy as measured by ROC and AUC are due to gains in specificity (Lobo et al., 2008). For models with 800–1000 presences, one draw of an equal number of pseudoabsences is sufficient, otherwise, at least 10 draws are needed (Barbet-Massin et al., 2012). Highly correlated predictors are identified using the ‘findCorrelation’ command of the ‘caret’ package, which identifies the predictor(s) with the highest mean correlation with all other predictors (Kuhn, 2022). A global risk map is produced at 1 km resolution based on historical climate conditions using ensemble modeling as described above. The spatial extent of the risk map is limited to Europe to reduce computational processing times. The risk map generated from the global model is used as input into the European model so that the placement of pseudoabsences is restricted to areas with a probability of presence less than 0.5.

2.4 European model

As emerging invasive alien species are unlikely to have many occurrences in a particular European country, WiSDM constructs SDMs using occurrences from all of Europe, and then country-level risk maps are a subset of the European model. The European occurrences are a subset of the cleaned global occurrence data used to build the global SDMs. The European level model incorporates the climatic suitability map generated by the global SDM to locate pseudoabsences in areas of predicted low habitat suitability. The pseudoabsences are randomly located in the same ecoregions inhabited by the alien species (as described above) that overlap with the areas of low habitat suitability predicted by the global model. While introduced species may not have had the chance to fully colonize the ecoregions into which they are introduced, restricting the invasive range of pseudoabsence selection to these regions minimizes the chance of selecting pseudoabsences

corresponding to inaccessible environmental conditions (Chapman et al., 2019). As with the global model, taxonomic occurrence grids are used to avoid locating pseudoabsences in areas with low sampling effort. The pseudoabsences are joined to the occurrences to create a European presence-pseudoabsence dataset. The baseline European level risk model uses the historical climate data for Europe described above, LULC cover data, anthropogenic pressure, and distance to water, described above. From this model, risk maps for specific European countries can be obtained. This model is then projected onto future climate data according to the three RCP scenarios only for the country of interest for faster computational processing times. Country-level risk maps are generated automatically for baseline conditions and the RCPs. Difference maps in the current baseline risk as compared to the future risk under the RCP scenarios are also generated.

2.4.1 Addressing multicollinearity

Multicollinearity in SDMs can increase uncertainty and obscure the most salient predictor in driving species distributions, as well as inhibit model transferability (Yates et al. 2018; Feng et al., 2019b; Liu et al., 2020). WiSDM records and removes highly correlated predictors in the European model using the same method described for the global model. After adding habitat and anthropogenic predictors (heretofore referred to as “habitat” predictors) to the filtered climate dataset, the climate and habitat predictors are examined together for multicollinearity. If an ecologically relevant predictor is flagged, we suggest users consult the correlation matrix to identify alternative predictors for removal as there could be a less crucial predictor contributing to the collinearity. While dimension reduction techniques such as principal component analysis can be used to reduce multicollinearity and improve model transferability for invasive species (Petitpierre et al., 2017), the effects of individual predictors on species distributions are obscured. An understanding of the relationships between invasive species and their environment can inform decision-making, hence our choice not to use data reduction methods.

2.4.2 Assessing spatial autocorrelation

WiSDM assesses the residuals from the European level ensemble model for spatial autocorrelation, using Moran's I. Values of Moran's I greater than 0.1 indicate that the occurrence data may be highly clustered and thinning before model fitting should be employed (Boria et al., 2014; Diniz-Filho and Bini, 2005). The option to thin occurrence data is provided in the workflow via the rarefy command from the Humboldt R package (Brown, 2023).

2.5 Model evaluation and validation

WiSDM reports both threshold-independent (AUC) and threshold-dependent (accuracy, sensitivity, specificity, and kappa) measures of model performance for each algorithm using cross-validation (Kuhn, 2008) for both the global and European-level models. The AUC (Area Under the Curve) statistic quantifies the overall ability of a binary classification model to distinguish between positive and negative classes, with an AUC of 0.5 indicating random

performance and an AUC of 1.0 representing perfect discrimination. Accuracy is measured as the number of True Positives + True Negatives/Total Observations. The kappa statistic is measured on a scale of -1 to +1, with 0 indicating the predictive ability of the model is no better than as expected by chance. Sensitivity (true positive rate) and specificity (true negative rate) are reported on a scale of 0-1, with a value of “1” indicating a perfect score. A variety of methods exist to choose a threshold to convert the predicted probabilities to classify a location as either “species present” or “species absent” (Liu et al., 2005). The threshold value can be determined based on ecological knowledge or by optimizing a specific evaluation metric, such as the true positive rate (sensitivity) or the true negative rate (specificity). WiSDM identifies and applies a threshold where sensitivity is equal to specificity with the assumption that the cost of predicting false presences and false absences is the same (Lobo et al., 2008).

2.6 Quantifying and visualizing confidence of model predictions

Uncertainty associated with model predictions and their transferability to new biogeographic areas or novel climate conditions presents another barrier to effective decision-making with SDMs (Brodie et al., 2022). Typically, the accuracy of SDMs is assessed based on how well the model has performed using cross-validation or independent data sets. The dominant methods for quantifying the uncertainty of SDMs are model averaging of the predictive outputs from different algorithms, reporting the standard deviation of the predictions, or using the consensus of the outputs (Thuiller et al., 2019). However, this does not show how good our individual-level predictions are, or how confident we are of them, especially outside the conditions that the model has been calibrated on (‘extrapolation’). Our workflow includes code developed by the authors to implement the conformal prediction algorithm to quantify confidence in model predictions. Conformal prediction is a method that leverages past experience to estimate the level of confidence associated with individual predictions, providing a measure of how likely a prediction is to be correct based on historical data. Conformal prediction is distribution-free and has a guaranteed error rate (Shafer and Vovk, 2008). Using a prediction from any method, conformal prediction produces a conformity measure so that strange or unlikely observations are assigned lower conformity scores as compared to more likely observations. The conformity score also known as a p-value (not to be confused with the P values used for statistical hypothesis testing), is the probability estimate that the observation belongs to a class label, with a $1 - \epsilon$ error rate prediction region, a set that contains y with a probability of at least $1 - \epsilon$ percent. The smaller the error rate, the larger the prediction region becomes. WiSDM defaults to an error rate of 20% to balance confidence levels with prediction region size. In binary classification problems, both classes are often included in the prediction region, regardless of the size of the error rate. To address this, probability estimates for each prediction belonging to a class are obtained separately, yielding p-value A that the prediction belongs to class A, and p-value B that it belongs to

class B. Thus the most likely class label is based on the class with the highest p-value. The confidence that the prediction belongs to that class is $1 - \text{the second highest p-value}$. (Vovk et al., 2005). Conformal prediction has been successfully applied in other fields including Computational biology (Norinder et al., 2014), Medicine (Pereira et al., 2017), and Drug discovery (Alvarsson et al., 2021) but is surprisingly absent from ecological applications. The confidence of each prediction can be visualized in maps, providing an intuitive understanding of how model confidence varies across space and climate scenarios. This can help to identify areas or scenarios where model predictions are less reliable, or where additional data are needed to improve the model. To facilitate the interpretation of the confidence maps, WiSDM can optionally show only those predictions that meet or exceed a user-defined minimum threshold of confidence.

2.7 Use case

We applied WiSDM to a case study species: *Vaccinium corymbosum* L. (North American blueberry, Ericaceae family). This species was identified as a species of potential conservation concern by the TrIAS automated early warning pipeline for prioritizing emerging alien species (Adriaens et al., 2022). North American blueberry is a deciduous shrub that typically grows in moist forests, bogs, and swamps, was introduced to Belgium in the early 1950s and was recently observed to escape from nurseries (Adriaens et al., 2019). This species and its hybrid *Vaccinium corymbosum* × *angustifolium* is considered invasive in Germany and the Netherlands and is known to be problematic in protected areas (wet heathlands, peatlands) there (Schepker and Kowarik, 1998; Penninkhof et al., 2018).

1678 georeferenced occurrences of *V. corymbosum* from 1971–2010 were downloaded from GBIF. After removing centroids, duplicates occurring in the same grid cell, and occurrences with a spatial uncertainty greater than 1 km, 1064 occurrences remained. The majority of these occurrences are located in North America (Figure 2). Of these, only 66 occurrences were located in Europe, with 3 occurrences in Belgium. To account for variability resulting

from the location of pseudoabsences, we ran 10 models for Europe, each with a different draw of pseudoabsences equal to the number of presences (Barbet-Massin et al., 2012). The 10 models were evaluated using 4-fold cross-validation. The model with the highest sensitivity, specificity, Kappa, and AUC was selected and projected onto the RCP scenarios. WiSDM automatically generated confidence maps using a minimum confidence threshold of 0.7 for the best predictive model and RCP scenarios. The R markdown document published from this workflow shows in detail all settings, data, algorithms, parameters used, and model validation results and is included as [Supplementary Information \(Supplementary S1\)](#).

3 Results

3.1 Global model

After filtering for multicollinearity, five climate predictors remained and were used in the models. Annual precipitation, the maximum temperature of the warmest month, amount of precipitation (mm) during the driest month, the annual variation of precipitation, and the range of annual temperature °C. The mean predictive accuracy assessed by 10-fold cross-validation for the algorithms ranged from 0.66 to 0.78 and kappa from 0.32 to 0.55. After ensembling, the final model had a mean accuracy of 0.77, and a Kappa of 0.54. Details regarding the performance of each algorithm, model correlation, and variable importance, as well as maps of the area used for sampling pseudoabsences, are available as [Supplementary Information \(Supplementary S1\)](#). The risk map is shown in [Figure 3](#).

3.2 European model

All occurrences found in Europe (n=66) were used in the European models. The following predictors were used: annual variation of precipitation, maximum temperature of the warmest month, range of annual temperature °C, and percent wetland per 1 km². 10 models were constructed with 10 unique draws of

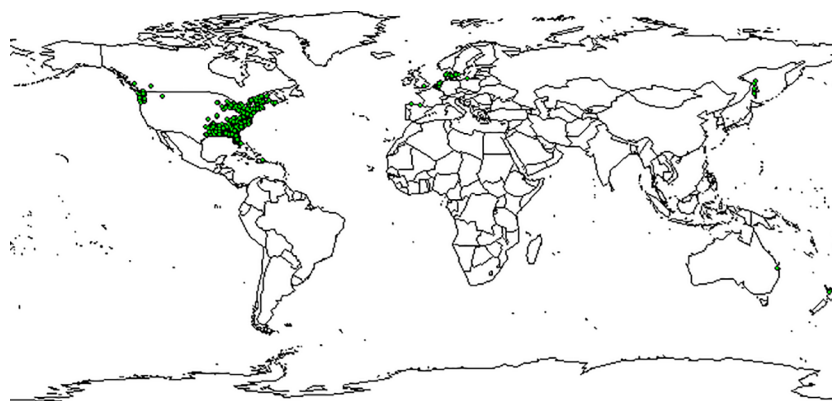
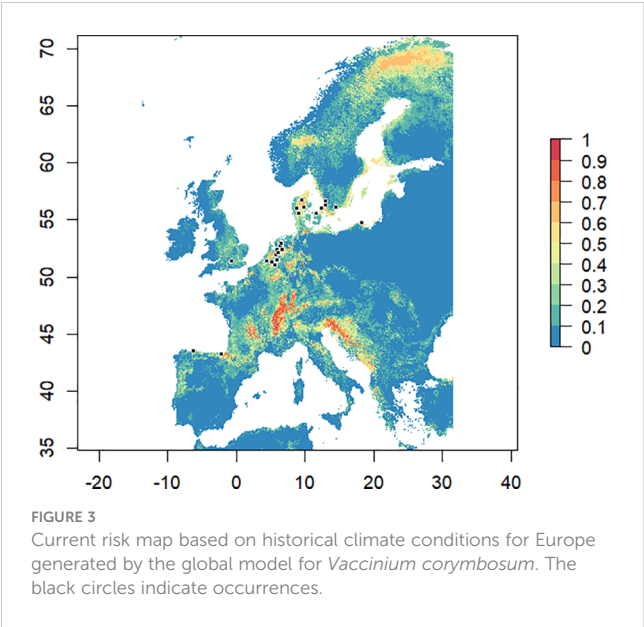


FIGURE 2

Global distribution of *V. corymbosum* occurrences (shown in green) used in the global model.



pseudoabsences. The results of the 10-fold cross-validation of these models and the mean of the predicted probabilities demonstrated consistently good performance, with model 6 having the best performance (Table 2). The Moran’s I of the residuals from model 6 was very low (- 0.007) indicating that the occurrences did not need thinning. To further test and evaluate the model, *Vaccinium corymbosum* occurrences located in Belgium from 2011-2021 were downloaded (n=111) from GBIF. We regard these data as independent, as they were not used in model building and date after model calibration (> 2010). This model correctly classified 90% of the occurrences as present (Supplementary S1). Model 6 was used for forecasting risk under the RCP scenarios and for the remaining steps in the workflow. The risk map for Europe generated from model 6 is shown in Figure 4.

The current risk map based on historical climate indicates that much of northern Belgium and the Ardennes region(located along the southeast border) is highly suitable for North American blueberry (Figure 5). Risk forecasts for RCP scenarios 2.6, 4.5,

and 8.5 suggest that environmental suitability for North American blueberry will greatly decrease in the future for Northern Belgium, but will remain for the Ardennes in RCPs 2.6 and 4.6 (Figures 5, 6).

Confidence in the predicted risk values is highest by area for the current risk map (Figures 7A, 8A). The majority of the predicted risk values under the RCP scenarios are of low confidence (< 0.4) (Figures 7B–D), with very few predicted risk values having high confidence (> 0.7) (Figures 8B–D).

The maximum temperature (°C) of the warmest month followed by the annual range in temperature (°C) had the highest overall variable importance (Table 3). The response curves indicate that the probability of occurrence decreases with increasing maximum temperature of the warmest month and that the species prefers habitats with both warm and cold seasons with yearly temperature differences of approximately 22°C (Figure 9).

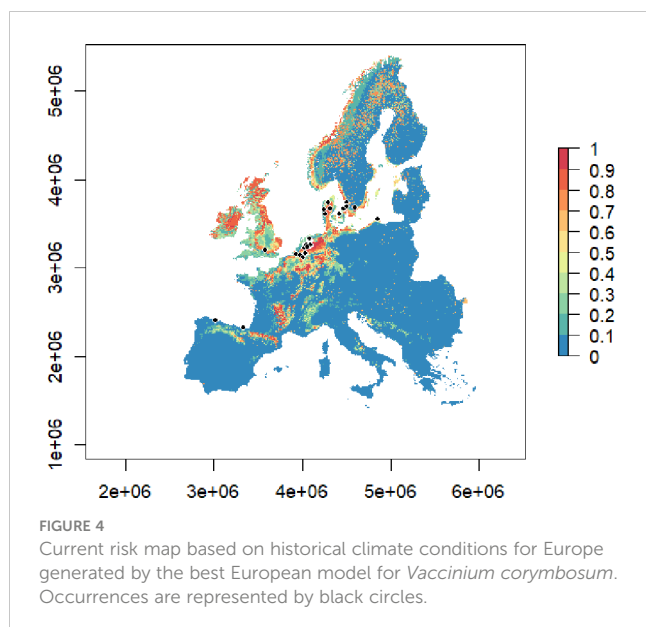
4 Discussion

WiSDM constitutes an open, reproducible, and flexible workflow for generating invasion risk forecasts for use in invasive species risk assessment and management. Our framework is ideally suited for agencies, consultants, or environmental planners where fast and easily updatable information on species invasion risk is needed, e.g., for answering to legal reporting requirements such as those mandated by the EU (1143/2014) regulation on invasive alien species or to identify areas where early-detection and rapid response measures preventing invader establishment should be prioritized. Uncertainties in the use of SDM outcomes can lead stakeholders to question the usefulness of invasion risk forecasts for conservation planning (Kujala et al., 2013).

Given the uncertainty associated with extrapolating risk to novel climates, WiSDM produces maps of confidence associated with each risk forecast, allowing identification of where and when model predictions are the most confident. Notably, the majority of predictions for the RCP scenarios have low confidence (Figure 7). For both the historical climate-based and RCP scenarios, the areas predicted as highly suitable for North American blueberry in

TABLE 2 Results of 4-fold cross-validation for European models.

model	threshold	sensitivity	specificity	Kappa	AUC
1	0.48	0.85	0.85	0.70	0.89
2	0.51	0.85	0.85	0.70	0.88
3	0.50	0.85	0.85	0.70	0.89
4	0.47	0.82	0.82	0.64	0.87
5	0.52	0.77	0.76	0.53	0.85
6	0.42	0.88	0.88	0.76	0.89
7	0.46	0.83	0.83	0.67	0.91
8	0.54	0.85	0.83	0.68	0.87
9	0.55	0.82	0.82	0.64	0.88
10	0.55	0.79	0.77	0.56	0.85



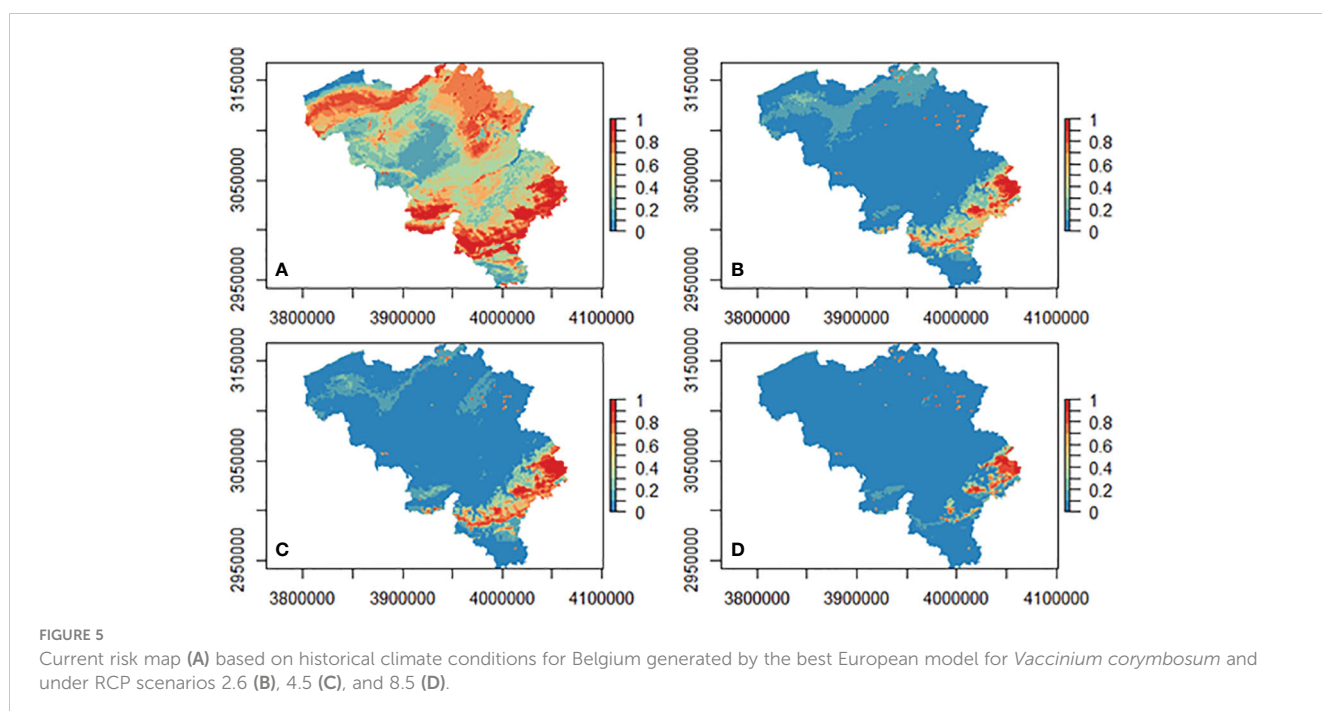
Belgium have high confidence and areas predicted to be absent or of low suitability have low confidence (Figures 5, 7). This suggests that in addition to monitoring high-risk areas, surveillance efforts should potentially also include “predicted to be absent but low confidence areas”, particularly if they overlap with protected areas or suspected dispersal pathways. Overall, the high uncertainty of the forecasts under the RCP scenarios observed in this study warrants future investigation to determine what steps, if any, can be taken to decrease it. For example, conformal prediction can be used to examine the impacts of variable selection or algorithm choice on model confidence in SDMs. Multivariate environmental similarity surface (MESS) maps (Elith et al., 2010) provide a spatially explicit

visualization of the correlation between different climate regimes or scenarios but leave the user to infer how robust their model is. Conformal prediction goes beyond mapping correlation by quantifying the confidence of predictions using a statistical framework with a guaranteed error rate (Vovk et al., 2005). Thus, the user can immediately assess the robustness of their model based on confidence rather than guessing based on climate (dis)similarity.

The European model predicted new areas (Ireland, northern UK, and the coast at risk of invasion as compared to the global model (Figures 2, 3) suggesting the existence of regional niches that would not be observed using only the global model. WiSDM uses the global model to decrease the likelihood of having false absences in the European model by not locating pseudoabsences in areas predicted as suitable by the global model. Furthermore, the European level model is constrained to regional climate and land use data which can help to uncover a regional niche (Gallien et al., 2012).

Response curves provided by WiSDM visualize the relationship between climate and habitat and invasion risk. They can be used to evaluate the ecological realism of the model forecasts as well as to help formulate optimal surveillance efforts in response to changing environmental conditions. For example, the response curve for the annual range of temperature and probability of North American blueberry occurrence shows that invasion is more probable as the difference between the coldest and warmest temperatures increases (Figure 9). This suggests that when annual climate extremes occur (i.e. an unusually cold winter and warm summer), additional monitoring is warranted (Johnstone, 1986).

It should be noted that an inherent limitation to correlative SDMs, including ours, is that the area at risk of invasion may be greater than what is predicted by the model due to the ability of the species to potentially occupy climates and regions that it does not currently inhabit. Failures to accurately predict the full invasive



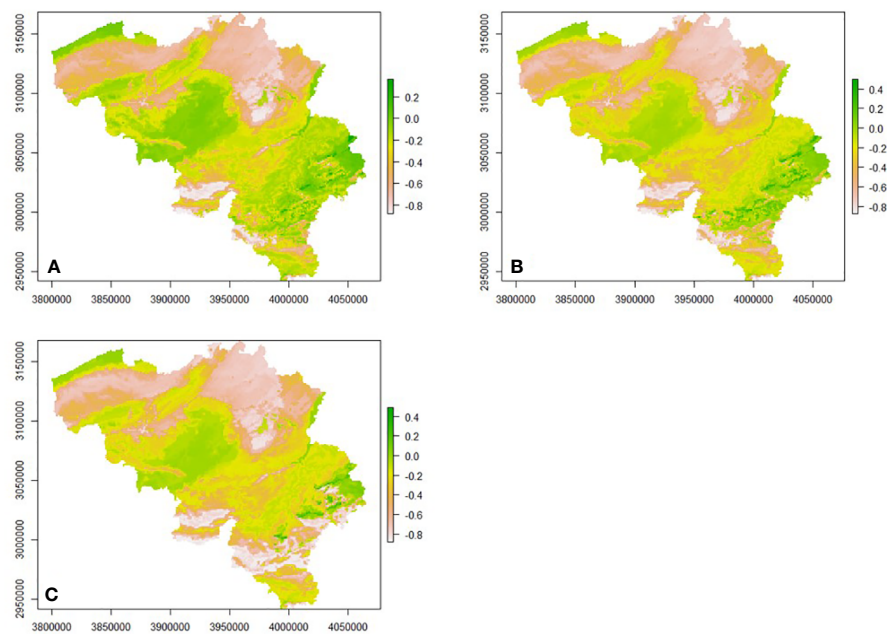


FIGURE 6

Difference maps illustrate the spatial difference between historical climate and RCP scenarios 2.6 (A), 4.5 (B), and 8.5 (C). Green areas indicate where the highest positive differences are observed, and beige and white areas indicate the highest negative differences.

distribution of introduced species are frequently attributed to the violation of a core assumption of SDMs: that the species being modeled is in equilibrium with the environment. The violation of this assumption can occur when the species realized niche is substantially different from its fundamental niche, or in the case of niche expansion when introduced species colonize 'novel' environments in their introduced range, which may not be

apparent during the early stages of invasion (Václavík and Meentemeyer, 2012). Apparent niche expansion can occur when eco-evolutionary changes (e.g., genetic adaptations) result in changes in species' fundamental niches, or because, for example, biotic interactions and dispersal limitations prevent species from occupying all suitable areas available to them across their native range. Characterizing species' fundamental niches is generally

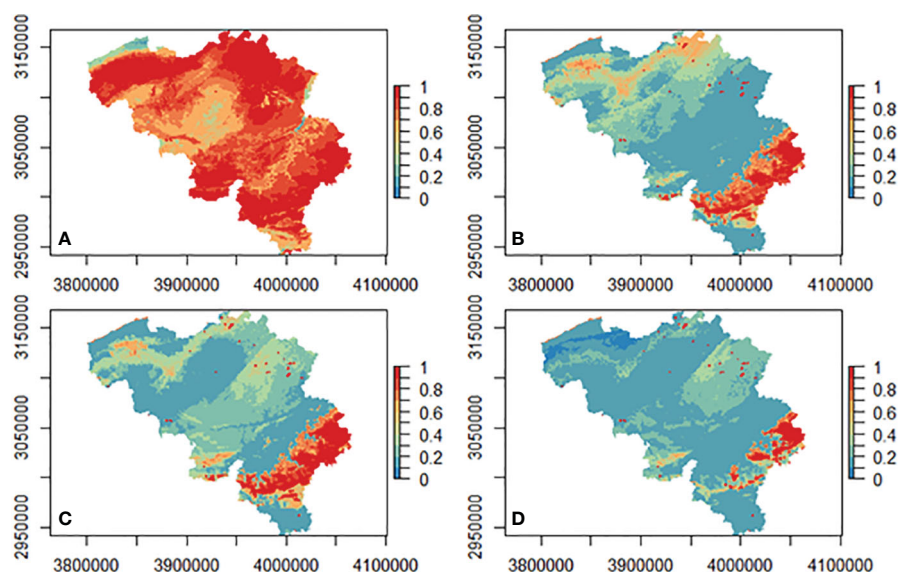
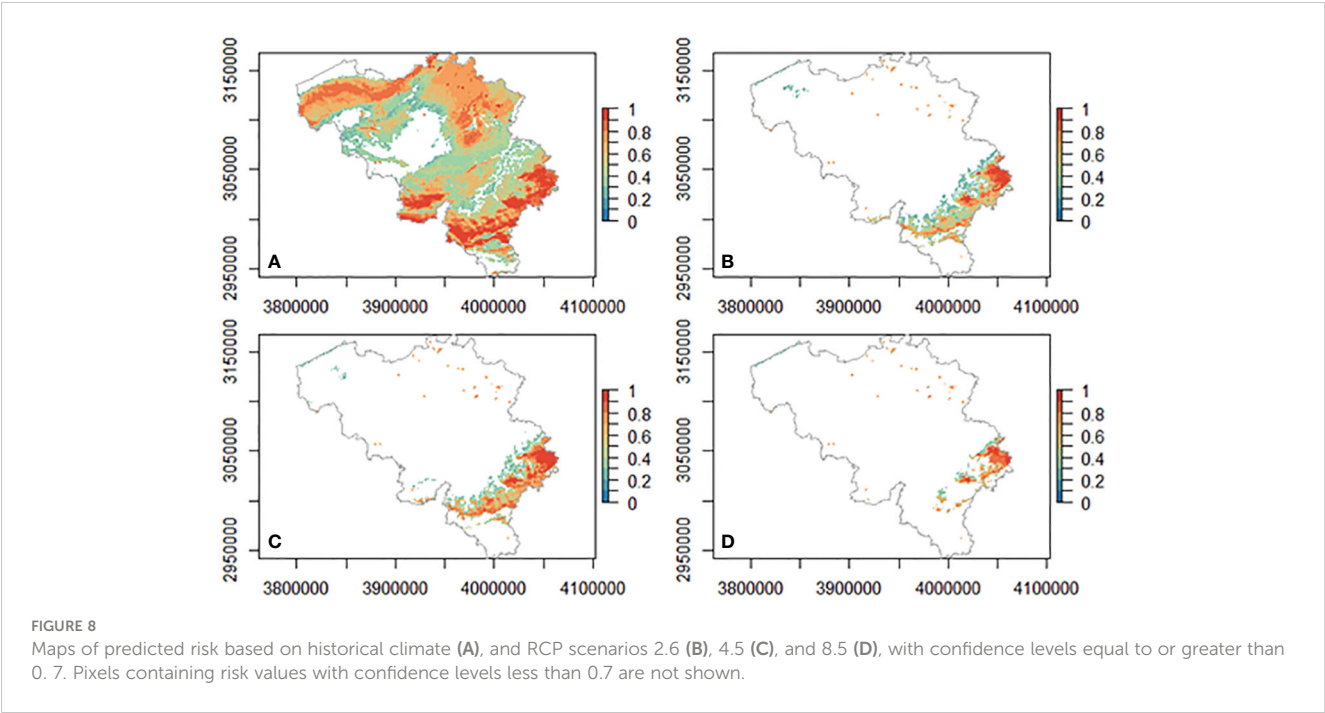


FIGURE 7

Confidence maps for the predicted distribution of *Vaccinium corymbosum* based on historical climate (A), and RCP scenarios 2.6 (B), 4.5 (C), and 8.5 (D), with confidence values ranging between 0 (no confidence) and 1 (maximum confidence). A value of 0 indicates that the prediction is completely nonconforming and not supported by previous data while 1 indicates the prediction is identical to a previous observation in the data.



considered impossible without information on ecophysiological tolerances. Still, there is an active debate about whether certain model settings or algorithms are better able to approximate fundamental niches – and thus species’ full potential distribution - using occurrence data only (Jiménez et al., 2019). In addition, missing ecological and/or anthropogenic predictors and gaps in occurrence data that span ecoregions or larger, can also lead to the under-prediction of the full distribution of a species. WiSDM is not set up as a bespoke ecological niche modeling framework to test hypotheses about the factors governing species distributions across native and invasive ranges and how to best model them. Instead, WiSDM produces data-driven SDMs, taking a pragmatic approach by combining GBIF occurrence data from both native and invasive ranges, into a single modeling framework.

The models underlying WiSDM do not account for dispersal, thus the maps produced by WiSDM indicate where a species can potentially colonize once introduced to a region. Furthermore, risk assessments for IAS are often conducted for species that are not yet (widely) present in a country or region thus quantifying the geographic area suitable for the species is an essential step in the risk assessment process. Until a consensus emerges on how potential distributions are best obtained using correlative SDMs

(e.g. algorithms, choice of background area, parameter settings) or until alternative (data demanding) process-based models (e.g. based on ecophysiological mechanisms and/or demography and dispersal) can be upscaled to apply to modeling large numbers of species, the WiSDM approach represents a robust and informed tool for use in invasive species risk assessment and management. Furthermore, the modeling workflow can easily be rerun when new occurrence data become available, e.g., through increased biodiversity monitoring, to potentially improve the prediction of the area at risk of invasion. Models can also be run using different baseline climate and habitat predictor layers. WiSDM currently defaults to using a 1976-2005 climate average for model training, which may lead to some uncertainty in estimating species occurrence - environment relationships especially for the most recent occurrence data (Milanesi et al., 2020). The amount of uncertainty introduced by our choice to use a ‘static’ baseline depends on the rate of change of the predictor variables over time and on how important individual predictor variables are for each species’ distribution (Bracken et al., 2022). While no consensus currently exists about how to best ensure optimal correspondence between available occurrence data and predictor variables (Steen et al., 2019), users may decide to use more detailed annual predictor

TABLE 3 Percent variable importance for each algorithm for the best European model (model 3).

	overall	GLM	GBM	RF	MARS
Percent wetland	1.3	8.6	0.0	0.0	0.3
Variation in annual precipitation (coefficient of variation)	28.2	53.2	36.0	32.1	0.0
Temperature annual range °C	32.3	0.0	32.1	27.0	63.1
Maximum temperature warmest month °C	38.2	38.2	31.9	40.9	36.6

The number corresponding to the most important variable is shown in bold for each algorithm.

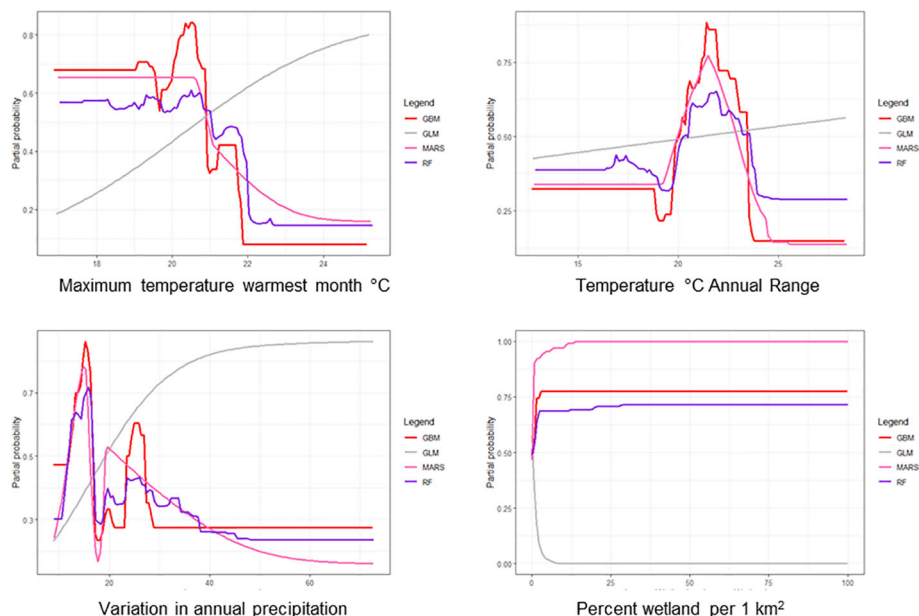


FIGURE 9
Response curves for each algorithm in the best model.

variables (e.g. such as the ERA5 and ERA5-Land time series), effectively turning WiSDM into a dynamic species distribution model (Abrahms et al., 2019).

The climate data currently used by WiSDM were generated using the RCP scenarios from the CMIP5 (Coupled Model Intercomparison Project Phase 5). The RCP scenarios have since been updated with the new SSP (Shared Socioeconomic Pathway) based scenarios from CMIP6 (Coupled Model Intercomparison Project Phase 6). The updated scenarios in CMIP6 that correspond to RCP2.6, RCP4.5, and RCP8.5 from CMIP5 are called SSP1-2.6, SSP2-4.5, and SSP5-8.5, respectively. The SSP scenarios result in similar 2100 radiative forcing levels used by their RCP counterparts, but use different assumptions and improved models with more recent emissions data (Tebaldi et al., 2021). In contrast to RCP scenarios, the SSP scenarios provide economic and social reasons for the assumed emission pathways and land use changes. The SSP scenarios start with emissions data from 2014 (the RCPs start with data from 2007), thus the scenarios start with a higher emissions level and also show a slower decline. When interpreting the results from SDMs using either RCP or SSP scenarios, it is important to consider the assumptions used such as the expected levels of greenhouse gases, population growth, and mitigation as these in addition to the climate models used can influence the results and introduce uncertainty into the projections (Thuiller et al., 2019).

Open, transparent, data-driven risk assessments, with clear indications of uncertainties, foster credibility, which is vital for acceptance by stakeholders and uptake by policy-makers (McGeoch et al., 2012; Groom et al., 2019; Sofaer et al., 2019). The WiSDM approach fits well with recent trends towards transparency and repeatability in ecological forecasting, such as encapsulated in the 'best practice standards' for SDM model development (e.g. Araújo

et al., 2019; Zurell et al., 2020). WiSDM further promotes the uptake of SDM modeling into policy and conservation actions by its adoption of the FAIR principles of 'Findability, Accessibility, Interoperability, and Reuse' by making the workflow freely available on GitHub and publishing all data layers needed to run the workflow on Zenodo. The flexible nature of WiSDM also makes it possible for users to customize our code to match the specific demands of the assessment under consideration (e.g. use of alternative climate scenarios and habitat predictors, or model algorithms and settings). The customized settings used are automatically recorded in an R markdown document that can be shared to ensure transparency and reproducibility. Thus, the reproducible workflow presented here maximizes the usefulness of available open data and provides a structured framework for obtaining and interpreting forecasts of the invasion risk of introduced species.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://zenodo.org/communities/trias/?page=1&size=20>.

Author contributions

DS and AD conceptualized the research. All authors contributed to the development of the modelling workflow, AD, DS and QG wrote the manuscript with input from all authors. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Belgian Science Policy Office under the TrIAS project (BR/165/A1/TrIAS).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Abrahams, B., Welch, H., Brodie, S., Jacox, M. G., Becker, E. A., Bograd, S. J., et al. (2019). Dynamic ensemble models to predict distributions and anthropogenic risk exposure for highly mobile species. *Divers. Distrib.* 25, 1182–1193. doi: 10.1111/ddi.12940
- Adriaens, T., Van Daele, T., Groom, Q., Vanderhoeven, S., Davis, A. J., Strubbe, D., et al. (2022). “Automated early warning: a pipeline for feeding headline indicators on the state of invasions and to prioritize emerging alien species. In Biological Invasions in a Changing World. Book of Abstracts,” in *Neobiota 2022–12th International Conference on Biological Invasions*, Tartu, Estonia, 12–16 September 2022. 34.
- Adriaens, T., Van Valkenburg, J., Verloove, F., and Groom, Q. (2019). Trosbosbes, probleemsoort in wording? *Natuur. Focus*. 2019 (2), 75–76.
- Alvarsson, J., McShane, S. A., Norinder, U., and Spjuth, O. (2021). Predicting with confidence: using conformal prediction in drug discovery. *J. Pharm. Sci.* 110 (1), 42–49. doi: 10.1016/j.xphs.2020.09.055
- Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., et al. (2019). Standards for distribution models in biodiversity assessments. *Sci. Adv.* 5, eaat4858. doi: 10.1126/sciadv.aat4858
- Baker, D. J., Maclean, I. M. D., Goodall, M., and Gaston, K. J. (2021). Species distribution modelling is needed to support ecological impact assessments. *J. Appl. Ecol.* 58, 21–26. doi: 10.1111/1365-2664.13782
- Barbet-Massin, M., Rome, Q., Villemant, C., and Courchamp, F. (2018). Can species distribution models really predict the expansion of invasive species? *PLoS One* 13 (3), e0193085. doi: 10.1371/journal.pone.0193085
- Barbet-Massin, M., Jiguet, F., Albert, C. H., and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.* 3, 327–338. doi: 10.1111/j.2041-210X.2011.00172.x
- Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., et al. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecol. Model.* 222 (11), 1810–1819. doi: 10.1016/j.ecolmodel.2011.02.011
- Beck, J., Holloway, J. D., and Schwanghart, W. (2013). Undersampling and the measurement of beta diversity. *Methods Ecol. Evol.* 4, 370–382. doi: 10.1111/2041-210X.12023
- Bellard, C., Jeschke, J. M., Leroy, B., and Mace, G. M. (2018). Insights from modeling studies on how climate change affects invasive alien species geography. *Ecol. Evol.* 8, 5688–5700. doi: 10.1002/ece3.4098
- Boria, R. A., Olson, L. E., Goodman, S. M., and Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecol. Modell.* 275, 73–77. doi: 10.1016/j.ecolmodel.2013.12.012
- Bracken, J. T., Davis, A. Y., O'Donnell, K. M., Barichivich, W. J., Walls, S. C., and Jezkova, T. (2022). Maximizing species distribution model performance when using historical occurrences and variables of varying persistency. *Ecosphere* 13 (3), e3951. doi: 10.1002/ecs2.3951
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Brodie, S., Smith, J. A., Muhling, B. A., Barnett, L. A. K., Carroll, G., Fiedler, P., et al. (2022). Recommendations for quantifying and reducing uncertainty in climate projections of species distributions. *Global Change Biol.* 28, 6586–6601. doi: 10.1111/gcb.16371
- Brown, J. L., and Carnaval, A. C. (2019). A tale of two niches: methods, concepts and evolution. *Front. Biogeogr.* 11, e44158. doi: 10.21425/F5FBG44158
- Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wüest, R. O., Wang, Z., et al. (2020). Model complexity affects species distribution projections under climate change. *J. Biogeogr.* 47, 130–142. doi: 10.1111/jbi.13734
- Chapman, D., Pescott, O. L., Roy, H. E., and Tanner, R. (2019). Improving species distribution models for invasive non-native species with biologically informed pseudo-absence selection. *J. Biogeogr.* 46 (5), 1029–1040. doi: 10.1111/jbi.13555
- Chauvier, Y., Descombes, P., Guéguen, M., Boulangeat, L., Thuiller, W., and Zimmermann, N. E. (2022). Resolution in species distribution models shapes spatial patterns of plant multifaceted diversity. *Ecography* 2022, e05973. doi: 10.1111/ecog.05973
- Cox, D. R. (1958). The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B (Methodological)* 20 (2), 215–232. doi: 10.1111/j.2517-6161.1958.tb00292.x
- Davis, A., Strubbe, D., and Groom, Q. (2023). Global taxonomic occurrence grids using GBIF data for species distribution models. (1.0.0) [Data set]. *Zenodo*. doi: 10.5281/zenodo.7556851
- Davis, A. J. S., Thill, J.-C., and Meentemeyer, R. K. (2017). Multi-temporal trajectories of landscape change explain forest biodiversity in urbanizing ecosystems. *Landscape Ecol.* 32, 1789–1803. doi: 10.1007/s10980-017-0541-8
- De Troch, R., Termonia, P., and Van Schaybroeck, B. (2020). High-resolution future climate data for species distribution models in Europe [Data set]. *Zenodo*. doi: 10.5281/zenodo.3694065
- Diniz-Filho, J. A. F., and Bini, L. M. (2005). Modelling geographical patterns in species richness using eigenvector-based spatial filters. *Global Ecol. Biogeogr.* 14 (2), 177–185. doi: 10.1111/j.1466-822X.2005.00147.x
- Elith, J., Kearney, M., and Phillips, S. (2010). The art of modelling range-shifting species. *Methods Ecol. Evol.* 1, 330–342. doi: 10.1111/j.2041-210X.2010.00036.x
- Elith, J. H., Graham, C. P., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., et al. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29, 129–151. doi: 10.1111/j.2006.0906-7590.04596.x
- European Environment Agency. (2011). *EEA reference grid for Europe (1km)*. Available at: <https://sdi.eea.europa.eu/catalogue/srv/api/records/d9d4684e-0a8d-496c-8be8-110f4b9465f6>.
- Feng, X., Park, D. S., Walker, C., Peterson, T., Merow, C., and Papeš, M. (2019a). A checklist for maximizing reproducibility of ecological niche models. *Nat. Ecol. Evol.* 3, 1382–1395. doi: 10.1038/s41559-019-0972-5
- Feng, X., Park, D. S., Liang, Y., Pandey, R., and Papeš, M. (2019b). Collinearity in ecological niche modeling: Confusions and challenges. *Nat. Ecol. Evol.* 3, 1382–1395. doi: 10.1002/ece3.5555
- Ferraz, K. M. P. M., Morato, R. G., Bovo, A. A. A., da Costa, C. O. R., Ribeiro, Y. G. G., de Paula, R. C., et al. (2021). Bridging the gap between researchers, conservation planners, and decision makers to improve species conservation decision-making. *Conserv. Sci. Pract.* 3, e330. doi: 10.1111/csp2.330
- Fourcade, Y. (2021). Fine-tuning niche models matters in invasion ecology. A lesson from the land planarian *Obama nungara*. *Ecol. Model.* 457, 109686. doi: 10.1016/j.ecolmodel.2021.109686
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Stat.* 19 (1), 1–67. doi: 10.1214/aos/1176347963
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29 (5), 1189–1232. doi: 10.1214/aos/1013203451

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fevo.2024.1148895/full#supplementary-material>

- Gallien, L., Douzet, R., Pratte, S., Zimmermann, N. E., and Thuiller, W. (2012). Invasive species distribution models – how violating the equilibrium assumption can create new insights. *Glob. Ecol. Biogeogr.* 21, 1126–1136. doi: 10.1111/j.1466-8238.2012.00768.x
- GBIF. (2023). *About species counts in GBIF* Copenhagen, Global Biodiversity Information Facility. Available at: <https://www.gbif.org/about-species-counts>.
- GBIF Secretariat. (2022a). *GBIF Backbone Taxonomy*. Copenhagen, Global Biodiversity Information Facility. doi: 10.15468/39omei
- GBIF Secretariat. (2022b). *GBIF Work Programme 2022: Annual Update to Implementation Plan 2017–2022* Copenhagen, Global Biodiversity Information Facility. doi: 10.35035/doc-jjrz-b144
- González-Moreno, P., Lazzaro, L., Vilà, M., Preda, C., Adriaens, T., Bacher, S., et al. (2019). Consistency of impact assessment protocols for non-native species. *NeoBiota* 44, 1–25. doi: 10.3897/neobiota.44.31650
- Gotelli, N. J., and Colwell, R. K. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4 (4), 379–391. doi: 10.1046/j.1461-0248.2001.00230.x
- Groom, Q., Strubbe, D., Adriaens, T., Davis, A. J. S., Desmet, P., Oldoni, D., et al. (2019). Empowering citizens to inform decision-making as a way forward to support invasive alien species policy. *Citizen Science: Theory Pract.* 4 (1), 33. doi: 10.5334/cstp.238
- Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., and Kueffer, C. (2014). Unifying niche shift studies: insights from biological invasions. *Trends Ecol. Evol.* 29, 260–269. doi: 10.1016/j.tree.2014.02.009
- Guisan, A., and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8, 993–1009. doi: 10.1111/j.1461-0248.2005.00792.x
- Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y., and Mackey, B. (2019). Species distribution models can be highly sensitive to algorithm configuration. *Ecol. Model.* 408, 108719. doi: 10.1016/j.ecolmodel.2019.108719
- Hao, T., Elith, J., Guillerá-Arroita, G., and Lahoz-Monfort, J. J. A. (2019). review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. *Divers. Distrib.* 25, 839–852. doi: 10.1111/ddi.12892
- Jeschke, J. M., and Strayer, D. L. (2008). Usefulness of bioclimatic models for studying climate change and invasive species. *Ann. N.Y. Acad. Sci.* 1134, 1–24. doi: 10.1196/annals.1439.002
- Jiménez, L., Soberón, J., Christen, J. A., and Soto, D. (2019). On the problem of modeling a fundamental niche from occurrence data. *Ecol. Model.* 397, 74–83. doi: 10.1016/j.ecolmodel.2019.01.020
- Johnstone, I. M. (1986). Plant invasion windows: a time-based classification of invasion potential. *Biol. Rev.* 61 (4), 369–394. doi: 10.1111/j.1469-185X.1986.tb00659.x
- Karger, D. N., Conrad, O., Böhrer, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., et al. (2017). Climatologies at high resolution for the Earth land surface areas. *Sci. Data.* 4, 170122. doi: 10.1038/sdata.2017.122
- Kass, J. M., Vilela, B., Aiello-Lammens, M. E., Muscarella, R., Merow, C., and Anderson, R. P. (2018). Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods Ecol. Evol.* 9, 1151–1156. doi: 10.1111/2041-210X.12945
- Kotlarski, S., Keuler, K., Christensen, O. B., Colette, A., Déqué, M., Gobiet, A., et al. (2014). Regional climate modeling on European scales: a joint standard evaluation of the EURO-CORDEX RCM ensemble. *Geosci. Model. Dev.* 7 (4), 1297–1333. doi: 10.5194/gmd-7-1297-2014
- Kuhn, M. (2008). Building predictive models in R using the caret package. *J. Stat. softw.* 28, 1–26. doi: 10.18637/jss.v028.i05
- Kujala, H., Moilanen, A., Araújo, M. B., and Cabeza, M. (2013). Conservation planning with uncertain climate change projections. *PLoS One* 8, e53315. doi: 10.1371/journal.pone.0053315
- Lee-Yaw, A. J., McCune, L. J., Pironon, S., and Sheth, N. S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography* 2022, e05877. doi: 10.1111/ecog.05877
- Liu, C., Berry, P. M., Dawson, T. P., and Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28 (3), 385–393. doi: 10.1111/j.0906-7590.2005.03957.x
- Liu, C., Wolter, C., Xian, W., and Jeschke, J. M. (2020). Species distribution models have limited spatial transferability for invasive species. *Ecol. Lett.* 23, 1682–1692. doi: 10.1111/ele.13577
- Lobo, J. M. (2008). Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodivers. Conserv.* 17 (4), 873–881. doi: 10.1007/s10531-008-9333-4
- Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr.* 17 (2), 145–151. doi: 10.1111/j.1466-8238.2007.00358.x
- Mazor, T., Doropoulos, C., Schwarzmüller, F., Gladish, D. W., Kumaran, N., Merkel, K., et al. (2018). Global mismatch of policy and research on drivers of biodiversity loss. *Nat. Ecol. Evol.* 2, 1071–1074. doi: 10.1038/s41559-018-0563-x
- McGeoch, M. A., Spear, D., Kleynhans, E. J., and Marais, E. (2012). Uncertainty in invasive alien species listing. *Ecol. Appl.* 22, 959–971. doi: 10.1890/11-1252.1
- Milanesi, P., Mori, E., and Menchetti, M. (2020). Observer-oriented approach improves species distribution models from citizen science data. *Ecol. Evol.* 10 (21), 12104–12114. doi: 10.1002/ece3.6832
- Mostert, P., Björkås, R., Bruls, A. J. H. M., Koch, W., Martin, E. C., and Perrin, S. W. (2023). intSDM: an R package for building a reproducible workflow for the field of integrated species distribution models. *bioRxiv*, 2022.09.15.507996. doi: 10.1101/2022.09.15.507996
- Muscatello, A., Elith, J., and Kujala, H. (2021). How decisions about fitting species distribution models affect conservation outcomes. *Conserv. Biol.* 35, 1309–1320. doi: 10.1111/cobi.13669
- Nguyen, D., and Leung, B. (2022). How well do species distribution models predict occurrences in exotic ranges? *Global Ecol. Biogeogr.* 31, 1051–1065. doi: 10.1111/geb.13482
- Norinder, U., Carlsson, L., Boyer, S., and Eklund, M. (2014). Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.* 54 (6), 1596–1603. doi: 10.1021/ci5001168
- Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., et al. (2001). Terrestrial ecoregions of the world: a new map of life on Earth. *Bioscience* 51 (11), 933–938. doi: 10.1641/0006-3568(2001)051[0933:TEOTWA]2.0.CO;2
- Pearson, R. G., Raxworthy, C. J., Nakamura, M., and Townsend Peterson, A. (2007). Predicting species distributions from small numbers of occurrence records: A test case using cryptic geckos in Madagascar. *J. Biogeogr.* 34, 102–117. doi: 10.1111/j.1365-2699.2006.01594.x
- Pearson, R. G., and Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Glob. Ecol. Biogeogr.* 12, 361–371. doi: 10.1046/j.1466-822X.2003.00042.x
- Penninkhof, J., Boosten, M., and de Groot, C. (2018). Effect bestrijding trosbosbes in de Pelen. Resultaten van de monitoring in de periode 2015–2017. *Probos Wageningen*. 41.
- Pereira, T., Cardoso, S., Silva, D., Mendonça, A. D., Guerreiro, M., and Madeira, S. C. (2017). Towards trustworthy predictions of conversion from mild cognitive impairment to dementia: a conformal prediction approach, in *International Conference on Practical Applications of Computational Biology & Bioinformatics*, eds. F. Fdez-Riverola, M. S. Mohamad, M. Rocha, J. F. De Paz and T. Pinto (Cham: Springer International Publishing, 155–163. doi: 10.1007/978-3-319-60816-7_19
- Peterson, A. T., Cobos, M. E., and Jiménez-García, D. (2018). Major challenges for correlational ecological niche model projections to future climate conditions. *Ann. N.Y. Acad. Sci.* 1429, 66–77. doi: 10.1111/nyas.13873
- Petitpierre, B., Broennimann, O., Kueffer, C., Daehler, C., and Guisan, A. (2017). Selecting predictors to maximize the transferability of species distribution models: lessons from cross-continental plant invasions. *Global Ecol. Biogeogr.* 26, 275–287. doi: 10.1111/geb.12530
- Phillips, S. J., Dudik, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., et al. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197. doi: 10.1890/07-2153.1
- Reyserhove, L., Desmet, P., Oldoni, D., Adriaens, T., Strubbe, D., Davis, A. J. S., et al. (2020). A checklist recipe: making species data open and FAIR. *Database* 2020, baaa084. doi: 10.1093/database/baaa084
- Roy, H., Rabitsch, W., Scalera, R., Stewart, A., Gallardo, B., Genovesi, P., et al. (2017). Developing a framework of minimum standards for the risk assessment of alien species. *J. Appl. Ecol.* 55, 526–538. doi: 10.1111/1365-2664.13025
- Schepker, J., and Kowarik, I. (1998). “Invasive North American blueberry hybrids *Vaccinium corymbosum* x *angustifolium* in Northern Germany,” in *Plant invasions. Ecology and human response*. Eds. U. Starfinger, K. Edwards, I. Kowarik and M. Williamson (Leiden: Backhuys Publisher).
- Schwartz, M. W., Cook, C. N., Pressey, R. L., Pullin, A. S., Runge, M. C., Salafsky, N., et al. (2018). Decision support frameworks and tools for conservation. *Conserv. Lett.* 11, e12385. doi: 10.1111/conl.12385
- Shafer, G., and Vovk, V. (2008). A tutorial on conformal prediction. *J. Mach. Learn. Res.* 9 (3), 371–421.
- Sillero, N., Campos, J. C., Arenas-Castro, S., and Barbosa, A. M. (2023). A curated list of R packages for ecological niche modelling. *Ecol. Model.* 476, 110242. doi: 10.1016/j.ecolmodel.2022.110242
- Simberloff, D., Martin, J. L., Genovesi, P., Maris, V., Wardle, D. A., Aronson, J., et al. (2013). Impacts of biological invasions: what’s what and the way forward. *Trends Ecol. Evol.* 28, 58–66. doi: 10.1016/j.tree.2012.07.013
- Sofaer, H. R., Jarnevich, C. S., Pearse, I. S., Smyth, R. L., Auer, S., Cook, G. L., et al. (2019). Development and delivery of species distribution models to inform decision-making. *BioScience* 69 (7), 544–557. doi: 10.1093/biosci/biz045
- Srivastava, V., Lafond, V., and Griess, V. C. (2019). Species distribution models (SDM): applications, benefits and challenges in invasive species management. *CABI Rev.* 2019, 1–13. doi: 10.1079/PAVSNNR201914020
- Steen, V. A., Elphick, C. S., and Tingley, M. W. (2019). An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity Distrib.* 25 (12), 1857–1869. doi: 10.1111/ddi.12985
- Tebaldi, C., Debeire, K., Eyring, V., Fischer, E., Fyfe, J., Friedlingstein, P., et al. (2021). Climate model projections from the scenario model intercomparison project (ScenarioMIP) of CMIP6. *Earth Syst. Dyn.* 12, 253–293. doi: 10.5194/esd-12-253-2021
- Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., and Zimmermann, N. E. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nat. Commun.* 10 (1), 1446. doi: 10.1038/s41467-019-09519-w

- Urban, M. C. (2015). Accelerating extinction risk from climate change. *Science* 348, 571–573. doi: 10.1126/science.aaa4984
- Václavík, T., and Meentemeyer, R. K. (2012). Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. *Diversity Distrib.* 18, 73–83. doi: 10.1111/j.1472-4642.2011.00854.x
- Vanderhoeven, S., Adriaens, T., Desmet, P., Strubbe, D., Backeljau, T., Barbier, Y., et al. (2017). Tracking Invasive Alien Species (TrIAS): Building a data-driven framework to inform policy. *Res. Ideas Outcomes* 3, e13414. doi: 10.3897/rio.3.e13414
- Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* 6 (1). doi: 10.2202/1544-6115.1309
- van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., and Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography* 39, 542–552. doi: 10.1111/ecog.01509
- Venter, O., Sanderson, E., Magrath, A., Allan, J. R., Beher, J., Jones, K. R., et al. (2016). Global terrestrial Human Footprint maps for 1993 and 2009. *Sci. Data* 3, 160067. doi: 10.1038/sdata.2016.67
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world* Vol. 29 (New York: Springer).
- Waller, J., Volik, N., Mendez, F., and Hahn, A. (2021). GBIF data processing and validation. *Biodivers. Inf. Sci. Standards* 5, e75686. doi: 10.3897/biss.5.75686
- Wenger, S. J., Som, N. A., Dauwalter, D. C., Isaak, D. J., Neville, H. M., Luce, C. H., et al. (2013). Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Change Biol.* 19 (11), 3343–3354. doi: 10.1111/gcb.12294
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3 (1), 1–9. doi: 10.1038/sdata.2016.18
- Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., et al. (2018). Outstanding challenges in the transferability of ecological models. *Trends Ecol. Evol.* 33 (10), 790–802. doi: 10.1016/j.tree.2018.08.001
- Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., et al. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods Ecol. Evol.* 10, 744–751. doi: 10.1111/2041-210X.13152
- Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., et al. (2020). A standard protocol for reporting species distribution models. *Ecography* 43, 1261–1277. doi: 10.1111/ecog.04960

Frontiers in Ecology and Evolution

Ecological and evolutionary research into our natural and anthropogenic world

This multidisciplinary journal covers the spectrum of ecological and evolutionary inquiry. It provides insights into our natural and anthropogenic world, and how it can best be managed.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



Frontiers in Ecology and Evolution

