The application of machine learning methods to satellite data for the management of invasive water hyacinth.

A thesis submitted to the Faculty of Science, The University of the Witwatersrand, Johannesburg in fulfilment of the requirements for the degree, Doctor of Philosophy

by

Geethen Singh

942909

and

Supervised by: Dr Chevonne Reynolds, Prof. Marcus Byrne and Prof. Benjamin Rosman

Acknowledgements

This research was funded by the University of the Witwatersrand and by the DST/NRF centre of excellence at the FitzPatrick Institute of African Ornithology at the University of Cape Town, grant number UID: 40470.

I would like to thank, the Centre for Biological Control (CBC), Rhodes University, South Africa for providing field and in-situ data that formed an essential component for both Chapter 2 and Chapter 4. Special thanks to Prof. Julie Coetzee for being willing to share the data, Ms. Esther Mostert who curated the locality data, and Mr. Benjamin Miller who shared the biocontrol release data and provided information on the current release strategies followed in South Africa.

I would like to thank my three supervisors for inspiring me to be a better academic while allowing me the freedom to pursue research questions that I felt most excited about and found the most meaningful. All of you have guided me along the path to fulfilling this PhD and in so doing have made the life of my family and I better. I will forever be grateful for you all seeing my potential when I sometimes could not. Marcus, after many of our work-related conversations I always felt motivated to push myself to do work that I sometimes wrote off as being unnecessary or too difficult. You have also shared many nuggets of life advice that I often recall when making decisions. Chevonne, you have selflessly supported your students, including me. I hope that I can have that attitude towards future colleagues and students. One day I will write concisely - today is probably not that day. Benji, I am still trying to figure out how you get so much done. Despite having so much going on, you were still able to provide the guidance I needed (I needed a lot).

I would also like to share my gratitude with Geraud Nangue Tasse and Shahil Mawjee for patiently helping me to fix code, inspiring me to improve my python programming skills, and troubleshoot model failures. Joseph White for sharing your wisdom, much needed perspectives, friendship and sharing in my excitement to speak about geospatial methods and machine learning. I am grateful for all our insightful conversations throughout our multiple trips to Kruger National Park. David Kinsler for always being available to chat about the latest advancements in remote sensing, and bounce ideas off, that helped shape the methods used in this work. Nikhail Arumoogum for the insightful discussions and conversations that served to remind me to keep the ecology in mind, or when not applicable, just keep my mind. Craig Mahlasi for the conversations and suggestions that helped me interpret the results of Chapter 2. Yashvir Bunwarie, Lennon Naidoo, and Jared Nadasen for the emotional support, friendship and for all the times they have been a part of the very important ‘winding back’ sessions.

I am also appreciative of the core GEE developers and the GEE community for the detailed documentation, examples, compute, storage and help forums. Moreover, I am thankful to the many individuals that have put time and effort into developing and maintaining open source and/or open-access machine learning and geospatial software. I hope to contribute to the community soon.

Finally, I would like to thank my brother (Reethen), sister (Sandhika), and now late mother (Mandhika) for the emotional support and for picking up the day-to-day chores when I needed to focus on work.

Disclaimer

The introduced tools should not be used in isolation as actual conditions may be different from the model outcomes and require additional verification.

University of the Witwatersrand

Animal, Plant and Environmental Sciences

Declaration Form for PhD Thesis

I, Geethen Singh, hereby declare that the work presented in this PhD thesis, titled “The application of machine learning methods to satellite data for the management of invasive water hyacinth.", is entirely my own effort and has been conducted under the supervision of Dr Chevonne Reynolds, Prof. Benjamin Rosman and Prof. Marcus Byrne.

I confirm that:

1. This thesis represents my original research work and has not been submitted, in whole or in part, for the award of any other degree or diploma at any university or institution.

2. Any sources of information, including published or unpublished works of others, have been duly acknowledged and referenced within the thesis.

3. The content of this thesis does not infringe upon the intellectual property rights of any individual or entity.

4. All ethical guidelines and regulations concerning research involving human subjects, animals, or any other form of data collection have been strictly adhered to.

5. The data, results, and conclusions presented in this thesis are accurate and unbiased to the best of my knowledge and abilities.

6. I have not received any unauthorized assistance or unfair advantage in the completion of this thesis.

7. Any contributions from co-authors or collaborators have been appropriately acknowledged and their respective roles in the research have been disclosed.

8. I understand that any violation of academic integrity or dishonesty in the completion of this thesis may result in severe consequences, including the revocation of my degree.

I hereby grant University of the Witwatersrand the non-exclusive right to reproduce and distribute copies of this thesis for scholarly purposes.

Signed: G.S. Date: 13/06/2023

Summary

Biological invasions are responsible for some of the most devastating impacts on the world's ecosystems, with freshwater ecosystems among the worst affected. Invasions threaten not only freshwater biodiversity, but also the provision of ecosystem services. Tackling the impact of invasive aquatic alien plant (IAAP) species in freshwater systems is an ongoing challenge. In the case of water hyacinth (*Pontederia crassipes*, previously *Eichhorniae crassipes*), the worst IAAP presents a long-standing management challenge that requires detailed and frequently updated information on its distribution, the context that influences its occurrence, and a systematic way to identify effective biocontrol release events. This is particularly urgent in South Africa, where freshwater resources are scarce and under increasing pressure.

This research employs recent advances in machine learning (ML), remote sensing, and cloud computing to improve the chances of successful water hyacinth management. This is achieved by (i) mapping the occurrence of water hyacinth across a large extent, (ii) identifying the factors that are likely driving the occurrence of the weed at multiple scales, from a waterbody level to a national extent, and (iii) finally identifying periods for effective biocontrol release. Consequently, the capacity of these tools demonstrates their potential to facilitate wide-scale, consistent, automated, pre-emptive, data-driven, and evidence-based decision making for managing water hyacinth.

The first chapter is a general introduction to the research problem and research questions. In the second chapter, the research combines a novel image thresholding method for water detection with an unsupervised method for aquatic vegetation detection and a supervised random forest model in a hierarchical way to localise and discriminate water hyacinth from other IAAP's at a national extent. The value of this work is marked by the comparison of the user (87%) and producer accuracy (93%) of the introduced method with previous small-scale studies. As part of this chapter, the results also show the sensor-agnostic and temporally consistent capability of the introduced hierarchical approach to monitor water and aquatic vegetation using Sentinel-2 and Landsat-8 for long periods (from 2013 - present). Lastly, this work demonstrates encouraging results when using a Deep Neural Network (DNN) to directly detect aquatic vegetation and circumvents the need for accurate water extent data. The two chapters that follow (Chapter 3 and 4 described below) introduce an application each that build off the South African water hyacinth distribution and aquatic vegetation time series (derived in Chapter 2).

The third chapter uses a species distribution model (SDM) that links climatic, socio-economic, ecological, and hydrological conditions to the presence/absence of water hyacinth throughout South Africa at a waterbody level. Thereafter, explainable AI (xAI) methods (specifically

SHapley Additive exPlanations or SHAP) are applied to better understand the factors that are likely driving the occurrence of water hyacinth. The analyses of 82 variables (of 140 considered) show that the most common group of drivers primarily associated with the occurrence of water hyacinth in South Africa are climatically related (41.4%). This is followed by natural land cover categories (32.9%) and socio-economic variables (10.7%), which include artificial land-cover. The two least influential groups are hydrological variables (10.4%) including water seasonality, runoff, and flood risk, and ecological variables (4.7%) including riparian soil conditions and interspecies competition. These results suggest the importance of considering landscape context when prioritising the type (mechanical, biological, chemical, or integrated) of weed management to use.

To enable the prioritisation of suitable biocontrol release dates, the fourth chapter forecasts 70-day open water proportion post-release as a reward for effective biocontrol. This enabled the simulation of the effect of synthetic biocontrol release events under a multiarmed bandit framework for the identification of two effective biocontrol release periods (late spring/early summer (mid-November) and late summer (late February to mid-March)). The latter release period was estimated to result in an 8-27% higher average open-water cover post-release compared to actual biocontrol release events during the study period (May 2018 - July 2020)., Hartbeespoort Dam, South Africa, is considered as a case study for improving the pre-existing management strategy used during the biocontrol of water hyacinth.

The novel frameworks introduced in this work go a long way in advancing IAAP species management in the age of both ongoing drives towards the adoption of artificial intelligence and sustainability for a better future. It goes beyond (i) traditional small-scale and infrequent mapping, (ii) standard SDMs, to now include the benefits of spatially explicit model explainability, and (iii) introduces a semi-automated and widely applicable method to explore potential biocontrol release events. The direct benefit of this work, or indirect benefits from derivative work outweighs both the low production costs or equivalent field and lab work.

To improve the adoption of modern ML and Earth Observation (EO) tools for invasive species management, some of the developed tools are publicly accessible. In addition, a human-AI symbiosis that combines strengths and compensates for weaknesses is strongly recommended. For each application, directions are provided for future research based on the drawbacks and limitations of the introduced systems. These future efforts will likely increase the adoption of EO-derived products by water managers and improve the reliability of these products.

Table of Contents

1. Chapter 1: Introduction

The incorporation of geospatial analyses are an essential part of landscape-level management, a major theme in landscape ecology (Newton et al. 2009). Landscape ecology includes the study of the effects of disturbance, ecological flows in landscape mosaics, land-use and land-cover change, landscape conservation and sustainable management (Forman 1995; Turner 2005; Turner et al. 2001; Wu and Hobbs 2002). Within these themes, landscape ecologists focus on the analysis of spatial patterns and their links to ecological processes, at a variety of spatio-temporal scales (Wu 2006). This, together with the landscape-level analysis enabled by aerial photography, satellite remote sensing and Geographical Information Systems (GIS) has stimulated strong connections that led to the development of landscape ecology as a discipline (Groom et al. 2006).

The management of Invasive Alien Aquatic Plants (IAAPs) represents a long-standing challenge that can benefit from a landscape-level management approach. Owing to the interconnected nature of freshwater ecosystems and their susceptibility to changes in their surrounding terrestrial landscapes, it is important that any holistic management efforts consider landscape-level interactions (Francis and Chadwick 2012). Invasive species refer to organisms that have established outside their native range with negative consequences for their recipient environments (Republic of South Africa 2004). These IAAP species impose pressure on freshwater systems, threatening their integrity, the biodiversity they support and the human well-being they promote (Carpenter, Stanley, and vander Zanden 2011; Francis and Chadwick 2012). Among IAAP species, water hyacinth (*Pontederia (=Eichhorniae) crassipes* Mart. (Pontederiaceae) (Pellegrini, Horn, and Almeida 2018)) is considered the worst weed, which, despite management efforts, is widespread and ecologically compromising.

1.1. An invasive alien aquatic plant: water hyacinth

Water hyacinth is native to the Amazon basin, South America and is considered an invasive species in more than 50 tropical, sub-tropical or temperate countries (Barrett and Forno 1982; Téllez et al. 2008; Villamagna and Murphy 2010). In South Africa, water hyacinth was first recorded on the Cape flats in the early 1900s and has established itself in numerous sites within the country (Ashton et al. 1979; Coetzee et al. 2011, 2021; Hill et al. 2020; Hill and Coetzee 2017). Currently, water hyacinth can be found throughout many waterways in the south-western winter rainfall areas, eastern subtropical reaches and on the Highveld plateau’s cool and temperate areas (Henderson and Cilliers 2002). This IAAP has been listed as a category 1b invasive on the Alien and Invasive species Regulations as part of the National

Environmental Management: Biodiversity Act, 2004 (Act no. 10 of 2004). Category 1b refers to an invasive species that must be controlled and, wherever possible, removed and destroyed. Any form of trade or planting is strictly prohibited (Maluleke, Fraser, and Hill 2021; Republic of South Africa 2004).

Although secondary dispersal is facilitated by waterbirds and fish (Bartodziej and Weymouth 1995; VonBank 2015), water hyacinth has been introduced primarily by humans as an ornamental plant because of its aesthetic flowers (Edwards and Musil 1975; Guillarmod 1979).(Bartodziej and Weymouth, 1995; VonBank, 2015). The secondary dispersal of the plant is facilitated by its ability to produce many small seeds (up to 5000-6000 seeds per plant, [17,18]) that may remain dormant in the substrate for long periods (up to 28 years (Sullivan and Wood 2012)). In addition, water hyacinth can propagate vegetatively and is therefore not limited by the presence of pollinators or seed dispersers. These factors, together with the occurrence of water hyacinth in neighbouring countries (for example, Zimbabwe, Mozambique, and Botswana (Kurugundla et al. 2016; Langa, Hill, and Compton 2020; Navarro and Phiri 2000)), increase the risk of dispersal through interconnected waterways (hydrochory). In many South African freshwater systems, the establishment of IAAP species are promoted by the simple plant community structures and the relatively low native species richness compared to their terrestrial counterparts (Santamaría 2002), and can therefore be characterised as vacant niches with low biotic resistance and a high vulnerability to alien plant invasions (Wu, Carrillo, and Ding 2017; Wu and Ding 2020).

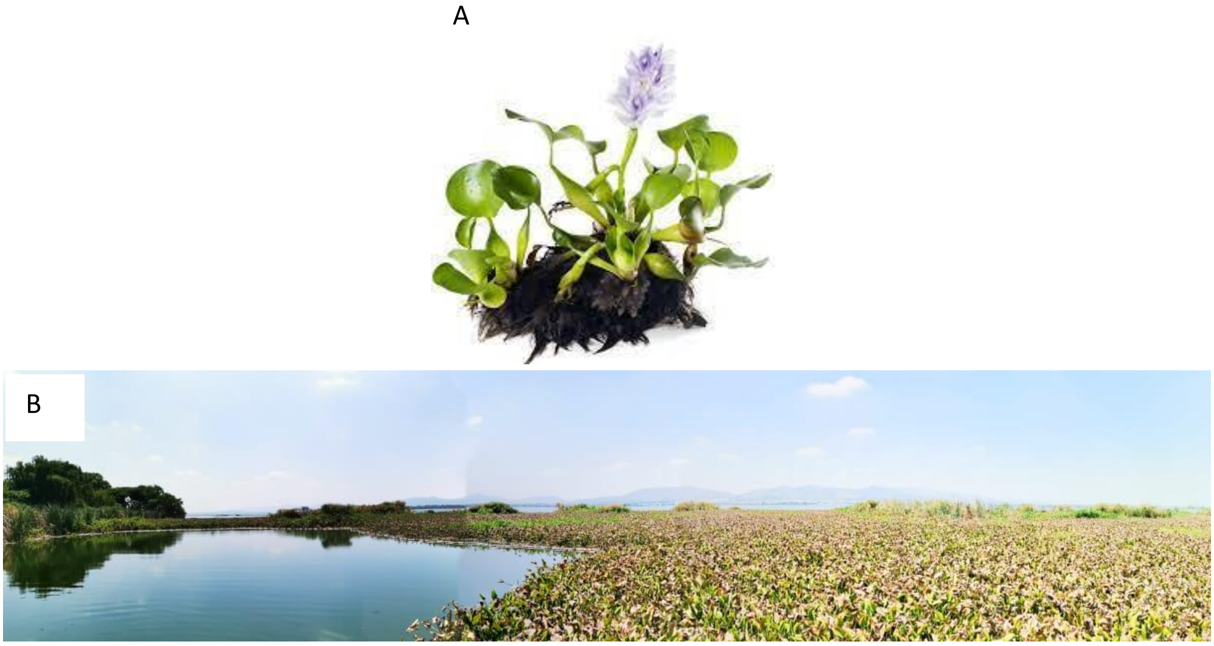


Figure 1. Water hyacinth with its flowers (**A**) and a water hyacinth infestation (**B**) at Hartbeespoort Dam, South Africa during January 2020. Photo (**A**) credit: Irina K - stock.adobe.com. Photo (**B**) credit: Cowie B.

1.2. Impacts

Throughout the invasive range of water hyacinth, numerous ecological and socio-economic impacts have been documented and are outlined below following the categories specified by the Risk Assessment for Alien Taxa (RAAT) guidelines (only applicable impact categories are considered) (Kumschick, Wilson, and Foxcroft 2020).

1.2.1. Environmental Impact

Water hyacinth modifies the water surface, altering the chemical, physical, and structural functioning of the water system. Consequently, this negatively influences the population numbers of numerous taxa and ecosystem functions (Villamagna and Murphy 2010).

Table 1. Categories (according to the Risk Assessment for Alien Taxa (RAAT) guidelines) and details of the environmental impacts of water hyacinth.

Competition	Water hyacinth can out-compete and/or suppress native vegetation and phytoplankton (Little 1965; Mitchell 1985). Free-floating plants (such as water hyacinth) can monopolise light and absorb nutrients from the water column, preventing phytoplankton and submerged vegetation from obtaining sufficient resources for photosynthesis (McVea and Boyd 1975). Refer to (Almeida et al. 2006; Brendonck et al. 2003; Lugo et al. 1998; Mangas-Ramírez and Elías-Gutiérrez 2004; Scheffer et al. 2003) for examples of free-floating water hyacinth out-competing other plant species.
Bio-fouling or other direct physical disturbance	Large dense mats of interlocking water hyacinth may form on the water surface (Villamagna and Murphy 2010), that may cause or aggravate flooding and increase siltation (van Wyk and van Wilgen 2002). These mats may also function to exclude certain aquatic species from important breeding, nursery, and feeding grounds (Twongo and Howard 1998).
Chemical, physical, or structural impact on ecosystem	The water hyacinth mats result in a decrease in dissolved oxygen by preventing the transfer of oxygen from the air to the waters' surface and by blocking light used by phytoplankton and submerged vegetation for photosynthesis (McVea and Boyd 1975). Other impacts on water quality include higher sedimentation rates within the plant's complex root structure and higher evapotranspiration rates from water hyacinth leaves in comparison to open water (Gopal 1987). Water hyacinth has also been found to stabilise pH levels and temperature within lotic systems, increasing mixing within the water column and potentially preventing stratification (Giraldo and Garzon 2002).

Indirect impacts through interactions with other species	Water hyacinth can change fish diets through changes in prey availability (for example, (Njiru et al. 2004)).

1.2.2. Socio-economic Impact

Water hyacinth can exacerbate poverty by affecting agriculture, fisheries, and biodiversity, which form a basis of people's livelihoods in developing countries (Güereña et al. 2015; Lu et al. 2007; Villamagna and Murphy 2010). Moreover, water hyacinth can increase the incidence of diseases such as malaria and schistosomiasis (Ofulla et al. 2010; Téllez et al. 2008).

Table 2. Categories (according to the Risk Assessment for Alien Taxa (RAAT) guidelines) and details of socio-economic impacts of water hyacinth.

Safety (including financial security)	Water hyacinth may reduce access to resources through reduced boating access and navigability; and disrupt the water flow for agricultural irrigation systems, industry, and municipal water supply (Villamagna and Murphy 2010). Mats may also block breeding, nursery and feeding grounds for economically important species, such as tilapia and nile perch (Kateregga and Sterner 2009; Twongo and Howard 1998), resulting in a decrease in population numbers for these economically important species. Water hyacinth cover may hinder effective use of hydropower (Wondie 2010).
Material and immaterial assets	Water hyacinth interrupts the flow of water and increases the risk of flooding (van Wyk and van Wilgen 2002).
Health	Water hyacinth can reduce water quality in various ways and encourage mosquitoes, snails, and other organisms associated with human illnesses, including malaria, bilharzia, schistosomiasis, encephalitis, filariasis, and cholera (Dersseh et al. 2019; Gopal 1987; Harley, Julien, and Wright 1996; Ofulla et al. 2010; Téllez et al. 2008).

To limit the negative impacts of water hyacinth, countries allocate large amounts of resources to limit and manage these invasive populations (Gaertner et al. 2016; Maluleke et al. 2021).

For example, 4.6 million Rand (~275, 448 $) was spent on the control of water hyacinth on three water systems in South Africa between 1977 and 2001 (van Wyk and van Wilgen 2002)

1.2.3. Socio-economic and environmental benefits of water hyacinth

Water hyacinth may decrease some fish populations, however, decreased catchability of certain overfished species can lead to increased fishery stocks (Kateregga and Sterner 2009) that benefit a fishery. Water hyacinth as a herbal medicine contains various bioactive compounds which have been reported to have antioxidant properties (Ali et al. 2010; Chantiratikul, Meechai, and Nakbanpotec 2009; Lata et al. 2010; Surendraraj, Farvin, and Anandan 2013), anticancer activity (Jayanthi and Lalitha 2013) as well as antimicrobial activity (Jayanthi and Lalitha 2013; Rufchaie et al. 2018). Despite these socio-economic benefits provided by water hyacinth, there are other plant species that offer these benefits.

Water hyacinth absorbs heavy metals (Tiwari, Dixit, and Verma 2007) and organic contaminants (Zimmels, Kirzhner, and Malkovskaja 2007), reducing nutrient concentrations and algal blooms (Villamagna and Murphy 2010). On a similar note, water hyacinth's capacity to absorb nutrients makes it a potential biological alternative to secondary and tertiary treatment for wastewater (Cossu et al. 2001; Ho and Wong 1994). Although there is potential for water hyacinth to provide phytoremediation in highly eutrophic systems (Rodriguez-Gallego et al. 2004), the nutrient reductions depend on the density and extent of water hyacinth coverage (Pinto-Coelho and Barcelos Greco 1999). The complex structure provided by macrophytes may provide more microhabitats for epiphytic zooplankton. Several studies have documented a positive correlation between epiphytic macroinvertebrate densities and the surface area of floating aquatic vegetation, including water hyacinth (Crowder and Cooper 1982; Schramm Jr, Jirka, and Hoyer 1987). Floating macrophytes, including water hyacinth, at the open-water interface supported a richer and more abundant macroinvertebrate community compared to the rooted emergent vegetation in Lake Victoria (Uganda). These floating macrophyte species harboured more macroinvertebrates than open water (Masifwa, Twongo, and Denny 2001). Submersed macrophytes and roots from floating macrophytes provide shelter and cover for juvenile and small fish that can enhance fish diversity (Johnson and Stein 1979). In addition, an increase in epiphytic invertebrate abundance and diversity associated with water hyacinth may support a more diverse and larger fish community; however, this is not always the case. Floating patches of water hyacinth in Lake Chapala (Mexico) have provided additional foraging habitat for large wading birds, such as the Great Egret, Snowy Egret, Tri-colored Heron and Great Blue Heron and a direct food source for American Coots (Villamagna and Murphy 2010). Despite the potential benefits of water hyacinth, its negative impacts outweigh these benefits and where possible can be used to incentivise the harvesting of the plant. In this way, a subset of the ecosystem services provided by the plant and its management can co-occur.

1.3. Management of water hyacinth

Biological, chemical, and mechanical control are the three main approaches used to manage water hyacinth populations. Biological control of water hyacinth involves the release of host-specific natural enemies. To date, there are eight different biocontrol agents of water hyacinth released in SA (Coetzee et al. 2021). Chemical control involves the application of herbicides to water hyacinth populations and mechanical control involves the manual removal of plants (refer to van Wyk and van Wilgen, 2002 for the history of chemical and biological control in South Africa). A fourth integrated management approach has become increasingly popular, in which two or all three types of management are implemented in a synergistic manner (Coetzee et al. 2021). For example, biocontrol can be released in succession to sub-lethal herbicidal applications (Coetzee et al. 2021).

Despite the large investment of resources into each of these management approaches, water hyacinth control has been met with variable success (Coetzee et al. 2011, 2021). Four main reasons have been suggested for this and include (Hill and Olckers 2000), antagonistic biocontrol and herbicidal application; cold winter conditions that prevent the build-up of biocontrol agent populations; nutrient-enriched water conditions that allow IAAPs to compensate for herbivory; and, in some cases, a small water body size that lacks the necessary wind fetch to break up mats of agent-infested weeds, limiting dispersal (Hill and Olckers 2000). Since these four factors were suggested more than two decades ago (Coetzee et al. 2021), a large amount of research has gone into further validating and better understanding these factors to improve the success of biocontrol (for example, (Hill et al. 2021; Miller, Coetzee, and Hill 2021)). However, research has been largely centred on costly field or lab experiments that are difficult to scale to the large number of water systems affected at a national extent. Therefore, computer-based analyses that involve remote sensing and machine learning may provide new, alternate, or supplementary approaches to lab and field-based methods. For example, understanding the drivers of IAAP population growth, impact assessment, post-release evaluation or more reliable species distribution mapping, in a cost-effective and scalable manner. To date, these methods have not yet been fully exploited (Parker et al. 2021). The value of these recent technological advances is supported by the exponential increase in the adoption of machine learning applications to satellite data, which aim to tackle environmental issues over the past decade (Sun et al. 2022). By incorporating computer-based approaches, the knowledge-doing gap in ecology can be bridged. The knowing-doing gap refers to the divide between what researchers focus on and what practitioners need to make informed management decisions.

1.4. Leveraging recent advances in technology to enhance IAAP management.

This thesis explores novel applications of machine learning methods to satellite data to provide new tools to assist in the monitoring and management of IAAP species, such as water hyacinth. It is useful to highlight the advantages of combining machine learning and freely available satellite data with the cloud computing infrastructure of Google Earth Engine - a platform that provides free access to many regularly updated geospatial datasets and the computational resources for their analysis (Gorelick et al. 2017).

Since the public release of Landsat data in 2008 (Woodcock et al. 2008; Zhu et al. 2019), there have been numerous additional satellite missions with open-data policies (for example, Landsat-8 (2013-present), Sentinel-1 (2015-present), and Sentinel-2 (2013-present)). The global coverage of this data, which is systematically acquired, processed, and made readily available via cloud computing platforms, such as Google Earth Engine, has accelerated the creation of derived data products at corresponding spatial extents and has driven a shift in remote sensing from mapping (once-off) towards monitoring (Radočaj et al. 2020). The role of machine learning to produce these satellite-derived data products has been pivotal, and in many cases (Camps-Valls 2009; Lary et al. 2016; Sun et al. 2022), represents the only way to extract wide-scale insight from large amounts of satellite data (Zhichkina et al. 2020). The reliance of satellite data on machine learning will continue to grow owing to the rapid advancement in machine learning methods and algorithms that will lead to accuracy gains for satellite-derived products (Camps-Valls 2009; Lary et al. 2016). In many ways, freely available satellite data, machine learning, and cloud computing represent the key ingredients, which have only just begun to provide the novel environmental insights that were previously inaccessible owing to technological limitations. A large part of the benefit of these tools yet to be realised is the open-source and/or open-access machine learning and spatial data analysis software that have a large growing community of contributors (Sun et al. 2022).

These benefits are especially advantageous for small teams that are tasked with managing biological invasions and have a strong focus on field work (Parker et al. 2021).This has translated to a limited number of water systems being actively managed with these tools. Subsequently, information extracted from satellite data through machine learning will be valuable in providing wide-scale information at low-cost, in comparison to the equivalent fieldwork (Parker et al. 2021). This may represent an invaluable opportunity for early detection, early warning, and optimised interventions, therefore improving the capacity for IAAP management. By adopting satellite-derived information and insights, management capacity to plan, make effective decisions under environmental uncertainty, and monitor the outcome of any management interventions will be greatly improved. However, such applications have been limited and have not reached their potential capacity in IAAP management compared to domains such as forestry and agriculture (Zhichkina et al. 2020) or for other freshwater management challenges such as harmful algal blooms (Wynne et al. 2020; Xu et al. 2022), and flood monitoring (Albertini et al. 2022; Mateo-Garcia et al. 2021; Schumann et al. 2022). Therefore, this thesis investigates novel applications of a wide variety of machine learning methods to earth observation data to answer four fundamental questions,

(i) Where is the water hyacinth in South Africa? (Chapter 2)

(ii) Why does water hyacinth (not) occur at a given location? (Chapter 3)

(iii) Which water systems are at a high risk for a water hyacinth infestation? (Chapter 2 and 3) and

(iv) What are likely optimal biological control release periods for a water hyacinth-infested site? (Chapter 4)

1.5. Major Contributions of the Thesis Towards the Management of Water Hyacinth

This thesis has introduced novel methods or combined existing methods for novel applications that enable the following.

1) Monitoring of water and aquatic vegetation based on freely available medium resolution (10-30 m) satellite data (Landsat-8, Sentinel-1 and Sentinel-2) over long periods (2013-present) for any water system (Singh et al. 2020).
2) Mapping of water hyacinth on a national scale (South Africa) at 30 m for 2013 (Singh et al. 2020).
3) Assessment of risk/suitability of water hyacinth infestation at all South African water systems detected at 30 m for 2013 (Chapter 3).
4) Identification and quantification of the relative contribution of the factors used to estimate the suitability/risk of water hyacinth infestation for a particular waterbody in 2013 (Chapter 3).
5) Identification of optimal biocontrol release periods for a single study site between 2018-2020 (Chapter 4).

While the introduced methods and tools are largely demonstrated for sites in South Africa, the low-cost attributes of the methods make them widely suitable, especially for resource-limited countries. For example, the (first) water and aquatic vegetation monitoring web tool can be used for any waterbody globally. This thesis is structured as three independent research papers. Subsequently, the reader will encounter moderate repetition within the introductions of each of the chapters as it relates to the introduction of the challenges surrounding water hyacinth and its management. Since this thesis presents inter-disciplinary research, effort has been taken to make it understandable to machine learning practitioners with a computer science background, remote sensing practitioners with a geography background, and water managers with an invasion ecology background. Chapter 2 introduces a single unsupervised machine learning method that can be applied to both Landsat-8 and Sentinel-2 to detect and monitor changes in the extent of water and aquatic vegetation. After applying this method for the national detection of water and then aquatic vegetation, water hyacinth discrimination is performed throughout South Africa for 2013 using a supervised machine learning algorithm. This chapters also investigates the use of a neural network based semantic segmentation model to directly detect aquatic vegetation. The following two chapters show two potential applications of the methods introduced in Chapter 2.

Chapter 3 focuses on Explainable Artificial Intelligence (xAI) as applied to a Species Distribution Model (SDM) to identify and quantify the relative importance of the likely factors determining the occurrence or, if absent, the risk of water hyacinth invasion on all waterbodies in South Africa. Spatial patterns are highlighted for the factors driving the occurrence of water hyacinth on a regional to a national scale and show the principal drivers for individual water systems in South Africa.

In Chapter 4 the multiarmed-bandit is applied- a machine learning framework to identify periods that are most suitable for the release of biocontrol agents at an individual water system (Hartbeespoort Dam, South Africa). Chapter 3 relies on the mapped water hyacinth distribution from Chapter 2 and Chapter 4 expands on the introduced aquatic vegetation monitoring methods in Chapter 2. Lastly, the synthesis (Chapter 5) provides 1) recommendations for management based on the findings in the three preceding chapters, 2) suggesting directions for future research and 3) providing reflections and recommendations for accessing and reproducing the results of the thesis.

Chapter 2 has been published in *Remote Sensing* (see (Singh et al. 2020)). The first author (Geethen Singh) was the main contributor, while the co-authors played a supervisory role. This thesis has been largely formatted in line with Remote Sensing guidelines to ensure consistency.

Singh, G.; Reynolds, C.; Byrne, M.; Rosman, B. A Remote Sensing Method to Monitor Water, Aquatic Vegetation, and Invasive Water Hyacinth at National Extents. *Remote Sensing* 2020, 12, 4021.

2. Chapter 2: A Remote Sensing Method to Monitor Water, Aquatic Vegetation, and Invasive Water Hyacinth at National Extents

2.1. Abstract

Diverse freshwater biological communities are threatened by Invasive Alien Aquatic Plant (IAAP) invasions and consequently, cost countries millions to manage. Effective management of these IAAP invasions necessitates their frequent and reliable monitoring across a broad extent and over a long-term. Here, a monitoring approach that meets these criteria is introduced and applied based on a three-stage hierarchical classification to first detect water, then aquatic vegetation and finally water hyacinth (*Pontederia crassipes*, previously *Eichhornia crassipes*), the most damaging IAAP species within many regions of the world. Our approach circumvents many challenges that restricted previous satellite-based water hyacinth monitoring attempts to smaller study areas. The method is executable on Google Earth Engine (GEE) extemporaneously and utilises free, medium resolution (10 – 30 m) multispectral Earth Observation (EO) data from either Landsat-8 or Sentinel-2. The automated workflow employs a novel simple thresholding approach to obtain reliable boundaries for open water, which are then used to limit the area for aquatic vegetation detection. Subsequently, a random forest modelling approach is used to discriminate water hyacinth from other detected aquatic vegetation using the eight most important variables. This study represents the first national-scale EO-derived water hyacinth distribution map. Based on our model, it is estimated that this pervasive IAAP covered 417.74 km^2 across South Africa in 2013. Additionally, encouraging results are shown for utilising the automatically derived aquatic vegetation masks to fit and evaluate a convolutional neural network-based semantic segmentation model, removing the need for detection of surface water extents that may not always be available at the required spatio-temporal resolution or accuracy. The water hyacinth species discrimination has a 0.80, or greater, overall accuracy (0.93), F1-score (0.87) and Matthews correlation coefficient (0.80) based on 98 widely distributed field sites across South Africa. The results suggest that the introduced workflow is suitable for monitoring changes in the extent of open water, aquatic vegetation, and water hyacinth for individual waterbodies or across national extents. The GEE code can be accessed here.

Keywords: Cloud computing; previously *Eichhornia crassipes*; Invasive weeds; Machine learning; Mapping; Remote sensing; Satellite.

2.2. Introduction

Biological invasions are responsible for some of the most devastating impacts on the world's ecosystems. Freshwater ecosystems are among the worst affected, with invasive alien aquatic plants (IAAPs) posing serious threats, not only to freshwater biodiversity (Strayer and Dudgeon 2010), but also to the important ecosystem services it provides (Ricciardi and MacIsaac 2011). Once an IAAP species has established itself in a landscape, it can be difficult, if not impossible, to stop or slow down the invasion (Rejmánek and Pitcairn 2002). Thus, regular monitoring of IAAP species is needed to promote targeted, feasible, and effective management (Nielsen et al. 2005; Pyšek and Hulme 2005; Vila and Ibáñez 2011; Wittenberg and Cock 2005). Consequently, there is an urgent need for techniques that enable consistent, frequent, and accurate monitoring of IAAP species (Richardson et al. 2020; Wallace et al. 2014). Semi-automated satellite image analysis techniques using open-access satellite imagery offer an ideal basis for an IAAP monitoring programme. Recent advances in the capture and processing of Earth Observation (EO) data have made these monitoring requirements more achievable.

The frequent capture of Landsat-8 and Sentinel-2 imagery on a planetary scale, for example, increases the capacity for data collection (Cohen and Goward 2004; Li and Roy 2017). For these EO data to be transformed into useful species distributions at correspondingly frequent and large extents, practitioners require costly computational infrastructure to execute algorithms. Fortunately, cloud computing platforms such as Google Earth Engine (GEE) provide the infrastructure (at no cost) to access and rapidly process large amounts of regularly updated EO data in a systematic and reproducible manner (Gorelick et al. 2017). These benefits allow for a large number of images to be summarized using mean, median, or percentile images, and have been reported to reduce data volume without any trade-off in predictive accuracy (Phan, Kuch, and Lehnert 2020). Moreover, semantic segmentation models based on convolutional neural networks (CNN), i.e., pixel-wise localization and classification models, which outperform traditional EO approaches (Persello and Stein 2017), may provide opportunities for more reliable monitoring networks. However, the operational benefits and drawbacks of such an approach, in comparison to the less popular hierarchical classification technique, have yet to be considered. Hierarchical classification refers to a coarse-to-fine class classification approach (Wolter et al. 1995). The use of CNN methods for mapping invasive species, like other biological fields, has been hindered by the time-consuming production of large quantities of high-quality labelled EO data.

Together, these technological and algorithm advancements may prove beneficial for the monitoring of typically highly variable aquatic vegetation cover (Byrne et al. 2010), and for the detection of newly spreading IAAP invasions, at a lower cost than equivalent field work (Bradley and Mustard 2005; Estes et al. 2008). Moreover, these advances may allow for the monitoring of satellite-detectable waterbodies and their associated infestations across national levels, surpassing the capacity of previous methods for large scale discrimination of water hyacinth (*Pontederia crassipes,* previously *Eichhornia crassipes*) (Dube et al. 2017a; Ingole et al. 2018; Mukarugwiro et al. 2019; Thamaga and Dube 2018b, 2019). Subsequently, if IAAP infestations and their associated management efforts are closely monitored across large areas, effective management that allows for the efficient allocation of limited resources may be promoted (Coetzee et al. 2011; Hill et al. 2020; Jones and Cilliers 2001).

The benefit of easily accessible EO data, in conjunction with improved satellite sensor capabilities, has led to a greater uptake of remote sensing in the field of invasion biology (Dvořák et al. 2015; Joshi, de Leeuw, and van Duren 2004; Manfreda et al. 2018; Müllerová, Pergl, and Pyšek 2013; Thamaga and Dube 2018a). However, in a review of remote sensing applications within invasion biology, (Vaz et al. 2018) showed that freshwater habitats have received far less research interest when compared to their terrestrial counterparts. The bias toward terrestrial systems persists despite freshwater ecosystems being under greater invasion risk per unit area (Rahel 2002; Sala et al. 2000). The respective magnitude of impacts was identified as one of the main motivations influencing the selection of organisms and habitats in invasion management and likely, research (Pyšek et al. 2008). Additionally, the difficulty to study aquatic systems and limited boat access may have also contributed to this pattern. Despite water hyacinth being a high impact IAAP species (Villamagna and Murphy 2010), there is a limited number of studies showcasing the use of freely available medium spatial resolution (<30 m) multispectral satellite data, to locate IAAP species (for example, Mukarugwiro et al. 2019; Ingole et al. 2018; Dube et al. 2017b; Cheruiyot et al. 2013). Ultimately, there are still a limited number of studies that forecast the risk of IAAP species spread (for example, (Truong, Hardy, and Andrew 2017)), or that aim to apply end-to-end satellite data approaches to investigate the drivers and impacts of IAAPs (for example, (Agutu, Gachari, and Mundia 2018)). Therefore, despite the benefits to be gained from EO-based IAAP monitoring at large extents (Thamaga and Dube 2018b; Zhang et al. 2017), such an approach has not yet been realized. This may be largely attributed to previous studies that rely on a single supervised model to simultaneously discriminate water hyacinth from multiple types of surrounding land cover (Dube et al. 2017b; Ingole et al. 2018; Mukarugwiro et al. 2019; Thamaga and Dube 2019). Consequently, these studies highlight the erroneous detection of surrounding land cover as water hyacinth (Dube et al. 2017b, 2017a; Ingole et al.

2018; Mukarugwiro et al. 2019; Thamaga and Dube 2018b). This source of error, together with the limited, local distribution of field locality data hinders the detection of any IAAP within waterbodies distributed across large areas in a reliable manner.

This chapter details a novel framework for the application of an IAAP monitoring network, which relies on freely available EO data. South Africa is selected because of the availability and access to locality data either collected by experts or under expert supervision and is used to illustrate the framework's suitability for national scale detection and discrimination of water hyacinth, which is of concern within the region and other parts of the world (Hill and Coetzee 2017). Moreover, the proposed three-stage hierarchical classification framework allows for the national scale distribution of surface water and aquatic vegetation to be determined, closest to a user-specified date. Or, if implemented on a site-basis, water extent and aquatic vegetation cover can be monitored at a frequency of up to three days. Firstly, a novel modified normalized difference water index (MNDWI, (Xu 2006)) thresholding approach is introduced to obtain surface water extents (Section 2.3.2.1.). Thereafter, the role of a hierarchical classification approach that allows for a repurposed Otsu thresholding and Canny edge detection (Otsu + Canny) algorithm to detect aquatic vegetation is discussed (Section 2.3.2.2.). Next, the use of these automatically derived aquatic vegetation masks to fit and evaluate a CNN-based semantic segmentation model is considered, for the first time, which circumvents labelling constraints and complexities associated with water detection (Section 2.3.2.4. and 2.3.2.5.). Thereafter, this chapter shows how a random forest model using selected variables can discriminate between water hyacinth and other detected aquatic vegetation pixels (Section 2.3.2.5.). Finally, guidelines, caveats, and limitations associated with the proposed workflow are discussed (Section 2.5.5.).

2.3. Materials and Methods

2.3.1. Study Area and Study Species

Currently, the Southern African Plant Invaders Atlas (SAPIA) offers the most comprehensive data to assess the national extent of IAAP invasions (Richardson et al. 2020). Unfortunately, the maps presented in the SAPIA only highlight general distribution patterns owing to their coarse quarter-degree spatial resolution, precluding the understanding of local-scale distribution patterns and population dynamics. Moreover, the GPS locality data underpinning the maps in the SAPIA combines data from different periods, may be outdated and is not spatially comprehensive. According to these data, South African aquatic weeds are found mainly in the southwestern winter rainfall areas, eastern subtropical reaches and on the Highveld plateau's cool and temperate areas (Henderson 2001).

The plant locality data for this study were collected during annual national field trips (unpublished data, (Coetzee and Mostert, E; (Rhodes University 2019)). A single study species (water hyacinth) was selected as the target species because of limited GPS reference localities for the remaining four species. These four species include water lettuce (*Pistia statiotes*), parrot's feather (*Myriophyllum aquaticum*), Kariba weed (*Salvinia molesta*), and red water fern (*Azolla filiculoides*). Together, these five species make up the most damaging aquatic IAAP species within South Africa (Henderson and Cilliers 2002; Richardson and van Wilgen 2004; van Wilgen et al. 2001). The non-water hyacinth localities were combined with the "none" class (i.e., other riparian or aquatic vegetation) to form the non-target or negative class during mapping. No 'none' locations are known to contain water hyacinth infestations. The disparity in the availability of locality data between water hyacinth and the remaining four species is a direct result of water hyacinth being more widespread throughout the country (Figure 1).

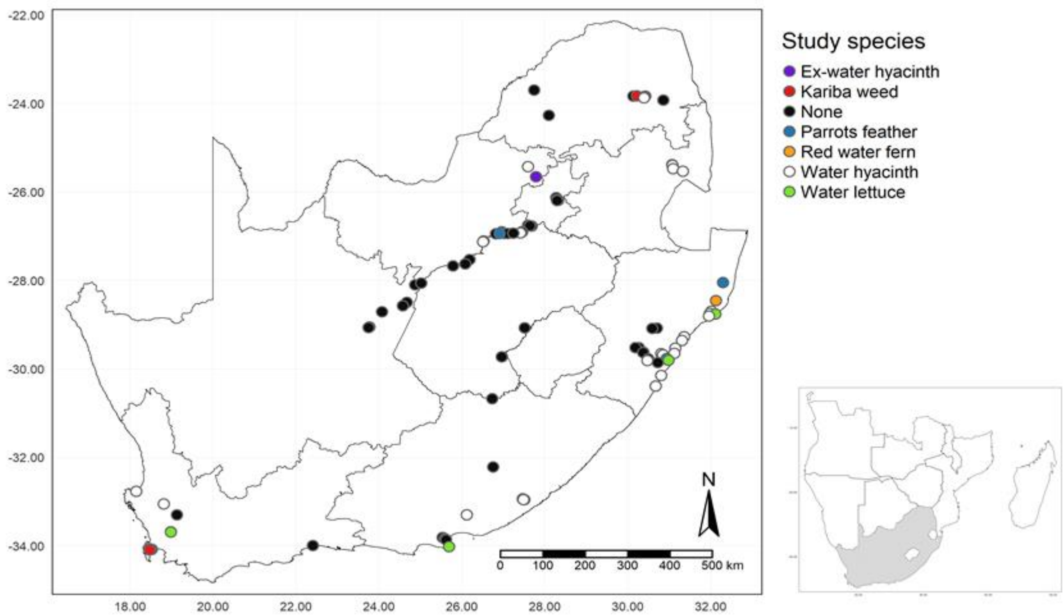


Figure 1. The field localities of five invasive alien aquatic plant (IAAP) species for 2013–2015 across 98 sites within South Africa including sites with detected aquatic vegetation but none of the five IAAPs (Unpublished data, (Coetzee and Mostert, E; (Rhodes University 2019)). The period, 2013–2015 corresponds to a temporal window of concurrent availability of both the Landsat-8 satellite data and the field IAAP locality data. Note, the strong imbalance between the seven classes in favour of water hyacinth and against the other IAAP species. As a result, the water hyacinth

localities were used as the target (positive) class while the remaining categories were combined to form the non-target (negative) class.

2.3.2. The Mapping Process

To map and discriminate between water hyacinth and other non-target species, a three-stage process was followed. During the first stage, surface water was detected using a novel MNDWI thresholding approach. In the second stage, aquatic vegetation was detected using two different methods (Otsu + Canny, and MultiResUnet semantic segmentation). In the third stage, the Otsu + Canny detected vegetation was discriminated into water hyacinth and other non-target vegetation pixels using a random forest model. Each of these stages are detailed in Figure 2. In addition, an example output from stages 1 and 2 is provided in Figure 3.

The spectral index equations of MNDWI (Xu 2006), normalised difference vegetation index (NDVI) (Tucker 1979), land surface water index (LSWI) and the green atmospherically resistant index (GARI, (Gitelson et al. 2002; Xiao et al. 2002)), normalized difference aquatic vegetation index (NDAVI) used in this study are listed as Equations (1) – (5) (Villa et al. 2013).

$$MNDWI = \frac{Green - SWIR1}{Green + SWIR1} \tag{1}$$

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{2}$$

$$LSWI = \frac{NIR - SWIR1}{NIR + SWIR1} \tag{3}$$

$$GARI = \frac{NIR - (Green - (Blue - Red))}{NIR - (Green + (Blue - Red))} \tag{4}$$

$$NDAVI = \frac{NIR - Blue}{NIR + Blue} \tag{5}$$

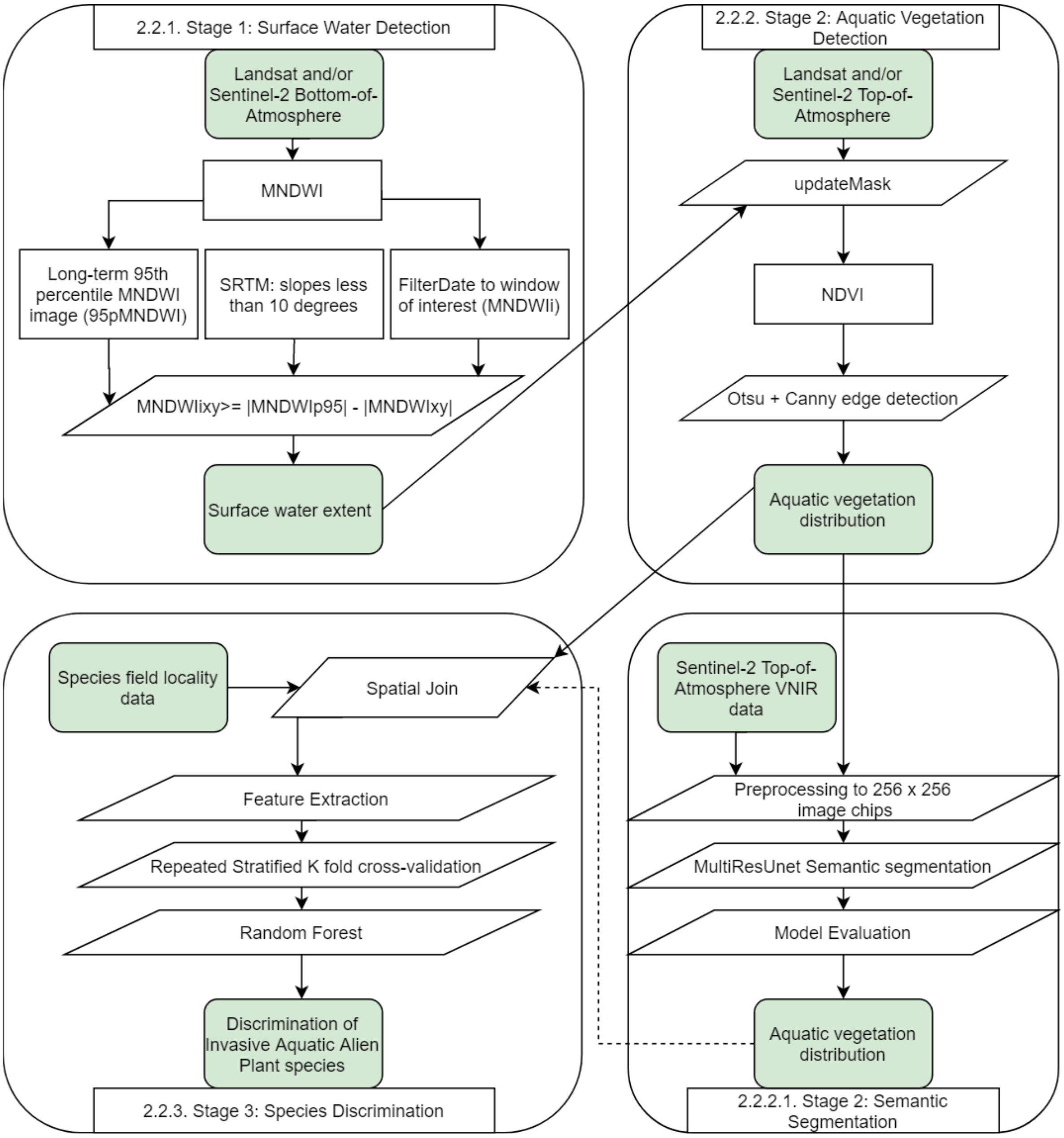


Figure 2. The general workflow used in this study to map and discriminate between invasive alien aquatic plant species across South Africa with input and terminal output data (**green rectangles**), intermediate outputs (**white rectangles**) and processes (**parallelograms**). The dashed arrow indicates the alternative (potential) use of the MultiResUnet-derived aquatic vegetation (Section 2.3.2.4) for water hyacinth discrimination (Section 2.3.2.5.). Refer to electronic color copy for the reference to color.

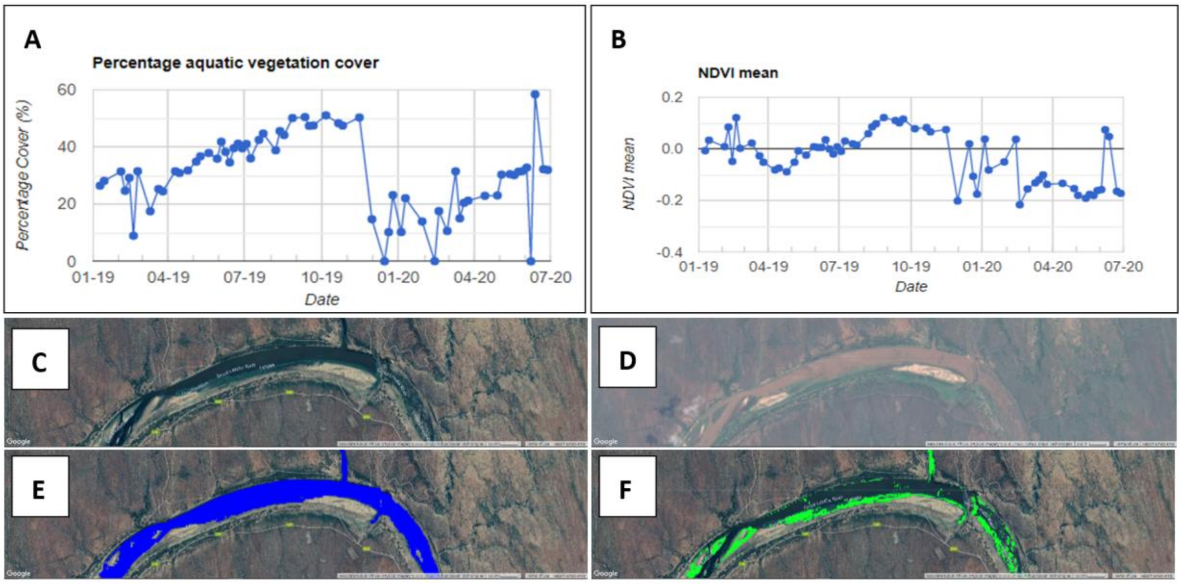


Figure 3. The ability of the introduced workflow's (Figure 2) stage 1 and 2 to monitor aquatic vegetation cover (**A**), associated mean aquatic vegetation normalized difference vegetation index (NDVI) (**B**) from January 2019 to July 2020 for the area depicted in the high resolution google earth image of Engelharddam, Kruger National Park, South Africa (**C**), a RGB Sentinel-2, level 1C (atmospherically uncorrected) image (**D**) with long-term (blue) surface water (**E**) and Otsu + Canny detected (**green**) aquatic vegetation (**F**) corresponding to the first Sentinel-2 observation of chart A and B (9 January 2019) for the same area (**F**). The application that provides similar charts for any waterbody of interest can be accessed here (script name: Aquatic vegetation and water monitor).

2.3.2.1. *Stage 1: Surface Water Detection*

The automatic detection of surface water across South Africa was achieved by first deriving a 95th percentile MNDWI image composite from the Landsat-8 surface reflectance image collection (*MNDWIp95*), which extended from the start of the archive (2013) to the end of 2018. A 95th percentile image is computed on a per-pixel basis, i.e., all pixel values for a location are sorted in ascending order and the value corresponding to the 95th percent index position is used as the pixel value. The 95th percentile image leverages the spectral properties of clouds, water, and land surfaces that are captured by the MNDWI index. Generally, the MNDWI value of water (MNDWIw) is greater than that of clouds (MNDWIc), but the MNDWI value of clouds is greater than those of non-water areas (MNDWIo). For example, irrigated agricultural areas and shadow areas that overlap (spectrally) with MNDWI water values (Donchyts et al. 2016). As a result of this relationship, i.e., MNDWIw > MNDWIc > MNDWIo, the contrast between water and cloud covered land surfaces that are frequently misclassified as water become more pronounced (Figure 4). This makes the 95th percentile MNDWI image useful for water detection. This *MNDWIp95* layer was then used within an inequality (6) to detect surface water. Additional pre-processing steps included the removal of sensor anomalies and steep slopes by removing pixels with *MNDWIp95* values greater than 0.9984 or with a slope greater than 10 degrees (Yang et al. 2020). The 0.9984 threshold was determined by manually inspecting the minimum value of scene overlap artefacts. The slope layer was derived from the Shuttle Radar Topography Mission (SRTM) elevation dataset. To produce a monthly/quarterly/annual water map, the detected water for each cloud-free image within the period of interest was superimposed.

A surface water pixel with unique coordinates x and y can be characterized by:

$$MNDWIxy \geq |MNDWIp95| - |MNDWIxy|, \tag{6}$$

where *MNDWIxy* is the MNDWI pixel value at coordinates x and y, from a single image. Since the inequality (6) utilizes a difference between two periods on the right-hand-side (RHS), this can be considered as a bi-temporal change detection component for which four scenarios need to be accounted for. These four scenarios include: (1) water remains water, i.e., LHS ≥ RHS; (2) water is gained, i.e., LHS ≥ RHS; (3) water is lost, i.e., LHS < RHS; and (4) non-water remains non-water, i.e., LHS < RHS. To satisfy all four of these scenarios, the use of absolute values on the RHS are necessary. For example, in scenario three, when water is lost at a pixel, the long term MNDWI value (*MNDWIp95*) will be high (closer to 1) and the MNDWI value (*MNDWIxy*) after the pixel has undergone a loss of water will have a low MNDWI value. Therefore, the difference between these two values (|*MNDWIp95*| − |*MNDWIxy*|) will be greater than the MNDWI value on the LHS (*MNDWIxy*), satisfying the inequality for scenario 3, i.e., LHS < RHS.

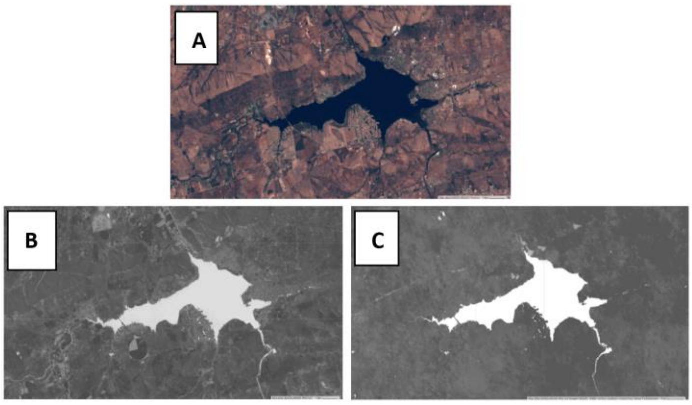


Figure 4. A portion of an RGB Landsat-8 image (**A**), the corresponding derived modified normalized difference water index (MNDWI) image (**B**) and the 95th percentile image derived from a long-term (2013-2019) MNDWI image collection (**C**) over Hartbeespoort Dam, South Africa. The benefit of using a 95th percentile image is shown by its ability to decrease the MNDWI pixel values over non-water areas (e.g., agriculture) while increasing the MNDWI values of water pixels. This helps enhance the contrast between water and non-water pixels.

The accuracy of the annual detected surface water was evaluated against the annual Joint Research Centers' (JRC) Global Surface Water product (GSW, (Pekel et al. 2016)), available on Google Earth Engine, using the Intersection over Union (IoU) metric. This metric represents the ratio of the shared pixel area to the total area after merging the two datasets. For this evaluation, the maximum extent of seasonal and permanent surface water from 2013 to 2015 was compared against the water extent derived using the introduced MNDWI thresholding approach for the same period. The second evaluation involved a comparison of the 2018 surface water layer from this study and the GSW dataset, each against the 2018 South African National Land Cover field validation data points (SANLC, (GEOTERRAIMAGE 2020)) using a confusion matrix. To enable the comparison, all water related class points were reclassified as water (class 12–21, n = 1316) or non-water (remaining classes, n = 5254). Lastly, Sentinel-2 derived surface water masks for France were used to assess the method’s transferability to different platforms and its ability to generalize to different countries (Yang et al. 2020).

2.3.2.2. *Stage 2: Aquatic Vegetation Detection*

The combination of the Otsu thresholding and Canny edge detection techniques (Otsu + Canny) have previously been used to detect water (Donchyts et al. 2016) by the calculation of a local, as opposed to a global, MNDWI threshold value through the Canny edge detection algorithm. This allows for a more accurate separation of water pixels from background pixels using Otsu thresholding (Otsu 1979) and is achieved by determining a local threshold value that maximizes the between class variance (interclass variance).

Here, the Otsu + Canny approach was repurposed to detect aquatic vegetation. The ability to detect aquatic vegetation as opposed to water was achieved by using the water extent generated from the previous water detection stage to limit the Landsat-8 or Sentinel-2, level 1C pixels that are passed to the Otsu + Canny detector. This results in a bimodal distribution for those sites that contain aquatic vegetation and therefore, allows for the computation of a NDVI threshold value that separates water from aquatic vegetation. By limiting vegetation detection to the extent of open water, the confusion between aquatic vegetation and surrounding landcover, including terrestrial and riparian vegetation, is reduced.

The benefit of using the Sentinel-2, Top-of-Atmosphere (TOA) product is its extended availability compared to the Bottom-of-Atmosphere (BOA) product—available outside of Europe only from December 2018. To determine if the increased atmospheric influence on the TOA derived spectral indices influenced the detection ability of aquatic vegetation, three different spectral indices with varying sensitivity to atmospheric interference were evaluated. These include, from least sensitive to most sensitive, the GARI (Gitelson et al. 2002), the NDVI

(Tucker 1979), and the NDAVI (Villa et al. 2013). The higher sensitivity of the NDAVI index can be attributed to the blue band that is used to derive the index.

2.3.2.3. *Stage 2: Semantic Segmentation: An Alternative Aquatic Vegetation Detection Method to Otsu + Canny*

The detected aquatic vegetation was then used to train a semantic segmentation model. Unlike the Otsu + Canny method, a semantic segmentation model directly detects aquatic vegetation, circumventing the need to first detect water and, in this way, preventing the propagation of errors from water layers that may not always be available at a desirable spatio-temporal resolution or accuracy. The MultiResUnet architecture applied here (Ibtehaz and Rahman 2020), like the popular Unet model, follows the encoder-decoder model structure (architecture, (Litjens et al. 2017)). During the encoder or contracting path, the image dimensions are reduced and encoded into feature representations at multiple levels. Thus, the model trades off spatial information for semantic information. These encoder layers consist of convolution layers that help to learn the contextual relationships, for example, between water and aquatic vegetation. During the decoder or expansive path, the feature information and spatial information are combined through a series of deconvolutions and concatenations with high-resolution features from the contracting path to allow for corresponding height and width dimensions between the input image and the output segmentation map.

The first improvement of the Unet architecture that led to the MultiResUnet architecture included two additional convolution steps, a 3 × 3 and a 7 × 7 kernel, in addition to the usual 5 × 5 convolution, to improve the model's ability to cope with the detection of variable sized aquatic vegetation mats. Second, the concatenation paths, which transfer high-resolution spatial information from the encoder layers to the decoder layers, contain skip connections between added layers (Ibtehaz and Rahman 2020). This helps align deep and shallow features being assembled from the encoder and decoder (Ibtehaz and Rahman 2020).

2.3.2.4. *Pre-Processing*

462 Sentinel-2A and 2B, level 1C image scenes (~2304 pixels × 2304 pixels) of Hartbeespoort Dam, South Africa from the start of the Sentinel-2 archive (23 June 2015) to the 22nd of June 2019 was selected as the test area for semantic segmentation because it contains a large known infestation of water hyacinth. Moreover, it is surrounded by a heterogenous landscape, including urban infrastructure, agricultural land, and shadow areas—all of which are frequently confused with water or aquatic vegetation. Therefore, the heterogeneous composition of the Hartbeespoort area can more closely represent other areas of interest with a similar or less complex landcover composition. The 462 visible and near-infrared (VNIR) images were filtered

to 275 images with less than 10% cloud cover. The 10 m VNIR images with corresponding Otsu + Canny aquatic vegetation masks were downloaded from GEE for each of the 275 images.

The downloaded images and corresponding masks were sliced and saved into 16-bit patches of 256 × 256 pixels to reduce the memory footprint during model training. The VNIR bands were further normalized between the range of zero to one to improve model convergence during learning. The slicing yielded 1700 patches that were then randomly divided into 80% training data, 10% validation data, and 10% test data. The validation data were used for halting model fitting (early stopping), i.e., when the validation loss stopped decreasing over seven consecutive epochs, a point at which model overfitting to data noise may occur. An epoch, for training a neural network, refers to a single cycle of parameter learning from the training dataset, calculation of error (loss) against the validation dataset and subsequent update of learnt parameters for the next learning cycle. The term loss is used instead of error since a neural network outputs a probability of belonging to the target class (aquatic vegetation). The validation loss (binary cross-entropy loss) represents the mean negative log probability for the predicted probabilities. This is comparable to an error score that may be reported for discrete predicted outputs.

2.3.2.5. *Stage 3: Species Discrimination*

The GPS localities (*n* = 513) that contain species name, waterbody name and year were linked to waterbodies within 250 m of a GPS point. This resulted in 458 waterbodies with labels, 98 of these waterbodies contained aquatic vegetation detected by Landsat-8. These final 98 waterbody extents were then used to clip the explanatory image variables (refer to Table 1 for the list of variables). The aquatic vegetation pixels within the clipped waterbody extents were identified using the Otsu + Canny approach. To reduce the computation for the aquatic vegetation detection stage required for deriving the national water hyacinth distribution, a near-maximum (95th percentile) NDVI image was selected instead of detecting and superimposing the aquatic vegetation for every NDVI image in 2013. This still provided an accumulated aquatic vegetation extent for the year while reducing the inclusion of algae with a lower NDVI than water hyacinth. A near-maximum NDVI image was used instead of a maximum image to avoid outlier NDVI values and false aquatic vegetation detection.

Subsequently, random forest models (number of trees = 100) were calibrated and evaluated using a repeated stratified k-fold cross-validation strategy (repeats = 20, k = 5, strata = one of seven classes in Figure 1) to calculate a summed confusion matrix and the mean and standard deviation overall accuracy, F1-score and Matthews correlation coefficient (MCC) metrics.

Since the initial calibration dataset contained point localities and not pixel-wise image masks, the popular mean Intersection over Union (mIOU) metric was not used to assess the species discrimination. During cross-validation, the aquatic vegetation pixels that belonged to a waterbody were constrained to either the training or testing partition, but not both partitions. This reduced the overestimation of accuracy owing to spatial autocorrelation and is referred to here as spatially constrained repeated stratified k-fold cross-validation. A random forest model consists of multiple decision trees that are grown from a random selection of explanatory variables. The final prediction is based on the most common (mode) prediction class from all decision trees.

Table 1. The initial 64 explanatory variables that were used to fit a random forest model that discriminated between water hyacinth and other aquatic vegetation pixels.

Variable	Data Source (Year)	Spatial Resolution (m)
Seasonal median spectral bands, excluding panchromatic band 8 Summer (12/01–02/28) Autumn (03/01–05/31) Winter (06/23–08/31) Spring (09/01–11/30)	Landsat-8 for the respective field locality year (2013–2015)	30
Percentiles (5, 25, 50, 75, 95) of spectral indices (NDVI, GARI, LSWI)	Landsat-8 for the respective field locality year (2013–2015)	30
Elevation	SRTM (2000)	90
Global human modification	Kennedy et al. (2019) (2016)	1000
Sum of solar radiation	Terraclimate (2013–2015)	~4670
Minimum temperature Temperature seasonality Precipitation seasonality	Worldclim: bio variables (1960–1990)	1000
Gross biomass water productivity Actual evapotranspiration Total biomass production	Wapor for the respective field locality year (2013–2015)	250

NDVI—Normalized Difference Vegetation Index (Tucker 1979), GARI—Green Atmospherically Resistant Index (Gitelson et al. 2002), LSWI—Land Surface Water Index (Xiao et al. 2002), Shuttle Radar Topography Mission (SRTM).

To determine the variables that were used in the final model, a combination of feature selection techniques was employed, recursive feature elimination, permutation importance, and highly correlated feature elimination. Feature selection removes redundant features, decreases computation, improves model interpretability, and may also improve model accuracy. Recursive feature elimination served to remove those variables that decreased the F1-score. Thereafter, these features were further refined by removing those variables that had a negative permutation importance score and finally, for two features that had a high correlation coefficient (>0.8), the correlated feature that had a lower permutation importance score was removed. The coarser resolution explanatory variables (human modification, minimum temperature, and precipitation seasonality) were resampled to 30 m (Landsat-8) by default within GEE using nearest neighbour resampling. The final model, fitted on the selected variables, was deployed on GEE to discriminate water hyacinth from other non-target species across South Africa.

2.4. Results

2.4.1. Evaluation of Surface Water Detection

When applying the introduced water detection approach to Landsat-8 data for South Africa and comparing the results to the maximum GSW extent for the same period of 2013 to 2015, a moderate 0.695 IoU score was obtained. A value of 1 corresponds to complete overlap between datasets. These results suggest the superior accuracy of the GSW dataset. However, when compared to the SANLC dataset, both approaches perform similarly with our method slightly outperforming the GSW dataset based on MCC (Figure 5). In addition, the confusion matrices highlight a greater commission error than omission error for both datasets.

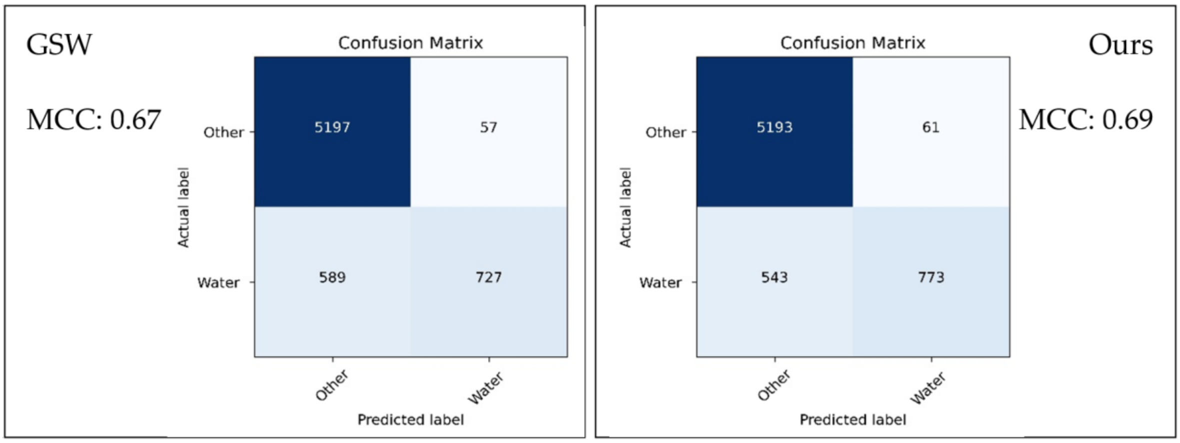


Figure 5. Comparison of misclassified pixels between the 2018 South African National Land Cover (SANLC) validation points and the 2018 Global Surface Water (GSW) water (**left**) and the 2018 detected surface water using the current method (**right**).

To evaluate the transferability of the introduced water detection methods to Sentinel-2 and a different country, our method was compared to a 10 m surface water dataset for France (Yang et al. 2020). The 10 m dataset was derived by a rule-based superpixel (RBSP) method (Yang et al. 2020). A visual comparison of the GSW, RBSP-derived water layer, and the water layer derived using the introduced method can be found here. Seasonal and smaller waterbodies were less accurately detected compared to the RBSP method and, to a lesser degree, the GSW layer. Note, in addition to the water detection method (Section 2.3.2.1.), an additional post-processing step for France was carried out to reduce the commission errors with agricultural areas. This was achieved by removing pixels that had a higher NDVI value than a normalized difference water index (NDWI) value.

2.4.2. Evaluation of Aquatic Vegetation Detection

The Sentinel-2 TOA image collection, available on GEE, can be used to detect aquatic vegetation with a superior accuracy to the BOA data despite the influence of atmospheric effects (Figure 6).

A visual comparison between the detected aquatic vegetation from the Otsu + Canny and the MultiResUnet approach showed a high correspondence with each other and their associated RGB image (Figure 7). In addition, the predicted outputs correspond to a low validation binary cross-entropy loss/error and a high validation accuracy against reference aquatic vegetation generated using the Otsu + Canny algorithm (Table 2).

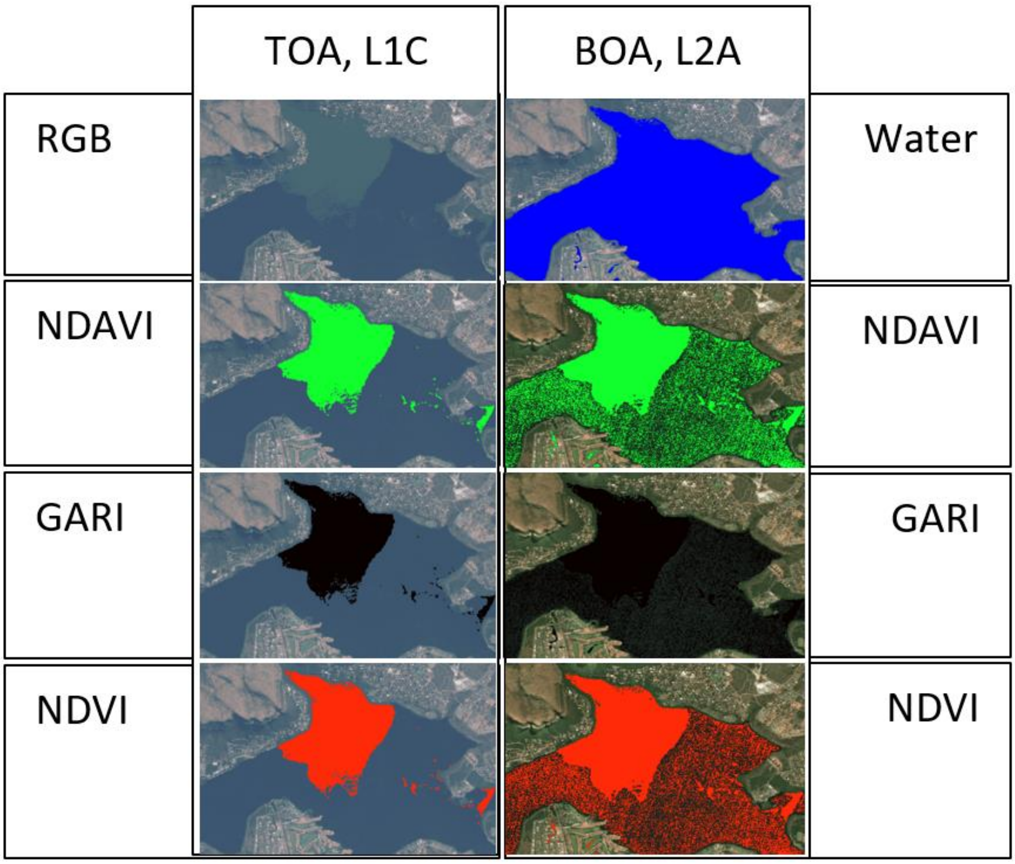


Figure 6. The detection of water (**blue**) and aquatic vegetation derived from three vegetation indices with a varying sensitivity to atmospheric conditions, i.e., normalized difference aquatic vegetation index (NDAVI) (**green**) the most sensitive, GARI (black) the least sensitive and NDVI (**red**) with relatively moderate sensitivity from Sentinel-2 Top-of-Atmosphere (left column, TOA, L1C) and Bottom-of-

Atmosphere (right column, BOA, L2A) data over a portion of Hartbeespoort Dam, South Africa.

Table 2. The training and validation mean (± SD) accuracy and binary cross entropy loss/error for the semantic segmentation (MultiResUnet) of aquatic vegetation (examples are shown in Figure 3). Note, accuracy values closer to one, and low loss values closer to zero are better and are based on Otsu + Canny reference masks. The scores of the last epochs are averaged over three iterations of model fitting (lasting 94, 76, and 83 epochs respectively).

Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
0.9814 ± 0.0030	0.9851 ± 0.0035	0.0655 ± 0.0017	0.0740 ± 0.0078

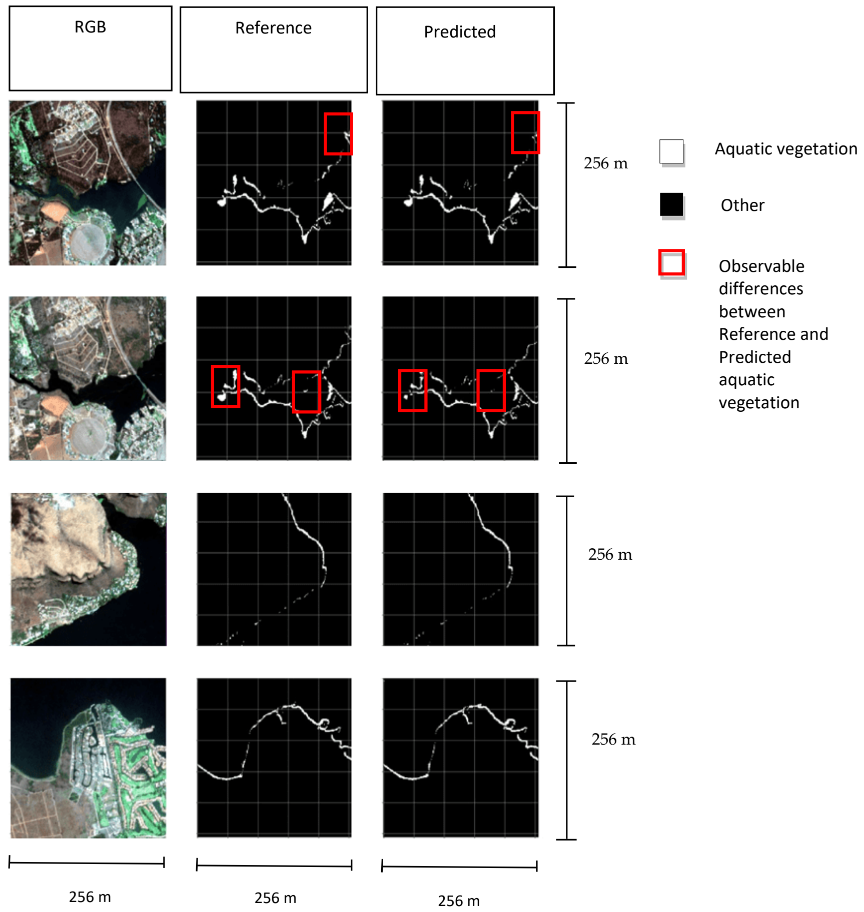


Figure 7. The visual correspondence between the true-color Sentinel-2 256 × 256 image patch, the associated reference mask of aquatic vegetation (**white**) generated using the Otsu + Canny method, and the MultiResUnet aquatic vegetation prediction mask over portions of Hartbeespoort Dam, South Africa. Each row represents an example of the test data. The first two rows cover the same portion of Hartbeespoort Dam viewed on different days.

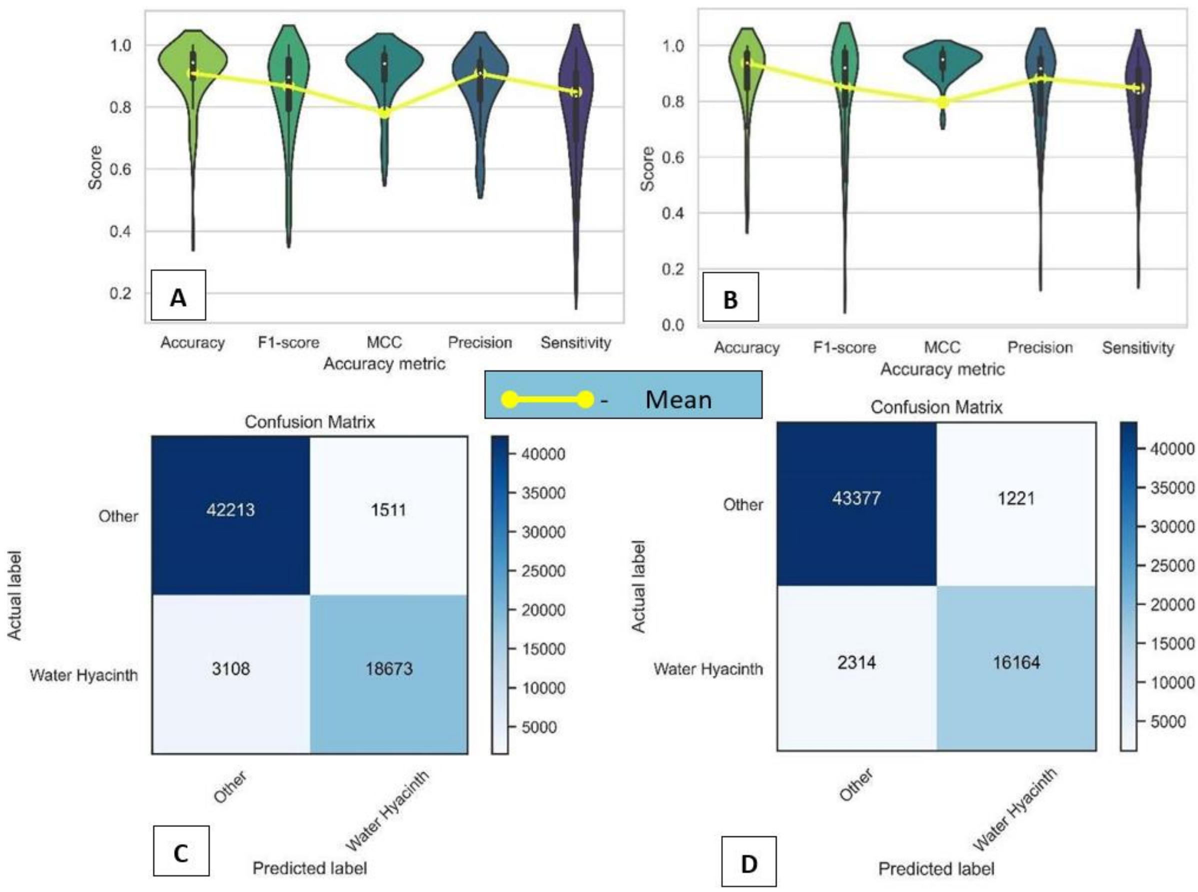


Figure 8. Slightly lower accuracy scores for water hyacinth discrimination when using all 64 features ((**A**), listed in Table 1) than when using the top eight most important features (**B**). Considering the results of using all 64 features (**A, C**) and the top eight features (**B, D**), the MCC shows the lowest variation among the five metrics in both scenarios. This points to its reduced sensitivity to the negative class imbalance evident in the summed confusion matrices (**C–D**). There is also a false positive ratio of 2:1 for water hyacinth and other aquatic vegetation, respectively.

2.4.3. Evaluation of Aquatic Vegetation Discrimination

The distribution values (mean and standard deviation) of the overall accuracy, F1-score, MCC, precision, and sensitivity metrics are based on spatially constrained repeated stratified k-fold (k=5) cross-validation (Figure 8A, B). The validation data, like the calibration data, show a negative class imbalance along with false positive and false negative occurrences for water hyacinth and non-water hyacinth pixels (Figure 8C, D). The F1-score is a weighted average of precision and recall, and thus takes into consideration both false positives and false negatives, respectively. The F1-score and MCC are much less sensitive to the class imbalances between water hyacinth (positive instances) and non-water hyacinth pixels (negative instances). However, for binary classification, especially with imbalanced data cases, the MCC metric is favored over other accuracy metrics, including the overall accuracy and F1-score metric (Chicco and Jurman 2020) because the MCC metric is high when the majority of both positive and negative instances have been successfully discriminated. Nevertheless, multiple metrics are reported for easier comparison with other studies. Based on MCC, the metric that is least sensitive to class imbalance, the discrimination accuracy of water hyacinth and non-water hyacinth pixels is 0.78 (0.17) when using all 64 variables (listed in Table 1) or 0.80 (0.16) if limited to the eight most important variables (Figure 9).

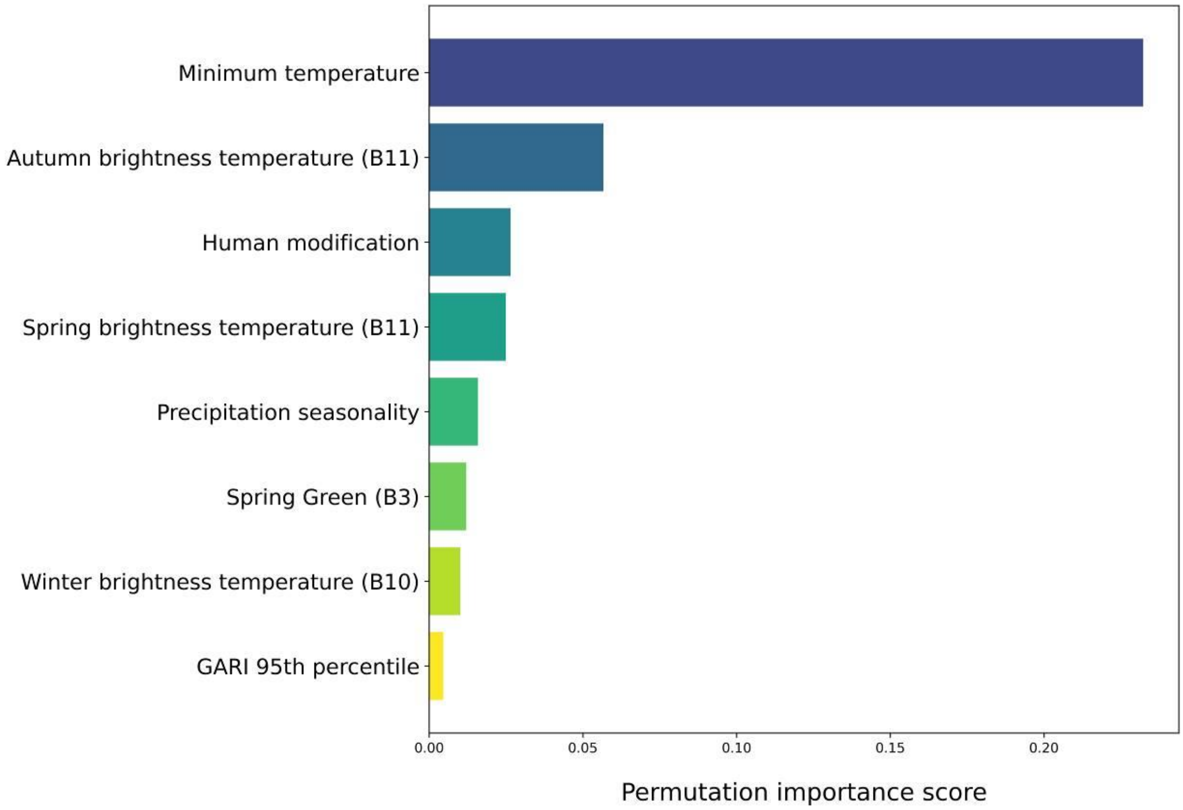


Figure 9. Mean permutation scores for the top eight most important variables used to map and discriminate water hyacinth from other aquatic vegetation from 2013–2015, across 98 reference sites. The permutation importance scores are calculated from the spatially constrained repeated stratified k-fold cross-validation. Note, Shuttle Radar Topography Mission (SRTM) elevation showed similar importance scores to minimum temperature but was removed owing to its high (>0.8) correlation with minimum temperature.

Four of the eight most important variables for discriminating between water hyacinth and other aquatic vegetation are thermally related; for example, the minimum temperature in the coldest month and autumn, spring and winter brightness temperature (top-of-atmosphere radiance). Optically active bands that feature among the top performing discriminative bands include the autumn median green band and the 95th percentile GARI (Figure 9). Moreover, by reducing the number of variables and keeping the eight most important variables, there is no loss in accuracy but rather a slight improvement in discrimination accuracy (Table 3). Lastly, owing to the wide distribution of sites used during cross-validation, these results suggest the effectiveness for water hyacinth discrimination across wide areas (Figure 10).

The user and producer accuracy of water hyacinth is used as a comparison to previous studies since this metric is reported by all five previous studies (Table 3). This study showed superior water hyacinth discrimination results for at least one of these metrics despite the (67x) larger inference area (Table 3).

Table 3. Comparison of user and producer accuracy scores between studies that have mapped water hyacinth. The area of imagery (km^2) that water hyacinth was mapped is provided. The area estimates (~) were not provided in the studies and were based on the bounding box of the georeferenced output classification map found within each study. The highest accuracy and greatest area shown in bold print.

Reference	**Accuracy (%)**		**Area (km^2)**	**Sensor (Season)**
	User	**Producer**		
(Dube et al. 2017a)	67.35	67.35	~958	Landsat-7
	91.67	89.8	~958	Landsat-8
(Dube et al. 2017b)	**100**	90	~5228	Landsat-8 (wet)
	92	90	~5228	Landsat-8 (dry)
(Thamaga and Dube 2018b)	44	50	~5228	Landsat-8
	89.3	61	~5228	Sentinel-2
(Thamaga and Dube 2019)	76.42	**94.44**	~5228	Sentinel-2 (wet)
	74.67	66.04	~5228	Sentinel-2 (dry)
(Mukarugwiro et al. 2019)	85	83	~18,180	Landsat-8
This study	87.48	92.98	**1,219,090**	Landsat-8

Roodekoppies dam and Sandvlei is shown (Figure 10) and was selected owing to the model's predicted occurrence of water hyacinth, the size (not too small waterbodies), and its location. Roodekoppies dam is near Hartbeespoort Dam that has a persistent water hyacinth infestation and is surrounded by agricultural land use. Sandvlei is surrounded by urban and commercial land use.

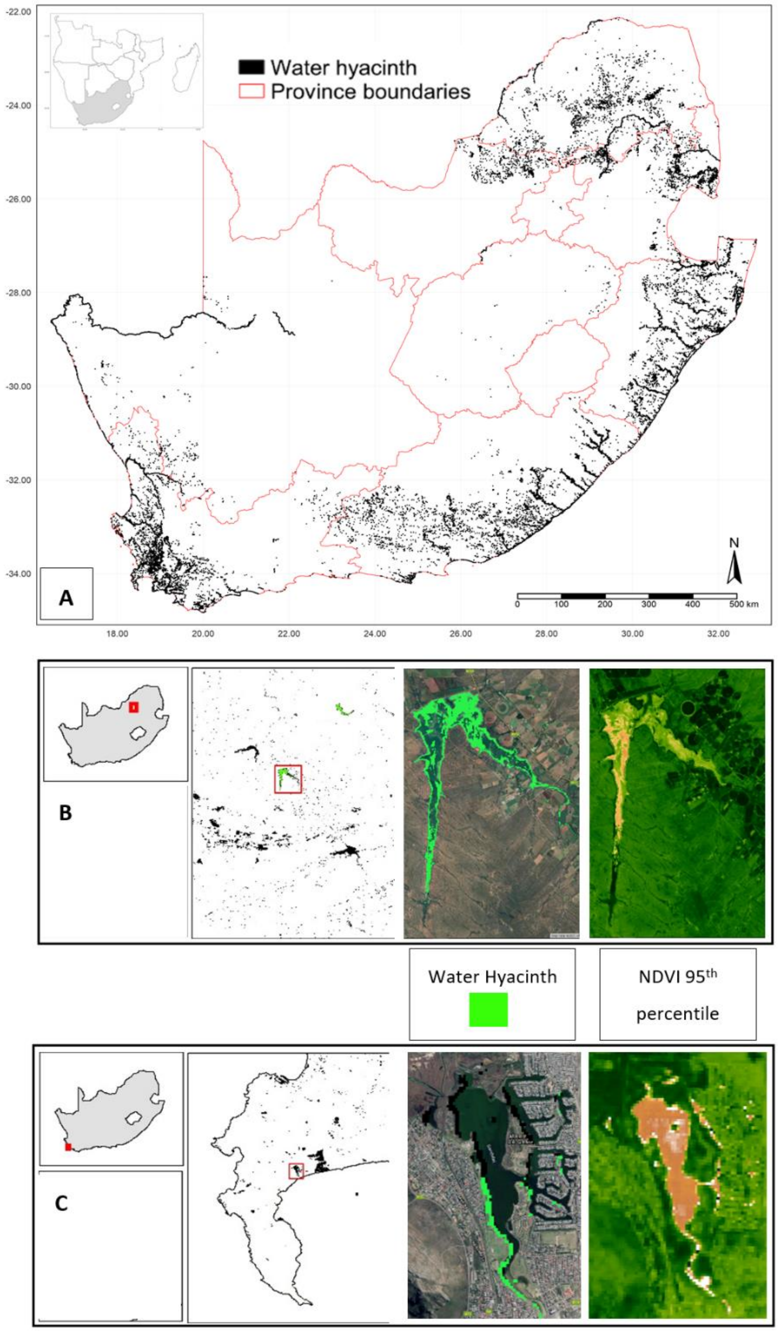


Figure 10. The water hyacinth pixel distribution for 2013 across South Africa (**A**), Roodekoppies dam (**B**) and Sandvlei (**C**) derived from a 95th percentile NDVI Landsat-8 image composite. The water hyacinth coverage corresponds to a near-maximum accumulated cover over 2013. The area of water hyacinth infestation is 417.74 km^2 or 0.03% of the total area of South Africa. This also corresponds to 2.69% of surface water (permanent and seasonal) within South Africa (15,552.72 km2 - based on the 2013 annual GSW data). Users can interactively explore this national distribution of water hyacinth here. (script name: 2013 water hyacinth distribution).

2.5. Discussion

This chapter demonstrates the benefit of utilizing a hierarchical classification approach to first map water, then aquatic vegetation and finally water hyacinth. Additionally, the challenges associated with the MultiResUnet method for large scale mapping is highlighted. The introduced hierarchical classification method relies on simple thresholding and allows for the near-instantaneous monitoring of water extent and aquatic vegetation cover over the long-term and across national extents. Subsequently, if water hyacinth infestations are closely monitored in large areas, this may assist in systematically evaluating the effectiveness of past and future management efforts. It is anticipated that this GEE workflow (Figure 2) can be applied to other countries infested with IAAPs in a cost-effective but easy and reliable manner.

2.5.1. Stage 1—Surface Water Detection

The introduced MNDWI thresholding approach requires a long-term 95th percentile MNDWI image and a slope layer derived from the SRTM elevation dataset for the detection of open-water from either a Landsat-8 or Sentinel-2 cloud-free image. When the open-water images are combined for a monthly, quarterly, or annual time-period, they allow for the open-water and water areas that may have been covered by aquatic vegetation, as described below, to be detected for the temporal window of interest. In the scenario that there is perennial aquatic vegetation cover, the use of the long-term (from 2013 to 2018) 95th percentile MNDWI image minimises the chance that a waterbody is missed during detection. If aquatic vegetation is present for more than 95% of a pixels' observations between 2013 to 2018 the water will be omitted

While the GSW products are an iconic dataset ascribed to its unique global and long-term coverage (1984–2018, (Pekel et al. 2016)), the introduced water detection approach provides several advantages over the GSW dataset. Both annual and monthly GSW products remain restricted to Landsat (30 m) and for the period extending from March 1984 to January 2018 (Pekel et al. 2016). While this dataset is updated, there is an update lag and a lack of publicly available update schedules. These restrictions limit the suitability of the GSW data for tracking water and aquatic vegetation on a current basis and at higher spatial resolutions when using Sentinel-2. Furthermore, the derivation of the GSW relies on an expert classification system, which has not yet been adapted for Sentinel-2. In addition to the expert classification system that relies on the SRTM dataset for terrain shadow removal, the GSW dataset depends on additional ancillary datasets (for example, glacier data and global human settlement layers (GHSL) for the removal of shadows cast by buildings, (Pekel et al. 2016)). Therefore, there is increased computation associated with handling these datasets and an increase in the propagation of errors within the final GSW dataset. In contrast, the introduced water detection approach only requires the SRTM layer and has been shown to be applicable to both Landsat-8 and Sentinel-2 data. Moreover, it can be used to monitor the current water dynamics of individual sites near-instantaneously and over multiple years owing to its computational efficiency (from 2013 for Landsat-8 and from 2015, within Europe, for Sentinel-2 or December 2018 outside Europe).

Other water extent datasets provided by, for example, OpenStreetMap are unsuitable because of its inability to capture seasonality in water extent. Historical OpenStreetMap data is irregularly updated and reflective of contribution effort as opposed to seasonal dynamics (Neis and Zielstra 2014; Witt, Loos, and Zipf 2021). Many other available water detection approaches have only been evaluated for individual satellite scenes in time and/or space (for example, (Acharya et al. 2016; Qiao et al. 2012; Wang et al. 2018)) and therefore show a limited generalization capacity when transferred to non-study time periods or areas. For example, supervised learning approaches are computationally expensive and sensitive to the quality and quantity of training samples, which may be difficult and costly to obtain. Our MNDWI thresholding approach overcomes these limitations and can generalize across time for the same area or to different areas while remaining simple to execute. This can be attributed to the nature of the thresholding method, i.e., it does not require labelled data or model fitting.

The moderate IoU score (0.695) between the 2013–2015 GSW layer (reference) and the water layer derived using this method may partly be attributed to the misalignment issue of the GSW data - due west across South Africa. Furthermore, omission and commission errors in both water products may have also contributed to the IoU score. For the GSW product, an omission error of less than 5% was reported, however, this was based on 40,000 reference points distributed globally (Pekel et al. 2016). Since the points are not a comprehensive evaluation of all pixels, it is likely that there are omission and commission errors that are not accounted for by the points. This is supported by the GSW omission errors reported in recent studies that compared satellite detected water from higher resolution data (Wu et al. 2019; Yang et al. 2020). Wu et al. 2019 (Wu et al. 2019) have shown considerable omission errors in the GSW product when compared to 1 m detected surface water. Furthermore, Yang et al. (2020) (Yang et al. 2020), reported a 25% omission error in the GSW product when compared to Sentinel-2 derived water (10m) for France. Lastly, based on a comparison with the 2018 SANLC dataset (Figure 5), there is a much higher omission error compared to the commission error for both the GSW dataset and the introduced dataset. These studies and results highlight the role of spatial resolution and the difficulty in mapping smaller seasonal surface water features that account for most misclassification errors between the GSW water product and the water layer derived using the introduced approach. Despite the moderate IoU score and the difficulty in identifying smaller seasonal waterbodies, the introduced water detection method is valuable for current and near-instantaneous monitoring of large, less seasonal surface waterbodies that are of higher priority for IAAP species monitoring, especially when the GSW data are unavailable. For periods with readily available GSW data, the GSW product is recommended since it will not require additional compute.

To compensate for the seasonal dynamics of water and improve aquatic vegetation detection, a more frequent water detection product is required. While the GSW product provides a monthly dataset, it is highly sensitive to the presence of aquatic vegetation, ascribed to the relatively low 16-day temporal resolution of Landsat (Pekel et al. 2016). As a result, it is less suitable for detecting aquatic vegetation based on the introduced hierarchical classification method. For this reason, Sentinel-2, with a higher temporal resolution (up to 2 to 3 days), is preferable for the detection of aquatic vegetation.

2.5.2. Stage 2—Aquatic Vegetation Detection

Aside from the occasional artefacts present when using the BOA data, it also overestimates the detected aquatic vegetation compared to the TOA product. Moreover, the aquatic vegetation derived from the different indices using either the L1C or L2A data product shows a high agreement within a product (L1C or L2A) and between the different indices. The results suggest that the TOA data with atmospheric influences have improved water and aquatic vegetation separability and therefore outperform the BOA data for aquatic vegetation detection (Figure 6). A basic analysis of the histograms for TOA- and BOA-derived NDVI images that contained water and aquatic vegetation revealed a consistent pattern whereby the BOA NDVI had a greater range than the TOA-derived NDVI. This has been attributed to the adjacency effect and contrast degradation (Roy et al. 2016; Zhang et al. 2018). Contrast degradation as part of the adjacency effect refers to the introduction of systematic noise when there is a strong water reflectance contrast (Frouin et al. 2019), in this case, with aquatic vegetation. Overall, this leads to a larger number of water pixels being falsely detected as aquatic vegetation in the BOA data because of the increased spectral overlap with aquatic vegetation pixels. As a result, the TOA data outperform the BOA data for the detection of aquatic vegetation using the Otsu + Canny method.

Semantic segmentation is proposed as an alternative approach to detect aquatic vegetation. A semantic segmentation model trained to detect aquatic vegetation benefits from the ability to directly detect aquatic vegetation, without the need to first detect water and then detect aquatic vegetation, like the Otsu + Canny method. This is valuable since EO-based water detection is confronted by numerous complexities and inaccuracies that are propagated from ancillary layers, placing an upper bound on the accuracy of aquatic vegetation detection when using Otsu + Canny in a hierarchical classification framework. When an inaccurate water boundary is provided for aquatic vegetation detection, the Otsu + Canny method results in less accurate aquatic vegetation maps compared to those derived using the MultiResUnet model.

The achievements of CNN-based semantic segmentation architectures can be ascribed to these models automating the extraction and organization of discriminative features at multiple levels of representation (from edges to entire objects) exclusively from input data, without the need for additional user input (Zhu et al. 2017). This ability of CNNs to learn high level features takes advantage of both the spatial context of land-cover and the spectral features in an invariant manner—unlike pixel-based and object-based approaches (Long, Shelhamer, and Darrell 2015). Furthermore, traditional approaches require additional features to be manually engineered and can thus be time consuming to design and validate, especially for large areas (Wurm et al. 2019) and may still result in sub-optimal discrimination features. For example, there is no consensus on the most suitable spectral water index to detect water—this points to manual features that may perform sub-optimally depending on the waters' optical properties, study area location, and the varying quality and properties of satellite imagery (Y. Zhou et al. 2017). Secondly, the deep nature (numerous layers) of CNN models, which allow for a variety of simple to highly complex, hierarchical, and non-linear features to be learnt from large amounts of data, allow for improved model prediction power (Ball, Anderson, and Chan 2017). Owing to the increase in model prediction power and the associated increases in model complexity and computational requirements, CNN models generally require costly hardware, which may not always be available. Subsequently, the presented hierarchical classification scheme used to detect aquatic vegetation offers a cost-effective, scalable, and accurate but simple alternative to detect aquatic vegetation for large volumes of EO data.

In terms of the accuracy of either approach, no independent and publicly available datasets exist with which to compare the detected area of aquatic vegetation. However, based on the comparative visual inspection (Figure 7) and calculated accuracy of the MultiResUnet detected aquatic vegetation against the reference masks generated by the Otsu + Canny algorithm (Table 2), both approaches show very high correspondence with each other and their associated RGB image. Lastly, the generated aquatic vegetation masks may be combined with existing water products to alleviate omission errors arising from aquatic vegetation cover; for example, the GSW monthly water product (Pekel et al. 2016).

2.5.3. Stage 3—Species Discrimination

The use of the percentile spectral index bands (95th percentile GARI) together with the median seasonal bands (spring green band—50th percentile) assisted in capturing the phenological characteristics of water hyacinth and have been used to accommodate the intra-species variance in phenological patterns across large extents. Moreover, median images (50th percentile) have been shown to be valuable in reducing data volume while providing important discrimination information (Phan et al. 2020). Simultaneously, four of the eight final variables used in the final model included thermal related bands and highlight the importance of including environmental variables that take into consideration the target plant species' thermal physiology when discriminating between IAAPs using EO data. The growth of water hyacinth has been shown to be strongly linked to temperature, where the plant starts to brown and die back at temperatures below 10 °C (Byrne et al. 2010). Therefore, the importance of the minimum temperature (in the coldest month) feature aligns with known water hyacinth thermal physiology and further suggests that the model has learnt valid relationships. Note, the four thermal-related variables had a correlation below 0.8 between each other most likely because they represent different temporal windows. The minimum temperature feature relates to climate while the remaining three temperature variables relate to seasonal weather. Moreover, the minimum temperature feature that had a correlation coefficient greater than 0.8 with elevation, represents a broader temporal (> 30 years) and spatial scale (1 km resampled to 30 m) as compared to the brightness temperature bands (seasonal median bands at 100 m resampled and provided at a 30 m spatial resolution). Note, the correlation coefficients are based on the entire training and validation dataset partitions.

It is difficult to reliably compare the water hyacinth discrimination results of this study against previous studies owing to the different goals and evaluation approaches. For example, previous studies aimed to generalize within fewer study sites distributed across smaller study areas (Dube et al. 2017b, 2017a; Thamaga and Dube 2018b, 2019), whilst this study aimed to generalize across different study sites distributed across a large study area. In addition, most previous studies (Dube et al. 2017b, 2017a; Thamaga and Dube 2018b, 2019) aimed to discriminate surrounding terrestrial land cover and water from water hyacinth and *Cyperus papyrus* (Mukarugwiro et al. 2019), whereas this study focused on discriminating water hyacinth from other aquatic vegetation. As a result of the easier classification goal, our method was able to map water hyacinth at superior accuracies for at least one accuracy metric (user accuracy or producer accuracy), despite the study area being 67 times greater than previous attempts (Table 3). However, there are still commission and omission errors (Figure 8) that are likely attributed to limited and/or noisy training data that led to confusion with dense algal mats, other IAAP species or riparian vegetation, especially in narrow river systems. This is the main reason that may then lead to any scepticism or concern of the water hyacinth distribution by users. Simultaneously, it is important to recognise that the pre-existing data locality data was not collected for the purpose of validating a satellite-derived map and as such some localities may have had infestations that were too small to be detected by Landsat-8, despite using the accumulated aquatic vegetation cover for the entire year of 2013. This, together with geolocation inaccuracy resulted in 81% of the field sites from the initial dataset not being used during model calibration and validation. Despite this, the remaining 98 field sites used during validation were widely spread across the country and the infested pixels at the 98 sites were combined and used in a larger dataset. Moreover, the reported accuracy estimates are competitive with those reported in previous studies that discriminate and localise water hyacinth.

The higher sensitivity (0.85) and users' accuracy (87.48%) (in comparison to the MCC score (0.8)) suggests that the MCC is a conservative accuracy estimate. The user's accuracy for water hyacinth suggests that if a map user were to go to any of the mapped water hyacinth locations during 2013, there would have been an 87.48% chance they would have seen water hyacinth (Stehman 1997). This is also referred to as the reliability of the map. Sensitivity refers to the ability of the model to correctly identify an infested water hyacinth pixel. The variance in the accuracy estimates across the five folds suggest that the model performs better in some areas over others (figure 8B). Based on the commission error (12.52%), 52.30 km^2 are estimated to be falsely detected water hyacinth (~2.5 times the area of Hartbeespoort dam distributed across the country). If all the water hyacinth detected in the Northern Cape (20.64 km^2) are false positives, this is still within the reported commission error. The false positives (commission error) could also correspond to pixels that are at high risk of a water hyacinth infestation. For example, the Orange River infestation detected by the model, which is currently not known to be infested by water hyacinth, may instead be at a high-risk of infestation because of the extensive agriculture along the river and its susceptibility to incoming plant material from the upstream Vaal River that is known to contain water hyacinth (Cilliers et al. 1996). Simultaneously, commission errors are preferred over omission errors since the environmental and socio-economic costs of an infestation going undetected is assumed to be greater than allowing it to consolidate and expand. For this reason, false positives may represent sites with a high-risk of water hyacinth infestation and should warrant closer (but potentially more costly) monitoring.

2.5.4. Management Tools

The minimum accuracy requirements for a satellite-derived invasive species' distribution to be incorporated into a management system are unknown owing to the unavailability of these datasets at extents relevant for management and because they are in their early stages of development and adoption. Considering the inevitable errors and spatially variable performance associated with satellite-based invasive species mapping, the derived datasets should not be used naively or in isolation, instead they should be coupled with model uncertainty estimates (that can be accessed here), local expert knowledge on the occurrence of a weed within a region of interest, and more up-to-date information from, for example, the aquatic vegetation monitoring application that will increase the user's confidence in the predicted occurrences (Singh et al. 2020). Similarly, to increase the certainty of an infestation, prior to field visits to predetermined areas, the introduced tools should be combined with higher spectral and/or spatial resolution imagery to discern past dynamics, waterbody extent, and possible confusion with riparian vegetation, especially for river systems. The consideration of such tools and products in future management efforts needs to be evaluated against the cost of omission and commission errors. If a field visit results in a false positive (non-water hyacinth) infestation being identified, this information will be useful to improve future mapping efforts. At present there is no systematic way of identifying and collating information on new infestations and is left to chance or concerned citizens. Simultaneously, satellite derived distributions are currently the only plausible way to obtain consistent, large extent, regularly updated species distribution information. Moreover, national-global land cover maps have provided a wide-array of benefits despite having lower overall accuracies (close to 80%, in the best cases) than the introduced water hyacinth occurrence map (García-Álvarez and Nanu 2022; Venter et al. 2022).

The monitoring tools generated in this study are made available within GEE (here) and will allow users to track the extent of water and aquatic vegetation within a waterbody from 2013 to present using Landsat-8 (30 m) or when using Sentinel-2, from 2015, within Europe or from December 2018 outside Europe at 10 m (aquatic vegetation) and 20 m (water) (script name: Aquatic vegetation and water monitor). Users will also be able to generate a larger national extent overview of aquatic vegetation cover using Landsat-8 (script name: National Extent Vegetation Detection L8) or Sentinel-2 (script name: National Extent Vegetation Detection S2). Lastly, the 2013 species-level distribution of water hyacinth across South Africa has been made available as a raster layer within GEE (script name: 2013 water hyacinth distribution) together with the field locality data used for model calibration and validation. It is anticipated that these data products will assist in management prioritization, control measure implementation, and subsequently the tracking of past, present, and future (in)effectiveness of these control efforts.

2.5.5. User Guidelines, Caveats, and Limitations

Since the proposed water detection approach relies on a long-term MNDWI percentile image, the introduced approach remains most suitable to optical imaging systems with archived imagery. In addition, the duration between successive water maps may vary from two to three days under optimum conditions, i.e., cloud-free conditions and vegetation-free water sites. Or may decline to monthly, quarterly, or annually in the presence of clouds or stationary aquatic vegetation mats. Both these factors may hinder the detection of the actual water surface extent and therefore, the ability to accurately monitor water and aquatic vegetation dynamics. These are limitations experienced by other water detection approaches that utilize optical data for surface water detection (Mueller et al. 2016; Pekel et al. 2016). Generally, the more frequently water can be mapped, the better the accommodation of changes in water extent, and the greater the accuracy of the aquatic vegetation detection. For example, if the detected waterbody extent is less than the actual extent, the area of aquatic vegetation is likely to be underestimated, and vice versa. If the detected waterbody is greater than the actual extent, then riparian vegetation or previously inundated land may falsely be identified as aquatic vegetation.

Since a pixel-level classification was used to detect and discriminate aquatic vegetation from explanatory variables with a 30 m spatial resolution, an isolated 30 m pixel may be identified as water hyacinth and represents the area of the minimum detectable water hyacinth canopy for Landsat-8 (i.e., 30 m × 30 m). Given this, smaller (<30 m × 30 m) isolated mats have a lower probability of being detected. Smaller undetected mats are important for an improved chance of successful management. As a result, there may be delays in early detection of water hyacinth mats until the mats grow to an EO-detectable size. This limitation should be considered when using the introduced workflow to monitor smaller water hyacinth infestations. At the same time, it may be very costly to monitor water hyacinth at higher spatial resolutions and at a similar frequency of Landsat-8 or Sentinel-2. Currently, it is feasible to monitor large infestations with the available medium resolution EO data from Landsat-8 and Sentinel-2.

Species reference localities from the field are valuable for reliably assessing the accuracy of EO derived water, aquatic vegetation, and IAAP species distribution maps. However, existing field locality data is often less suitable for high spatial resolution distribution maps compared to other EO applications, e.g., habitat suitability modelling (Vorster et al. 2018). Field data may be clustered in some areas and sparse or not available in other areas (Figure 1). This is evident when comparing the EO derived water hyacinth distribution (Figure 10) against the field locality data (Figure 1). Other compatibility challenges that were encountered with the field locality data in this study (similarly encountered and outlined by Vorster et al., 2018

(Vorster et al. 2018)) included geo-registration errors (i.e., points were located outside the boundary of waterbodies) and/or opportunistic data capture errors (i.e., GPS points may have been taken outside waterbodies, on nearby roads). The inaccuracy effects of these points were minimized by linking them to waterbodies that were up to a maximum distance of 250 m away. Despite this, of the initial 513 locality points considered in this study only 19% (98) were suitable. Additionally, those field locality points that were recorded for moving IAAP mats could reduce the spatial correspondence between imagery and field data owing to the time lag between image capture and field data capture. By linking these field localities to entire waterbodies, this was avoided.

2.6. Recommendations and Conclusions

A hierarchical classification approach that enables the near-instantaneous monitoring of water and aquatic vegetation on a frequent basis is introduced. This approach has also allowed for the first EO-derived distribution of water hyacinth on a national extent. The opportunities introduced by the proposed workflow to circumvent labelling costs associated with CNN-based aquatic vegetation detection is also discussed. This would consequently remove the need for surface water masks that may not always be available at the required spatio-temporal resolution or accuracy. Whilst this chapter has shown encouraging results for the MultiResUnet-based detection of aquatic vegetation within a single heterogenous study area, the computational infrastructure required for the operational CNN-based mapping of aquatic vegetation across large extents remains a challenge. Therefore, this study suggests that the hierarchical classification approach is advantageous over a CNN based approach for the final goal of mapping water hyacinth since it greatly reduces commission errors in comparison to previous approaches.

Future work should aim to have a larger field-based validation dataset that could involve regular contributions from invasion ecologists, water managers and informed locals into a system that allows them to collaboratively label data with the assistance of satellite images, derived spectral signatures and AI suggested class assignments based on spectral similarity to expert identified infestations. This system may also prioritise labelling for high uncertainty (low probability) sites/pixels. For those areas that still exhibit low-confidence and that are out-of-distribution (ecologically and statistically) could be excluded during inference time to avoid erroneous detection. For example, areas in the South African Karoo. This live dataset may also provide a high-quality reference dataset that will enable a more consistent and reliable comparison of IAAP mapping methods and their associated results in a systematic manner across studies. The value of such a dataset may also translate to advances in the mapping of other IAAP species since our results, similar to previous studies (Everitt et al. 2009, 2011), highlight the spectral separability of water hyacinth from the other IAAP species that were represented in the non-target species class.

Lastly, the value of using species appropriate variables, such as minimum temperature, was shown to be highly valuable. While the map's utility in some areas is limited owing to misclassification errors (~20%), it still has applicability for other areas (~80%). Despite these shortcomings, the value of the derived distribution should be evaluated against the data we currently have, as we do not have a comprehensive and accurate water hyacinth occurrence dataset. The introduced water hyacinth distribution is more comprehensive, spatio-temporally scalable, cheaper to obtain and, in some cases, has a higher geolocation accuracy than the water hyacinth distribution data we currently have (SAPIA and the expert collected data used in this study). Overall, this study suggests that EO data is highly suitable for an IAAP monitoring programme. The recommendations made in this study are likely to be applicable to EO-based monitoring of invasive terrestrial alien plants.

3. Chapter 3: The Factors Influencing the Occurrence of Invasive Water Hyacinth: An Earth Observation and Interpretable Machine Learning Method.

3.1. Abstract

Despite a long history of efforts to manage water hyacinth, successful control of the weed has been variable. Moreover, the most appropriate management strategy is not always apparent and depends on the weeds' environmental and socio-economic context. This, together with the advent of numerous Earth Observation (EO) derived data products, and recent advances in explainable Artificial Intelligence (xAI) prompted an investigation into the environmental and socio-economic contexts that likely influence the occurrence of water hyacinth at multiple scales. This was achieved by using SHaply Additive exPlanations (SHAP) to interpret a Species Distribution Model (SDM) that modelled the response of water hyacinth to numerous biotic and abiotic features. The spatially explicit results for all South African detectable waterbodies can assist risk-prioritisation of sites and the selection of context-specific management strategies through a data-driven and evidence-based approach. There is a consensus between many of the modelled plant responses and known physiological constraints limiting water hyacinth growth, such as minimum temperature. In addition, the results may also stimulate future lab and field-based experiments that aim to validate and elucidate the underlying mechanisms of the suggested novel weed-environment relationships that show non-linear trends and reversed responses because of interaction effects.

Keywords: satellite, remote sensing, Google Earth Engine, explainable artificial intelligence, habitat suitability.

3.2. Introduction

Across tropical, sub-tropical, and warm-temperate regions (Coetzee et al. 2017; Julien, M.H. and Orapa 1999), the Invasive Alien Aquatic Plant (IAAP), water hyacinth *Pontederia* (previously *Eichhornia*) *crassipes* (Pellegrini et al. 2018) has frequently been associated with disturbed aquatic environments, subsequently disrupting ecological function and constraining ecosystem service provision (Vila and Ibáñez 2011; Villamagna and Murphy 2010). Water hyacinth was first recorded in South Africa during 1908. Since the 1960s, there has been a large investment into the development of water hyacinth management strategies (for example, (Bick et al. 2020; Miller et al. 2021; Tipping et al. 2020)). However, despite many efforts to eradicate and/or control the weed, water hyacinth was estimated to cover 417.74 square kilometres of South Africa during 2013 (Singh et al. 2020). This is partly attributed to management strategies that are typically reactive (Hill and Coetzee 2017), despite higher chances of successful invasive species management when small alien populations are targeted at introduction (Hussner et al. 2017; Mehta et al. 2007).

The development of a pre-emptive management strategy relies on understanding the ecological and socio-economic contexts that limit or promote water hyacinth's establishment and spread (Wilson, Holst, and Rees 2005). For the effective use of limited resources, this knowledge should be available in a cost-effective manner, at local to national scales. This will support data-driven decision making, which will underpin the pre-selection and prioritisation of different control strategies on a site-by-site basis (King 2011; Rainford et al. 2020). Moreover, the adoption of recent technological advances in Earth Observation (EO), cloud computing and machine learning may stimulate the generation of actionable insight from early IAAP incursion warnings and timely responses to new incursions across large extents of the plant's introduced range in a cost-effective manner.

IAAP management strategies centered around the early response to new incursions can benefit from EO for early detection (through surveillance), and early warnings (through modelling) (Bradley 2014; Huang and Asner 2009; Joshi et al. 2004; Kilroy et al. 2008; Müllerová et al. 2013; Zhang et al. 2018). While monitoring may provide an indication of new incursions and the actual IAAP distribution (Kilroy et al. 2008), modelling habitat susceptibility and habitat suitability for an IAAP may provide an indication of probable future incursions and areas that should be closely monitored. Well established habitat suitability models or Species Distribution Models (SDMs) represent the relative likelihood of an IAAP establishing should the species be introduced or dispersed to each location in the modelled landscape (Barbet-Massin et al. 2018; Briscoe Runquist, Lake, and Moeller 2021; Elith 2017). To model the combined risk of susceptibility and suitability, SDM's have been extended to include features that influence the introduction and spread of species (for example, distance to roads, and the presence/absence of other species as proxies for biotic interaction) (Kumar, Neven, and Yee 2014; Srivastava, Lafond, and Griess 2019; Wisz et al. 2013).

While a comprehensive Risk Analysis for Alien Taxon (RAAT) has been established (Kumschick et al. 2020), it is not suited towards fine-scaled spatially explicit risk quantification. Previous studies, that have investigated the risk of water hyacinth expanding its distribution under the varying climate change scenarios made use of mechanistic climate matching models (CLIMEX) (Kriticos and Brunel 2016) and correlative SDMs (Kriticos and Brunel 2016; Rodriguez-Merino et al. 2018). However, numerous challenges persist that limit the applicability of the results of these studies. For example, 1) a limited number of features are often used during modelling (Ahmed, Atzberger, and Zewdie 2020; Elith and Leathwick 2009),

2) presence-only data are used 3) model selection, feature selection, and (Elith and Leathwick 2009) model interpretability has largely neglected spatial autocorrelation and spatially explicit analyses of feature importance (Domisch et al. 2019; Elith and Leathwick 2009). This study therefore seeks to overcome these limitations by harnessing 1) numerous diverse EO-derived datasets and 2) recent advances in explainable Artificial Intelligence (xAI) while adhering to recommended SDM modelling guidelines for the exploration of spatially variable species-environment relationships captured by an SDM (Kass et al. 2021; Ryo et al. 2021).

Unlike field-based approaches, the use of EO-derived datasets coupled with recent advances in xAI are directly suited to multi-scale SDM analyses across large areas (for example, (Cha et al. 2021)). This can be ascribed to 1) free EO data policies along with their global, systematic and frequent acquisition (for example, (Grill et al. 2019; Kennedy et al. 2019; Pekel et al. 2016)), 2) the recent availability of free cloud computing infrastructure that has lowered the barrier for large-scale EO data analysis, 3) open-source libraries (Jordahl 2014) and 4) the availability of numerous fine-scale but large-extent EO-derived data products that represent ecological, hydrological, climatological, social, and topographical phenomenon across the earth's surface (Bradie and Leung 2017; Braunisch et al. 2013). These derived variables, together with species occurrence records represent inputs used to determine the environmental range of a species via SDMs (Bradie and Leung 2017). The popularity of (machine) learning-based SDM algorithms' can be attributed to their ability to model non-linear species responses to environmental gradients with high predictive performance, despite high dimensionality (Effrosynidis and Arampatzis 2021; Smith and Santos 2020).

SHaply Additive exPlanations (SHAP) is used (Cha et al. 2021) to identify and explore the 1) relative feature (comprised of biotic and abiotic variables) importance of probable water hyacinth drivers and, 2) SDM-captured species-environment responses across environmental gradients. The model-agnostic (i.e., applicable to any machine learning model) SHAP stem from Shapley values, a concept from game theory used to determine the contribution of each player (for example, an environmental predictor of water hyacinth) in a coalition, towards the (weak or strong) outcome of the game (reproducing the outcome of the model - the probability of water hyacinth occurrence). Here, each variable's (player's) contribution to the prediction (outcome) is evaluated (Lundberg et al. 2020). Because of its theoretical justification and analytical gains over other proposed xAI tools (for example, Local Interpretable Model-agnostic Explanations (LIME) and Mean Decrease in Impurity (MDI)), SHAP is preferred and increasingly being adopted for model interpretation across a variety of domains (Roscher et al. 2020), including ecology (for example, (Cha et al. 2021; R. Wang, Kim, and Li 2021; H. Yu, Cooper, and Infante 2020)). Some of the benefits of SHAP include local and global interpretations that align whereas other methods only provide either local or global explanations. Theoretically, SHAP benefits from the long standing and well understood economic value theory that identifies the factors that determine the distribution of income and can guarantee that feature importance is distributed fairly (Lundberg et al. 2020).

This study demonstrates, for the first time, the utility of SHAP, EO-derived data products and cloud computing made available through Google Earth Engine (GEE) and Google Collaboratory to; 1) determine the relative importance of biotic and abiotic factors that likely influence the occurrence of water hyacinth for a single waterbody and/or across multiple waterbodies at provincial to national extents, 2) capture the relationship between socio-economic and environmental factors, and the probability of water hyacinth occurrence, and 3) elucidate the influence of estimated interaction effects between variables on the probability of water hyacinth occurrence. Together, these contributions represent a low-cost desktop-based method to support risk management of potential incursions and an evidence-based approach to guide management strategy selection for existing invasions across large extents.

3.3. Methods

Four main steps were followed (Figure 1). These include: 3.3.1. water hyacinth occurrence data preparation, 3.3.2. feature selection and preparation, 3.3.3. model tuning and cross-validation and 3.3.4. model explainability. In addition, the details of the software used in the analysis can be found in section 3.3.5.

3.3.1. Water hyacinth occurrence data preparation

Based on previous work (Chapter 2) (Singh et al. 2020), the occurrence of water hyacinth (Figure 2) across South Africa was determined with an accuracy of 80% (16%) based on the Matthews correlation coefficient (MCC). Since this accuracy estimate suggests that that there are more correct instances (~80%) than incorrect (~20%) instances, it is assumed that the occurrence errors will be diluted by the correct instances and therefore, the analysis would still prove useful. The weeds distribution was derived from a combination of in-situ reference data and satellite-derived topographic, climate, weather, and Landsat-8 spectral variables (Singh et al. 2020). The distribution was mapped at a spatial resolution of 30 m and was found to cover 2.69% of the total (permanent and seasonal) surface water area within the country during 2013. The areas of main concern (high water area and high water hyacinth area) are areas in the western cape, north eastern KZN and areas along the northern Gauteng boundary (Figure 2). This distribution corresponds to the detection of 27 206 infested waterbodies within the Joint Research Commissions' Global Surface Water (GSW) dataset for 2013 (Pekel et al. 2016). These infested waterbodies were subsequently labelled as the positive class (1, n = 27,206). The remaining waterbodies were all labelled as the negative class (0, n = 221,164). The satellite-derived distribution was used instead of the underlying 98 field sites because of the dataset size (98 instances). In this case, the number of initial features (140) would have been greater than the number of instances used to calibrate the SDM, a challenging scenario for modelling that may lead to model failure, overfitting, and unreliable predictions (Johnstone and Titterington 2009; Kwon and Sim 2013). It is also important to note that Chapter 2 relied on each pixel in the 98 sites, whereas here we are interested in a site-level analysis and therefore summarise data at a corresponding scale.

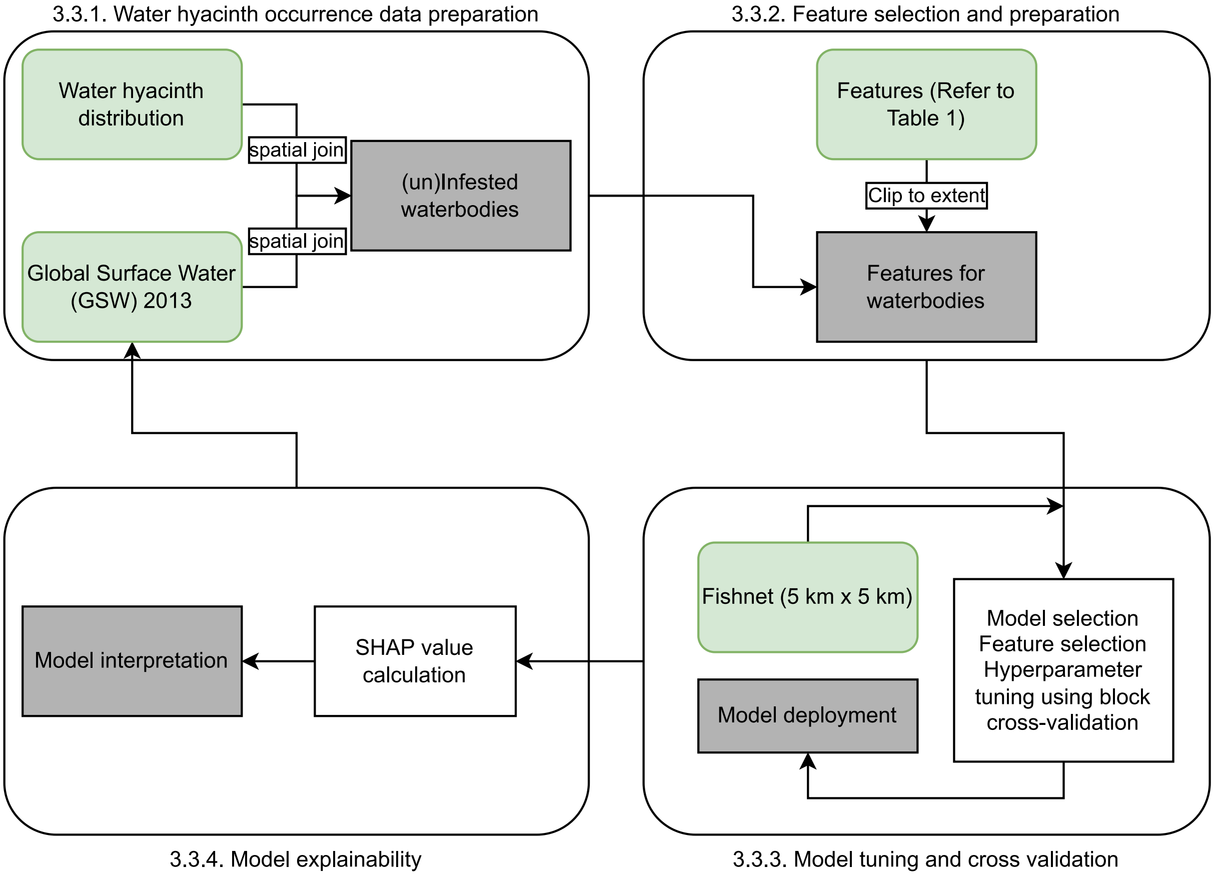


Figure 1. The general workflow used in this study to determine the relative importance of likely drivers of water hyacinth occurrence with inputs (**green**) pre-processing and modelling processes (**white**) and outputs (**grey**).

3.3.2. Feature selection and preparation

140 features were compiled for this study based on their public availability as EO data products for South Africa and known influence on water hyacinth. Table 1 provides a list of the EO derived products used as features, their temporal coverage, and sources. The 140 potential drivers can be grouped as follows (Note: each land cover category is considered as a single variable. 73 of the 140 variables stem from the South African National Landcover dataset). For reported links between driver and water hyacinth occurrence, refer to citation(s) below:

- Climatic: Temperature (Byrne et al. 2010; Miskella and Madsen 2019; Owens and Madsen 1995), precipitation, frost (Byrne et al. 2010), wind (Wilson et al. 2005), and topo-climate (Lawrence, Hoffmann, and Beierkuhnlein 2020; Miskella and Madsen 2019).

- Socio-economic: Artificial land cover (Essl et al. 2019; Westphal et al. 2008), human modification and wealth index (Hill 2003; VonBank et al. 2018).
- Ecological: Natural land cover, interspecies competition (Agami and Reddy 1990; Gopal 1987), distance to coastline (Bick et al. 2020; Coetzee et al. 2017; Muramoto, Aoyama, and Oki 1991), runoff, flood risk and riparian nutrients (Bick et al. 2020; Coetzee et al. 2017).
- Topographical: Elevation (Kriticos and Brunel 2016; Lawrence et al. 2020).
- Hydrological: River connectivity, stream power and water seasonality (Gopal 1987; Wilson et al. 2005).

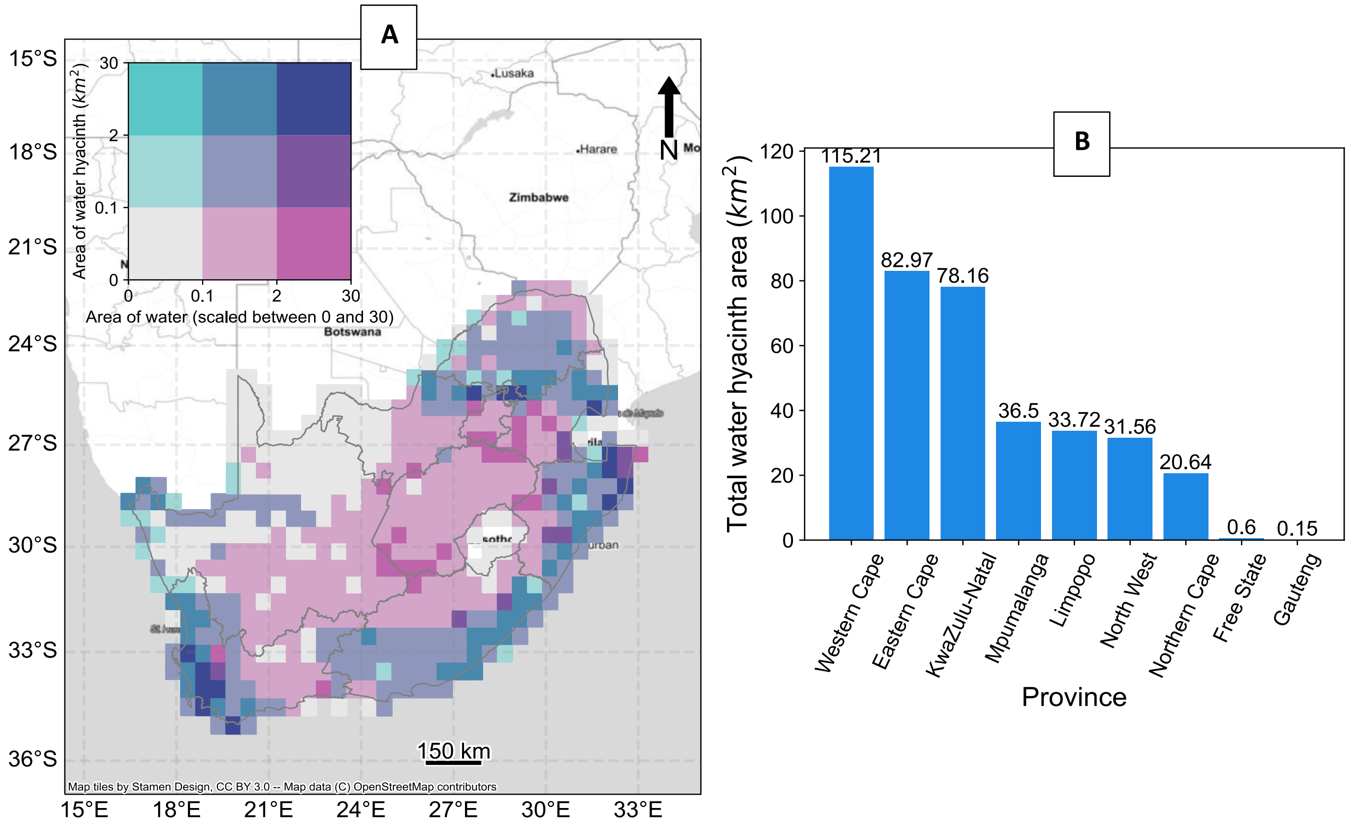


Figure 2. The relationship between the area of water (source: Global Surface Water 2013) and the area of water hyacinth across South Africa (source: Chapter 2, **A**) and per province (**B**). MinMax (min = 0, max = 30) scaling was applied to the water area to improve visualisation.

Some variables were not available for 2013, therefore the closest temporally available data was used instead. For those features that were not readily available within the GEE data catalogue or GEE community datasets, the datasets were downloaded (refer to Table 1 for data sources) for South Africa and uploaded to GEE.

Considering the large number of available environmental layers (n = 140), a standardized and reproducible procedure to select features to be used during modelling was implemented. The benefits of this include; 1) the reduced redundancy between features, 2) a more parsimonious and computationally efficient model, 3) the decreased the risk of over-fitting, 4) avoidance of a biased assessment of feature importance, 5) prevention of a high dimensionality from negatively affecting the model performance and 6) easier model interpretation (Chandrashekar and Sahin 2014; Guyon and Elisseeff 2003). The feature selection procedure consisted of three steps: first, the removal of unsuitable features based on low variance (3.3.2.1.), second, the removal of redundant features that were identified by an analysis of their correlation with all remaining features (Section 3.3.2.2.) and the removal of features with a low predictive power through Recursive Feature Elimination (Section 3.3.2.3.).

3.3.2.1. Step 1: *Removal of irrelevant features*

The number of consecutive nights that experienced temperatures less than 10°C and the number of upstream rivers were removed owing to low variance across waterbodies and therefore, their lower predictive power for modelling. Other irrelevant features such as snow and moss cover fraction from the global broad class landcover data were removed since they are not applicable to South Africa.

3.3.2.2. *Step 2: De-correlation analysis*

De-correlation analysis was carried out to reduce the redundancy of information between the remaining 136 environmental layers. Only features that had a low pairwise correlation coefficient <= 0.7 (Dormann et al. 2013) with all other features were included in the subsequent analyses. For each pair of features correlated above this threshold, a single feature was manually selected. Taking a manual selection approach allowed for the prioritisation of features with higher spatio-temporal resolution, future scalability, and availability during selection. The resultant features (n = 103) were passed to the final stage of feature selection.

3.3.2.3. *Step 3: Recursive Feature Elimination*

Recursive feature elimination and cross-validated selection (RFECV) is a method that has proven effective in selecting an optimal set of features for tree-based and linear models (for example, (Gomes et al. 2019; Pullanagari, Kereszturi, and Yule 2018)). The algorithm starts by fitting a model using all features, assessing model performance, and ranking the features importance. The least important features are then removed from the pool, and again the model is fitted, assessed, and the least important covariates removed. RFECV resulted in 82 final covariates for modelling as indicated (by a *) in Table 1. RFECV was preferred over other methods such as elastic net regularisation because it is appropriate for non-linear problems, unlike elastic net regularisation. Moreover, it allows for incorporating a block CV approach which accounts for spatial autocorrelation with a trade-off of computation time.

3.3.3. Cross-validation, model selection and model tuning.

Owing to the high spatial variation in (un)infested waterbody density and inherent spatial autocorrelation in the feature data, randomly splitting the (un)infested waterbody instances into k validation folds (groups) may result in a large difference between invaded and uninvaded instances (i.e., severe class imbalance) between folds and high levels of spatial autocorrelation within folds, and therefore inflated model accuracy (Meyer et al. 2019; Ploton et al. 2020; Valavi et al. 2019). Therefore, a block cross-validation (CV) strategy was adopted whereby the presence/absence observations are spatially grouped by 5 km blocks. These blocks were randomly assigned to one of ten validation folds. In this way, all instances (presences and absences) belonging to a block were always in the same fold (train or validation, not both) during model calibration and validation. This ensured a more balanced positive-negative class distribution and a lower level of spatial autocorrelation in each fold. During 10-fold block CV, training occurs iteratively on 9 folds while validation is carried out on the remaining fold. This is performed until each of the 10 folds are used as a validation fold and is hereafter referred to as block CV. Model selection, feature selection, hyperparameter tuning, model evaluation and model interpretation were all performed using block CV.

Precision, Recall, F1-score, Matthews correlation coefficient (MCC) and balanced accuracy are preferred metrics for the evaluation of models (refer to Figure 2 for a list of the 15 models evaluated) fitted to imbalanced datasets since these metrics are less sensitive to differences in the number of positive and negative instances i.e., class imbalance (Chicco and Jurman 2020). All metrics, except MCC (-1 to 1) range from zero to one, where a higher value is indicative of a more accurate model. Precision represents the ratio of true positives (correctly classified water hyacinth infestations) to total positives (correctly classified infestations and incorrectly classified infestations). Recall represents the ratio between true positives and the total actual positives (a combination of correctly classified infestations and incorrectly classified uninfested waterbodies). Precision is maximised when the cost of falsely classified infestations is greater than the cost of falsely classified uninfested waterbodies. In contrast, recall is maximised when the cost of falsely classified uninfested waterbodies are greater than falsely classified infested waterbodies. In this study, the precision score was selected as the metric to optimise during model selection and hyperparameter tuning since this corresponds to precautionary water hyacinth management i.e., the cost of missing an infestation is minimised but with a tradeoff of higher chances of false infestations being identified. In the scenario that both false positives and negatives have similar costs, F1 score should be maximised and if true and false positives and true and false negatives are of equal importance the MCC should be maximised.

During model selection, the 15 candidate models available in the pycaret python package were evaluated, and the best performing model selected (Table 2). The selected random forest model had the highest mean precision score (Table 2 and Figure 3). Random forest is an ensemble algorithm that endeavours to make an accurate classifier from multiple weak classifiers (decision trees) (Breiman 2001). This is done by randomly selecting multiple subsets of explanatory variables and using each subset to fit different decision tree models that each infer the presence or absence of a water hyacinth infestation. Thereafter, the inferred infestation status for a waterbody made by each decision tree are combined using a majority vote (mode infestation status). The optimal random forest model hyperparameters were evaluated using a sequential model-based optimisation (SMBO) strategy. During SMBO, the range of values for each hyperparameter is defined. This forms the search space from which hyperparameters are selected by a selection function. Those hyperparameters with the highest expected improvement are selected for the next trial. In this way, the hyperparameters that optimise model performance can be found faster.

3.2.1. Model explainability

To determine the presumed contribution of each predictor or a predictor value to the occurrence of water hyacinth, SHAP is applied to the entire dataset using the block CV strategy. This provides a (SHAP) contribution value for each predictor value in the dataset. SHAP values are measured in log odds and were not converted to probability scores owing to technical challenges i.e., setting the output mode to probability still returns log odds and a lack of consensus on the correct post processing conversion. However, a log odds score of 0 roughly corresponds to a probability of 0.5. while a log odds score of -5 or 5 correspond to probability scores of 0 and 1 respectively. SHAP values allow us to:

1. to determine if a feature increases or decreases the suitability of a waterbody for the occurrence of water hyacinth and quantify its contribution towards the occurrence of water hyacinth at that waterbody (Figure 4),

2. rank the overall importance of features for South Africa (Figure 5) or for cohorts (for example, provinces, Figure 6) through feature-wise aggregation and,

3. visualise the response of water hyacinth occurrence to a feature and visualise interaction effects between features (Figure 7).

Table 1. Feature descriptions, associated units, temporal coverages, and data sources of the features considered to investigate the likely drivers of water hyacinth occurrence. All explanatory variables were downloaded at a 30 m spatial resolution. Those variables that were only available at a coarser spatial resolution (> 30 m) were automatically resampled to 30 m using bilinear interpolation within Google Earth Engine.

Variable	Units	Temporal coverage	Source
Number of consecutive nights with less than 10 degrees Celsius	days	2013	MODIS LST- night
Area of waterbody	m^2	2013	(Singh et al. 2020)
Median river width at mean discharge	m	1984-2018	Global River Width from Landsat (GRWL) (Allen and Pavelsky 2018)
*Area of aquatic vegetation	m^2	2013	(Singh et al. 2020)
*Minimum Temperature in the coldest month between1970 and1990	°C	1970 - 2000	WorldClim
Connectivity- upstream and downstream river count	river count	2000	WWF hydrosheds (Grill et al. 2019)
*Water seasonality- number of months water is present	month count	**1984 - 2019	JRC GSW (Pekel et al. 2016)
*Total precipitation	mm	**1958 - 2021	TerraClimate
*Distance to nearest coastline	m	2009	NOAA (https://oceancolor.gsfc.nasa.gov/docs/distfromcoast/)
*Distance to roads	m	1979 - 2010	Global Road Infrastructure Project (GRIP) (Meijer et al. 2018)
Elevation	m	2000	NASADEM
*Continuous Heat Insolation Load Index (0=cool, 255=warm). Surrogate for effects of shading and topographic insolation	0-255	2006 - 2011	SRTM, (Theobald et al. 2015)
*Global Human Modification (1= high modification)	0 - 1	2016	(Kennedy et al. 2019)
*Topographic diversity- Surrogate for the variety of temperature and moisture conditions available to species as local habitats	-1323 - 8.81	2006 - 2011	SRTM, (Theobald et al. 2015)
Frost duration - Median duration of frost	day count	2007	(Schulze and Maharaj 2007)
Number of days below 0°C (0-365)	day count	**2000-present	MODIS- LST, night temperature

Number of days below 10°C (0-365)	day count	**2000-present	MODIS- LST, night temperature
*South African National Land Cover (73 class)	km^2	2013-2014	GeoTerraimage
*Mean Annual Runoff	mm/year	2005	Strategic Water Source Areas (SWSA)
*Broad (10) class Landcover	% (coverFraction)	2015	Corine Landcover
*Flood hazard with a 10-year return period	water depth (m)	2016	Joint Research Commission (Dottori et al. 2016)
Riparian soil nitrogen (5km buffer)	g/kg	1905 - 2016	Soil grids
Riparian soil ph (5km buffer)	pH	1905 - 2016	Soil grids
*Riparian soil carbon (5km buffer)	g/kg	1905 - 2016	Soil grids
*Stream Power Index (SPI)	-1 - 1	1987-2017	GeoMorpho90 GeoMorphometric layers (Amatulli et al. 2020)
*Relative wealth index	-1 - 1	2001-2019	Facebook (https://dataforgood.fb.com/tools/relative-wealth-index/)
13 iSDA Soil layers (includes *total nitrogen, etc.)	various	2013-2019	(Hengl et al. 2021)
*Mean wind speed for a 10-year period	m/s	2008-2017	Global wind atlas

* Included in final model, ** Multi-temporal data - 2013 or temporally closest data was selected

Table 2. Evaluation metrics for 15 candidate models calculated based on a 10-fold block cross-validation strategy and sorted by precision. The performance scores of the models are reported after feature selection but prior to hyperparameter tuning. The highest score for each metric is highlighted in yellow (refer to Section 3.4.1 for metric descriptions).

Model	Recall	Prec.	F1	MCC	Balanced accuracy
Random Forest Classifier	0.5823	0.8269	0.6831	0.6621	0.7832
Extra Trees Classifier	0.5827	0.8182	0.6801	0.6580	0.7829
CatBoost Classifier	0.6252	0.7906	0.6981	0.6696	0.8018
Gradient Boosting Classifier	0.4338	0.7737	0.5555	0.5417	0.7086
Extreme Gradient Boosting	0.6100	0.7725	0.6815	0.6511	0.7933
Light Gradient Boosting Machine	0.5762	0.7660	0.6573	0.6276	0.7766
Ridge Classifier	0.1760	0.6941	0.2798	0.3134	0.5828
Ada Boost Classifier	0.4433	0.6757	0.5351	0.5017	0.7077
Logistic Regression	0.3094	0.6272	0.4136	0.3930	0.6426
Decision Tree Classifier	0.6127	0.6009	0.6065	0.5547	0.7798
Linear Discriminant Analysis	0.3869	0.5979	0.4691	0.4286	0.6764
SVM - Linear Kernel	0.4365	0.4695	03901	0.3464	0.6662
K Neighbours Classifier	0.2698	0.4636	0.3409	0.2923	0.6145
Quadratic Discriminant Analysis	0.5887	0.4151	0.4865	0.4148	0.7402
Naive Bayes	0.5244	0.3852	0.4438	0.3643	0.7075

The mean SHAP value across all sites represents a measure of overall feature importance (Figure 5). Features are listed in the order of global (overall) importance, from top to bottom. Each point in a SHAP summary plot (Figure 6) represents a SHAP contribution value towards the probability of water hyacinth occurrence at a site.

A SHAP dependence plot (Figure 7a, b) consists of points that represent a variable's SHAP value for a given site. Dependence plots provide information on the variance of feature importance in response to changes in the feature's value. One of the benefits of SHAP dependence plots (Figure 7) over traditional partial dependence plots is the ability to provide insight into the magnitude of the interaction through the vertical dispersion of the scatter plot at a given feature value. The points are coloured by the feature that explains the most variation among the remaining modelled features.

3.2.1. Implementation details

All analysis were done using python-based packages and/or the GEE python Application Programming Interface (API). The environmental features were automatically extracted (in batches) for all GSW detected waterbodies within South Africa using the Google Earth Engine python API. This was facilitated by the geemap package (Wu 2020). The scikit-learn package was used for RFECV feature selection. The pycaret library was used during model selection (Ali 2020). Hyperopt was used for hyperparameter tuning (Bergstra et al. 2015). The shap and fasttreeshap packages were used for model explainability (Lundberg et al. 2020). All figures and maps were created using matplotlib (Barrett et al. 2005), seaborn (Waskom 2021), geopandas (Jordahl 2014) and contextily (Arribas-Bel 2021) for basemaps.

3.3. Results

3.3.1. Model evaluation

Model accuracy has implications on the reliability and consistency of model explanations between different xAI methods (Liu and Udell 2020). There is increasing alignment between different xAI methods as accuracy increases. The model shows overall good performance (Figure 2, F1 score > 0.7, as defined by (Liu and Udell 2020)). However, the model does not perform as well based on its Matthews correlation coefficient (Figure 2, MCC, mean = 0.688) since the model finds it difficult to correctly classify uninfested sites that have water but do not have water hyacinth based on the suite of features considered in this study. Moreover, MCC shows more variability (narrow and long blue area) in contrast to the other metrics. The model precision is the highest scored metric (Figure 2, mean = 0.890), corresponding to 89.0% of water hyacinth infested sites being correctly identified. This is advantageous when trying to understand the occurrence (as opposed to the absence) of a species. The high precision score and high false negatives are a result of optimising the model for precision.

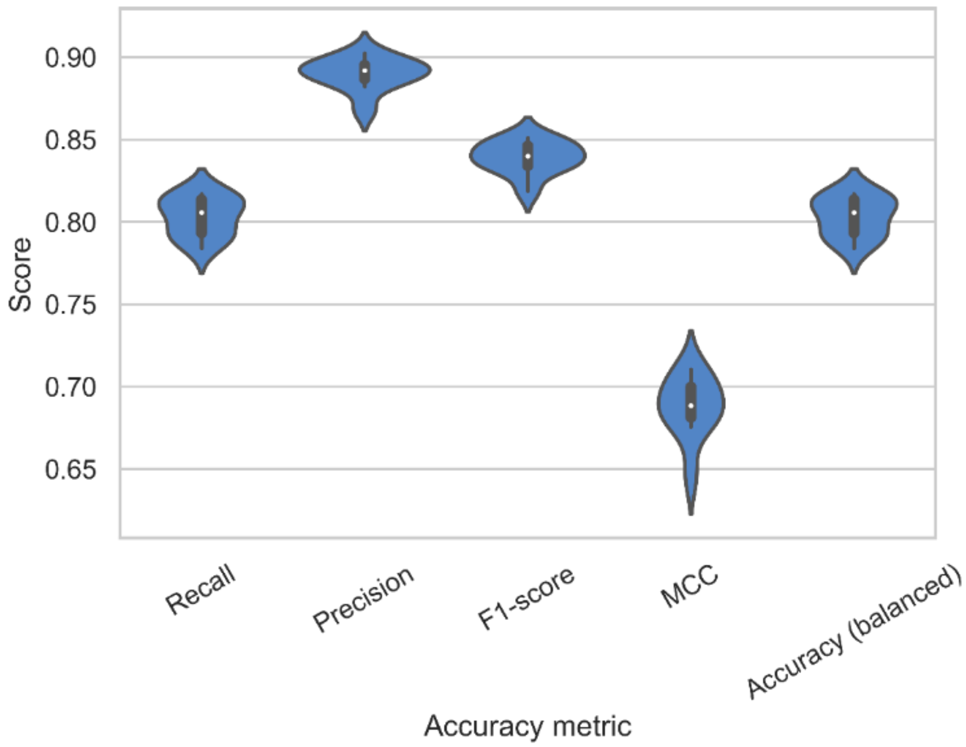


Figure 3. The distribution (blue) and boxplots of the random forest model evaluation metrics, post-hyperparameter optimisation. Scores are based on 10-fold block cross-validation. The high precision and F1-score suggest good (F1> 0.7) model performance.

3.3.2. Local (per waterbody) interpretation

To determine the contribution of each of the 82 selected variables to the occurrence of water hyacinth at a site, a waterfall plot was used (Figure 4). In this example, the water hyacinth infested Roodekoppies dam has a probability of 0.517 in water hyacinth occurrence, a value much greater than the expected base value (0.114) i.e., the average probability across all sites in South Africa. The minimum temperature (2.9 °C) in the coldest month and the persistence of surface water (surface water present for more than 9 months) are the two main factors promoting (dark pink) the occurrence of water hyacinth at Roodekoppies dam. In contrast, the near-average precipitation (474.21 mm) (when compared to average rainfall for South Africa – 463.42 mm (World bank group 2021) reduces the suitability of the dam for the occurrence of water hyacinth. Roodekoppies dam was selected owing to its use in the previous chapter, the size of the waterbody (not too small) and its location. It is downstream of Hartbeespoort dam and is surrounded by agricultural land use.

3.3.3. Global feature importance

Sixty of the 82 selected features (out of an initial 140 features) (>72%) used during model fitting are related to land use and land cover (LULC, Figure 5). This highlights the importance of the surrounding land cover in influencing the occurrence of water hyacinth. A large portion (38 features, 63%) of the 60 LULC features are related to human modified LULC. However, the best (top-most) predictor of water hyacinth presence is the minimum temperature (Figure 4). In addition to the commonly acknowledged critical role of low temperatures on limiting water hyacinth distribution (Byrne et al. 2010; Gettys, Haller, and Petty 2014), insights into less considered drivers of water hyacinth are highlighted (Figures 5 and 6). For example, surrounding (5 km buffer) shrub cover is more important than tree and grass cover (Figure 5). Water systems associated with low topographic shading effects, heterogenous temperature and moisture conditions, as indicated by low Continuous Heat Insolation Load Index (CHILI) and high topographic diversity values, improve the suitability for water hyacinth occurrence (Figure 5).

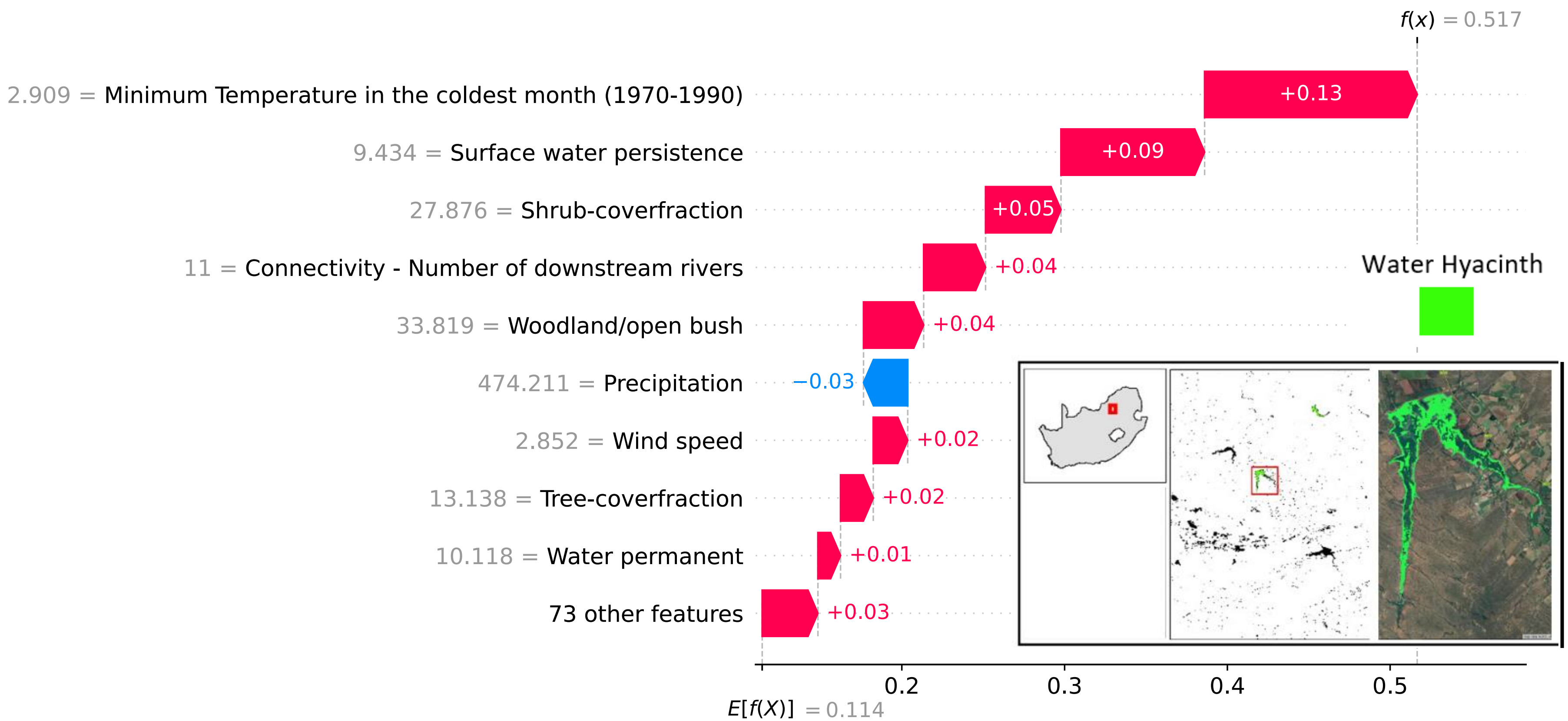


Figure 4. The main factors that influence the occurrence of water hyacinth at Roodekoppies dam, South Africa. Grey values to the left of the y axis labels correspond to actual feature values for the dam. The length of a bar corresponds to the contribution of a feature towards the prediction (i.e., the probability of occurrence, f(x) = 0.517) and away from the average probability across all sites (E[f(x)] = 0.114). Longer bars translate to a larger contribution towards the model outcome/prediction. Note, the near-average precipitation (474.21 mm) (when compared to average rainfall for South Africa – 463.42 mm (World bank group 2021) recorded for the dam is the only feature that reduces its suitability (**blue**) for water hyacinth. The 2013 accumulated water hyacinth cover is shown in the map inset.

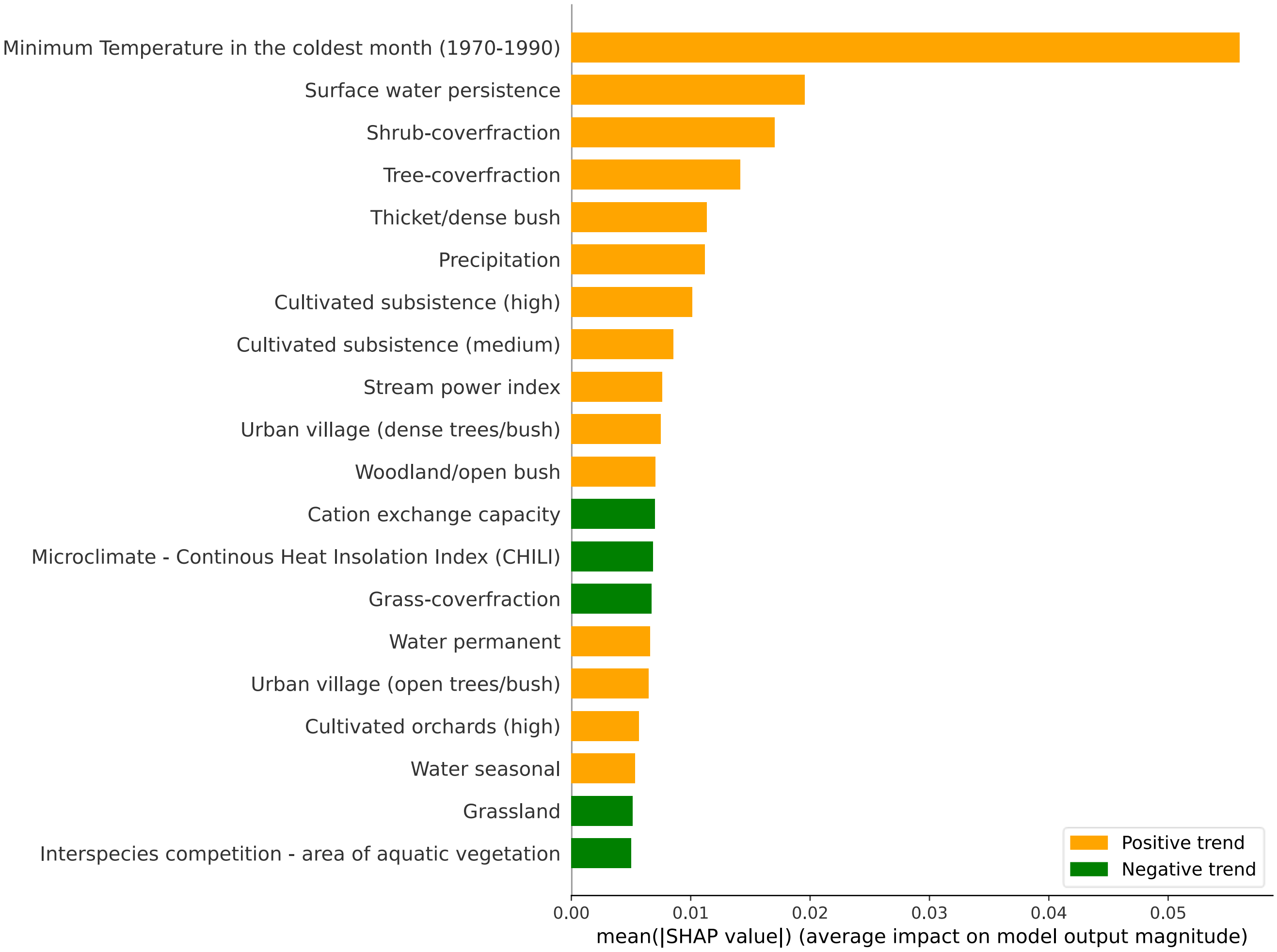
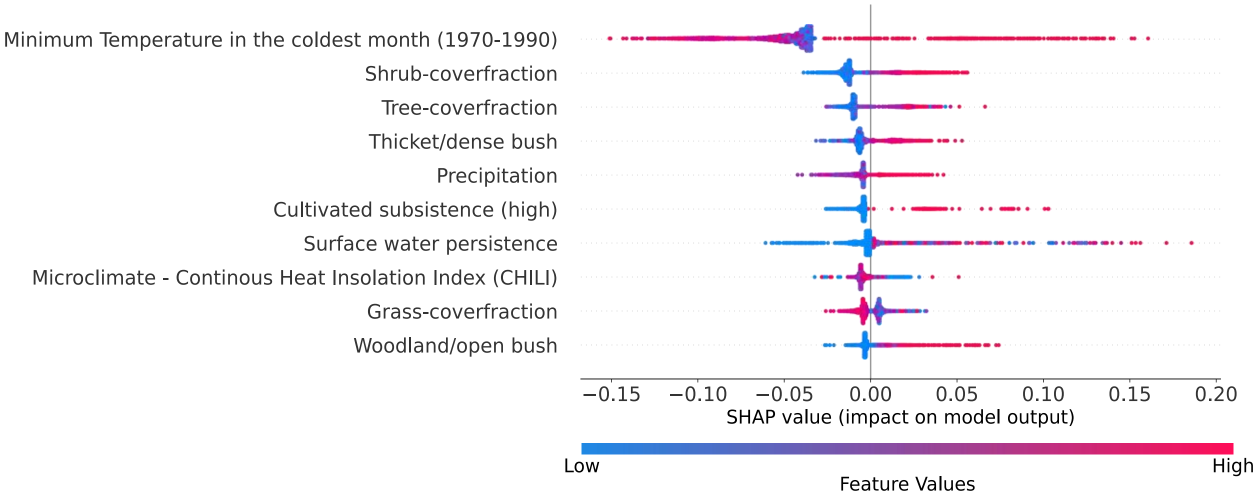


Figure 5. Feature importance for the top 20 features used to classify the occurrence of water hyacinth, sorted by their mean contribution (absolute SHAP values) to predict water hyacinth presence across South Africa for 2013. Features show either a positive (**orange**) or a negative effect (**green**) on the probability of water hyacinth occurrence**.**

The differences in ranking, of overall feature importance, between the two South African provinces reinforce the need to consider the social and ecological contexts that influence the probability of successful management (Figure 6). Within the Western Cape, for example, surface water persistence is the second most important factor determining the occurrence of water hyacinth while in the Highveld region of Gauteng, it ranks lower at seventh position. Similarly, tree cover fraction was identified as the third most important feature in Gauteng while, in the Western Cape tree cover fraction was ranked at ninth position. The Western Cape corresponds to the province with the highest water hyacinth area whilst Gauteng corresponds to the province with the lowest area of water hyacinth.

A



B)

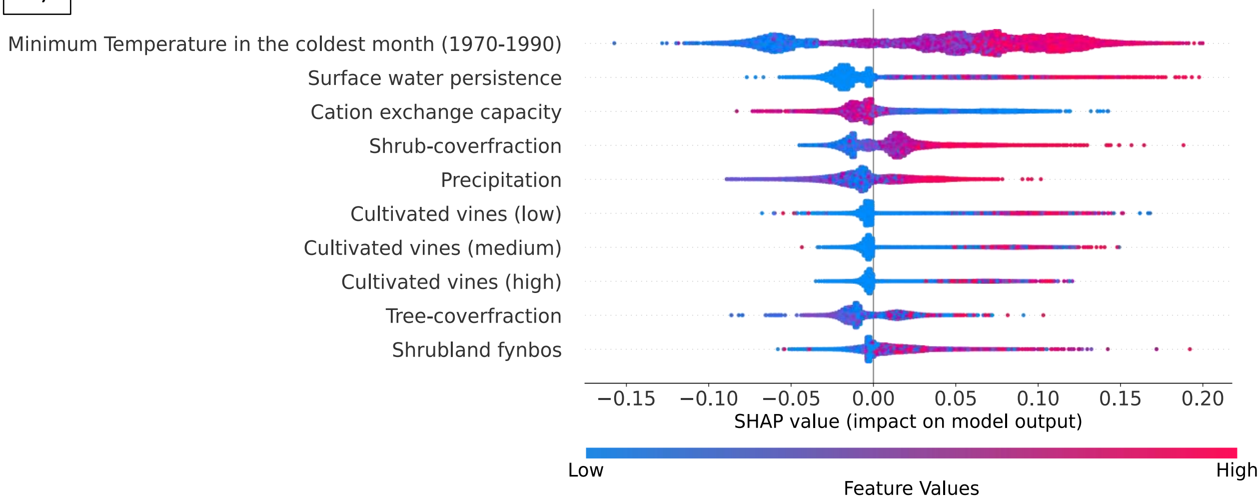


Figure 6. SHAP summary plots for the top 10 features and their direction (positive or negative) of effect for the Gauteng province (**A**) and the Western Cape province (**B**) in order of importance, from top-to-bottom, for the prediction of water hyacinth presence. Each point corresponds to an individual site. High values (**pink**) to the right and low values (**blue**) to the left of the chart indicate a positive trend while high values (**pink**) to the left and low values (**blue**) to the right of a chart indicate a negative trend with water hyacinth occurrence. The overlapping points are displaced along the y-axis to give a sense of the distribution of the SHAP values per feature. Each point represents one site and the gaps between points are indicative of a province not experiencing those environmental conditions.

3.3.4. Feature dependence and interactions

The SHAP dependency plots are useful to identify species-environment responses, inflection points that correspond to abrupt changes in feature importance and interactions between factors that drive water hyacinth occurrence. For example, the abrupt changes in the minimum temperature feature importance at ~ 2.5 °C and 5 °C both favour the occurrence of water hyacinth. The inflection point at 2.5 °C is a more important limiting factor for determining the presence/absence of water hyacinth i.e., the SHAP values exceed 0 at temperatures greater than 2.5 °C. At the former threshold (~ 2.5 °C), high shrub cover fraction is more important to promote water hyacinth occurrence compared to the latter inflection point (~5 °C, Figure 7a). Surface water persistence has a gradual positive trend with the occurrence of water hyacinth (Figure 7b). Surface water persistence (values along the x axis) combined with warmer temperatures (pink-red colours) increases the suitability of a water system for water hyacinth. For shrub cover, percentages greater than 15% within the 5 km surrounding a waterbody promote the occurrence of water hyacinth (Figure 7c).

3.3.5. Spatial distribution of a features' importance

To capture the spatial distribution of a variable's importance, a single variables' SHAP values were aggregated at a 5 km block level (Figure 8) by computing the mean SHAP value of the variable for that block. Low temperatures are associated with a lower probability of water hyacinth occurrence. Moreover, the odds of water hyacinth occurrence have a much more abrupt change (Figure 8b) in relation to the temperature gradient (Figure 8a). The unsuitability of the interior of the country, based on minimum temperature, can also be noted (Figure 8b).

Within the interior of the country, there is greater variability in the most dominant factor that determines the occurrence of water hyacinth in comparison to the coastal regions (Figure 9a). The warmer temperatures (as indicated by higher minimum temperature values) is most important in the coastal regions, while natural land cover is more important for the interior regions of the country. The hydrological and natural land cover categories are the dominant categories of drivers influencing the occurrence of water hyacinth in the Free State (Figure 9). This is attributed to the tree, shrub cover and precipitation variables.

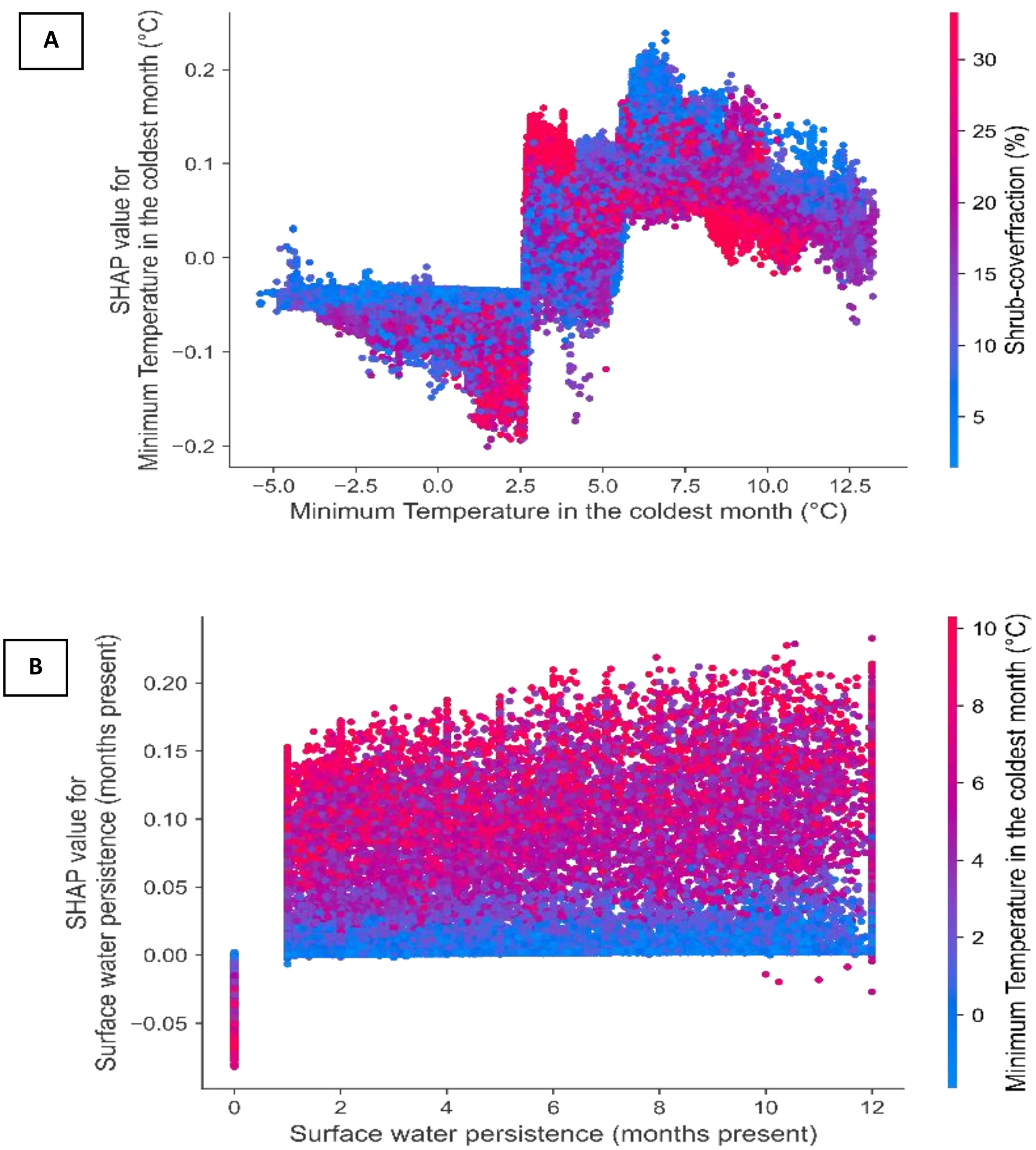


Figure 7. SHAP dependence plots for the top two predictors of water hyacinth with feature interactions (**blue - red)**. Minimum temperature (**A**) refers to the coldest temperature in the coldest month. Water persistence (**B**) indicates the average number of months the pixels of a waterbody contain water.

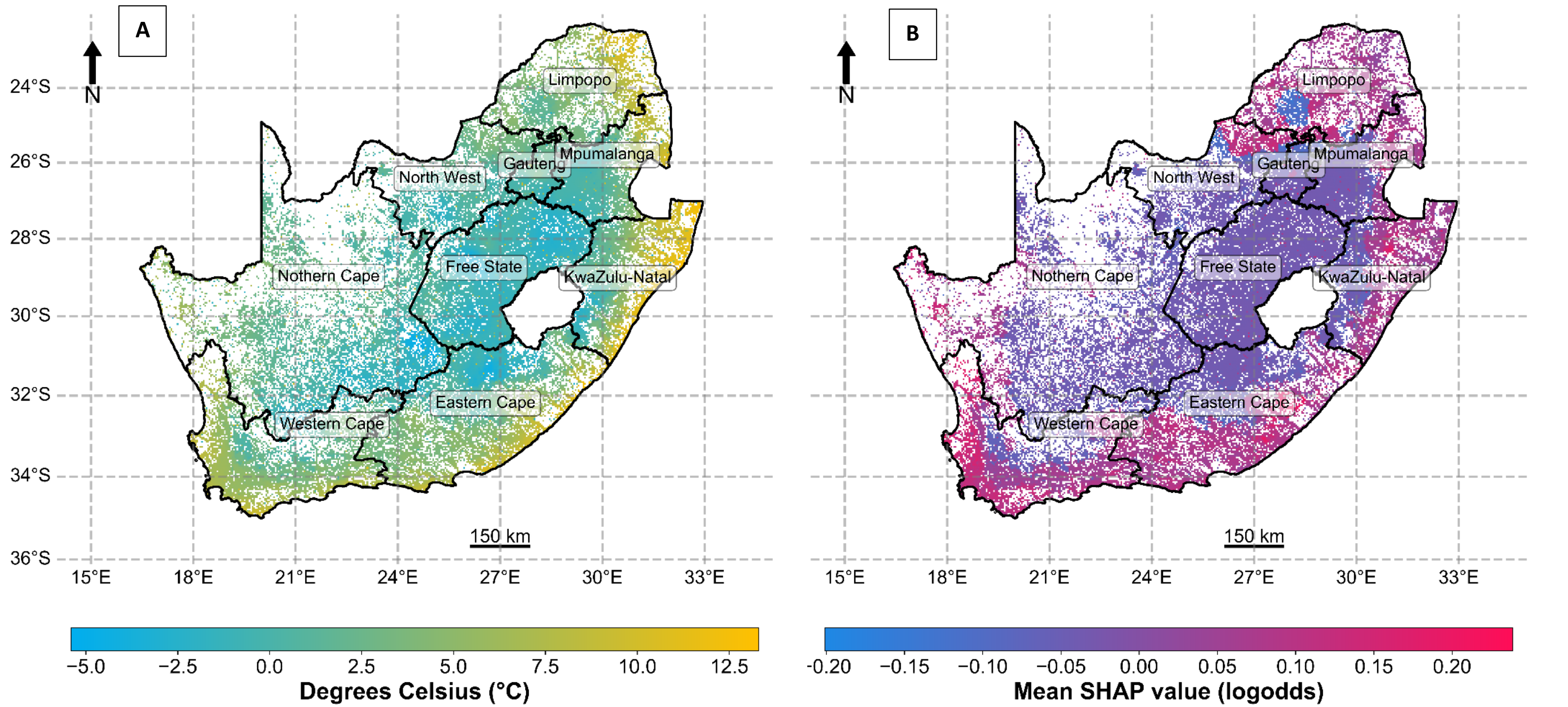


Figure 8. The distribution of minimum temperature (**A**) - the top national predictive feature of water hyacinth occurrence and the SHAP values for minimum temperature (**B**) across South Africa. SHAP values greater than 0 correspond to areas where less extreme minimum temperature promotes the occurrence of water hyacinth (limited mainly to coastal areas). While the interior of the country experiences extreme cold temperatures reducing the suitability for water hyacinth.

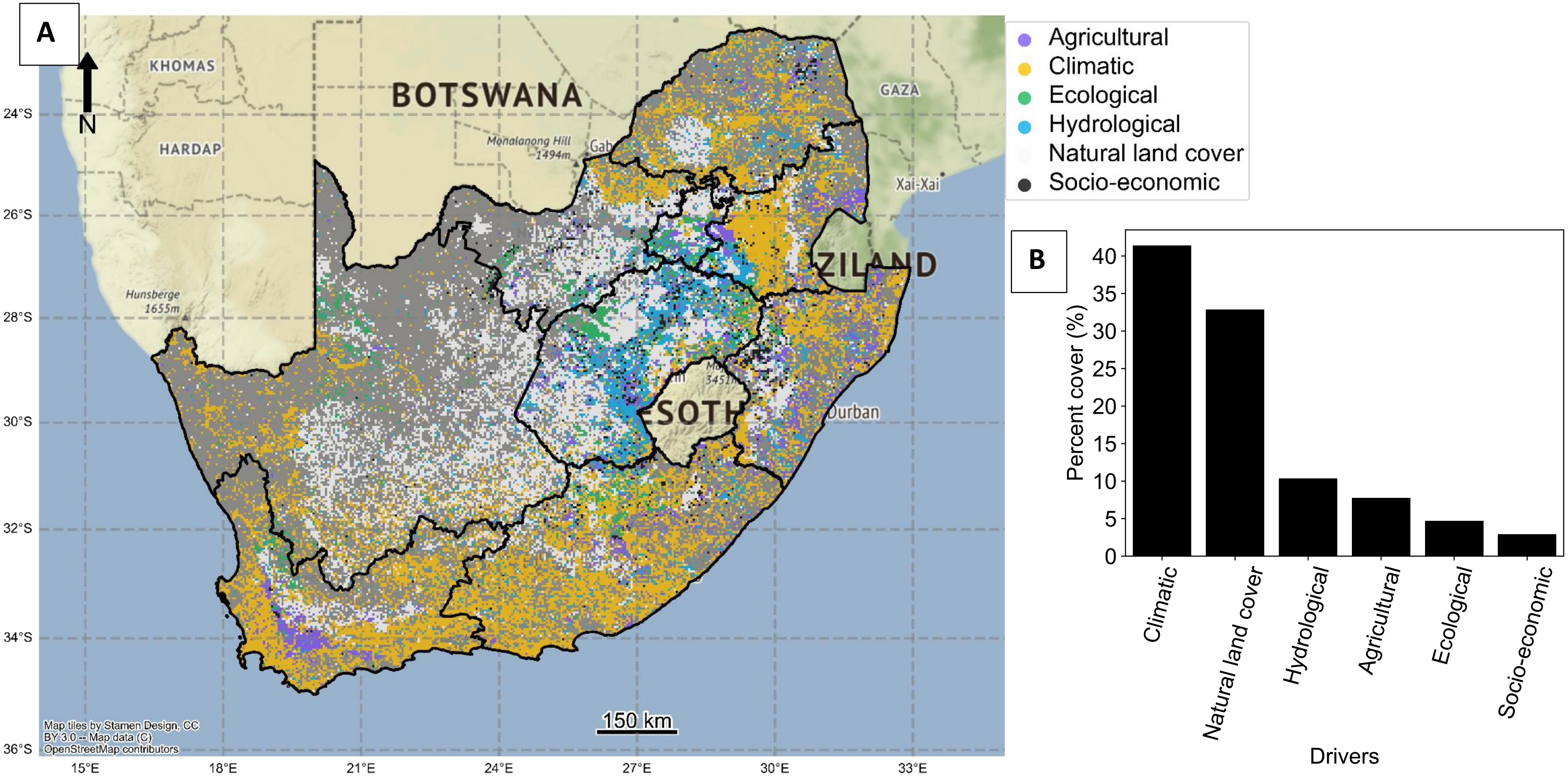


Figure 9. The distribution of the most important feature (feature with the highest mean SHAP value) promoting water hyacinth occurrence at South African water systems per (5 x 5 km) block across South Africa (**A**) and the proportion cover of each group (**B**). Grey areas correspond to areas of no surface water and therefore no data. For a list of the factors included in each of the six groups refer to Table S1. Climatic factors dominate in the coastal regions whereas non climate variables dominate in the interior of the country.

3.4. Discussion

The most appropriate management strategy is dependent on how an IAAP behaves under different environmental and socio-economic contexts (Wilson et al. 2005). This study used pre-existing EO-derived datasets, SDMs and xAI to infer the relative importance of likely distribution defining features of water hyacinth within South Africa at multiple spatial scales. Many of the species-environment interactions of water hyacinth align with known thermal limits of the plant, and other environmental plant preferences. In addition, novel quantitative insights into plant-environment interactions based on the national extent analysis of water hyacinth occurrence is reported for the first time in a spatially explicit manner. Beyond revealing general correlations between water hyacinth occurrence and environmental and socio-economic variables, the high variation of these species-environment responses between provinces, and sites show the value of spatially explicit model interpretability for site-specific management in a cost-effective and data-driven manner.

3.4.1. The influence of climatic factors on water hyacinth distribution

The interaction between temperature and water hyacinth has been the most widely investigated feature among those considered in this study (Gopal 1987; Owens and Madsen 1995; Wilson et al. 2005) and provides useful information for verifying changes in feature importance at varying physiologically relevant temperatures.

Below 2.5°C, the persistence of water hyacinth through winter is extremely unlikely. Air temperatures < 0°C significantly increase the mortality of the plant by killing above-water portions (Madsen, Luu, and Getsinger 1993). This is supported by the negative SHAP values and the sharp decrease in the log odds value of water hyacinth presence below a minimum temperature of 2.5°C (Figure 7a). Such areas likely experience repeated exposure to severe cold that limit the plant's range (Miskella and Madsen 2019). This may result in smaller populations of overwintering water hyacinth surviving to the following season, such that nuisance populations may not form in areas along the limiting cold (boundary) edge of the IAAPs distribution (Kriticos and Brunel 2016). The cold edge is likely to lie at the boundary between contrasting blocks of negative and positive SHAP values. For example, the boundary separating coastal regions from inland areas (Figure 8b).

Between 2.5°C and 5°C, water hyacinth may be able to persist (indicated by the positive SHAP values, Figure 7a). However, these sites will contain smaller populations since short periods (< 2 weeks, (Owens and Madsen 1995)) of cold (5 °C) exposure may not kill the plant (Owens and Madsen 1995). In addition, the persistence of water hyacinth at these low temperatures may be attributed to the ineffectiveness of biological control agents under these conditions (Coetzee et al. 2011). Low temperatures and the lower plant quality may prevent agent establishment or suppress insect population growth by slowing down their development (Byrne et al. 2010; Hill and Cilliers 1999; Hill and Olckers 2000).

At temperatures between 5 - 8°C a dramatic vertical increase in suitability for water hyacinth occurs (Fig 6a), suggesting the suitability of warmer temperatures for water hyacinth. Interestingly, beyond this range, there is a gradual decline in the log odds of water hyacinth presence. This may be attributed to the proliferation of water hyacinth, its subsequent management prioritisation, and the increased effectiveness of biocontrol agents at warmer temperatures (for example, the lower developmental threshold for the two weevil biocontrol agents of water hyacinth is between 10-15 °C (Julien 2000)). This suggests the benefit of prioritising biological control at sites with a minimum temperature above 8°C when biocontrol agents are effective. This also suggests that mechanical, chemical, or integrated control should be prioritised for areas with minimum temperatures (in the coldest month) lower than 8 °C.

3.4.2. Microclimate effects

Continuous Heat Insolation Load Index (CHILI, 0-255) is a surrogate effect of topographic shading. Higher CHILI values are associated with warmer topo-climates attributed to the lack of topographic shading. Topographic diversity corresponds to the availability of local moisture and temperature topo-climatic habitat conditions. High topographic diversity values correspond to heterogenous temperature-moisture conditions and has been associated with species resilience under climate change (Lawrence et al. 2020).

Topographic shading effects have a variable effect on the occurrence of water hyacinth (Figure S1a). This may be attributed to the high mobility associated with free-floating aquatic vegetation, resulting in temporary exposure to unsuitable shading effects. In comparison, topographic diversity shows a less variable effect on water hyacinth occurrence (Figure S1b). Moreover, the water hyacinth suitability response to topographic diversity and urban area (Figure S1b) suggests that uninvaded waterbodies with a combined high topographic diversity (> ~0.5) and a large urban area (2.5 – 3 km^2) within a 5 km radius of the waterbody have a high susceptibility to invasion by water hyacinth and should be prioritised for early detection of water hyacinth (Hill 2003; VonBank et al. 2018).

3.4.3. Precipitation

New Year's dam is a small (150 ha), shallow, oligotrophic dam within the Eastern Cape. Above-average rainfall (> 350 – 550 mm, (Palmer 2004; Zengeni, Kakembo, and Nkongolo 2016)) in this semi-arid area has been suggested to have initiated the weed's resurgence in 1998 by increasing the nutrient input into the dam (Hill and Olckers 2000). This may represent one of the mechanisms that explain the gradual positive relationship between precipitation and water hyacinth occurrence (Figure S1c). Based on the dependence plot for precipitation (Figure S1c), an annual accumulated precipitation greater than ~560 mm (much higher than the average (463 mm, (World bank group 2021)) rainfall in South Africa) is linked with an increased chance of water hyacinth occurrence (Figure S1c), supporting the likely role of increased rainfall at New Year's dam.

Precipitation also has a direct influence on the presence and persistence of surface water. The dependence plot for water persistence shows a positive trend with the occurrence of water hyacinth (Figure 7b). While water hyacinth is resilient to water level changes (Venter et al. 2017), waterbodies with high seasonality are less likely to support water hyacinth (Figure 7b). Permanent waterbodies have 2-3 times the (log) odds of supporting water hyacinth compared to seasonal surface water conditions (present for 1-3 months of 2013). In the dry karoo areas of the Northern Cape, there is a reduced availability of permanent water to support water hyacinth (Figure 9a) explaining the limited occurrence of WH in the area despite suitable temperature.

For Roodekoppies dam the slightly above average rainfall (474 mm vs the average 463 mm) at the site was estimated to reduce the suitability for the occurrence of water hyacinth. The suitability of sites with slightly above average rainfall to support water hyacinth is dependent on shrub cover (Figure S1C). Based on figure S1C, the decision is partly determined by an interaction with shrub cover (%) within the 5km riparian zone considered. Generally, high shrub cover (>25%, 27.8% for Roodekoppies dam) together with near- average rainfall conditions are associated with reduced water hyacinth suitability. This may be linked to the buffering effect of shrubs that reduce the amount of nutrients that enter water systems through runoff or subsurface flows (Aguiar Jr et al. 2015; Jiang et al. 2020). However, when there is high rainfall and accelerated runoff, the buffering effect of shrubs is reduced through increases in carbon mineralisation (the conversion of soil organic carbon to carbon dioxide), and subsequent releases of nitrogen into the adjacent water, especially in (soil) nitrogen saturated systems (Sabater et al. 2003; Taylor and Townsend 2010). Moreover, since the tree, and shrub cover variables comprise of both natural vegetation and agricultural land use, there are likely cases whereby the fertiliser input for tree crops may combine with the precipitation to result in an increased nutrient input into the waterbodies. This is supported by the tree, shrub cover and precipitation variables being the dominant factors influencing the occurrence of water hyacinth within the Free State province (Figure 9). Specifically, the east-west rainfall gradient across the Free State, which corresponds with the occurrence of agriculture resulting in more precipitation and therefore agriculture in the eastern region than the western side of the province. This may also explain why the tree cover and shrub cover variables show higher overall predictive power than either of the agricultural or natural vegetation variables alone that only contribute partial information to the model (Figure 5).

3.4.4. Socio-economic effects

The presence of alien and invasive species is strongly linked to human activity through human aided dispersal for introduction and spread (Hill 2003) and, human-induced disturbance for establishment (VonBank et al. 2018). Based on the human modification feature (global human modification, ranging between zero (low human modification) and one (high human modification)) that captures human influence from 13 datasets representing human settlement, agriculture, transportation, mining, and energy production (Kennedy et al. 2019), there is a parabolic relationship with the presence of water hyacinth (Figure S1d), whereby the odds of water hyacinth occurring peaks at a human modification score of 0.25. Human modification values on either extreme are shown to limit the distribution of water hyacinth occurrence. Greater human modification may cause more disturbance than water hyacinth is able to tolerate, or alternatively, the plant may persist and because of its associated socio-economic negative consequences, the plant may be targeted for removal. In addition, the occurrence of artificial surfaces (urban villages), and agricultural fields, (cultivated subsistence and cultivated orchards) are all associated with a higher probability of water hyacinth likely because of the increased nutrients from runoff (Figure 5b). In addition to agricultural water nutrient inputs, poorly functioning wastewater treatment facilities are also a known contributor to the eutrophic water conditions in South Africa (Harding 2015; Oberholster and Ashton 2008).

3.4.5. Environmental effects

The higher overall importance of shrub cover to increase the suitability of water hyacinth, over tree and grass cover, may be ascribed to the distribution of tree cover, land use effects or buffer and interaction effects (Figure 5, (Aguiar Jr et al. 2015; Jiang et al. 2020; Sabater et al. 2003; Taylor and Townsend 2010)). For example, microclimate (CHILI) shows a strong interaction effect with tree cover (Figure S1A), which may influence nutrient cycling between the riparian zone, a waterbody and the atmosphere and may induce a shading effect that reduces the temperature and sunlight available for photosynthesis. However, this is not consistently the case (high (data) aleatoric uncertainty – that is, characterized by chance or indeterminate elements) and there are likely other confounding variables such as elevation and temperature (Cilliers et al. 1996). Minimum temperature, the strongest predictor of water hyacinth occurrence (Figure 5) has the strongest interaction effect with shrub cover (Figure 7A, as opposed to tree or grass cover). Within the narrow band of 2.5°C and 4°C, high percentage shrub cover (>~25%) increases the suitability of water hyacinth (Figure 8A). This may be attributed to shrubs moderating the cold winter temperatures. Whereas at all other temperatures, shrub cover reduces the suitability for water hyacinth, which may be attributed to the nutrient-buffer effects of shrub cover. Some studies have found that shrubs are more effective than trees and/or grass cover at reducing nutrients from runoff (Aguiar Jr et al. 2015; Cole, Stockan, and Helliwell 2020). In high rainfall areas (575-600mm), shrub cover increases the suitability for water hyacinth (Figure 7b). This may be because of the reduced effectiveness of shrub cover to filter nutrients under high rainfall regimes and high nutrient runoff loads that could lead to the saturation of nutrients in the riparian zone and a reduction in the vegetation nutrient-buffer effectiveness (Jiang et al., 2020). Next, agriculture is a major source of nutrients for waterbodies and is encompassed in the tree and shrub cover variables (for example, woody perennial crops <5m are likely included in shrub cover and if >5m are likely included in tree cover) and may partly explain the lower ranked importance of grass cover that does not comprise agriculture. In addition, the western cape has more water hyacinth but less tree cover (Figure 6) and could therefore suggest that tree cover is not as general a predictor of water hyacinth as shrub cover, especially across large extents and why shrub cover is ranked much higher than tree or grass cover.

Interspecies competition has no effect on water hyacinth occurrence at low temperatures (Figure S1F). Moreover, between 0 and 730 m^2 of aquatic vegetation cover, there are no sites. This suggests that water hyacinth outcompetes other aquatic vegetation to the point of exclusion. It is only at warmer temperatures (>3°C minimum temperature in the coldest month) that interspecies competition occurs and allows for multiple species to co-occur, reducing the probability of water hyacinth establishment (Wundrow et al. 2012). These circumstances may also be attributed to the increased effectiveness of biocontrol at warmer temperatures, the consequent reduced competitiveness of water hyacinth and secondary invasions (Coetzee et al. 2005; Strange et al. 2019). Owing to the plant's free-floating characteristics and not being limited to shallow water depths, water hyacinth is able to outcompete most littoral vegetation (Coetzee et al. 2017). The weeds' phenotypic plasticity has also been reported to aid its competitiveness. Established mats tend to extend their leaf area and increase in height by lengthening their petioles, shading out other IAAPs (Agami and Reddy 1990; Center and Spencer 1981). This coupled with the high vegetative reproduction on the leading edge of newly formed mats are probably the mechanisms that have aided in the ability of water hyacinth to outcompete other floating weeds (Agami and Reddy 1990; Coetzee et al. 2005). The minor proportion of instances that are in contradiction i.e., the instances that suggest interspecies competition improves the suitability for water hyacinth may be attributed to false negative classification (omission) errors. This study assumes that the impact of the incorrect instances of water hyacinth occurrences on the overall SHAP values would be negligible and can still provide reliable insights considering the high F1 score (F1 score >0.8).

Comprehensive satellite-based or field-based estimations of water nutrient levels are not publicly available for South Africa and are difficult to estimate from satellite imagery. Therefore, the soil nutrients, runoff, and agriculture variables for the (5 km) area surrounding a waterbody were used instead. Eutrophic water conditions are known to be a dominant driver of water hyacinth infestations and are expected to be increased by runoff of nutrients in adjacent soil (Bick et al. 2020). However, in contradiction to existing research (Bick et al. 2020), the total riparian soil nitrogen has a variable effect on water hyacinth occurrence and suggests that riparian soil nitrogen is not an adequate proxy for water nitrogen content (Figure S1f). This is also supported by the increased water hyacinth suitability with the occurrence of agricultural and urban land cover (Figure 6). The limited utility of riparian nutrient levels as a proxy for water nutrient levels may be ascribed to the biogeochemical processes occurring in the riparian zone and the buffer extent used in this study. The wide buffer used (5km) includes soil nutrient conditions of the riparian vegetation (typically between 10-100m) and surrounding land use land cover that, when combined does not closely reflect the water nutrient levels (Bredin and Macfarlane 2017). In addition, riparian vegetation and microorganisms may transform, retain or remove nitrate by conversion to nitrogen gas through denitrification (Aguiar Jr et al. 2015; Sabater et al. 2003; Taylor and Townsend 2010). Future work should therefore consider multiple buffer widths and other proxies for water nitrogen, such as time series of NDVI from irrigated agricultural fields, that can be scaled to large extents. Based on the variables considered in this study, the area of urban village and soil organic carbon showed the highest correlation (~0.4) with in-situ water nitrogen data for 2013 and should therefore be top candidates for water quality proxies.

Floods are associated with an increased chance of water hyacinth occurrence, especially if there is urban land use surrounding the site (Pérez et al. 2011b) (Figure S1g). Floods increase the dispersal ability of water hyacinth, allow for seeds buried in the substrate to germinate under the now open water with access to sunlight and the inflow of nutrients (Neiff, Poi, and Casco 2001; Pérez et al. 2011b), and may also decrease the effectiveness of biological control at these sites (Cilliers 1991) because biocontrol agent populations take much longer to recover after a flood than their host plant. However, in near-coastal infestation scenarios, floods have been suggested to force plants into intolerable saline water conditions (Coetzee et al. 2017). The results (Figure S1g) do not support the ability of floods to act as a regulatory mechanism for water hyacinth populations during 2013. This is despite many water hyacinth infestations occurring near coastal areas (Figure 2).

3.4.6. Benefits and drawbacks of the used methods

Correlative SDMs remain valuable for IAAP species risk mapping and management but are often criticised for a lack of biological underpinning (Srivastava et al. 2019). Modelling experts partly address this concern by using prior expert knowledge of species' requirements or tolerances to define the set of likely and potential variables to use during modelling. Nevertheless, SDMs are correlative in nature and therefore (Smith and Santos 2020; H. Yu et al. 2020), all insights drawn from this study need to be considered with typical caution, especially if species-environment feedback mechanisms occur. For example, minimum temperature strongly influences the occurrence of water hyacinth and water hyacinth mats can decrease surface water temperature through canopy shading effects and by encouraging mixing of water to moderate temperatures in lotic systems (Giraldo and Garzon 2002).

Mechanistic models require costly and labour-intensive field or lab experiments to obtain sufficient data for a representative set of drivers that enable the creation of SDMs that may still not hold across large national extents with varying climate regimes (for example, (Wilson et al. 2005)). While many EO-derived datasets provide numerous readily available explanatory variables that can be used to create SDMs (Srivastava et al. 2019), the availability of these variables are constrained by technical limitations, spatio-temporal extensibility, spatially varying accuracy, and the overall recent establishment of the open EO sector. Moreover, it is not always possible to directly compare EO-derived values with instantaneous values derived through direct observations. For example, the minimum temperature data, used in this study is dependent on long-term climate data and may not always correspond to fine-scale, direct and instantaneous in-situ measurements. Nevertheless, the results in this study for minimum temperature fit known critical temperatures governing water hyacinth growth (Figure 5a).

By using EO-derived distribution data, the benefits of using presence-absence data over presence-background data used with Maxent or Ecological Niche Factor Analyses (ENFAs) or presence-pseudo-absence data used with other ML algorithms are gained (Chapman et al. 2019). This includes reduced uncertainty compared to using pseudo-absence data (Chapman et al. 2019). High quality EO-derived distribution maps are more comprehensive and less prone to sample bias compared to costly field collected samples. These biases violate the assumption of independence among species records (Guillera-Arroita 2017). Consequently, the output habitat suitability maps based on biased data may not only correspond to the species' observed distribution, but to the distribution of sampling effort. Equally, all EO-derived distributions are prone to mapping errors that may include residual spatial autocorrelation. Residual spatial autocorrelation refers to spatial autocorrelation in the distribution of model error, causing the model to be spatially-differentially reliable. As a result, these errors and limitations may be propagated to the interpretability of SDMs.

The overall performance of the SDM used in this study results in less than 20% error based on the F1-score. This error may partly be attributed to the environmental conditions being suitable for water hyacinth occurrence, but water hyacinth not yet being introduced at these sites. If this is the case, it would suggest that the distribution of water hyacinth is strongly controlled by dispersal constraints, and thus might not be in complete equilibrium with its recipient environment throughout South Africa (see also, (Normand et al. 2011)). The error may also be an outcome of not including all variables (such as turbidity, water nutrient levels, history of management interventions, and water depth (Venter et al. 2017)) that are also known to influence the occurrence of water hyacinth.

3.4.7. Future work

To reduce the probability of including spurious correlative species-environment features, causal feature selection which will reduce the probability of including spurious features may be useful (via the pyCausalFS package) if certain criteria are met (for example the absence of unmeasured confounding variables) (K. Yu et al. 2020). Lastly, considering the predictive power of non-spectral (i.e., climatic, ecological, etc) features used in this study, their inclusion is recommended during species mapping exercises that rely on EO spectral data to improve the accuracy and generalisation capability of SDMs. The SDM generated in this study can easily be used to identify high risk waterbodies where water hyacinth is likely to extend its range in the future. The identification of features contributing to the increased susceptibility of a water hyacinth infestation, their spatial distribution and associated importance may aid in risk-prioritisation, justifying costly but more data rich monitoring exercises or traditional lab-controlled and artificial outdoor experiments. The knowledge gained from these experiments may, in turn, assist the development and improvement of model-based approaches. In this way, such an active learning-based approach may leverage the benefits from both traditional experimental setups and model-based methods that together promote the effective local management of introduced IAAP populations at an early stage, across the wide water hyacinth distribution.

3.5. Conclusion

This study points towards spatially varying drivers of water hyacinth and subsequently suggests that effective management efforts will be context dependent. To this end, the utility of ML, xAI, EO and GEE to assist in understanding the factors that limit and promote the establishment of water hyacinth in a data-driven and evidence-based manner is demonstrated. This information may prove useful for the pre-selection and prioritisation of management strategies on a site-by-site basis. Owing to the negligible costs of carrying out this analysis, in comparison to similar large scale field studies, computer-based studies and additional research into the development of EO-derived species distributions that are foundational for the development of such studies are encouraged.

3.6. Data and code availability

The dataset used in this study with water hyacinth occurrence and the extracted features has been published on Kaggle. The notebooks used in this study that allow for the complete reproducibility of this study, including figures are available from this GitHub repository.

3.7. Supplementary information

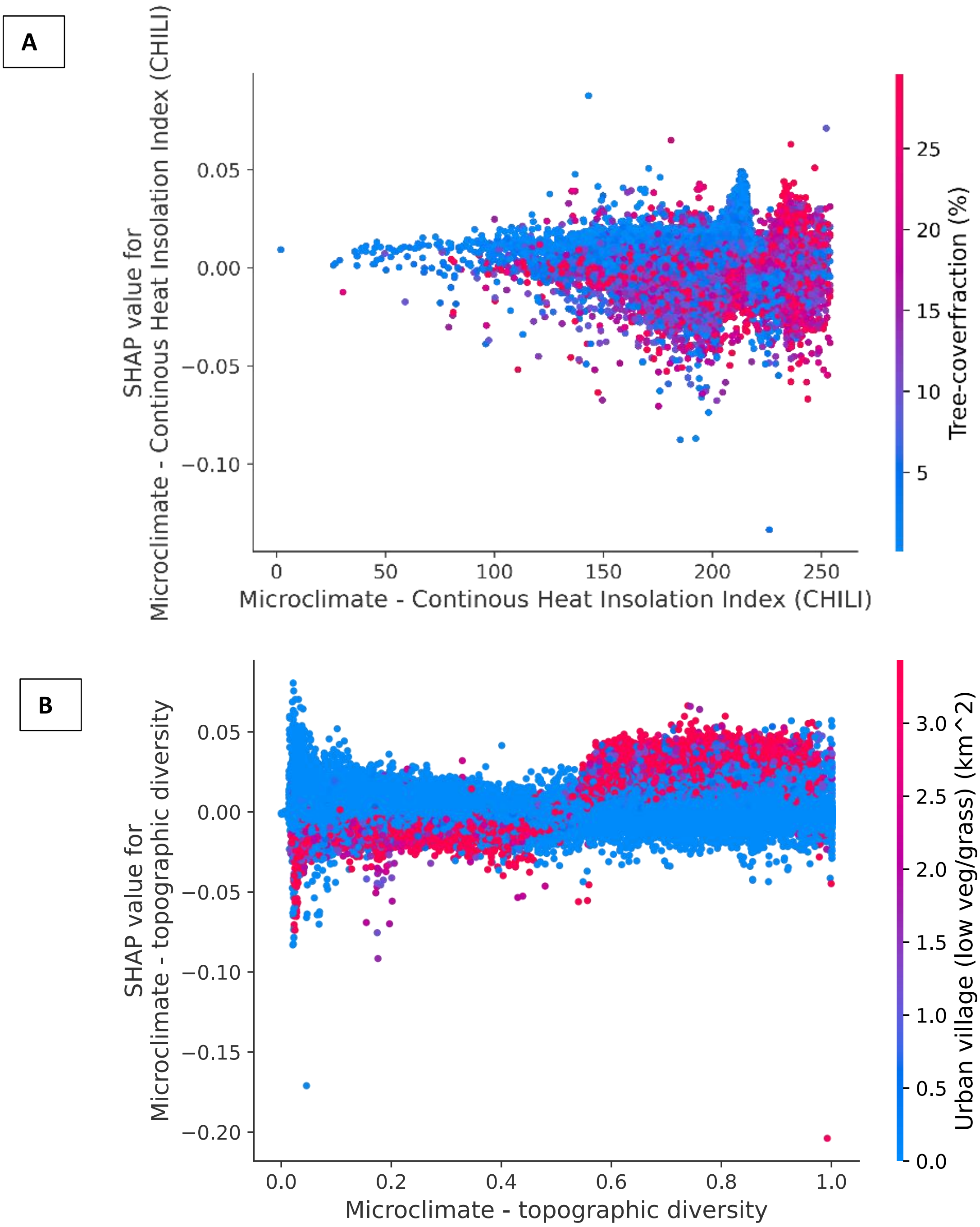
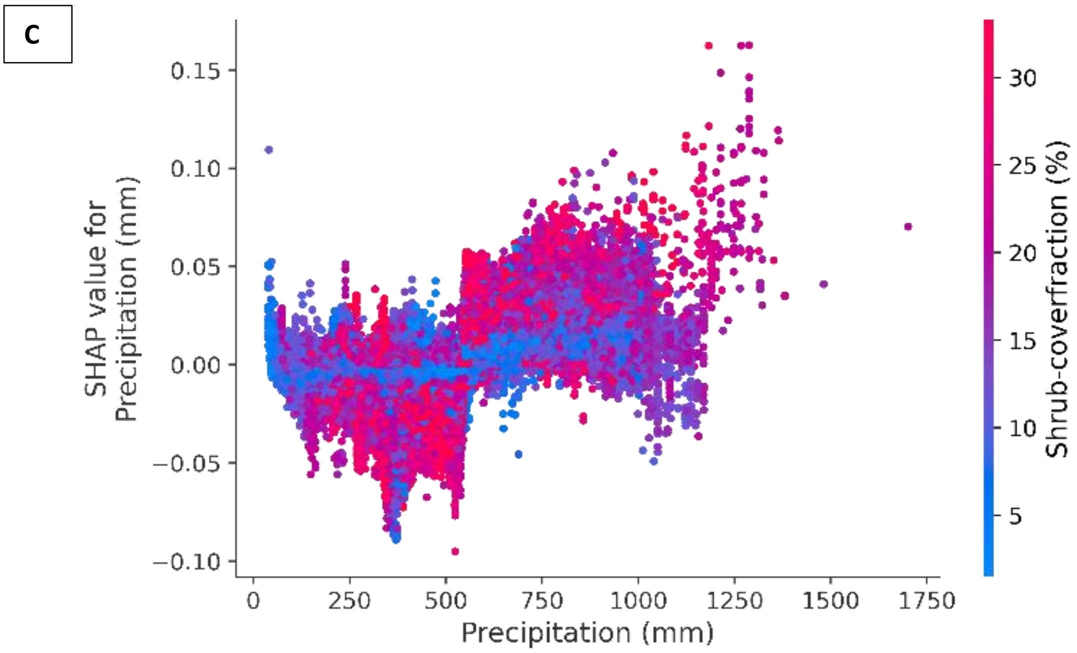


Figure S1: SHAP dependence plots for (**A**) Continuous Heat Insolation Index (CHILI) and (**B**) topographic diversity- a proxy for topo-climate niches and micro-climate. Also includes feature interactions indicated by the blue-red colour.



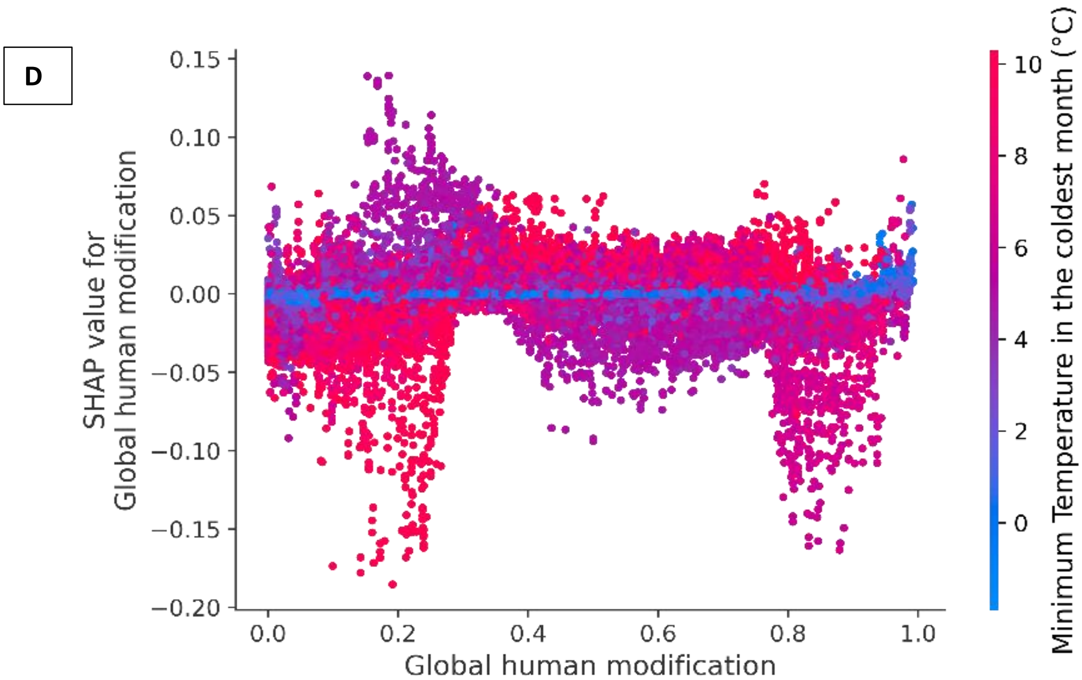
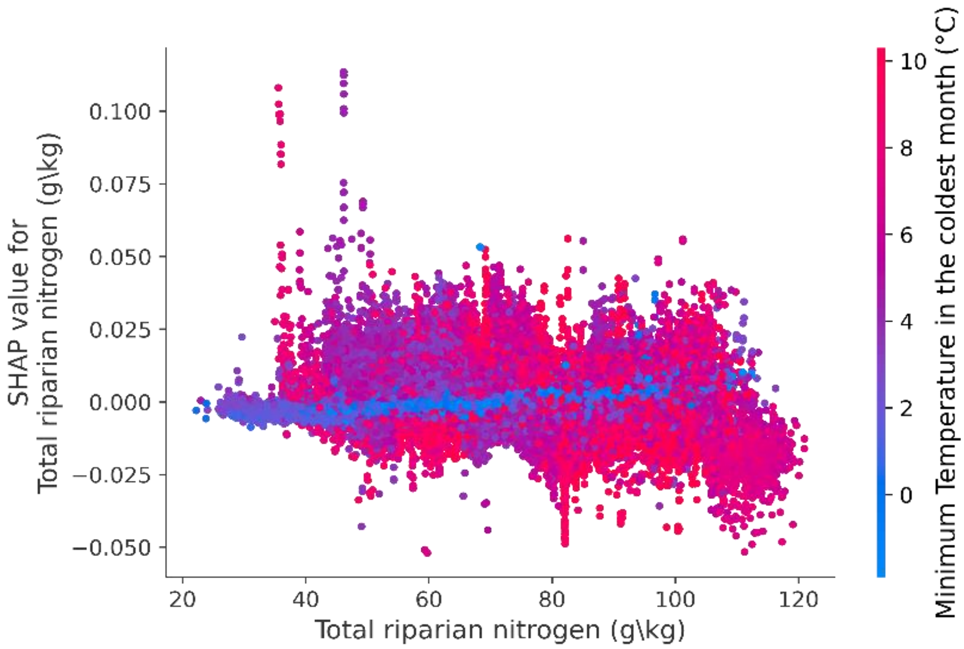


Figure S1 (continued): SHAP dependence plots for € total precipitation (mm) for 2013, (**D**) global human modification – an index that reflects anthropogenic influences and landscape alteration. Also includes feature interactions indicated by the blue-red colour.

E



F

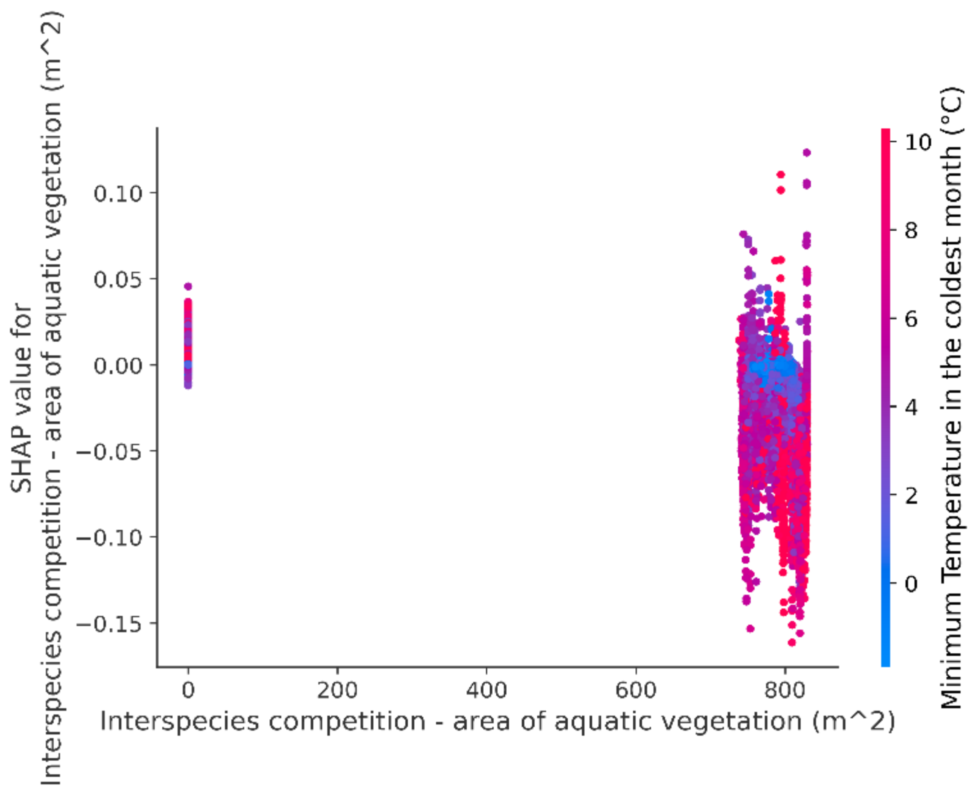


Figure S1 (continued): SHAP dependence plots f€(**E**) Riparian total soil nitrogen (–/kg) - a proxy of water nitrogen levels (**F**) Aquatic vegetation area – a proxy for interspecies competition that excludes water hyacinth area . Also includes feature interactions indicated by the blue-red colour.

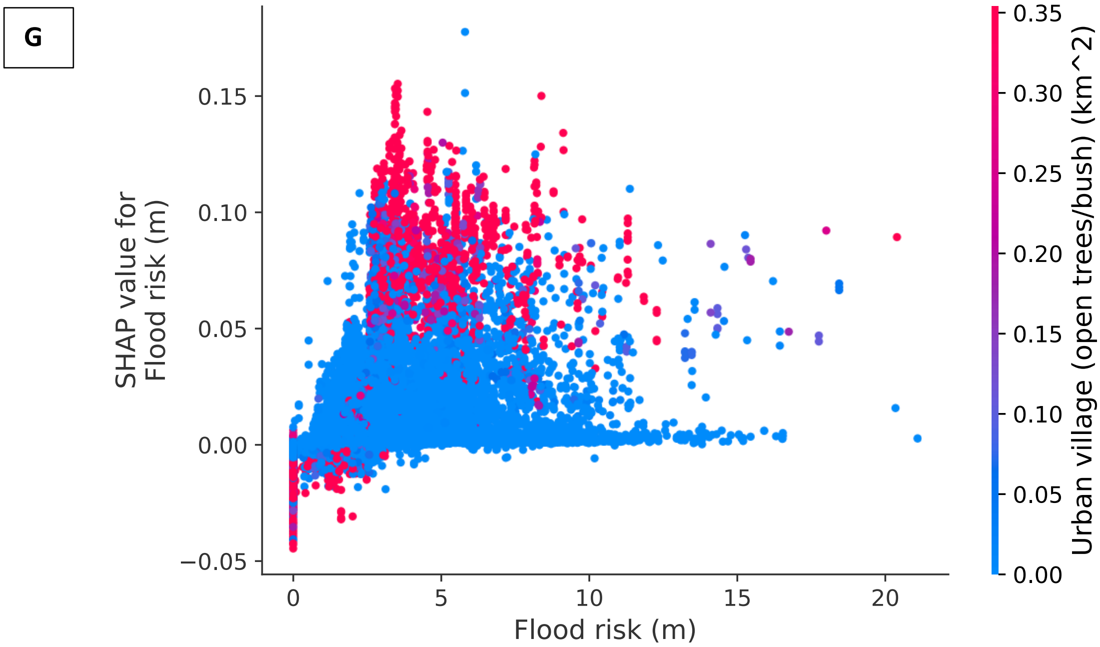


Figure S1 (continued): SHAP dependence plots for (**G**) 10-year return flood hazard (metres). Also includes feature interactions indicated by the blue-red colour. Sites with high urban area are associated with a greater risk of water hyacinth.

Table S1. The 82 (of 140) features used in the final Species Distribution Model. The variables have been grouped into five categories referred to in Figure 9. A large portion of the variables relate to area of a Land Use and/or Land Cover within a 5 km buffer surrounding a waterbody (km^2).

Group	**Features (82 significant predictors of water hyacinth occurrence)**
Hydrological	• Downstream river count • Surface water persistence (up to 12 months) • Percentage and area of permanent water area within a 5km buffer • Percentage and area of seasonal water area within a 5 km buffer • Stream power index • Risk of flood (metres of flood water) • Runoff (mm) • Natural surface water (pans flooded at observation time)
Natural land cover	• Percentage grass cover • Percentage shrub cover • Percentage tree cover • Percentage bare cover • Forest (woodland and indigenous forest) • Shrubland (shrubland fynbos, low shrubland, thicket) • Grassland (sparsely wooded grassland)
Climatic	• Minimum temperature • Precipitation

	• Wind speed • CHILI (microclimate, see text for explanation) • Topographic diversity (microclimate, see text for explanation)
Ecological	• Interspecies competition (area of non-water hyacinth aquatic vegetation) • Riparian soil (organic carbon, nitrogen content and cation exchange capacity) • Riparian stone, sand and clay content • Riparian soil fertility capability classification • Erosion • Fertility capability classification • Bare non-vegetated
Socio-economic	• Plantation forest • Global human modification (see text for explanation) • Mines • Residential • Urban (smallholding, built-up, schools and sp0rts ground, township) • Industrial • Commercial • Urban cover within a 5 km buffer • Distance to roads
Agricultural	• Agriculture (subsistence, commercial, cane, orchards, vines and crop cover fraction)

4. Chapter 4: A Machine Learning Framework for Identifying Optimal Biocontrol Release Actions: A Case Study of Aquatic Weeds on a South African Dam.

4.1. Abstract

Limited data and resources hinder Invasive Alien Aquatic Plant (IAAP) management and monitoring. This study explores a widely applicable and low-cost decision-making framework to forecast effective IAAP management interventions, that includes the release date of biological control on water hyacinth in a South African water system. Specifically, a multi-armed bandit (MAB) balances the exploitation[1] of biocontrol release dates, which appear to maximize performance, with the exploration[2] of uncertain, but potentially more effective release dates. In this chapter, three MAB action selection strategies[3] are tested using limited biocontrol release data, satellite-based weather data and open water extents derived from Earth Observation (EO) data. All three action selection strategies identified a spatio-temporal locally relevant effective biocontrol release period between late summer (February and mid-March) for Hartbeespoort dam. However, this is because of an algal bloom. Therefore, an early summer (mid-November) release is also suggested to ensure the availability of water hyacinth throughout summer, which may also prevent an algal bloom. This can be followed up with remote or in-person monitoring of the dam and if water hyacinth is present during late February an augmented biocontrol release could be made. The results suggest an 8-27% improvement in the average (70-day) proportion of open water for the identified optimal biocontrol release over the actual biocontrol releases made during the study period (May 2018 - July 2020) or the worst simulated open water responses within the study period. In addition, future extensions are proposed to increase the applicability and utility of the introduced framework to a wider range of infested sites and factors that influence the effectiveness of biocontrol release events. It is anticipated that the novel application of MABs to optimise biocontrol release under a changing climate, land use change, or on sites with limited biocontrol management success will stimulate new research directions and assist such releases to become automated and pre-emptive, and therefore more effective in the future.

[1] Exploitation- the selection of an arm (biocontrol release strategy) that is known to maximise immediate performance (Russo et al. 2018).

[2] Exploration – to invest in the selection of an arm not previously selected and with unknown performance but may improve long-term future performance(Russo et al. 2018).

[3] Action selection strategies – the heuristic used to decide between exploration and exploitation. In this study, each action was comprised of four factors, namely, biocontrol release date, species of biocontrol, the population size of the released biocontrol agent and the number of points at which the agents were released.

Keywords: monitoring, post-release evaluation assessment, water hyacinth, reinforcement learning, Thompson Sampling, Google Earth Engine.

4.2. Introduction

Biological control of water hyacinth is one approach to manage the weed and involves the release of host-specific insects or fungi from the plants' native range for the purpose of population control (Harley and Forno 1992; Hussner et al. 2017). Some IAAPs require regular releases of biological control agents to prevent the IAAPs from thriving in the absence of insect-feeding pressure (Hill et al. 2021). For example, in South Africa, the suitability of water hyacinth populations for biocontrol fluctuates throughout the year because of cold winters, floods, herbicide application that is antagonistic to biocontrol or the mechanical removal of plant populations (Coetzee et al. 2011; Miller et al. 2021). This leads to the concurrent collapse or removal of any resident biological control agents, necessitating supplementary releases (Hill and Olckers 2000; Miller et al. 2021). However, there is uncertainty regarding the optimal (i) date of release, (ii) biocontrol species, (iii) release population size and the (iv) number of localities at which to release biocontrol (Memmottlz et al. 1996; Shea and Possingham 2000). Moreover, it is difficult to weight the relative importance of these four factors using field-based approaches across many water systems, which experience unique environmental conditions (Shea et al. 2002; Westgate, Likens, and Lindenmayer 2013).

Unfortunately, for low-biocontrol-priority managed sites (due to limited resources and/or environmental unsuitability for biocontrol), the decision to release biocontrol agents into resurgent IAAP populations often relies on field observations of the IAAP by informed citizens (B. Miller, CBC, pers. comm). Consequently, alerts of resurgent or new infestations are reactive, not systematic, and may be ill-timed. In contrast, for high-biocontrol-priority sites these consequences are alleviated, to some degree, by close-monitoring, systematic release and expert knowledge regarding local seasonality and the growth cycles of the target weed. With this knowledge, biocontrol agents are continuously mass-reared[4] (Hill et al. 2021), and owing to population growth periods and resource constraints, agent releases are prioritised during periods of peak weed population growth - typically during spring (September-November), or after herbicidal application (Coetzee et al. 2021). This ensures the availability of a large quantity of high-quality food for the biocontrol agents to thrive (Hill et al. 2021). Despite this inundative biocontrol release approach, current biocontrol releases possibly result in sub-optimal weed control and are likely to remain sub-optimal due to the cost, and duration of adaptive management. Specifically, numerous cycles of trial and error are typically required under successful adaptive management to identify effective biocontrol release actions (Walters and Hilborn 1978). For biocontrol, a single cycle encompasses mass-rearing biocontrol agents, their release, and the monitoring of the weed’s response to the release. After a single cycle, the insight gained can be used to suggest incremental improvements to subsequent cycles that result in gradual optimisation of a biocontrol release for the study site under investigation. The number of cycles, time taken to reach optimal biocontrol release, uncertainty attributed to changing environmental conditions (for example, weather conditions), and the required amount of in-situ data collection may be reduced with the incorporation of machine learning methods and satellite monitoring data. This will allow for the accelerated exploration and evaluation of biocontrol release actions, that can therefore assist in the prioritisation of effective release events (Walters and Hilborn 1978).

The release and post-release evaluation assessment can be combined under a multi-armed Bandit (MAB) framework. The MAB framework can be used to identify the optimal biocontrol release strategy from a set of candidates. This represents the exploration-exploitation dilemma in decision making (Berger-Tal et al. 2014), that is, how to acquire knowledge about the set of available biocontrol release actions while exploiting the most profitable ones (Srivastava, Reverdy, and Leonard 2013). Generally, MAB techniques explore the environment during the early stages of the decision-making framework to find the best actions (or ‘arms’). After this exploration phase, the algorithm has sufficient knowledge to exploit the actions that will maximize the cumulative reward (in this study, open water proportion) over the complete number of trials. During each trial in the decision-making framework, an agent selects one arm among the available arms and receives a reward (or payoff) proportional to the value of the proposed action’s outcome.

In the literature, different action selection strategies have been studied to solve the exploration-exploitation dilemma (for example, (Carpentier et al. 2011; Russo and van Roy 2016)), with Thompson Sampling (TS), being one of the most used and successful (Russo and van Roy 2016). There have been many practical applications of MABs that justify the exploration of MABs for biocontrol in this study (Bouneffouf and Rish 2019). For instance, the MAB has been applied in the field of clinical trials (Durand et al. 2018), online advertising, recommendation systems (Q. Zhou et al. 2017), or within ecology to better understand insect and animal foraging behaviour (Keasar et al. 2002; Srivastava et al. 2013). The benefit of modelling the prioritisation of biocontrol release actions as a MAB problem includes the use of pre-existing algorithms and best practices from the MAB domain. Moreover, trial and error within a simulation environment enables many more potential release actions to be evaluated at a faster rate than actual field releases, with a reduced risk of failure and loss on investment.

This study applies the MAB framework to a novel application - the identification of effective biocontrol release actions whilst using open-water cover data, satellite-based weather data and limited biocontrol release data. This method is demonstrated for biological control of water hyacinth on Hartbeespoort Dam, South Africa. This research explores, 1) the combined use of Landsat-8, Sentinel-1, and Sentinel-2 data to monitor the extent of open-water and 2) the novel application of a MAB framework to assess and identify candidate biocontrol release actions.

4.3. Methods

Hartbeespoort Dam is located at the border of the semi-arid Gauteng and North-West provinces of South Africa. The dam covers ~18.8 km^2 at an altitude of ~1158 m in the highveld region of South Africa and experiences cold-winter conditions (< 8°C air temperature based on satellite data used in this study), and hypertrophic water conditions (Auchterlonie, Eden, and Sheridan 2021; Matthews and Bernard 2015). Since the occurrence of water hyacinth on Hartbeespoort Dam in the 1960s, and the start of biocontrol in the 1990s, biocontrol was first documented to decrease water hyacinth cover independently and significantly during October 2019 and March 2020 (Coetzee et al. 2022) (refer to (Cilliers 1991; Coetzee et al. 2021, 2022; Hill and Coetzee 2017; van Wyk and van Wilgen 2002) for details on the history of water hyacinth management at Hartbeespoort Dam). This reduction has been associated with frequent releases of large numbers of the water hyacinth plant hopper (*Megamelus. scutellaris*) (Coetzee et al. 2021). The various species of water hyacinth biocontrol agents perform most effectively in tropical to sub-tropical conditions and therefore experience population bottlenecks during the South African highveld winter, contributing to their variable effectiveness across the country (Coetzee et al. 2021; Hill and Olckers 2000).

To assess the effectiveness of a potential biocontrol release event using a MAB framework (Figure 1, section 4.3.4), a supervised machine learning model (referred to as the simulation environment) is used to simulate the response of open water to a biocontrol release event (section 4.3.3.). The simulation environment can forecast the likely response of the proportion of open water to the selected biocontrol release date (section 4.3.1.) based on satellite-based weather data and a given set of biocontrol release parameters (section 4.3.2.).

4.3.1. Generate a time-series of open water proportion.

Both optical (Landsat-8 and Sentinel-2) and radar (Sentinel-1) data were combined to derive the response variable - a time-series of open-water fluctuations on Hartbeespoort Dam between May 2018 – July 2020 (n=951). To extract the proportion of open-water cover (open water/ total waterbody extent) from the optical and radar images in a spatio-temporally extensible and automatic manner, a sensor-agnostic approach that involved clustering spectrally similar satellite image pixels over Hartbeespoort Dam by applying an unsupervised machine learning algorithm was used.

Prior to generating a time-series of open-water area from Sentinel-1 for the period May 2018 – July 2020, image quality was improved by pre-processing (refer to [15] for technical details). This involved applying, 1) a mono-temporal refined-lee speckle filter, 2) border-noise removal and 3) terrain correction based on NASADEM (readily available from Google Earth Engine (GEE)). For the optical systems (Landsat-8 and Sentinel-2), surface reflectance images with less than five percent cloud cover and images with complete coverage across the area of interest were selected. Thereafter, the unsupervised K-means (K=2) clustering algorithm was applied in a pixel-based manner for the detection of open-water within the boundaries of the waterbody (as recommended by (Singh et al. 2020)) by using the waterbody extent closest to the date of open-water detection. To compute the proportion of open water, the open water cluster was selected by identifying the cluster with the lowest mean NDVI value for Landsat-8 and Sentinel-2 or the lowest median backscatter value for Sentinel-1.

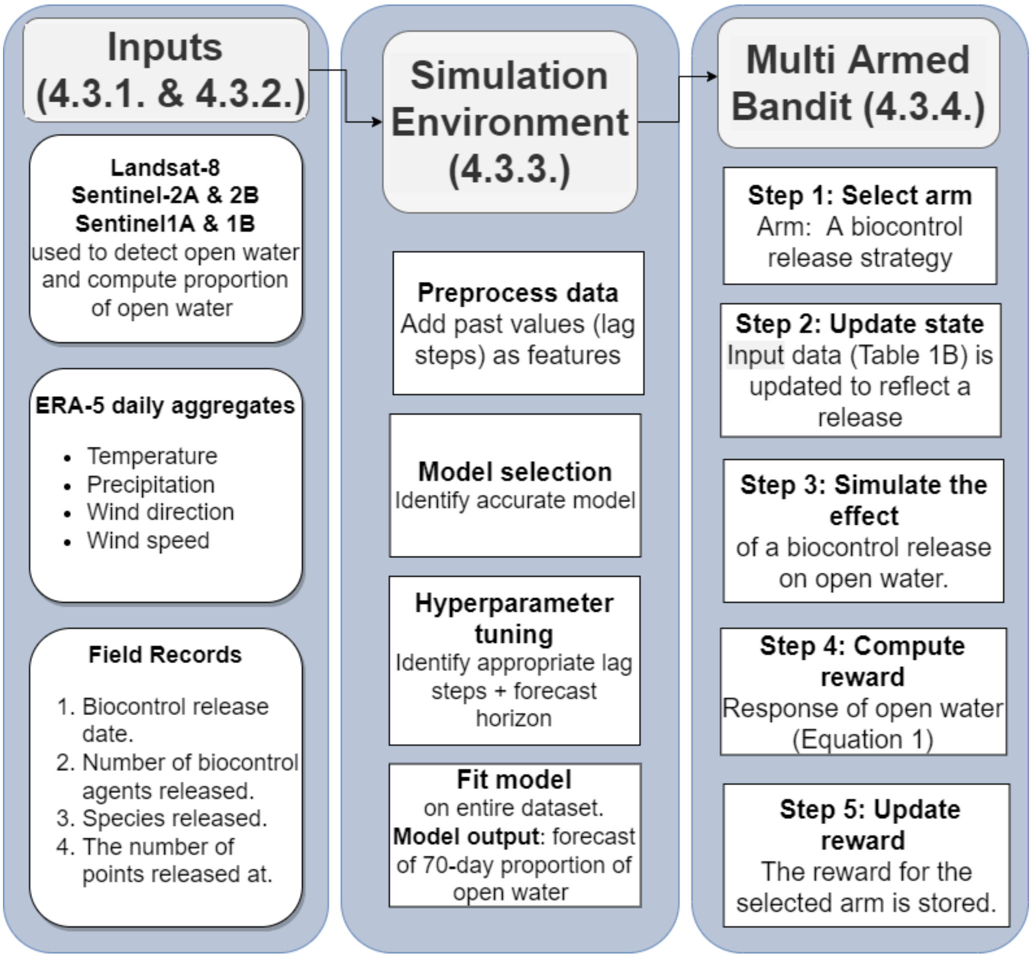


Figure 1. The multi-armed bandit framework used in this study to prioritise biocontrol release actions for Hartbeespoort Dam, South Africa. Steps 1 - 5 are taken during a single trial – action selection followed by the evaluation (based on forecasts made by the simulation environment) of the selected actions effectiveness.

This method was used to process all (Landsat-8, Sentinel-1, and Sentinel-2) satellite images over Hartbeespoort Dam between 2013 (the start of the landsat-8 archive) and 2021 (n=951). For days with multiple (up to 3) satellite observations, the mean proportion of open water was computed. Sixty seven percent of all observations used in the final analysis (May 2018 – July 2020, n=310) had a temporal interval of 2 days or less between consecutive observations (Figure 2A). Despite this there were still 486 days with missing values ascribed to satellite revisit periods and cloud cover. To obtain a regular time series of open water cover with a daily interval (n=796), smoothing was applied to the time series and the missing values were imputed using a Kalman filter (Figure 2B). Smoothing is often applied to time series data to remove short term variability while maintaining the long-term trends and improving model-forecasting-power (Pérez-Ortiz et al. 2003). In this case, short term variability may be attributed to wind conditions (Wilson et al. 2005), noise attributed to cloud occurrence, image artefacts or detection errors.

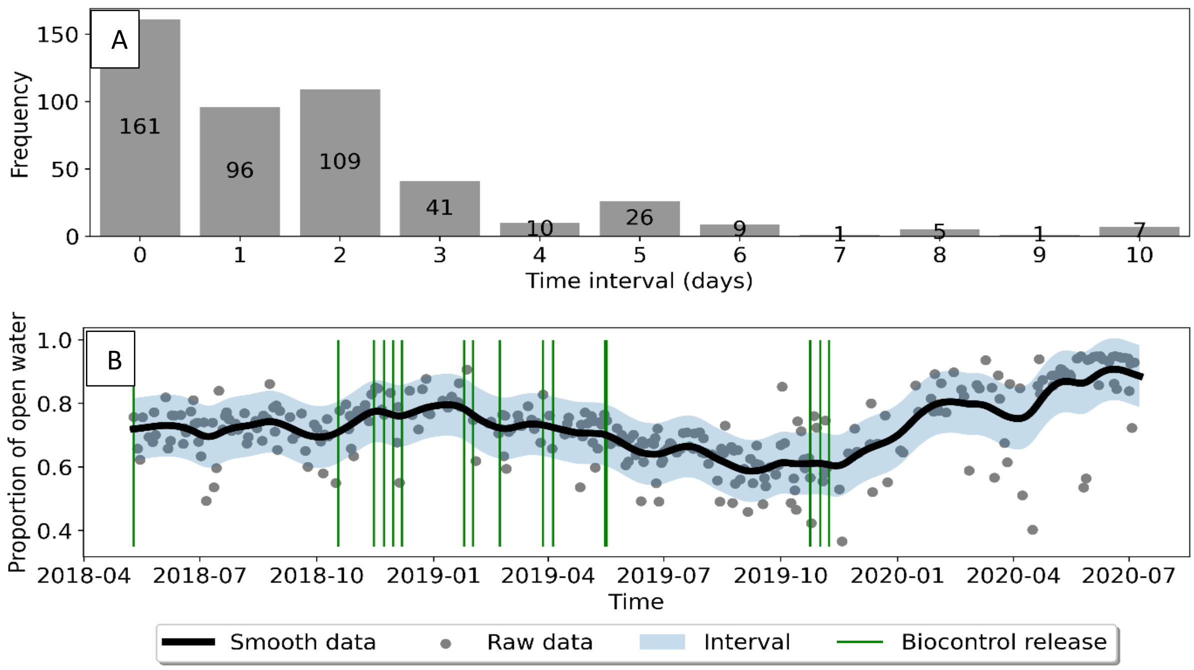


Figure 2. The observation interval between all (n = 310) optical (<5% cloud cover) and radar observations over Hartbeespoort Dam, South Africa for May 2018 to July 2020 used to detect open water (**A**). The smoothed daily open water data (n= 796) (**black line**) derived by applying a Kalman filter to the raw open water data (**grey dots, B**). Green vertical lines indicate the 16 actual biocontrol release events (Source: Centre for Biological Control, Rhodes University). The heavy green line is indicative of two biocontrol releases. Note, an interval of zero days indicates the total number of days that had more than one observation (up to three satellite observations) for the same day (n = 161). The **grey** (proportion of open water) points outside the **blue** region indicate outliers.

4.3.2. Extract time-series feature data

Environmental features used as the explanatory variables for forecasting the proportion of open water were extracted from the daily aggregated ERA-5 dataset (available from GEE). The ERA-5 climate data product provided daily satellite-based weather measurements of temperature, precipitation, wind speed and direction within the area of interest (Table 1). These weather variables were selected based on their high (daily) temporal frequency, and role in influencing the occurrence of water hyacinth and biocontrol agents (Hill and Olckers 2000; Miskella and Madsen 2019; Owens and Madsen 1995; Wilson et al. 2005). In addition, records of previous biocontrol releases at Hartbeespoort Dam were provided by the Centre for Biological Control (CBC), Rhodes University and were used as features. Biocontrol release features include the (i) date that a (ii) population (size) of (iii) *Megamelus scutellaris* or *Eccritotarsus catarensis* were released at a (iv) number of points around the dam. Note, during the study period (May 2018 – July 2020) the two mentioned biocontrol agents were the only two agents that were recorded as being released. All unique combinations of the four parameters composing a biocontrol release resulted in 4608 different biocontrol release actions (refer to Table 1B for the ranges of each feature considered).

Table 1. Details of features used for forecasting open water cover on Hartbeespoort Dam, Northwest province, South Africa. The features can be grouped into (**A**)- weather phenomenon sourced from ERA-5 daily aggregates or (**B**) biocontrol release parameters of actual release events obtained from the Centre for Biological Control, Rhodes University, South Africa.

Feature name	
A	Maximum air temperature at 2 m above the ground (°C)
	Mean air temperature at 2 m above the ground (°C)
	Minimum air temperature at 2 m above the ground (°C)
	Total precipitation (mm/day)
	Mean wind speed (km/hr) for the day
	Mean wind direction (degrees) for the day
B	Week of year (1-52), thereafter sine and cosine transformed (-1-1)
	The total number of biocontrol agents released (from 5000 to 11000 in intervals of 2000)
	The species of biocontrol agent released (either *Megamelus scutellaris* or *Eccritotarsus catarensis*)
	The number of points at which the biocontrol agents were released on a given date (from 1 to 3)
	Day of year (1-365), thereafter sine and cosine transformed (-1-1)
	Month of year (1-12), thereafter sine and cosine transformed (-1-1)

Additional features were created from the date of release such as, day of year, week of year and month. A sine and cosine transformation was applied to these cyclical variables, together with wind direction to capture their cyclic properties. For example, week 1 is closer to week 52 than it is to week 10. Since the number of unique biocontrol events within the study period were limited (n=16), the biocontrol release dates were encoded as the number of days since a biocontrol release. This allowed the data to be represented with a daily frequency (via interpolation between points), which increased the size of the dataset, and could potentially provide additional information on latent variables, such as dispersal distance or population density since there is a strong correlation between time of release and population density (Miller et al. 2021). In addition, the encoding ensured the biocontrol release date matched the daily temporal frequency of the climatic features.

4.3.3. The design and setup of the MAB

A MAB selects an arm (biocontrol release strategy), which is passed to the simulation environment that assesses the selected arms' effectiveness (sum of 70-day open water proportion, equation 1). The 70-day period is used since it corresponds to the minimum forecasting error (for details refer to section 4.4.1). Over multiple iterations, the MAB recognises actions that are likely to result in higher rewards (greater increases in open water or declines in IAAP cover).

$$reward = \sum_{t=0}^{t=69} proportion\ of\ open\ water \tag{1}$$

Where, t is the open water cover at a time step from day 0 (release day) to 69.

4.3.3.1. The simulation environment

In this study, the simulation environment - a Light Gradient Boosting Machine (LGBM) regressor (a supervised machine learning model) was used to forecast the proportion of open water (response variable, Section 2.1) and is dependent on lag time-steps (past values) of open-water data, weather, and biocontrol release data (explanatory variables, Section 2.2). Lag time steps allow for temporal dependencies among variables to be learnt. The LGBM model simulates the effect of a proposed biocontrol release strategy on open water cover for 70 steps (10 weeks) ahead of the proposed intervention. The 70-step open water cover forecast is then used to compute the reward (Equation 1).

Prior to selecting the LGBM model, 16 supervised regression models were evaluated in a time-based cross validation setting (Table 1). Time-based cross validation followed a prequential sliding window approach as recommended for datasets with a kurtosis less than two (For this study, kurtosis = -0.21) (Cerqueira, Torgo, and Mozetič 2020). A prequential sliding window method entails setting a specified number of train and test steps (window size). For each train and test window, a model is fitted and then its accuracy evaluated. The regressors' performance (mean absolute percentage error (MAPE) and inference speed) were used to select the final model (Table 2). Thereafter, the optimal forecast horizon and number of lag features (history) between the range of 14 days to 105 days (15 weeks) were investigated using time-based cross validation (Figure 3). The 14–105-day range investigated was dependent on the amount of available data. The greater the number of lag features used, the smaller the dataset available for modelling and typically the higher the error. Moreover, the investigated range was not extended since the error landscape (Figure 3) did not suggest lower error rates immediately outside the selected region and therefore did not warrant investigation. Once the optimal number of lag features and the forecast horizon that minimised the error was identified, the final model was trained on the entire dataset and was used to simulate the response of open water to biocontrol release.

4.3.3.2. Behind the scenes

Once an arm is selected by the MAB, the input data is updated with the selected biocontrol release strategy and is then fed as an input to the simulation environment (represented by the LGBM regression model) to forecast open-water for the subsequent 70 time-steps and compute the reward. The final reward and the associated action (known as action-value pairs or q values) are stored and influence future action selection. This process is repeated for T = 5000 trials. At the start of each trial, an arm needs to be selected by either exploring a new biocontrol release strategy or by exploiting a previously selected strategy. This exploration-exploitation dilemma is solved differently by different types of action selection strategies.

4.3.3.3. The action selection strategies

In this study two main types of action selection strategies were explored, epsilon-greedy and Thompson Sampling (TS). The straightforward epsilon-greedy strategy attempts to explore with a probability epsilon (ε) and exploit with a 1- ε probability. In this way, the arm with the highest recorded rewards is selected during exploitation, while during exploration an arm is randomly selected. However, as more information about the benefits (rewards) of the available arms becomes available, exploration should be discouraged. This is achieved by the epsilon-greedy with decay strategy whereby epsilon decays as the number of trials increase. In this study, ε is reduced by 1% after every 100 trials (Equation 2). Moreover, the effect of varying the value of ε on the overall reward is investigated.

$$\varepsilon \leftarrow \varepsilon\,(1 - 0.01) \qquad (2)$$

An agent following the TS policy applies Bayesian inference (Russo and van Roy 2016), in that the agent updates its beliefs (confidence of obtaining a reward) as more evidence becomes available. During each trial two parameters that determine the reward distribution of an arm, are updated. These two parameters are the reward mean and reward precision (1/variance) - an estimate of confidence. This contrasts with the epsilon-greedy policy, where at the start of each trial, the action with the highest estimated reward is selected, even if the confidence in that estimate is low. Thompson Sampling instead samples from the reward distribution of each action, selecting the action with the highest returned reward. Since actions that have been tried infrequently have wide mean reward distributions, they have a larger range of possible reward values. In this way, a biocontrol release strategy that currently has a low estimated mean reward but has been tested less than a biocontrol release strategy with a higher estimated mean, can return a larger sample value and can therefore become the selected biocontrol release strategy (arm). After repeatedly selecting an arm, the mean reward of the arm converges towards the actual reward and the confidence in the estimated reward increases.

4.3.4. Implementation details

Satellite-based variables (open-water cover and weather data) were prepared and extracted from GEE using the python API. Tabular data pre-processing was done using pandas v1.3.4 (McKinney 2011). The LGBM regressor for forecasting was built using the lightgbm package and scikit-learn v1.02 (Ke et al. 2017; Pedregosa et al. 2011). Figures were created using matplotlib v3.4.3 (Ari and Ustazhanov 2014). All code is available on GitHub here.

4.4. Results

4.4.1. Model performance: simulation environment

After assessing 16 candidate models to forecast open water proportion (Table 2), a LGBM regression model was selected owing to its low mean absolute percentage error (MAPE) (39.98%) and faster training and inference time (TT(s), ~ 16 times faster) compared to the next most accurate model (Natural Gradient Boosting). Note, lower error values result in more realistic simulated responses of open water proportion to a biocontrol release action, reward calculation and the subsequent ranking of optimal biocontrol release dates. Overall, decision tree-based boosting models outperformed both linear and kernel-based models suggesting the non-linear relationship between open water and the features considered (Table 1). With the exception of the LGBM model, all models had a MAPE error > 50 % which corresponds to inaccurate forecasts (Lewis 1982). Importantly, small increases in the error caused unrealistic simulation outcomes i.e., open water proportions greater than one and temporally incoherent forecasts or sudden spikes. For this reason, it was important to prioritise accuracy and select a forecast horizon that minimised the error as opposed to prioritising time dependent biocontrol agent dispersal ranges, or life-cycle durations. In a scenario where the simulation environment is more accurate and less sensitive to changes in accuracy, it may still not be ideal to directly incorporate the biology of the biocontrol agent into the model, especially if the forecast horizon can be reliably extended to 365 days (water hyacinth cover shows a seasonal cycle). This can be attributed to scenarios where the use of biological data may be more harmful or constraining than it is beneficial. Simultaneously, the biology of the weed and biocontrol should not be and, in many cases, will represent the lens through which the MAB results can be interpreted. The use of this information (for example, a statistical-based dispersal model) at a later stage of the decision-making process will be less sensitive to the generalisability, availability or accuracy of the model and will not limit the MABs' results to areas with suitable data, nor degrade the MABs' performance through the propagation of errors, nor increase the risk of using inappropriate models under different climates or local climates.

Table 2. The time-based cross validation (CV) MAPE performance of 16 candidate models assessed to forecast open water. The time-based CV used to obtain mean (standard deviation) MAPE scores are averaged over five iterations. For each iteration, there were three folds that each had 230 training steps and 185 test steps. All models were trained with 56 lag steps of each feature and thereafter, forecasted 70 steps-ahead. The time taken (TT, seconds) to calibrate and validate each model for five iterations is also reported. Results are sorted by descending MAPE values i.e., best performing models (lowest error in bold print) to poorest performing models (highest error).

MAPE	Model name	TT(s)
0.3998	*LGBM Regressor	117.7
0.606 (0.114)	**NGB Regressor	1684.4
0.676 (0.051)	Random Forest Regressor	43.4
0.679 (0.100)	AdaBoost Regressor	63.6
0.792 (0.151)	Gradient Boosting Regressor	73.5
0.799	CatBoost Regressor	769.5
0.947	***XGB Regressor	153.3
1.092 (0.778)	Decision Tree Regressor	44.4
1.117 (0.067)	Extra Trees Regressor	83.2
1.701	Lasso	43.8
1.788	ElasticNet	46.0
5.463	Bayesian Ridge	40.2
5.586	Kernel Ridge	49.8
5.587	Ridge	44.1
5.606	Linear Regression	39.8
7.299	Orthogonal Matching Pursuit	40.8

*LGBM: Light Gradient Boosting Machine, **NGB: Natural Gradient Boosting, ***XGB: Extreme Gradient Boosting.

After model selection, the number of lag step features used during model fitting and the forecast horizon were investigated to determine the optimal performance (Figure 3). Fewer lag step features are preferred because of the reduced input data requirements for a forecast, whilst longer forecast horizons are preferred since they are more likely to capture long term responses of open water to a proposed biocontrol release event. The results indicated that 56 lag step features and a forecast horizon of 70 days resulted in the lowest MAPE error (39.98%) and a medium range forecast horizon (10 weeks). The 70-day forecast horizon was prioritised over longer forecast horizons, since it resulted in the lowest MAPE error and more realistic simulations with reduced spikes in open water and increased temporal coherency.

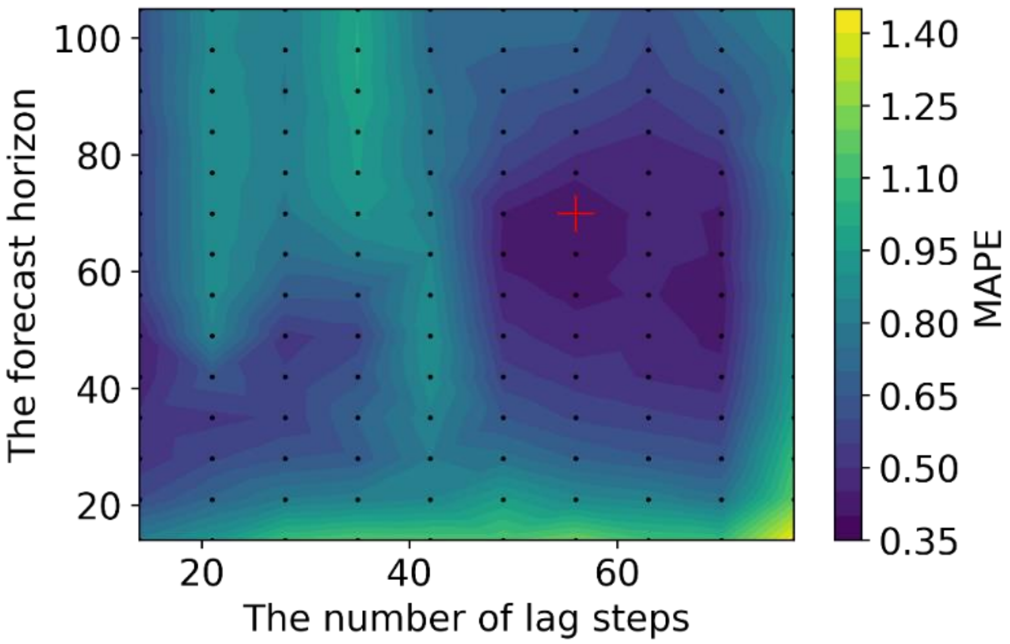


Figure 3. The Mean Absolute Percentage Error (MAPE) landscape for the Light Gradient Boosting Machine (LGBM) regressor when applied in a time-based cross validation setting to forecast open water proportion. Note, lower error values are indicative of more accurate models. The optimal combination of lag steps and forecast horizon is 56 and 70 respectively (red plus symbol). The LGBM model shows reasonably accurate results (20<MAPE<50, (Lewis 1982)) for all combinations of lag steps and forecast horizons tested.

When using the LGBM regressor to forecast the proportion of open water 70 steps ahead, a moderate error is observed for the time-based CV (Figure 4) that corresponds to a mean MAPE test-score of 39.98%. This suggests the model has accurately learnt to forecast open water and is therefore likely to accurately simulate responses of open water to hypothetical biocontrol releases.

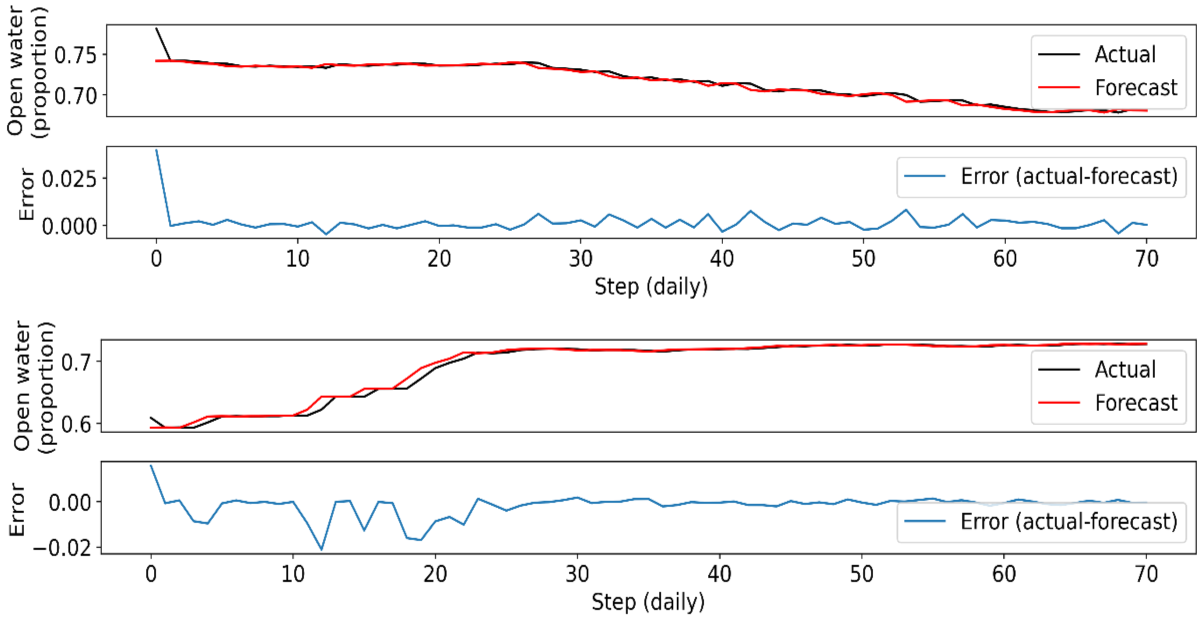


Figure 4. A comparison of LGBM forecasts of proportion of open water (area of open water/ area of waterbody) at Hartbeespoort Dam 70 days into the future. The difference between the actual and forecasted open water proportion has been shown - indicating minor forecasting errors. The forecasts are based on the test data (discussed in section 4.3.3.1.).

4.4.2. MAB performance

Three types of action selection strategies were investigated, both epsilon greedy and epsilon greedy with decay considered the effect of different epsilon values on overall reward - higher overall reward values correspond to the earlier identification of more effective biocontrol release actions. The epsilon (0.1) greedy with decay agent (Figure 5, purple line) that outperformed the remaining agents was able to identify an optimal biocontrol release very early as indicated by the high average reward at the start (Figure 5).

A Thompson Sampling agent, which tends to explore many more actions, resulted in the lowest overall reward among the seven action selection strategies, but still had a much higher reward than a random action selection strategy that randomly selected a biocontrol release strategy at the start of each trial. A random selection corresponds to a scenario whereby no context is considered and whereby the agent does not follow a protocol for exploitation, and therefore any difference in the overall reward between this random strategy and any of the remaining strategies corresponds to the benefit of strategic exploitation. The lower overall reward of the Thompson Sampling action selection strategy suggests that the agent exploits sub-optimal actions much more frequently than the other MAB agents.

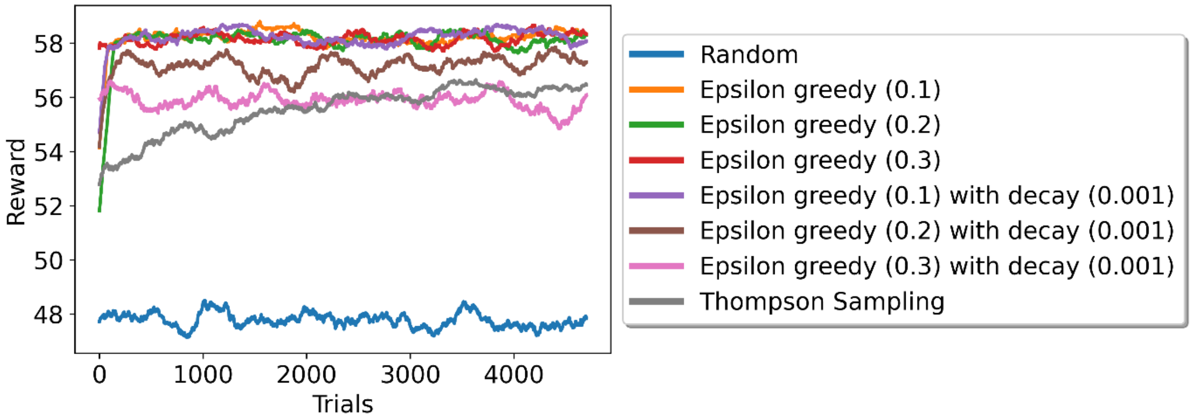


Figure 5. Running (window size = 300 trials) mean rewards across T = 5000 trials for seven action selection strategies in relation to a baseline agent that randomly selects a biocontrol release strategy at the start of each trial. The epsilon (0.1) greedy with decay agent (**purple line**) provides the highest overall reward. Reward refers to the accumulated 70-day open water proportion after a biocontrol release. The overall reward for each of the seven action selection strategies are also indicated (Table 3**, below**).

Table 3. Summary statistics for chart (above) showing overall mean (std) reward across four 5000 trial iterations. Bold text corresponds to the agents with the highest rewards.

Action selection strategy	Overall reward
Random	239,555 (517)
Epsilon (0.1) greedy	289,987 (893)
Epsilon (0.2) greedy	285,257 (550)
Epsilon (0.3) greedy	279,743 (752)
Epsilon (0.1) greedy with decay (0.001)	**290,840 (714)**
Epsilon (0.2) greedy with decay (0.001)	285,951 (547)
Epsilon (0.3) greedy with decay (0.001)	284,145 (776)
Thompson Sampling (TS)	256,109 (20344)

4.4.1. Evaluating potential biocontrol release actions

To better understand the biological implications of the suggested biocontrol release actions (Table 4), three questions were posed.

Question 1: What is the optimal biocontrol release strategy for each of the best action selection strategies in each of the three categories? (Table 4)

By considering the suggested release date, resource planning and management can be adapted to meet the requirements of the proposed action (Table 4). The highest open water reward obtained by the Thompson Sampling agent (57.407) is slightly lower than that obtained by the epsilon greedy agent (59.412). This implies that the TS agent has not converged on the optimal biocontrol release action (Table 4).

Table 4. Most optimal biocontrol release dates (arms) identified by the top-performing action selection strategies in each of the three categories. Optimal actions correspond to the most selected biocontrol release actions made by each of the three agents. These actions are between 8-27% more effective than actual biocontrol releases in terms of accumulated open water proportion. The average (sum) reward corresponds to the average (sum) of open water proportion over 70-days after the date of release.

Action selection strategy	Reward	Date of release
Epsilon (0.3) greedy	0.85 (59.412)	27 Feb 2020
Epsilon (0.1) greedy with decay (0.001)	0.85 (59.412)	27 Feb 2020
Thompson Sampling	0.82 (57.407)	14 March 2019
Least effective actual release	0.63 (44.437)	24 Oct 2019
Most effective actual release	0.78 (54.568)	23 Nov 2018

Question 2: How does open water response vary with the optimal (highest overall reward) strategy, the worst identified strategy, and the actual implemented strategy?

The second question juxtaposes the response of open water under three scenarios, (i) actual circumstances (Figure 6, black line), (ii) implementing the least effective biocontrol release strategy (orange line) and (iii) implementing the optimal biocontrol action from the 4608 investigated actions (Figure 6, blue line). Biocontrol advocates may be able to justify the cost of a biocontrol release in relation to its projected benefit (decline in IAAP cover). The reward for the most effective strategy is ~8-25% greater than the reward for actual biocontrol releases (green vertical lines, Figure 6) during the study period.

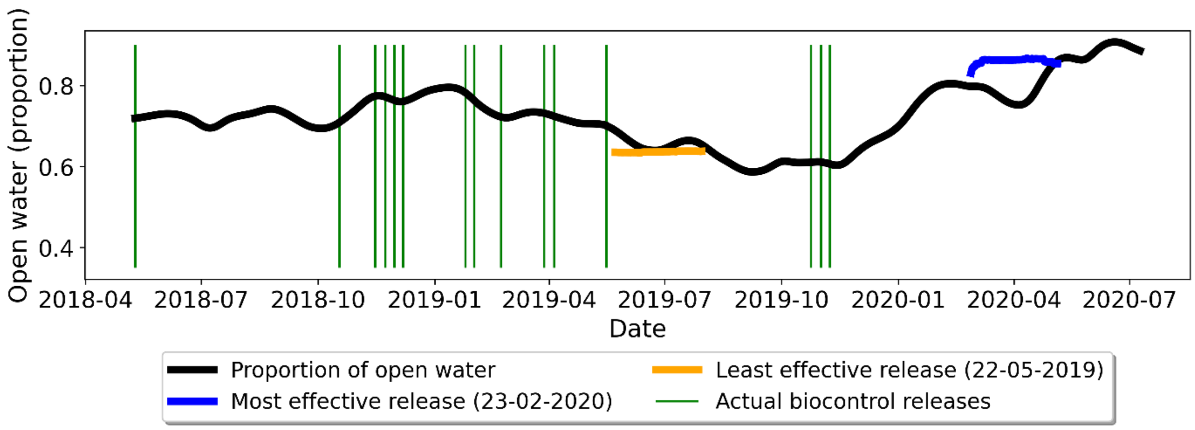


Figure 6. The (non-linear) response of open water for the suggested most effective biocontrol release strategy (**blue**, 23 Feb 2020) and the least effective biocontrol release strategy (**orange**, 22 May 2019) for 70 days after the respective release in comparison to the actual releases (**green lines**) and actual open water cover (**black line**) on Hartbeespoort Dam.

Question 3: For an optimal biocontrol release strategy, how does the reward vary by release week, species of agent released, number of agents released and the number of points on the dam at which the biocontrol agents are released?

This question explores the sensitivity of the reward (the forecasted 10 weeks of open water proportion) to changes in the parameters of a biocontrol release strategy. From an operational perspective, this highlights the cost of delayed interventions or if insufficient resources are available, the acceptable trade-offs that can be made and their associated costs.

For the optimal actions, the population size of biocontrol agents released, the number of release points, and the species of agent released had negligible influence on the reward (Figure 7). Consequently, the effect of the release week (week of the year) on the reward is only considered (Figure 7). The release week shows a cyclical pattern, whereby early in the year (late February- mid-March, late summer) - the most effective summer release period is followed by a period of declining release effectiveness and then an end-of-year period (mid-November), which gradually becomes more effective going into the start of the subsequent year. The effective summer period corresponds to a period of high insect reproductive rates and large population sizes (Byrne et al. 2010; Miller et al. 2021). The most effective release week (week 8 - end of Feb) results in a more than 23% increase in average open water across the 70-day period compared to the least effective release weeks (week 28 (end of June) – week 45 (mid-November)).

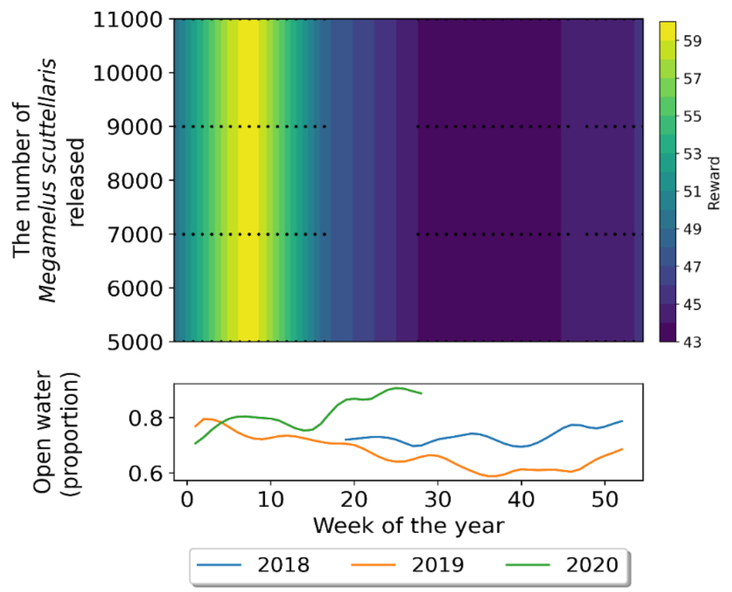


Figure 7. The cyclical effect of release week (week of the year) and the negligible effect of the number of released *Megamelus scutellaris* on the accumulated open-water proportion for the subsequent 70 time-steps after a release (reward) averaged across two to three years (2018-2020). Higher rewards are indicative of more effective biocontrol release weeks. The year-on-year actual open water proportion is also shown at the bottom.

While the effect of previous biocontrol releases on simulated releases may not always have been considered directly, the occurrence of pre-existing biocontrol agents may have been incorporated by the inclusion of the ‘days since the last biocontrol release’ variable. For example, for each day of the 70-day forecast, 56 prior days data is used. This may include information on prior actual releases. However, for the most optimal release period identified (late Feb to mid-March 2020), there are no previous releases that fall within the 56-day input data that the model uses to make a forecast.

4.5. Discussion

A combination of a simulation environment (regression model) and a MAB framework is proposed for a novel application to identify potential biocontrol release dates that are likely to result in the maximum proportion of open water. An analysis of the response of open water to different simulated biocontrol release actions (n=4608), highlights the effect of biocontrol release date on open water. These results suggest a decrease in the number of potential biocontrol release weeks that should be investigated in a real-world scenario, representing a shorter path to the successful implementation of an effective biocontrol release strategy.

Whilst the proposed biocontrol release actions’ actual effects on open water are yet to be verified within a real-world scenario, based on simulated responses, the framework suggests that late summer (February to mid-March) was the optimal period to release biocontrol during the study period (10 May 2018 – 09 July 2020), however, this was attributed to the collapse of water hyacinth and a subsequent algal bloom. This points toward the need for systems-level management approaches (discussed later). This period is followed by a steady decline in biocontrol effectiveness until week 27 (early June, the start of Winter) and an increase in effectiveness from ~week 45 or mid-November - close to the start of Summer to mid-March of the subsequent year. These results partly align with current knowledge and recommendations, that biocontrol releases should be made in spring just after the first sign of host-plant recovery post-winter (September-October) (Coetzee et al. 2021; Miller et al. 2021). This is in response to a (~2 month) lag in the timing of peak population growth of the over-wintering insect in relation to the weed (Miller et al. 2021). However, the increase in effectiveness of biocontrol releases (Figure 7) from mid-November onwards suggest that releasing biocontrol slightly later than at the start of spring may translate to more favourable outcomes since the warmer conditions will be more suitable to the thermal requirements of *M. Scutellaris* (May and Coetzee 2013), the dominant species released at Hartbeespoort Dam (Coetzee et al. 2022). This is supported by the lower developmental threshold ($t_0 = 11.458$) being higher than the cold temperature conditions experienced at the dam until October (Coetzee, Byrne, and Hill 2007). This suggests that an earlier release (September- October) may not be suitable for the development of the insect and that a mid-November release, ~2 month prior to the peak water hyacinth coverage and ~5 months prior to peak insect density (late Summer – early Autumn) should allow for eggs to develop to adults at least one and three times before the respective peaks (May and Coetzee 2013; Miller et al. 2021). The later November biocontrol release may ensure that water hyacinth is present throughout summer and reduces the chances of a secondary invasion. If there is water hyacinth present, a second augmented release may occur between late February and mid-March. In this way, the release will take advantage of the warm temperature conditions for optimal insect growth and can augment the population that will overwinter (at the start of June). During winter a natural decline in plant quality and quantity occur, combining with low temperature to slow or halt insect reproduction and prolong the biocontrol agents' development period (Miller et al. 2021).

The suggested two releases assume that the current inundative biocontrol release regime is halted. In this scenario there is likely to be water hyacinth present during late Feb to mid-March period. This is supported by the presence of 15-20% aquatic vegetation cover during late Feb to mid- March 2017, 2018, and 2019, prior to the start of the current inundative release regime in October 2019 (Coetzee et al. 2022). The repeated annual summer resurgence of water hyacinth (Figure 6) suggests that the pre-existing biocontrol populations have likely had minimal effects on the occurrence of water hyacinth because they take long to build up to damaging numbers after the winter population bottlenecks if not augmented. Currently biocontrol agent populations are augmented from early Spring (Coetzee et al. 2022).

The results indicated negligible differences between the responses of open water to the two biocontrol agents (*Megamelus scutellaris* and *Eccritotarsus catarensis*), or the benefits of releasing biocontrol at multiple points on the dam, nor the effect of the number of biocontrol agents released during the study period (10 May 2018 – 09 July 2020). This is indicated by, for example, the number (5000-11000) of biocontrol agents not influencing the open water reward (Figure 5). Since the model is only sensitive to a changing release date, despite having an equal number of samples for all four factors could suggest; 1) ineffective variable encoding, 2) insufficient data to understand the remaining three complex variables or alternatively, 3) the effect of the species of agent released and their population size are much less important than week of release under temperate climatic conditions such as those experienced at Hartbeespoort Dam. Further research aimed at verifying this outcome may prove valuable, owing to the high insect rearing and release costs.

To improve the applicability of this framework to other sites, the forecast model (simulation environment) would benefit from increased spatial generalisability with the inclusion of data available for climatically similar water systems and/or improved ability to use limited data i.e., improved sample efficiency by data augmentation. In this study, some of the efforts that were taken to ensure the generalisability to other water systems include, 1) the use of proportion of open water as the basis for computing reward, 2) the use of a sensor-agnostic open water detection algorithm that is widely applicable, and 3) the use of globally available satellite weather data.

The benefit of using open water as the response is three-fold; 1) the extent of surface water may vary for small to large waterbodies or waterbodies with strong seasonal fluctuations in water extent, thus using the proportion of open-water allows for these values to be scaled between 0 and 1 and therefore improves the spatio-temporal applicability across sites, 2) open water detection benefits from the large body of research on surface water detection and 3) since open-water proportion is the reciprocal of proportion of aquatic vegetation, maximising the reward based on open water proportion, subsequently minimises the proportion of aquatic vegetation and therefore identifies biocontrol release actions that correspond to the minimum cover of other IAAP species. Therefore, this corresponds more closely to a water-system level approach to managing IAAP infestations, as opposed to species-level management whereby the area occupied by a single species is minimised and may inadvertently assist to address the succession of secondary invasions by other IAAPs (Coetzee et al. 2021; Strange et al. 2019).

To improve the utility of the proposed framework for water hyacinth management, future biocontrol releases could capture the variance across the factors that need to be optimised. For example, varying amounts of insects should be released in increments of 2000. This will provide representative data to train simulation models. Subsequently, the predictive accuracy will increase, and the epistemic (model) uncertainty will decrease (Kaur, Pannu, and Malhi 2019; Yang et al. 2021). The simulation environment only provides point estimates, without any measures of uncertainty. Recent work has introduced conformal prediction for time series and will be a useful addition to improve user trust and adoption (Xu and Xie 2020). Conformal prediction provides prediction intervals with statistical guarantees (viz. coverage guarantees) for example, that the actual prediction lies within the provided intervals with a 95% guarantee. These uncertainty estimates may then be combined and used by the multiarmed bandit. Users may then be able to gauge when the simulated output should be considered with extra caution.

Additionally, any confounding factors that may influence the effectiveness of a biocontrol release event need to be measured in a consistent manner across sites and made publicly accessible. These factors may include the presence and the density of pre-existing biocontrol agents at the time of release, the date, duration, and extent of herbicidal spray events and/or mechanical removal. There have been numerous studies that have reported synergistic effects between different species of water hyacinth biocontrol agents (Ajuonu et al. 2007; Firehun et al. 2013; Marlin, Hill, and Byrne 2013; Moran 2005). Hartbeespoort Dam has had multiple species of biocontrol agents released prior to 2018 that may partially be responsible for the open water responses modelled in this study. However, owing to the lack of data of multi-species co-occurrences at each release for Hartbeespoort Dam during the study period, this was not considered. The availability of such in-situ data will be complementary to the proposed framework and has the potential to make the simulation model more accurate and applicable to real-world scenarios.

Whilst the model does not directly include population density of the biocontrol agents owing to the lack of consistent high-quality population density during the study period, this may be a preferable scenario. Population density data is likely to be challenging to collect for a large number of infested sites, in a consistent manner and on a regular basis. Instead, the limited data should be used to identify variables that function as proxies for biocontrol agent population density and are scalable to large extents. For example, satellite-based temperature and the 'time since the last biocontrol release' variable used in this study.

4.6. Future work

As with the adoption of any new systems, there are initial challenges that will need to be overcome prior to wide-scale adoption. Similarly, ML based systems will require active development before being widely adopted by government. These initial challenges and the lack of understanding of the inner workings of the introduced system may hinder its adoption. However, it is currently the only option that has desirable characteristics for scalability across many sites, and that can render tailored biocontrol release actions in a near-real-time manner. For this reason, water managers that are also heavily research focused (for example, the CBC, Rhodes university) are the ideal candidates to trial and carry out additional experiments to provide the empirical evidence to verify and to support the development of the models. At the same time, this should not be trialled for sites where the costs of failing are high. For example, Hartbeespoort dam has a lot of public attention, and any failed biocontrol trials may also be perceived by the public as the failure of biocontrol. Therefore, a site with less public attention may be more conducive to scientific experimentation.

Two extensions to the MAB that may be useful in future work include the use of a cost and cost weighting factor, and improvements to the reward function. A cost factor may be useful to mimic the behaviour of only selecting a release strategy if the benefit outweighs the cost. A cost-weighting factor (a user specified value) would allow the user to take into consideration any fluctuations in the cost attributed to different development periods for different insect species, higher rearing expenditure during winter, different transportation costs for specific water bodies or to account for inflation. The decision to release biocontrol for an algal bloom during late February 2019 (based on the manual inspection of satellite imagery for this period), the optimal release period, is a symptom of the well-known reward misspecification problem in reinforcement learning and can be addressed by accounting for algal cover in the reward or by creating a simulation environment that forecasts algae, macrophytes and open water separately (Pan, Bhatia, and Steinhardt 2022). NDVI could be incorporated with open water to provide a more stable reward function than the use of open water alone. Open water shows drastic day-night, and daily fluctuations. It may also allow for a systems-level intervention approach, whereby, the water hyacinth present only occurs to sustain biocontrol populations but at a low-medium health that can slow-down vegetative reproduction and prevent secondary invasions.

Since the model can only forecast 70-days ahead (20% of the year), the reward only accounts for short-term declines in open water, the MAB, in its current form, may only be useful to pre-empt a biocontrol release up to 70-days ahead as opposed to identifying a long-term optimal period that should be adopted for annual releases. This is supported by Figure 6, whereby a (temporally) local decline in open water was identified as an effective time for an intervention. This can be addressed by a longer forecast horizon and will require increasing the spatio-temporal coverage of the dataset used to calibrate the simulation environment. Next, seeing that different circumstances may present itself on a yearly basis for the same site, the MAB will need to be used in an online system to be useful. Here, the MAB will need to be run more frequently, up to a daily basis to suggest interventions. Results may then be summarised across multiple years of data to identify periods that are overall/generally more suitable for biocontrol at a site. Future work should therefore prioritise increasing the dataset size and improving the reward function Given that different circumstances may present on a yearly basis for the same site, the MAB will need to be used in an online system to be useful. Where it will run more frequently, up to daily, to suggest interventions. Results may then be summarised across multiple years of data to identify periods that are overall/generally more suitable for biocontrol at a given site. Future work should therefore prioritise increasing the dataset size and improving the reward function.

This chapter aimed to formulate a framework to determine an optimal release week across multiple years. However, a more holistic strategy that spans multiple releases may be better adapted towards effective biocontrol. In this scenario, a move from the MAB framework towards a framework that will directly suggest a multi-release strategy should be investigated. By modelling this as a reinforcement learning problem, feedback between an action and the environment are taken into consideration. This contrasts with the MAB framework, that only considers the state of the environment on the selection of an action and does not consider knock-on effects of an action on future decision-making. Owing to this characteristic, a complete reinforcement learning framework may likely be able to learn a multi-release solution - a potentially better solution (Coetzee et al. 2022).

4.7. Conclusion

This study applied a low cost and water-system agnostic framework, which combines a multi-armed bandit with a simulation environment. The multiarmed bandit proposes a biocontrol release while the simulation environment simulates the response of open water to the biocontrol release and therefore, allows for the evaluation and ranking of the proposed biocontrol release's effectiveness to increase open water. Through this framework, the multi-armed bandit was able to capture the effect of release week on biocontrol release effectiveness and when coupled with knowledge of the weeds' growth cycle (domain knowledge) was interpreted to suggest two periods (between late Feb – mid-March, and mid-November) that should be prioritised for biocontrol release. Simultaneously, the results suggest that additional release data is required to optimise other biocontrol release parameters, such as the population size and species of biocontrol agent to be released. While these results remain limited to Hartbeespoort Dam, the framework remains applicable to any water body, provided there is prior biocontrol release data available. In lieu of this, recommendations to improve the utility of the proposed framework is provided.

5. Chapter 5: Synthesis

5.1. Summary and research implications

Water hyacinth is widespread, it is currently considered the worst Invasive Alien Aquatic Plant (IAAP) and has been projected to expand its distribution (Kariyawasam, Kumar, and Ratnayake 2021), adding pressure to already scarce freshwater resources in countries like South Africa. The main factor encouraging prolific water hyacinth infestation is the eutrophic-hypertrophic water conditions (Coetzee et al. 2022). While countries combat nutrient-enriched water, water hyacinth management success needs to improve. Improved environmental management can be spurred on by the numerous developments in Artificial Intelligence (AI), remote sensing, Earth Observation (EO) and cloud computing platforms, as shown here. However, the application of these technologies in invasion ecology are considered emerging in comparison to forestry (Boyd and Danson 2005; Larson et al. 2020). Therefore, this thesis sought to introduce new image processing (computer vision) methods and combine pre-existing AI methods with strong spatio-temporal extensible characteristics to shift management interventions on water hyacinth towards a more pre-emptive and automated approach. Spatio-temporal extensibility refers to the utility of computer vision methods across regions and time because of their reduced sensitivity to spectral shifts attributed to changing atmospheric, and environmental conditions (Thamaga and Dube 2018a). This is important because any satellite derived insight that could be used for IAAP management needs to be provided repeatedly across wide extents in a consistent manner. Through the subsequent projected increases in management effectiveness, a decline in water hyacinth infestation across the country could be realised.

This thesis takes advantage of freely available satellite data, free and open software, and free cloud computing infrastructure to localise water hyacinth and monitor changing aquatic vegetation extents. Knowing the current location of water hyacinth is the basis for any intervention. To localise water hyacinth, a system was built (Chapter 1) that takes a satellite image and returns the areas (pixels) that correspond to water hyacinth. This thesis built on previous work by providing a scalable method for mapping water hyacinth and monitoring aquatic vegetation. The distribution of water hyacinth was mapped, for the first time, across a large extent (> 1.2M km^2, Chapter 2) 67 times greater than the extent of previous studies, but with comparative accuracy. Providing water managers with the unprecedented capability to remotely identify, monitor infestations and evaluate the effectiveness of management interventions.

The inferred water hyacinth distribution only provides a snapshot of the weeds' distribution. To supplement the temporally sparse information, the sensor agnostic (Sentinel-1, Sentinel-

2, and Landsat-8) and spatially extensible method based on thresholding (Chapter 2) and unsupervised clustering (Chapter 4) will be useful. This will allow managers to monitor the extent of water and aquatic vegetation on a near-daily frequency and with greater trust than the species-level water hyacinth distribution which is difficult to visually verify. This reduced verification ability (for the water hyacinth distribution) is of less concern in South Africa because incorrectly classified water hyacinth pixels likely correspond to other IAAP species that are also targeted for management. From an operational perspective, identifying IAAPs at a species level and on a per-pixel basis poses an adoption challenge owing to reasonable scepticism from water managers. In this scenario, the aquatic vegetation mapping provides a suitable alternative since it is easily verified (visually) against the raw satellite data from which the aquatic vegetation was derived and therefore, more easily adopted. In addition, the mandate of individual water managers is typically limited to a small portion of local water systems for which the problem IAAP species is typically known a priori on a per site-basis, rendering the species-level information redundant in some local scenarios but still useful at higher scales of management. While these arguments oppose the need for a satellite-derived species distribution, if such distributions are developed in a way that promotes trust and adoption, then research, and development of state-of-the-art species-level monitoring systems, early detection, and early warning systems will accelerate. At present, the availability of this information from this thesis can assist risk prioritisation, resource allocation and form a basis for the development of additional methods (for example, Species Distribution Models (Chapter 3)) and datasets that rely on species occurrence information at the waterbody level.

In Chapter 3, the response of water hyacinth to a large number (82 of a potential 140 considered variables) of socio-economic and ecological features was assessed in a spatially explicit manner to identify the drivers influencing the occurrence of water hyacinth in South Africa. This was achieved by using explainable Artificial Intelligence (xAI) that makes the behaviour and predictions of machine learning algorithms or systems understandable to humans. In this case, a Species Distribution Model (SDM). While this work remains the first of its kind for an invasive species, there have been similar applications in other environmental domains (for example, landslide susceptibility (Zhou et al. 2022), climate change adaptation (Chakraborty et al. 2021), air pollution and water quality (Stirnberg et al. 2021; F. Wang et al. 2021)). This study emphasized the importance for management interventions to consider the broader ecological context of an infestation. The relative importance of the factors driving water hyacinth occurrence at a waterbody level can be used to guide managers in developing an expert rule-based system to prioritise the type of management that might be required, in advance.

This analysis also revealed the overall importance of drivers and groups of drivers at a national scale. For example, the minimum temperature feature is by far the most important (strongest predictor) factor limiting the occurrence of water hyacinth in South Africa. Nevertheless, more than 60% of the significant land use land cover predictors (38/60) of water hyacinth occurrence are related to socio-economic features. This emphasizes the negative role of anthropogenic influences that improve the habitat suitability for the occurrence of water hyacinth and increase the difficulty of managing water hyacinth within a changing world, characterised by increasing levels of human modification, and climate change (Kriticos and Brunel 2016). The role of climate change is supported by the climate related species-feature interactions. For example, increases in temperature and rainfall generally correspond to increased water hyacinth suitability and susceptibility (Chapter 3, Figure 6a & S1c). To reinforce some of the conclusions reached here, more research is needed to verify these landscape level interactions identified in this study.

Lastly, Chapter 4 adapted a reinforcement learning based framework, the multiarmed bandit, for the novel application of evaluating simulated biocontrol release events, subsequently narrowing down the potential biocontrol release events that should be investigated in the real-world to optimise a biocontrol programme. This was achieved by modelling the response of open water proportion to climatic factors and actual biocontrol release events. Thereafter, the calibrated model was used to simulate the open-water response to more than 4000 synthetic biocontrol releases. The results indicate the importance of release date for effective management using biocontrol. Moreover, this work represents a step towards an automated and pre-emptive management method for selecting optimal biocontrol release dates. While the framework remains widely applicable to sites with the required training data, it was only applied at a single water system - Hartbeespoort Dam, South Africa. In future work, the simulation model should be calibrated on many more sites and updated with recent climate and biocontrol release data. Thereafter, it should be applied, and the results verified in the real-world. The outcomes from real-world experiments may then be fed into the introduced framework to form an active learning loop.

Since the MAB framework was configured to maximises open water cover, the optimal biocontrol release dates should minimise the likelihood of any aquatic vegetation cover. This includes *Salvinia minima*, or cyanobacteria blooms at Hartbeespoort Dam that result in secondary IAAP invasions (Ballot et al. 2013; Richards 2022; Strange et al. 2019). While such a reinforcement learning based system is novel in invasion biology, there is a vast body of knowledge, algorithms, and best practices to adopt from the broader reinforcement machine learning community that may result in more holistic biocontrol management actions, or new integrated management interventions. For example, a short-term accumulation of water hyacinth followed by a mass-release of biocontrol may hypothetically result in a sharper decline of water hyacinth cover.

Developments in AI, increased data availability and interdisciplinary collaboration (for example, ecologists, remote sensing practitioners, data scientists and policy makers) are three of the main factors that will likely lead to developments/improvements in the introduced methods, landscape ecology, and addressing global environmental challenges (such as, natural disaster management, climate change, invasive species management and conservation) and future life changing technologies (such as autonomous vehicles, or the next generation of wireless connectivity (6G) and applications such as Google search).

The hierarchical nature of the water hyacinth discrimination method (Chapter 2) is valuable across a wide variety of mapping problems, including the mapping, and monitoring of terrestrial invasive species. In the future, Chapter 2 and chapter 3 (xAI) should be combined into a single system. Chapter 4 (the MAB) has the greatest capacity for development and management. The MAB framework relies on the previous chapters monitoring system and therefore provides a useful conjunction of different systems that can be incorporated into a single real-time system for managing water hyacinth and has a clear path to benefit aquatic system management. Moreover, it is applicable to any other ecological setting that deals with decision making under uncertainty.

5.2. Combining the developed tools

The introduced tools should not be used in isolation as actual conditions may be different from the model outcomes and require additional verification. By combining Landsat and Sentinel image data (Chapter 4) gaps between successive water and aquatic vegetation observations were reduced, resulting in a dense time series. In this way, the utility of using freely available medium resolution data was amplified (Pahlevan et al. 2019). This is pivotal for the development of early warning systems that can be developed in conjunction with the probability of occurrence data (Chapter 3) and water hyacinth distribution map (Chapter 2) to identify areas that are not yet infested and have a high risk of water hyacinth invasion. These high-risk areas can then be prioritised for monitoring (Chapter 2 or Chapter 4) or early detection.

For pre-existing infestations, the type of management may be chosen based on the drivers of water hyacinth occurrence (Chapter 3). For these sites, a multiarmed bandit framework may be used to narrow down effective biocontrol release periods (Chapter 4). If the introduced methods are used in this way, and if the outcomes are subsequently verified by field observations the management of water hyacinth could become significantly more successful.

While the research of AI-powered satellite image analysis techniques is numerous, fast evolving and can benefit invasion biology, they have not yet been adopted at a similar rate during management. There are an encouraging number of success cases that provide evidence of the utility and adoption in operational settings. Examples of AI-powered satellite image analysis include tracking progress towards environmental-related targets and goals (for example, the Montreal protocol (Cracknell and Varotsos 2009)) and the Sustainable Development Goals (SDGs). 30 of the 231 unique socio-ecological indicators used to monitor progress towards achieving the SDGs are based on EO data and range from conservation (Ashutosh and Roy 2021; Joshi et al. 2016), disease prevention (Frake et al. 2020), reducing deforestation (Pro, Watcher, and Atlases 2014), monitoring coastlines (Ding et al. 2021; Vos et al. 2019), disaster management and urban planning (Deliry, Avdan, and Avdan 2021; Kaku 2019; Manfré et al. 2012).

5.3. Reproducibility and accessibility

To assist the dissemination of the derived products to water managers and encourage the adoption of the introduced aquatic vegetation monitoring tools, a freely accessible and easy-to-use Google earth engine-based application is provided. This will enable the provision of timely, accurate and reliable information for mangers to remotely track an infestation and its response to management interventions. Since the introduction of the tool (Singh et al. 2020), there have been three (developer - myself) updates that fixed minor bugs and improved efficiency with, for example, the recent dynamic world landcover dataset (Brown et al. 2022). Future work should include the separation of macrophytes from algae and move towards an early warning system that will provide alerts to the early onset of algal or aquatic vegetation occurrence.

Open source and open access software, data, and its derived products made this work possible and should be encouraged because it accelerates research progress by removing barriers. In line with this, the code for this thesis has been open sourced on GitHub.

- https://github.com/Geethen/PhD--EO_and_ML_for_water_hyacinth_management

Despite the code being publicly accessible, it may still be difficult for some interested end-users, that are not schooled programmers to setup a python environment with the necessary packages. This need for a python environment can be circumvented by executing the provided notebooks in the freely available Google Collaboratory[1]. Moreover, to increase the accessibility of the developed tools to a wider user group many of the tools have been packaged into applications with a Graphical User Interfaces (GUI). These tools include:

1) A GEE application to monitor the extent of water, aquatic vegetation, and Normalised Difference Vegetation Index (NDVI) (a proxy for aquatic vegetation health) for any waterbody globally (accessible here).
2) A web application to extract more than a hundred environmental features for any South African waterbody from GEE (here).
3) A GEE script to interactively explore the 2013 South African water hyacinth distribution (here).
4) A notebook to extract water and aquatic vegetation proportions from Landsat-8, Sentinel-1 and Sentinel-2 for any waterbody (globally) as a time series (accessible here).
5) A notebook to identify the local drivers influencing the occurrence of water hyacinth on any South African water system (accessible here).

The links to the code and tools have been embedded into a website (available at https://geethen.github.io/). In addition, future updates to these tools will be reported on the website. The site will also allow for user feedback and suggestions. While these initiatives are helpful to increase the adoption of similar applied ecology research, this points to the benefit to be gained by the next generation of biologists that invest in programming and embrace computational methods. Moreover, it points to the benefit to be gained if remote sensing researchers are equipped with strong programming skills to ensure their work is accessible to a wider audience.

5.4. Looking ahead

ML, RS and GEE are a powerful combination, and it has been shown to be useful in generating information that can improve the management of water hyacinth at large spatio-temporal scales. With the rapid advancement in ML and future satellite missions that will provide frequent and systematic acquisition of free EO data, there is the ability to improve the accuracy and efficiency of the three applications shown in this study and to develop entirely new systems (for example, early warning systems and early detection systems). This suggests the need for the continuous development and maintenance of any EO-derived products. Perhaps, as important as developing and maintaining these systems, is the deployment of these systems in a manner that is accessible to concerned stakeholders in a way that can bring about change and ultimately combat invasive plants. While the availability of GEE as a freely available cloud computing platform has allowed RS practitioners to perform analysis at greater spatial-temporal scales, there are ongoing efforts to develop computer vision methods that support analyses at similar scales. To support these scalability efforts and bridge the knowledge-doing gap[1] between different researchers, researchers should consider the following points that were inspired by my experiences during this thesis which may assist in overcoming some of the barriers to the interdisciplinary research required to benefit the management of water hyacinth.

5.4.1. Suggestions: Water managers

Water managers, in this case, refers to individuals who are responsible for managing IAAP invasions and who may be involved in collecting in-situ field data on the occurrence and status of IAAP invasions.

- Water managers and invasion biologists are urged to support and encourage the adoption of recent technological advances that incorporate machine learning and EO in invasion ecology.
- During field surveys or biocontrol release events, information about the factors that influence the effectiveness of biocontrol should be measured and documented. These factors may include the presence and the density of pre-existing biocontrol agents at the time of release, the date, duration, and extent of herbicidal spray events and/or mechanical removal. If this information is to be useful for calibrating and validating ML models, they will need to be collected in a systematic and consistent manner. For example, the duration between satellite image capture and in-situ sampling must be minimised.
- Identifying a list of satellite-derived variables that will be useful for managing invasive plants will prove invaluable for aligning and prioritising remote sensing and machine learning research related to invasive species. For example, the list of [278,279] essential variables for SDGs and environmental monitoring [270,271]. Based on this thesis, suggested variables that require more research include species-level distribution information, large extent risk-of-invasion maps, waterbody nutrient status that are consistently derived across large spatial extents and plant density.
- In each of the three main applications of this thesis, emphasis was largely placed on the role of EO data and ML to assist water hyacinth management, however, the role of field experience and in-situ data must also be emphasised as it is critical for each of the three applications to be incorporated into operational systems. Therefore, water managers should look to integrate EO-derived insights into their management efforts, and foster collaborations between remote sensing and ML experts.
- There are unique management applications yet to be developed. This can be achieved by supporting and encouraging the cultivation of strong technical skills in ecology students. Moreover, research projects requiring interdisciplinary skills and

knowledge should be supervised by a multi-disciplinary team of supervisors with a background in invasion biology, remote sensing, and computer science. Moreover, research and open ecological problems should be presented at ML workshops to bring about awareness and encourage collaborations between disciplines.

5.4.2. Suggestions: Remote sensing practitioners

Remote sensing practitioners, in many scenarios, are an intermediary between ecologists and the application of machine learning to address ecological questions. Therefore, they typically require an understanding of the role of remote sensing to address an ecological problem and the correct protocol to implement and evaluate ML algorithms.

- The field of machine learning is rapidly advancing; however, the adoption of these advances by remote sensing practitioners have been slow. More practitioners need to move towards monitoring (for example, global forest watch for rainforest destruction) (as opposed to mapping) and need to ensure the methods they employ are suitable across large geographical areas and across time (i.e., spatio-temporally extensible). This will increase the utility of EO-based solutions for environmental management purposes.
- Programming is an essential skill for remote sensing practitioners that intend to do remote sensing at large scales. For this GEE is the ideal platform to become familiar with because it allows users to easily scale their analysis and learn programming at the same time. These skills provide the scaffolding to advance the application of remote sensing for invasion science, including other species and terrestrial systems.

5.4.3. Suggestions: Machine learning researchers

Machine learning researchers with a computer science background tend to focus on algorithm development, the curation of benchmark datasets for ML algorithm comparison and/or determining ML protocols and best practices to be used during implementation by practitioners.

- Free, low-medium (>10 m) spatial resolution satellite data should be prioritised over commercial high spatial resolution data (<10 m) during ML research. Whilst high spatial resolution satellite datasets are more easily interpreted than coarse spatial resolution data, coarse spatial resolution datasets are more likely to be used for operational monitoring of large areas. The accelerated transition to free data will directly benefit downstream ecological applications (for example, (Brown et al. 2022; Isikdogan, Bovik, and Passalacqua 2019)) because the developed ML methods will be able to optimally leverage the unique characteristics of the medium resolution data i.e., the higher spectral characteristics, systematic and regular image capture. Moreover, there are far more educational resources and research studies for getting started with the analysis of medium resolution satellite data in comparison to commercial satellite data.

- GEE has a python Application Programming Interface (API) that can easily be integrated with already available (primarily) python-based ML methods. The actively developed geemap python package provides useful functionality that can simplify the exploration of GEE datasets and the development of python-based workflows and, in this way, assist in the adoption of the more flexible python API (compared to the GEE JavaScript API).
- Whilst the natural environment may be a complex system from an applied ML perspective, aquatic systems are likely easier to trial novel ML applications than terrestrial systems owing to their smaller coverage, spectral contrast, and easier boundary delineation. This, together with the importance of freshwater resources and the pressures on freshwater resources suggest these systems should be prioritised for ML-EO research.

5.5. Conclusion

Through the three applications (Chapter 2 - 4), the capacity of machine learning and EO data to benefit IAAP management has been shown. The three applications provide the necessary information to answer four important questions for the biocontrol of water hyacinth, including;

(i) where is the water hyacinth in South Africa? (Chapter 2)

(ii) why is water hyacinth (not) occurring at a given location? (Chapter 3)

(iii) which water systems are at a high risk of a water hyacinth infestation? (Chapter 3) and

(iv) what are likely optimal biological control release periods for a water hyacinth infested site? (Chapter 4)

In addition, emphasis is placed on the role of in-situ data, interdisciplinary research, and the prioritisation of aquatic systems over terrestrial systems for the creation, validation, improvement, and adoption of ML-based tools that can assist in the management of IAAPs. Aside from advancing landscape ecology, EO and ML applications will continue to broaden its impact on the way invasive species are managed.

6. References

Acharya, Tri Dev, Dong Ha Lee, In Tae Yang, and Jae Kang Lee. 2016. “Identification of Water Bodies in a Landsat 8 OLI Image Using a J48 Decision Tree.” *Sensors* 16(7):1075.

Agami, M., and K. R. Reddy. 1990. “Competition for Space between *Eichhornia Crassipes* (Mart.) Solms and *Pistia Stratiotes* L. Cultured in Nutrient-Enriched Water.” *Aquatic Botany* 38(2–3):195–208.

Aguiar Jr, T. R., K. Rasera, L. M. Parron, A. G. Brito, and M. T. Ferreira. 2015. “Nutrient Removal Effectiveness by Riparian Buffer Zones in Rural Temperate Watersheds: The Impact of No-till Crops Practices.” *Agricultural Water Management* 149:74–80.

Agutu, Peter Odoyo, Moses Karoki Gachari, and Charles N. Mundia. 2018. “An Assessment of the Role of Water Hyacinth in the Water Level Changes of Lake Naivasha Using GIS and Remote Sensing.” *American Journal of Remote Sensing* 6(2):74–88.

Ahmed, Nurhussen, Clement Atzberger, and Worku Zewdie. 2020. “Integration of Remote Sensing and Bioclimatic Data for Prediction of Invasive Species Distribution in Data-Poor Regions: A Review on Challenges and Opportunities.” *Environmental Systems Research* 9(1):1–18.

Ajuonu, O., M. Byrne, M. Hill, P. Neuenschwander, and S. Korie. 2007. “Survival of the Mirid *Eccritotarsus Catarinensis* as Influenced by *Neochetina Eichhorniae* and *Neochetina Bruchi* Feeding Scars on Leaves of Water Hyacinth *Eichhornia Crassipes*.” *BioControl* 52(2):193–205.

Albertini, Cinzia, Andrea Gioia, Vito Iacobellis, and Salvatore Manfreda. 2022. “Detection of Surface Water and Floods with Multispectral Satellites.” *Remote Sensing* 14(23):6005.

Ali, Huma, N. Lata, Janak Ahi, and N. Ganesh. 2010. “Evaluation of Wound-Healing Activity of *Eichornia Crassipes*: A Novel Approach.” *Drug Invention Today* 2(3):212–14.

Ali, Moez. 2020. “PyCaret: An Open Source, Low-Code Machine Learning Library in Python.” *PyCaret Version* 2.

Allen, George H., and Tamlin M. Pavelsky. 2018. “Global Extent of Rivers and Streams.” *Science* 361(6402):585–88.

Almeida, Ana S., Ana M. M. Goncalves, Joana L. Pereira, and Fernando Goncalves. 2006. “The Impact of *Eichhornia Crassipes* on Green Algae and Cladocerans.” *Fresenius Environmental Bulletin* 15(12 A):1531–38.

Amatulli, Giuseppe, Daniel McInerney, Tushar Sethi, Peter Strobl, and Sami Domisch. 2020. “Geomorpho90m, Empirical Evaluation and Accuracy Assessment of Global High-Resolution Geomorphometric Layers.” *Scientific Data* 7(1):1–18.

Ari, Niyazi, and Makhamadsulton Ustazhanov. 2014. “Matplotlib in Python.” *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)* 1–6.

Arribas-Bel, Dani. 2021. “Contextily.”

Ashton, P. J., W. E. Scott, D. J. Steyn, and R. J. Wells. 1979. “Chemical Control Programme against the Water Hyacinth *Eichhornia Crassipes* (Mart.) Solms on Hartbeespoort Dam: Historical and Practical Aspects.” *South African Journal of Science* 75(3):303–6.

Ashutosh, Subhash, and P. S. Roy. 2021. “Three Decades of Nationwide Forest Cover Mapping Using Indian Remote Sensing Satellite Data: A Success Story of Monitoring Forests for Conservation in India.” *Journal of the Indian Society of Remote Sensing* 49(1):61–70.

Auchterlonie, Jessica, Christin-Leigh Eden, and Craig Sheridan. 2021. “The Phytoremediation Potential of Water Hyacinth: A Case Study from Hartbeespoort Dam, South Africa.” *South African Journal of Chemical Engineering* 37:31–36.

Ball, John E., Derek T. Anderson, and Chee Seng Chan. 2017. “Comprehensive Survey of Deep Learning in Remote Sensing: Theories, Tools, and Challenges for the Community.” *Journal of Applied Remote Sensing* 11(4):42609–42609.

Ballot, Andreas, Morten Sandvik, Thomas Rundberget, Christo J. Botha, and Christopher O. Miles. 2013. “Diversity of Cyanobacteria and Cyanotoxins in Hartbeespoort Dam, South Africa.” *Marine and Freshwater Research* 65(2):175–89.

Barbet-Massin, Morgane, Quentin Rome, Claire Villemant, and Franck Courchamp. 2018. “Can Species Distribution Models Really Predict the Expansion of Invasive Species?” *PloS One* 13(3):e0193085.

Barrett, Paul, John Hunter, J. Todd Miller, J. C. Hsu, and Perry Greenfield. 2005. “Matplotlib--A Portable Python Plotting Package.” *Astronomical Data Analysis Software and Systems XIV* 347:91.

Barrett, S. C. H., and I. W. Forno. 1982. “Style Morph Distribution in New World Populations of *Eichhornia Crassipes* (Mart.) Solms-Laubach (Water Hyacinth).” *Aquatic Botany* 13:299–306.

Bartodziej, WILLIAM, and G. Weymouth. 1995. “Waterbird Abundance and Activity on Water Hyacinth and *Egeria* in the St. Marks River, Florida.” *Journal of Aquatic Plant Management* 33:19–22.

Berger-Tal, Oded, Jonathan Nathan, Ehud Meron, and David Saltz. 2014. “The Exploration-Exploitation Dilemma: A Multidisciplinary Framework.” *PloS One* 9(4):e95693.

Bergstra, James, Brent Komer, Chris Eliasmith, Dan Yamins, and David D. Cox. 2015. “Hyperopt: A Python Library for Model Selection and Hyperparameter Optimization.” *Computational Science & Discovery* 8(1):14008.

Bick, Emily, Elvira S. de Lange, Cindy R. Kron, Lorena da Silva Soler, Jessie Liu, and Hoang Danh Nguyen. 2020. “Effects of Salinity and Nutrients on Water Hyacinth and Its Biological Control Agent, *Neochetina Bruchi*.” *Hydrobiologia* 847:3213–24.

Bock, Jane H. 1969. “Productivity of the Water Hyacinth *Eichhornia Crassipes* (Mart.) Solms.” *Ecology* 50(3):460–64.

Bouneffouf, Djallel, and Irina Rish. 2019. “A Survey on Practical Applications of Multi-Armed and Contextual Bandits.” *ArXiv Preprint ArXiv:1904.10040*.

Boyd, D. S., and F. M. Danson. 2005. “Satellite Remote Sensing of Forest Resources: Three Decades of Research Development.” *Progress in Physical Geography* 29(1):1–26.

Bradie, Johanna, and Brian Leung. 2017. “A Quantitative Synthesis of the Importance of Variables Used in MaxEnt Species Distribution Models.” *Journal of Biogeography* 44(6):1344–61.

Bradley, Bethany A. 2014. “Remote Detection of Invasive Plants: A Review of Spectral, Textural and Phenological Approaches.” *Biological Invasions* 16(7):1411–25.

Bradley, Bethany A., and John F. Mustard. 2005. “Identifying Land Cover Variability Distinct from Land Cover Change: Cheatgrass in the Great Basin.” *Remote Sensing of Environment* 94(2):204–13.

Braunisch, Veronika, Joy Coppes, Raphaël Arlettaz, Rudi Suchant, Hans Schmid, and Kurt Bollmann. 2013. “Selecting from Correlated Climate Variables: A Major Source of Uncertainty for Predicting Species Distributions under Climate Change.” *Ecography* 36(9):971–83.

Bredin, Ian, and Douglas Macfarlane. 2017. *Buffer Zone Guidelines for Wetlands, Rivers and Estuaries. Part 1: Technical Manual*.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45(1):5–32.

Brendonck, Luc, Joachim Maes, Wouter Rommens, Nzwirashe Dekeza, Tamuka Nhiwatiwa, Maxwell Barson, Veerle Callebaut, Crispen Phiri, Kelle Moreau, and Brian Gratwicke. 2003. “The Impact of Water Hyacinth (*Eichhornia Crassipes*) in a Eutrophic Subtropical Impoundment (Lake Chivero, Zimbabwe). II. Species Diversity.” *Archiv Fur Hydrobiologie* 158(3):389–405.

Briscoe Runquist, Ryan D., Thomas A. Lake, and David A. Moeller. 2021. “Improving Predictions of Range Expansion for Invasive Species Using Joint Species Distribution Models and Surrogate Co-occurring Species.” *Journal of Biogeography*.

Brown, Christopher F., Steven P. Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Brooks Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie J. Pasquarella, Robert Haertel, and Simon Ilyushchenko. 2022. “Dynamic World, Near Real-Time Global 10 m Land Use Land Cover Mapping.” *Scientific Data* 9(1):1–17.

Byrne, M., M. Hill, M. Robertson, A. King, N. Katembo, J. Wilson, R. Brudvig, J. Fisher, and A. Jadhav. 2010. “Integrated Management of Water Hyacinth in South Africa.” *WRC Report* (454/10).

Camps-Valls, Gustavo. 2009. “Machine Learning in Remote Sensing Data Processing.” Pp. 1–6 in *2009 IEEE international workshop on machine learning for signal processing*. Grenoble, France: IEEE.

Carpenter, Stephen R., Emily H. Stanley, and M. Jake vander Zanden. 2011. “State of the World’s Freshwater Ecosystems: Physical, Chemical, and Biological Changes.” *Annual Review of Environment and Resources* 36:75–99.

Carpentier, Alexandra, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. 2011. “Upper-Confidence-Bound Algorithms for Active Learning in Multi-

Armed Bandits." Pp. 189–203 in *International Conference on Algorithmic Learning Theory*. Springer.

Center, Ted D., and Neal R. Spencer. 1981. "The Phenology and Growth of Water Hyacinth (*Eichhornia Crassipes* (Mart.) Solms) in a Eutrophic North-Central Florida Lake." *Aquatic Botany* 10:1–32.

Cerqueira, Vitor, Luis Torgo, and Igor Mozetič. 2020. "Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods." *Machine Learning* 109(11):1997–2028.

Cha, YoonKyung, Jihoon Shin, ByeongGeon Go, Dae-Seong Lee, YoungWoo Kim, TaeHo Kim, and Young-Seuk Park. 2021. "An Interpretable Machine Learning Method for Supporting Ecosystem Management: Application to Species Distribution Models of Freshwater Macroinvertebrates." *Journal of Environmental Management* 291:112719.

Chakraborty, Debaditya, Arafat Alam, Saptarshi Chaudhuri, Hakan Başağaoğlu, Tulio Sulbaran, and Sandeep Langar. 2021. "Scenario-Based Prediction of Climate Change Impacts on Building Cooling Energy Consumption with Explainable Artificial Intelligence." *Applied Energy* 291:116807.

Chandrashekar, Girish, and Ferat Sahin. 2014. "A Survey on Feature Selection Methods." *Computers & Electrical Engineering* 40(1):16–28.

Chantiratikul, Piyanete, P. Meechai, and WJRJoBS Nakbanpotec. 2009. "Antioxidant Activities and Phenolic Contents of Extracts from *Salvinia Molesta* and *Eichornia Crassipes*." *Research Journal of Biological Sciences* 4(10):1113–17.

Chapman, Daniel, Oliver L. Pescott, Helen E. Roy, and Rob Tanner. 2019. "Improving Species Distribution Models for Invasive Non-native Species with Biologically Informed Pseudo-absence Selection." *Journal of Biogeography* 46(5):1029–40.

Cheruiyot, E., M. Menenti, B. Gerte, and R. Koenders. 2013. "Accuracy and Precision of Algorithms to Determine the Extent of Aquatic Plants: Empirical Sealing of Spectral Indices vs. Spectral Unmixing." P. 85 in *ESA Special Publication*. Vol. 722.

Chicco, Davide, and Giuseppe Jurman. 2020. "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation." *BMC Genomics* 21(1):1–13.

Cilliers, C. J. 1991. "Biological Control of Water Hyacinth, *Eichhornia Crassipes* (Pontederiaceae), in South Africa." *Agriculture, Ecosystems & Environment* 37(1–3):207–17.

Cilliers, C. J., P. L. Campbell, DSSN Naude, and S. Neser. 1996. "An Integrated Water Hyacinth Control Programme on the Vaal River, in a Cool, High Altitude Area in South Africa." *Strategies for Water Hyacinth Control* 87–103.

Coetzee, J. A., A. Bownes, G. D. Martin, B. E. Miller, R. Smith, P. S. R. Weyl, and M. P. Hill. 2021. "A Review of the Biocontrol Programmes against Aquatic Weeds in South Africa." *African Entomology* 29(3):935–64.

Coetzee, J. A., M. P. Hill, M. J. Byrne, and A. Bownes. 2011. "A Review of the Biological Control Programmes on *Eichhornia Crassipes* (C. Mart.) Solms (Pontederiaceae), *Salvinia*

Molesta DS Mitch.(Salviniaceae), *Pistia Stratiotes* L.(Araceae), *Myriophyllum Aquaticum* (Vell.) Verdc.(Haloragaceae) and *Azolla Filiculoides* L." *African Entomology* 19(2):451–68.

Coetzee, J., and centre for biological control) Mostert, E; (Rhodes University. 2019. "GPS Locality Data for Invasive Aquatic Alien Plants (IAAPs)."

Coetzee, Julie A., Marcus J. Byrne, and Martin P. Hill. 2007. "Predicting the Distribution of *Eccritotarsus Catarinensis*, a Natural Enemy Released on Water Hyacinth in South Africa." *Entomologia Experimentalis et Applicata* 125(3):237–47.

Coetzee, Julie A., Ted D. Center, Marcus J. Byrne, and Martin P. Hill. 2005. "Impact of the Biocontrol Agent *Eccritotarsus Catarinensis*, a Sap-Feeding Mirid, on the Competitive Performance of Water Hyacinth, *Eichhornia Crassipes*." *Biological Control* 32(1):90–96.

Coetzee, Julie A., Martin P. Hill, Trinidad Ruiz-Téllez, Uwe Starfinger, and Sarah Brunel. 2017. "Monographs on Invasive Plants in Europe N° 2: *Eichhornia Crassipes* (Mart.) Solms." *Botany Letters* 164(4):303–26.

Coetzee, Julie A., Benjamin E. Miller, David Kinsler, Keneilwe Sebola, and Martin P. Hill. 2022. "It's a Numbers Game: Inundative Biological Control of Water Hyacinth (*Pontederia Crassipes*), Using *Megamelus Scutellaris* (Hemiptera: Delphacidae) Yields Success at a High Elevation, Hypertrophic Reservoir in South Africa." *Biocontrol Science and Technology* 1–10.

Cohen, Warren B., and Samuel N. Goward. 2004. "Landsat's Role in Ecological Applications of Remote Sensing." *Bioscience* 54(6):535–45.

Cole, Lorna J., Jenni Stockan, and Rachel Helliwell. 2020. "Managing Riparian Buffer Strips to Optimise Ecosystem Services: A Review." *Agriculture, Ecosystems & Environment* 296:106891.

Cossu, Raffaello, Ketil Haarstad, M. Cristina Lavagnolo, and Paolo Littarru. 2001. "Removal of Municipal Solid Waste COD and NH4–N by Phyto-Reduction: A Laboratory–Scale Comparison of Terrestrial and Aquatic Species at Different Organic Loads." *Ecological Engineering* 16(4):459–70.

Cracknell, Arthur P., and Costas A. Varotsos. 2009. "The Contribution of Remote Sensing to the Implementation of the Montreal Protocol and the Monitoring of Its Success." *International Journal of Remote Sensing* 30(15–16):3853–73.

Crowder, Larry B., and William E. Cooper. 1982. "Habitat Structural Complexity and the Interaction between Bluegills and Their Prey." *Ecology* 63(6):1802–13.

Deliry, Sayed Ishaq, Zehra Yiğit Avdan, and Uğur Avdan. 2021. "Extracting Urban Impervious Surfaces from Sentinel-2 and Landsat-8 Satellite Data for Urban Planning and Environmental Management." *Environmental Science and Pollution Research* 28(6):6572–86.

Dersseh, Minychl G., Assefa M. Melesse, Seifu A. Tilahun, Mengiste Abate, and Dessalegn C. Dagnew. 2019. "Water Hyacinth: Review of Its Impacts on Hydrology and Ecosystem Services—Lessons for Management of Lake Tana." *Extreme Hydrology and Climate Variability* 237–51.

Ding, Yaxin, Xiaomei Yang, Hailiang Jin, Zhihua Wang, Yueming Liu, Bin Liu, Junyao Zhang, Xiaoliang Liu, Ku Gao, and Dan Meng. 2021. “Monitoring Coastline Changes of the Malay Islands Based on Google Earth Engine and Dense Time-Series Remote Sensing Images.” *Remote Sensing* 13(19):3842.

Domisch, Sami, Martin Friedrichs, Thomas Hein, Florian Borgwardt, Annett Wetzig, Sonja C. Jähnig, and Simone D. Langhans. 2019. “Spatially Explicit Species Distribution Models: A Missed Opportunity in Conservation Planning?” *Diversity and Distributions* 25(5):758–69.

Donchyts, Gennadii, Fedor Baart, Hessel Winsemius, Noel Gorelick, Jaap Kwadijk, and Nick van de Giesen. 2016. “Earth’s Surface Water Change over the Past 30 Years.” *Nature Climate Change* 6(9):810–13.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Marquéz, Bernd Gruber, Bruno Lafourcade, and Pedro J. Leitão. 2013. “Collinearity: A Review of Methods to Deal with It and a Simulation Study Evaluating Their Performance.” *Ecography* 36(1):27–46.

Dottori, Francesco, Lorenzo Alfieri, Peter Salamon, Alessandra Bianchi, Luc Feyen, and Valerio Lorini. 2016. “Flood Hazard Map for Europe, 200-Year Return Period.” *European Commission, Joint Research Centre (JRC)*.

Dube, Timothy, Onisimo Mutanga, Mbulisi Sibanda, Victor Bangamwabo, and Cletah Shoko. 2017a. “Evaluating the Performance of the Newly-Launched Landsat 8 Sensor in Detecting and Mapping the Spatial Configuration of Water Hyacinth (*Eichhornia Crassipes*) in Inland Lakes, Zimbabwe.” *Physics and Chemistry of the Earth, Parts A/b/c* 100:101–11.

Dube, Timothy, Onisimo Mutanga, Mbulisi Sibanda, Victor Bangamwabo, and Cletah Shoko. 2017b. “Testing the Detection and Discrimination Potential of the New Landsat 8 Satellite Data on the Challenging Water Hyacinth (*Eichhornia Crassipes*) in Freshwater Ecosystems.” *Applied Geography* 84:11–22.

Durand, Audrey, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. 2018. “Contextual Bandits for Adapting Treatment in a Mouse Model of de Novo Carcinogenesis.” Pp. 67–82 in *Machine learning for healthcare conference*. PMLR.

Dvořák, Petr, Jana Müllerová, Tomáš Bartaloš, and Josef Brůna. 2015. “Unmanned Aerial Vehicles for Alien Plant Species Detection and Monitoring.” *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* 40:83–90.

Edwards, D., and C. J. Musil. 1975. “*Eichhornia Crassipes* in South Africa-a General Review.” *Journal of the Limnological Society of Southern Africa* 1(1):23–27.

Effrosynidis, Dimitrios, and Avi Arampatzis. 2021. “An Evaluation of Feature Selection Methods for Environmental Data.” *Ecological Informatics* 61:101224.

Elith, Jane. 2017. “Predicting Distributions of Invasive Species.” Pp. 94–129 in *Invasive species: Risk assessment and management*. Vol. 10. Cambridge University Press Cambridge.

Elith, Jane, and John R. Leathwick. 2009. “Species Distribution Models: Ecological Explanation and Prediction across Space and Time.” *Annual Review of Ecology, Evolution and Systematics* 40(1):677–97.

Essl, Franz, Wayne Dawson, Holger Kreft, Jan Pergl, Petr Pyšek, Mark van Kleunen, Patrick Weigelt, Thomas Mang, Stefan Dullinger, and Bernd Lenzner. 2019. “Drivers of the Relative Richness of Naturalized and Invasive Plant Species on Earth.” *AoB Plants* 11(5):plz051.

Estes, L. D., G. S. Okin, A. G. Mwangi, and H. H. Shugart. 2008. “Habitat Selection by a Rare Forest Antelope: A Multi-Scale Approach Combining Field Data and Imagery from Three Sensors.” *Remote Sensing of Environment* 112(5):2033–50.

Everitt, J. H., C. Yang, K. R. Summy, C. S. Owens, L. M. Glomski, and R. M. Smart. 2011. “Using in Situ Hyperspectral Reflectance Data to Distinguish Nine Aquatic Plant Species.” *Geocarto International* 26(6):459–73.

Everitt, James H., K. Rodney Summy, L. M. Glomski, C. S. Owens, and Chenghai Yang. 2009. “Spectral Reflectance and Digital Image Relations among Five Aquatic Weeds.” *Subtropical Plant Sciences* 61:15–23.

Firehun, Y., P. C. Struik, E. A. Lantinga, and T. Taye. 2013. “Joint Use of Insects and Fungal Pathogens in the Management of Water Hyacinth (*Eichhornia Crassipes*): Perspectives for Ethiopia.” *J Aquat Plant Manage* 51:109–21.

Forman, Richard T. T. 1995. “Some General Principles of Landscape and Regional Ecology.” *Landscape Ecology* 10(3):133–42.

Frake, April N., Brad G. Peter, Edward D. Walker, and Joseph P. Messina. 2020. “Leveraging Big Data for Public Health: Mapping Malaria Vector Suitability in Malawi with Google Earth Engine.” *Plos One* 15(8):e0235697.

Francis, Robert A., and Michael A. Chadwick. 2012. “Invasive Alien Species in Freshwater Ecosystems: A Brief Overview.” Pp. 23–41 in *A Handbook of Global Freshwater Invasive Species*. Routledge.

Frouin, Robert J., Bryan A. Franz, Amir Ibrahim, Kirk Knobelspiesse, Ziauddin Ahmad, Brian Cairns, Jacek Chowdhary, Heidi M. Dierssen, Jing Tan, and Oleg Dubovik. 2019. “Atmospheric Correction of Satellite Ocean-Color Imagery during the PACE Era.” *Frontiers in Earth Science* 7:145.

Gaertner, Mirijam, Brendon M. H. Larson, Ulrike M. Irlich, Patricia M. Holmes, Louise Stafford, Brian W. van Wilgen, and David M. Richardson. 2016. “Managing Invasive Species in Cities: A Framework from Cape Town, South Africa.” *Landscape and Urban Planning* 151:1–9.

García-Álvarez, David, and Sabina Florina Nanu. 2022. “Land Use Cover Datasets: A Review.” Pp. 47–66 in *Land Use Cover Datasets and Validation Tools*, edited by D. García-Álvarez, M. T. Camacho Olmedo, M. Paegelow, and J. F. Mas. Springer.

GEOTERRAIMAGE. 2020. “SANLC 2018 Accuracy Assessment Points.” Retrieved March 28, 2020 (https://egis.environment.gov.za/data_egis/data_download/current).

Gettys, Lyn A., William T. Haller, and D. G. Petty. 2014. *Biology and Control of Aquatic Plants*.

Giraldo, E., and A. Garzon. 2002. “The Potential for Water Hyacinth to Improve the Quality of Bogota River Water in the Muña Reservoir: Comparison with the Performance of Waste Stabilization Ponds.” *Water Science and Technology* 45(1):103–10.

Gitelson, Anatoly A., Yoram J. Kaufman, Robert Stark, and Don Rundquist. 2002. “Novel Algorithms for Remote Estimation of Vegetation Fraction.” *Remote Sensing of Environment* 80(1):76–87.

Gomes, Lucas Carvalho, Raiza Moniz Faria, Eliana de Souza, Gustavo Vieira Veloso, Carlos Ernesto G. R. Schaefer, and Elpídio Inácio Fernandes Filho. 2019. “Modelling and Mapping Soil Organic Carbon Stocks in Brazil.” *Geoderma* 340:337–50.

Gopal, B. 1987. “Water Hyacinth Elsevier.” *Amsterdam, The Netherlands* 471.

Gorelick, Noel, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. 2017. “Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone.” *Remote Sensing of Environment* 202:18–27.

Grill, Günther, B. Lehner, Michele Thieme, B. Geenen, D. Tickner, F. Antonelli, S. Babu, Pasquale Borrelli, L. Cheng, and H. Crochetiere. 2019. “Mapping the World’s Free-Flowing Rivers.” *Nature* 569(7755):215–21.

Groom, Geoff, C. A. Mücher, Margareta Ihse, and Thomas Wrbka. 2006. “Remote Sensing in Landscape Ecology: Experiences and Perspectives in a European Context.” *Landscape Ecology* 21(3):391–408.

Güereña, David, Henry Neufeldt, Julia Berazneva, and Sam Duby. 2015. “Water Hyacinth Control in Lake Victoria: Transforming an Ecological Catastrophe into Economic, Social, and Environmental Benefits.” *Sustainable Production and Consumption* 3:59–69.

Guillarmod, A. Jacot. 1979. “Water Weeds in Southern Africa.” *Aquatic Botany* 6:377–91.

Guillera-Arroita, Gurutzeta. 2017. “Modelling of Species Distributions, Range Dynamics and Communities under Imperfect Detection: Advances, Challenges and Opportunities.” *Ecography* 40(2):281–95.

Guyon, Isabelle, and André Elisseeff. 2003. “An Introduction to Variable and Feature Selection.” *Journal of Machine Learning Research* 3(Mar):1157–82.

Harding, William. 2015. “Living with Eutrophication in South Africa: A Review of Realities and Challenges, Transactions of the Royal Society of South Africa 70 (2) 2015: Pp. 155-171.” *Transactions of the Royal Society of South Africa* 70(3):299–303.

Harley, K. L. S., M. H. Julien, and A. D. Wright. 1996. “Water Hyacinth: A Tropical World Wide Problem and Methods for Its Control.” in *proceedings of the first meeting of the international water hyacinth consortium*. World Bank.

Harley, Kenneth Leigh Scott, and Iris Wendy Forno. 1992. “Biological Control of Weeds: A Handbook for Practitioners and Students.” P. 74 in. Melbourne: Inkata Press.

Henderson, L. 2001. “Alien Weeds and Invasive Plants. A Complete Guide to Declared Weeds and Invaders in South Africa.” *Plant Protection Research Institute Handbook No. 12 (South Africa)*.

Henderson, Lesley, and Carina J. Cilliers. 2002. *Invasive Aquatic Plants: A Guide to the Identification of the Most Important and Potentially Dangerous Invasive Aquatic and Wetland Plants in South Africa; Also Featuring the Biological Control of the Five Worst Aquatic Weeds*. ARC-Plant Protection Research Inst.

Hengl, Tomislav, Matthew A. E. Miller, Josip Križan, Keith D. Shepherd, Andrew Sila, Milan Kilibarda, Ognjen Antonijević, Luka Glušica, Achim Dobermann, and Stephan M. Haefele. 2021. "African Soil Properties and Nutrients Mapped at 30 m Spatial Resolution Using Two-Scale Ensemble Machine Learning." *Scientific Reports* 11(1):1–18.

Hill, M. P. 2003. "The Impact and Control of Alien Aquatic Vegetation in South African Aquatic Ecosystems." *African Journal of Aquatic Science* 28(1):19–24.

Hill, M. P., and C. J. Cilliers. 1999. "A Review of the Arthropod Natural Enemies, and Factors That Influence Their Efficacy, in the Biological Control of Water Hyacinth, *Eichhornia Crassipes* (Mart.) Solms-Laubach (Pontederiaceae), in South Africa." *In T. Olckers & M.P. Hill (Eds.), Biological Control of Weeds in South Africa (1990–1998), African Entomology Memoir* 1:103–12.

Hill, M. P., D. Conlong, C. Zachariades, J. A. Coetzee, I. D. Paterson, B. E. Miller, L. Foxcroft, and L. van der Westhuizen. 2021. "The Role of Mass-Rearing in Weed Biological Control Projects in South Africa." *African Entomology* 29(3):1030–44.

Hill, M. P., and T. Olckers. 2000. "Biological Control Initiatives against Water Hyacinth in South Africa: Constraining Factors, Success and New Courses of Action." Pp. 33–38 in *ACIAR proceedings*. ACIAR; 1998.

Hill, Martin P., and Julie Coetzee. 2017. "The Biological Control of Aquatic Weeds in South Africa: Current Status and Future Challenges." *Bothalia-African Biodiversity & Conservation* 47(2):1–12.

Hill, Martin P., Julie A. Coetzee, Grant D. Martin, Rosali Smith, and Emily F. Strange. 2020. "Invasive Alien Aquatic Plants in South African Freshwater Ecosystems." Pp. 97–114 in *Biological Invasions in South Africa*. Springer.

Ho, Y. B., and Wai-kin Wong. 1994. "Growth and Macronutrient Removal of Water Hyacinth in a Small Secondary Sewage Treatment Plant." *Resources, Conservation and Recycling* 11(1–4):161–78.

Huang, Cho-ying, and Gregory P. Asner. 2009. "Applications of Remote Sensing to Alien Invasive Plant Studies." *Sensors* 9(6):4869–89.

Hussner, Andreas, Iris Stiers, MJJM Verhofstad, E. S. Bakker, B. M. C. Grutters, Jacques Haury, JLCH van Valkenburg, G. Brundu, J. Newman, and J. S. Clayton. 2017. "Management and Control Methods of Invasive Alien Freshwater Aquatic Plants: A Review." *Aquatic Botany* 136:112–37.

Ibtehaz, Nabil, and M. Sohel Rahman. 2020. "MultiResUNet: Rethinking the U-Net Architecture for Multimodal Biomedical Image Segmentation." *Neural Networks* 121:74–87.

Ingole, Nandkishor A., A. S. Nain, Prem Kumar, and Ramesh Chalal. 2018. "Monitoring and Mapping Invasive Aquatic Weed *Salvinia Molesta* Using Multispectral Remote Sensing

Technique in Tumaria Wetland of Uttarakhand, India." *Journal of the Indian Society of Remote Sensing* 46(5):863–71.

Isikdogan, Leo F., Alan Bovik, and Paola Passalacqua. 2019. "Seeing through the Clouds with Deepwatermap." *IEEE Geoscience and Remote Sensing Letters* 17(10):1662–66.

Jayanthi, P., and P. Lalitha. 2013. "Antimicrobial Activity of Solvent Extracts of *Eichhornia Crassipes* (Mart.) Solms." *Der Pharma Chemica* 5(3):135–40.

Jiang, Fei, Heather E. Preisendanz, Tamie L. Veith, Raj Cibin, and Patrick J. Drohan. 2020. "Riparian Buffer Effectiveness as a Function of Buffer Design and Input Loads." *Journal of Environmental Quality* 49:1599–1611.

Johnson, David Lawrence, and Roy A. Stein. 1979. *Response of Fish to Habitat Structure in Standing Water*. (Indianapolis, IN)(USA) AFS, North Central Div.: American Fisheries Society.

Johnstone, Iain M., and D. Michael Titterington. 2009. "Statistical Challenges of High-Dimensional Data." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367(1906):4237–53.

Jones, R. W., and C. J. Cilliers. 2001. "Integrated Control of Water Hyacinth on the Nseleni/Mposa Rivers and Lake Nsezi in KwaZulu-Natal, South Africa." *Biological and Integrated Control of Water Hyacinth, Eichhornia Crassipes. Australian Centre for International Agricultural Research, Canberra* 123–29.

Jordahl, K. 2014. "GeoPandas: Python Tools for Geographic Data." *URL: Https://Github. Com/Geopandas/Geopandas*.

Joshi, Anup R., Eric Dinerstein, Eric Wikramanayake, Michael L. Anderson, David Olson, Benjamin S. Jones, John Seidensticker, Susan Lumpkin, Matthew C. Hansen, and Nigel C. Sizer. 2016. "Tracking Changes and Preventing Loss in Critical Tiger Habitat." *Science Advances* 2(4):e1501675.

Joshi, Chudamani, Jan de Leeuw, and Iris C. van Duren. 2004. "Remote Sensing and GIS Applications for Mapping and Spatial Modelling of Invasive Species." P. B7 in *Proceedings of ISPRS*. Vol. 35.

Julien, M. H. 2000. "Biological Control of Water Hyacinth with Arthropods: A Review to 2000." Pp. 8–20 in *Aciar Proceedings*. Citeseer.

Julien, M.H. and Orapa, W. 1999. "Successful Biological Control of Water Hyacinth (*Eichhornia Crassipes*) in Papua New Guinea by the Weevils *Neochetina Bruchi* and *Neochetina Eichhorniae* (Coleoptera: Curculionidae)." P. 1027 in *International Symposium on Biological Control of Weeds*. Bozeman, MT, USA: Montana State University.

Kaku, Kazuya. 2019. "Satellite Remote Sensing for Disaster Management Support: A Holistic and Staged Approach Based on Case Studies in Sentinel Asia." *International Journal of Disaster Risk Reduction* 33:417–32.

Kariyawasam, Champika S., Lalit Kumar, and Sujith S. Ratnayake. 2021. "Potential Distribution of Aquatic Invasive Alien Plants, *Eichhornia Crassipes* and *Salvinia Molesta* under Climate Change in Sri Lanka." *Wetlands Ecology and Management* 1–15.

Kass, Jamie M., Robert Muscarella, Peter J. Galante, Corentin L. Bohl, Gonzalo E. Pinilla-Buitrago, Robert A. Boria, Mariano Soley-Guardia, and Robert P. Anderson. 2021. “ENMeval 2.0: Redesigned for Customizable and Reproducible Modeling of Species’ Niches and Distributions.” *Methods in Ecology and Evolution* 12(9):1602–8.

Kateregga, Eseza, and Thomas Sterner. 2009. “Lake Victoria Fish Stocks and the Effects of Water Hyacinth.” *The Journal of Environment & Development* 18(1):62–78.

Kaur, Harsurinder, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2019. “A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions.” *ACM Computing Surveys (CSUR)* 52(4):1–36.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. “Lightgbm: A Highly Efficient Gradient Boosting Decision Tree.” *Advances in Neural Information Processing Systems* 30.

Keasar, Tamar, Ella Rashkovich, Dan Cohen, and Avi Shmida. 2002. “Bees in Two-Armed Bandit Situations: Foraging Choices and Possible Decision Mechanisms.” *Behavioral Ecology* 13(6):757–65.

Kennedy, Christina M., James R. Oakleaf, David M. Theobald, Sharon Baruch-Mordo, and Joseph Kiesecker. 2019. “Managing the Middle: A Shift in Conservation Priorities Based on the Global Human Modification Gradient.” *Global Change Biology* 25(3):811–26.

Kilroy, Cathy, Ton H. Snelder, Oliver Floerl, Christina C. Vieglais, and Katie L. Dey. 2008. “A Rapid Technique for Assessing the Suitability of Areas for Invasive Species Applied to New Zealand’s Rivers.” *Diversity and Distributions* 14(2):262–72.

King, Anthony Michael. 2011. “The Effect of Temperature on Biological Control of Water Hyacinth, *Eichhornia Crassipes* (Pontederiaceae) in South Africa.” Master’s Thesis, University of the Witwatersrand, Johannesburg.

Kriticos, Darren J., and Sarah Brunel. 2016. “Assessing and Managing the Current and Future Pest Risk from Water Hyacinth, (*Eichhornia Crassipes*), an Invasive Aquatic Plant Threatening the Environment and Water Security.” *PloS One* 11(8):e0120054.

Kumar, Sunil, Lisa G. Neven, and Wee L. Yee. 2014. “Assessing the Potential for Establishment of Western Cherry Fruit Fly Using Ecological Niche Modeling.” *Journal of Economic Entomology* 107(3):1032–44.

Kumschick, Sabrina, John R. U. Wilson, and Llewellyn C. Foxcroft. 2020. “A Framework to Support Alien Species Regulation: The Risk Analysis for Alien Taxa (RAAT).” *NeoBiota* 62:213–39.

Kurugundla, C. N., B. Mathangwane, S. Sakuringwa, and G. Katorah. 2016. “Alien Invasive Aquatic Plant Species in Botswana: Historical Perspective and Management.” *The Open Plant Science Journal* 9(1):1–40.

Kwon, Ohbyung, and Jae Mun Sim. 2013. “Effects of Data Set Features on the Performances of Classification Algorithms.” *Expert Systems with Applications* 40(5):1847–57.

Langa, Silvia D. F., Martin P. Hill, and Stephen G. Compton. 2020. “Agents sans Frontiers: Cross-Border Aquatic Weed Biological Control in the Rivers of Southern Mozambique.” *African Journal of Aquatic Science* 45(3):329–35.

Larson, Eric R., Brittney M. Graham, Rafael Achury, Jaime J. Coon, Melissa K. Daniels, Daniel K. Gambrell, Kacie L. Jonasen, Gregory D. King, Nicholas LaRacuente, and Tolulope I. N. Perrin-Stowe. 2020. "From EDNA to Citizen Science: Emerging Tools for the Early Detection of Invasive Species." *Frontiers in Ecology and the Environment* 18(4):194–202.

Lary, David J., Amir H. Alavi, Amir H. Gandomi, and Annette L. Walker. 2016. "Machine Learning in Geosciences and Remote Sensing." *Geoscience Frontiers* 7(1):3–10.

Lata, N., Ali Huma, Das Sumana, and Dubey Veenapani. 2010. "Antioxidants of *Eichhornia Crassipes*: The World's Worst Aquatic Plant." *Journal of Pharmacy Research* 3(9):2105–6.

Lawrence, Alexandra, Samuel Hoffmann, and Carl Beierkuhnlein. 2020. "Topographic Diversity as an Indicator for Resilience of Terrestrial Protected Areas against Climate Change." *Global Ecology and Conservation* e01445.

Lewis, Colin David. 1982. *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*. London, UK: Butterworth-Heinemann.

Li, Jian, and David P. Roy. 2017. "A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring." *Remote Sensing* 9(9):902.

Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. "A Survey on Deep Learning in Medical Image Analysis." *Medical Image Analysis* 42:60–88.

Little, E. C. S. 1965. "The World Wide Distribution of the Water Hyacinth." *Hyacinth Control J* 4:30–32.

Liu, Brian, and Madeleine Udell. 2020. *Impact of Accuracy on Model Interpretations*.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." Pp. 3431–40 in *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Lu, Jianbo, Jianguo Wu, Zhihui Fu, and Lei Zhu. 2007. "Water Hyacinth in China: A Sustainability Science-Based Management Framework." *Environmental Management* 40(6):823–30.

Lugo, A., L. A. Bravo-Inclán, J. Alcocer, M. L. Gaytán, Ma G. Oliva, Ma del R. Sánchez, M. Chávez, and G. Vilaclara. 1998. "Effect on the Planktonic Community of the Chemical Program Used to Control Water Hyacinth (*Eichhornia Crassipes*) in Guadalupe Dam, Mexico." *Aquatic Ecosystem Health & Management* 1(3–4):333–43.

Lundberg, Scott M., Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2(1):56–67.

Madsen, John D., Kien T. Luu, and Kurt D. Getsinger. 1993. "Allocation of Biomass and Carbohydrates in Waterhyacinth (*Eichhornia Crassipes*): Pond-Scale Verification."

Technical Report A93-3. Vicksburg, MS: US Army Corps of Engineers, Waterways Experiment Station.

Maluleke, Mary, Gavin C. G. Fraser, and Martin P. Hill. 2021. “Economic Evaluation of Chemical and Biological Control of Four Aquatic Weeds in South Africa.” *Biocontrol Science and Technology* 31(9):896–911.

Manfré, Luiz A., Eliane Hirata, Janaína B. Silva, Eduardo J. Shinohara, Mariana A. Giannotti, Ana Paula C. Larocca, and José A. Quintanilha. 2012. “An Analysis of Geospatial Technologies for Risk and Natural Disaster Management.” *ISPRS International Journal of Geo-Information* 1(2):166–85.

Manfreda, Salvatore, Matthew F. McCabe, Pauline E. Miller, Richard Lucas, Victor Pajuelo Madrigal, Giorgos Mallinis, Eyal ben Dor, David Helman, Lyndon Estes, and Giuseppe Ciraolo. 2018. “On the Use of Unmanned Aerial Systems for Environmental Monitoring.” *Remote Sensing* 10(4):641.

Mangas-Ramírez, Ernesto, and Manuel Elías-Gutiérrez. 2004. “Effect of Mechanical Removal of Water Hyacinth (*Eichhornia Crassipes*) on the Water Quality and Biological Communities in a Mexican Reservoir.” *Aquatic Ecosystem Health & Management* 7(1):161–68.

Marlin, Danica, Martin P. Hill, and Marcus J. Byrne. 2013. “Interactions within Pairs of Biological Control Agents on Water Hyacinth, *Eichhornia Crassipes*.” *Biological Control* 67(3):483–90.

Masifwa, Wanda Fred, Timothy Twongo, and Patrick Denny. 2001. “The Impact of Water Hyacinth, *Eichhornia Crassipes* (Mart) Solms on the Abundance and Diversity of Aquatic Macroinvertebrates along the Shores of Northern Lake Victoria, Uganda.” *Hydrobiologia* 452(1):79–88.

Mateo-Garcia, Gonzalo, Joshua Veitch-Michaelis, Lewis Smith, Silviu Vlad Oprea, Guy Schumann, Yarin Gal, Atılım Güneş Baydin, and Dietmar Backes. 2021. “Towards Global Flood Mapping Onboard Low Cost Satellites with Machine Learning.” *Scientific Reports* 11(1):1–12.

Matthews, Mark W., and Stewart Bernard. 2015. “Eutrophication and Cyanobacteria in South Africa’s Standing Water Bodies: A View from Space.” *South African Journal of Science* 111(5):1–8.

May, Bronwen, and Julie Coetzee. 2013. “Comparisons of the Thermal Physiology of Water Hyacinth Biological Control Agents: Predicting Establishment and Distribution Pre-and Post-release.” *Entomologia Experimentalis et Applicata* 147(3):241–50.

McKinney, Wes. 2011. “Pandas: A Foundational Python Library for Data Analysis and Statistics.” *Python for High Performance and Scientific Computing* 14(9):1–9.

McVea, Charles, and Claude E. Boyd. 1975. “Effects of Water Hyacinth Cover on Water Chemistry, Phytoplankton, and Fish in Ponds.” *Journal of Environmental Quality* 4:375–78.

Mehta, Shefali v, Robert G. Haight, Frances R. Homans, Stephen Polasky, and Robert C. Venette. 2007. “Optimal Detection and Control Strategies for Invasive Species Management.” *Ecological Economics* 61(2–3):237–45.

Meijer, Johan R., Mark A. J. Huijbregts, Kees C. G. J. Schotten, and Aafke M. Schipper. 2018. “Global Patterns of Current and Future Road Infrastructure.” *Environmental Research Letters* 13(6):64006.

Memmottlz, J., S. v Fowlerq, H. M. Harman, and L. M. Hayes. 1996. “How Best to Release a Biological Control Agent.” Pp. 291–96 in *Proceedings of the IX International Symposium on Biological Control of Weeds*, edited by V.C. Moran and J.H. Hoffmann. Cape Town, South Africa.: University of Cape Town.

Meyer, Hanna, Christoph Reudenbach, Stephan Wöllauer, and Thomas Nauss. 2019. “Importance of Spatial Predictor Variable Selection in Machine Learning Applications–Moving from Data Reproduction to Spatial Prediction.” *Ecological Modelling* 411:108815.

Miller, Benjamin E., Julie A. Coetzee, and Martin P. Hill. 2021. “Mind the Gap: The Delayed Recovery of a Population of the Biological Control Agent Megamelus Scutellaris Berg.(Hemiptera: Delphacidae) on Water Hyacinth after Winter.” *Bulletin of Entomological Research* 111(1):120–28.

Miskella, John J., and John D. Madsen. 2019. “The Effect of Temperature on Water Hyacinth Stem Base Regrowth.” *Journal of Aquatic Plant Management* 57:99–102.

Mitchell, David S. 1985. “Surface-Floating Aquatic Macrophytes.” Pp. 109–24 in *The ecology and management of African wetland vegetation*. Springer.

Moran, Patrick J. 2005. “Leaf Scarring by the Weevils *Neochetina Eichhorniae* and *N. Bruchi* Enhances Infection by the Fungus *Cercospora Piaropi* on Waterhyacinth, *Eichhornia Crassipes*.” *BioControl* 50(3):511–24.

Mueller, Norman, Adam Lewis, Dale Roberts, Steven Ring, R. Melrose, J. Sixsmith, Leo Lymburner, A. McIntyre, P. Tan, and S. Curnow. 2016. “Water Observations from Space: Mapping Surface Water from 25 Years of Landsat Imagery across Australia.” *Remote Sensing of Environment* 174:341–52.

Mukarugwiro, Jeanne d’Arc, S. W. Newete, E. Adam, F. Nsanganwimana, K. A. Abutaleb, and M. J. Byrne. 2019. “Mapping Distribution of Water Hyacinth (*Eichhornia Crassipes*) in Rwanda Using Multispectral Remote Sensing Imagery.” *African Journal of Aquatic Science* 44(4):339–48.

Müllerová, Jana, Jan Pergl, and Petr Pyšek. 2013. “Remote Sensing as a Tool for Monitoring Plant Invasions: Testing the Effects of Data Resolution and Image Classification Approach on the Detection of a Model Plant Species *Heracleum Mantegazzianum* (Giant Hogweed).” *International Journal of Applied Earth Observation and Geoinformation* 25:55–65.

Mullissa, Adugna, Andreas Vollrath, Christelle Odongo-Braun, Bart Slagter, Johannes Balling, Yaqing Gou, Noel Gorelick, and Johannes Reiche. 2021. “Sentinel-1 Sar Backscatter Analysis Ready Data Preparation in Google Earth Engine.” *Remote Sensing* 13(10):1954.

Muramoto, S., I. Aoyama, and Y. Oki. 1991. "Effect of Salinity on the Concentration of Some Elements in Water Hyacinth (*Eichhornia Crassipes*) at Critical Levels." *Journal of Environmental Science & Health Part A* 26(2):205–15.

Navarro, Luis A., and George Phiri. 2000. "Water Hyacinth in Africa and the Middle East: A Survey of Problems and Solutions." Pp. 72–76 in *ACIAR PROCEEDINGS. ACIAR*. Ottawa: International Development Research Centre.

Neiff, Juan Jose, Alicia Susana G. Poi, and Sylvina Lorena Casco. 2001. "The Effect of Prolonged Floods on *Eichhornia Crassipes* Growth in Paraná River Floodplain Lakes." *Acta Limnologica Brasiliensia* 13:51–60.

Neis, Pascal, and Dennis Zielstra. 2014. "Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap." *Future Internet* 6(1):76–106.

Newton, Adrian C., Ross A. Hill, Cristian Echeverría, Duncan Golicher, Jose M. Rey Benayas, Luis Cayuela, and Shelley A. Hinsley. 2009. "Remote Sensing and the Future of Landscape Ecology." *Progress in Physical Geography* 33(4):528–46.

Nielsen, Charlotte, Hans Peter Ravn, Wolfgang Nentwig, and Max Wade. 2005. *The Giant Hogweed Best Practice Manual. Guidelines for the Management and Control of an Invasive Weed in Europe*.

Njiru, M., J. B. Okeyo-Owuor, M. Muchiri, and I. G. Cowx. 2004. "Shifts in the Food of Nile Tilapia, *Oreochromis Niloticus* (L.) in Lake Victoria, Kenya." *African Journal of Ecology* 42(3):163–70.

Normand, Signe, Robert E. Ricklefs, Flemming Skov, Jesper Bladt, Oliver Tackenberg, and Jens-Christian Svenning. 2011. "Postglacial Migration Supplements Climate in Determining Plant Species Ranges in Europe." *Proceedings of the Royal Society B: Biological Sciences* 278(1725):3644–53.

Oberholster, Paul J., and Peter J. Ashton. 2008. "State of the Nation Report: An Overview of the Current Status of Water Quality and Eutrophication in South African Rivers and Reservoirs." *Parliamentary Grant Deliverable. Pretoria: Council for Scientific and Industrial Research (CSIR)* 2006.

Ofulla, A. V. O., D. Karanja, R. Omondi, T. Okurut, A. Matano, T. Jembe, R. Abila, P. Boera, and J. Gichuki. 2010. "Relative Abundance of Mosquitoes and Snails Associated with Water Hyacinth and Hippo Grass in the Nyanza Gulf of Lake Victoria." *Lakes & Reservoirs: Research & Management* 15(3):255–71.

Otsu, Nobuyuki. 1979. "A Threshold Selection Method from Gray-Level Histograms." *IEEE Transactions on Systems, Man, and Cybernetics* 9(1):62–66.

Owens, Chetta S., and J. D. Madsen. 1995. "Low Temperature Limits of Water Hyacinth." *Journal of Aquatic Plant Management* 33:63–68.

Pahlevan, Nima, Sandeep K. Chittimalli, Sundarabalan v Balasubramanian, and Vincenzo Vellucci. 2019. "Sentinel-2/Landsat-8 Product Consistency and Implications for Monitoring Aquatic Systems." *Remote Sensing of Environment* 220:19–29.

Palmer, Tony. 2004. *Vegetation of Makana (Technical Report)*.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt. 2022. “The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models.” *ArXiv Preprint ArXiv:2201.03544*.

Parker, Kelsey, Arthur Elmes, Peter Boucher, Richard A. Hallett, John E. Thompson, Zachary Simek, Justin Bowers, and Andrew B. Reinmann. 2021. “Crossing the Great Divide: Bridging the Researcher–Practitioner Gap to Maximize the Utility of Remote Sensing for Invasive Species Monitoring and Management.” *Remote Sensing* 13(20):4142.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine Learning Research* 12:2825–30.

Pekel, Jean-François, Andrew Cottam, Noel Gorelick, and Alan S. Belward. 2016. “High-Resolution Mapping of Global Surface Water and Its Long-Term Changes.” *Nature* 540(7633):418–22.

Pellegrini, Marco O. O., Charles N. Horn, and Rafael F. Almeida. 2018. “Total Evidence Phylogeny of Pontederiaceae (Commelinales) Sheds Light on the Necessity of Its Recircumscription and Synopsis of *Pontederia* L.” *PhytoKeys* (108):25.

Pérez, E. Albano, J. A. Coetzee, T. Ruiz Tellez, and M. P. Hill. 2011a. “A First Report of Water Hyacinth (Eichhornia Crassipes) Soil Seed Banks in South Africa.” *South African Journal of Botany* 77(3):795–800.

Pérez, E. Albano, J. A. Coetzee, T. Ruiz Tellez, and M. P. Hill. 2011b. “A First Report of Water Hyacinth (*Eichhornia Crassipes*) Soil Seed Banks in South Africa.” *South African Journal of Botany* 77(3):795–800.

Pérez-Ortiz, Juan Antonio, Felix A. Gers, Douglas Eck, and Juergen Schmidhuber. 2003. “Kalman Filters Improve LSTM Network Performance in Problems Unsolvable by Traditional Recurrent Nets.” *Neural Networks* 16(2):241–50.

Persello, Claudio, and Alfred Stein. 2017. “Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images.” *IEEE Geoscience and Remote Sensing Letters* 14(12):2325–29.

Phan, Thanh Noi, Verena Kuch, and Lukas W. Lehnert. 2020. “Land Cover Classification Using Google Earth Engine and Random Forest Classifier—The Role of Image Composition.” *Remote Sensing* 12(15):2411.

Pinto-Coelho, Ricardo Motta, and Magda Karla Barcelos Greco. 1999. “The Contribution of Water Hyacinth (*Eichhornia Crassipes*) and Zooplankton to the Internal Cycling of Phosphorus in the Eutrophic Pampulha Reservoir, Brazil.” *Hydrobiologia* 411:115–27.

Ploton, Pierre, Frédéric Mortier, Maxime Réjou-Méchain, Nicolas Barbier, Nicolas Picard, Vivien Rossi, Carsten Dormann, Guillaume Cornu, Gaëlle Viennois, and Nicolas Bayol. 2020. “Spatial Validation Reveals Poor Predictive Performance of Large-Scale Ecological Mapping Models.” *Nature Communications* 11(1):1–11.

Pro, Global Forest Watch, Forest Watcher, and Forest Atlases. 2014. “Global Forest Watch.” *World Resources Institute (WRI)*. Retrieved July 13, 2022 (http://www.globalforestwatch.org/).

Pullanagari, Rajasheker R., Gabor Kereszturi, and Ian Yule. 2018. “Integrating Airborne Hyperspectral, Topographic, and Soil Data for Estimating Pasture Quality Using Recursive Feature Elimination with Random Forest Regression.” *Remote Sensing* 10(7):1117.

Pyšek, Petr, and Philip E. Hulme. 2005. “Spatio-Temporal Dynamics of Plant Invasions: Linking Pattern to Process.” *Ecoscience* 12(3):302–15.

Pyšek, Petr, David M. Richardson, Jan Pergl, Vojtěch Jarošík, Zuzana Sixtová, and Ewald Weber. 2008. “Geographical and Taxonomic Biases in Invasion Ecology.” *Trends in Ecology & Evolution* 23(5):237–44.

Qiao, Cheng, Jiancheng Luo, Yongwei Sheng, Zhanfeng Shen, Zhiwen Zhu, and Dongping Ming. 2012. “An Adaptive Water Extraction Method from Remote Sensing Image Based on NDWI.” *Journal of the Indian Society of Remote Sensing* 40(3):421–33.

Radočaj, Dorijan, Jasmina Obhođaš, Mladen Jurišić, and Mateo Gašparović. 2020. “Global Open Data Remote Sensing Satellite Missions for Land Monitoring and Conservation: A Review.” *Land* 9(11):402.

Rahel, Frank J. 2002. “Homogenization of Freshwater Faunas.” *Annual Review of Ecology and Systematics* 33(1):291–315.

Rainford, James, Andrew Crowe, Glyn Jones, and Femke van den Berg. 2020. “Early Warning Systems in Biosecurity; Translating Risk into Action in Predictive Systems for Invasive Alien Species.” *Emerging Topics in Life Sciences* 4(5):453–62. doi: 10.1042/ETLS20200056.

Rejmánek, M., and M. J. Pitcairn. 2002. “When Is Eradication of Exotic Pest Plants a Realistic Goal.” *Turning the Tide: The Eradication of Invasive Species* 249–53.

Republic of South Africa. 2004. “National Environmental Management: Biodiversity Act 10 of 2004.” *Proc. R47/Government Gazette* (26887/20041008).

Ricciardi, Anthony, and Hugh J. MacIsaac. 2011. “Impacts of Biological Invasions on Freshwater Ecosystems.” *Fifty Years of Invasion Ecology: The Legacy of Charles Elton* 1:211–24.

Richards, Nica. 2022. “A New Aquatic Weed Threatens to Suffocate Hartbeespoort Dam.” *The Citizen*. Retrieved February 20, 2023 (https://www.citizen.co.za/news/south-africa/local-news/a-new-aquatic-weed-threatens-to-suffocate-hartbeespoort-dam/).

Richardson, David M., Llewellyn C. Foxcroft, Guillaume Latombe, David C. le Maitre, Mathieu Rouget, and John R. Wilson. 2020. “The Biogeography of South African Terrestrial Plant Invasions.” Pp. 67–96 in *Biological Invasions in South Africa*. Springer.

Richardson, David M., and Brian W. van Wilgen. 2004. “Invasive Alien Plants in South Africa: How Well Do We Understand the Ecological Impacts?: Working for Water.” *South African Journal of Science* 100(1–2):45–52.

Rodriguez-Gallego, L. R., N. Mazzeo, J. Gorga, M. Meerhoff, JCFS Clemente, C. Kruk, F. Scasso, G. Lacerot, J. García, and F. Quintans. 2004. "The Effects of an Artificial Wetland Dominated by Free-floating Plants on the Restoration of a Subtropical, Hypertrophic Lake." *Lakes & Reservoirs: Research & Management* 9(3-4):203–15.

Rodriguez-Merino, Argantonio, Pablo Garcia-Murillo, Santos Cirujano, and Rocio Fernandez-Zamudio. 2018. "Predicting the Risk of Aquatic Plant Invasions in Europe: How Climatic Factors and Anthropogenic Activity Influence Potential Species Distributions." *Journal for Nature Conservation* 45:58–71.

Roscher, Ribana, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. "Explainable Machine Learning for Scientific Insights and Discoveries." *Ieee Access* 8:42200–216.

Roy, David P., V. Kovalskyy, H. K. Zhang, Eric F. Vermote, L. Yan, S. S. Kumar, and A. Egorov. 2016. "Characterization of Landsat-7 to Landsat-8 Reflective Wavelength and Normalized Difference Vegetation Index Continuity." *Remote Sensing of Environment* 185:57–70.

Rufchaie, Rudabeh, Alireza Mirvaghefi, Seyed Hossein Hoseinifar, and Alireza Valipour. 2018. "Anti-Microbial Activity of *Echhornia Crasippes* Aquatic and Hydromethanolic Leaves Extract." *Journal of Fisheries* 71(1):31–41.

Russo, Daniel J., Benjamin van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2018. "A Tutorial on Thompson Sampling." *Foundations and Trends® in Machine Learning* 11(1):1–96.

Russo, Daniel, and Benjamin van Roy. 2016. "An Information-Theoretic Analysis of Thompson Sampling." *The Journal of Machine Learning Research* 17(1):2442–71.

Ryo, Masahiro, Boyan Angelov, Stefano Mammola, Jamie M. Kass, Blas M. Benito, and Florian Hartig. 2021. "Explainable Artificial Intelligence Enhances the Ecological Interpretability of Black-box Species Distribution Models." *Ecography* 44(2):199–205.

Sabater, Sergi, Andrea Butturini, Jean-Christophe Clement, Tim Burt, David Dowrick, Mariet Hefting, Véronique Matre, Gilles Pinay, Carmen Postolache, and Marek Rzepecki. 2003. "Nitrogen Removal by Riparian Buffers along a European Climatic Gradient: Patterns and Factors of Variation." *Ecosystems* 6(1):20–30.

Sala, Osvaldo E., F. Stuart Chapin, Juan J. Armesto, Eric Berlow, Janine Bloomfield, Rodolfo Dirzo, Elisabeth Huber-Sanwald, Laura F. Huenneke, Robert B. Jackson, and Ann Kinzig. 2000. "Global Biodiversity Scenarios for the Year 2100." *Science* 287(5459):1770–74.

Santamaría, Luis. 2002. "Why Are Most Aquatic Plants Widely Distributed? Dispersal, Clonal Growth and Small-Scale Heterogeneity in a Stressful Environment." *Acta Oecologica* 23(3):137–54.

Scheffer, Marten, Sándor Szabo, Alessandra Gragnani, Egbert H. van Nes, Sergio Rinaldi, Nils Kautsky, Jon Norberg, Rudi M. M. Roijackers, and Rob J. M. Franken. 2003. "Floating Plant Dominance as a Stable State." *Proceedings of the National Academy of Sciences* 100(7):4040–45.

Schramm Jr, Harold L., Kurt J. Jirka, and Mark v Hoyer. 1987. “Epiphytic Macroinvertebrates on Dominant Macrophytes in Two Central Florida Lakes.” *Journal of Freshwater Ecology* 4(2):151–61.

Schulze, R. E., and M. Maharaj. 2007. “Median First and Last Dates of Heavy Frost, Their Variability, and the Duration of the Frost Period.” *Schulze, RE (Ed)*.

Schumann, Guy, Laura Giustarini, Angelica Tarpanelli, Ben Jarihani, and Sandro Martinis. 2022. “Flood Modeling and Prediction Using Earth Observation Data.” *Surveys in Geophysics* 1–26.

Shea, K., and Hugh Philip Possingham. 2000. “Optimal Release Strategies for Biological Control Agents: An Application of Stochastic Dynamic Programming to Population Management.” *Journal of Applied Ecology* 37(1):77–86.

Shea, Katriona, Hugh P. Possingham, William W. Murdoch, and Rick Roush. 2002. “Active Adaptive Management in Insect Pest and Weed Control: Intervention with a Plan for Learning.” *Ecological Applications* 12(3):927–36.

Singh, Geethen, Chevonne Reynolds, Marcus Byrne, and Benjamin Rosman. 2020. “A Remote Sensing Method to Monitor Water, Aquatic Vegetation, and Invasive Water Hyacinth at National Extents.” *Remote Sensing* 12(24):4021.

Smith, Adam B., and Maria J. Santos. 2020. “Testing the Ability of Species Distribution Models to Infer Variable Importance.” *Ecography* 43(12):1801–13.

Srivastava, Vaibhav, Paul Reverdy, and Naomi E. Leonard. 2013. “On Optimal Foraging and Multi-Armed Bandits.” Pp. 494–99 in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE.

Srivastava, Vivek, Valentine Lafond, and Verena C. Griess. 2019. “Species Distribution Models (SDM): Applications, Benefits and Challenges in Invasive Species Management.” *CAB Rev* 14(10.1079).

Stehman, Stephen v. 1997. “Selecting and Interpreting Measures of Thematic Classification Accuracy.” *Remote Sensing of Environment* 62(1):77–89.

Stirnberg, Roland, Jan Cermak, Simone Kotthaus, Martial Haeffelin, Hendrik Andersen, Julia Fuchs, Miae Kim, Jean-Eudes Petit, and Olivier Favez. 2021. “Meteorology-Driven Variability of Air Pollution (PM 1) Revealed with Explainable Machine Learning.” *Atmospheric Chemistry and Physics* 21(5):3919–48.

Strange, Emily F., Pietro Landi, Jaclyn M. Hill, and Julie A. Coetzee. 2019. “Modeling Top-down and Bottom-up Drivers of a Regime Shift in Invasive Aquatic Plant Stable States.” *Frontiers in Plant Science* 10:889.

Strayer, David L., and David Dudgeon. 2010. “Freshwater Biodiversity Conservation: Recent Progress and Future Challenges.” *Journal of the North American Benthological Society* 29(1):344–58.

Sullivan, Paul R., and Rod Wood. 2012. “Water Hyacinth (*Eichhornia Crassipes* (Mart.) Solms) Seed Longevity and the Implications for Management.” Pp. 37–40 in *Eighteenth Australasian weeds conference*. Weed Society of Victoria Inc Melbourne, Australia.

Sun, Ziheng, Laura Sandoval, Robert Crystal-Ornelas, S. Mostafa Mousavi, Jinbo Wang, Cindy Lin, Nicoleta Cristea, Daniel Tong, Wendy Hawley Carande, and Xiaogang Ma. 2022. "A Review of Earth Artificial Intelligence." *Computers & Geosciences* 105034.

Surendraraj, Alagarsamy, K. H. Sabeena Farvin, and R. Anandan. 2013. "Antioxidant Potential of Water Hyacinth (*Eichornia Crassipes*): In Vitro Antioxidant Activity and Phenolic Composition." *Journal of Aquatic Food Product Technology* 22(1):11–26.

Taylor, Philip G., and Alan R. Townsend. 2010. "Stoichiometric Control of Organic Carbon–Nitrate Relationships from Soils to the Sea." *Nature* 464(7292):1178–81.

Téllez, Trinidad Ruiz, EMDR López, Gloria Lorenzo Granado, Eva Albano Pérez, Ricardo Morán López, and Juan Manuel Sánchez Guzmán. 2008. "The Water Hyacinth, *Eichhornia Crassipes*: An Invasive Plant in the Guadiana River Basin (Spain)." *Aquatic Invasions* 3(1):42–53.

Thamaga, Kgabo H., and Timothy Dube. 2019. "Understanding Seasonal Dynamics of Invasive Water Hyacinth (*Eichhornia Crassipes*) in the Greater Letaba River System Using Sentinel-2 Satellite Data." *GIScience & Remote Sensing* 56(8):1355–77.

Thamaga, Kgabo Humphrey, and Timothy Dube. 2018a. "Remote Sensing of Invasive Water Hyacinth (*Eichhornia Crassipes*): A Review on Applications and Challenges." *Remote Sensing Applications: Society and Environment* 10:36–46.

Thamaga, Kgabo Humphrey, and Timothy Dube. 2018b. "Testing Two Methods for Mapping Water Hyacinth (*Eichhornia Crassipes*) in the Greater Letaba River System, South Africa: Discrimination and Mapping Potential of the Polar-Orbiting Sentinel-2 MSI and Landsat 8 OLI Sensors." *International Journal of Remote Sensing* 39(22):8041–59.

Theobald, David M., Dylan Harrison-Atlas, William B. Monahan, and Christine M. Albano. 2015. "Ecologically-Relevant Maps of Landforms and Physiographic Diversity for Climate Adaptation Planning." *PloS One* 10(12):e0143619.

Tipping, Philip W., Melissa C. Smith, Ellen C. Lake, Carey R. Minteer, Ashley B. C. Goode, Jeremiah R. Foley, and Lyn A. Gettys. 2020. "Classical Biological Control and Apparent Competition: Evaluating a Water Hyacinth Invaded Community Module." *Journal of Applied Ecology* 57(5):926–35.

Tiwari, Suchi, Savita Dixit, and Neelam Verma. 2007. "An Effective Means of Biofiltration of Heavy Metal Contaminated Water Bodies Using Aquatic Weed *Eichhornia Crassipes*." *Environmental Monitoring and Assessment* 129(1):253–56.

Truong, Tuyet T. A., Giles E. St J. Hardy, and Margaret E. Andrew. 2017. "Contemporary Remotely Sensed Data Products Refine Invasive Plants Risk Mapping in Data Poor Regions." *Frontiers in Plant Science* 8:770.

Tucker, Compton J. 1979. "Red and Photographic Infrared Linear Combinations for Monitoring Vegetation." *Remote Sensing of Environment* 8(2):127–50.

Turner, Monica G. 2005. "Landscape Ecology: What Is the State of the Science?" *Annu. Rev. Ecol. Evol. Syst.* 36:319–44.

Turner, Monica G., Robert H. Gardner, Robert v O'neill, and Robert v O'Neill. 2001. *Landscape Ecology: In Theory and Practice*. Vol. 401. New York: Springer.

Twongo, T., and G. Howard. 1998. “Ways with Weeds.” *New Scientist* 159(2142):57.

Valavi, Roozbeh, Jane Elith, José J. Lahoz-Monfort, and Gurutzeta Guillera-Arroita. 2019. “Block CV: An r Package for Generating Spatially or Environmentally Separated Folds for K-fold Cross-validation of Species Distribution Models.” *Methods in Ecology and Evolution* 10(2):225–32.

Vaz, Ana Sofia, Domingo Alcaraz-Segura, Joao C. Campos, Joana R. Vicente, and Joao P. Honrado. 2018. “Managing Plant Invasions through the Lens of Remote Sensing: A Review of Progress and the Way Forward.” *Science of The Total Environment* 642:1328–39.

Venter, Nic, Blair W. Cowie, Ed T. F. Witkowski, Gavin C. Snow, and Marcus J. Byrne. 2017. “The Amphibious Invader: Rooted Water Hyacinth’s Morphological and Physiological Strategy to Survive Stranding and Drought Events.” *Aquatic Botany* 143:41–48.

Venter, Zander S., David N. Barton, Tirthankar Chakraborty, Trond Simensen, and Geethen Singh. 2022. “Global 10 m Land Use Land Cover Datasets: A Comparison of Dynamic World, World Cover and Esri Land Cover.” *Remote Sensing* 14(16):4101.

Vila, Montserrat, and Inés Ibáñez. 2011. “Plant Invasions in the Landscape.” *Landscape Ecology* 26(4):461–72.

Villa, Paolo, Alex Laini, Mariano Bresciani, and Rossano Bolpagni. 2013. “A Remote Sensing Approach to Monitor the Conservation Status of Lacustrine Phragmites Australis Beds.” *Wetlands Ecology and Management* 21(6):399–416.

Villamagna, A. M., and B. R. Murphy. 2010. “Ecological and Socio-economic Impacts of Invasive Water Hyacinth (*Eichhornia Crassipes*): A Review.” *Freshwater Biology* 55(2):282–98.

VonBank, Jay A. 2015. *An Assessment of Aquatic Invasive Plants in the Illinois River: Water Hyacinth Surveillance, Mapping, Persistence, and Potential Seed Dispersal*. Western Illinois University.

VonBank, Jay A., Andrew F. Casper, Jennifer E. Pendleton, and Heath M. Hagy. 2018. “Water Hyacinth (*Eichhornia Crassipes*) Invasion and Establishment in a Temperate River System.” *River Research and Applications* 34(10):1237–43.

Vorster, Anthony G., Brian D. Woodward, Amanda M. West, Nicholas E. Young, Robert G. Sturtevant, Timothy J. Mayer, Rebecca K. Girma, and Paul H. Evangelista. 2018. “Tamarisk and Russian Olive Occurrence and Absence Dataset Collected in Select Tributaries of the Colorado River for 2017.” *Data* 3(4):42.

Vos, Kilian, Kristen D. Splinter, Mitchell D. Harley, Joshua A. Simmons, and Ian L. Turner. 2019. “CoastSat: A Google Earth Engine-Enabled Python Toolkit to Extract Shorelines from Publicly Available Satellite Imagery.” *Environmental Modelling & Software* 122:104528.

Wallace, Rebekah D., Charles T. Bargeron, L. Ziska, and J. Dukes. 2014. “Identifying Invasive Species in Real Time: Early Detection and Distribution Mapping System (EDDMapS) and Other Mapping Tools.” *Invasive Species and Global Climate Change* 4:219.

Walters, Carl J., and Ray Hilborn. 1978. “Ecological Optimization and Adaptive Management.” *Annual Review of Ecology and Systematics* 9(1):157–88.

Wang, Feier, Yixu Wang, Kai Zhang, Ming Hu, Qin Weng, and Huichun Zhang. 2021. “Spatial Heterogeneity Modeling of Water Quality Based on Random Forest Regression and Model Interpretation.” *Environmental Research* 202:111660.

Wang, Runzi, Jun-Hyun Kim, and Ming-Han Li. 2021. “Predicting Stream Water Quality under Different Urban Development Pattern Scenarios with an Interpretable Machine Learning Approach.” *Science of The Total Environment* 761:144057.

Wang, Zifeng, Junguo Liu, Jinbao Li, and David D. Zhang. 2018. “Multi-Spectral Water Index (MuWI): A Native 10-m Multi-Spectral Water Index for Accurate Water Mapping on Sentinel-2.” *Remote Sensing* 10(10):1643.

Waskom, Michael L. 2021. “Seaborn: Statistical Data Visualization.” *Journal of Open Source Software* 6(60):3021.

Westgate, Martin J., Gene E. Likens, and David B. Lindenmayer. 2013. “Adaptive Management of Biological Systems: A Review.” *Biological Conservation* 158:128–39.

Westphal, Michael I., Michael Browne, Kathy MacKinnon, and Ian Noble. 2008. “The Link between International Trade and the Global Distribution of Invasive Alien Species.” *Biological Invasions* 10(4):391–98.

van Wilgen, B. W., D. M. Richardson, D. C. le Maitre, C. Marais, and D. Magadlela. 2001. “The Economic Consequences of Alien Plant Invasions: Examples of Impacts and Approaches to Sustainable Management in South Africa.” *Environment, Development and Sustainability* 3(2):145–68.

Wilson, John R., Niels Holst, and Mark Rees. 2005. “Determinants and Patterns of Population Growth in Water Hyacinth.” *Aquatic Botany* 81(1):51–67.

Wisz, Mary Susanne, Julien Pottier, W. Daniel Kissling, Loïc Pellissier, Jonathan Lenoir, Christian F. Damgaard, Carsten F. Dormann, Mads C. Forchhammer, John-Arvid Grytnes, and Antoine Guisan. 2013. “The Role of Biotic Interactions in Shaping Distributions and Realised Assemblages of Species: Implications for Species Distribution Modelling.” *Biological Reviews* 88(1):15–30.

Witt, Raphael, Lukas Loos, and Alexander Zipf. 2021. “Analysing the Impact of Large Data Imports in OpenStreetMap.” *ISPRS International Journal of Geo-Information* 10(8):528.

Wittenberg, Rüdiger, and Matthew J. W. Cock. 2005. “Best Practices for the Prevention and Management of Invasive Alien Species.” *Scope-Scientific Committee on Problems of the Environment International Council of Scientific Unions* 63:209.

Wolter, Peter T., David J. Mladenoff, George E. Host, and Thomas R. Crow. 1995. “Using Multi-Temporal Landsat Imagery.” *Photogrammetric Engineering & Remote Sensing* 61(9):1129–43.

Wondie, Ayalew. 2010. “Improving Management of Shoreline and Riparian Wetland Ecosystems: The Case of Lake Tana Catchment.” *Ecohydrology & Hydrobiology* 10(2–4):123–31.

Woodcock, Curtis E., Richard Allen, Martha Anderson, Alan Belward, Robert Bindschadler, Warren Cohen, Feng Gao, Samuel N. Goward, Dennis Helder, and Eileen Helmer. 2008. “Free Access to Landsat Imagery.” *Science* 320:1011.

World bank group. 2021. “Climate Change Knowledge Portal, Current Climate.” Retrieved July 14, 2022 (https://climateknowledgeportal.worldbank.org/country/south-africa/climate-data-historical).

Wu, Hao, Juli Carrillo, and Jianqing Ding. 2017. “Species Diversity and Environmental Determinants of Aquatic and Terrestrial Communities Invaded by *Alternanthera Philoxeroides*.” *Science of the Total Environment* 581:666–75.

Wu, Hao, and Jianqing Ding. 2020. “Abiotic and Biotic Determinants of Plant Diversity in Aquatic Communities Invaded by Water Hyacinth [*Eichhornia Crassipes* (Mart.) Solms].” *Frontiers in Plant Science* 1306.

Wu, Jianguo, and Richard Hobbs. 2002. “Key Issues and Research Priorities in Landscape Ecology: An Idiosyncratic Synthesis.” *Landscape Ecology* 17(4):355–65.

Wu, Jianguo Jingle. 2006. “Landscape Ecology, Cross-Disciplinarity, and Sustainability Science.” *Landscape Ecology* 21(1):1–4.

Wu, Qiusheng. 2020. “Geemap: A Python Package for Interactive Mapping with Google Earth Engine.” *Journal of Open Source Software* 5(51):2305.

Wu, Qiusheng, Charles R. Lane, Xuecao Li, Kaiguang Zhao, Yuyu Zhou, Nicholas Clinton, Ben DeVries, Heather E. Golden, and Megan W. Lang. 2019. “Integrating LiDAR Data and Multi-Temporal Aerial Imagery to Map Wetland Inundation Dynamics Using Google Earth Engine.” *Remote Sensing of Environment* 228:1–13.

Wundrow, Emily J., Juli Carrillo, Christopher A. Gabler, Katherine C. Horn, and Evan Siemann. 2012. “Facilitation and Competition among Invasive Plants: A Field Experiment with Alligatorweed and Water Hyacinth.” *PloS One* 7(10):e48444.

Wurm, Michael, Thomas Stark, Xiao Xiang Zhu, Matthias Weigand, and Hannes Taubenböck. 2019. “Semantic Segmentation of Slums in Satellite Images Using Transfer Learning on Fully Convolutional Neural Networks.” *ISPRS Journal of Photogrammetry and Remote Sensing* 150:59–69.

van Wyk, E., and B. W. van Wilgen. 2002. “The Cost of Water Hyacinth Control in South Africa: A Case Study of Three Options.” *African Journal of Aquatic Science* 27(2):141–49.

Wynne, Timothy, Andrew Meredith, Richard Stumpf, Travis Briggs, and Wayne Litaker. 2020. “Harmful Algal Bloom Forecasting Branch Ocean Color Satellite Imagery Processing Guidelines, 2020 Update.” *NOAA Technical Memorandum NOS NCCOS* 252:48.

Xiao, Xiangming, Stephen Boles, Steve Frolking, William Salas, B. Moore Iii, C. Li, L. He, and R. Zhao. 2002. “Observation of Flooding and Rice Transplanting of Paddy Rice Fields at the Site to Landscape Scales in China Using VEGETATION Sensor Data.” *International Journal of Remote Sensing* 23(15):3009–22.

Xu, Chen, and Yao Xie. 2020. “Conformal Prediction for Dynamic Time-Series.” *ArXiv Preprint ArXiv:2010.09107*.

Xu, Hanqiu. 2006. “Modification of Normalised Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imagery.” *International Journal of Remote Sensing* 27(14):3025–33.

Xu, Shengjun, Ping Lyu, Xiaoxu Zheng, Haijun Yang, Bing Xia, Hui Li, Hao Zhang, and Shuanglong Ma. 2022. “Monitoring and Control Methods of Harmful Algal Blooms in Chinese Freshwater System: A Review.” *Environmental Science and Pollution Research* 1–20.

Yang, Jingkang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2021. “Generalized Out-of-Distribution Detection: A Survey.” *ArXiv Preprint ArXiv:2110.11334*.

Yang, Xiucheng, Qiming Qin, Hervé Yésou, Thomas Ledauphin, Mathieu Koehl, Pierre Grussenmeyer, and Zhe Zhu. 2020. “Monthly Estimation of the Surface Water Extent in France at a 10-m Resolution Using Sentinel-2 Data.” *Remote Sensing of Environment* 244:111803.

Yu, Hao, Arthur R. Cooper, and Dana M. Infante. 2020. “Improving Species Distribution Model Predictive Accuracy Using Species Abundance: Application with Boosted Regression Trees.” *Ecological Modelling* 432:109202.

Yu, Kui, Xianjie Guo, Lin Liu, Jiuyong Li, Hao Wang, Zhaolong Ling, and Xindong Wu. 2020. “Causality-Based Feature Selection: Methods and Evaluations.” *ACM Computing Surveys (CSUR)* 53(5):1–36.

Zengeni, Rebecca, Vincent Kakembo, and Nsalambi Nkongolo. 2016. “Historical Rainfall Variability in Selected Rainfall Stations in Eastern Cape, South Africa.” *South African Geographical Journal* 98(1):118–37.

Zhang, Hankui K., David P. Roy, Lin Yan, Zhongbin Li, Haiyan Huang, Eric Vermote, Sergii Skakun, and Jean-Claude Roger. 2018. “Characterization of Sentinel-2A and Landsat-8 Top of Atmosphere, Surface, and Nadir BRDF Adjusted Reflectance and NDVI Differences.” *Remote Sensing of Environment* 215:482–94.

Zhang, Yunlin, Erik Jeppesen, Xiaohan Liu, Boqiang Qin, Kun Shi, Yongqiang Zhou, Sidinei Magela Thomaz, and Jianmin Deng. 2017. “Global Loss of Aquatic Vegetation in Lakes.” *Earth-Science Reviews* 173:259–65.

Zhichkina, L. N., V. v Nosov, K. A. Zhichkin, H. T. Aydinov, V. N. Zhenzhebir, and V. v Kudryavtsev. 2020. “Satellite Monitoring Systems in Forestry.” P. 032043 in *Journal of Physics: Conference Series*. Vol. 1515. IOP Publishing.

Zhou, Qian, XiaoFang Zhang, Jin Xu, and Bin Liang. 2017. “Large-Scale Bandit Approaches for Recommender Systems.” Pp. 811–21 in *International Conference on Neural Information Processing*. Springer.

Zhou, Xinzhi, Haijia Wen, Ziwei Li, Hui Zhang, and Wengang Zhang. 2022. “An Interpretable Model for the Susceptibility of Rainfall-Induced Shallow Landslides Based on SHAP and XGBoost.” *Geocarto International* (just-accepted):1–27.

Zhou, Yan, Jinwei Dong, Xiangming Xiao, Tong Xiao, Zhiqi Yang, Guosong Zhao, Zhenhua Zou, and Yuanwei Qin. 2017. “Open Surface Water Mapping Algorithms: A Comparison of Water-Related Spectral Indices and Sensors.” *Water* 9(4):256.

Zhu, Xiao Xiang, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. 2017. “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources.” *IEEE Geoscience and Remote Sensing Magazine* 5(4):8–36.

Zhu, Zhe, Michael A. Wulder, David P. Roy, Curtis E. Woodcock, Matthew C. Hansen, Volker C. Radeloff, Sean P. Healey, Crystal Schaaf, Patrick Hostert, and Peter Strobl. 2019. “Benefits of the Free and Open Landsat Data Policy.” *Remote Sensing of Environment* 224:382–85.

Zimmels, Y., F. Kirzhner, and A. Malkovskaja. 2007. “Advanced Extraction and Lower Bounds for Removal of Pollutants from Wastewater by Water Plants.” *Water Environment Research* 79(3):287–96.