DEVELOPMENT OF THE WATER RESEARCH OBSERVATORY AND CASE STUDIES ON MACHINE LEARNING APPLICATIONS

Report to the

WATER RESEARCH COMMISSION

by

MICHAEL VAN DER LAAN & CINDY VIVIERS (Editors)

with

Simphiwe Maseko, Christiaan Schutte, Aimee Thomson, Pitso Khoboko, Michael Silberbauer, Jay le Roux, Leushantha Mudaly, Harold Weepener, Gerrit Hoogenboom, Srinivasan Raghavan, David Clark, Richard Kunz

WRC Report No. 3121/1/23

ISBN 978-0-6392-0593-9

March 2024

Obtainable from
Water Research Commission
Bloukrans Building, Lynnwood Bridge Office Park
4 Daventry Street
Lynnwood Manor
PRETORIA

orders@wrc.org.za or download from www.wrc.org.za

This is the final report of WRC project no. C2020/2021-00440.

The research presented in this report was possible because of the existence of extensive datasets containing the results of many years of fine-resolution fieldwork. The modelling processes would have been mere conjecture without data on groundwater, streamflow, crop yield and soil type, while neural machine translation can only function when previously existing texts are available, painstakingly translated by multilingual experts in technical language. Artificial intelligence enhances human intelligence, it does not replace it.

- **Dr Michael Silberbauer (5 December 2023)**

This page was intentionally left blank

Executive Summary

Motivation

South Africa (SA) is facing mounting water scarcity, infrastructure, and pollution challenges. Southern Africa has been identified as one of the most vulnerable regions in the world to anthropogenic climate change and faces increased temperatures, droughts, and more intensive rainfall and floods. In the current Fourth Industrial Revolution (4IR) era of rapidly evolving technology, the need for not only a data repository but also a cloud-based platform that can enable big data processing and analytics, artificial intelligence (AI) applications, and data visualisation was recognised by the Water Research Commission (WRC). The Water Research Observatory (WRO) has now been developed to harness 4IR tools in support of water research and management in SA.

WRO Platform

The following document reports on the design of the WRO platform, including data storage architecture, metadata requirements and the data access management system. An open-source data portal, Comprehensive Knowledge Archive Network (CKAN), was selected for the user interface (data uploading and discovery), and the Google Cloud Platform (GCP) for data asset storage and various applications of the data. The system is interoperable with numerous other commercial cloud platforms such as Microsoft Azure, Amazon Web Services, and IBM Cloud.

Metadata

Metadata is information about data, and is essential for making data FAIR – Findable, Accessible, Interoperable and Reusable. In designing the metadata capture form, careful attention was paid to SANS (South African National Standard) 1878 and ISO (International Organization for Standardization) 19115/19139 and 19115-1/2 standards. The metadata required for a user to upload data assets are: (1) Dataset title, (2) Author details, (3) Contact person details, (4) Dataset description (max 500 words), (5) Organisation, (6) Private/Public, (7) Recommended citation, (8) Organization/Publisher, (9) Publication date, (10), Project number, (11) License (under which data is to be released, e.g. CC 4.0), (12) Keywords, (13) Geographic co-ordinates (14) Data reference dates, (15) Alternate identifier, and (16) Vertical extent data. The WRO user manual in the form of a WRC technology transfer document can be consulted for more information on the metadata required and how to use the platform.

Additional selections are made by the user when uploading data to enable a unique structured storage system in the GCP bucket named WRC-WRO. According to this storage architecture the user first selects the 'Topic category' under which the dataset falls (e.g. Agriculture, Biodiversity, Groundwater, Food Security, Wetlands, etc.), whether the data are structured, semi-structured or unstructured, the extent to which the data have been processed (Raw, Still being processed, Refined, Access), and the temporal nature of the data (Time series, Static, Both). This will enable future users to also search for data within the GCP storage bucket by navigating to the folder of interest and being able to preview the data.

Digitisation of historical information

New and improved software are enabling the efficient digitisation of historical research, such as old pdf research files that were written on typewriters decades ago. Making these documents machine readable allows the more granular search for relevant information thereby increasing researcher efficiency. It can also allow the application of artificial intelligence tools at previously unimaginable levels such as was seen with the launch of Chat-GPT the lifespan of this project. Several machine learning (ML) tools were tested to convert WRC final report pdfs to machine readable text, but most proved to be very expensive and often inaccurate. In the end Google Docs was used to convert a test batch of 200 final report pdfs to doc format, which are machine readable. It has been demonstrated that searching these machine-readable documents can find useful information faster by not only identifying relevant documents but also the places within the documents that contain the information being sought. It is recommended that all WRC documents be converted so they can be searched in this way, and that a 'WRC research-GPT' model be built to better harness the value of previous research and improve research efficiency.

Access management

For access management, data that is uploaded to the WRO can be made available to the public or can be private to (1) an individual, (2) an organisation, or (3) to a group comprising selected individuals and/or organisations. An individual, organisation or group may want to keep data private, for example, if there are still documents such as scientific articles that need to be published before release of the data, or if an embargo has been placed on the dataset's release. Funders can then be assured that the data has been uploaded to a suitable and sustainable repository and has a unique URL (Uniform Resource Locator). In all cases, it will be especially important to comply with the Protection of Personal Information Act (Act 4 of 2013) (often called the POPI Act or POPIA). If a data asset does contain potentially sensitive information, an option is to just upload the metadata for the dataset onto the WRO, and then interested parties can email the contact person to request the data and sharing arrangements made between the two entities.

Any new users who register need to first be added to an organisation. Due to the wide array of potential data contributors, the WRO will use organisations to assist with managing the platform by holding their employees accountable for responsible use of the WRO. To this end, each organisation actively using the platform will need to appoint an administrator.

WRC research project archiving protocol

Recommendations have now been made regarding a protocol for current and future WRC research projects to archive important data assets and other information collected in the WRO for subsequent use by others. The platform is not intended to be for only WRC project research, however, and other water data and information that is relevant may also be uploaded onto the platform. It is noted that research projects and other activities conducted or funded by, for example, the Department of Water and Sanitation (DWS), the Department of Agriculture, Land

Reform and Rural Development, the Department of Science and Innovation, and many other institutions also produce data assets that are suitable for uploading onto the WRO.

Estimating groundwater levels using remote sensing and machine learning

The innovative use of remotely sensed data to address gaps in *in situ* groundwater measurements was investigated. Monthly estimates of groundwater storage anomalies (GWSA) at a resolution of 0.25°, assimilating Gravity Recovery and Climate Experiment (GRACE) measurements through the Global Land Data Assimilation System Version 2.2 (GLDAS-2.2) were studied. The hypothesis was that two open-source, higher-resolution datasets, namely Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) for precipitation and Moderate Resolution Imaging Spectroradiometer (MODIS) for evapotranspiration (ET), could effectively be used downscale GLDAS-2.2 GWSA to 0.05°. To test this hypothesis, a random forest (RF) model was trained and tested across the Steenkoppies Catchment in SA. Additional experiments incorporating optimizing temporal lags estimated using machine learning (ML) demonstrated enhanced model performance. The intergranular and fractured aquifer achieved the highest correlation coefficient (*r*) and the lowest Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values at 0.62, 43 mm, and 33 mm, respectively. This downscaled GWSA product holds significant potential for informing IWRM, especially *in situ*ations where groundwater data is limited. These high-resolution GWS estimates, comparable to borehole observations, are also valuable for estimating aquifer recharge and the impacts of climate change on groundwater.

Predicting streamflow using machine learning

Understanding how much water is flowing in our rivers is essential for managing water resources effectively. There is a growing concern globally because the number of active measurement stations is decreasing, with an estimated two-thirds no longer operating or in decline. Deep learning (a type of ML) techniques, like Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) networks, show promise in estimating streamflow but their use in SA has not been well explored. This study tested the use of GRU and LSTM networks to predict river flow in two specific areas of the Steelpoort River in SA. The models used combinations of rainfall, temperature, and past river flow rates to make predictions. Using data from commercial weather stations produced dependable predictions, while using freely available gridded weather data resulted in only moderately accurate predictions. This approach shows excellent potential for both long-term data gap filling in streamflow records and for short term daily time-step forecasts, which could be beneficial for flood risk management and planning purposes. Look back windows of 10 to 30 days were found to be suitable for achieving accurate predictions with both GRU and LSTM models.

Application of machine learning in precision agriculture

The application of ML in predicting maize yield in a precision agriculture-managed commercial field was assessed. Different ML models were evaluated using Data Intensive Farm Management (DIFM) datasets. Additionally, the research aimed to investigate the potential of ML models to identify limiting factors and determine optimal input rates in a spatially variable field. Conducted in Henneman, Free State, the study utilized data from an 80 ha

commercial field for the 2019/2020 and 2020/2021 seasons. Four ML models – multiple linear regression (MLR), multilayer perceptron, decision tree (DT), and random forest (RF) – were trained and tested using crop management, soil, and remotely sensed data. Results showed a complex relationship between maize yield and various environmental attributes for the two seasons. The RF model consistently outperformed other models in sub-field yield predictions. Urea application was identified as the most critical variable across all scenarios, with its importance varying based on seasonal rainfall. Other influential factors included soil phosphorus, plant population, soil pH, clay content, sodium (in 2020) and plant population (in 2021). The research highlighted the importance of considering temporal and spatial variations in environmental factors for accurate ML models in agricultural systems. The study underscored the potential of RF models for accurate yield predictions in SA and emphasized the need to adapt and train ML with diverse weather, soil, and the environment variables to further increase accuracy.

Translation of farming guidelines from English to isiZulu, isiXhosa, and Sepedi

The WRC, ARC and various national and provincial government departments possess various farming guideline documentation in English, with limited translations into native languages like isiZulu, isiXhosa, and Sepedi. These languages are often spoken by smallholder farmers in specific regions. While these guidelines could significantly benefit farmers, language barriers hinder comprehension, particularly impacting smallholders who may struggle with English. Given climate change and evolving farming techniques, access to this information is crucial for effective farm management and increased yields. To address this, a study used a bilingual corpus (a collection of parallel texts or linguistic data in two languages) and Transformer models to achieve machine translation through natural language processing. Positive preliminary results were achieved, though Sepedi presented challenges due to loanwords from English affecting translation quality. Future work should involve obtaining more translated texts for agriculture, refining the translation models, and evaluating translation quality using suitable metrics. The transfer learning technique demonstrated potential for translating with limited data, offering promise for further applications in this domain.

Digital soil maps for South Africa

While there is an ever-increasing amount of earth surface measurement data being made freely available for water resources research and management, reliable high-quality soil maps remain a challenge in big data applications. In addition to the Agricultural Research Council's Land Type soil maps, several digital soil maps (DSMs) have been made available for SA but are of unknown accuracy. Five DSMs [SWAT-SA, Innovative Solutions for Decision Agriculture (iSDA), Africa SoilGrids 250m (AfSG250), Harmonised World Soil Database version 1.2 (HWSD), and SoilGrids-for-DSSAT-10 km (SG-DSSAT)] were tested using measured data from across SA within the context of obtaining reliable soil parameters to run crop and hydrological models. The study found that globally or continentally created DSMs often lose effectiveness in the diverse SA escarpment. This underscores the need to assess site-specific map accuracy before use, and highlights the importance of developing or refining locally tailored DSMs for SA. Despite associated errors, DSMs have proven valuable in enhancing soil surveys with

incomplete data, particularly in data-scarce regions like SA. Using a combination of DSMs is currently recommended: clay data from AfSG250, silt data from SWAT-SA, SOC from SG-DSSAT, bulk density for 0-0.3m from SWAT-SA and AfSG250 estimates for 0.3-1.0m, and pH information from iSDA.

User Manual

A separate Water Research Observatory User Manual accompanies this report, and the latest version can be downloaded at https://www.waterresearchobservatory.org/.

WRO management needs

Recommendations have been made for the future management of the WRO. Human capital resources will be needed to manage the administrative tasks such as adding newly registered users to organisations, managing groups, monitoring costs and overseeing the regular software updates. Ensuring compliance with SA legislation will also be important. The maintenance of the platform will either need to be provided in-house or outsourced to an information technology service provider. It is noted that the DWS is currently undergoing a digitalisation transition.

Capacity building

One honours student, one masters student, and three PhD students were directly involved and financially supported in this project. An ARC intern was also trained as well as multiple workshop attendees. In total three scientific publications have been submitted, five presentations given at national conferences, three presentations at international conferences, and one poster at national and another poster at an international conference. Two popular press articles and one interview for an international media outlet were also done. Four invited presentations were given on the WRO.

Outlook for the WRO

The long-term success of the WRO will be intricately linked to the quality of the data assets and models added, and this will require meticulously captured metadata and a well-designed curation system. A suitable method for the scientific community to rate the quality of the data should also be explored.

The prospect of having all previous water research material in machine readable formats, combined with data such as that from traditional research, citizen science, internet of things and social media posts could revolutionise the information available to improve water resource management. The WRO will complement other national and global initiatives of this nature.

In the next phase of the project, it is suggested that meetings be held regularly with other institutions with similar platforms or intentions to create similar platforms to explore ways to ensure interoperability as well as cost savings.

Steady investment will be needed to keep advancing the WRO with technological developments in the 4IR and beyond.

Acknowledgements

Name and Surname	Role
Dr Luxon Nhamo	**Water Research Commission (Chairman)**
Dr Shafick Adams	**Water Research Commission (Co-Project Manager and Co-Chairperson)**
	Research Project Team Members
Dr Michael van der Laan	**Agricultural Research Council/University of Pretoria (Project Leader)**
Mr Pitso Khoboko	University of Pretoria (Research team member)
Ms Aimee Thomson	University of Pretoria (Research team member, MSc Student)
Ms Leushantha Mudaly	University of Pretoria (Research team member)
Mr Christiaan Schutte	University of Pretoria (Research team member, PhD Student, Co-Secretary)
Mr Simphiwe Maseko	University of Pretoria (Research team member, PhD Student, Co-Secretary)
Ms Cindy Viviers	University of Pretoria (Research team member, PhD Student, Co-Secretary)
Mr Harold Weepener	Agricultural Research Council (Research team member)
Dr Mike Silberbauer	Department of Water and Sanitation (Retired) (Research team member)
Dr David Clark	University of KwaZulu-Natal (Research team member)
Mr Richard Kunz	University of KwaZulu-Natal (Research team member)
	Water Research Commission Reference Group Team Members
Ms Mpho Kapani	Water Research Commission (Intern)
Dr Samkelisiwe Hlophe-Ginindza	Water Research Commission (Reference Group member)
Mr Bonani Madikizela	Water Research Commission (Reference Group member)
Mr Deon Thirumalai	Water Research Commission (Reference Group member)
Mr Marius Snyman	Water Research Commission (Reference Group member)
Mr Wandile Nomquphu	Water Research Commission (Reference Group member)
Prof. Clint Mawing	Water Research Commission (Reference Group member)
Prof. Sylvester Mpandeli	Water Research Commission (Reference Group member)
	Reference Group Team Members from other Institutions
Dr Bongani Ncube	Cape Peninsula University of Technology (Reference Group member)
Prof. Bongani Ncube	Cape Peninsula University of Technology (Reference Group member)
Prof. Moses Azong Cho	Council of Scientific & Industrial Research (CSIR) (Reference Group member)
Mr Makhawana Mxolisi	Department of Water and Sanitation (Reference Group member)
Mr Terry Newby	GeoTerraImage (Reference Group member)
Mr James Takawira Magidi	National University of Science and Technology (Reference Group member)
Dr Joel Botai	South African Weather Service (Reference Group member)
Prof. Adriaan van Niekerk	Stellenbosch University (Reference Group member)
Dr James Takawira Magidi	Tshwane University of Technology (Reference Group member)
	Other Institutional Support
Dr Maneshree Jugmohan-Naidu	Department of Science and Technology, South Africa
Mr Divan Vermeulen	Kartoza (Pty) Ltd Consulting Company (Software Developer)
Mr Mohab Khaled	Kartoza (Pty) Ltd Consulting Company (Software Developer)
Ms Juanique Voogt	Kartoza (Pty) Ltd Consulting Company (Software Developer)
Ms Nancy Job	South Africa National Biodiversity Institute (SANBI)

The research team would like to thank the Reference Group members for all their advice and words of encouragement. We also thank various stakeholders for meeting with us and sharing ideas and insights for building this platform, both in formal and informal discussions. Finally, we express our gratitude to Ms. Sandra Fritz for her exceptional administration of this project

This page was intentionally left blank

Table of Contents

1. Introduction 1
   1.1. Report outline 1
   1.2. Background 2
2. Overview of the Water Research Observatory and management recommendations 4
   2.1. The building blocks of the WRO 4
   2.2. Proposed data archiving protocol for future WRC projects 7
   2.3. The Hydrological Model for South Africa (HAMSA) 9
   2.4. The Decision Support System for Agrotechnology Transfer (DSSAT) 12
   2.5. WRO continuity 12
3. Enhancing the resolution of GRACE-assimilated groundwater storage anomalies across two aquifer types in the Steenkoppies Catchment 14
   3.1. Introduction 14
   3.2. Materials and Methods 16
   3.3. Results and discussions 24
   3.4. Conclusions 32
4. Predicting streamflow in South Africa using deep learning 33
   4.1. Introduction 33
   4.2. Materials and Methods 35
   4.3. Results and Discussion 41
   4.4. Conclusions 49
5. Big data analytics in precision agriculture 51
   5.1. Introduction 51
   5.2. Materials and Methods 53
   5.3. Results and Discussion 58
   5.4. Conclusions 66
6. Digital soil maps for South Africa 67
   6.1. Introduction 67
   6.2. Materials and Methods 67
   6.3. Results and Discussion 72
   6.4. Conclusions 78
7. Transformer-based Neural Machine Translation for Native South African Languages 79
   7.1. Introduction 79
   7.2. Materials and Methods 80
   7.3. Results and discussion 81

7.4. Conclusion 84
8. Conclusions 86
9. References 87
Appendix 99
Appendix I: Capacity Building 99
Appendix II: Research Outputs 102
Appendix III: Streamflow Prediction with Deep learning 105

List of Tables

Table 1: Proposed themes or topic categories under which relevant data and information will be stored in the Water Research Observatory 6

Table 2: Model performance metrics calculated for Models 1 and 2 before residual correction. 27

Table 3: Comparison metrics between the in situ groundwater level storage and the downscaled Groundwater Storage Anomaly (GWSA) for the respective zones. 30

Table 4: Combinations of weather sources and input variables for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and weather station (ARC) data. 41

Table 5: P-values for differences in Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) values between the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council weather station data (ARC) for Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). 43

Table 6: P-values for differences in Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) values between Combination 1 and 2 for National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council weather station data (ARC) for Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM). 44

Table 7: Nash-Sutcliffe Efficiency values for the best Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) and ensemble predictions for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data. 45

Table 8: Kling-Gupta Efficiency values for the best Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) and ensemble predictions for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data. 46

Table 9: The agronomic management, soil and remotely sensed variables used in model development. 56

Table 10: Descriptive statistics of maize yield for the 2019/2020 and 2020/2021 seasons. 58

Table 11: Statistical analysis comparison of machine learning regression models on DIFM trial Uitsny maize field for 2019/2020, 2020/201, and combined dataset with and without NDVI evaluated using the 80/20 training and testing analysis (MAPE: mean absolute percentage error, RMSE: root mean square error). 63

Table 12: The root mean square error (RMSE) across all sites for silt estimates (%) of the respective DSMs under study. A heat map of how well the DSMs perform relative to one another has been included, where green is indicative of predictions with smaller error, and red is indicative of predictions with larger error. (n can be found in Table 14) 70

Table 13: Number of instances in which the respective DSMs produced under-estimates (measured > estimated), over-estimates (measured < estimated), and precise (measured = estimated) silt values 72

Table 14: Number of data points available for comparison between measured silt and estimated silt according to depth for each DSM under study. 72

Table 15: The normalised root mean square error (nRMSE) values for clay, silt, soil organic carbon (SOC), pH, and bulk density (BD) parameters. The DSMs showing the lowest average accuracy per parameter have been indicated with a bold outline. 73

Table 16: The average relative root mean square error (RRMSE) displayed by the DSMs for each parameter (SOC = soil organic carbon, BD = bulk density). The overall DSM accuracy is given as well as the overall accuracy at which parameters are predicted. 74

Table 17: BLEU score results of joeyNMT and mBART parent models before being fine-tuned with an agriculture domain dataset. 81

Table 18: BLEU score results of joeyNMT and mBART child models before being fine-tuned with an agriculture domain dataset. 81

Table 19: Examples of joeyNMT model translations for an agriculture domain sentence before being fine-tuned to the agriculture domain dataset. 82

Table 20: Examples of joeyNMT model translations for agriculture domain sentence after being fine-tuned for the agriculture domain dataset. 82

Table 21: Examples of mBART model translations for agriculture domain sentence before being fine-tuned for the agriculture domain dataset. 83

Table 22: Examples of mBART model translations for agriculture domain sentence after being fine-tuned for the agriculture domain dataset. 84

List of Figures

Figure 1: Schematic of an annual water balance for a hydrological response unit (HRU) simulated in the Hydrological Model for South Africa (HAMSA)..... 10

Figure 2: Interface for the Lower Vaal basin (top) and example of how different scenarios can be compared (bottom) 11

Figure 3: Location and surface geology of the Steenkoppies quaternary catchment and dolomite compartment, as well as the distribution of the downscaling validation in situ data in relation to the 0.25° Global Land Data Assimilation System (GLDAS)-2.2 Groundwater Storage (GWS) and the 0.05° downscale target resolution. 18

Figure 4: Flowchart of the downscaling model design in this study (CHIRPS - Climate Hazards Group InfraRed Precipitation with Station, MODIS - Moderate Resolution Imaging Spectroradiometer, GLDAS-2.2 Global Land Data Assimilation System (GLDAS) Version 2.2, GWSA - Groundwater Storage Anomaly). 22

Figure 5: Temporal variation of monthly Climate Hazards Group InfraRed Precipitation With Station Data (CHIRPS) precipitation, Moderate Resolution Imaging Spectroradiometer (MODIS) evapotranspiration (ET) and Global Land Data Assimilation System (GLDAS-2.2) Groundwater Storage Anomaly (GWSA), using the spatial aggregated average across quaternary catchment A21F: (a) Time-series monthly changes (b) and fitted harmonic model to characterise seasonal variability. 25

Figure 6: Spatial distribution of the temporal correlation between the Global Land Data Assimilation System (GLDAS-2.2) Groundwater Storage Anomaly (GWSA) to the a) three months lagged Climate Hazards Group InfraRed Precipitation With Station Data (CHIRPS) precipitation estimates and b) two months lagged Moderate Resolution Imaging Spectroradiometer (MODIS) evapotranspiration (ET) estimates at a monthly temporal and 0.05° spatial resolution. 26

Figure 7: Performance assessment of the respective machine learning (ML) models for predicting 0.25° Groundwater Storage Anomaly (GWSA) using Model 1 input variables without temporal lags, and Model 2 with input variables adjusted with the respective temporal lags...... 27

Figure 8: Monthly sum of the groundwater storage anomalies across catchment A21F for the downscaled product and the original coarse-resolution Global Land Data Assimilation System Version 2.2 (GLDAS-2.2) product...... 28

Figure 9: The geographic distribution of the Groundwater Storage Anomaly (GWSA) standard deviation (2003-2021) before (a) and after (b) downscaling. 28

Figure 10: Map of aquifer specific yield overlaid on a groundwater hydraulic head map. 29

Figure 11: Comparison between groundwater level storage and the downscaled Groundwater Storage Anomaly (GWSA) for Zone 1 (a), Zone 2 (b) and Zone 3 (c) (Zone 4 is not illustrated as it is visually not highly distinguishable from Zone 3). 31

Figure 12: Locality map of the catchments, streamflow stations, the Agricultural Research Council (ARC) weather station and the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) grid point. Catchment shapefiles were uploaded to the CHIRPS website to download catchment averaged rainfall. 36

Figure 13: A comparison of the internal architecture of a standard Long Short-Term Memory (LSTM) cell (left) and a Gated Recurrent Unit (GRU) cell (right). The LSTM cell is characterized by its input, output, and forget gates as well as its cell state. The GRU cell uses a more streamlined architecture with update and reset gates but without a separate cell state...... 37

Figure 14: Schematic representation of the deep learning architecture with four input nodes, five hidden layers, each comprising 25 Gated Recurrent Unit (GRU) or 25 Long Short-Term Memory (LSTM) cells, where each of

these cells utilizes a look-back window of 30 days. A dense layer follows the hidden layers, leading to the final prediction. 38

Figure 15: Based on the preceding 30 days of past weather and streamflow values in the look-back window (LBW), the network predicts a streamflow value for the next day. The model then moves one time step forward, ingesting the predicted streamflow value, along with observed weather values for that day, into the LBW to make the next prediction. 39

Figure 16: Cumulative frequency distribution (CDF) for Nash-Sutcliffe Efficiency (NSE) (top) and Kling-Gupta Efficiency (KGE) values (bottom) for Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council weather station data (ARC). Values smaller than -1 are not shown, with -10.9 being the lowest. 42

Figure 17: Best model of the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data plotted against observed streamflow for two years of the testing set. 47

Figure 18: The maize (Zea mays L.) field study was conducted in Henneman in the Free State province of South Africa. An example of different seeding rate treatments is shown. 54

Figure 19: Rainfall distribution for the two seasons (2019/20 and 2020/21) covering the period from planting to harvesting. 59

Figure 20: Correlation analysis for the relationship between agronomic management, soil, remotely sensed, and weather data and maize yields for the 2019/2020 and 2020/2021 seasons and combined data for the seasons. 61

Figure 21: Comparison of the performance of machine learning algorithms for season 1 (2019/2020) and season 2 (2020/2021) and combining the data from the two seasons with and without NDVI (MLR: multiple linear regression, MLP: multilayer perceptron, DT: decision tree, RF: random forest). 62

Figure 22: Feature importance from the random forest for 2019/2022, 2020/2021, and the combination of the two-season data with and without NDVI using the 80/20 training and testing analysis (Plant_pop: plant population, Urea: urea application, ph_top: soil pH in topsoil, bray_top: phosphorus in topsoil, K_top: potassium in topsoil, Mg_top: magnesium in topsoil, Na_top: sodium in topsoil, S_top: sulphur in topsoil, Clay_top: clay content in topsoil, Bray_sub: phosphorus in sub soil, K_sub: potassium in sub soil, Mg_sub: magnesium in sub soil, Na_sub: sodium in sub soil, S_sub: Sulphur in sub soil, Clay_sub: Clay content in sub soil, Soil_d: soil depth). 65

Figure 23: The study area distribution across South Africa, where SOSA refers to sites obtained from Soils of South Africa (Fey, 2010). 68

Figure 24: Comparison between measured and estimated silt content (%) for SWAT-SA, Innovative Solutions for Decision Agriculture (iSDA), Africa SoilGrids 250 m (AfSG250), Harmonised World Soil Database version 1.2 (HWSD), and SoilGrids-for-DSSAT-10 km (SG-DSSAT) DSMs at different soil depths. Only six of 25 study sites shown. 69

Figure 25: The spread of error at different depths for the various DSM silt content (%) predictions, where at 0%, measured silt = DSM-estimated silt. Thus, values clustered around the y axis reflect less error (n can be found in Table 14). 71

Figure 26: Negative streamflow predictions in Catchment A. The dotted line indicates 0 m^3s^{-1}. 105

Figure 27: Nash-Sutcliffe Efficiency (NSE) values for different look-back window sizes. 106

List of Acronyms

Acronym	Definition
4IR	Fourth Industrial Revolution
AfSP	Africa Soil Profiles
AGMIP	Agricultural Model Intercomparison and Improvement Project
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARC	Agricultural Research Council
ASPD	African Soil Profile Database
ATP	Agricultural Technology Provider
BD	Bulk Density
BLEU	BiLingual Evaluation Understudy
CART	Classification and Regression Trees
CKAN	Comprehensive Knowledge Archive Network
CLM	Community Land Model
CNN	Convolutional Neural Networks
CSIR	Centre for Scientific and Industrial Research
CWS	Canopy Water Storage
DIFM	Data Intensive Farm Management
DL	Deep Learning
DOI	Digital Object Identifier
DSM	Digital Soil Map
DSSAT	Decision Support System for Agrotechnology Transfer
DT	Decision Trees
DWAF	Department of Water and Forestry
DWS	Department of Water and Sanitation
E-GIS	Environmental Geographic Information System
EML	Ecological Metadata Language
EN-NSO	English to Northern Sesotho
EN-XH	English to isiXhosa
EN-ZUL	English to isiZulu
ET	Evapotranspiration
FBIS	Freshwater Biodiversity Information System
GCP	Google Cloud Platform
GEE	Google Earth Engine
GLDAS	Global Land Data Assimilation System
GPS	Global Positioning System
GRACE	Gravity Recovery and Climate Experiment
GRU	Gated Recurrent Unit
GWS	Groundwater Storage
GWSA	Groundwater Storage Anomaly
HAMSA	HydrologicAI Model for South Africa
HAWQS	Hydrologic And Water Quality System
HWSD	Harmonised World Soil Database

Acronym	Definition
IoT	Internet of Things
iSDA	Innovative Solutions for Decision Agriculture
ISO	International Organization for Standardization
IT	Information Technology
IWRM	Integrated Water Resource Management
JoeyNMT	Joey Neural Machine Translation
LAI	Leaf Area Index
LSM	Land Surface Model
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
mBART	Multilingual Denoising Pre-training for Neural Machine Translation
ML	Machine Learning
MLP	Multi Layer Perceptron
MLR	Multiple Linear Regression
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NIWIS	National Integrated Water Information System
NLC	National Land Cover
NMT	Neural Machine Translation
nRMSE	Normalised Root Mean Square Error
ODP	Open Data Platform
OFP	On-Farm Experimentation
OPUS	Open Parallel Corpora
PA	Precision Agriculture
PTF	Pedotransfer Function
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RQI	Resource Quality Information
SAEON	South African Earth Observation Network
SANBI	South African National Biodiversity Institute
SANS	South African National Standards
SAPD	South African Profile Database
SASDI	South African Spatial Data Infrastructure
SCYM	Scalable satellite-based Crop Yield Mapper
SG-DSSAT	Soil Grids for DSSAT
SMS	Soil Moisture Storage
SOC	Soil Organic Carbon
SOSA	Soils of South Africa
SOTERSAF	Soil and Terrain Southern Africa
SVM	Support Vector Machines
SWAT	Soil and Water Assessment Tool
SWS	Surface Water Storage
TWS	Terrestrial Water Storage

Acronym	Definition
UAV	Unmanned Aerial Vehicle
USGS	United States Geological Survey
WEFE	Water-Energy-Food-Ecosystems
WMT22	2022 Seventh Conference On Machine Translation
WRC	Water Research Commission
WRO	Water Research Observatory

This page was intentionally left blank

1. Introduction

Michael van der Laan, Cindy Viviers, Simphiwe Maseko, Christiaan Schutte, Aimee Thomson, Pitso Khoboko, Michael Silberbauer, Jay le Roux, Leushantha Mudaly, Harold Weepener, Gerrit Hoogenboom, Srinivasan Raghavan, Richard Kunz, David Clark

South Africa (SA) is facing major water challenges in the form of increasing water scarcity and a declining quality of the resource. Data for water resources management is being collected at unprecedented levels, but our ability to store and analyse the data and to gain actionable information is lagging far behind. Big data refers to large data volumes from heterogenous sources that are growing exponentially over time, and often cannot be analysed using conventional techniques. The potential of combining different types of data to gain deeper insights could be extremely useful in Integrated Water Resource Management (IWRM) and Water-Energy-Food-Ecosystems (WEFE) Nexus decision making.

Recognising the need to not only archive the data collected and digital assets produced during Water Research Commission (WRC)-funded research projects, but also to allow the more widespread storage and application of water-related data in SA, a project was commissioned to build a big data platform called the Water Research Observatory (WRO). In this final report together with the WRO User Manual, the authors report on the design of the platform, including the data storage architecture, metadata standards and data access management system.

There have been two guiding principles in the development of the WRO. The first was 'data democratisation', which refers to the accessibility of digital information to the average end-user. It entails empowering non-specialists to independently gather and analyse data without the need for external assistance. The second guiding principle was 'inter-operability', which refers to the ability of computerised systems to connect and communicate with one another readily.

1.1. Report outline

Following this first introductory section, in **Section 2**, an overview of the WRO is provided, including the digital platforms used to build it, metadata requirements, data storage and taxonomy information, and a proposed protocol for archiving data from completed WRC projects. Important aspects to ensure WRO continuity are lastly discussed. In **Section 3,** Viviers and co-authors explored how remotely sensed groundwater storage estimates can be downscaled to a high resolution using open-source rainfall and ET data in Google Earth Engine. In **Section 4,** Schutte and co-authors investigated whether Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks could be developed in South Africa, a country which is currently experiencing a decline in streamflow measurements, to generate reliable streamflow estimations for predictive and data gap filling purposes. The second aim investigated whether freely available gridded weather data could be used as input to the models to produce reasonably accurate streamflow estimates. **Section 5** discusses how Maseko and co-authors investigated

the use of big data analytics in precision agriculture by evaluating machine learning approaches for subfield yield predictions using DIFM datasets. In **Section 6**, Thomson and co-authors investigated the accuracy of several digital soil maps (DSMs) in a South African context by comparing DSM-estimated and measured values for clay, silt, soil organic carbon (SOC), pH, and bulk density. In **Section 7,** Khoboko and co-authors experimented on creating a neural machine translation model (NMT) that can translated agricultural information from English to native South African languages. In efforts of translating bring new agricultural research to South African small-scale farmers in their home languages.

1.2. Background

Heterogeneous potentially useful data include, physically measured hydrological data such as rainfall, streamflow, and groundwater levels and quality, imagery and derived outputs from satellite and unmanned aerial vehicle (UAV) based sensors, information from agricultural and industrial sectors, municipal records, as well as emerging sources like internet of things (IoT) and social media data. Additionally, large amounts of data are also being generated by predictive models, including those focused on climate change analysis and prediction.

If access to adequate computer storage and processing capabilities are available, big data can be analysed using artificial intelligence (AI), for example, to reveal trends and associations and inform decision-making. Application of big data in environmental and agricultural management is lagging far behind business intelligence (Sun and Scanlon, 2019). The application of data-intensive farming has been estimated to increase global profits from crops by tens of billions of US dollars each year (Kamilaris et al., 2017). Machine Learning (ML) is a component of AI, where algorithms learn unapparent relationships from large training datasets (Huntingford et al., 2019). A ML model can be descriptive (gaining knowledge from collected data and describing what had happened) or predictive (making predictions for the future) depending on the research problem (Van Klompenburg et al., 2020). As an example, ML-based groundwater models provide an alternative to physics-based numerical process models and use strongly correlated physical parameters as input variables (Gaffoor et al., 2020; Kanda et al., 2018).

Supervised ML uses labelled datasets (input-output pairs) to train models using an appropriate learning algorithm (such as neural networks) that typically works through some optimization routine to minimise a loss or errors function (DeepAI, 2021). The 'loss' or 'error' refers to the average squared difference between the estimated values and the actual value. Land cover classification is a simple example. On the other hand, unsupervised ML learns patterns from untagged data with the objective of identifying underlying structures or distribution in the data without prior labels or groups. ML algorithms primarily use structured, labelled data to make predictions; generally, the unstructured data used requires some pre-processing to organise it into a structured format. DL on the other hand eliminates some of the data pre-processing required for ML (IBM, 2020).

Deep Learning (DL), also known as deep neural learning or deep neural network, is a type of ML in AI (DeepAI, 2021). DL is different from classical ML with consideration to the type of data it processes and the methods in which it learns (IBM, 2020). DL is an AI technique that has been shown to better exploit spatial and temporal

structures in earth data, compared to traditional ML methods (Reichstein, 2019). DL is essentially a neural network with three or more layers. As appose to organizing data to run through predefined equations, deep learning sets up basic parameters about the data and trains the computer to learn on its own by recognizing patterns using many layers of processing (IBM, 2020).

That said, there lies challenges in managing and analysing diverse datasets which require advanced technological proficiency. Additionally, extracting valuable scientific insights for decision-making from these datasets involve expertise from specialists in relevant scientific or environmental domains. Hence either computer scientists must increasingly integrate environmental skills or environmental scientists must upskill in advanced computational abilities. Because of this, mainstream cloud companies like Google, Amazon, Microsoft, IBM and more have focused on enabling non-data experts to find and utilise data through the Fourth Industrial Revolution (4IR) proliferation by developing interfaces to visualise and process heterogeneous data. All around the world including SA, a wide range of software and databases have been used to store, process and access data. Selection of databases was driven by factors such as cost, existing institutional licences, user needs, and in-house information technology (IT) skills, capacity and familiarity. As a result, important water-related data is hosted on a range of systems, including Microsoft Access and SQL Server, Oracle, MySQL, PostgreSQL, SAP HANA, MongoDB, IBM Db2, and Firebird.

The current challenge is to migrate these databases to one or more centralised cloud platform when it makes sense, and to integrate these different sources of data and information for the more efficient, holistic and sustainable management of South Africa's water resource.

2. Overview of the Water Research Observatory and management recommendations

Michael van der Laan, Cindy Viviers, Simphiwe Maseko, Christiaan Schutte, Aimee Thomson, Pitso Khoboko, Michael Silberbauer, Jay le Roux, Leushantha Mudaly, Harold Weepener, Gerrit Hoogenboom, Srinivasan Raghavan, Richard Kunz, David Clark

2.1. The building blocks of the WRO

During the first phase of the project, meetings were held with various important stakeholders in the SA water industry to gauge what platforms were being used at that time and collect recommendations and lessons learnt from other groups.

The project team reviewed many high-quality big data cloud platforms available, including Amazon Web Services, Microsoft Azure, Oracle, IBM, Hydroshare, ERDAS APOLLO and ArcGIS Online. The pricing for these cloud systems is complex and depends on factors such as volume of data and region stored in, frequency of access, and processing power required during analytics. Based on this review and online discussion with key stakeholders, the team selected the Google Cloud Platform (GCP) as the main cloud platform for the WRO. Data will still be accessible for import and application in other platforms for user-specific needs, and the system can be migrated to another platform in the future if required.

The following criteria were of key importance to the selection of the GCP: The GCP is highly compatible with Google Earth Engine (GEE) which is increasingly applied for environment analysis applications. Transient satellite imagery and other important baseline datasets are accessible through GEE. In addition to making provision for the use of popular programming languages such as Java and Python, the GCP also accommodates 'no-code' tools. The GCP provides an 'end-to-end' platform for data science and ML for multiple skill levels through cloud or virtual processing, allowing stakeholders with limited computer hardware resources to run big data analytics. The GCP accommodates various open-source ML frameworks, continually contributed to by data scientists and computer engineers worldwide. The GCP is perceived as highly secure and low risk, while being highly interoperable with other cloud platforms. The GCP utilises a 'pay as you go system' based on the resource quantity used, although special offers are available for 'environmental' cases, for example, the free storage of baseline datasets. Google Cloud Platform and GEE training and illustration material are available in abundance, including YouTube videos.

CKAN

The Comprehensive Knowledge Archive Network (CKAN) is an open-source data management system that is widely used by governments, organizations, and communities around the world to manage and share datasets (https://docs.ckan.org/en/2.9/, 27 November 2023). CKAN provides a web-based interface for managing data, as

well as APIs that allow users to interact with the system programmatically. CKAN is designed to power data hubs and portals, and facilitates the worldwide publication, sharing, and use of data. CKAN's codebase is maintained by the Open Knowledge Foundation (https://en.wikipedia.org/wiki/CKAN, accessed 27 November 2023). Since it is free yet well supported, its selection is expected to lead to long term savings in hosting the WRO. There is extensive documentation and support available on the internet regarding the platform, and programmers can learn the operating system relatively quickly.

Users can discover data and information using search terms, and the platform uses fuzzy logic to identify and rank the data and information It is also possible to indicate the geographic region of interest on a map and all the data available for the point or polygon will appear in the results. Additionally, it is also possible to preview data and create simple data visualization dashboards within CKAN.

Metadata

The WRO data repository is located at https://data.waterresearchobservatory.org.

Metadata is captured according to ISO 19115/19139 and SANS 1878 standards and datasets can be kept private to an individual, organisation or group of individuals from different organisations, or a group of different organisations. For more information on metadata, please see the WRO User Manual.

Google Cloud Platform

The GCP presents a high-performance infrastructure for cloud computing, data analytics and ML, and offers elevated levels of security and reliability. Code can be written in multiple languages, including Java, C++, Python, Go and Ruby. Diverse types of storage systems are available, depending on factors such as the volume of data stored, access frequency, and speed of access. These factors also largely determine the cost of using the platform to store data.

A major advantage to the SA scientific community offered by cloud computing is that operators can use virtual machines (VMs) that run on other servers, meaning that even people with a standard computer and internet connection are able to perform big data analytics. The GCP is highly interoperable with other cloud computing and storage platforms, and data sharing with a wide range of stakeholders, including those using other software platforms, is envisaged.

The first storage 'bucket' has been created within the WRO GCP. It is a basic container to store data and control access. Unlike directories and folders, a user cannot create other buckets within the main bucket. As the project evolves and more data gets stored, the management structure and access, and editing permissions, of different buckets within the WRO 'Data Lake' will be carefully described.

Data storage and taxonomy

The entire WRO data repository is called a 'data lake'. Data can be uploaded in structured, semi-structured and unstructured formats.

Structured data is clearly labelled and in a standardised format, for example, a daily weather data table containing the variables date, rainfall, and maximum and minimum temperature in separate columns. Semi-structured data does not fully conform to the tabular format of structured data, but may contain tags or markers identifying properties to arrange it into an organisational framework. Unstructured data cannot be stored in relational databases and is often stored in its raw format, such as photographs of rivers or social media posts. In the GCP, a project labelled 'wrc-wro' represents the lake. Users may request to join the WRO GCP project, and upload their data according to recommendations where it will become citable and discoverable. Users may also choose to have their own 'data warehouse' within the WRO, or can establish their own external cloud platform that can interoperate with the WRO.

The organisation of data and information is made more understandable using a hierarchical structure of folders, and sub-folders can be nested up to 10 levels deep. Different 'roles' (owner, editor, viewer, administrator, curator) and levels of accessibility can be granted to different users. The proposed approach for water data organisation makes use of themes as proposed in Table 1. More themes can be added as needed, and existing themes can be divided into sub-themes, for example, agriculture can be further divided into aquaculture, crops and livestock and so on. Data that cuts across themes can be stored in the most suitable folder, as judged by the provider.

Table 1: Proposed themes or topic categories under which relevant data and information will be stored in the Water Research Observatory

Agriculture	Floods	Streamflow
Biodiversity	Food security	Transboundary water
Citizen Science	Groundwater	Water quality
Dam level	Hydrological data & modelling	Water scarcity
Drought	Legislation	Water user associations
Economics	Marine water	Weather & climate data
Ecosystems	Mine water	Wetlands
Estuaries	Social	…

Within each theme (Table 1) there will be structured, semi-structured and unstructured data sub-folders; within each of these folders will be 'raw', 'refined', 'processing' and 'access' zones. Finally, within each of these zones will be static and time series data sub-folders. Static data comprises datasets that do not change frequently (if ever) over time, for example, SA geology maps. Time series data comprises data that changes frequently. For example, streamflow and water quality measurements for a particular Department of Water and Sanitation (DWS) station in SA, to which recent data is continually added or where values of a certain constituent concentration may

be backcalibrated. The WRO will have a 'transient loading' folder for data being uploaded and still being checked by the user. Best practices will be recommended to partners uploading data to ensure the sustainability of the platform.

2.2. Proposed data archiving protocol for future WRC projects

To ensure efficient data archiving for future WRC projects within the WRO, a well-defined data archiving protocol should be established. This protocol will outline the procedures and guidelines for storing, organizing, and accessing data, models, and information on the WRO. The recommendations for data archiving for existing and future WRC projects are as follows:

1. Future project proposal should already include a brief description of how data will be archived and made available for future researchers (considered prior to archiving, an item for validating and verifying the accuracy, completeness, and integrity of the data to ensure its reliability for future researchers).
2. The project contract should bind the hosting institution and project leader to uploading the data before the final contract amount (20% of total budget) is paid out.
3. Depending on the nature of the project, all research output should be archived. Research output encompasses the broad range of materials produced during the research process, including raw data, processed data, experimental results, analytical models, software code, algorithms, and documentation. A concern here is digitally large outputs due to the cost of storage (link to point on cold storage below).
4. Encourage the use of open and standard data formats to maximize interoperability and ease of data integration. Provide guidelines (WRO User upload document?) and resources for researchers to convert their data into appropriate formats and ensure proper documentation of data formats used.
5. Depending on the nature of the research outputs, the files can be stored in a single zipped folder, or as a single dataset. The first option may be more straightforward, but the second option will be preferential for important standalone outputs/datasets that should be made FAIR (Findable, Accessible, Interoperable and Reusable).
6. If the research team would like to propose that any outputs are not stored in the cloud, this should be approved by the Reference Group.
7. The research outputs should be archived under the organisation that was the host institution and signatory of the contract of the WRC project.
8. Metadata must be recorded as meticulously as possible to ensure the data is FAIR.
9. Even after the WRC research project has come to an end contractually, researchers should continue to upload any outputs that have been generated or refined.
10. Uploaders of research outputs should familiarise themselves with the Protection of Personal Information Act 4 of 2013 (POPI Act) (https://www.gov.za/documents/protection-personal-information-act) and ensure they are not breaking the law. Some way to handle personal information is to anonymise or aggregate the data.
11. For datasets containing sensitive or personal information, the option exists to only upload the metadata. People interested in the data can then approach the contact person for private access to the data.

12. It is possible for an individual or organisation to store their data in the own cloud for which they cover the fees and then to link relevant data to the WRO.

    Each organisation that stores data on the WRO platform will need to appoint an Administrator. This person / These persons are responsible for approving other members within the organisation as either an Editor or Member. Editors can edit metadata and upload and download files, while members can only view metadata and download files. The Administrator must therefore be someone well-trusted by the organisation.

    The following aspects should be considered by a wider range of stakeholders regarding the future development and refinements of the WRO:

1. A set document format to be used going forward can greatly simplify text scraping. By adhering to a consistent format, the structure and organization of the document becomes more predictable, promoting the efficiency and accuracy of text scraping.
2. A data management policy specifically outlining how data should be organised, documented, and stored throughout its life cycle including date formats, decimal delimitators, graphic formats, column headings (merging cells should be avoided and units included with heading) and more.
3. Whether or not a Digital Object Identifier (DOI) (https://www.doi.org/) should be assigned to each dataset. It may not be prohibitively expensive, but he main 'cost' of using DOIs is not the money, but the administration as the platform needs to take responsibility for ensuring the integrity of its DOI database and will need to assign a human resource to that, and they have to work meticulously (Ms Tamsyn Sherwill, personal communication 21 October 2022). It is noted that the WRO currently assigns a unique URL to each metadata record and dataset, for example, https://data.waterresearchobservatory.org/metadata-form/a-south-african-national-input-database-to-run-the-swat-model-in-a-gis-soil-map/resource/10ca242d-f54d-4b09-8ce7-f9d5bb0f127b.
4. If assigning a DOI is desired, it is noted that CKAN, the open-source data portal that hosts the WRO, does have a plugin to assist with this matter.
5. Developing a system to decide what research outputs can be placed in 'cold storage', thereby saving monetary costs in cloud storage. Currently the storage of digital objects in the cloud is also associated with an environmental impact, primarily through the requirement for electricity and resources needed to build the servers. Consider a data retention policy that outlines the minimum duration for data retention after project completion. This policy should consider legal and ethical obligations, as well as the potential value of long-term data preservation for future (links to point 3 above).
6. Currently the CKAN platform for uploading and downloading research outputs is only protected using a username and password. A two-factor authentication system may want to be introduced in the future.
7. While access to data can be strictly controlled in the CKAN environment, it is more difficult to control when someone is given access to the GCP storage bucket. A separate bucket for private data may be needed if there were many users who would like to work in the GCP platform itself.
8. A committee should be established to oversee the WRO. Such a committee's responsibilities would include:

9. Overseeing the security of the platform and conducting regular security audits;
10. Negotiating cloud storage costs with service providers and deciding on contracts to commit to;
11. Ensuring compliance and governance, including adherence to relevant data protection regulations, institutional policies, and ethical guidelines. This includes data retention periods and consent management;
12. Renewal of any contracts with service providers;
13. Directing the vision and strategy of the WRO, especially with regards to adapting technological advances;
14. Implementing periodic checks of data redundancy, replication, and integrity;
15. Regularly review and update supporting documentation and protocols to ensure effectiveness and alignment with evolving practices. This includes guides that explain the data archiving protocol, including procedures for data submission, retrieval, and data documentation.

2.3. The Hydrological Model for South Africa (HAMSA)

The **H**ydrologic**A**l **M**odel for **S**outh **A**frica (HAMSA) has now been developed as part of the WRO project (https://hamsa.hawqs.tamu.edu/#/). It is based on the Hydrologic and Water Quality System (HAWQS) developed by Texas A&M who are also providing technical support to HAMSA. Its core modelling engine is the internationally used Soil and Water Assessment Tool (SWAT) (Arnold et al., 2012, Neitsch et al., 2002). SWAT is a watershed scale and physically based model, operating on a daily time step and was developed to predict the impact of land management practices on water, sediment, and agricultural yields in large, complex watersheds over long periods of time (see Figure 1 for an example of a simulated water balance). HAWQS substantially enhances the usability of SWAT and allows users to upload, share and run simulations in the cloud, eliminating the need for extensive processing capabilities and hardware. Various scenarios can be set up and tested using a user-friendly interface, and the output can be visually compared between two scenarios (Figure 2). HAMSA could potentially be used to simulate the effects of management practices considering various crops, soils, natural vegetation types, land uses, and other scenarios for hydrology, as well as for water quality parameters such as sediment, pathogens, nutrients, biological oxygen demand, dissolved oxygen, pesticides, and water temperature.

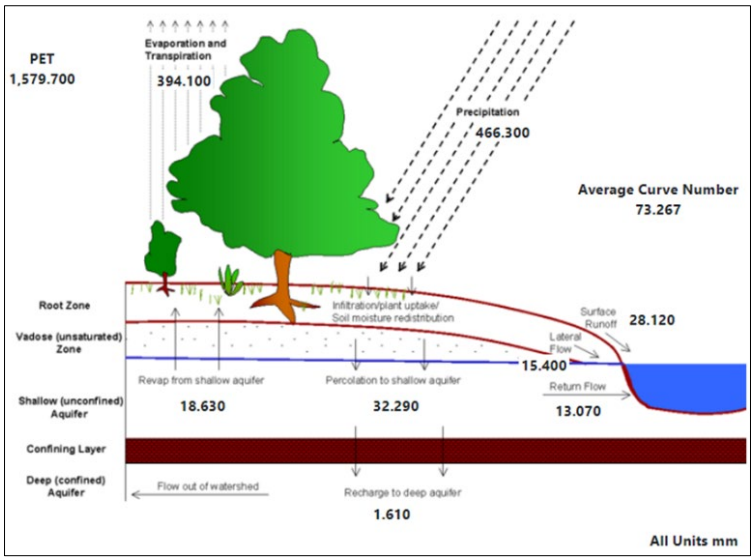


Figure 1: Schematic of an annual water balance for a hydrological response unit (HRU) simulated in the Hydrological Model for South Africa (HAMSA)

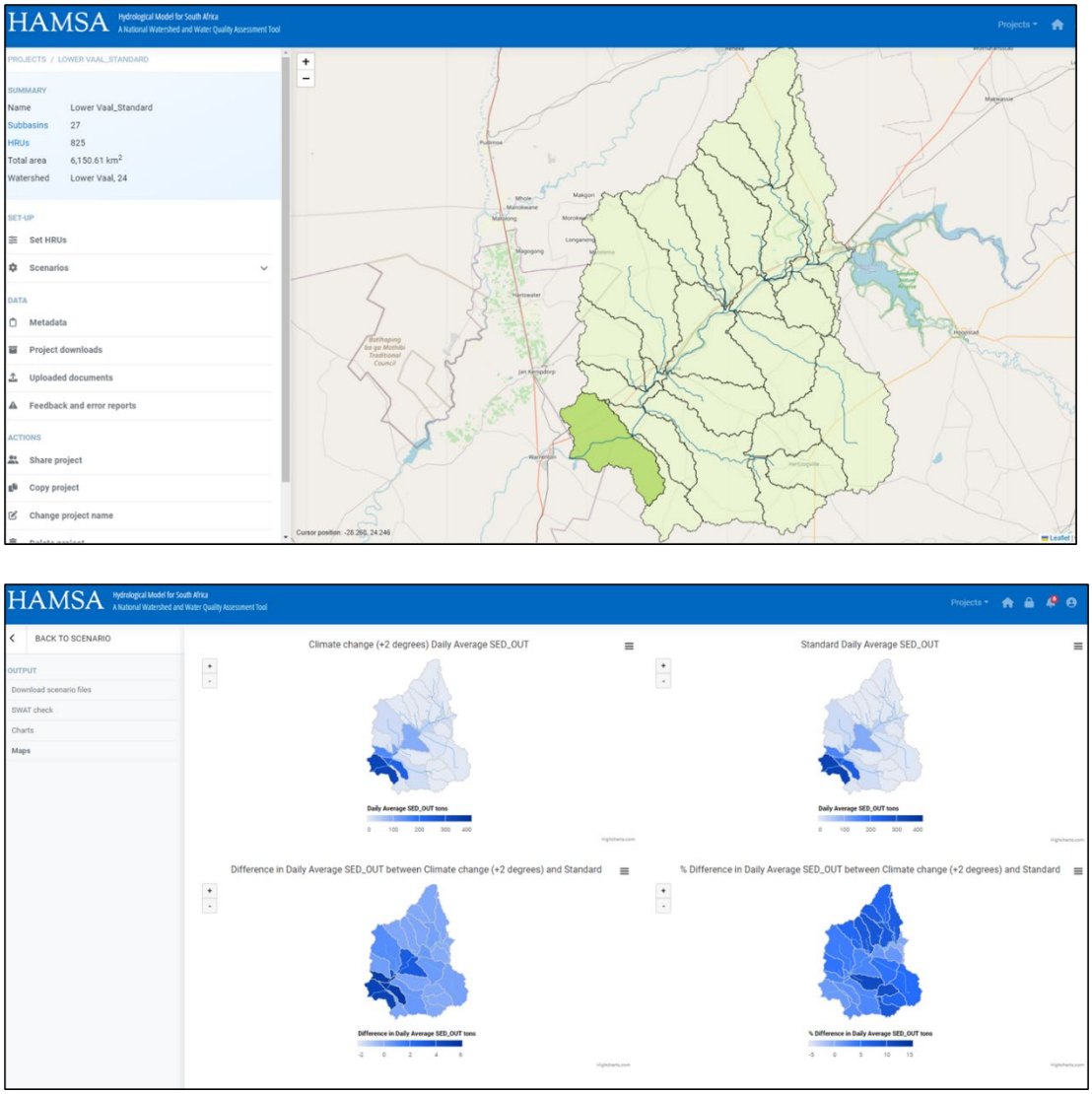


Figure 2: Interface for the Lower Vaal basin (top) and example of how different scenarios can be compared (bottom)

Regarding the future, IBM is working with Texas A&M in the application of an online HAWQS, as well as Deltares with their farmer/citizen science water monitoring app. IBM also has an ongoing collaboration with the ARC in developing a drought early warning system. An agreement has now been established to work with Texas A&M who already have an IBM Sustainability Accelerator project to combine the hydrological monitoring, water quality monitoring, and weather forecasting data streams into a centralised digital ecosystem decision support system for SA. The system is aimed at (a) water resource managers (Department of Water and Sanitation, Catchment Managements Agencies, municipalities), (b) farmers using weather and water forecasts to guide their operations, and (c) groups who would like to monitor the quality of their water supply and possibly develop early warning systems. IBM are providing support until the end of 2024 in the form of cloud infrastructure and IT support to develop the system.

2.4. The Decision Support System for Agrotechnology Transfer (DSSAT)

The decision support system for agro-technology transfer (DSSAT) is a Windows-based computer program that comprises crop simulation models for over 42 crops. The model was established by database management programmers for soil, weather, crop management and experimental data, by utilities and application programs (Hoogenboom et al., 2010). The DSSAT model is a multipurpose model that has been applied for the evaluation of crop development, such as crop phenology, biomass, and yield production (Abedinpour and Sarangi 2018). This model simulates crop growth and development, including yield, by utilizing a specific dataset that includes information on crop management, minimum weather data, and soil profile parameters using a set of independent programs which streamlines the simulation of cropping systems. The DSSAT model has been applied to investigate improved irrigation water and nitrogen fertiliser management under SA conditions (van der Laan et al., 2011), as well as numerous other studies on various crops (Jones et al., 2015, Ajilogba and Walker, 2020). It has also been used extensively in Agricultural Model Intercomparison and Improvement Project (AGMIP) studies to predict the impact of climate change on crop production, often being best or one of the best performing models during both step 1 blind and step 2 full calibration (Rosenzweig et al., 2018)

Model application results showed that the model can be accurately calibrated and applied under local conditions if suitable input and calibration data are available. A major challenge for application in SA has been related to challenges in obtaining the required input data and preparing it in the correct format to parameterise and run the model. Researchers have often made use of different data sources and approaches, adding uncertainties to results and making comparisons difficult. The evaluation of Digital Soil Maps (DSMs) available for SA (see Chapter 6 Digital soil maps for South Africa) has now identified the best sources of data for different parameters when measured data is not available, for example, in crop suitability studies. A sensitivity analysis by Thomson has also been conducted to demonstrate the effect of inaccurate parameters on model outputs. DSSAT has also been used in this project to test the predictive capabilities of the model for maize (*Zea mays* L.) in a Data Intensive Farm Management (DIFM) commercial precision agriculture trial (see Chapter 5 Big data analytics in precision agriculture). In a comparison with a ML approach to predict yield and optimise spatially variable inputs, a clear advantage of the model is being able to predict yields using forecasted weather data.

2.5. WRO continuity

Systems change over time, so we can expect the WRO to evolve with technological advances and user requirements. For example, the WRC's Waterlit Collection was a hard copy-digital catalogue research collection that lasted for 25 years (1974-1999) (Tempelhoff, 2015), before being superseded by technology that was scarcely imaginable in 1974. Therefore, some stability in the hosting arrangements is important. This WRO is now an important part of the WRC corporate strategy, and all WRC Thrusts will support the platform over the next five years. Much depends on succession planning around key IT positions.

Important decisions now need to be made on whether to migrate databases to one or more cloud-based systems or leave the systems in the current format. The rapid advance of technology and potential benefits of migrating to the cloud, include:

- IT resource (e.g. storage, processing power) scalability and being able to pay only for what is needed leading to cost efficiency,
- Advanced and automated maintenance, backup and recovery options,
- High levels of interoperability, flexibility and agility,
- Can enable higher levels of access and collaboration,
- Access to more advanced security measures provided by the service provider,
- Enhanced data analytics and machine learning capabilities,
- Wider availability of programming skills beyond in-house,
- Cloud hosting provides additional resilience to force majeure, for example, climate, conflict and disease,
- Environmental sustainability when service providers take measures to physical host servers in ways that reduce the environmental footprint.

Disadvantages, concerns, and risks in migrating to the cloud include:

- Some lack of knowledge and control on where exactly the data is stored,
- Data security and privacy concerns relating to data breaches, cyber-attacks,
- Compliance issues, for example, data sovereignty laws that inform where the data is stored geographically in the case of service providers that operate at a global scale,
- Contracting is a complex task involving IT, intellectual property, legal and other service involvement. Contraction termination data and migration options are likely the most difficult parts to negotiate.
- Cost management, including unpredictable costs and the need for careful monitoring and management that is required,
- Data transfer costs, including data downloading that exceeds a threshold or for data stored in different, such as standard access versus 'cold storage' that is cheaper to store but may be more expensive to access,
- Service provider lock-in when it is difficult to migrate to another platform. This can also lead to reduced negotiation power,
- Steep learning curve for especially older IT specialists who are familiar with legacy systems, and additional training costs may be incurred,
- Less control over cloud infrastructure for specialised applications.

The lists above are not exhaustive, and more challenges and opportunities can be expected as technology continues to evolve. It is important for organizations to carefully plan and execute database migrations to the cloud, considering factors such as data security, compliance, and the specific needs of their applications.

3. Enhancing the resolution of GRACE-assimilated groundwater storage anomalies across two aquifer types in the Steenkoppies Catchment

Cindy Viviers, Michael van der Laan, Zaheed Gaffoor and Matthys Dippenaar

3.1. Introduction

A combination of mitigation and adaptation measures are essential to overcome the increasing pressure on global water resources amidst changing climate conditions (Sen, 2015; Sadath et al., 2023). Decision making for effective and sustainable groundwater resource management necessitates continuous, accurate and timely information of aquifer conditions (De Bruin et al., 2023). Groundwater level measurements are the primary information for evaluating hydrogeologic stresses acting on aquifers (Kenda et al., 2018). In places like South Africa (SA), the distribution of hydrogeological monitoring stations is, however, inadequate or has decreased over the years, and where monitoring stations with publicly accessible data are active, the data storage infrastructure and formats are not standardised, obstructing data retrieval, sharing and integration into resource management (Gaffoor et al., 2020). The scarcity of reliable data at adequate spatial and temporal resolutions has prompted research into the viability of remotely sensed data as a potential alternative when high quality *in situ* data is unavailable (Milewski et al., 2019; Joseph et al., 2020).

The Gravity Recovery And Climate Experiment (GRACE) mission presented the first opportunity to directly measure water storage changes from space (https://grace.jpl.nasa.gov/applications /groundwater/ 2015). The distance between the twin satellites are measured meticulously, and fluctuations are attributed to the variations in the Earth's gravitational field, specifically reflecting mass redistribution in the hydrosphere (Swenson and Wahr, 2002). When the influence of mass transports from the ocean, atmosphere and Earth's interior are excluded, the temporal gravity field variations over land are primarily ascribed to Terrestrial Water Storage (TWS) change (Feng et al., 2013; Yan et al., 2022). Terrestrial water storage consists of all water on the land surface and subsurface combined (Yeh et al., 2006). To isolate a single component from the TWS, such as groundwater storage (GWS), the other estimated or measured components are subtracted from the total TWS (Feng et al., 2013).

Multiple studies have relied on the 0.25° Global Land Data Assimilation System (GLDAS) products to isolate the groundwater component from TWS (Alghafli et al., 2023; Strassberg et al., 2007; Feng et al., 2013; Liesch and Ohmer, 2016; Gaffoor et al., 2022; Ramjeawon et al., 2022). In early 2020, the GLDAS Version 2.2 (GLDAS-2.2) data products, assimilating GRACE data from the Centre for Space Research (CSR) into the Catchment Land Surface Model (CLSM) within the NASA Land Information System (LIS), were released. The method involves incorporating GRACE TWS observations into the model driven by meteorological data using a Kalman smoother approach to update and enhance the accuracy of forecasted model states (Li et al., 2019; Rui et al., 2022). Li et

al. (2019) assessed how effectively GRACE Data Assimilation (GRACE-DA) enhances groundwater simulation across different regions worldwide. Of the African regions analysed in greater depth, Uganda serves as a proximate case study as the authors do not address SA specifically. For Uganda, the authors found that GRACE-DA effectively reduced the dynamic range of CLSM-simulated GWS and improved interannual variability. This resulted in notable root mean square error (RMSE) and correlation improvements when compared to *in situ* observations, affirming that GRACE-DA significantly improved the accuracy of the GWS product, particularly in areas with substantial recharge. Overall, on a global scale, Li et al. (2019) observed that GRACE-DA improved simulation of GWS in GLDAS based on data from nearly 4 000 boreholes. Nevertheless, to be useful for local scale applications, coarse products often require downscaling (Gemitzi et al., 2021).

The primary objective of downscaling is to enhance the spatial resolution of a coarse dataset by integrating high-resolution information obtained from other sources. This can be achieved using either a dynamical or a statistical approach. Dynamical downscaling relies on assimilating boundary conditions from regional or global models and combining data from multiple sources to construct a process-based physical model at higher grid resolutions. Statistical downscaling is the process of analysing and establishing the empirical relationship between the coarse-resolution, regional variables and the corresponding more detailed, fine-resolution, local variables observed simultaneously in historical data. High-resolution, monthly time-series GWS estimates comparable to borehole observations could have significant benefits for groundwater resource assessments and management in SA.

Statistical methods have been applied extensively in previous studies. For example, using the partial least squares regression statistical method and assimilating 0.5° water storage datasets from the WaterGAP hydrology model (WGHM) with precipitation, ET and runoff from other models, Vishwakarma et al. (2021) refined the GRACE TWS resolution from 3° to a 0.5°. The study demonstrated that the WGHM accurately redistributed water mass spatially while preserving mass conservation principles and signal amplitude. Yin et al. (2018) presented the correlative statistical downscaling relation method to downscale GRACE GWS anomalies (GWSA) from 110 to 2 km using only high-resolution Moderate Resolution Imaging Spectroradiometer (MODIS) ET data. The authors suggest the downscaling method may be applied for local water resource planning in areas where there is a strong relationship between GRACE TWS and ET. Given the importance of precipitation in TWS changes, Gemitzi et al. (2021) applied a spatial statistical downscaling method to downscale GRACE TWS from 1.0° to 0.1° relying solely on the Integrated Multi-satellite Retrievals for Global Precipitation Measurement precipitation dataset. The study, which considered temporal lags between precipitation and TWSA, found that the downscaled data closely matched independently modelled TWSA from 2005 to 2015, exhibiting strong performance across all evaluation metrics.

Several studies have applied machine learning (ML) algorithms to downscale GRACE data. Rahaman et al. (2019) and Ali et al. (2021) developed random forest (RF) models with multiple hydrological input variables [precipitation, evapotranspiration (ET), temperature, soil moisture storage, topography, surface runoff, plant canopy water storage and others] to downscale GRACE GWSA from 1° to 0.25°. Sabzehee et al. (2023) applied a RF model to forecast GRACE-derived GWSA and enhance the spatial resolution from 0.25° to 0.1° resolution, using

precipitation, ET, land surface temperature (LST) and normalized difference vegetation index (NDVI) as predictors. The respective studies concluded that the RF models effectively increased the resolution of GRACE data and generated estimates that accurately capture groundwater level storage anomalies. None of these studies, however, investigated how incorporating temporal lags in the training process could minimise residuals.

Gaffoor et al. (2022) included local groundwater data, hydroclimatic parameters, and land-surface characteristics to train Gradient Boosting Decision Tree (GBDT) models. These models predicted groundwater level changes at a 10 km resolution for the karstic aquifers of the Ramotswa/Northwest/Gauteng Dolomites region, which encloses the Steenkoppies Dolomitic Compartment (SDC) used as a case study in this investigation. To assess the impact of hydroclimatic variables over different timeframes, the authors incorporated arbitrary 30, 60 and 90-day increments as well as 30-60, 60-90, and 30-90 days intervals. Although the study did not specify the optimal temporal lag, it highlighted the importance of a 30-day gap for LST and GRACE-derived GWSA for model predictions.

The study used the SDC, SA, as the study area to investigate the hypothesis that GLDAS-2.2 (GRACE-DA) GWSA can be downscaled to a 0.05° resolution using only the freely available Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) rainfall and MODIS ET products. The study also proposed that incorporating the computed temporal lags between the input variables to the GWSA would improve the model performance by minimising residuals. Furthermore, this study explored the potential association between the accuracy of the downscaled product and groundwater level behaviour with karst and intergranular and/or fractured aquifers, respectively. Effective statistical downscaling of GRACE-DA GWS can offer an innovative solution for understanding aquifer conditions when spatial or temporal data distribution is insufficient.

3.2. Materials and Methods

Study area

The SDC is situated approximately 75 km south-west from the SA capital city, Pretoria, primarily within the boundaries of quaternary catchment A21F (Figure 3). Intensive irrigated agriculture has more than doubled across the SDC in the last 20 years (Vahrmeijer et al., 2013). The dominant crops cultivated throughout the year under pivot or sprinkler irrigation across the SDC are beetroot (Beta vulgaris), carrots (Daucus carota), lettuce (Lactuca sativa), cabbage and broccoli (Brassica oleracea), as well as maize (Zea mays) in the summer and wheat (Triticum aestivum) in the winter (Le Roux et al., 2016). With an estimated surface area of 213 km², the SDC is a highly valuable groundwater resource for SA, in terms of economic significance as well as employment (Holland and Wiegmans, 2009; Wiegmans et al., 2013). Since the spring flow decreased to 5.49 Mm^3a^{-1} in 2008, significantly lower than the long-term average flow of 14.38 Mm^3a^{-1} (Meyer, 2014), there has been an ongoing debate on whether decrease in GWS in the SDC is due to excessive groundwater abstraction or reduced precipitation (Holland and Wiegmans, 2009)

Based on the monthly data from the local automatic weather station (AWS) [Deodar (30619), -26.1427°S; 27.5743°E] (Figure 3), with data from 2004 to 2021, the average temperatures range from a minimum of 15.9°C up to a maximum of 27°C in the summer, and the average winter temperatures range from a minimum of 2.2°C up to a maximum of 18.5°C. Roughly 80% of the mean annual precipitation (MAP) of 677 mm is received between November and March.

Geology and hydrogeology

The regional geology is primarily made up of the Transvaal Sequence, which includes a major karst aquifer formed within the dolomite rich Malmani Subgroup, along with the sequential intergranular and/or fractured aquifer formed in the eight Pretoria Group formations, namely the Rooihoogte Formation to the Magaliesberg Formation. The Pretoria Group formations comprise of alternating layers of mudrock and quarzitic sandstone units (Keyser, 1986; Eriksson et al., 2009; Moore et al., 2001). The region has been significantly deformed by post-Transvaal diabase and large-scale tectonic events.

The absence of surface water across the SDC suggests extensive recharge, where the surface water that would typically flow across the landscape is intercepted and infiltrates the dolomite aquifer characterised by numerous sinkholes (Wiegmans et al., 2013). Depending on specific geological conditions, the dividing dykes can form isolated hydrological compartments with a flat groundwater level table and relatively uniform groundwater conditions, with less significant abstraction and flow impact between compartments in comparison to within a compartment. Dolomite compartment boundaries are typically associated with spring lines and seepages as the groundwater is forced to the surface (Holland and Wiegmans, 2009; Bredenkamp et al., 1986; Kuhn, 1986).

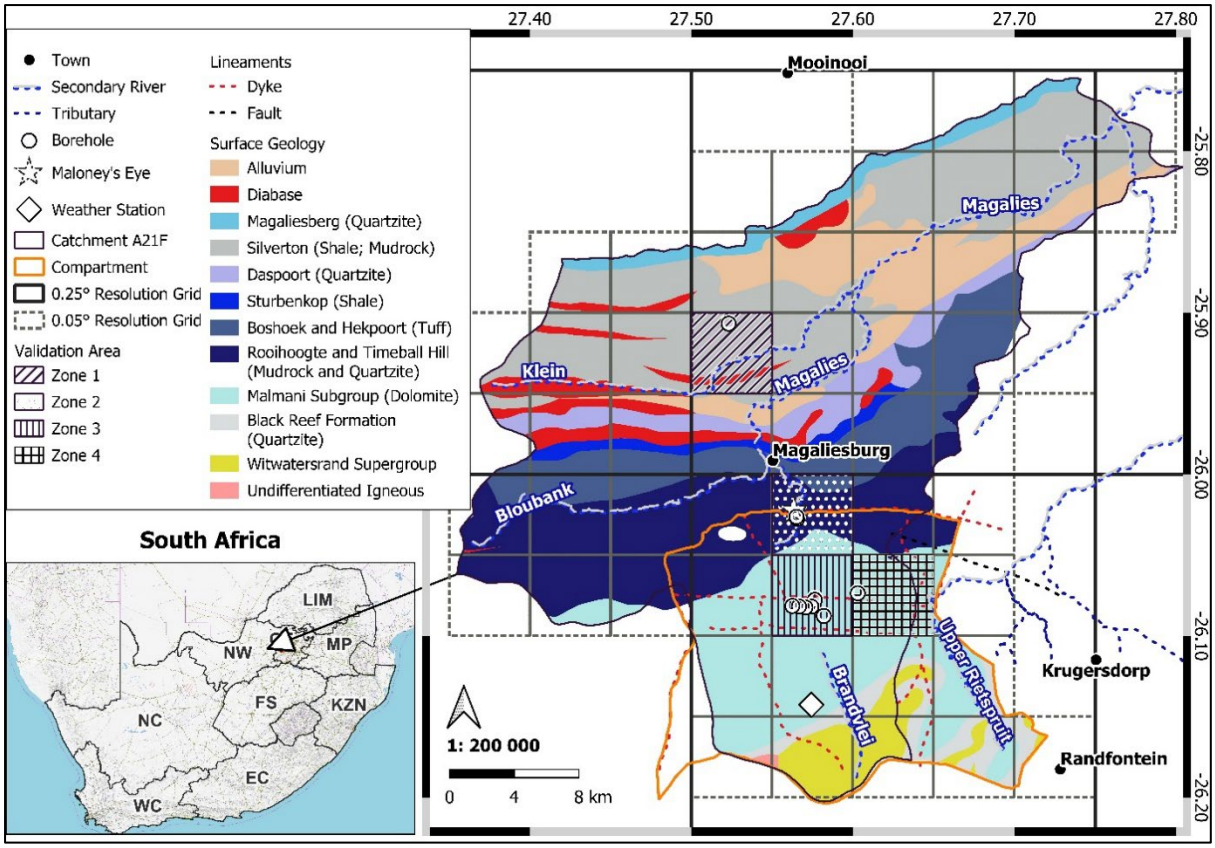


Figure 3: Location and surface geology of the Steenkoppies quaternary catchment and dolomite compartment, as well as the distribution of the downscaling validation in situ data in relation to the 0.25° Global Land Data Assimilation System (GLDAS)-2.2 Groundwater Storage (GWS) and the 0.05° downscale target resolution.

Data and pre-processing

GLDAS-2.2 groundwater storage data

Vishwakarma et al. (2018) determined that a minimum catchment size of 63 000 km² is observable using filtered GRACE fields and more advanced techniques are necessary to detect mass variations at finer resolutions. The CSR mascon solution relies solely on GRACE data, independent from TWS and external models. The solutions are computed using a custom hexagonal grid (approximately 120 km or 1° at the equator), which offers improved spatial resolution over traditional spherical harmonics (Chen et al., 2017; Save et al., 2016).

Dynamic downscaling assimilates coarse-resolution data to develop a model applicable to finer resolution data (Rahaman et al., 2019). The GLDAS-2.2 is specifically engineered to integrate various observational data products to enhance its land surface models. Data Assimilation has been applied to combine remotely sensed GRACE TWS data with the fine 2.5 km grid CLSM in Land Information System (LIS) Version 7 estimates driven by meteorological inputs, land cover, soil, topography, and other model-specific parameters to estimate fluctuations in water content at a 0.25° spatial resolution.

The GLDAS-2.2 product provides disaggregated and continuous water storage estimates, which include GWS computed as TWS minus the root zone soil moisture to a depth of 1 m, the snow water equivalent, and any canopy interception (Rui et al., 2022). Unlike the GRACE TWS data, the GLDAS-2.2 TWS are not provided as anomalies. To make the GRACE TWSA measurements comparable to GLDAS-2.2 TWS, the 2003-2015 simulated TWS open-loop run (averaged to each 0.5° GRACE grid) temporal mean was added to the GRACE TWSA prior to assimilating GRACE data (Li et al., 2019). Hence to convert the GLDAS-2.2 data back to anomalies comparable to GRACE measurements, the 2003-2015 temporal mean must be subtracted. The GLDAS-2.2 (NASA_GLDAS_V022_CLSM_G025_DA1D) data extending from March 2003 to December 2021 was accessed and processed through Google Earth Engine (GEE).

Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) precipitation data

The satellite-gauge gridded precipitation product, CHIRPS, extending from 50°S to 50°N (and all longitudes) and dating from 1981 to near-present, is obtained from an algorithm combining precipitation datasets from different sources into a single product (Funk et al., 2015). Released at a daily temporal resolution, the monthly precipitation was calculated using the 0.05° CHIRPS precipitation (UCSB-CHG_CHIRPS_DAILY) product. Because the GRACE data are provided as anomalies, specifically deviations relative to a 2004-2009 baseline temporal average, and the 2003-2015 GLDAS-2.2 temporal mean was subtracted to make the GLDAS-2.2 measurements comparable to GRACE, the monthly CHIRPS estimates were standardised to anomalies by computing the 2004-2009 mean and then subtracting that mean.

Moderate Resolution Imaging Spectroradiometer (MODIS) evapotranspiration

The MODIS ET data product is estimated using a modified (Mu et al., 2011) Penman-Monteith equation (Monteith, 1965) that improves the accuracy and spatial resolution of ET estimates by integrating daily meteorological reanalysis data, as well as eight-day MODIS remotely sensed data products that capture the effect of land cover and vegetation dynamics on ET. The MOD16 ET is the sum of canopy transpiration, soil evaporation and interception evaporation (Mu et al., 2007 and Mu et al., 2011). Accessed and processed through GEE, the 0.005° spatial resolution and temporal resolution ET data were sourced from the open-source MOD16A2 v6 (MODIS_006_MOD16A2) data product (Running et al., 2021). Prior to quantifying the monthly mean ET from the eight-day composite, the data quality was filtered for a 'clear' cloud state and a quality control confidence state of 'best result possible with no saturation'. To standardise the ET data to anomalies comparable to the GLDAS-2.2 GWSA, the 2004-2009 temporal mean was calculated and then subtracted.

Input variable selection and correlation analysis

High precipitation coupled with lower ET generally results in increased water availability for infiltration and percolation into the groundwater (Cao et al., 2013; Moiwo et al., 2011). In this study, precipitation was selected because aquifer recharge is primarily from precipitation, and the relationship with groundwater level fluctuations and discharge from the Maloney's Eye spring is well documented (Wiegmans et al., 2013; Vahrmeijer et al., 2013).

Evapotranspiration was selected because it is considered a key process within the hydrological cycle which reflects varying factors such as weather conditions, landscape and topography, vegetation and land use cover, and soil properties (Mu et al., 2007).

The impact of precipitation and ET is not immediately reflected in GWS and lags due to factors such as the unsaturated zone thickness, and the transmissivity and storage characteristics of the underlying lithology (Kotchoni et al., 2019). To evaluate whether precipitation and ET can be applied as independent variables for downscaling, the temporal correlation coefficient (r) was quantified per 0.05° pixel using cross-correlation. The *r* was calculated by shifting the precipitation and the ET values with a range of time lags spanning from one to twelve months prior to calculating the *r* against the reference GWS at every lag. The lag time periods that produced the highest aggregated mean *r* across the study area was selected and integrated into the downscaling model.

Random Forest Model and downscaling model design

Studies have shown that RF models outperform support vector regression (SVR), artificial neural network (ANN) and multivariate linear regression (MLR) ML models in downscaling GRACE data to a higher spatial resolution (Chen et al., 2019, Ali et al., 2021 and Sabzehee et al., 2023). A decision tree is trained by branching out into increasingly homogeneous subsets from the root node until reaching leaves that are nodes without further sub-division. The root node begins with all the training data, repeatedly splitting the data according to the independent, predictor variables. At every node, the algorithm calculates the prediction improvement for each variable's split point, choosing the split which results in the highest improvement. To evaluate the 'improvement' made per threshold (node condition), the model calculates the sum of the squared residuals between the node and its child nodes after the split (Hoare, 2023). The RF supervised learning algorithm is a non-linear statistical learning method which deals with multiple decision trees combinations and is used to overcome the bias and overfitting problem associated with single decision trees (Shelestov et al., 2017). The RF approach uses the Bootstrap Aggregation ensemble technique to create many individual, uncorrelated trees all trained on different parts (random subset) of the same training set (bagging). The average prediction for the ensemble of trees, is the output (step known as aggregation). The effectiveness of the RF algorithm lies in its principle of leveraging the 'collective wisdom of a committee of predictions', which is more accurate than any individual tree (Yiu, 2019; Srivastava et al., 2023).

In this study, monthly GLDAS GWSA data extending over 18 years from 2003 to 2021 were downscaled from a spatial resolution of 0.25° to a higher spatial resolution of 0.05° using a RF approach, as illustrated in Figure 4.

The respective independent variables were first aggregated to 0.25° in line with the target GLDAS GWSA variable using pixel averaging. Two respective RF models were trained and tested using these data in 'Model 1', which had no temporal lag implemented for the input variables (precipitation and ET), versus 'Model 2' for which the input variables were adjusted with the optimised temporal lags as described in Section 2.4. The respective models were trained on 0.25° data spanning from January 2003 to December 2016 (thirteen years or 72% of the available time-series). To evaluate the model's prediction performance, the respective trained RF models were applied to predict

0.25° GWSA for the separate testing dataset extending from January 2017 to December 2021 (five years or 28%). To optimise the respective models during hyperparameterisation, two parameters were adjusted, namely the number of forest trees in increments of 10 (10 to 150) and the bag fraction in increments of 0.1 (0.1 to 0.9). The parameters were set to what combination resulted in the lowest RMSE when comparing the new 0.25° predicted GWSA against the 0.25° original GWSA before any residual correction, thus minimizing residuals.

After predicting the 0.25° resolution GWSA, pixel-wise residuals were computed by subtracting the new predicted GWSA from the original GWSA, essentially the low-resolution model error. Subsequent to model training and residual calculation, the best performing RF model was applied to predict high-resolution GWSA estimates using the 0.05° high-resolution input variables. Following the established approach (Chen et al., 2019; Ali et al., 2021 and Sabzehee et al., 2023) for downscaling using the RF algorithm, residual correction was applied to the new high-resolution GWSA estimates. This included resampling (nearest neighbour) the coarse residuals before adding the residuals to the predicted 0.05° GWSA, resulting in monthly GWSA estimates at a finer resolution. For each respective zone, the 0.05° downscaled GWSAs were then evaluated against the *in situ* groundwater level storage anomalies.

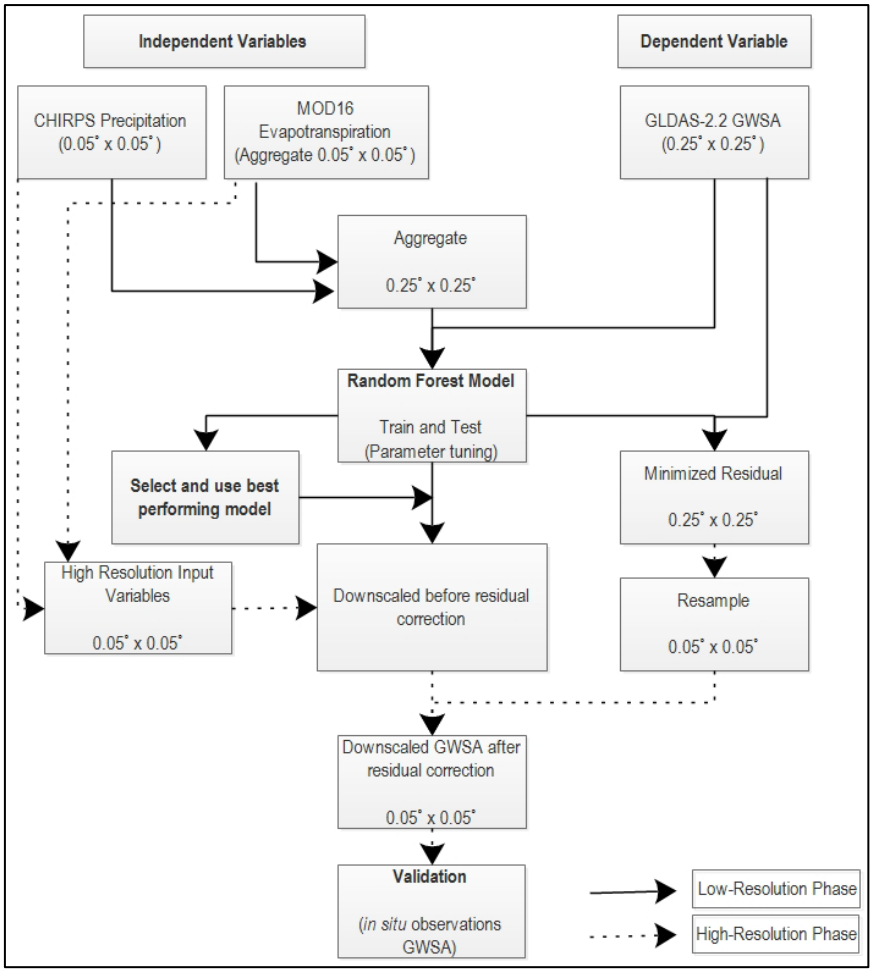


Figure 4: Flowchart of the downscaling model design in this study (CHIRPS - Climate Hazards Group InfraRed Precipitation with Station, MODIS - Moderate Resolution Imaging Spectroradiometer, GLDAS-2.2 Global Land Data Assimilation System (GLDAS) Version 2.2, GWSA - Groundwater Storage Anomaly).

Data validation and error analysis

To conduct a quantitative evaluation of the time-series datasets, the metrics used include the *r* (Equation 3.1), RMSE (Equation 3.2) and mean absolute error (MAE) (Equation 3.3). The *r* was calculated by determining the covariance and dividing the covariance by the product of the variables' standard deviations. The *r* can vary between -1 and 1, where 1 signifies a model that perfectly captures the synchronised rise and fall of the two variables. The RMSE and MAE metrics gauge how close the model predictions are to the actual values. A value near 0 indicates a highly accurate model.

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\ (Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\ \sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \qquad (3.1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i-Y_i)^2}{n}} \qquad (3.2)$$

$$MAE = \frac{1}{n} * \left(\sum_{i=1}^{n} |\, Y_i - X_i \,|\right) \qquad (3.3)$$

Where $X_i \; and \; Y_i$ indicate the input variable and the predicted variable, respectively, and $\bar{X} \; and \; \bar{Y}$ indicate the respective means of the dataset with 'n' representing the sample count.

***In situ* groundwater level monitoring data**

Groundwater level monitoring data suitable for comparison with remotely sensed GWSA estimates should be sensitive to atmospheric conditions, accurately reflect seasonal fluctuations, and the record should span across at least five years to ensure strong statistical analyses (Rodell et al., 2007, Li et al., 2019). Groundwater level monitoring data across the study area were sourced from the Hydstra Data Management platform (https://www.dws.gov.za/Groundwater/data.aspx, received March 2022), which includes data from the DWS Groundwater Monitoring Network.

During data pre-processing, the records received were limited to within the specified study time frame of 2002 to 2021, resulting in approximately 1 750 observations distributed over thirteen boreholes. Two boreholes were excluded for having ten or less observations over the 2004 to 2009 timeframe. Outliers were identified using the interquartile range method (IQR) and visualisation, after which outliers were either retained or removed. The result was 1 480 observations distributed over 11 boreholes (Figure 3).

To quantitatively compare *in situ* measurements to downscaled GWSA, the temporal 2004-2009 groundwater level mean ($GWL_{2004-2009\,mean}$), recorded as depth to groundwater level (in metres) from the topographic surface per monitoring borehole, was first calculated before being subtracted from the respective measurements (GWL_i) to create a time-series anomaly of above (positive) or below (negative) mean depth to groundwater level (Equation 3.4). The respective groundwater level anomaly measurements ($GWL_{Anomaly}$) were then converted from m to mm (unit comparable to GLDAS GWSA) and multiplied by the aquifer specific yield (S_y) to change the time-series groundwater level anomalies to storage anomalies ($GWS_{Anomaly}$) (Equation 3.5).

$$GWL_{Anomaly} = GWL_i - \; GWL_{2004-2009\,mean} \tag{3.4}$$

$$GWS_{Anomaly} = (GWL_{Anomaly} \times 1000) \times S_y \tag{3.5}$$

Utilising a 0.05° resolution grid, the observation boreholes were distributed over four distinct verification zones as visualised in Figure 3. Proceeding from north to south, Zone 1 corresponds with the intergranular and fractured Pretoria Group Aquifer and Zone 2 with two boreholes situated near Maloney's Eye. Zone 3 and Zone 4, which encompassed a cluster of seven boreholes and a solitary borehole, respectively, were positioned over the SDC.

In contrast to the *r* metric, both RMSE and MAE are considerably dependent on the S_y value. To ensure a representative and accurate comparison between the downscaled GWSA estimates and the *in situ* observation derived GWSA, the authors methodically adjusted the S_y values at intervals of 0.01, spanning from 0.01 to 0.1. The optimal $S_y$value which resulted in the lowest RMSE and MAE across the SDC and the intergranular and/or fractured aquifer, respectively, was selected.

3.3. Results and discussions

Temporal correlation analyses

The variation of the aggregated mean monthly precipitation (CHIRPS), MODIS ET and GWSA across quaternary catchment A21F from 2017 to 2021 is displayed in Figure 5a. The highest *r* was achieved when accounting for a three-month precipitation lag (0.61) and a two-month ET lag (0.65) until the response in GWSA was reflected. Using a harmonic analysis, the average seasonal variability of precipitation, ET and GWSA were characterised (Figure 5b), and concluded that the monthly precipitation peak in January, as is anticipated for a summer rainfall region, aligned with ET peaking in January and February. The seasonal GWSA trends indicated aquifer recovery and an increase in GWS starting from the lowest point in October reaching a maximum peak in April. However, subsequent to this peak, the aquifer storage decreased as outflow surpassed inflow when crop irrigation intensified during the dry autumn and winter months (May to August).

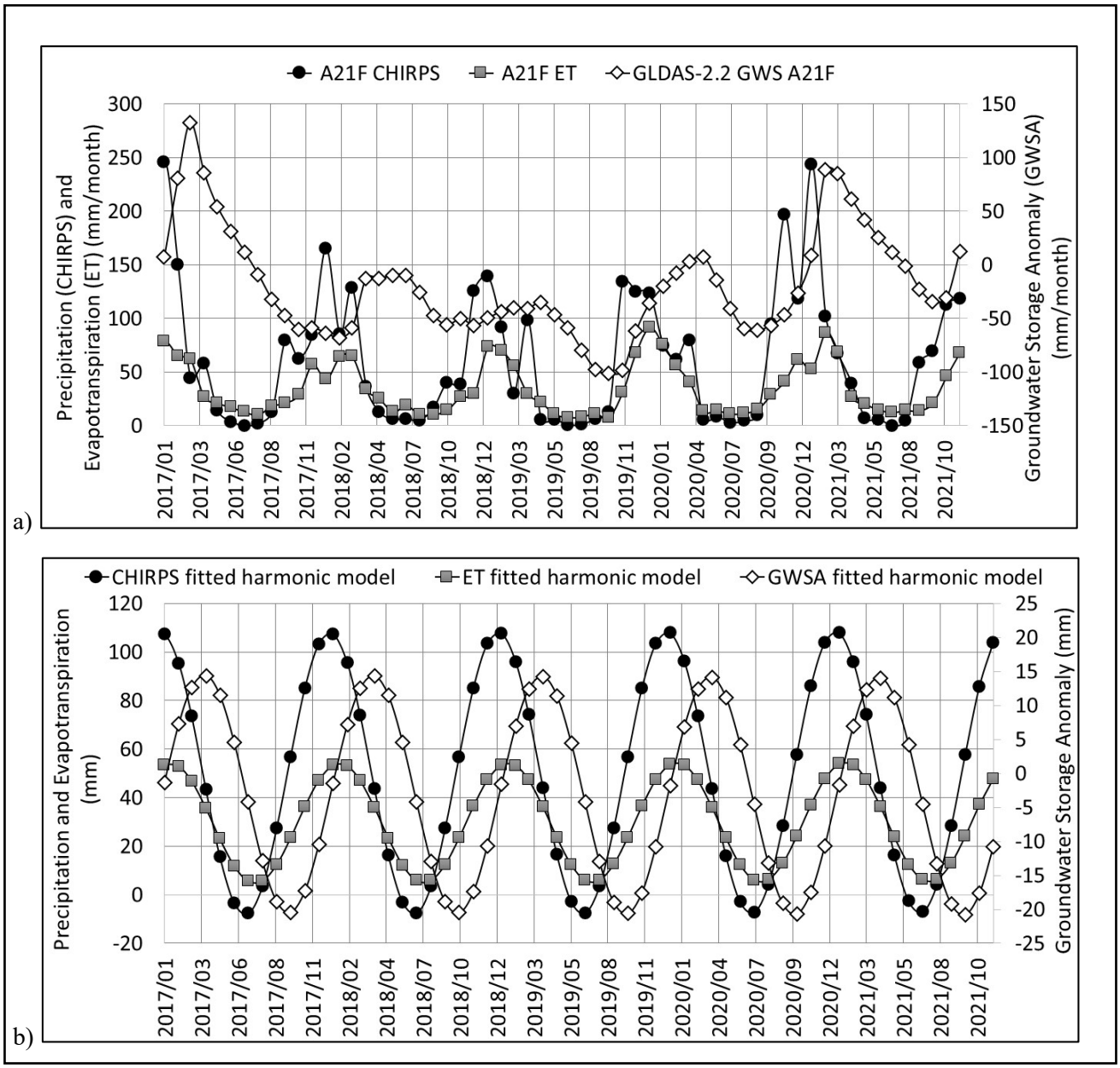
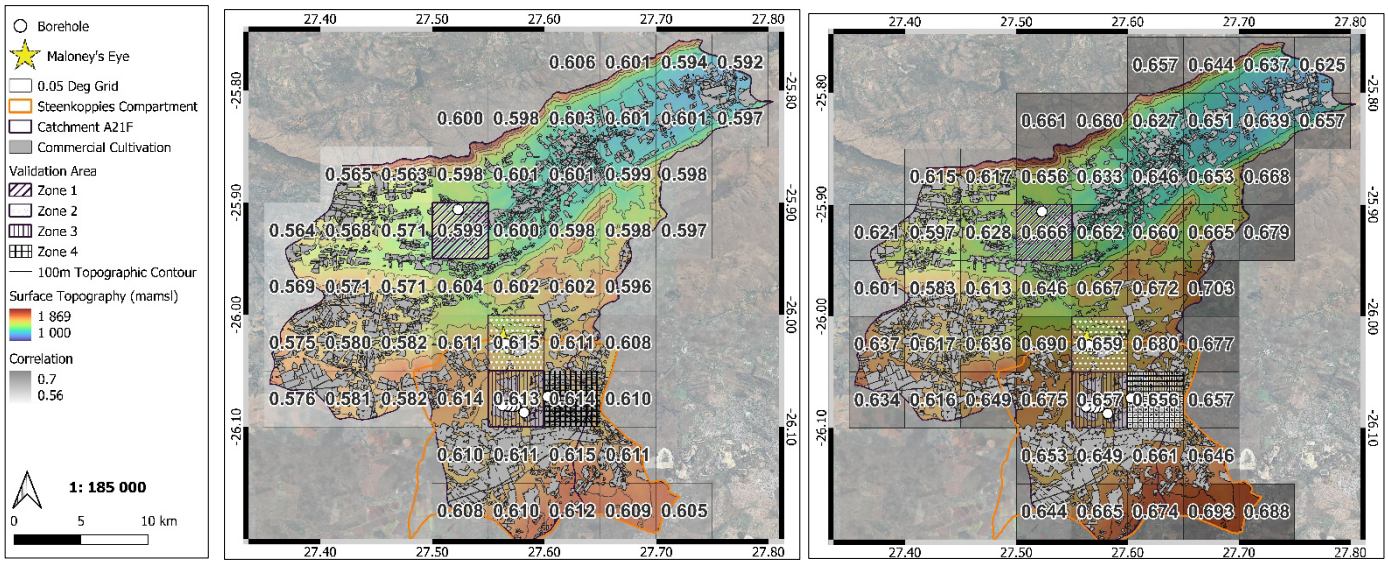


Figure 5: Temporal variation of monthly Climate Hazards Group InfraRed Precipitation With Station Data (CHIRPS) precipitation, Moderate Resolution Imaging Spectroradiometer (MODIS) evapotranspiration (ET) and Global Land Data Assimilation System (GLDAS-2.2) Groundwater Storage Anomaly (GWSA), using the spatial aggregated average across quaternary catchment A21F: (a) Time-series monthly changes (b) and fitted harmonic model to characterise seasonal variability.

The spatial distribution of the GWSA and lagged precipitation *r* ranged from 0.56 to 0.62, as illustrated in Figure 6a, whereas the *r* between GWSA and lagged ET varied between 0.58 and 0.7 (Figure 6b). Statistical analysis confirmed the significance of these correlations at the 1% level, with pixel-wise p-values well below 0.01. This suggests that the observed correlation is highly unlikely to have occurred by chance.

An insignificant *r* between the GWSA and precipitation would indicate that GWS does not respond and recover following precipitation events, and that the aquifer is being unsustainably exploited, or that land use, such as irrigation, is resulting in GWS recovery in the absence of a precipitation event. A stronger *r* in some parts can be

ascribed to the spatial heterogeneity of factors that affect groundwater recharge (Rukundo and Doğan, 2019). The *r* is somewhat stronger across the SDC distinguished by numerous sinkholes and a flat undulating plain, and weaker towards the western section of the study area characterised by steeper slopes.



a) b)

Figure 6: Spatial distribution of the temporal correlation between the Global Land Data Assimilation System (GLDAS-2.2) Groundwater Storage Anomaly (GWSA) to the a) three months lagged Climate Hazards Group InfraRed Precipitation With Station Data (CHIRPS) precipitation estimates and b) two months lagged Moderate Resolution Imaging Spectroradiometer (MODIS) evapotranspiration (ET) estimates at a monthly temporal and 0.05° spatial resolution.

Random Forest model performance analysis

Machine learning model performance metrics (Table 2) were computed by comparing the 0.25° predicted GWSA to the original GLDAS GWSA for the separate testing dataset before any residual correction, as illustrated in Figure 7. The initial comparison, before any residual adjustments, demonstrates the inherent predictive capabilities of each model based on parameters and the training data (Jain et al., 2023). The model performance metrics demonstrate that Model 2 outperformed Model 1 as is evident from the higher r, combined with the lower RMSE and MAE values. Figure 7 demonstrates that Model 2 has lower residuals and fewer outliers compared to Model 1.

Residual correction applied to the predictions addresses low-resolution model inaccuracies and corrects for the inaccuracies in the high-resolution product. Residual correction ensures that the downscaled GWSA estimates align with the original data, adhering to the conservation of mass principle while mitigating any prediction biases.

Table 2: Model performance metrics calculated for Models 1 and 2 before residual correction.

	Model 1	Model 2
r	0.35	0.6
RMSE (mm)	48	43
MAE (mm)	40	36

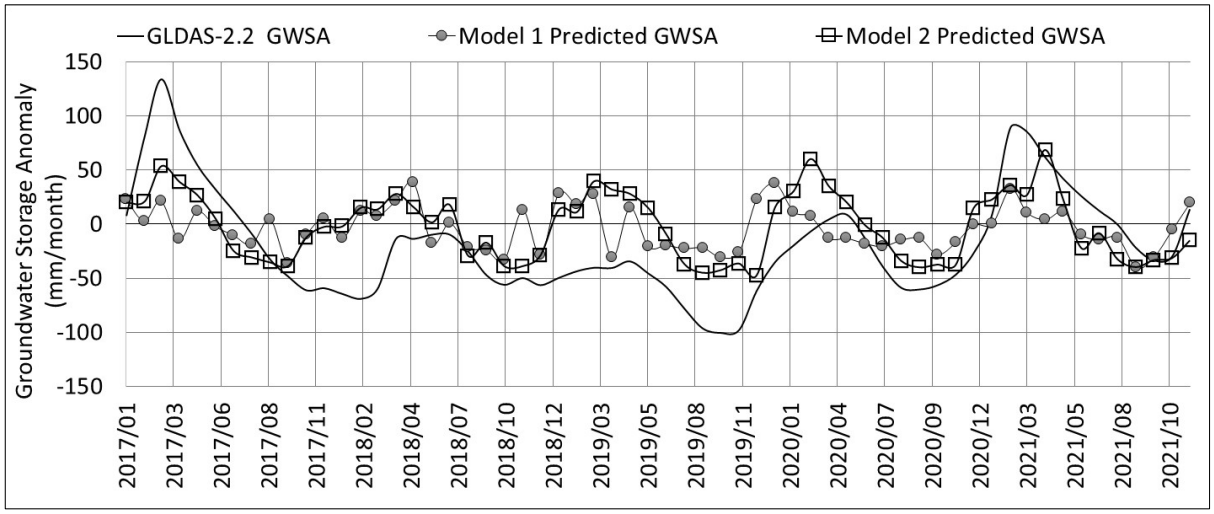


Figure 7: Performance assessment of the respective machine learning (ML) models for predicting 0.25° Groundwater Storage Anomaly (GWSA) using Model 1 input variables without temporal lags, and Model 2 with input variables adjusted with the respective temporal lags.

Mass conservation and validation of downscaled GWSA data

Downscaling aims to redistribute groundwater mass using additional, high-resolution information while upholding the principle of mass conservation. Vishwakarma et al. (2021) validated the efficiency of downscaled outputs by assessing mass conservation at catchment scale. To confirm mass conservation at catchment scale and the ability of the downscaled product to provide mass change estimates at higher spatial resolution, the authors compared the monthly total of the coarse-resolution GLDAS-2.2 GWSA across the catchment with the monthly total of downscaled GWSA estimates (Figure 8). The assessment yielded values of 0.99 for r, 205 mm for RMSE, and 161 mm for MAE.

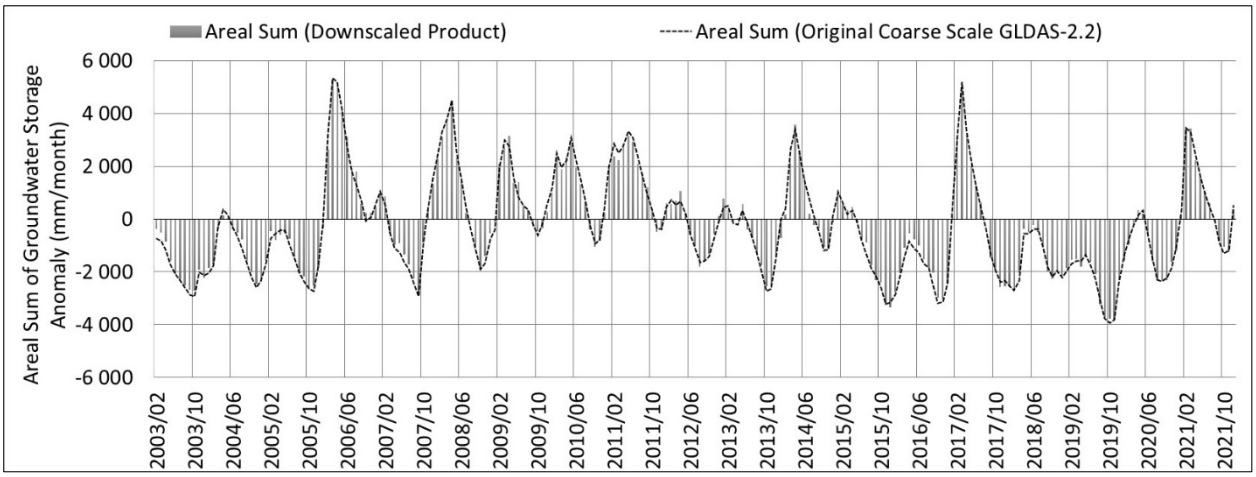
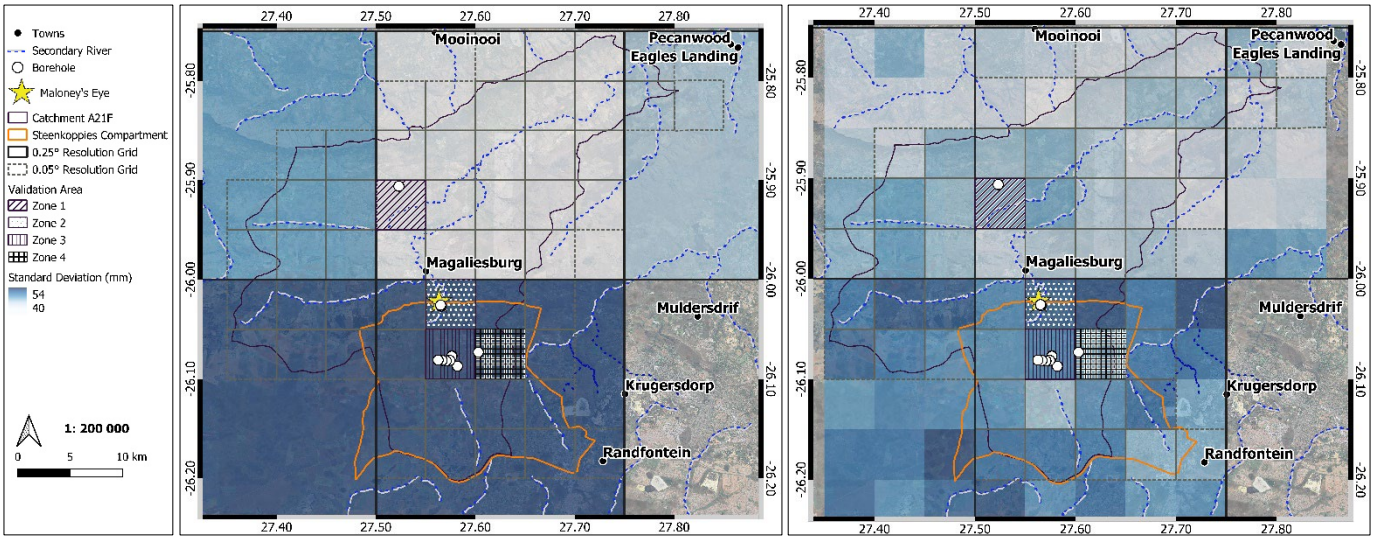


Figure 8: Monthly sum of the groundwater storage anomalies across catchment A21F for the downscaled product and the original coarse-resolution Global Land Data Assimilation System Version 2.2 (GLDAS-2.2) product.

GRACE data characteristically have a low-resolution with strong spatial correlation between neighbouring GRACE grids (Seyoum et al., 2019). The long-term standard deviation (2003-2021) GWSA before (Figure 9a) and after (Figure 9b) downscaling was compared to assess the similarity and whether the downscaled data preserves the spatial distribution of the original data. The similarity of the spatial distribution between the coarse and finer resolution datasets are apparent in Figure 9. The greater range of downscaled GWSA values indicate that the high-resolution output effectively integrated the input variables precipitation and ET. It is therefore hypothesised that the downscaled product captured spatial differences and details, such as temperature, slope, and land cover combined into the ET variable as well, to provide a more comprehensive reflection of groundwater dynamics.



a) b)

Figure 9: The geographic distribution of the Groundwater Storage Anomaly (GWSA) standard deviation (2003-2021) before (a) and after (b) downscaling.

The hydraulic head map illustrated in Figure 10 was generated using the most recent groundwater level data available for boreholes across the study area, dating from 1908 to 2021. Analysis of the depth to groundwater level data since 2000 indicate significantly deeper groundwater levels in the karst aquifer (averaging around 64 m) compared to the northern intergranular and/or fractured aquifers, where the average depth is approximately 9 m. Figure 10 also include the S_y values which resulted in the lowest RMSE and MAE, specifically 0.02 across the intergranular and/or fractured aquifer and 0.04 for the SDC boreholes. These S_y values correspond to the average S_y of 0.01 for the intergranular and/or fractured aquifer and the average of 0.035 for the dolomitic aquifer, which were used by Wiegmans et al. (2013) to calibrate their numerical model for the study area.

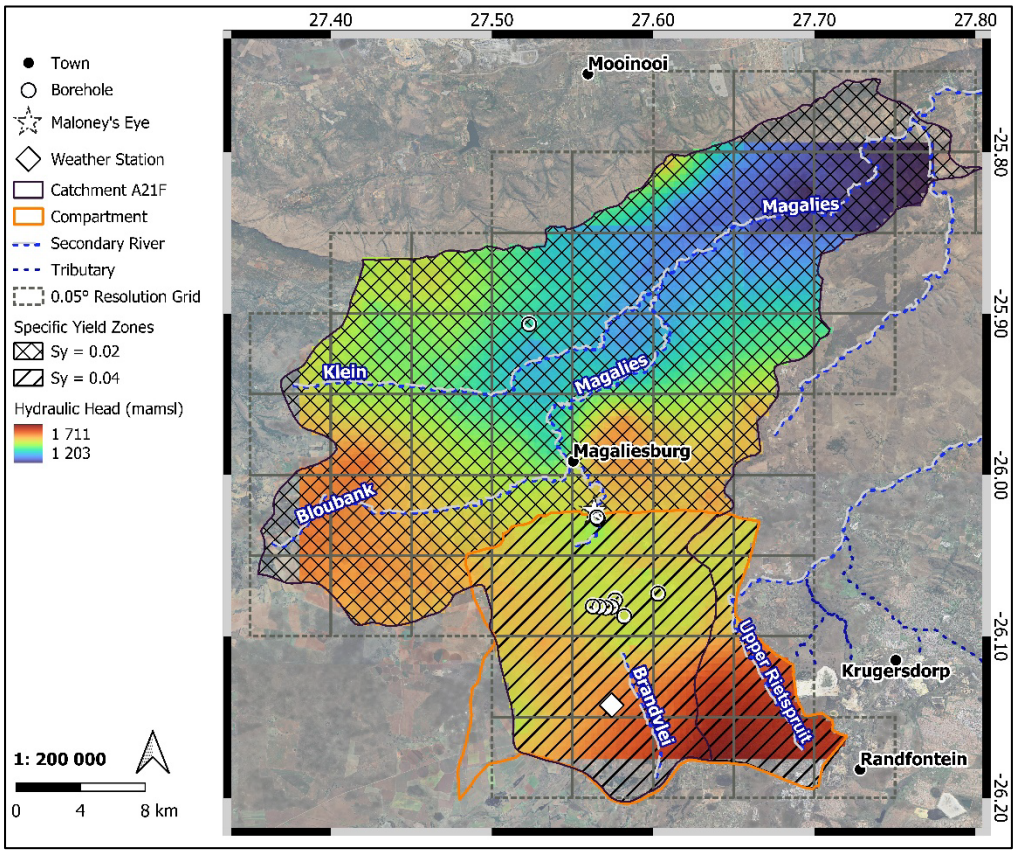


Figure 10: Map of aquifer specific yield overlaid on a groundwater hydraulic head map.

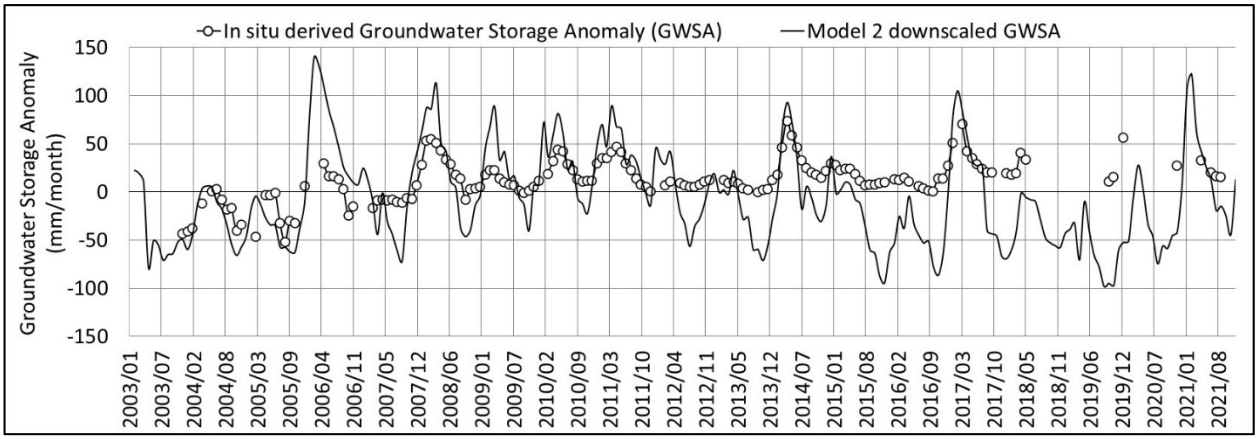
The validation results presented in Table 3, were achieved by comparing the downscaled GWSA to the *in situ* derived GWSA (Figure 11). Apart from having the highest r, Zone 1 exhibited the lowest RMSE and MAE, in contrast to Zone 2 and Zone 3, which displayed higher RMSE and MAE values. The *r* for Zone 4 was comparable to that of Zone 2, and Zone 4 exhibited the lowest RMSE and MAE across the SDC. This is ascribed to the larger GWSA amplitudes, or seasonality, depicted in the *in situ* groundwater level data for Zone 1, but less so for Zones 2, 3 and 4. This phenomenon can be attributed to the groundwater levels in karst aquifers and near springs remaining more stable due to the highly permeable nature of the aquifer, and the flow through the dolomitic subsurface into the less conductive shales and quartzites.

It is important to recognize that S_y varies horizontally and vertically within the same unconfined aquifer, and hence relying on a single value to characterise the entire aquifer simplifies a complex system (Chen et al., 2010). A higher S_y value of around 0.1 would be required for the *in situ* derived GWSA to reach the downscaled GWSA highs (shallower groundwater level), but this did not result in the lowest RMSE and MAE values and overestimated the GWSA lows (deeper groundwater levels). While high S_y values may be feasible in karst aquifers, they are likely not characteristic of the entire karst system but rather specific zones within the aquifer (Yu et al., 2022; Rose et al., 2018).

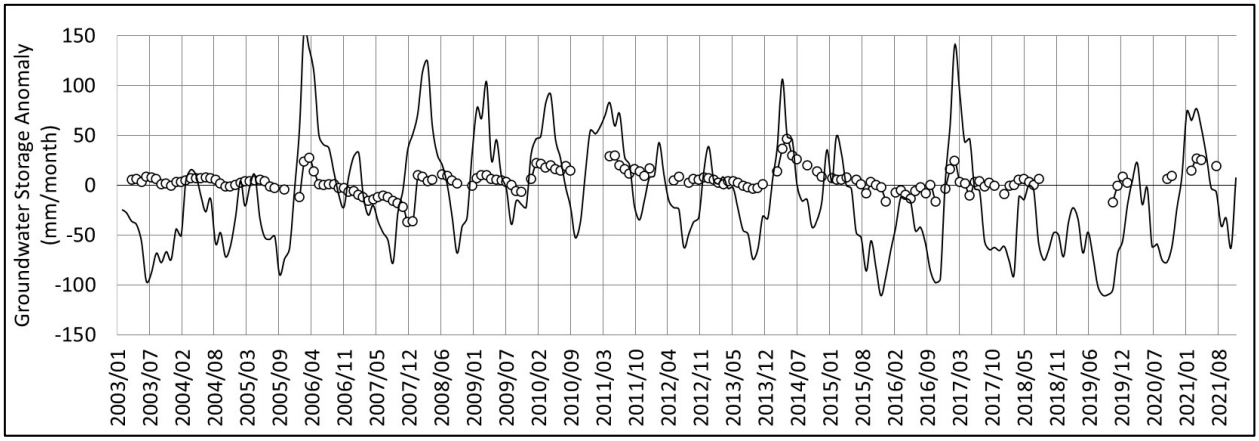
Table 3: Comparison metrics between the in situ groundwater level storage and the downscaled Groundwater Storage Anomaly (GWSA) for the respective zones.

	Zone 1	**Zone 2**	**Zone 3**	**Zone 4**
r	0.62	0.46	0.30	0.44
RMSE (mm)	41	51	51	45
MAE (mm)	32	42	39	36

a)



b)



c)

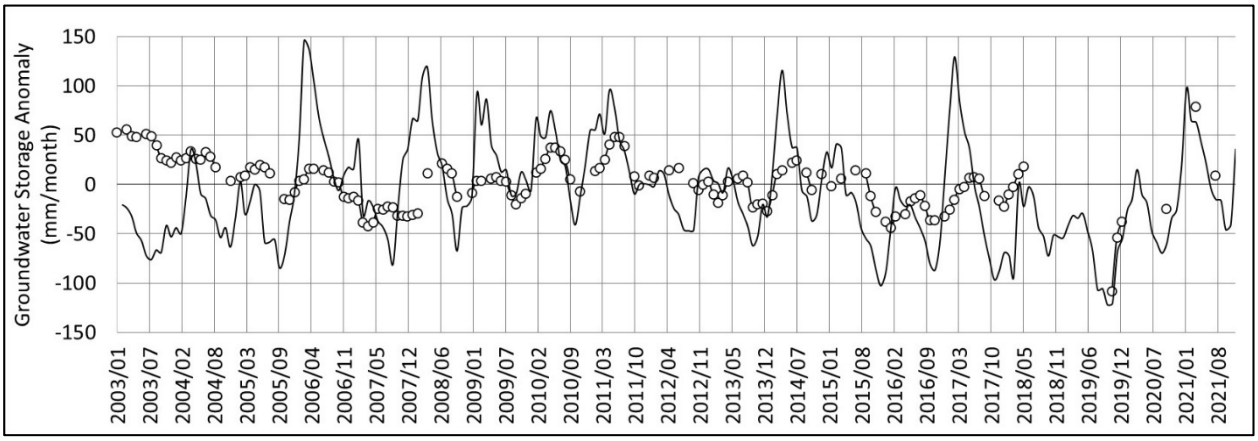


Figure 11: Comparison between groundwater level storage and the downscaled Groundwater Storage Anomaly (GWSA) for Zone 1 (a), Zone 2 (b) and Zone 3 (c) (Zone 4 is not illustrated as it is visually not highly distinguishable from Zone 3).

Gaffoor et al. (2022) achieved a MAE of approximately 170 mm in predicting current groundwater level changes for the SDC by incorporating thirteen variables and pre-determined temporal lags into GBDT models. Using only two input variables adjusted with the optimal lag to GWSA resulted in an average MAE of 37 mm across the SDC in this study. While the *r* strength differed across the respective zones, the downscaled data suggests that the groundwater levels do not only increase in depth but recover following precipitation events. This indicates that the aquifers are currently being managed sustainably, rather than being overexploited.

If the downscaling was done across an aquifer experiencing substantial land cover changes, such as expanded irrigated cultivation, or enduring consecutive years of below average precipitation, these dynamics could be highlighted and detected in the downscaled GWS product, even if not initially apparent in the low-resolution original data.

When evaluating the downscaled GWSA, it is important to acknowledge key uncertainties. These uncertainties include the limiting distribution of observed measurements and the challenges of directly validating high-resolution ET data against observed data. Additionally, local GWSA decreases unrelated to low precipitation and high ET, such as abstraction for domestic or industrial purposes, may not be fully accounted for by the downscaled product.

3.4. Conclusions

The study presented a simple framework for downscaling GWSA from 0.25° to 0.05° in the GEE cloud computing platform using the RF ML algorithm, and only precipitation and ET as input variables. The comparison between the monthly total, coarse-resolution GLDAS-2.2 GWSA and the monthly total downscaled GWSA estimates across the catchment, concluded that the downscaled product captures groundwater redistribution across the catchment well. This is because the model incorporates details about precipitation and ET, revealing local variations in recharge and discharge processes while adhering to the mass conservation principle.

The use of high spatial resolution, monthly time-series GWSA estimates can be applied to develop an efficient validation network, and to enhance the determination of high-resolution hydrogeological parameters such as groundwater recharge. Since the accuracy assessment and application of the downscaled GWSA necessitates careful consideration of how groundwater levels respond in different aquifers, future work could include an even more refined downscaling outcome by applying an adaptive approach where a separate RF model is trained for each aquifer type, and where viable *in situ* observations are included as an input variable. The use of varying Sy values should also be considered. Nevertheless, this study demonstrated the practical application of using open-source remote sensing products combined with ML to produce localised groundwater information to enhance groundwater management, particularly in regions with limited data availability.

4. Predicting streamflow in South Africa using deep learning

Christiaan Schutte, Michael van der Laan, Barend van der Merwe

4.1. Introduction

Developing countries around the world are facing significant water challenges, including limited water resources in low rainfall areas, water pollution, and inadequate infrastructure, all of which are exacerbated by population growth, urbanization and climate change (Oyebande, 2001). Hydrological information is crucial for ensuring the sustainable management of water resources, and understanding the complex dynamics of water availability, flow patterns, and groundwater recharge under a changing climate (Beven, 2011). Accurate and reliable hydrological data form the foundation for informed decision-making, enabling stakeholders to gain insights into water system dynamics, identify trends, and make informed choices on water allocation, infrastructure development and conservation. Streamflow data, a critical component of hydrological information, is vital for monitoring ecosystem health, determining sustainable water abstraction, and assessing flood and drought potential (Beven, 2011).

Various governmental and private sector institutions operate streamflow gauging stations to enable informed water resource assessment and planning (Rogers et al., 2019). The number of active gauging stations is, however, declining in many catchments around the world (Rogers et al., 2019). In South Africa for instance, the decrease is due to factors such as lack of funding and maintenance, wear and tear, vandalism and theft (DWS, 2021). Moreover, the availability of weather station data has also been declining in South Africa since around 1970 (Engelbrecht et al., 2009). This is a major concern as these data are critical inputs to process-based and deep learning (DL) hydrological models, with rainfall being the most important input to water resource studies (Pitman and Bailey, 2021). The decreased availability of streamflow and weather data hampers the ability to make informed decisions on water resource management and planning, and this presents a long-term threat to water security (Odendaal, 2021).

Where the availability of freshwater is highly variable and where resources required to sustain long-term monitoring programmes are constrained, hydrological models are of particular importance (Hughes, 2004). Process-based models used for streamflow prediction face limitations due to the lack of comprehensive information about system properties, including topography, soil characteristics and vegetation cover. These properties are highly heterogeneous and can change over time, while detailed knowledge of subsurface hydrological processes, where much of hydrology takes place, remains scarce (Kratzert et al., 2019a). In addition to data for model parameterization, data for initialization and calibration is also often limited. In contrast, a data-driven approach such as deep learning (DL), which is based on Artificial Neural Networks (ANNs), offers an alternative approach by predicting streamflow without explicitly defining the underlying physical processes. Additionally, the ability of ANNs to learn complex, non-linear relationships directly from data is particularly important in hydrology, as many

hydrological processes exhibit intricate and non-linear behaviours that process-based models struggle to represent (Kratzert et al., 2019a).

Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Cho et al., 2014) networks are DL architectures that were specifically designed to analyse sequential data. Streamflow data can be considered sequential and recent benchmarking studies have illustrated that these LSTM networks can rival and even outperform process-based and conceptual hydrological models in streamflow prediction (Kratzert et al., 2019c, Lees et al., 2021). Subsequently, GRU networks, as well as hybrid approaches combining LSTM with GRU networks (Muhammad et al., 2019) or combining Convolutional Neural Networks (CNN) with either LSTM or GRU networks (Ghimire et al., 2021, Anderson and Radić, 2022), were also found to be useful for streamflow prediction.

Despite their advances, Deep Learning (DL) techniques face challenges such as lack of interpretability (often termed 'black box'), ensuring physical consistency in predictions, and managing multi-dimensional data sets (Reichstein, 2019). The concept of theory-guided data science emerges in this context, aiming to merge well-established scientific theories with data-driven findings (Karpatne et al., 2017). Hybrid approaches, blending physical models' predictability with DL's adaptability, has been advocated in recent discussions (Reinstein, 2019). In line with this trend, there is a growing interest in physics- and hydrologically-informed machine/deep learning (Nearing et al., 2021).

Another major challenge when implementing pure DL or physics- and hydrologically-informed approaches, is the availability of sufficient, good quality training data, especially in developing countries where data collection and storage can be limited (Gaffoor et al., 2022). For example, high-resolution, spatially-distributed weather datasets may not be readily or freely available (du Plessis and Kibii, 2021), which presents a challenge for streamflow prediction. It should also be considered that a significant portion of existing ML/DL and physics- or hydrologically-informed research in hydrology has depended on high-quality, spatially-distributed datasets such as Catchment Attributes for Large-Sample Studies (CAMELS) (Addor et al., 2017). Such datasets are predominantly based in developed countries and thus not always representative of conditions in less-resourced settings. The performance of LSTM networks has also been shown to be less accurate in drier catchments (Kratzert et al., 2019c, Lees et al., 2021, Anderson and Radić, 94 2022), a characteristic of much of southern Africa. Such challenges must be considered when evaluating the feasibility of using DL models for streamflow prediction in the hydrological context of semi-arid, data-scarce and developing countries.

The main research question was whether LSTM and GRU networks could be developed in South Africa, a country which is currently experiencing a decline in streamflow measurements, to generate reliable streamflow estimations for predictive and data gap filling purposes? The increased availability of open-source gridded weather data provides a potential opportunity to address the issue of insufficient weather station data (Reichstein et al., 2019). The second research question was, therefore, whether freely available gridded weather data could be used as

input to the models to produce reasonably accurate streamflow estimates? Two catchments near Lydenburg, South Africa, were used as a case study to answer the research questions.

4.2. Materials and Methods

Study area

Situated in the north-east of South Africa, the study area consisted of two catchments in the headwaters of the Steelpoort River located in the Olifants River basin (Figure 12). Catchment A is approximately 100 km^2 and consist of quaternary catchment B42D as defined by the Department of Water and Sanitation (DWS). Catchment B is larger (300 km^2) and consists of quaternary catchments B42A and B42B. The average elevation ranges between 1 336-2 263 metres above mean sea level (mamsl). Based on the Köppen-Geiger climate classification system, the study area is classified as Cwb, indicating a warm temperate climate with cool and dry winters, and warm and wet summers. In winter, average daily temperatures range from 6-22°C, and from 22-32°C in summer (Herold and Bailey, 2016). The study area is in one of the country's higher rainfall areas, with most of the rainfall occurring between October and April. The average annual rainfall is approximately 800 mm yr^{-1}, but highly variable, and the average annual evaporation for an S-Pan is approximately 1 500 mm yr^{-1} (Herold and Bailey, 2016). The land cover of Catchment A mainly consists of grasslands, woodlands, and thicket areas. In addition to those land cover types, Catchment B also includes the town of Lydenburg, agricultural activities, and some minor pine plantations.

Data

Daily weather data, consisting of rainfall (mm), minimum and maximum temperature (℃), and streamflow data (m^3 s^{-1}) were used as input variables. The data included both in-situ measured data and gridded products. Daily weather data were obtained from three sources: 1) a weather station operated by the Agricultural Research Council (referred to as ARC in this study), 2) the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) data (Funk et al., 2015) (https://www.chc.ucsb.edu/data/chirps), and 3) the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) project (https://power.larc.nasa.gov/). The ARC weather station had a record spanning from 1979 to 2002. CHIRPS and NASAP data were downloaded for the same period. To download data from CHIRPS, a geojson file for each catchment was uploaded to the CHIRPS portal (https://climateserv.servirglobal.net/). The NASAP data were downloaded from the Water Research Observatory website (https://www.waterresearchobservatory.org/), where the data is available as a grid of points for South Africa. The grid point closet to each catchment was identified, and in this case, it was the same grid point for both catchments. Streamflow data were obtained from the DWS website (https://www.dws.gov.za/Hydrology/Verified/hymain.aspx) for stations B4H007 and B4H001 for Catchment A and Catchment B, respectively.

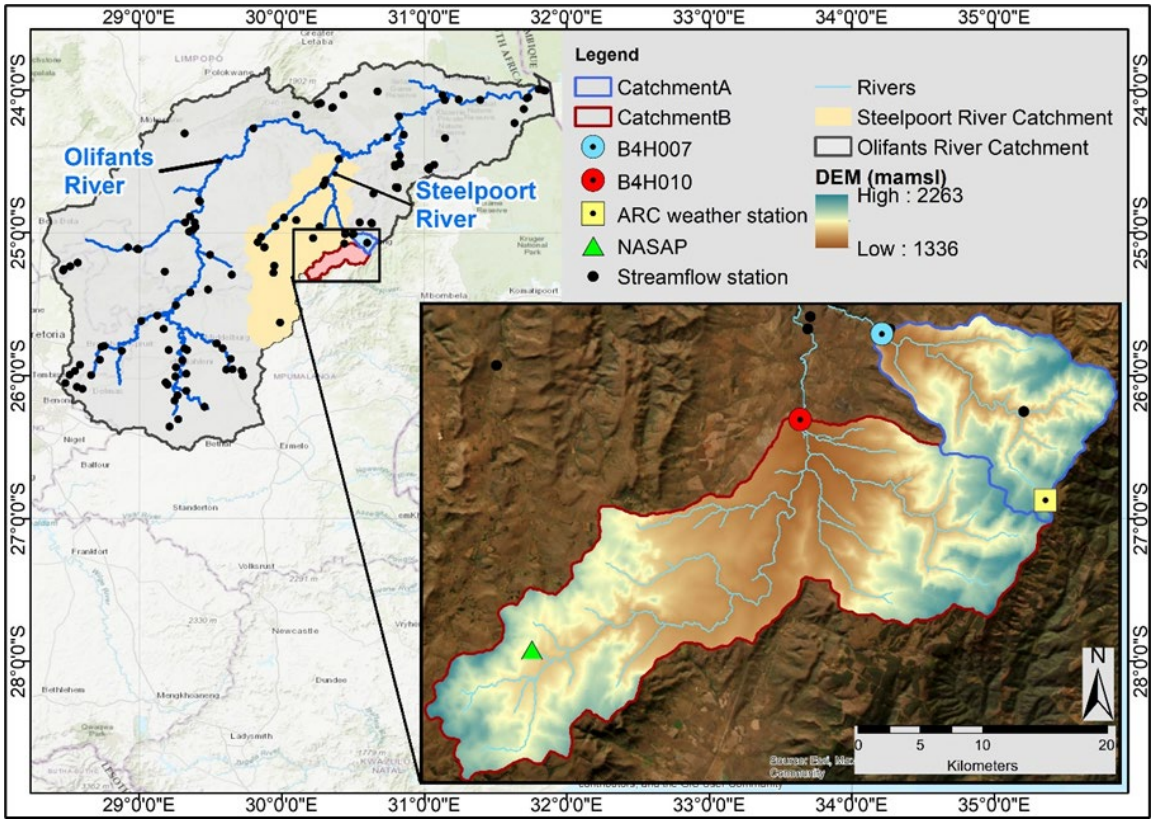


Figure 12: Locality map of the catchments, streamflow stations, the Agricultural Research Council (ARC) weather station and the National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) grid point. Catchment shapefiles were uploaded to the CHIRPS website to download catchment averaged rainfall.

The streamflow data were combined with each of the weather data sources (ARC, CHIRPS and NASAP) to create three datasets for each catchment. Each dataset was divided into a training set that was used to derive the optimal network weights, and a testing set, that was used to assess the prediction accuracy. In hydrological modelling, the first 70-80% of the data is often used to calibrate the model and the last 20-30% to validate the model. In this context, calibration and validation is analogous to training and testing a DL model. The training period spanned from 1 October 1979 to 30 September 1997, making it 18 years. The testing period was from 1 October 1997 to 22 February 2002 (date when ARC weather station was discontinued), which is 4 years, 4 months, and 22 days. The entire dataset covers 22.4 years.

Model development

High-level overviews of the workings of the GRU and LSTM have been provided elsewhere (Hochreiter and Schmidhuber, 1997, Cho et al., 2014). LSTM networks use a memory cell, which is a long-term memory of the network and is controlled by three gates: the input gate, the output gate, and the forget gate (Figure 13). These gates control the flow of information into and out of the memory cell, allowing the network to selectively retain or disregard information (Hochreiter and Schmidhuber, 1997). GRUs do not have a separate cell state, only a hidden state (Figure 13) (Cho et al., 2014). While LSTMs utilize three gates to manage the flow of information, GRUs only

use two gates. The update gate, that determines the extent to which the previous hidden state is updated, and the reset gate that determines the extent to which the previous hidden state is reset.

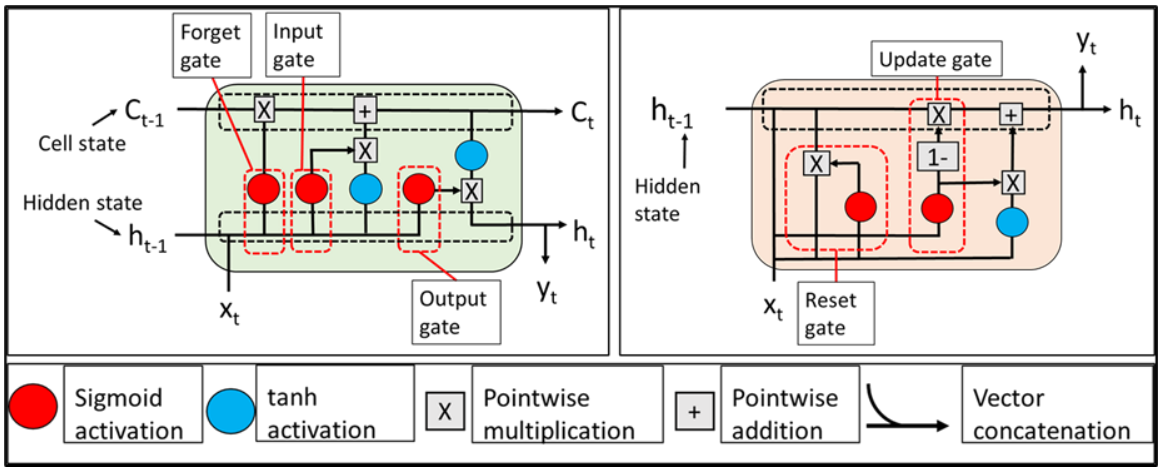


Figure 13: A comparison of the internal architecture of a standard Long Short-Term Memory (LSTM) cell (left) and a Gated Recurrent Unit (GRU) cell (right). The LSTM cell is characterized by its input, output, and forget gates as well as its cell state. The GRU cell uses a more streamlined architecture with update and reset gates but without a separate cell state.

Hyperparameters were selected based on a combination of experimentation and previous literature (Kratzert et al., 2018, Kratzert et al., 2019b, Fan et al., 2020, Nifa et al., 2023). An LBW of 30 days was selected, which has been proven effective in a semi-arid catchment Nifa et al. (2023) that has a climate more similar to the catchments in this study compared to those in snow-dominated or tropical basins (see Appendix I for additional information on the choice of LBW). The model architecture consisted of five hidden layers with twenty-five GRU or LSTM cells per layer, and a dense layer for the final streamflow prediction (Figure 14). Each of the 25 GRU or LSTM cells in the first hidden layer, processes four input features: rainfall, streamflow, maximum temperature, and minimum temperature. From the second hidden layer onward, each GRU or LSTM cell receives a 25-dimensional input, being the hidden state output from each cell in the previous layer. The dense layer then receives the outputs from all 25 GRU or LSTM cells in the fifth hidden layer to produce the final streamflow prediction. No activation function was used for the dense layer, which was linear.

A challenge in DL models is the generation of physically inconsistent predictions (Reichstein et al., 2019). In many instances during the dry seasons, when streamflow was close to 0 m^3s^{-1}, the models would start simulating slightly negative streamflow values, and then return to normal in the wet seasons (Appendix I). To prevent the models from predicting these negative streamflow values, two modifications to the GRU and LSTM network architectures were required. First, the hyperbolic tanh activation function Goodfellow et al. (2016) of each of the 25 GRU or LSTM cells, contained in the fifth hidden layer of the network, was changed to the rectified linear unit (ReLU) activation function (Krizhevsky et al., 2017) (refer to Figure 14 for the position of the tanh within a standard GRU or LSTM cell). Secondly, a non-negative constraint (Abadi et al., 2016) was used in the dense layer. Neither the ReLU activation function or non-negative constraint on its own prevented negative streamflow predictions, and both modifications were necessary and complemented each other effectively. The ReLU activation function in the

fifth layer could control the internal representations and mitigate the occurrence of negative values within the model layers, while the non-negative constraint in the dense layer served as an additional measure to enforce non-negativity in the final predictions.

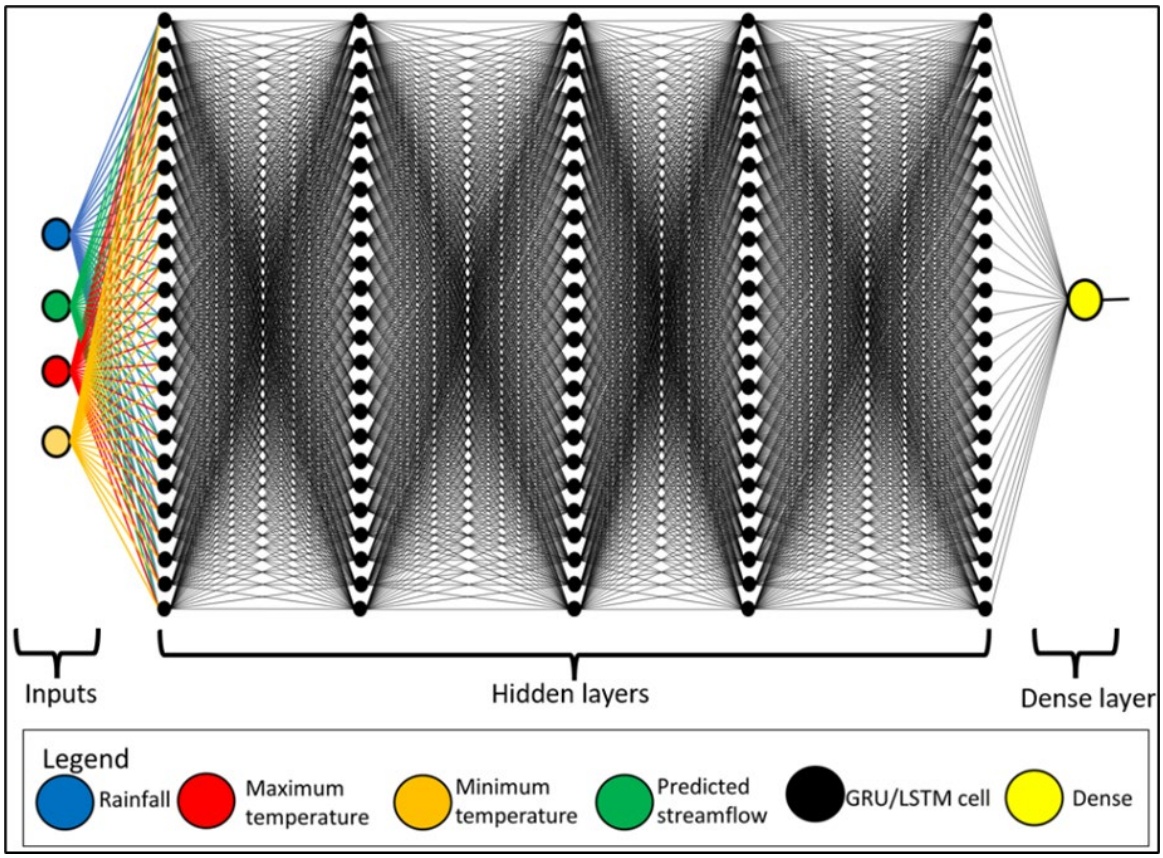


Figure 14: Schematic representation of the deep learning architecture with four input nodes, five hidden layers, each comprising 25 Gated Recurrent Unit (GRU) or 25 Long Short-Term Memory (LSTM) cells, where each of these cells utilizes a look-back window of 30 days. A dense layer follows the hidden layers, leading to the final prediction.

Before training, the MinMaxScaler function of the Scikit Learn Library (Pedregosa et al., 2011), was used to scale the training set and the testing set to a range between 0 and 1. The models were trained for fifty-five epochs (number of iterations) with a batch size of 256 (number of samples extracted from the training set at a time) (Kratzert et al., 2019b). To prevent overfitting, a dropout rate of 0.1 was used for regularization. The Mean Squared Error (MSE) (Bishop and Nasrabadi, 2006) was used as the loss function and the Adam optimizer (Kingma and Ba, 2014) was used for gradient descent. The weather variables of the testing set were used as input to the trained GRU and LSTM models to predict a time series consisting of 1 605 days of streamflow values. An essential aspect to note is that the models used the preceding 30 days of weather and streamflow values to predict the next day of streamflow (Figure 15). As only the measured weather variables (not the measured streamflow) contained in the testing set were used, the models had to incorporate the predicted streamflow values into the LBW to predict the entire time series. To do this, the model used the preceding 30 days of measured weather and predicted streamflow values to predict streamflow on the next day. The model then moved one time step forward and incorporated the

predicted streamflow value and the measured weather values of that day into the LBW. This process continued for the entire testing set and was achieved with a for loop (Shirzadi, 2023).

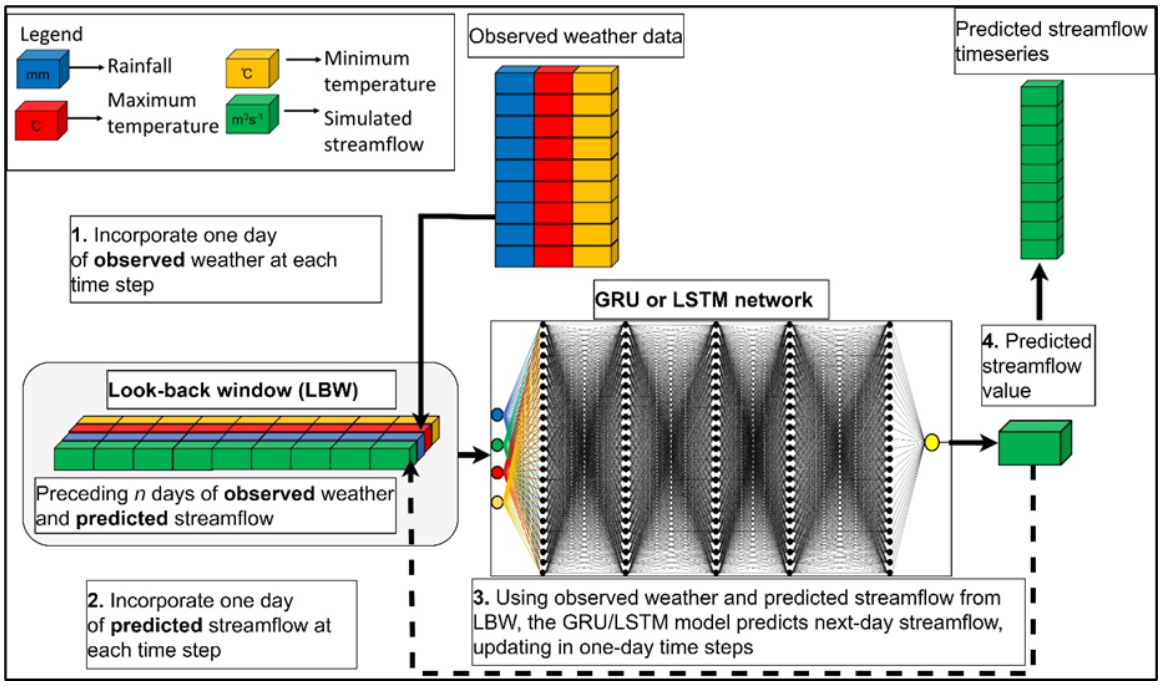


Figure 15: Based on the preceding 30 days of past weather and streamflow values in the look-back window (LBW), the network predicts a streamflow value for the next day. The model then moves one time step forward, ingesting the predicted streamflow value, along with observed weather values for that day, into the LBW to make the next prediction.

Ensemble predictions refer to the practice of combining the predictions of multiple individual models to obtain a more accurate and robust prediction (Goodfellow et al., 2016). For each catchment and architecture, 20 individual models were trained and a streamflow timeseries was predicted with each model. The ensemble predictions were created by computing the average across a set of 20 models. This approach allowed accounting for the variability in model performance due to random initialization and different training runs.

As the streamflow predictions were in a range between 0 and 1, they were scaled back to the original range with the inverse function of the Scikit Learn Library, in order to compare with measured streamflow contained in the original unscaled testing set. The Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970) and Kling-Gupta Efficiency (KGE) (Kling et al., 2012) were used for the evaluation of model performance, and was carried out in two stages. The first part of model evaluation involved assessing variation in prediction accuracy for each set of 20 GRU or LSTM models, while in the second stage, the evaluation focused on a comparative analysis between the best-performing models of each set of 20 models and the corresponding ensemble predictions within each set.

The NSE and KGE both range from negative infinity to one, with one indicating perfect model fit and values closer to 0 indicating poorer model performance. In hydrology an acceptable NSE/KGE value often depends on the nature of the specific application, however, an NSE/KGE value of 0.5 or higher is generally considered an acceptable

model, while a value of 0.8 or higher is considered a very good model (Moriasi et al., 2007). The Hydrostats package (Roberts et al., 2018) was used to compute the NSE/KGE values.

Experiment: Testing different sources and combinations of weather input data

Three different sources of weather data, and two different combinations of weather input variables, were tested as input for the GRU and LSTM models to explore the effect on model performance (Table 4). The first combination (Combination 1) consisted only of rainfall and streamflow as input variables, with observed streamflow used during training and model predicted streamflow during testing. Only rainfall was obtained for CHIRPS that was only included under Combination 1, while ARC and NASAP were used in both combinations. The second combination (Combination 2) included minimum and maximum temperature in addition to rainfall and streamflow. For each catchment and each combination, 20 GRU and 20 LSTM networks were trained, and ensemble predictions calculated for each set of 20 models.

In conventional ML paradigms, manual feature engineering is often a prerequisite to effectively model complex systems (Zheng and Casari, 2018). This frequently involves the calculation of lagged variables, especially in hydrological applications. However, one of the unique strengths of DL models lies in their capacity to automatically learn and identify relevant features from the data (Goodfellow et al., 2016). Leveraging these capabilities, the present study opted not to incorporate lagged variables for runoff prediction. Instead, we employed deep recurrent neural networks, which are inherently designed to capture essential temporal relationships and have been demonstrated to be effective in streamflow prediction (Kratzert et al., 2019c).

The Combination 1 predictions for each catchment, architecture and weather data source, were compared to the corresponding Combination 2, to determine if adding minimum and maximum temperature improved prediction accuracy. Next, the three weather data sources for each catchment and architecture were compared to each other, to determine which weather data source performed the best overall. The Shapiro-Wilk test (Shapiro and Wilk, 1965) confirmed that the NSE values violated assumptions of normality, therefore, the need for non-parametric tests. The non-parametric Mann-Whitney U test (Mann and Whitney, 1947) was used to test for significant differences between each of the three weather data sources and combinations.

Table 4: Combinations of weather sources and input variables for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and weather station (ARC) data.

Dataset	Combination 1	Combination 2
CHIRPS	Rainfall	-
	Streamflow	
NASAP	Rainfall	Rainfall
	Streamflow	Streamflow
		Maximum temperature
		Minimum temperature
ARC	Rainfall	Rainfall
	Streamflow	Streamflow
		Maximum temperature
		Minimum temperature

4.3. Results and Discussion

Experiment: Testing input variables and weather data sources

The NSE values varied from -9.4 (very poor) to 0.74 (very good), while the KGE values varied from -4.4 (very poor) to 0.77 (very good) (Figure 16). This range in the NSE and KGE values indicated a wide variability in model performance. The models also showed different performance spreads in cumulative frequency distribution (CDF) curves, specifically, the ARC models (both GRU and LSTM) tended to cluster around higher NSE and KGE values (their CDF curves shifted towards the right), implying better performance. A steeper slope, as observed for ARC data-driven models, implied that a significant portion of those models had closely clustered performance values. In contrast, the CHIRPS and NASAP models displayed a flatter curve, indicating a wider spread in performances. While the CDF plots for both catchments exhibited some differences, a consistent observation was the relative superiority of ARC data-driven models. Both in terms of NSE and KGE values, ARC models consistently outperformed those based on CHIRPS or NASAP data. There were notable differences between the CDF plots of Catchment A and Catchment B. While ARC remained superior in both cases, the degree of superiority and the relative performance of CHIRPS and NASAP models differed.

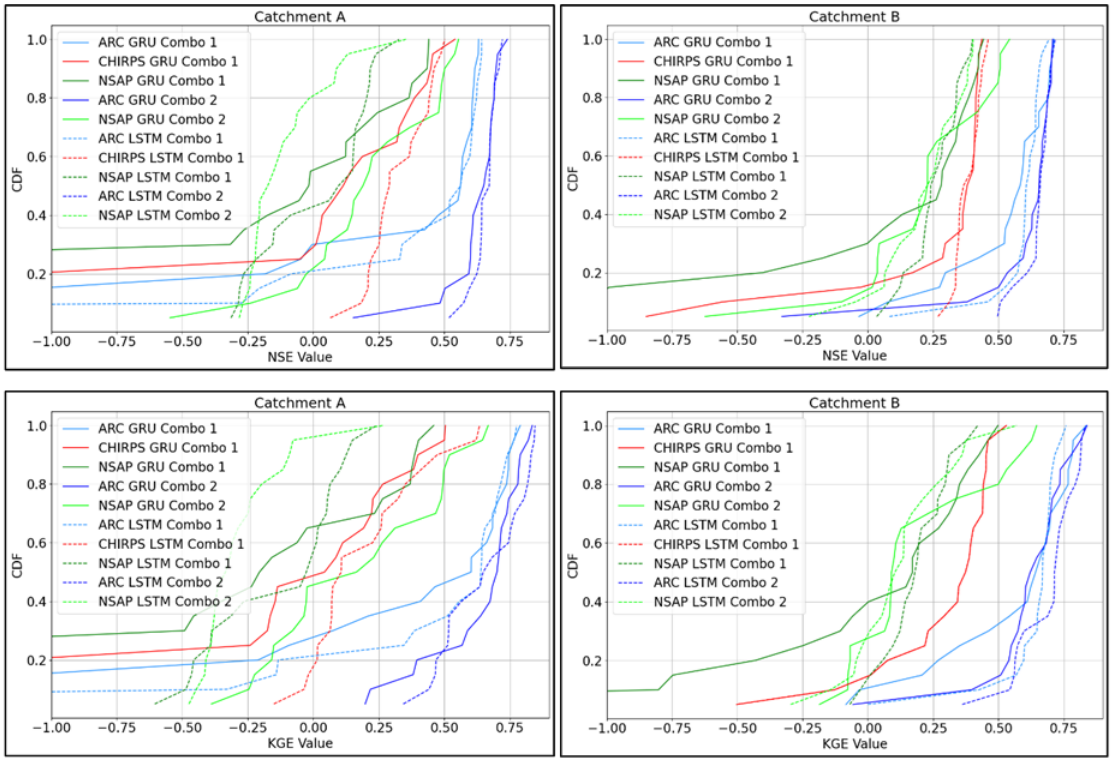


Figure 16: Cumulative frequency distribution (CDF) for Nash-Sutcliffe Efficiency (NSE) (top) and Kling-Gupta Efficiency (KGE) values (bottom) for Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) models for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council weather station data (ARC). Values smaller than -1 are not shown, with -10.9 being the lowest.

The CHIRPS and NASAP data-driven models were capable of moderately accurate predictions in a few instances with maximum NSE and KGE values greater than 0.5. The ARC weather station data driven models achieved considerably higher NSE and KGE values of 0.74 and 0.77, respectively, and the predictions were considered satisfactory. The Mann-Whitney U test revealed significant differences in both NSE and KGE values between the ARC models and both the CHIRPS and NASAP models for Combination 1 and for Combination 2 in both catchments (Table 5). The test also revealed significant differences in both NSE and KGE values between CHIRPS and NASAP for the LSTMs for Catchment A, and for both the GRUs and LSTMs for Catchment B (Table 5). There were no significant differences in NSE or KGE values between the CHIRPS and NASAP GRU of Catchment A. While significant differences exist between CHIRPS and NASAP in some scenarios, it's challenging to decisively point out which data source leads to better performance consistently across all configurations and catchments. This suggests that while they may differ, neither CHIRPS nor NASAP consistently outperforms the other.

The test revealed significant differences in NSE values between combinations 1 and 2 for the NASAP GRUs and LSTMs of Catchment A, while no differences in NSE were observed for the GRUs or LSTMs of Catchment B (Table 6). The test revealed significant differences in KGE values between Combinations 1 and 2 for both the NASAP

GRUs and LSTMs of Catchment A, while no differences in KGE values were observed for the GRUs and LSTMs of Catchment B (Table 6). The test also revealed significant differences in NSE values between Combination 1 and 2 for the ARC GRUs and LSTMs for Catchment A, as well as for the LSTM models of Catchment B (Table 6). No differences in NSE values were observed between combinations 1 and 2 for the GRUs of Catchment B. No significant difference in KGE values were observed between combinations 1 and 2 for the ARC data driven models. While the addition of temperature in Combination 2 led to increased prediction accuracy in certain scenarios, particularly in Catchment A, it did not consistently enhance performance across all model configurations and catchments.

No statistical significance tests were conducted for NSE (Table 7) and KGE (Table 8) values between the best-performing models and the corresponding ensemble predictions, due to the fact that only one ensemble was produced for each combination. The performance of the ensemble predictions varied across different data sources and combinations. The ensemble predictions for the CHIRPS data driven models did not exceed NSE or KGE values of 0.5, however, the best CHIRPS GRU model for Combination 1 achieved an NSE value of 0.54 for Catchment A. While the best CHIRPS GRU models for Combination 1 achieved KGE values of 0.58 and 0.57 for Catchment A and B, respectively.

Table 5: P-values for differences in Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) values between the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council weather station data (ARC) for Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

RNN	ARC vs CHIRPS	ARC vs NASAP	CHIRPS vs NASAP	ARC vs NASAP
	Combination 1			Combination 2
Catchment A				
NSE				
GRU	$1.1\times10^{-2*}$	$2.3\times10^{-3*}$	0.2	$1.1\times10^{-6*}$
LSTM	$9.8\times10^{-3*}$	$2.5\times10^{-4*}$	$6.7\times10^{-6*}$	$6.8\times10^{-8*}$
KGE				
GRU	$7.7\times10^{-3*}$	$1.7\times10^{-3*}$	0.38	$1.6\times10^{-5*}$
LSTM	$3.3\times10^{-3*}$	$1.3\times10^{-4*}$	$5.6\times10^{-4*}$	$1.1\times10^{-6*}$
Catchment B				
NSE				
GRU	$3.4\times10^{-4*}$	$5.9\times10^{-5*}$	$4.7\times10^{-2*}$	$4.0\times10^{-6*}$
LSTM	$1.4\times10^{-6*}$	$9.1\times10^{-7*}$	$2.9\times10^{-5*}$	$6.8\times10^{-8*}$
KGE				
GRU	$1.0\times10^{-3*}$	$4.1\times10^{-5*}$	$1.9\times10^{-2*}$	$1.8\times10^{-5*}$
LSTM	$2.4\times10^{-6*}$	$9.1\times10^{-8*}$	$7.7\times10^{-3*}$	$1.7\times10^{-7*}$

** Significantly different at values of $p \leq 0.05$.*

Table 6: P-values for differences in Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) values between Combination 1 and 2 for National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council weather station data (ARC) for Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM).

Catchment A		Catchment B	
GRU	LSTM	GRU	LSTM
NSE			
ARC			
$4.2\times10^{-4*}$	$2.0\times10^{-5*}$	0.12	$3.3\times10^{-3*}$
NASAP			
$2.6\times10^{-2*}$	$7.2\times10^{-2*}$	0.61	0.58
KGE			
ARC			
0.06	0.11	0.82	0.06
NASAP			
$1.7\times10^{-2*}$	0.09	0.62	0.19

** Significantly different at values of $p \leq 0.05$.*

For Combination 1, the best NASAP GRU models achieved NSE values of 0.44 in Catchment A and B. For the NASAP LSTM models, the NSE values were 0.48 and 0.46 for Catchment A and B, respectively. The ensemble predictions achieved NSE values of 0.19 and 0.27 for GRU and LSTM models in Catchment A, while for Catchment B, they were both 0.40. When observing the KGE values, the best NASAP GRU models achieved values of 0.51 in Catchment A and 0.45 in Catchment B. The LSTM models had KGE values of 0.50 and 0.46 for Catchment A and B, respectively. The ensemble predictions for KGE values were noticeably lower, at -0.15 and 0.22 for the GRU and LSTM models in Catchment A, and values of 0.20 and 0.30 for Catchment B, respectively.

The best NASAP models for Combination 2 achieved higher NSE (Table 7) and KGE (Table 8) values than Combination 1. The best NASAP models for Catchment A achieved NSE values of 0.55 and 0.35 for the GRU and LSTM models, respectively. The corresponding NSE values for the ensemble predictions were 0.39 and -0.09, respectively. The best NASAP models for Catchment A had maximum KGE values of 0.61 and 0.39 for the GRU and LSTM models, respectively. The corresponding KGE values for the ensemble predictions were 0.45 and 0.06, respectively. The best NASAP GRU and LSTM models of Catchment B achieved maximum NSE values of 0.54 and 0.40, respectively, and 0.38 and 0.26 for the corresponding ensemble predictions. The best GRU and LSTM models for Catchment B achieved maximum KGE values of 0.59 and 0.44, respectively, and 0.45 and 0.36 for the corresponding ensemble predictions.

For Combination 1, the best ARC models for Catchment A had maximum NSE values of 0.64 and 0.63 for the LSTM and GRU models, respectively. The corresponding NSE values for the ensemble predictions were 0.52 and 0.51, respectively. The best ARC models for Catchment A had maximum KGE values of 0.72 and 0.74 for the GRU and LSTM models, respectively. The corresponding KGE values for the ensemble predictions were 0.25 and 0.35, respectively. The best ARC GRU and LSTM models of Catchment B achieved maximum NSE values of 0.71 and

0.69, respectively, and 0.66 and 0.65 for the corresponding ensemble predictions. The best GRU and LSTM models for Catchment B achieved maximum KGE values of 0.81 and 0.76, respectively, and 0.62 and 0.68 for the corresponding ensemble predictions.

The best ARC models of Combination 2 for Catchment A achieved maximum NSE values of 0.74 and 0.72 for the GRU and LSTM models, respectively, and NSE values of 0.73 and 0.72, respectively, for the corresponding ensemble predictions. The increase in ensemble prediction NSE values for Catchment B were smaller than for Catchment A. The best ARC models of Combination 2 for Catchment B achieved NSE values of 0.72 and 0.71 for the LSTM and GRU models respectively, and NSE values of 0.72 for both ensemble predictions. The best ARC models of Combination 2 for Catchment A achieved KGE values of 0.77 and 0.76 for the GRU and LSTM models, respectively, and KGE values of 0.86 and 0.79 for the corresponding ensemble predictions. The increase in ensemble prediction KGE values for Catchment B were smaller than for Catchment A, with the best ARC models of Combination 2 for Catchment B achieved KGE values of 0.73 and 0.77 for the GRU and LSTM models, respectively, and KGE values of 0.79 and 0.83 for the corresponding ensemble predictions.

Table 7: Nash-Sutcliffe Efficiency values for the best Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) and ensemble predictions for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data.

Dataset	RNN	Catchment A		Catchment B	
		Best model	Ensemble	Best model	Ensemble
		Combination 1			
CHIRPS	GRU	0.54	0.38	0.44	0.43
	LSTM	0.48	0.42	0.46	0.41
NASAP	GRU	0.44	0.19	0.44	0.40
	LSTM	0.24	0.27	0.40	0.31
ARC	GRU	0.63	0.51	0.71	0.66
	LSTM	0.64	0.52	0.69	0.65
		Combination 2			
NASAP	GRU	0.55	0.39	0.54	0.38
	LSTM	0.35	-0.09	0.40	0.26
ARC	GRU	0.74	0.73	0.71	0.72
	LSTM	0.72	0.72	0.72	0.72

Table 8: Kling-Gupta Efficiency values for the best Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) and ensemble predictions for the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data.

Dataset	RNN	Catchment A		Catchment B	
		Best model	Ensemble	Best model	Ensemble
		Combination 1			
CHIRPS	GRU	0.58	0.22	0.57	0.38
	LSTM	0.50	0.47	0.46	0.49
NASAP	GRU	0.51	-0.15	0.47	0.20
	LSTM	0.14	0.21	0.35	0.31
ARC	GRU	0.72	0.25	0.81	0.62
	LSTM	0.74	0.35	0.76	0.68
		Combination 2			
NASAP	GRU	0.61	0.45	0.59	0.45
	LSTM	0.39	0.06	0.44	0.36
ARC	GRU	0.77	0.86	0.73	0.79
	LSTM	0.76	0.79	0.77	0.83

For all Combination 1 models, the NSE (Table 7) and KGE (Table 8) values of the ensemble predictions were lower than the NSE and KGE values of any single best model in all cases. For Combination 2 (only NASAP and ARC data driven models), this was still the case for the NASAP models, however, for the ARC models, the NSE values of the ensemble predictions were equal to the best models, and the KGE values of the ensemble predictions, were better than the best models in all cases.

Hydrographs for two years from the testing set for Catchment A are provided as an illustration (Figure 17). The model predictions generally showed a good fit with the observed streamflow, although they struggled to predict the very high levels of extreme peak flow events during 2000. Devastating floods in Mozambique, Zimbabwe, and South Africa (including the study area) during February to March 2000 brought about by Tropical Cyclone Eline in late February and a tropical depression early in March (Reason and Keibel, 2004). The extreme conditions of the cyclone provided a valuable opportunity to assess the ability of the models to extrapolate to unobserved conditions (such extreme flood events were not present in the training set), a crucial aspect in the context of climate change.

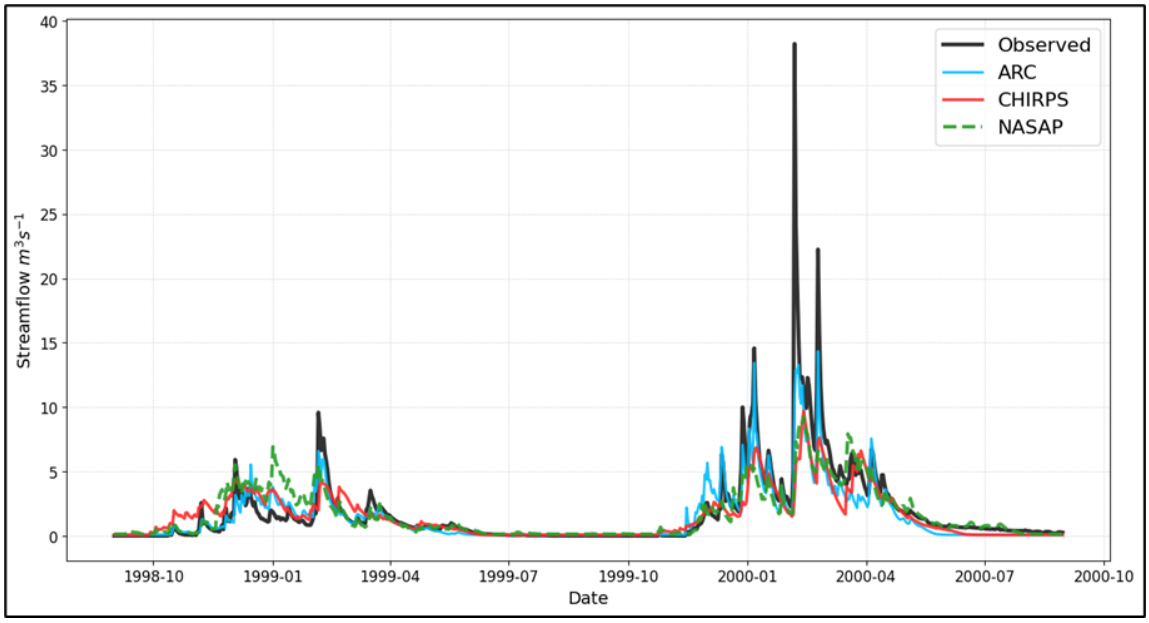


Figure 17: Best model of the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS), National Aeronautics and Space Administration Prediction of Worldwide Energy Resources (NASAP) and Agricultural Research Council (ARC) weather station data plotted against observed streamflow for two years of the testing set.

The method used in this study to predict streamflow involved two key aspects that distinguish it from previously published approaches. Firstly, the incorporation of simulated streamflow values into the look-back window (LBW) for predicting the streamflow of the testing set, was a novel approach that has not, to the best knowledge of the authors, been explored in the context of streamflow prediction. Secondly, the required modifications to the GRU and LSTM architectures, specifically the change in activation function of the cells in the final hidden layer of the network and the non-negative constraint used in the dense layer, played a critical role in making this method effective. The unique combination of these aspects has contributed to the advancement of DL techniques in hydrological modelling.

One notable observation is that researchers using DL in hydrology often lack details in their reporting methodologies (Sit et al., 2020), making it difficult to determine the exact methods used. While the use of calculated lagged streamflow variables has been observed in numerous articles, the specific approach adopted in this study, to incorporate predicted streamflow into the LBW, appears to be distinct. Similar approaches may, however, have been used for hourly data (Sit et al., 2021). Contrary to the approach used in this study, the Sit et al. (2021) model also utilized streamflow data from gauging stations situated in the upstream river network, which might be a limitation when such information is not available.

The approach of constraining the architectures is in line with the principles of theory-guided data science, specifically the concept of theory-guided design of model architecture (Karpatne et al., 2017). The targeted modifications to the architectures ensured physically consistent predictions, and were aimed at ensuring that the models adhered to certain physical principles, albeit in a simplified form, and proved effective in mitigating physically unrealistic predictions (negative streamflow values), allowing for the prediction of long term streamflow

time series. However, the 'black box' nature of DL models remains a challenge for the hydrological community (Anderson and Radić, 2022). The fact that it is not completely understood exactly how the two modifications prevented negative streamflow predictions, it is recommended for future work.

Wang and Karimi (2022) recently utilized the CAMELS dataset to analyze the effects of catchment mean rainfall and spatially distributed rainfall data on LSTM networks, and found that including spatial distribution information of rainfall could improve performance. However, as in this case, high-quality, spatially-distributed catchment weather datasets are not always available in less developed countries. The method developed here addresses the limitation by incorporating past streamflow data into the LBW of the model. This served as an implicit but effective proxy for hydrological processes that go unaccounted for and a lack of more spatially representative weather data. Specifically, the inclusion of historical streamflow captured some of the aggregated impacts of spatial and temporal variations within the catchment area, albeit indirectly.

The best models trained using weather station data could be used to generate reliable streamflow estimates with both GRU and LSTM networks. The best models trained using data from CHIRPS and NASAP were only moderately accurate. However, given the decline in weather station data availability, it is still encouraging that moderately accurate predictions were achieved using free data sources that are available for most of the surface of the earth. Future work could also explore combining CHIRPS and NASAP as input data to take advantage of the strengths of each data product while minimizing limitations (Kratzert et al., 2021). For example, CHIRPS may provide better accuracy for precipitation data, while NASAP could provide additional variables, such as temperature.

The results showed that rainfall together with streamflow were sufficient to obtain accurate streamflow predictions. Including minimum and maximum temperatures into the models did not consistently increase the accuracy, contrasting with (Fan et al., 2020). The reason for this might be attributed to the fact that streamflow predictions are primarily driven by rainfall and its subsequent runoff, with temperature playing a secondary role in influencing the evapotranspiration rates and soil moisture dynamics.

The ARC ensemble predictions for Combination 2 outperformed the best single models in terms of KGE for both catchments. For the NASAP dataset, which relied on remote sensing products, the ensemble did improve in some instances, but not as much as the ARC-based ensemble, and decreased in others. This discrepancy is considered to be related to the quality of the input data, specifically, the noise inherent in remote sensing products like NASAP, which lead to more varied model outcomes and, consequently, lower average accuracy in the ensemble. It is also essential to consider the nature of the catchments. Catchment characteristics, such as topography, soil type, and land use, can play a significant role in how different models perform (Beven, 2011). It seems that Catchment A has some specific features that make the ARC model ensemble, especially when temperature data is incorporated, perform exceptionally well, while the models with the NASAP dataset struggled.

Rainfall and temperature only were used in this study to test the applicability of this method for data-scarce regions, but further research with higher numbers of input variables is recommended. Future work could also investigate:

1) the integration of other relevant data, such as evapotranspiration and NDVI (Normalized Difference Vegetation Index), and 2) feature design (such as lagged variables), to further enhance the predictive capabilities of the hydrological models.

Although accurate streamflow predictions were achieved, there is room for improvement in capturing extreme peak flows, particularly in the context of climate change with an expected increase in extreme flood events (Allan, 2021). Future work could explore if combining this method with process-based models in a hybrid approach (Senent-Aparicio et al., 2019), or through informing the DL networks with physical principles (Reichstein et al., 2019), could increase accuracy in extreme peak flow prediction. The fact these models were able to exploit patterns in historic streamflow and weather data only to predict streamflow accurately, while not requiring information of geophysical properties of the catchments, is both an advantage and a limitation. This is because these models cannot be used to account for or assess the impact of land cover changes. Future work should explore the addition of land cover / land use change data, especially in more impacted catchments, where these changes can have significant impacts on streamflow (Beven, 2011). Finally, the data of many catchments could be combined together in an attempt to develop a model to predict streamflow in ungauged basins (Kratzert et al., 2019b).

4.4. Conclusions

In this study, the research questions were: 1) if LSTM and GRU networks could be successfully developed in semi-arid South African catchments to generate reliable streamflow estimations, and 2) whether freely available gridded weather data could be used to produce reasonably accurate streamflow estimates.

A challenge was obtaining spatially distributed weather data for the catchments. To compensate, 30 days' worth of past streamflow was used alongside rainfall and temperature to train the models to predict the next day's streamflow. Predicted streamflow was incorporated into the model input data during model testing to simulate long term time series. This resulted in negative streamflow predictions, mainly during the dry seasons. The GRU and LSTM were constraint in two ways to eliminate these predictions. This allowed for the successful training of GRU and LSTM networks using historic streamflow and weather data that could be used to generate satisfactory streamflow estimates based on two statistical criteria.

While the approach does not fully integrate hydrological or physical theories in the way that recent advances in physics- or hydrologically informed ML/DL have, it does employ domain-specific adjustments to improve the physical plausibility of model outputs, thereby offering a robust solution especially well-suited for regions where spatially comprehensive meteorological data are not readily available.

Both GRU and LSTM networks could be used as a fast and efficient technique to generate streamflow information. These techniques showed great potential and could be used, for example, to generate missing streamflow data for streamflow stations that were monitored in the past but are no longer monitored, or to fill gaps in records — a crucial need for long-term climate and hydrological studies. There is also potential to further develop and apply these models for short-term (daily time step) streamflow forecasts, which could have important implications for

flood risk management and water resource planning. Testing this method in more catchments under varying characteristics, where sufficient historic streamflow and weather data is available, is recommended to further validate this method.

Given that LSTM and GRU networks are inherently designed for sequential data, the method, though initially developed for streamflow prediction, holds promise for a broad range of time-series applications within hydrological modelling. These could include forecasting variables such as evapotranspiration (ET), soil moisture, and groundwater levels.

In conclusion, this study contributes to the growing body of literature affirming the capabilities of DL in hydrological modelling. It also provides valuable insights into the specific challenges and solutions associated with applying these models to streamflow prediction.

5. Big data analytics in precision agriculture

Simphiwe Maseko, Michael van der Laan, Eyob Tesfamariam, Marion Delport, Helga Otterman

5.1. Introduction

The recent emergence of big data and its analytics to revolutionize the agricultural industry has been an area of great interest globally. The development and deployment of big data analytics in agriculture involves the interaction between people, institutional and regulatory settings, and the technologies involved. Increased productivity and improved operational efficiency can be achieved with accurate data analysis in farming systems (Elijah et al., 2018). Having access and analytic capabilities on increasing numbers of farming datasets enables a more effective approach to prescriptive and predictive analysis. It is now possible to combine large datasets and analyses including climate and weather forecasts, crop simulation models, GIS mapping technology, consumer consumption data, pest management (Tantalaki et al., 2019).

The success of big data analytics application in agriculture largely depends on the implementation of this technology being able to respond to stakeholder dynamics and needs within the agricultural sector. For big data application to be effectively applied, stakeholders must be willing to share and integrate data, and new technologies need to be understood, adopted, and accepted by especially farmers and decision makers. Furthermore, protocols must be in place to protect farmers rights to privacy, data ownership and control (Sonka, 2016, Wolfert et al., 2017, Jakku et al., 2019). A careful consideration of social and technical implications needs to be in place since big data application can transform roles and power relationships between various stakeholders in the agricultural sector (Ryan, 2020).

The past decade has seen the emergence of agricultural technology provider (ATP) business component in big agricultural companies. These ATPs can install a range of data retrieving devices on the farm, farmer's machinery, and surrounding areas to monitor and obtain persistent datasets. This, however, might come with novel but immature technologies, that would need time to get validated, and farmers to efficiently handle and use them for precise and reliable data (Jakku et al., 2019). This means farmers might be made to work with newly introduced imperfect systems, requiring time to be able to get better with them and obtain accurate datasets. Numerous technologies for related problems may be offered to farmers by different technology providers, with different levels of complexities and recommendations, thus leaving farmers unsure of a better solution from these different ATPs.

The advent of advanced technologies, such as variable rate planters and applicators, has made site-specific data acquisition more cost-effective and has opened up avenues for the development of new decision support systems that can handle more intricate and data-heavy tasks than the conventional systems currently in use (Nyéki et al., 2017). The last two decades have seen the rapid expansion of on-farm research, especially in developing countries, owing to the increased adoption and use of PA technologies (Kyveryga, 2019). On-farm precision

experimentation is another type of on-farm experimentation (OFE) that enables the collection of large amounts of crop and soil data in a relatively short period of time and can be of special interest to large-scale farmers aiming to improve site-specific crop input management (Bullock et al., 2019). A multidisciplinary research project initiative called the Data Intensive Farm Management (DIFM) (Bullock et al., 2019), enables researchers to develop data-intensive, site specific input management advice and collaborate with farmers to provide guidance on how to make OFE systems pay back in their operations. This data collection may result to increased cost-effectiveness, but if adequately analysed, can have a huge potential to refine the current knowledge of agricultural systems.

Case study: Evaluating Machine Learning Models For Sub-Field Maize Yield Predictions In Precision Agriculture

To fully utilize big data analytics in the field of agriculture, it is necessary to advance scientific methodologies. Thus, the application of artificial intelligence (AI), particularly machine learning (ML) techniques, is highly relevant. As the amount of large, geo-referenced on-farm data becomes increasingly available, there is a need for analytical AI frameworks that can provide crop management recommendations and yield predictions. With the help of large datasets, it is now possible to conduct inductive research methodologies and investigate the complex interactions between crop management practices, environmental factors, and crop yield. This approach offers a practical and effective way to conduct large-scale agronomic research (Silva et al., 2020). By leveraging AI-powered data analysis, forecasting, and prediction techniques outlined by (Basso et al., 2016) farmers can make informed decisions that increase productivity and ensure the success of their crops.

Over the last decade, ML approaches have become increasingly prevalent in agriculture because of their ability to effectively address complex agricultural problems and nonlinear relationships leading to more accurate results (Pantazi et al., 2016, Tantalaki et al., 2019). One area that has seen particular growth is the use of ML to forecast crop yields, although the research community still debates on the most effective techniques for various data types and situations (Ransom et al., 2019, Van Klompenburg et al., 2020). Precision agriculture data, combined with ML techniques has proven to be particularly helpful in estimating crop biomass and yield (Näsi et al., 2018, Li et al., 2020), thanks to the ability of ML to handle large datasets with numerous variables, such as those created using PA equipment with data collection capabilities (Li et al., 2022). The advancement of AI applications has led to a broad range of applications for ML in agriculture, benefiting data gathering and selection to improve agricultural practices (Nawar et al., 2017).

The success of agricultural production heavily relies on farmers implementing advanced techniques, at every level of crop production, to increase yield per unit area. To aid in this endeavour, farmers can be assisted by an accurate model for crop yield prediction, which enables them to make informed decisions on when to produce certain crops. By predicting the yield of a specific site, farmers can also adjust the application of farm inputs, such as fertilizers, based on the anticipated crop and soil needs. This study employed ML algorithms to forecast maize yield in a PA farmed field. The objectives of this research are twofold: (i) to assess the predictive ability of the ML model using DIFM datasets and determine the best-performing ML model and (ii) to investigate the potential of ML models to

identify optimal input rates and limiting factors in a spatially variable field. We hypothesized that the performance of ML models in subfield predictions would be enhanced through the utilization of spatially representative DIFM datasets for training and testing.

5.2. Materials and Methods

Experimental site

The study was conducted in Henneman in the Free State province of South Africa (27˚51'16"S, 27˚01'15"E, 1 412 m.a.s.l.). This region is characterized by commercial medium-to large-scale farming of crops and livestock. It has a cold semi-arid climate, with hot and wet summer days and cooler, dry winters, an average temperature of approximately 18°C, and an annual average rainfall of approximately 600 mm yr^{-1}. Seasonal rainfall usually starts in October and ends in April, with more than 80% of rainfall occurring from December to March.

Trial design and management

This study was based on datasets collected in a DIFM maize field trial conducted in the 2019/2020 and 2020/2021 growing seasons. The DIFM trials are designed to generate data for localized crop response to site specific input factors (Bullock et al., 2019). The experiment had two management input factors, seeding rate (S) and nitrogen fertilizer application rate (Urea). The treatments were set up in a completely randomized factorial design, with nine seeding rate factors (10 000, 15 000, 18 000, 21 500, 27 000, 32 000, 38 000, 44 000 and 50 000 seeds ha^{-1}) for each of the two seasons. There were eight levels of Urea fertilizer rates (90, 120, 150, 170, 200, 225, 250, and 270 kg ha^{-1} urea in 2019/2020, and 105, 120, 150, 170, 200, 225, 250, and 270 and 300 kg ha^{-1} urea in 2020/2021) assigned randomly throughout the field in each of the two seasons. The procedure used for fertilizer application was an initial uniform 200 kg/ha 15.10.6 (31) NPK mixture applied throughout the field at planting. The N variation treatments were then implemented by applying the different urea rates banded 15 cm offset to the row and 10 cm deep. The standard practice of the farmer was to apply 18 000 seeds ha^{-1} and 224 kg ha^{-1} of urea. The plots were designed to be 15 m wide and 73 m long. The plot width is typically determined by the width of the planter used, such that every plot hosted one pass of the planter and one pass of the fertilizer applicator, the 7 3 m length is determined by the distance it takes for the planter to change input application rates as determined by the trial design. All treatments were implemented in the field using variable-rate-enabled seed planters and fertilizer applicators. A medium season cultivar from seed company Dekalb (DKC 78-77 BR) was used consistently throughout the field over the two seasons. A rate of 18 000 seeds ha^{-1} was assigned to a buffer zone around the perimeter of the trial, but observations from the buffer zone were not included in the subsequent analysis.

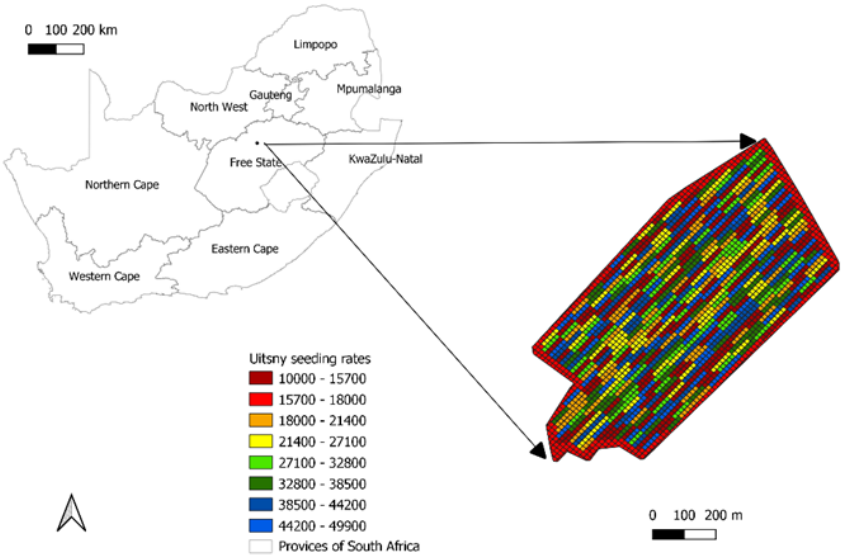


Figure 18: The maize (Zea mays L.) field study was conducted in Henneman in the Free State province of South Africa. An example of different seeding rate treatments is shown.

Data collection and processing

The final dataset used consisted of maize grain yield (20% moisture content) as a dependent variable and 24 georeferenced management, soil, and remotely sensed data as independent variables (see Table 9 for all list). The data were processed to represent multiple small plots across the field, with each plot having a unique yield value linked to the management, soil and remotely sensed data.

Yield data

Yield data were collected using a calibrated yield-monitoring system mounted on the combined harvester, which recorded the yield data every two seconds during the harvesting process. The data cleaning procedures of DIFM trial data which were followed in this study were discussed in detail by Bullock et al. (2019). The farmer utilized a strategic harvesting technique in which during every other pass, the combined harvester traversed through the center of the plot. As a result, yield data were gathered solely from the middle 50% of each plot. Raw 'as applied' and harvest data were retrieved directly from the variable rate applicators and yield monitors. The raw data were cleaned to remove observations with extreme yield or as applied rates ('outliers'). Additionally, data points were excluded from the headlands due to varying sun exposure, fluctuations in machinery driving speed, and the possibility of application overlaps, which made the data less reliable. The DIFM strategy also involved the placement of about 10 m 'transitional buffer zones' at the end of each plot where the planter could be changing from one application rate to another. The distance between points, swath width, and headings recorded in the raw yield data was used to make yield polygons, and subplots were made by combining yield polygons with similar N rates into groups (yield polygons combined made a subplot of about 12 m in length). The average value of all yield

points within each subplot was calculated and used in the analysis as a single observational unit. After undergoing data cleaning, 5 748 and 3 409 observational units were analyzed for the respective seasons. The first season yielded more data points than the second due to more heterogeneous yields throughout the field in the 2020/2021 season, and some plots had yields that were too low for meaningful analysis.

Soil data

Data on soil distribution were collected by a commercial fertilizer company (OMNIA) for soil analysis, which was conducted on-site before the start of each planting season. The variables measured at each location for both the topsoil (0-0.3 m) and subsoil (0.3-0.6 m) included physical properties such as clay percentage and soil depth as well as chemical properties such as soil pH, potassium (K), Bray P, calcium (Ca), magnesium (Mg), and sodium (Na). The soil was sampled in 62 locations across the field on a 100 m grid, and five samples were taken in each sampling point about 3 m apart. This investigation utilized 24 variables for yield predictions, of which 14 were primarily focused on the physical and chemical properties of both the topsoil and subsoil.

Remotely sensed data

The normalized difference vegetation index (NDVI), which gives an indication of canopy development and health, is a highly effective tool for assessing crop yield potential. To incorporate real-time data into our model training and testing, NDVI data were used to track crop growth at various stages of the growing season. NDVI values were calculated using Sentinel 2A images with a 10 m resolution in QGIS, creating raster files from which the NDVI was calculated for each pixel. These images were captured between 1 November 2019 and 30 April 2020 as well as 20 November 2020 and 30 April 2021, and downloaded through the Copernicus hub (https://scihub.copernicus.eu). The NDVI was calculated in QGIS by using raster files containing near infrared and red bands from Sentinel 2A images. The centroids for each plot from the yield data shapefile were used to sample the corresponding NDVI value. The NDVI data was extracted at seven different intervals, beginning 11 days after emergence (DAE) and continuing until the crop reached physiological maturity (135 DAE). The images were carefully selected to focus on the area of interest (the maize field), and only those with less than 5% cloud cover were included in the analysis. Finally, the raster files were sampled using the centroids from each yield polygon to extract the NDVI time series for corresponding yield points from emergence to harvest.

Table 9: The agronomic management, soil and remotely sensed variables used in model development.

	Variable name	Description	Units
Agronomic	Plant_pop	Plant population	seeds ha^{-1}
	Urea	Urea application	kg ha^{-1}
Soil	pH_top	Soil pH in topsoil	-
	Bray_top	Phosphorus in topsoil	mg kg^{-1}
	K_top	Potassium in topsoil	mg kg^{-1}
	Mg_top	Magnesium in topsoil	mg kg^{-1}
	Na_top	Sodium in topsoil	mg kg^{-1}
	S_top	Sulphur in topsoil	mg kg^{-1}
	Clay_top	Clay content in topsoil	%
	Bray_sub	Phosphorus in sub soil	mg kg^{-1}
	K_sub	Potassium in sub soil	mg kg^{-1}
	Mg_sub	Magnesium in sub soil	mg kg^{-1}
	Na_sub	Sodium in sub soil	mg kg^{-1}
	S_sub	Sulphur in sub soil	mg kg^{-1}
	Clay_sub	Clay content in sub soil	%
	Soil_d	soil depth	m
Remotely sensed	11DAE_ndvi	NDVI at 11 DAE	unitless
	25DAE_ndvi	NDVI at 25 DAE	
	60DAE_ndvi	NDVI at 60 DAE	
	85DAE_ndvi	NDVI at 85 DAE	
	100DAE_ndvi	NDVI at 100 DAE	
	110DAE_ndvi	NDVI at 110 DAE	
	120DAE_ndvi	NDVI at 120 DAE	
	135DAE_ndvi	NDVI at 135 DAE	

Machine learning maize yield predictions

The ML models were built using Python Keras libraries in a Google Collaboratory cloud computing environment. The ML algorithms were implemented using a multi-step process. The data file included attributes from the soil, agronomic management practices, and remotely sensed data as independent variables, with maize yield as the dependent variable. The processed data was utilized for both model training and testing purposes. The data used for model training and testing were from 2019/2020, 2020/2021, and a merged dataset of the two seasons. The data points used for model training and testing was 4 025, 2 387 and 6 410 observational units for the 2019/2020, 2020/2021, and merged datasets for both seasons, respectively. An 80/20% ratio data split was used on the data for model training and testing in each of the four ML models.

To predict maize yield, four ML algorithms (MLR, MLP, DT, and RF) were trained and tested. The MLR models the relationship between two or more explanatory variables and a dependent variable, assuming a linear relationship. For predicting crop yields, MLR has been a popular technique (Drummond et al., 2003, Van Klompenburg et al., 2020). Multiple linear regression uses least-squares optimization to determine the dependent variable that best fits each independent variable (measured yield). The MLR model was developed according to Equation 5.1.

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots\cdots + \beta_n X_n + \varepsilon_i \quad (5.1)$$

where y_i is the grain yield, β_0 represents bias, $\beta_1 - \beta_n$ are the coefficients of regression, $X_1 - X_n$ are the input variables, and ε is the error associated with the *i*th observation.

The MLP model was created using Keras, a deep learning application programming interface (API) implemented in Python. Because the models in Keras are described as a series of layers, a sequential model was initially created, and then four additional layers were added using the Rectified Linear Unit (ReLu) activation function. An Adams gradient descent optimizer was chosen with default hyperparameters, as tests have shown that this is a good optimizer when used with adaptive learning rates (Ruder, 2016). We implemented a mean squared error (MSE) loss function and a maximum of 500 epochs.

The DT is a type of supervised learning model that can be used for both classification and regression tasks. It is capable of selecting an outcome from a tree of potential decisions (Maimon and Rokach, 2014, Perez-Alonso et al., 2017). The tree structure resembles a flowchart and is used to evaluate issues by considering numerous features and attributes. In this study, the Scikit-learn Python module "DecisionTreeRegressor" class was applied, with a maximum depth of 30 trees.

Random Forest is a tree-based ensemble model built on the concept of bagging, which averages final predictions from different training subsets made by sample training data with replacement in an effort to reduce prediction variation (Breiman, 2001). Random forest adds a new feature to bagging by randomly selecting a set of features, building a tree with those features, repeating this process numerous times, and then averaging all the predictions made by the trees (Shahhosseini et al., 2021). The Gini index was used to identify the key characteristics that

significantly influence the yield based on the various independent variables. This selection process of features is crucial for identifying significant variables that explain yields and could highlight limiting factors in agronomic terms.

Model evaluation

Performance evaluation measures were used to assess the level of accuracy of each prediction model and select the most suitable algorithm for supervised learning regression exercises. To evaluate our ML models, we used the root mean square error (RMSE) and mean absolute error percentage (MAPE). The degree of correlation between the expected and actual values was gauged using the coefficient of determination (R^2). These metrics can be used for both regression and classification tasks (Naser and Alavi, 2020). The closer R^2 is to 1, the higher the prediction performance of the model. Small RMSE and MAPE values indicate a small discrepancy between the observed and predicted yields. It is generally accepted that the model with the smallest estimation error is the best.

5.3. Results and Discussion

Descriptive yield analyses and model training and testing

This study focused on comparing ML techniques to define the relationship between soil, agronomic management, and remotely sensed data for sub-field scale maize yield predictions and the identification of influential attributes explaining yields. The descriptive statistics of seeding rates, urea application, and grain yield during the 2019/2020 and 2020/2021 seasons are presented in Table 10. Maize yields varied from 6.8-12.4 t ha^{-1} and 2-13.7 t ha^{-1} in the 2019/2020 and 2020/2021 seasons, respectively. The 2019/2020 season had a higher average yield (9.7 t ha^{-1}) and lower yield standard deviation (1.2 t ha^{-1}) than the 2020/2021 season (8.6 t ha^{-1} average yield and 2.1 t ha^{-1} standard deviation). Although the dataset presented a wide range of yield values, the spatial structure of variability was still unclear despite the rich dataset.

Table 10: Descriptive statistics of maize yield for the 2019/2020 and 2020/2021 seasons.

index	**Seeding rate**		**Urea**		**Yields**	
	seeds ha^{-1}		kg ha^{-1}		t ha^{-1}	
Season	2019/202	2020/2021	2019/2020	2020/2021	2019/2020	2020/2021
std	10527	9594	55	62	1.2	2.1
min	5613	13676	94	104	6.8	2
mean	30434	28862	181	204	9.7	8.6
max	49881	49381	276	308	12.4	13.7
count	5748	3409	5748	3409	5748	3409

Based on the descriptive statistics of the measured yield data, it was clear that there was variation in the sub-plot yields within a single season and between the two seasons. The yield variation between the two seasons can be

explained by differences in the total rainfall received from planting to harvesting, which was 674 mm and 439 mm for the 2019/2020 and 2020/2021 seasons, respectively (Figure 19). It is noted that there was also a high incidence of tillering from the planted cultivar in various treatments, especially in a combination of low plant populations and high fertilizer rate treatments, which could have contributed to final grain yield but not accounted directly by the seeding rate.

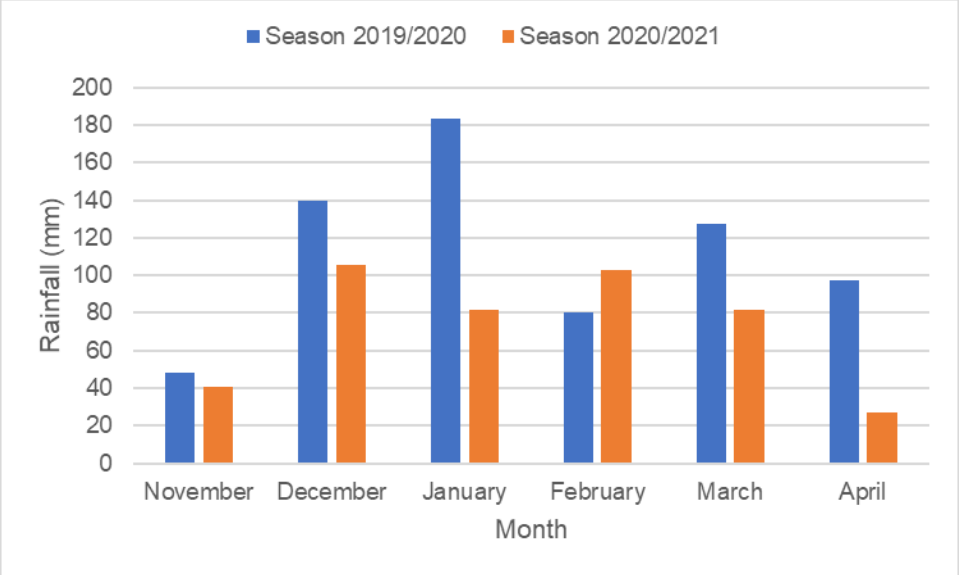
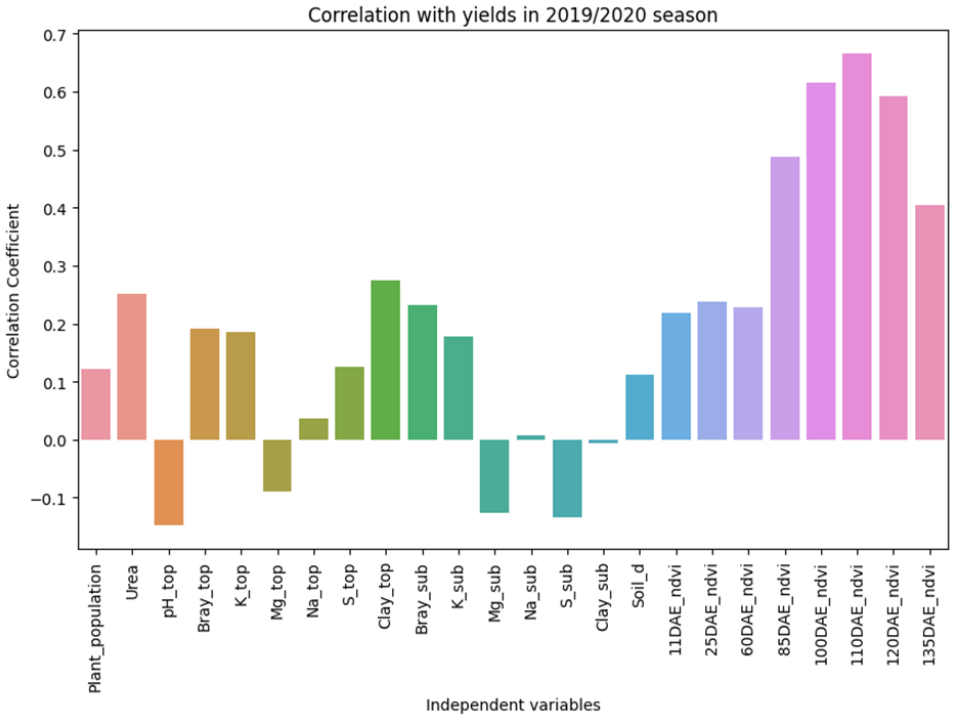
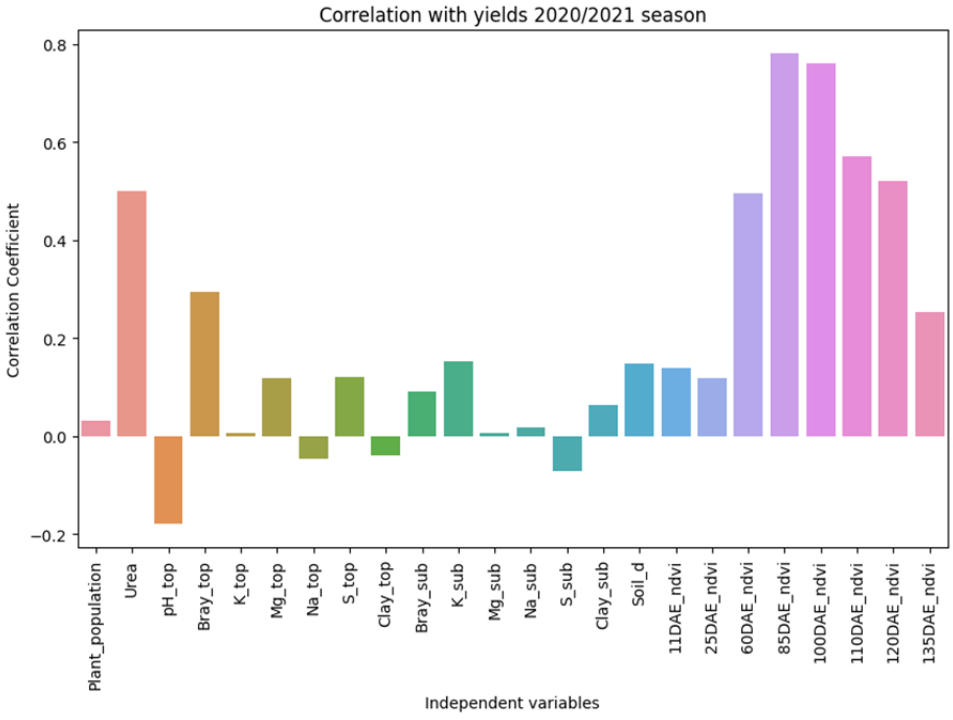


Figure 19: Rainfall distribution for the two seasons (2019/20 and 2020/21) covering the period from planting to harvesting.

It is important to first learn about the data using statistical tools for successful training of ML algorithms. The correlation analysis results for all the datasets are presented in Figure 20. The results indicated that the extent of the relationship between yield and individual yield-influencing attributes varied between seasons. Despite seasonal variations, a distinct positive correlation between yield and agronomic management and NDVI at all growth stages in both seasons was observed. Urea application had a stronger relationship with yield compared to the plant population for each of the two seasons and when seasonal data were combined. Plant population and urea application also helped explain the yield differences between the two seasons, with urea application having a higher correlation with yields in the 2020/2021 season than in the 2019/2020 season, while the opposite could be observed for plant population, with a higher correlation in the 2019/2020 season than in the 2020/2021 season. This could be a result of the higher rainfall received and better distribution in the 2019/2020 season than in the 2020/2021 season. The NDVI at 110 and 85 DAE had the highest correlation with yield in the 2019/2020 and 2020/2021 seasons, respectively. The extent to which some soil attributes were related to yield could be negative (pH_top, and S_sub), positive (Bray_top, Bray_sub, K_top, K_sub and clay_sub), or both (Mg_top, Mg_sub, clay_top and Na_sub) over the two seasons. There was no clear positive or negative interpretation of these variables and yields, suggesting that the relationships were nonlinear.


Correlation with yields in 2019/2020 season
Correlation Coefficient
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
Plant_population
Urea
pH_top
Bray_top
K_top
Mg_top
Na_top
S_top
Clay_top
Bray_sub
K_sub
Mg_sub
Na_sub
S_sub
Clay_sub
Soil_d
11DAE_ndvi
25DAE_ndvi
60DAE_ndvi
85DAE_ndvi
100DAE_ndvi
110DAE_ndvi
120DAE_ndvi
135DAE_ndvi
Independent variables


Correlation with yields 2020/2021 season
Correlation Coefficient
0.8
0.6
0.4
0.2
0.0
−0.2
Plant_population
Urea
pH_top
Bray_top
K_top
Mg_top
Na_top
S_top
Clay_top
Bray_sub
K_sub
Mg_sub
Na_sub
S_sub
Clay_sub
Soil_d
11DAE_ndvi
25DAE_ndvi
60DAE_ndvi
85DAE_ndvi
100DAE_ndvi
110DAE_ndvi
120DAE_ndvi
135DAE_ndvi
Independent variables

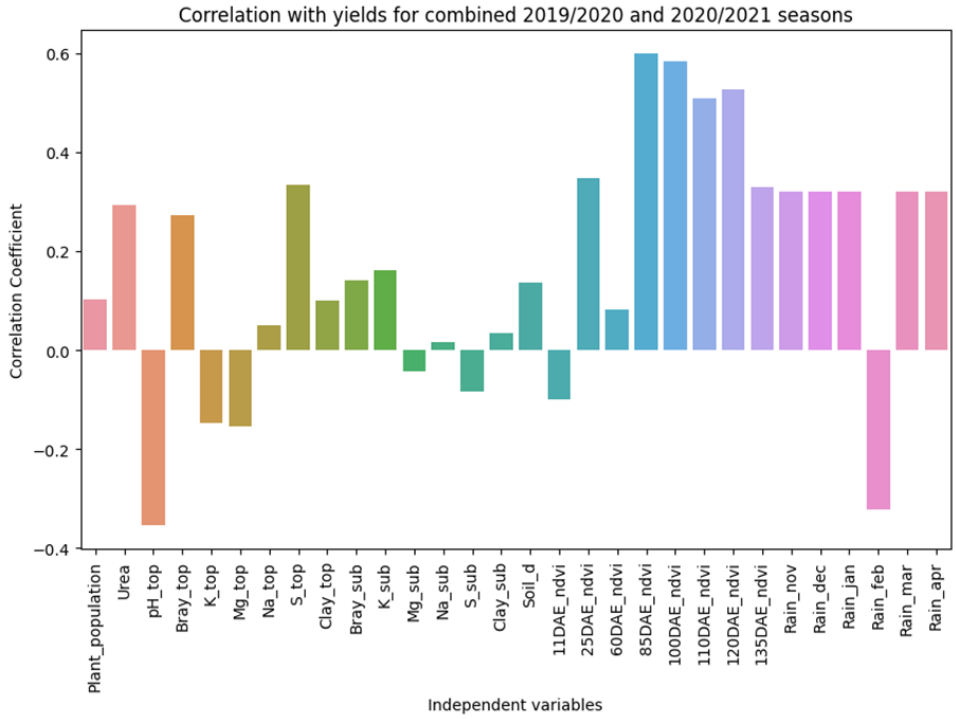


Figure 20: Correlation analysis for the relationship between agronomic management, soil, remotely sensed, and weather data and maize yields for the 2019/2020 and 2020/2021 seasons and combined data for the seasons.

After combining the data of the two seasons into one dataset and including monthly rainfall as variables, urea application still proved to be more important than plant population in explaining yield. The precipitation pattern during the 2019/2020 season was more favourable than that of the 2020/2021 season, with a higher total rainfall amount and better-distributed precipitation from planting to harvest. Most months in 2019/2020 recorded higher monthly rainfall, apart from February. This discrepancy in February rainfall explains the negative correlation between February rainfall and crop yield. Other notable attributes explaining yields were S and Bray P in the topsoil and soil pH, which had the highest correlation, whereas S_sub and Na_sub had the lowest correlation with yields over the two seasons. The negative correlation between soil pH and yield indicates that, for this field, there may be a compounding effect of other yield-influencing factors that are associated with an increase in soil pH, which results in certain parts of the field having lower yields at higher soil pH. Through the correlation analysis, it was evident that there was a complex relationship between maize yield and the different yield attributes during the two seasons. The varying correlations between yield and yield-influencing attributes across seasons reflect the complexity of agricultural systems. Varying weather and the interaction between multiple yield-influencing variables can contribute to varying correlations, such as soil properties, which may have a more positive effect on yield in a wet season than in a subsequent drier season.

Evaluation of ML algorithms for sub-field maize yield predictions

The R^2 values between the actual and predicted yields for the four ML models using datasets with and without NDVI are shown in Figure 21. An analysis of the overall trend revealed that when NDVI data were factored in, the

ML models, in most instances, demonstrated improved predictive accuracy in both seasons. The incorporation of NDVI strengthened the linear associations between yield and yield prediction attributes within the dataset, which was effectively captured by the models in most cases, particularly MLR. In contrast, when comparing model performance with data without NDVI, the MLR, DT, and RF exhibited either increased or equal prediction accuracy from the 2019/2020 to 2020/2021 seasons, while MLP showed a decline in prediction accuracy from the same period when trained and tested without NDVI data. The model predictions using NDVI data revealed that MLR (R2 = 0.55 and 0.74) had superior prediction accuracy compared to MLP (R2 = 0.5 and 0.65) across both seasons. Furthermore, this was true when combined data were used, and NDVI was included. This was likely a result of the higher correlation between NDVI and yield, as indicated in the descriptive statistics, being able to capture more linear relationships between NDVI and yields in the MLR model, while the MLP also considered the nonlinear relationships that lowered the prediction accuracy. Previous studies have suggested that linear regression analysis is less effective in agronomic studies because yield data are a result of multiple interacting factors (Kitchen et al., 2003).

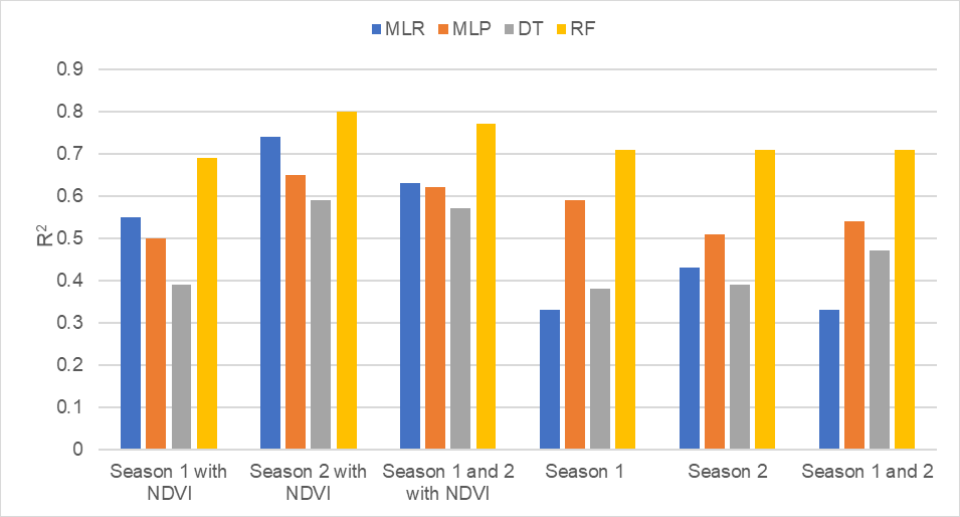


Figure 21: Comparison of the performance of machine learning algorithms for season 1 (2019/2020) and season 2 (2020/2021) and combining the data from the two seasons with and without NDVI (MLR: multiple linear regression, MLP: multilayer perceptron, DT: decision tree, RF: random forest).

The model performance was further evaluated using the MAPE and RMSE, as shown in Table 11. Similar to the R^2 analysis, the inclusion of NDVI reduced the error of the ML predictions, as indicated by the low MAPE and RMSE values with NDVI. The MAPE and RMSE accuracy trends displayed inconsistencies during the 2019/2020 season, with lower values observed for MLP and RF models without NDVI. Although model accuracy varied depending on the ML model and dataset used, the DT was the least accurate model, with MLP and MLR scoring reasonable accuracies, and RF was the most accurate model when NDVI was included in the training and testing data. The MLR was the least accurate, DT and MLP were reasonably accurate, and RF was the most accurate ML model without NDVI data included in the training and testing datasets. The lower accuracy of the MLR model in predicting subfield yields without NDVI data emphasizes the difficulty of capturing the intricate and nonlinear

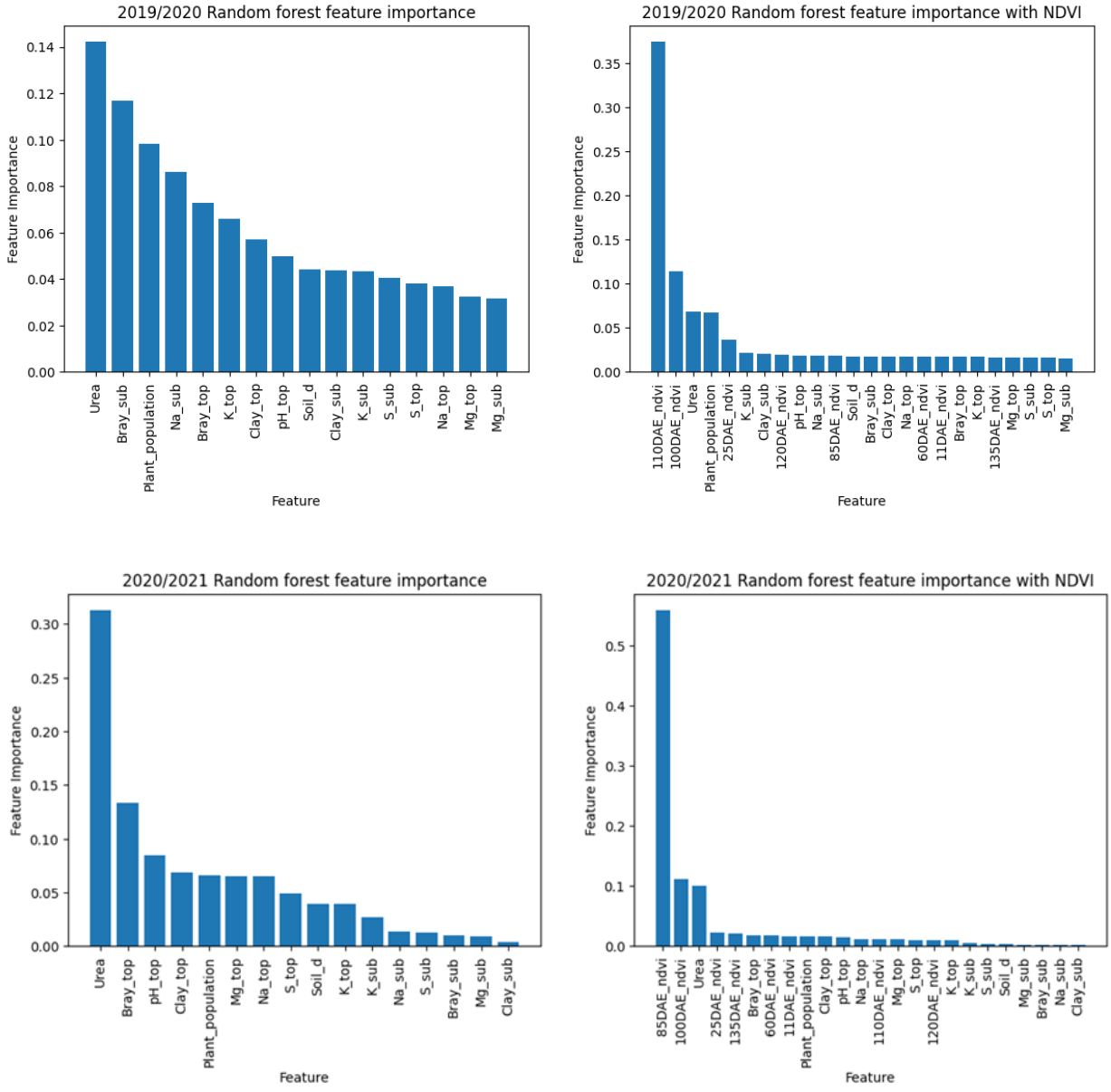
connections present in the dataset. The inclusion of NDVI data improved the initially poor prediction accuracy of the MLR model more than that of the other models, indicating that NDVI data can enhance the predictive performance of the model. The results of this study indicate that although varying ML techniques are applicable for sub-field maize yield predictions, the accuracy of these models can be influenced by the specific datasets used and may vary with temporal changes between successive seasons. Incorporating NDVI data into the training and testing datasets led to improved model predictions and a decrease in prediction errors in all models, although the extent of these improvements varied among the different models and datasets used. The application of ML models in two distinct seasons yielded different prediction accuracy results for all models, although there was consistency in the overall comparison between the models. This could be an indication of the shift in the impact of temporal and spatial variables on yield between seasons (Kravchenko and Bullock, 2000), and these interactions are more important in rain-fed cropping systems.

Table 11: Statistical analysis comparison of machine learning regression models on DIFM trial Uitsny maize field for 2019/2020, 2020/201, and combined dataset with and without NDVI evaluated using the 80/20 training and testing analysis (MAPE: mean absolute percentage error, RMSE: root mean square error).

Season	ML algorithm	MAPE (%)	RMSE (t ha^{-1})
With NDVI data			
2019/2020	MLR	6.7	0.83
	MLP	6.5	0.85
	DT	7.4	0.96
	RF	5.4	0.69
2020/2021	MLR	9.8	1.08
	MLP	10.5	1.22
	DT	12.0	1.35
	RF	8.4	0.95
Combined seasons	MLR	8.7	1.00
	MLP	7.7	0.93
	DT	8.7	1.10
	RF	6.6	0.81
Without NDVI data			
2019/2020	MLR	8.4	1.01
	MLP	6.2	0.79
	DT	7.4	0.97
	RF	5.3	0.66
2020/2021	MLR	14.4	1.59
	MLP	12.4	1.40
	DT	13.3	1.65
	RF	10.0	1.14
Combined seasons	MLR	11.6	1.36
	MLP	7.9	0.98
	DT	8.9	1.21
	RF	7.0	0.89

The important factors for maize yield predictions

The RF model was utilized not only to assess the predictive abilities of the models but also to evaluate the interaction between seeding and fertilizer rate combinations and various attributes within each plot to ascertain the most influential attributes for yield prediction (Figure 22). As the RF model emerged as the top performer, this section delves into the implications of important variables extracted from that model. Although the relative impact of a solitary variable cannot be measured independently of other variables, a measure for assessing the relative importance of factors on prediction outcomes was provided.



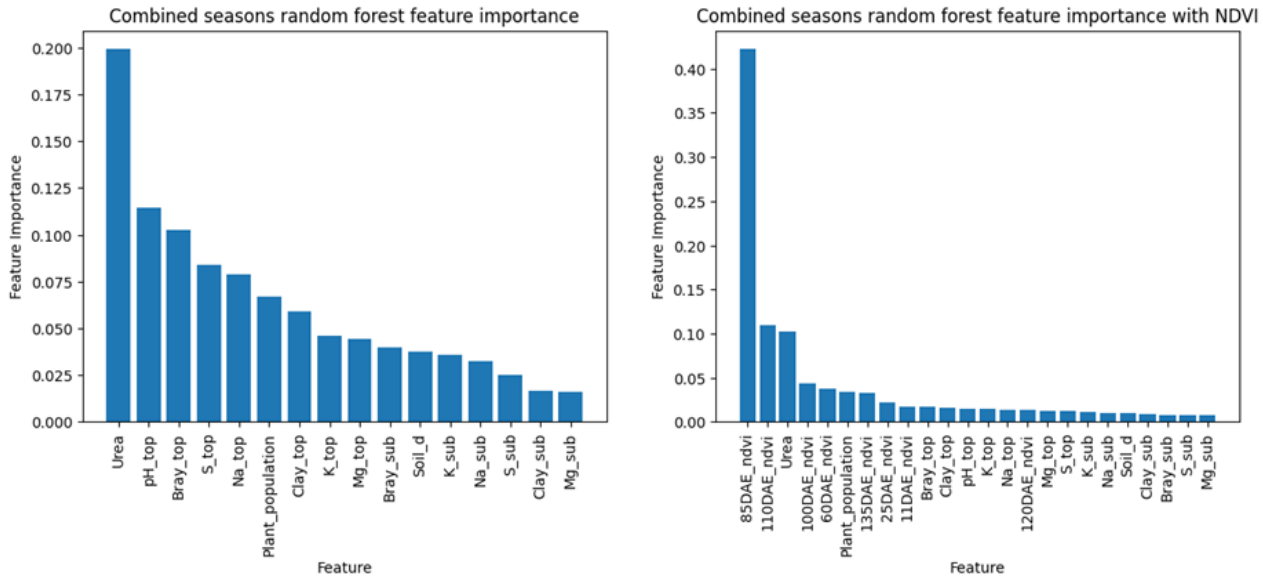


Figure 22: Feature importance from the random forest for 2019/2022, 2020/2021, and the combination of the two-season data with and without NDVI using the 80/20 training and testing analysis (Plant_pop: plant population, Urea: urea application, ph_top: soil pH in topsoil, bray_top: phosphorus in topsoil, K_top: potassium in topsoil, Mg_top: magnesium in topsoil, Na_top: sodium in topsoil, S_top: sulphur in topsoil, Clay_top: clay content in topsoil, Bray_sub: phosphorus in sub soil, K_sub: potassium in sub soil, Mg_sub: magnesium in sub soil, Na_sub: sodium in sub soil, S_sub: Sulphur in sub soil, Clay_sub: Clay content in sub soil, Soil_d: soil depth).

The variable importance generated from the RF model on data without NDVI indicated that urea application and plant population, which are variables that the farmer can more easily control, explained 24, 40 and 27% of yield variation in the first, second, and combined season data. The inclusion of NDVI data in model training resulted in urea application and plant population explaining 15, 13 and 14% of yield variation, respectively, for the three datasets used. Despite the changes in variables (with and without NDVI) used for model development, soil pH and clay content in the topsoil consistently explained 5-6% and 2-3% of the yield. The NDVI variables had a stronger impact on maize yield predictions than the soil and management variables. This explains the improved model performance when NDVI data are included, as indicated by the higher prediction accuracies in all ML models used. Specific NDVI measurements on different days after emergence demonstrated a significant explanatory power. For example, NDVI at 110 DAE in the 2019/2020 season explained 38% of the yield variation, while NDVI at 85 DAE in both the 2020/2021 season and the combined seasons explained 55% and 45% of the yield variation, respectively. The addition of NDVI data can only be crucial for improving the yield prediction accuracy of the models, as this can be an attribute that a farmer cannot control.

The feature importance shows that urea application, plant population, soil pH, and clay content in the topsoil were agronomic management and soil attributes that led to larger information gains on yield variability, whereas subsoil chemical properties explained yield variability the least in the low rainfall season or combined season analysis. In

the wet season, however, urea application, plant population and sub-soil properties (subsoil P, Na, K and clay) explained yield variability the most. The inclusion of NDVI data is important for increasing model prediction capabilities and could be crucial for in-season yield prediction exercises. This study showed that the important features explaining yield variability differed between seasons and were influenced by the inclusion of other variables in the dataset.

5.4. Conclusions

The results showed that despite differences in accuracy, all four ML techniques could be effectively trained and tested on DIFM data to develop models that can be used to estimate sub-field maize yields in a highly heterogeneous field. This could be important for helping farmers and agronomists to estimate the yield profit margins and determine the cost-benefit of an intervention. The RF model was the best for spatial yield predictions using DIFM datasets and the inclusion of NDVI data improved model performance drastically. The RF feature importance statistics provided insights into the variables that most likely limit grain yield and those that can explain yield variability, which may prove to be useful for on-farm decision-making processes. Integrating more within-season data sources, such as NDVI, rainfall distribution in relation to plant growth stage, and timing of split fertilizer application, should be considered in future research because it can potentially improve sub-field yield predictions. Machine learning models in agricultural systems require the consideration of variations in weather, soil types, and other environmental factors that can vary in space and time. The DIFM trials can play an important role in advancing the field of precision agriculture by providing valuable insights into the complex interactions between crops, soils, weather, and management practices, and by identifying new opportunities for improving crop yields and environmental sustainability. The approach used in this study can be refined and adapted for application in different farming environments.

6. Digital soil maps for South Africa

Aimee Thomson, Michael van der Laan, Leushantha Mudaly, Garry Paterson

6.1. Introduction

Although major advancements in digital soil mapping (DSM) have led to the increased availability of soil information (Rossiter, 2016), several studies have shown that internationally developed DSMs, or DSMs at global scales, are not highly accurate representations of actual soil data (Rossiter et al., 2021, Bodenstein et al., 2022). As soil-landscape interactions vary at different geographic locations, the methodologies used to create global DSMs may include soil forming processes that might be unsuitable to apply to areas in South Africa (van Zijl, 2019). Additionally, high quality, accessible soil data in actionable form has historically been lacking for southern Africa, and the quality of legacy soil data used to construct the DSMs has a direct effect on DSM accuracy (Paterson et al., 2015, Van Tol and Van Zijl, 2020, Miller et al., 2021).

Previously, the investigation into the accuracy of the SWAT-SA, Innovative Solutions for Decision Agriculture (iSDA), Africa SoilGrids 250 m (AfSG250), Harmonised World Soil Database version 1.2 (HWSD), and SoilGrids-for-DSSAT-10 km (SG-DSSAT) showed that the DSMs had differing ranges of accuracy for a study area (The ARC-Roodeplaat Experimental Farm), but that AfSG250 was the most representative for the soil parameters under study. This study aimed to assess the accuracy of the same five DSMs in a South African context by observing the differences between the measured and DSM-estimated soil parameters across varying regions and soil types. It is hypothesised that, as seen previously, the DSM values will be substantially different to in-field observations. Additionally, locally produced DSMs will be more accurate than internally produced DSMs, but it is likely that the most reliable representation of South African soils will be a combination of the DSMs.

6.2. Materials and Methods

A total of 25 studies were selected for data comparison (Figure 23). The data were sourced from previous studies and Prof Martin Fey's Soils of South Africa (Fey, 2010). Overall, the sites were selected with the intention of having a broad coverage of South Africa as well as ensuring a variety of soil types. Sites occurring in areas with agriculturally significant soils were also favoured.

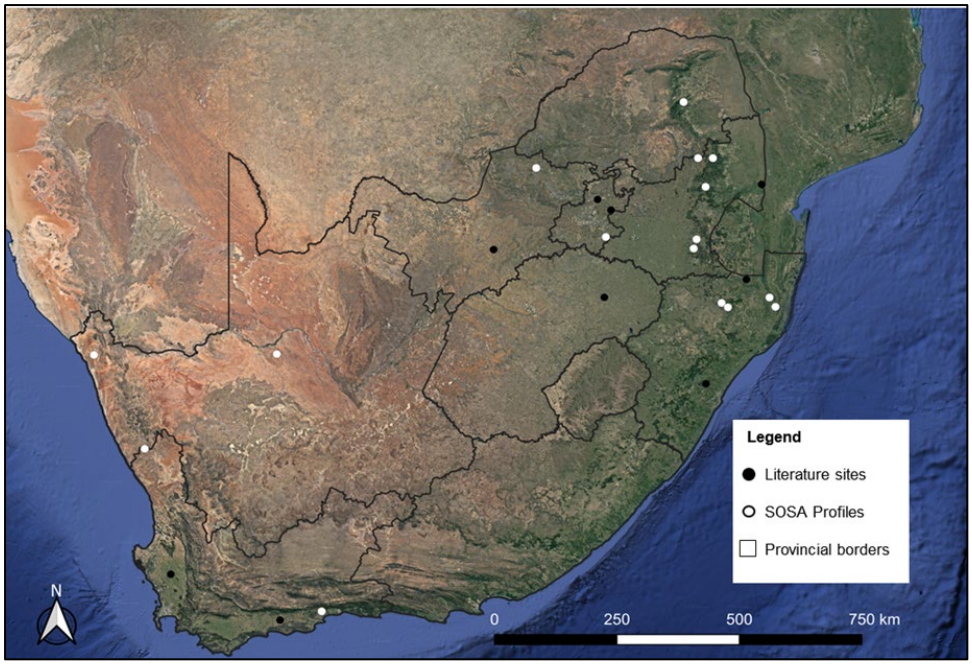


Figure 23: The study area distribution across South Africa, where SOSA refers to sites obtained from Soils of South Africa (Fey, 2010).

Soil data from the respective DSMs under study were collected for each location. To explore the accuracy of the DSMs, the parameters under study were evaluated in independent sections (clay, silt, SOC, pH, and bulk density) by comparing measured data to DSM soil parameter data. In the comparison, graphs representing the measured and DSM parameter values at different soil depths, graphs showing the relative spread of error exhibited by the DSMs, and tables with Root Mean Square Error (RMSE) values were given. The graphs and tables representing all the parameters were extensive and not included in this document in the interest of space. However, the section on silt is given as an example of what the data looked like:

Silt

As seen in Figure 24, the general trend was that DSMs over-estimated silt values, with exception of SWAT-SA, which, in most cases, produced either under-estimates or values very similar to measured values. Although SWAT-SA had silt values that were in some instances very different to the measured values, it remained the most frequently accurate amongst the group. As with clay, iSDA and AfSG250 present similar accuracy both DSMs rarely produced estimates that were far out of range of the measured values. SG-DSSAT performed poorly in general but still produced silt values close to measured values in some instances. HWSD produced silt estimates that were highly erroneous most frequently.

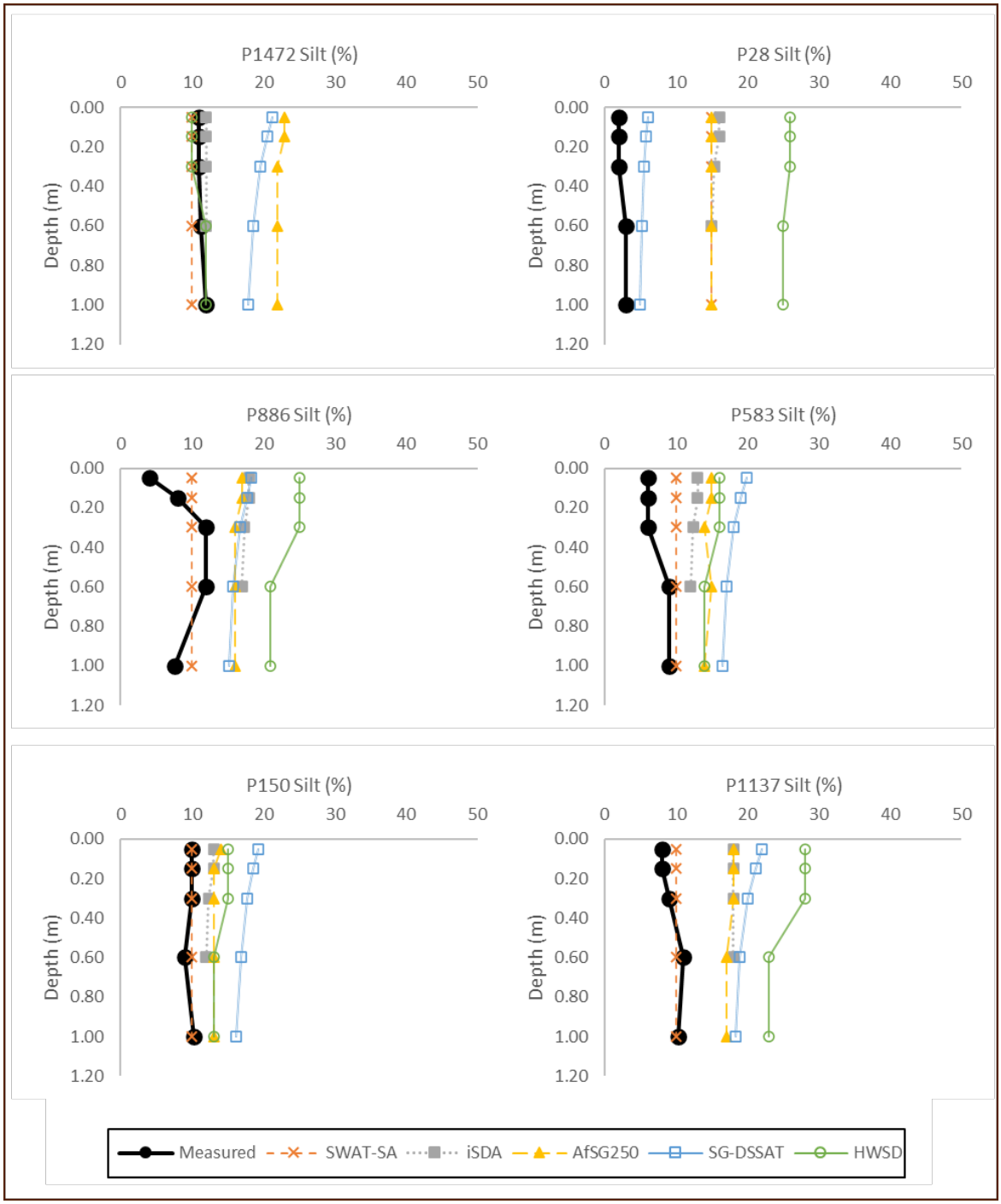


Figure 24: Comparison between measured and estimated silt content (%) for SWAT-SA, Innovative Solutions for Decision Agriculture (iSDA), Africa SoilGrids 250 m (AfSG250), Harmonised World Soil Database version 1.2 (HWSD), and SoilGrids-for-DSSAT-10 km (SG-DSSAT) DSMs at different soil depths. Only six of 25 study sites shown.

The RMSE values seen in Table 12 indicate that SWAT-SA was the most accurate of the DSMs for the first 0.05 m and 0.6-1 m, and iSDA and AfSG250 had similarly accurate values at depth 0.05-0.6 m. Conversely to clay, silt values appeared to become more accurate with increasing depth for all DSMs.

Table 12: The root mean square error (RMSE) across all sites for silt estimates (%) of the respective DSMs under study. A heat map of how well the DSMs perform relative to one another has been included, where green is indicative of predictions with smaller error, and red is indicative of predictions with larger error. (n can be found in Table 14)

Silt RMSE (%)					
Depth (m)	SWAT-SA	iSDA	AfSG250	SG-DSSAT	HWSD
0.00-0.05	7.31	7.62	7.90	10.41	12.01
0.05-0.15	7.29	7.26	7.26	9.59	11.56
0.15-0.30	6.53	6.27	6.51	8.40	11.57
0.30-0.60	6.16	6.14	5.95	7.37	10.34
0.60-1.00	5.66		6.76	7.08	9.45

The general tendency for DSMs to over-estimate silt values can be observed in Figure 25. When considering a range of -20% to 20% deviation from measured silt values, 100% of SWAT-SA, iSDA, AfSG250, and SG-DSSAT datapoints values fell within range, whereas only 91.5% of HWSD datapoints fell within range. HWSD had the largest spread of error when compared to other DSMs, followed by SG-DSSAT and then SWAT-SA. Despite the large range of error seen in SWAT-SA, it had a substantial concentration of estimates around the y-axis, indicating higher accuracy compared to other DSMs and, as shown in *Table 13*, had the most even ratio of under- to over-estimates and also produced the most instances in which silt estimates matched silt measurements. The iSDA and AfSG250 DSMs produced estimates with a relatively low ranges of error.

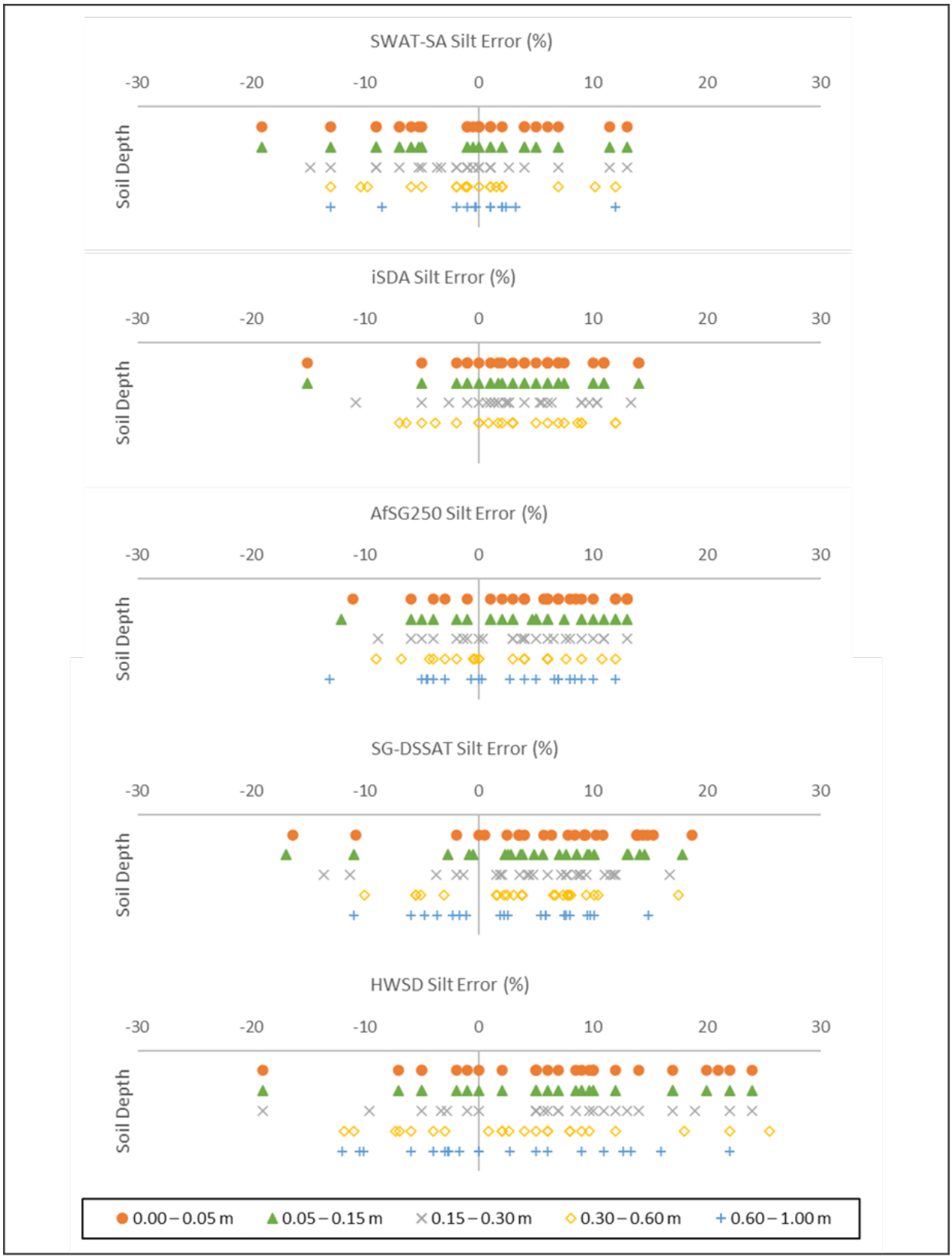


Figure 25: The spread of error at different depths for the various DSM silt content (%) predictions, where at 0%, measured silt = DSM-estimated silt. Thus, values clustered around the y axis reflect less error (n can be found in Table 14).

Table 13: Number of instances in which the respective DSMs produced under-estimates (measured > estimated), over-estimates (measured < estimated), and precise (measured = estimated) silt values

Silt error (%)	SWAT-SA	iSDA	AfSG250	SG-DSSAT	HWSD
Under-estimates	56	19	33	26	34
Over-estimates	44	74	81	92	79
Precise	8	5	4	0	5

Table 14: Number of data points available for comparison between measured silt and estimated silt according to depth for each DSM under study.

	Number of observations for silt comparisons				
Depth (m)	SWAT	iSDA	AfSG250	SG-DSSAT	HWSD
0.00-0.05	25	25	25	25	25
0.05-0.15	25	25	25	25	25
0.15-0.30	25	25	25	25	25
0.30-0.60	20	23	23	23	23
0.60-1.00	13	0	20	20	20
Total	108	98	118	118	118

6.3. Results and Discussion

As can be seen in Table 15, of the five DSMs, SWAT-SA had the lowest average error for silt and bulk density, iSDA had the lowest average error for pH, AfSG250 had the lowest average error for clay, and SG-DSSAT had the lowest average error for SOC. HWSD had the highest error for all parameters except SOC, making it the least accurate DSM of the group. While AfSG250 only had the lowest average error for one parameter, it had the lowest total average error (Table 16). The next most accurate DSM overall was SWAT-SA, followed by iSDA, SG-DSSAT, and lastly HWSD. Also seen in Table 16 is the relative parameter accuracy. The DSMs appeared to predict clay most accurately overall, followed by SOC, pH, silt, then bulk density

Table 15: The normalised root mean square error (nRMSE) values for clay, silt, soil organic carbon (SOC), pH, and bulk density (BD) parameters. The DSMs showing the lowest average accuracy per parameter have been indicated with a bold outline.

Database and Depth	nRMSE (%)				
	Clay	Silt	SOC	pH	BD
SWAT-SA					
0.05 m	19.03	22.76	21.73		28.73
0.15 m	18.81	22.70	21.66		28.73
0.30 m	18.73	23.45	26.54		39.01
0.60 m	23.83	27.53	35.97		36.83
1.00 m	27.23	22.53	25.37		72.17
iSDA					
0.05 m	19.09	23.73	21.23	21.12	80.01
0.15 m	18.87	22.63	21.27	21.97	80.01
0.30 m	17.77	22.52	28.66	23.25	93.51
0.60 m	19.61	27.47	25.20	24.80	66.37
AfSG250					
0.05 m	18.88	24.62	20.45	23.17	66.58
0.15 m	18.19	22.62	18.99	23.19	60.35
0.30 m	17.60	23.39	24.82	23.30	65.52
0.60 m	19.36	26.60	25.94	24.44	34.83
1.00 m	19.27	26.92	28.18	25.90	17.64
SG-DSSAT					
0.05 m	35.27	32.44	19.48	24.45	98.44
0.15 m	21.26	29.89	17.78	24.17	92.03
0.30 m	20.22	30.18	20.93	24.79	95.30
0.60 m	21.23	32.94	21.57	24.96	39.11
1.00 m	21.62	28.18	23.44	26.58	68.89
HWSD					
0.05 m	25.72	37.42	23.15	28.48	100.69
0.15 m	25.84	36.01	23.19	29.34	100.69
0.30 m	25.36	41.57	27.16	31.11	123.53
0.60 m	26.66	46.23	28.26	31.83	72.43
1.00 m	28.16	37.60	24.75	36.82	53.41

Table 16: The average relative root mean square error (RRMSE) displayed by the DSMs for each parameter (SOC = soil organic carbon, BD = bulk density). The overall DSM accuracy is given as well as the overall accuracy at which parameters are predicted.

	RRMSE (%)					
DSM	Clay	Silt	SOC	pH	BD	DSM accuracy
SWAT-SA	21.53	23.79	26.25		41.09	28.17
iSDA	18.83	24.09	24.09	22.79	79.98	33.95
AfSG250	18.66	24.83	23.68	24.00	48.99	28.03
SG-DSSAT	23.92	30.73	20.64	24.99	78.75	35.81
HWSD	26.35	39.77	25.30	31.51	90.15	42.62
Parameter accuracy	21.86	28.64	23.99	25.82	67.79	

It was hypothesised that local DSMs would be more accurate compared to global DSMs. DSMs are more accurate when the processes used to build the DSM are specifically tailored to a region, as soil forming processes (and thus soil properties) change substantially geographically (van Zijl, 2019). SWAT-SA was constructed using soil data from the national Land Type Survey (Land Type Survey Staff 1972-2002) and was developed specifically for use in South Africa, whereas AfSG250 and iSDA were created on a continental scale for use across Africa. Both AfSG250 and iSDA were made using soil information from the Africa Soil Profiles (AfSP) database (amongst others), which is a collection of legacy soil profiles compiled from approximately 540 sources, including the Agricultural Research Council – Institute for Soil, Climate, and Water (ARC-ISCW) (Leenaars et al., 2013). The soil information sources used to supplement the global soil maps SG-DSSAT and HWSD overlap those used to supplement the continental and local DSMs. SG-DSSAT was constructed using information from SoilGrids1km (Hengl et al., 2014), which includes data from AfSP and the Soil and Terrain Southern Africa (SOTERSAF), in combination with HarvestChoice Generic soil profiles (Han et al., 2015), and HWSD was constructed through the harmonisation of data from SOTERSAF, ISRIC-WISE, and the national Land Type Map (Bodenstein et al., 2022). The ARC-ISCW is the custodian of much of the soil information in South Africa, including data acquired from the national Land Type Survey, and is a mutual soil information data source for all the DSMs under study. This means that all the DSMs are linked by the same primary soil observations to some degree. A similar observation was noted by Bodenstein et al. (2022), who acknowledged that known and unknown relationships exist between the DSMs they studied (AfSG250, HWSD, and SOTERSAF). Although the DSMs have inherent data associations, they clearly display different levels of accuracy. This is likely linked to the different DSM techniques and the quality of the covariate data used to create the maps (Hengl et al., 2015, Balla et al., 2016, Zhang et al., 2017, Hengl et al., 2021, Zhi et al., 2022). For example, iSDA and AfSG250 were compiled using similar input soil data, but iSDA makes use of more advanced machine learning techniques and uses higher resolution remote sensing data than AfSG250, allowing for a higher resolution soil map (Hengl et al., 2021).

The larger scale maps (iSDA and AfSG250) were of the most accurate in the group, but despite the small scale of SWAT-SA, it still produced relatively low error compared to the other DSMs. Additionally, SWAT-SA is the only

DSM that had unique soil depth intervals, a factor that influences soil forming processes and soil-water-plant interactions (Suleiman and Ritchie, 2004, Minasny et al., 2016). The importance of SWAT-SA having depth intervals more representative of the South African landscape will be clearer after investigating how sensitive models are to differences in soil depths. Should model outcomes be largely influenced by soil depth, SWAT-SA could ultimately be the DSM of preference but would need to be supplemented with pH data from another DSM (ideally iSDA). The strong suits and performance of SWAT-SA relative to the larger scale maps challenges the common misconception that that higher resolution DSMs are more reliable and accurate (Arrouays et al., 2020) and emphasises the importance of creating DSMs that consider local conditions, as they have thus far proven to be more reliable than continental and global DSMs.

There was a diverse set of results when considering the general representation of soil parameters by the DSMs. There were many cases in which there was a relatively even distribution of under- to over-estimates for clay DSM values. Given this pattern, it might be of interest to explore whether the average clay of the five DSMs outperforms any individual DSM. Silt values were largely over-estimated by the DSMs, with exception of SWAT-SA. This was also noted in the previous chapter and was pinpointed to the fact that the southern hemisphere does not have glacial processes, a process that contributes largely to silt formation, to the same extent that the northern hemisphere does and was likely not accounted for in the DSMs that were not made for use specifically within South Africa (Assallay et al., 1998, Boelhouwers and Meiklejohn, 2002). As with the silt data, all DSMs except for SWAT-SA under-estimated bulk density. These observations make SWAT-SA appear to be a likely candidate for the most reliable DSM to use, but there were still instances for which SWAT-SA did not represent silt and bulk density values accurately, whereas the other DSMs all over-estimated silt and under-estimated bulk density fairly consistently. Would it, then, be more reliable to adjust the respective DSMs according to the average deviation from the measured values, or use the mostly accurate SWAT-SA values?

Bulk density is a scarcely measured soil property in South Africa (Myeni et al., 2021), as illustrated by the limited number of bulk density data points available to be used in this study. This is likely because taking bulk density samples can be time consuming, expensive, and often impractical (Benites et al., 2007, Abdelbaki, 2018). The lack of bulk density data is also shown in the number of training points used to create the iSDA map, where there were approximately 13 500 bulk density samples versus approximately 122 000 samples for clay, silt, and SOC data (Hengl et al., 2021). Estimating bulk density using pedotransfer functions (PTFs) is a potential solution to the lack of data, however PTFs have limited extrapolation value for environments outside of which they were developed to use (Van Tol et al., 2016). In a similar way to silt, bulk density values may be under-estimated by the DSMs (with exception of AfSG250 and SWAT-SA) in general because the processes used to calculate them might not be globally applicable.

The DSMs showed no inclination to over- or under-estimate SOC values, and it was noted that SOC was a parameter that all DSMs produced relatively large error for, with the lowest average deviation from measured values being approximately 0.6% (SG-DSSAT). These differences do not seem substantial at first but considering

the small range that SOC values typically fall into (approximately 0-4% for this dataset), these seemingly small discrepancies are speculated to have substantial impacts on model outputs, as it has been shown that crop modelling systems are sensitive to SOC data (Gasanov et al., 2020). Accumulation and loss of SOC is affected by factors such as climate, land use changes, and agricultural practices and can change over short periods of time (Post et al., 2001). The sites that were assessed were sampled over a long period of time, some of which dated back to as early as 1970. In that period, SOC content was likely influenced by the aforementioned factors, especially agricultural practices as some of the sites were located on experimental farms. DSM data compared to more recent sites (from 2016 onwards) were not more accurate than when compared older sites, although the number of recent sites was limited and thus might not be adequate to make reliable observations from.

Similarly to SOC, pH is a dynamic soil property. Key factors responsible for spatial and temporal variations in pH are mean annual precipitation, clay content, SOC, nitrogen fertilisers, and basic cation leaching/uptake (Uwiragiye et al., 2023). South Africa is a large country with different climatological zones, a rich geology, and has a large area of land under agricultural management. It is therefore expected that South African soils exhibit a wide pH range, with drier areas (western parts of the country) consisting of more alkaline soils and wetter regions (eastern part of the country) consisting of more acidic soils (Courtman et al., 2012). This distribution of acidic and alkaline soils was supported by the regional analysis but was not reflected by the DSMs. It was noted that DSM values were more similar to measured data when measured pH values fell between a range of approximately 5 and 6, often over-estimating pH in more acidic cases and under-estimating pH in more alkaline cases. Given this trend, it was thought that the DSM creators may have restricted the predictions to these typical pH values, however, the DSMs had values that were outside of this range on some occasions, and investigating the processes used to predict pH in each of the DSMs did not reveal conservative approaches. The specific algorithms each DSM used to calculate pH values are not publicly available, but the iSDA and AfSG250 DSMs use machine learning techniques that consider remotely sensed data in combination with measured sample data to predict pH values, while SG-DSSAT and HWSD make use linear statistical models to predict pH values. Given that SOC and pH are relatively dynamic soil parameters and that the data used in this study is quite dated, the feasibility of comparing the historical data to DSM estimates that were calculated using more recent data is questionable, and there is room for further investigation in future studies.

Calculating the normalised root mean square error (nRMSE) values made it possible to directly compare the error between parameters. The DSMs predicted clay with the highest accuracy and bulk density with the lowest accuracy overall. As previously reported, it is hypothesised that non-South African DSMs did not consider local silt forming processes, which might explain why the DSMs predict clay well, but not silt. The low bulk density accuracy may be attributed the infamous low sample availability for the DSMs to be constructed from.

Of the maps investigated, only iSDA and AfSG250 explicitly describe the expected range of error of parameters. Hengl et al. (2021) described iSDA and AfSG250 as having an average data fit (R^2) of 0.8 and 0.6, respectively. RMSE values for individual parameters are also described for AfSG250 and iSDA (Hengl et al., 2017, Hengl et al.,

2021). The error information for SWAT-SA is not directly specified, however, le Roux et al. (2020) used the DSM to parameterise the Soil and Water Assessment Tool ArcSWAT interface to predict streamflow for four catchments in South Africa, and got relatively high R^2 values. Error for the SG-DSSAT DSM was also not directly quantified as the parameter accuracy evaluation relied on the visual comparison of predicted parameters to other soil DSM values (Han et al., 2019). The HWSD also does not give any indication of parameter accuracy.

Comparing the average RMSE values determined in this study to RMSE values described for parameters of the specific DSMs in other studies showed analogous results, despite this study's relatively small sample size. The iSDA RMSE values compared to this study's were: clay 9% versus 13% (study), silt 9% versus 7% (study), SOC 0.03% and 1% (study), and pH 0.5 versus 1 (study) (Hengl et al., 2021). For AfSG250, RMSE values were closer to the study's than iSDA with clay being 14% (13% study), silt 8% (7% study), SOC 1.6% (0.8% study), pH 0.6 (1 study), and bulk density 0.14 kg m^{-3} (0.11 kg m^{-3} study) (Hengl et al., 2015). A study by Bodenstein et al. (2022) might be more telling than the previously mentioned studies as it is, much like this study, a local investigation to DSM accuracy in South Africa. Comparing HWSD and AfSG250 to the historical soil profiles within the South African Profile Database (SAPD) and the African Soil Profile Database (ASPD), Bodenstein et al. (2022) achieved mean RMSE values very similar to those found in this study.

It was noted that for clay, pH, and bulk density there were increases in error deeper in the profile. A possible explanation for these depth trends could be that there is limited data for deeper in the soil profiles. The density of soil observations has been shown to be a key determinant for prediction accuracy (Loiseau et al., 2021), thus having fewer samples for deeper in the soil profile would likely yield higher error in the digital soil maps for those depths. Conversely, the error displayed by the silt and SOC values decreased with increasing profile depth. It was initially thought that this trend could be random, but data provided in the paper by Bodenstein et al. (2022) show depth trends similar to this study for all the parameters. Silt error decreasing with increasing depth could be attributed to modelling approaches, and that silt-forming processes may be simpler deeper in the soil profile. As for SOC values, the associated error might decrease with depth simply because SOC content is typically lower deeper in the soil profile and there is thus less room for error (Schulze and Schütte, 2020, Wang et al., 2022).

The HWSD map had notably poor accuracy for all parameters compared to the other DSMs. Although the HWSD was constructed by the combination and harmonisation of several soil maps, the map that contributed soil information to South Africa specifically is known as 'SOTER for South Africa' (SA-SOTER) (Batjes, 2004). SA-SOTER was harmonised with other southern African soil maps to create SOTER-SAF, which was then harmonised again to create the HSWD. The SA-SOTER map, which the ARC-ISCW was responsible for compiling, could not be constructed using the same methodology as defined in the SOTER methodology, and had an unwanted level of detail (Dijkshoorn, 2003), which may have affected the DSM accuracy.

Although RMSE values similar to other studies were obtained despite a relatively small pool of data to work with, the conclusions drawn from the study would be more reliable with more datapoints. It was particularly challenging to comment on bulk density as well as SOC trends with confidence given the low number of samples to work with.

Additionally, a greater number of datapoints would likely have allowed for more obvious regional trends. Another factor to consider is that SWAT-SA did not have any data for pH and iSDA did not give data for depths over 0.5 m, which could have influenced their overall performance. The standardisation of soil data as well as the fact that the measured data of the sample sites was not analysed using uniform methods (except for the SOSA profiles) may have also contributed to potential data skewing. Another factor worth noting is that some of the sites occurred on experimental farms which were likely disturbed at some point. The cultivation of soils impacts SOC content considerably (Swanepoel et al., 2016) and the measured data might therefore not represent values typically found under undisturbed conditions.

6.4. Conclusions

Values from the DSMs under study are not always representative of soils across South Africa. A wide range of accuracy existed within the DSMs and some DSMs were more accurate than others for certain parameters. Although all the DSMs except for HWSD had individual strengths, AfSG250 was the most accurate in general. The choice of which DSM to use ultimately depends on user requirements. In time-constrained circumstances, it might be best to use AfSG250, otherwise integration of the different datasets according to their respective strengths would likely be the most optimal solution. For example, one might use pH information from iSDA, clay data from AfSG250, silt data from SWAT-SA, SOC from SG-DSSAT, and potentially use bulk density estimates for 0-0.3 m from SWAT-SA and AfSG250 estimates for 0.3-1.0 m.

The results obtained showed that DSMs created on a global and, perhaps to a lesser degree, continental scale often lose their utility in the heterogenous South African escarpment, emphasising the importance of both investigating map accuracy in the area of interest before application as well as calling for the creation/improvement of DSMs local to South Africa. However, it has been shown that despite the error associated with DSMs, they facilitate the significant improvement of existing soil surveys that have missing or incomplete data (Mora-Vallejo et al., 2008). DSMs of varying resolution are thus still valuable in data scarce countries, such as South Africa.

7. Transformer-based Neural Machine Translation for Native South African Languages

Pitso Walter Khoboko, Vukosi Marivate, Joseph Sefara, Michael van der Laan, Michael Silberbauer

7.1. Introduction

Translating text using a human translator can cost a lot of money and consume a lot of time, thus a lot of research is being conducted to create Natural Language Processing (NLP) models that can translate text in large quantities at high quality and speed. Transfer learning in Neural Machine Translation (NMT) models have shown to be able to attain quality scores on BLEU (Bilingual Evaluation Understudy Score) evaluation metric for low-resourced corpora (Lakew et al., 2018, Aji et al., 2020, and Zhang and Zong, 2016). A large resourced parallel corpus, which is a dataset that has large number of parallel sentences of source and target languages, is used to train the first NMT model in transfer learning settings. The previously mentioned model is referred to as the parent model. Then a low-resourced language, which have limited bilingual corpora, is used to fine-tune, by training from learned parent model's parameters. The resulting model is called the child model (Ostling and Tiedemann, 2017). Fine tuning Transformer based NMT models has improved the BLEU metric score when compared to baseline NMT models on low-resourced languages, especially in multilingual-models' settings for European low-resourced languages (Mohamed et al., 2019 and LIU et al., 2020). Two machine translation ideals have been conducted in this research report, namely 1) translation tasks on low-resourced parallel corpora of some of native South African languages and 2) learning technique for translations of extremely low-resourced agriculture domain specific parallel corpora. The WRC and ARC datasets contain valuable information that could help smallholder farmers in South Africa who may not fully understand the literature in English. This language barrier can prevent these farmers from applying scientific knowledge to increase their crop yields. The guidelines provide solutions for maximizing crop yields in the current South African climate.

Most of the South African smallholder farmers in the rural areas still use farming methods that have been passed down to them by their forefathers. However, climate change has affected ancient methods of farming and lack of information for smallholder farmers hinder their adaptation capacity to climate change (Benhin, 2006). In the rural areas do have job opportunities, as result crop farming is the only way people can get food and money. It has been reported in the annual reports of Department of Rural Development and Land Reform (DRDLR) that the current climate change has negatively affected the rain pattern and the land in the rural areas (Thinda, 2020), thus smallholder farmers' crop yield production is being negatively impacted. Small-holder farmers are left without knowledge of knowing when the best time is to plant crops. The Agriculture Research Council and Water Research Council have been working with smallholder farmers in terms of providing translated guidelines on how to adapt crop production to climate change in the farmers' native languages, however these translated guidelines are limited and provide extremely low parallel corpora to train on a bilingual NMT model. This is supported by (Ahmadnia,

2019) who have highlighted in their paper that bilingual NMT require a significant number of bilingual corpora to learn. In this research report we have fine-tuned multilingual BART in efforts of automatically translating more of these guidelines from English to native South African languages – isiZulu, isiXhosa and Sepedi, respectively. The translated guidelines may aid smallholder farmers in increasing their yield, as they will gain insights of the negative impact of climate change on rainfall patterns (Benhin, 2006).

Fine tuning a mBART(multilingual Bidirectional Auto-Regressive Transformers) model has successfully been used to translate language pairs that have low parallel corpora in European settings (Liu et al., 2020). For this research report, mBART has been fine-tuned to translate low parallel corpora for the prior mentioned South African languages, and further fined tuned and trained on extremely low resourced agriculture domain dataset for the same languages. We have also used the joeyNMT minimal toolkit model to perform the same process of WMT22 (Jiao, 2022), and OPUS (Tiedemann, 2020) datasets for the parent models. And for the child model we have used extracted sentences from the farming gridlines we have acquired from the WRC and ARC. for 11 official South African languages. The performance of parent and child models, respectively for JoeyNMT and mBART models have been compared. On average BLEU score of all created mBART models have beaten the JoeyNMT models beside the one for Sepedi language which got 4 times more BLEU score than the mBART model.

7.2. Materials and Methods

In this research, we used a collected bilingual corpus from Autshumato, WTM22, and OPUS datasets to translate between English and the native South African languages of Sepedi, isiXhosa, and isiZulu. We also manually extracted parallel corpora from guideline documents acquired from the WRC and ARC for these languages. Our goal was to use two distinct Transformer models to translate an extremely low domain-specific dataset. To achieve this, we used a transfer learning technique where a parent model was trained on combined bilingual corpora from various domains, and a child model was trained on both the combined bilingual dataset and the extremely low domain-specific dataset.

We conducted experiments where mBART was fine-tuned and JoeyNMT was customized to create parent models in supervised settings for English-Sepedi, English-isiXhosa, and English-isiZulu. We then fine-tuned the child models of mBART and JoeyNMT to be trained on the WRC and ARC domain-specific datasets using learned parameters from the pre-trained parent models. Both models are Transformer-based, as this type of NMT architecture has shown state-of-the-art performance for low-resourced languages.

We evaluated our models using the BLEU metric and compared their scores to determine which one performed better for our datasets. This research report presents our experimental methodology and findings on the effectiveness of our proposed models in translating our datasets.

7.3. Results and discussion

In previous work, a multilingual Transformer model was compared to a joeyNMT model. The results showed that the Transformer model outperformed joeyNMT for all languages except isiZulu (Sefara et al., 2021). Our results show that our mBART parent model achieved a higher BLEU score than all the bilingual joeyNMT parent models except for Sepedi (Table 17). The Sepedi bilingual joeyNMT model achieved a BLEU score of 25.34, which was 4 times higher than the scores for isiZulu and isiXhosa, which were 6.28 and 5.12, respectively. The mBART parent model achieved a BLEU score of 6.54 for all languages combined.

Table 17: BLEU score results of joeyNMT and mBART parent models before being fine-tuned with an agriculture domain dataset.

	BLEU score		
Model	en-zul	en-xh	en-nso
joeyNMT	6.28	5.12	25.34
mBART	6.54		

The BLEU scores for all child models decreased significantly, by an average of more than 90%. The Sepedi joeyNMT child model achieved the highest BLEU score of 1.06, which was higher than both the mBART child model and all other joeyNMT child models (Table 18).

Table 18: BLEU score results of joeyNMT and mBART child models before being fine-tuned with an agriculture domain dataset.

	BLEU score		
Model	en-zul	en-xh	en-nso
joeyNMT	0.071	0.13	1.06
mBART	3.33		

Table 19 shows the translation quality of the parent models when translating sentences from the agriculture domain. The quality is measured by how closely the machine-generated translation matches the words and context of the reference sentence. For the joeyNMT parent models, the isiXhosa model had the highest translation quality, followed by isiZulu, while Sepedi had the lowest quality. This may be due to the presence of many loanwords from English in the Sepedi dataset, which can make it difficult to produce high-quality translations.

Table 19: Examples of joeyNMT model translations for an agriculture domain sentence before being fine-tuned to the agriculture domain dataset.

Bilingual Pair	Translations: JoeyNMT
en-zul	source: these ditches increase access to and availability of water in intensive food production reference: lemisele ikhuphula ukufinyelela nokutholakala kwamazi uma kukhiqizwa ukudla nokuningi endaweni encane hypothesis: ≪unk≫ ukufinyelela kwamanzi nase≪unk≫ ekukhiqizni kwamanzi
en-xh	source: these ditches increase access to and availability of water in intensive food production reference: le misele yandisa ukufikeleleka ukuya, kunye nokufumaneka kwa manzi ekuveliseni ukutya kanzulu hypothesis: ezi zixhobo zifikeleka kwa zifikelele ekudleni kwamanzi
en-nso	source: these ditches increase access to and availability of water in intensive food production reference: doforo tse oketsa phihlelelo ya le go hwetsagala ga meetse mo hlokegong ya go tsewletsa ga dijo. hypothesis: se se ra meetse le

The mBART parent model had good translation quality for isiZulu and isiXhosa, but poor quality for Sepedi. This is consistent with the issues we previously highlighted with the Sepedi dataset.

When we compare the child models to their parent models, we see that their BLEU scores decreased significantly. However, in terms of translation quality for sentences from the agriculture domain, the child models performed slightly better than their parent models. The mBART child model had higher translation quality than all the joeyNMT child models, but it failed to translate any English sentences into any of the South African languages.

Our results show that a high BLEU score does not always mean good translation quality. This supports previous research that has identified issues with using the BLEU metric to evaluate machine translation models (Wołk and Marasek, 2015).

Table 20: Examples of joeyNMT model translations for agriculture domain sentence after being fine-tuned for the agriculture domain dataset.

Bilingual Pair	Translations: JoeyNMT
en-zul	source: these ditches increase access to and availability of water in intensive food production reference: lemisele ikhuphula ukufinyelela nokutholakala kwamazi uma kukhiqizwa ukudla nokuningi endaweni encane

	hypothesis: ≪unk≫ ukufinyelela kwamanzi nase ≪unk≫ ekukhiqizni kwamanzi
en-xh	source: these ditches increase access to and availability of water in intensive food production reference: le misele yandisa ukufikeleleka ukuya, kunye nokufumaneka kwa manzi ekuveliseni ukutya kanzulu hypothesis: ezi zixhobo zifikeleka kunye nukusetyenziswa kwimveliso yokutya
en-nso	source: these ditches increase access to and availability of water in intensive food production reference: doforo tse oketsa phihlelelo ya le go hwetsagala ga meetse mo hlokegong ya go tsweletsa ga dijo. hypothesis: yo le changed a la dijo

Table 21: Examples of mBART model translations for agriculture domain sentence before being fine-tuned for the agriculture domain dataset.

Bilingual Pair	Translations: mBart
en-zul	source: these ditches increase access to and availability of water in intensive food production reference: lemisele ikhuphula ukufinyelela nokutholakala kwamazi uma kukhiqizwa ukudla nokuningi endaweni encane hypothesis: Le nqubo ikhula ukusetshe nziswa nokusetshenz
en-xh	source: these ditches increase access to and availability of water in intensive food production reference: le misele yandisa ukufikeleleka ukuya, kunye nokufumaneka kwa manzi ekuveliseni ukutya kanzulu hypothesis: Ezi ditshi zikhula kunye nokufumanek a kwemanzi
en-nso	source: these ditches increase access to and availability of water in intensive food production reference: doforo tse oketsa phihlelelo ya le go hwetsagala ga meetse mo hlokegong ya go tsweletsa ga dijo. hypothesis: Nkabe bahu ba tsoga be re tlo bitsa Babirwale

Table 22: Examples of mBART model translations for agriculture domain sentence after being fine-tuned for the agriculture domain dataset.

Bilingual Pair	Translations: JoeyNMT mBart?
en-zul	source: these ditches increase access to and availability of water in intensive food production reference: lemisele ikhuphula ukufinyelela nokutholakala kwamazi uma kukhiqizwa ukudla nokuningi endaweni encane hypothesis: Ngenxa yoko sesit hathu ukuthi ungakwazi ukuthi
en-xh	source: these ditches increase access to and availability of water in intensive food production reference: le misele yandisa ukufikeleleka ukuya, kunye nokufumaneka kwa manzi ekuveliseni ukutya kanzulu hypothesis: Ngenxa yoko sesit hathu ukuthi nagewaza kufanele
en-nso	source: these ditches increase access to and availability of water in intensive food production reference: doforo tse oketsa phihlelelo ya le go hwetsagala ga meetse mo hlokegong ya go tsweletsa ga dijo. hypothesis: O tlo gabotse go batho ba batho ba tlo gabotse

This report has discussed the ability of the machine learning model, mBART, to translate between English and some South African languages. The report shows that mBART can be used with low-resourced languages as the parent models and transfer their learned parameters to a child model with an extremely low dataset. When compared to another model, joeyNMT, mBART achieved higher translation accuracy for most languages, except for Sepedi. However, the report concludes that a high translation accuracy score does not always mean good translation quality.

In future work, we plan to obtain more translated datasets in the agriculture domain. Our results show that our transfer learning technique has the potential to translate sentences and datasets in this domain, even when the available data is extremely limited. We also plan to implement the mBART model from scratch to improve its BLEU score and evaluate its translation quality using not only the BLEU metric but also the METEOR metric. Additionally, for both the joeyNMT and mBART parent models, we will use datasets from a specific domain rather than from various domains, as we have done in this study, to see how this affects the BLEU score and translation quality.

7.4. Conclusion

We have shown that mBART is able to translate between English some of the South African language other than isiXhosa that is already pre-trained with it. We have further shown that transfer learning technique we are able to start with low-resourced languages as the parent models and transferring their learned parameters to a child with extremely low dataset. When comparing mBART model to the joeyNMT models, mBART got a higher BLEU than

all of the joeyNMT model beside the one for Sepedi. Mean mBART's child model got higher BLEU score than all the joeyNMT's child models. In regard to translation the former failed to give good translations and latter was able give fairly good translation for the agriculture domain sentence. However, we have concluded that a high BLEU score does not always mean good translation quality.

For future work we intend to get more translated agriculture domain dataset as our results show there is potential of our transfer learning technique to be able to translate this domain's sentence and dataset for it was extremely low. Furthermore, we intend to code mBART from scratch to increase its BLEU score and not only evaluate it on the BLEU metric alone but also the METEOR 18 to check the quality of its translation. For both joeyNMT and mBART parent models we will acquire dataset from a specific to domain and not from various domains like we have done to see how that well the models will affect the BLEU score and quality of translation.

8. Conclusions

Michael van der Laan, Cindy Viviers, Simphiwe Maseko, Christiaan Schutte, Aimee Thomson, Pitso Khoboko, Michael Silberbauer, Jay le Roux, Leushantha Mudaly, Harold Weepener, Gerrit Hoogenboom, Srinivasan Raghavan, Richard Kunz, David Clark

Recognising the need to strategically coordinate data collection and archiving, South Africa's Water Research Commission (WRC) has developed a cloud-based big data platform called the Water Research Observatory (WRO). A centralised and secure platform for water data asset storage, analytics and visualisation is expected to have major benefits for research in the form of increasing research efficiency, an advanced ability for new projects to build on previous research. Enabling data processing, analytics, and modelling in the cloud means that institutions no longer need powerful hardware and expensive software to achieve high quality research, effectively levelling the playing field between all teaching and research institutions across South Africa.

The flexibility of the Google Cloud Platform (GCP) and other similar platforms allows very high levels of data acquisition and interoperability. Although beyond the scope of this project, data from citizen science and social media can now be ingested and utilised in ways that promote the equitable and sustainable use of water in South Africa.

Real-time dashboards driven by reliable data can enhance decision-making at various levels. It can also assist in ensuring that hydrological data is made available to the public as stipulated in South African legislation. Users are encouraged to integrate tools and develop new ones that promote data democratisation. Practitioners in the sector are strongly encouraged to register and uploading valuable datasets that have been collected over their careers.

Examples of the use of big data analytics and machine learning were done for streamflow and groundwater level prediction, precision agriculture, and language translation case studies. Unique insights and answers were gained that would not have been possible without these tools, and the potential of artificial intelligence is remarkable.

Ongoing work will be to ensure the platform remains secure as its use and value increases, that a data quality control system is implemented, and that new technology is harnessed in such a way that it remains financially viable over the long term as data volumes and processing and application requirements grow.

9. References

ABADI M, AGARWAL A, BARHAM P, BREVDO E, CHEN Z, CITRO C, CORRADO GS, DAVIS A, DEAN J and DEVIN M (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv 1. https://doi.org/10.48550/arXiv.1603.04467.

ABDELBAKI AM (2018) Evaluation of pedotransfer functions for predicting soil bulk density for US soils. *Ain Shams Engineering Journal* 9 (4) 1611-1619.

ADDOR N, NEWMAN AJ, MIZUKAMI N and CLARK MP (2017) The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences* 21 5293-5313. https://doi.org/10.5194/hess-21-5293-2017.

ALGHAFLI K, SHI X, SLOAN W, SHAMSUDDUHA M, TANG Q, SEFELNASR A and EBRAHEEM AA (2023) Groundwater recharge estimation using in-situ and GRACE observations in the eastern region of the United Arab Emirates. *Science of the Total Environment* 867 161489. http://dx.doi.org/10.1016/j.scitotenv.2023.161489

ALHAM A, NIKOLAY B, KENNETH H and RICO S (2002) In neural machine translation, what does transfer learning transfer? Association for Computational Linguistics.

ALI S, LIU D, FU Q, CHEEMA MJM, PHAM QB, RAHAMAN MM, DANG TD and ANH DT (2021) Improving the Resolution of GRACE Data for Spatio-Temporal Groundwater Storage Assessment. *Remote Sensing* 13 3513. https://doi.org/10.3390/rs13173513

ALLAN RP (2021) Climate Change 2021: The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press. https://doi.org/10.1017/9781009157896.

ANDERSON S and RADIĆ V (2022) Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling. *Hydrology and Earth System Sciences* 26 795-825. https://doi.org/10.5194/hess-26-795-2022.

ANDUALEM TG, DEMEKE GG, AHMED I, DAR MA and YIBELTAL M (2021) Groundwater recharge estimation using empirical methods from rainfall and streamflow records. *Journal of Hydrology: Regional Studies* 37 100917.

ARROUAYS D, MCBRATNEY A, BOUMA J, LIBOHOVA Z, RICHER-DE-FORGES AC, MORGAN CL, ROUDIER P, POGGIO L and MULDER VL (2020) Impressions of digital soil maps: The good, the not so good, and making them ever better. *Geoderma Regional* 20 e00255.

ASSALLAY A, ROGERS C, SMALLEY I and JEFFERSON I (1998) Silt: 2-62 μm, 9-4φ. *Earth-Science Reviews* 45(1-2) 61-88.

ATKINSON PM (2003) Downscaling in remote sensing. *International Journal of Applied Earth Observation and Geoinformation* 22 106-114. https://doi.org/10.1016/j.jag.2012.04.012.

BALLA D, VARGA O and ZICHAR M (2016) Accuracy assessment of different soil databases concerning WRB reference soil groups. *Landscape & Environment* 10 1-12.

BARUA S, CARTWRIGHT I, DRESEL PE and DALY E (2020) Using multiple methods to understand groundwater recharge in a semi-arid area. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-2020-143.

BASHEER I A, HAJMEER M. (2000) Artificial neural networks: fundamentals, computing, design, and application. *Journal of Microbiological Methods* 43 3-31.

BASSO B, FIORENTINO C, CAMMARANO D, SCHULTHESS U. (2016) Variable rate nitrogen fertilizer response in wheat using remote sensing. *Precision Agriculture* 17 168-182.

BATJES NH (2004) SOTER-based soil parameter estimates for Southern Africa. Wageningen: ISRIC-World Soil Information.

BECK HE, ZIMMERMANN NE, MCVICAR TR, VERGOPOLAN N, BERG A and WOOD EF (2018) Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Nature Scientific Data* 5 180214. 10.1038/sdata.2018.214.

BENITES VM, MACHADO PL, FIDALGO EC, COELHO MR and MADARI BE (2007) Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. *Geoderma* 139 (1-2) 90-97.

BENYAMIN, A and DORR B (2019) Bilingual low-resource neural machine translation with round-tripping: The case of PersianSpanish. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (RANLP 2019), pp. 18-24.

BEVEN KJ (2011) Rainfall-runoff modelling: the primer, Sussex, UK, John Wiley & Sons. ISBN: 1119951011.

BISHOP CM and NASRABADI NM (2006) Pattern recognition and machine learning, New York, NY, Springer. ISBN: 978-0-387-31073-2.

BODENSTEIN D, CLARKE C, WATSON A, MILLER J, VAN DER WESTHUIZEN S and ROZANOV A (2022) Evaluation of global and continental scale soil maps for southern Africa using selected soil properties. *Catena* 216 106381.

BOELHOUWERS J and MEIKLEJOHN K (2002) Quaternary periglacial and glacial geomorphology of southern Africa: review and synthesis: Periglacial and Permafrost Research in the Southern Hemisphere. *South African Journal of Science* 98(1) 47-55.

BREDENKAMP DB, VAN DER WESTHUIZEN C, WIEGMANNS FE and KUHN CM (1986) Groundwater supply potential of dolomite compartments west of Krugersdorp. Technical Report GH3440. Directorate Geohydrology. Department of Water Affairs and Forestry, Pretoria.

BREIMAN L (2001) Random forests. *Machine Learning* 45 5-32.

BULLOCK DS, BOERNGEN M, TAO H, MAXWELL B, LUCK J D, SHIRATSUCHI L, PUNTEL L and MARTIN NF (2019) The data-intensive farm management project: Changing agronomic research through on-farm precision experimentation. *Agronomy Journal* 111 2736-2746.

CAO G, HAN D and SONG X (2013) Evaluating actual evapotranspiration and impacts of groundwater storage change in the North China Plain. *Journal for Hydrological Processes* 28 1797-1808. https://doi.org/10.1002/hyp.9732.

CHADALAWADA J, HERATH H and BABOVIC V (2020) Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. *Water Resources Research* 56 e2019WR026933. https://doi.org/10.1029/2019WR026933.

CHEN JL, WILSON CR, TAPLEY BD, SAVE H, CRETAUX JF (2017) Long-term and seasonal Caspian Sea level change from satellite gravity and altimeter measurements. *Journal of Geophysical Research: Solid Earth* 122(3) 2274-2290. https://doi.org/ 10.1002/2016JB013595.

CHEN L, HE Q, LIU K, LI J and JING C (2019) Downscaling of GRACE-derived Groundwater Storage based on the random forest model. *Remote Sensing* (11) 2979. doi:10.3390/rs11242979

CHEN X, SONG J, WANG W. (2010) Spatial variability of specific yield and vertical hydraulic conductivity in a highly permeable alluvial aquifer. *Journal of Hydrology* 388 3-4. https://doi.org/10.1016/j.jhydrol.2010.05.017.

CHO K, VAN MERRIËNBOER B, BAHDANAU D and BENGIO Y (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv.* https://doi.org/10.48550/arXiv.1409.1259.

CLEMMENS A and MOLDEN D (2007) Water uses and productivity of irrigation systems. *Irrigation Science* 25 247-261.

COURTMAN C, VAN RYSSEN J and OELOFSE A (2012) Selenium concentration of maize grain in South Africa and possible factors influencing the concentration. *South African Journal of Animal Science* 42(5) 454-458.

DE BRUIN K, RADEMAN Z and TOWERS L (2023) Guidance document for management of a groundwater scheme. Water Research Commission (WRC) Report No. TT 906/22. ISBN 978-0-6392-0374-4. Pretoria, South Africa.

DEPARTMENT OF FORESTRY, FISHERIES AND THE ENVIRONMENT [DFFE] (2021) South African National Land Cover (SANLC) 2020. Available online: https://egis.environment.gov.za/sa_national_land_cover_datasets

DIJKSHOORN J (2003) SOTER database for Southern Africa. International Soil Reference and Information Centre, Wageningen, The Netherlands.

DRUMMOND S T, SUDDUTH K A, JOSHI A, BIRRELL S J and KITCHEN NR (2003) Statistical and neural methods for site-specific yield prediction. *Transactions of the American Society of Agricultural Engineers* 46 5.

DU PLESSIS J and KIBII J (2021) Applicability of CHIRPS-based satellite rainfall estimates for South Africa. *Journal of the South African Institution of Civil Engineering* 63 43-54. http://dx.doi.org/10.17159/2309-8775/2021/v63n3a4.

DUNN OJ (1964) Multiple comparisons using rank sums. *Technometrics* 6 241-252. 10.1080/00401706.1964.10490181.

DWS (2021) National State of Water Report for South Africa – Hydrological year 2019/20. Pretoria, South Africa.

ENGELBRECHT F, MCGREGOR J and ENGELBRECHT C (2009) Dynamics of the Conformal-Cubic Atmospheric Model projected climate-change signal over southern Africa. *International Journal of Climatology: A Journal of the Royal Meteorological Society* 29 1013-1033. https://doi.org/10.1002/joc.1742.

ERIKSSON PG, ALTERMANN W and HARTZER FJ (2009) The Transvaal Supergroup and its precursors. In: Johnson, M.R.; Anhaeuser CR and Thomas RJ. The Geology of South Africa. The Geological Society of South Africa, Johannesburg, pp. 237-260.

FAN H, JIANG M, XU L, ZHU H, CHENG J and JIANG J (2020) Comparison of long short-term memory networks and the hydrological model in runoff simulation. Water, 12 175. https://doi.org/10.3390/w12010175.

FENG W, ZHONG M, LEMOINE J, BIANCALE R, HSU H and XIA J (2013) Evaluation of groundwater depletion in North China using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based measurements. *Water Resources Research* 49 2110-2118. doi:10.1002/wrcr.20192.

FEY M (2010) Soils of South Africa. Cambridge University Press, Cape Town, South Africa.

FUNK C, PETERSON P, LANDSFELD M, PEDREROS D, VERDIN J, SHUKLA S et al. (2015) The climate hazards infrared precipitation with stations: A new environmental record for monitoring extremes. *Scientific Data* 2 1-21. doi: 10.1038/sdata.2015.66.

FUNK C, PETERSON P, LANDSFELD M, PEDREROS D, VERDIN J, SHUKLA S, HUSAK G, ROWLAND J, HARRISON L and HOELL A (2015) The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data* 2 1-21. https://doi.org/10.1038/sdata.2015.66.

GAFFOOR Z, GRITZMAN A, PIETERSEN K, JOVANOVIC N, BAGULA A and KANYERERE T (2022) An autoregressive machine learning approach to forecast high-resolution groundwater-level anomalies in the Ramotswa/North West/Gauteng dolomite aquifers of Southern Africa. *Hydrogeology Journal* 30 575-600. https://doi.org/10.1007/s10040-021-02439-4.

GAFFOOR Z, PIETERSEN K, JOVANOVIC N, BAGULA A and KANYERERE T (2020) Big data analytics and its role to support groundwater management in the Southern Africa development community. *Water* 2796, 1-28. doi:https://doi.org/10.3390/w12102796

GAFFOOR Z, PIETERSEN K, JOVANOVIC N, BAGULA A, KANYERERE T, AJAYI O and WANANGWA G (2022) A Comparison of Ensemble and Deep Learning Algorithms to Model Groundwater Levels in a Data-Scarce Aquifer of Southern Africa. *Hydrology* 9. https://doi.org/10.3390/hydrology9070125.

GASANOV M, PETROVSKAIA A, NIKITIN A, MATVEEV S, TREGUBOVA P, PUKALCHIK M and OSELEDETS I (2020) Sensitivity analysis of soil parameters in crop model supported with high-throughput computing. Springer, Berlin, Germany, pp. 731-741.

GEMITZI A, KOUTSIAS N and LAKSHMI VA (2021) Spatial downscaling methodology for GRACE Total Water Storage Anomalies using GPM IMERG precipitation estimates. *Remote Sensing* 13. doi.org/10.3390/rs13245149

GHIMIRE S, YASEEN ZM, FAROOQUE AA, DEO RC, ZHANG J and TAO X (2021) Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Scientific Reports* 11 17497. https://doi.org/10.1038/s41598-021-96751-4.

GONZALEZ-SANCHEZ A, FRAUSTO-SOLIS J, OJEDA-BUSTAMANTE W (2014) Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research* 12 313-328.

GOODFELLOW I, BENGIO Y and COURVILLE A (2016) Deep learning, Michigan Institute of Technology press, Michigan, USA. ISBN: 0262337371.

HAN E, INES A and KOO J (2015) Global high-resolution soil profile database for crop modeling applications. *Harvard Dataverse* 1 1-37.

HAN E, INES AV and KOO J (2019) Development of a 10-km resolution global soil profile dataset for crop modeling applications. *Environmental Modelling & Software* 119 70-83.

HENGL T, DE JESUS JM, MACMILLAN RA, BATJES NH, HEUVELINK GB, RIBEIRO E, SAMUEL-ROSA A, KEMPEN B, LEENAARS JG and WALSH MG (2014) SoilGrids1km—global soil information based on automated mapping. *PloS One* 9(8) e105992.

HENGL T, HEUVELINK GB, KEMPEN B, LEENAARS JG, WALSH MG, SHEPHERD KD, SILA A, MACMILLAN RA, MENDES DE JESUS J and TAMENE L (2015) Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PloS One* 10(6) e0125814.

HENGL T, MENDES DE JESUS J, HEUVELINK GB, RUIPEREZ GONZALEZ M, KILIBARDA M, BLAGOTIĆ A, SHANGGUAN W, WRIGHT MN, GENG X and BAUER-MARSCHALLINGER B (2017) SoilGrids250m: Global gridded soil information based on machine learning. *PLoS One* 12(2) e0169748.

HENGL T, MILLER MAE, KRIŽAN J, SHEPHERD KD, SILA A, KILIBARDA M, ANTONIJEVIĆ O, GLUŠICA L, DOBERMANN A, HAEFELE SM, MCGRATH SP, ACQUAH GE, COLLINSON J, PARENTE L, SHEYKHMOUSA M, SAITO K, JOHNSON J-M, CHAMBERLIN J, SILATSA FBT, YEMEFACK M, WENDT J, MACMILLAN RA, WHEELER I and CROUCH J (2021) African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports* 6130.

HEROLD C and BAILEY A (2016) Water Resources of South Africa, 2012 Study (WR2012). Water Research Commission. Pretoria, South Africa.

HOARE J (2023) How is Splitting Decided for Decision Trees? Advanced Analysis. *Machine Learning*. https://www.displayr.com/how-is-splitting-decided-for-decision-trees/

HOCHREITER S and SCHMIDHUBER J (1997) Long short-term memory. *Neural Computation* 9 1735-1780. doi:10.1162/neco.1997.9.8.1735.

HOLLAND M and WIEGMANS F (2009) Geohydrology Guideline Development: Implementation of Dolomite Guideline – Phase 1: Activity 19 & 28. Department of Water Affairs. Project No. 14/14/5/2.

HUGHES D (2004) Three decades of hydrological modelling research in South Africa. *South African Journal of Science* 100 638-642. https://hdl.handle.net/10520/EJC96172.

ISLAM S, SINGH RK and KHAN RA (2015) Methods of Estimating Ground water Recharge. *International Journal of Engineering* Associates 5(2) 6.

ISO Focus+ (2010) (International Standards Organisation) Interoperability. https://www.iso.org/files/live/sites/isoorg/files/news/magazine/ISO%20Focus%2B%20(2010-2013)/en/2010/ISO%20Focus%2B%2C%20February%202010.pdf (accessed 23 November 2023).

JAIN K, KAUSHIK K, GUPTA SK, MAHAJAN S, KADRY S (2023) Machine learning-based predictive modelling for the enhancement of wine quality. Scientific Reports 13(1) 17042. doi: 10.1038/s41598-023-44111-9.

JAKKU E, TAYLOR B, FLEMING A, MASON C, FIELKE S, SOUNNESS C, THORBURN P (2019) "If they don't tell us what they do with it, why would we trust them?" Trust, transparency and benefit-sharing in Smart Farming. *NJAS-Wageningen Journal of Life Sciences* 90 100285.

JAMES G, WITTEN D, HASTIE T, TIBSHIRANI R (2013) *An Introduction to Statistical Learning*, vol. 112. Springer.

JAMES KB (2006) Climate change and South African agriculture: Impacts and adaptation options. Technical report, Centre for Environmental Economics and Policy in Africa, University of Pretoria, South Africa.

JEONG JH, RESOP JP, MUELLER ND, FLEISHER DH, YUN K, BUTLER E E, TIMLIN D J, SHIM K-M, GERBER JS, REDDY VR (2016) Random forests for global and regional crop yield predictions. *PloS One* 11 e0156571.

JET PROPULSION LABORATORY [JPL] (2015) Groundwater: tracking groundwater changes around the world. Gravity Recovery and Climate Experiment (GRACE) Tellus. Available online: https://grace.jpl.nasa.gov/applications/groundwater/

JIAJUN Z and CHENGQING Z (2016) Bridging neural machine translation and bilingual dictionaries. arXiv preprint arXiv:1610.07272.

JIAO W, TU Z, LI J, WANG W, HUANG J and SHUMING S (2022) Tencent's multilingual machine translation system for wmt22 large-scale African languages. ArXiv preprint arXiv:2210.09644.

JÖRG T and SANTHOSH T (2020) Opus-mt-building open translation services for the world. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. European Association for Machine Translation.

JOSEPH JE, AKINROTIMI OO, RAO KPC, RAMARAJ A (2020) The usefulness of gridded climate data products in characterising climate variability and assessing crop production. CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). Working Paper No. 322. DOI:10.13140/RG.2.2.27548.31367

KARPATNE A, ATLURI G, FAGHMOUS JH, STEINBACH M, BANERJEE A, GANGULY A, SHEKHAR S, SAMATOVA N and KUMAR V (2017) Theory-guided data science: A new paradigm for scientific discovery from data. Institute of Electrical and Electronics Engineers Transactions on Knowledge and Data Engineering, 29, 2318-2331. doi: 10.1109/TKDE.2017.2720168.

KAYAD A, SOZZI M, GATTO S, MARINELLO F, PIROTTI F. (2019) Monitoring within-field variability of corn yield using Sentinel-2 and machine learning techniques. *Remote Sensing* 11 2873.

KENDA K, CERIN M, BOGATAJ M, SENOZETNIK M, KLEMEN K, PERGAR P, LASPIDOU C, MLADENIC D (2018) Groundwater modelling with machine learning techniques: Ljubljana polje aquifer. Multidisciplinary Digital Publishing Institute Proceedings 2(11) 697. https://doi.org/10.3390/proceedings2110697.

KEYSER N (1986) 1:250 000 Geological Map of the Wes Rand, 2626. South African Committee for Stratigraphy, Council for Geoscience, Pretoria.

KINGMA DP and BA J (2014) Adam: A method for stochastic optimization. arXiv. doi:1412.6980.

KITCHEN N, DRUMMOND S, LUND E, SUDDUTH K, BUCHLEITER G (2003) Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems. *Agronomy Journal* 95 483-495.

KLEYNHANS C, THIRION C and MOOLMAN J (2005) A level I river ecoregion classification system for South Africa, Lesotho and Swaziland. Department of Water Affairs and Forestry, Pretoria, South Africa.

KOTCHONI DOV, VOUILLAMOZ JM, LAWSON FMA et al. (2019) Relationships between rainfall and groundwater recharge in seasonally humid Benin: a comparative analysis of long-term hydrographs in sedimentary and crystalline aquifers. Hydrogeology Journal 27 447-457. https://doi.org/10.1007/s10040-018-1806-2.

KRATZERT F, HERRNEGGER M, KLOTZ D, HOCHREITER S and KLAMBAUER G (2019a) NeuralHydrology-interpreting LSTMs in hydrology. Explainable AI: Interpreting, explaining and visualizing deep learning, 347-362. https://doi.org/10.1007/978-3-030-28954-6_19.

KRATZERT F, KLOTZ D, BRENNER C, SCHULZ K and HERRNEGGER M (2018) Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences* 22 6005-6022. https://doi.org/10.5194/hess-22-6005-2018.

KRATZERT F, KLOTZ D, HERRNEGGER M, SAMPSON AK, HOCHREITER S and NEARING GS (2019b) Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research* 55 11344-11354. https://doi.org/10.1029/2019WR026065.

KRATZERT F, KLOTZ D, HOCHREITER S and NEARING GS (2021) A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall-runoff modeling. *Hydrology and Earth System Sciences* 25 2685-2703. https://doi.org/10.5194/hess-25-2685-2021.

KRATZERT F, KLOTZ D, SHALEV G, KLAMBAUER G, HOCHREITER S and NEARING G. (2019c) Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences* 23 5089-5110. https://doi.org/10.5194/hess-23-5089-2019.

KRAVCHENKO A N, BULLOCK D G. (2000) Correlation of corn and soybean grain yield with topography and soil properties. *Agronomy Journal* 92 75-83.

KRIZHEVSKY A, SUTSKEVER I and HINTON GE (2017) Imagenet classification with deep convolutional neural networks. *Communications of the Association for Computing Machinery* 60 84-90. https://doi.org/10.1145/3065386.

KUHN CM (1986) Geohydrological investigation of the western and central Steenkoppies compartment. Technical Report Gh3446. Directorate Geohydrology. Department of Water Affairs and Forestry, Pretoria, South Africa.

KYVERYGA PM (2019) On-farm research: experimental approaches, analytical frameworks, case studies, and impact. *Agronomy Journal* 111 2633-2635.

LAKEW SM, EROFEEVA A, NEGRI M, FEDERICO M and TURCHI M (2018) Transfer learning in multilingual neural machine translation with dynamic vocabulary. arXiv preprint arXiv:1811.01137.

LAND TYPE SURVEY STAFF (1972-2002) Land types of South Africa: Digital map (1: 250 000 scale) and soil inventory databases. ARC-Institute for Soil, Climate and Water, Pretoria, South Africa.

LE ROUX B, VAN DER LAAN M, VAHRMEIJER T, ANNANDALE JG and BRISTOW KL (2016) Estimating Water Footprints of Vegetable Crops: Influence of Growing Season, Solar Radiation Data and Functional Unit. *Water Journal* 8 473. doi:10.3390/w8100473.

LE ROUX J, MARARAKANYE N, LEUSHANTHA M, WEEPENER H and VAN DER LAAN M (2020) A South African National Input Database to Run the SWAT model in a GIS. WRC Report, Pretoria, South Africa.

LEENAARS J, VAN OOSTRUM A and GONZALEZ MR (2013) Africa Soil Profiles Database, Version 1.2. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). International Soil Reference and Information Centre (ISRIC) Report, ISRIC Report No. 2014/01.

LEES T, BUECHEL M, ANDERSON B, SLATER L, REECE S, COXON G and DADSON SJ (2021) Benchmarking Data-Driven Rainfall-Runoff Models in Great Britain: A comparison of LSTM-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*. https://doi.org/10.5194/hess-25-5517-2021.

LENG G, HALL JW (2020) Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental Research Letters* 15 044027.

LEWIS M, LIU Y, GOYAL N, GHAZVININEJAD M, MOHAMED A, LEVY O, STOYANOV V and ZETTLEMOYER L (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461.

LI B, RODELL M, KUMAR S, BEAUDOING HK, GETIRAN, A, ZAITCHIK BF et al. (2019) Global GRACE data assimilation for groundwater and drought monitoring: Advances and challenges. *Water Resources Research* 55 7564-7586 https://doi.org/10.1029/2018WR024618.

LI C, MA C, PEI H, FENG H, SHI J, WANG Y, CHEN W, LI Y, FENG X, SHI Y (2020) Estimation of Potato Biomass and Yield Based on Machine Learning from Hyperspectral Remote Sensing Data. *Journal of Agricultural Science and Technology* 10 195-213.

LI K-Y, SAMPAIO DE LIMA R, BURNSIDE NG, VAHTMÄE E, KUTSER T, SEPP K, CABRAL PINHEIRO V H, YANG M-D, VAIN A and SEPP K. (2022) Toward Automated Machine Learning-Based Hyperspectral Image Analysis in Crop Yield and Biomass Estimation. *Remote Sensing* 14 1114.

LIESCH T and OHMER M (2016) Comparison of GRACE data and groundwater levels for the assessment of groundwater depletion in Jordan. Hydrology Journal (24) 1547-1563. doi: 10.1007/s10040-016-1416-9

LIU Y, GU J, GOYAL N, LI X, EDUNOV S, GHAZVININEJAD M, LEWIS M and ZETTLEMOYER L (2020) Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* 8 726-742.

LOISEAU T, ARROUAYS D, RICHER-DE-FORGES AC, LAGACHERIE P, DUCOMMUN C and MINASNY B (2021) Density of soil observations in digital soil mapping: A study in the Mayenne region, France. *Geoderma Regional* 24 e00358.

MAIMON OZ and ROKACH L (2014) Data mining with decision trees: theory and applications. Vol. 81. World Scientific.

MANN HB and WHITNEY DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 50-60.

MEYER R (2014) Hydrogeology of Groundwater Region 10: The Karst Belt. Water Research Commission (WRC) Report No. TT 553/14, Pretoria, South Africa.

MIKE L, YINAHAN L, NANAM G, MARJAN G, ABDELRAHMAN M, OMER L, VES S and LUKE Z (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

MILEWSKI AM, THOMAS MB, SEYOUM WM, RASMUSSEN TC (2019) Spatial downscaling of GRACE TWSA data to identify spatiotemporal groundwater level trends in the Upper Floridan Aquifer, Georgia, USA. *Remote Sensing* 11 2756. doi: 10.3390/rs11232756.

MILLER MA, SHEPHERD KD, KISITU B and COLLINSON J (2021) iSDAsoil: The first continent-scale soil property map at 30 m resolution provides a soil information revolution for Africa. *PLoS Biology* 19(11) e3001441.

MINASNY B, STOCKMANN U, HARTEMINK AE and MCBRATNEY AB (2016) Measuring and modelling soil depth functions. *Digital Soil Morphometrics* 225-240.

MOIWO JP, YANG Y, HAN S, LU W, YAN N and WU B (2011) A method for estimating soil moisture storage in regions under water stress and storage depletion: a case study of Hai River Basin, North China. *Hydrological Processes* 25 2275-2287. doi: 10.1002/hyp.7991.

MONTEITH JL (1965) Evaporation and environment. *Symposium for the Society for Experimental Biology* 19 205-234.

MOORE JM, TSIKOS H and POLTEAU S (2001) Deconstructing the Transvaal Supergroup, South Africa: implications for Palaeoproterozoic palaeoclimate models. *Journal of African Earth Sciences* 33 437-444. doi: 10.1016/S0899-5362(01)00084-7.

MORA-VALLEJO A, CLAESSENS L, STOORVOGEL J and HEUVELINK GB (2008) Small scale digital soil mapping in Southeastern Kenya. *Catena* 76(1) 44-53.

MORIASI DN, ARNOLD JG, VAN LIEW MW, BINGNER RL, HARMEL RD and VEITH TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the American Society of Agricultural and Biological Engineers* 50 885-900. doi: 10.13031/2013.23153.

MU Q, HEINSCH FS, ZHAO M and RUNNING SW (2007) Development of a global evapotranspiration algorithm based on MODIS and global meteorology data. *Remote Sensing of the Environment* (111) 519-536. doi: 10.1016/j.rse.2007.04.015.

MU Q, ZHAO M and RUNNING SW (2011) Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of the Environment* 115 1781-1800. https://doi.org/10.1016/j.rse.2011.02.019

MUHAMMAD AU, LI X and FENG J (2019) Using LSTM GRU and hybrid models for streamflow forecasting (2019) Machine Learning and Intelligent Communications: 4th International Conference, MLICOM 2019, Nanjing, China, August 24-25, 2019, Proceedings 4. Springer, 510-524. https://doi.org/10.1007/978-3-030-32388-2_44.

MYENI L, MDLAMBUZI T, PATERSON DG, DE NYSSCHEN G and MOELETSI ME (2021) Development and evaluation of pedotransfer functions to estimate soil moisture content at field capacity and permanent wilting point for South African soils. *Water* 13(19) 2639.

NASER M and ALAVI A (2020) Insights into performance fitness and error metrics for machine learning. *arXiv preprint arXiv:2006.00887*.

NASH JE and SUTCLIFFE JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology* 10 282-290. https://doi.org/10.1016/0022-1694(70)90255-6.

NÄSI R, VILJANEN N, KAIVOSOJA J, ALHONOJA K, HAKALA T, MARKELIN L and HONKAVAARA E. (2018) Estimating biomass and nitrogen amount of barley and grass using UAV and aircraft based spectral and photogrammetric 3D features. *Remote Sensing* 10 1082.

NAWAR S, CORSTANJE R, HALCRO G, MULLA D and MOUAZEN AM (2017) Delineation of soil management zones for variable-rate fertilization: A review. *Advances in Agronomy* 143 175-245.

NEARING GS, KRATZERT F, SAMPSON AK, PELISSIER CS, KLOTZ D, FRAME JM, PRIETO C and GUPTA HV (2021) What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research* 57. https://doi.org/10.1029/2020WR028091.

NIFA K, BOUDHAR A, OUATIKI H, ELYOUSSFI H, BARGAM B and CHEHBOUNI A (2023) Deep Learning Approach with LSTM for Daily Streamflow Prediction in a Semi-Arid Area: A Case Study of Oum Er-Rbia River Basin, Morocco. *Water* 15 262. https://doi.org/10.3390/w15020262.

NYÉKI A, MILICS G, KOVÁCS A and NEMÉNYI M (2017) Effects of soil compaction on cereal yield: A review. *Cereal Research Communications* 45 1-22.

ODENDAAL N (2021) Govt needs to focus on securing reliable hydrological information to ensure water security. Available: https://www.engineeringnews.co.za/article/govt-needs-to-focus-on-securing-reliable-hydrological-information-to-ensure-water-security-2021-02-24/rep_id:4136.

OSTLING R and TIEDEMANN J (2017) Neural machine translation for low-resource languages. arXiv preprint arXiv:1708.05729.

OYEBANDE L (2001) Water problems in Africa—how can the sciences help? *Hydrological Sciences Journal* 46 947-962. https://doi.org/10.1080/02626660109492888.

PANTAZI XE, MOSHOU D, ALEXANDRIDIS T, WHETTON L, MOUAZEN AM (2016) Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture* 121 57-65.

PATERSON G, TURNER D, WIESE L, VAN ZIJL G, CLARKE C and VAN TOL J (2015) Spatial soil information in South Africa: Situational analysis, limitations and challenges. *South African Journal of Science* 111(5-6) 1-7.

PEDREGOSA F, VAROQUAUX G, GRAMFORT A, MICHEL V, THIRION B, GRISEL O, BLONDEL M, PRETTENHOFER P, WEISS R and DUBOURG V (2011) Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* 12 2825-2830.

PEREZ-ALONSO D, PEÑA-TEJEDOR S, NAVARRO M, RAD C, ARNAIZ-GONZÁLEZ Á, DÍEZ-PASTOR J-F (2017) Decision Trees for the prediction of environmental and agronomic effects of the use of Compost of Sewage Slugde (CSS). *Sustainable Production and Consumption* 12 119-133.

PITMAN WV and BAILEY AK (2021) Can CHIRPS fill the gap left by the decline in the availability of rainfall stations in southern Africa? *Water SA* 47(2) 162-171. https://doi.org/10.17159/wsa/2021.v47.i2.10912.

POST WM, IZAURRALDE RC, MANN LK and BLISS N (2001) Monitoring and verifying changes of organic carbon in soil. *Climatic Change* 51(1) 73-99.

PROTECTION OF PERSONAL INFORMATION ACT NO. 4 OF 2013 (2013) South Africa. https://www.gov.za/documents/protection-personal-information-act (accessed 28 November 2023).

RAHAMAN MM, THAKUR B, KALRA A, LI R and MAHESHWARI P (2019) Estimating High-Resolution Groundwater Storage from GRACE: A Random Forest Approach. *Environments* (6)63. https://doi.org/10.3390/environments6060063.

RAMJEAWON M, DEMLIE M and TOUCHER M (2022) Analyses of groundwater storage change using GRACE satellite data in the Usutu-Mhlatuze drainage region, north-eastern South Africa. *Journal of Hydrology: Regional Studies* 42 101118. https://doi.org/10.1016/j.ejrh.2022.101118.

RANSOM CJ, KITCHEN NR, CAMBERATO JJ, CARTER PR, FERGUSON RB, FERNÁNDEZ FG, FRANZEN DW, LABOSKI CA, MYERS DB, NAFZIGER ED (2019) Statistical and machine learning methods evaluated for incorporating soil and weather into corn nitrogen recommendations. *Computers and Electronics in Agriculture* 164 104872.

REASON C and KEIBEL A (2004) Tropical cyclone Eline and its unusual penetration and impacts over the southern African mainland. Weather and Forecasting 19 789-805. https://doi.org/10.1175/1520-0434(2004)019<0789:TCEAIU>2.0.CO;2.

REICHSTEIN M, CAMPS-VALLS G, STEVENS B, JUNG M, DENZLER J and CARVALHAIS N (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566 195-204. https://doi.org/10.1038/s41586-019-0912-1.

ROBERTS W, WILLIAMS GP, JACKSON E, NELSON EJ and AMES DP (2018) Hydrostats: A Python package for characterizing errors between observed and predicted time series. *Hydrology* 5 66. https://doi.org/10.3390/hydrology5040066.

RODELL M, CHEN J, KATO H, FAMIGLIETTI JS, NIGRO J and WILSON CR (2007) Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. *Hydrogeology Journal* 15(1) 159-166. https://doi.org/10.1007/s10040-006-0103-7.

ROGERS DP, TSIRKUNOV VV, KOOTVAL H, SOARES A, KULL D, BOGDANOVA A-M and SUWA M (2019) Weathering the change: how to improve hydromet services in developing countries? World Bank.

ROSE MD, FIDELIBUS C, MARTANO P (2018) Assessment of Specific Yield in Karstified Fractured Rock through the Water-Budget Method. *Geosciences* 8(9) 344. https://doi.org/10.3390/geosciences8090344

ROSSITER DG (2016) Digital soil resource inventories: status and prospects in 2015. *Digital soil mapping across paradigms, scales and boundaries* 275-286.

ROSSITER DG, POGGIO L, BEAUDETTE D and LIBOHOVA Z (2021) How well does Predictive Soil Mapping represent soil geography? An investigation from the USA. *Soil Discuss* 1-35.

RUDER S (2016) An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

RUI H, BEAUDOING H and LOESER C (2022) README Document for NASA GLDAS Version 2 Data Products. National Aeronautics and Space Administration (NASA) and Goddard Earth Sciences Data and Information Services Centre (GES DISC).

RUKUNDO E and DOĞAN A (2019) Dominant Influencing Factors of Groundwater Recharge Spatial Patterns in Ergene River Catchment, Turkey. *Water* 11 653. doi:10.3390/w11040653

RUNNING S, MU Q and ZHAO M. (2021) MODIS/Terra Net Evapotranspiration 8-Day L4 Global 500m SIN Grid V061. Distributed by NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MODIS/MOD16A2.061.

RYAN M (2020) Agricultural big data analytics and the ethics of power. *Journal of Agricultural and Environmental Ethics* 33 49-69.

SABZEHEE F, AMIRI-SIMKOOEI AR, IRAN-POUR S, VISHWAKARME BD and KERACHIAN R (2023) Enhancing spatial resolution of GRACE-derived groundwater storage anomalies in Urmia catchment using machine learning downscaling methods. *Journal of Environmental Management* 330 117180. https://doi.org/10.1016/j.jenvman.2022.117180.

SADATH PVR, KARTHEESHWARE MR and ELANGO L (2023) Sustainable groundwater management under global climate change: mitigation and adaptation measures. In: Li, P., Elumalai, V. (eds) Recent Advances in Environmental Sustainability. EESIWC 2021. Environmental Earth Sciences. Springer, Cham. https://doi.org/10.1007/978-3-031-34783-2_10.

SAHOUR H (2020) Statistical Downscaling Techniques to Enhance the Spatial Resolution of the Grace Satellite Data and to Fill Temporal Gaps. Dissertations. 3634. https://scholarworks.wmich.edu/dissertations/3634.

SAVE H, BETTADPUR S and TAPLEY BD (2016) High-resolution CSR GRACE RL05 mascons. *Journal of Geophysical Research: Solid Earth* 121(10) 7547-7569. https://doi.org/10.1002/ 2016JB013007.

SCHULZE RE and SCHÜTTE S (2020) Mapping soil organic carbon at a terrain unit resolution across South Africa. *Geoderma* 373 114447.

SEFARA TJ, ZWANE SG, GAMA N, SIBISI H, SENOAMADI PN and MARIVATE, V (2021) Transformer-based machine translation for low-resourced languages embedded with language identification. 5th Conference on Information Communications Technology and Society, Durban, South Africa, 10-11 March 2021.

SEN Z (2015) Chapter 6 – Groundwater Management. Practical and Applied Hydrogeology. pp. 341-397. https://doi.org/10.1016/B978-0-12-800075-5.00006-6

SENENT-APARICIO J, JIMENO-SÁEZ P, BUENO-CRESPO A, PÉREZ-SÁNCHEZ J and PULIDO-VELÁZQUEZ D (2019) Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction. *Biosystems Engineering* 177 67-77. https://doi.org/10.1016/j.biosystemseng.2018.04.022.

SEYOUM WM, KWON D and MILEWSKI M (2019) Downscaling GRACE TWSA Data into High-Resolution Groundwater Level Anomaly Using Machine Learning-Based Models in a Glacial Aquifer System. *Remote Sensing* 11 824. doi:10.3390/rs11070824.

SHAHHOSSEINI M, HU G, HUBER I and ARCHONTOULIS S V. (2021) Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Scientific Reports* 11 1-15.

SHAPIRO SS and WILK MB (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52 591-611.

SHELESTOV A, LAVRENIUK M, KUSSUL N, NOVIKOV A and SKAKUN S (2017) Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Classification. *Frontiers in Earth Science* 5(17). doi: 10.3389/feart.2017.00017.

SHIRZADI N (2023) Time Series Analysis and Forecasting with Python [Online]. Available: https://www.udemy.com/course/time-series-analysis-and-forecasting-with-python/ [Accessed 2022-06-06].

SILVA JV, TENREIRO TR, SPÄTJENS L, ANTEN NP, VAN ITTERSUM MK, REIDSMA P (2020) Can big data explain yield variability and water productivity in intensive cropping systems? *Field Crops Research* 255 107828.

SIT M, DEMIRAY B and DEMIR I (2021) Short-term hourly streamflow prediction with graph convolutional gru networks. arXiv. https://doi.org/10.48550/arXiv.2107.07039.

SIT M, DEMIRAY BZ, XIANG Z, EWING GJ, SERMET Y and DEMIR I (2020) A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology* 82 2635-2670. https://doi.org/10.2166/wst.2020.369.

SONKA S (2016) Big data: fueling the next evolution of agricultural innovation. *Journal of Innovation Management* 4 114-136.

SRIVASTAVA R, KUMAR S and KUMAR B (2023) Classification model of machine learning for medical data analysis. Chapter 7. Statistical Modeling in Machine Learning. Academic Press.

STRASSBERG G, SCANLON BR and RODELL, M. (2007) Comparison of seasonal terrestrial water storage variations from GRACE with groundwater-level measurements from the High Plains Aquifer (USA). *Geophysical Research Letters* 34 L14402, doi:10.1029/2007GL030139.

SULEIMAN AA and RITCHIE JT (2004) Modifications to the DSSAT vertical drainage model for more accurate soil water dynamics estimation. *Soil Science* 169(11) 745-757.

SWANEPOEL C, VAN DER LAAN M, WEEPENER H, DU PREEZ C and ANNANDALE JG (2016) Review and meta-analysis of organic matter in cultivated soils in southern Africa. *Nutrient Cycling in Agroecosystems* 104 107-123.

SWENSON S and WAHR J (2002) Methods for inferring regional surface-mass anomalies from Gravity Recovery and Climate Experiment (GRACE) measurements of time-variable gravity. *Journal of Geophysical Research: Solid Earth* 107 (B9). https://doi.org/10.1029/ 2001JB000576.

TANTALAKI N, SOURAVLAS S and ROUMELIOTIS M (2019) Data-driven decision making in precision agriculture: the rise of big data in agricultural systems. Journal of Agricultural & Food Information 20 344-380.

THINDA K, OGUNDEJI A, BELLE J and OJO T (2020) Understanding the adoption of climate change adaptation strategies among smallholder farmers: Evidence from land reform beneficiaries in South Africa. *Land Use Policy* 99 104858.

UWIRAGIYE Y, KHALAF QAW, ALI HM, NGABA MJY, YANG M, ELRYS AS, CHEN Z and ZHOU J (2023) Spatio-Temporal Variations in Soil pH and Aluminum Toxicity in Sub-Saharan African Croplands (1980-2050). *Remote Sensing* 15(5) 1338.

VAHRMEIJERA JT, ANNANDALEA JG, BRISTOWB KL, STEYN JM and HOLLAND M (2013) Drought as a catalyst for change: A case study of the Steenkoppies Dolomite Aquifer. In: Schwabe K, Albiac J, Connor J, Hassan R, Meza Gonzalez L editors. Drought in arid and semi-arid regions: A multi-disciplinary and cross-country perspective. Dordrecht: Springer Publications. pp. 251-268.

VAN KLOMPENBURG T, KASSAHUN A and CATAL C (2020) Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture* 177 105709.

VAN TOL J and VAN ZIJL G (2020) Regional soil information for hydrological modelling in South Africa. Water Wheel 19 (2) 43-45.

VAN TOL J, DZVENE A, LE ROUX P and SCHALL R (2016) Pedotransfer functions to predict Atterberg limits for South African soils using measured and morphological properties. *Soil Use and Management* 32(4) 635-643.

VAN ZIJL G (2019) Digital soil mapping approaches to address real world problems in southern Africa. *Geoderma* 337 1301-1308.

VISHWAKARMA BD, DEVARAJU B, SNEEUW N (2018) What Is the Spatial Resolution of GRACE Satellite Products for Hydrology? *Remote Sensing* 10 852. doi:10.3390/rs10060852

VISHWAKARMA BD, ZHANG J, SNEEUW N (2021) Downscaling GRaCE total water storage change using partial least squares regression. *Scientific Data* 8 95. https://doi.org/10.1038/s41597-021-00862-6

VIVIER JJP, BULASIGOBO JR, WIETHOFF A and KRIEK C (2005) Groundwater flow management model of the Maloney's Eye Catchment Area. Africa Geo-Environmental Services (Pty) Ltd. Technical report: AG/R/05/11/30.

WANG L, LI Z, WANG D, LIAO S, NIE X and LIU Y (2022) Factors controlling soil organic carbon with depth at the basin scale. *Catena* 217 106478.

WIEGMANS FE, HOLLAND M and JANSE VAN RENSBURG H (2013) Groundwater Resource Directed Measures for Maloney's Eye Catchment. Water Research Commission Project No. K8/970, Pretoria, South Africa.

WOLFERT S, GE L, VERDOUW C and BOGAARDT M-J (2017) Big data in smart farming – a review. *Agricultural Systems* 153 69-80.

WOLK K and MARASEK K (2015) Unsupervised comparable corpora preparation and exploration for bi-lingual translation equivalents. In: Proceedings of the 12th International Workshop on Spoken Language Translation: Papers. Da Nang, Vietnam, 118-125.

YAN X, ZHANG B, YAO Y, YIN J, WANG H and RAN Q (2022) Jointly using the GLDAS-2.2 model and GRACE to study the severe Yangtze flooding of 2020. *Journal of Hydrology* 610(3-4) 127927. doi: 10.1016/j.jhydrol.2022.127927.

YEH PJF, SWENSON S, FAMIGLIETTI J and RODELL M (2006) Remote sensing of groundwater storage changes in Illinois using the Gravity Recovery and Climate Experiment (GRACE). *Water Resources Research* 42 (12). doi: 10.1029/2006WR005374

YI D, AHN J and JI S (2020) An effective optimization method for machine learning based on ADAM. *Applied Sciences* 10 1073.

YIN W, HU L, ZHANG M, WANG J and HAN SC (2018) Statistical downscaling of GRACE-derived groundwater storage using ET data in the North Chine Plain. Journal of Geophysical Research: *Atmospheres* 123 5973-5987. https://doi.org/10.1029/2017JD027468

YIU T (2019) Understanding Random Forest: How the Algorithm Works and Why it Is So Effective. Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

YU S, DING H and ZENG Y (2022) Evaluating water-yield property of karst aquifer based on the AHP and CV. Scientific Reports 12 3308. https://doi.org/10.1038/s41598-022-07244-x.

ZHANG G-L, LIU F and SONG X-D (2017) Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture* 16 2871-2885.

ZHENG A and CASARI A (2018) Feature engineering for machine learning: principles and techniques for data scientists, O'Reilly Media, Inc. ISBN: 9781491953242.

ZHI J, CAO X, WUGU E, ZHANG Y, WANG L, QU LA and WU J (2022) Effects of Soil Map Scales on Estimating Soil Organic Carbon Stocks in Southeastern China. *Land* 11(8) 1285.

Appendix

Appendix I: Capacity Building

	Current degree	Co-supervisor(s)	Research Outputs
Cindy Viviers	PhD in Water Resource Management (University of Pretoria)	Prof. Matthys Dippenaar Engineering Geology and Hydrogeology	• Machine Learning for Earth Observation (with application in Agriculture) course (1 April-4 July 2022). • On the 20th of September 2023, Cindy presented at the 50th International Association for Hydrogeologists (IAH) Congress in Cape Town, on "Can open-source remote sensing data be used to accurately downscale groundwater storage estimates?". • **Research article:** "*Assessing the accuracy of downscaled GRACE-assimilated groundwater storage estimates across two different aquifer types*".
Christiaan Schutte	PhD in Water Resource Management (University of Pretoria)	Dr Barend van der Merwe Department of Geography, Geoinformatics and Meteorology	• Kirkham Soil Physics Conference (September 2022). Title: Streamflow forecasting with deep learning: a case study in a South African semi-arid catchment. • South African Hydrological Society (SAHS) Conference (October 2022). Title: Can freely available weather data and deep learning accurately predict stream flow in a South African semi-arid catchment? • WRC Model-a-thon: Participated in the first catchment hydrology model-a-thon in SA. Used the SWAT+ model. • Research article: "Leveraging historic streamflow and weather data with deep learning for enhanced streamflow predictions" accepted by the Journal for Hydroinformatics (IWA).

Simphiwe Maseko	PhD Agronomy (University of Pretoria)	Prof E.H Tesfamariam Associate Professor – Department of Plant and Soil Sciences Faculty of Natural and Agricultural Sciences	• International Conference for On-farm Precision Experimentation 2024 Title: Analyzing the performance of machine learning models for Sub-field maize yield prediction in precision agriculture (08-11 Jan 2024) • SANCID symposium 2023 Title: The Water Research Observatory: Discovering and uploading data for hydrological modelling and big data analytics (21-23 Feb 2023) • Combined Congress 2023 Title: Evaluating machine learning approaches for sub-field maize yield predictions to inform precision agriculture (24-26 Jan 2023) • Knowledge Transfer Partnership (KTP) presentations 2022 Title: Yield prediction using precision agriculture big data: A comparison of process-based and machine learning models (30 November 2022) • Combined Congress 2022 Title: Maize (Zea mays L.) yield response to varying input application rates in on-farm precision agriculture trials (25-27 Jan 2022) • Agricultural Research Council (ARC) – Water Science Student Presentations Title: Yield prediction using precision agriculture big data: A comparison of process-based and machine learning models (23 November 2021) • Research article: "Evaluating machine learning models for subfield maize yield predictions in precision agriculture"
Aimee Thomson	MSc Soil Science (University of Pretoria)	Leushantha Mudaly Department of Plant and Soil Sciences Faculty of Natural and Agricultural Sciences	• February 2022 – ARC Water Seminar (attendee) • August 2022 – Kirkham Conference (poster presenter) • January 2023 – Combined Congress (presenter)

		Garry Paterson Agricultural Research Council – Natural Resources and Engineering	• February 2023 – SANCID conference (presenter) • Research project: "The accuracy of digital soil maps for South Africa using historical and newly measured data and crop model sensitivity analysis."
Pitso Khoboko	BSc Hons Computer Science (University of Pretoria)	Prof V Marite University of Pretoria EBIT J Sefara Centre for Scientific and Industrial Research	• Research project: "Transformer-based Neural Machine Translation for Native South African Languages"
Ntando Mthembu	Agricultural Research Council Intern	n/a – Internship for work experience.	• Uploading datasets to the WRO and capturing metadata.

Appendix II: Research Outputs

Conference presentations

Maseko, S; van der Laan, M; Meyer, F; Delport, M; Otterman, H; Edge, B; Tesfamariam, E. Maize yield responses to varying input application in on-farm precision experimentation. Combined Congress, 25-27 January 2022, Virtual (www.combinedcongress.org.za).

Van der Laan, M; Maseko, S & Thompson, A. The Water Research Observatory: Discovering and uploading data for hydrological modelling and big data analytics. South African National Committee on Irrigation and Drainage (SANCID) Symposium, 21-23 February 2023, Tzaneen, South Africa.

Van der Laan, M; le Roux, J; Viviers, C; Schutte, C; Maseko, S; Silberbauer, M; Mudaly, L; Weepener, H; Clark, D; Mabhaudhi, T; Adams, S; Nhamo, L; Mpandeli, S; Thomson, A; Hoogenboom, G; Srinivasan, *r* & Khoboko, P. The Water Research Observatory: Developing a cloud-based data platform for water research and hydrological modelling in South Africa. South African Hydrological Society Conference, 10-12 October, Muldersdrift.

Schutte, C; van der Laan, M & van der Merwe, B. Can freely available weather data and deep learning accurately predict stream flow in a South African semi-arid catchment? South African Hydrological Society Conference, 10-12 October, Muldersdrift.

Van der Laan, M; Silberbauer, M; Schutte, C; Maseko, S; Viviers, C; Thomson, A; Khoboko, P; le Roux, J; Mudaly, L; Weepener, H; Mabhaudhi, T., Hoogenboom G., Srinivasan, R. The Water Research Observatory: a cloud-based platform for agrohydrological modelling and big data applications. Combined Congress, 22-24 January 2022. Pretoria, South Africa.

Viviers, C; van der Laan, M; Z, Gafoor; Dippenaar M. Can open-source remote sensing data be used to accurately downscale groundwater storage estimates. International Association of Hydrogeologist Worldwide Groundwater Congress, 18-22 September 2023, Cape Town, South Africa.

Maseko S, van der Laan M, Tesfamariam E H, Delport M, Otterman H. Analyzing the performance of machine learning models for sub-field maize yield prediction in precision agriculture, International Conference for On-farm Precision Experimentation 2024, 08-11 Jan 2024, Texas, USA. (Virtual)

Schutte, CE; van der Laan, M & van der Merwe, B. Streamflow prediction with deep learning: a South African semi-arid catchment case study. Soil Science Society of America Kirkham Conference, 28 August-2 September 2022, Skukuza, South Africa.

Posters

Maseko, S; van der Laan, M; Tesfamariam, E; Delport, M & Otterman, H. Evaluating machine learning approaches for sub-field maize yield predictions to inform precision agriculture. Combined Congress, 22-24 January 2023. Pretoria, South Africa.

Thomson, A; van der Laan, M; Mudaly, L; Paterson, G. Evaluating the reliability of multiple soil digital maps for crop model parameterisation in South Africa. Soil Science Society of America Kirkham Conference, 28 August-2 September 2022, Skukuza, South Africa.

Invited talks

Van der Laan, M et al. (2023) The WRC's Water Research Observatory and ARC's digital assets: Collaboration opportunities to help the DWS in its digitalisation strategy. Department of Water and Sanitation Digitisation of Water Monitoring Systems Workshop. Online, 1 November 2023.

Van der Laan, M et al. University of KwaZulu-Natal Centre for Water Resources Research 'Research on Tap' webinar series. Invited to talk on the Water Research Observatory. Online, 24 May 2023.

Van der Laan, M et al. Introduction to the Water Research Observatory. South African National Biodiversity Institute and Department of Science and Innovation Fresh Water Ecosystem (FEN) Annual meeting. Online, 17 November 2023.

Journal articles

Schutte, C.E., van der Laan, M. & van der Merwe B. Leveraging historic streamflow and weather data with deep learning for enhanced streamflow predictions. *Journal of Hydroinformatics*, accepted.

Viviers, C., van der Laan, M., Gaffoor, Z. & Dippenaar M. A machine learning based approach for downscaling groundwater 1 storage anomalies and accuracy assessment in two different aquifer types. *Journal of Hydrology*: Regional Studies, submitted.

Maseko, S; van der Laan, M; Tesfamariam, E; Delport. Evaluating machine learning models for sub-field maize yield predictions in precision agriculture. *European Journal of Agronomy*, submitted.

Popular press articles

Van der Laan, M. Water research in South Africa: Getting ready for big data analytics. Water Wheel, March-April 2022.

Van der Laan, M. The Water Research Observatory: Unlocking the power of water data. Water Wheel, March/April 2024.

Awards

The WRO project was awarded the 'Best paper on new technology' award from the Soil Science Society of South Africa' at the Combined Congress (2023).

Appendix III: Streamflow Prediction with Deep learning

Negative streamflow predictions

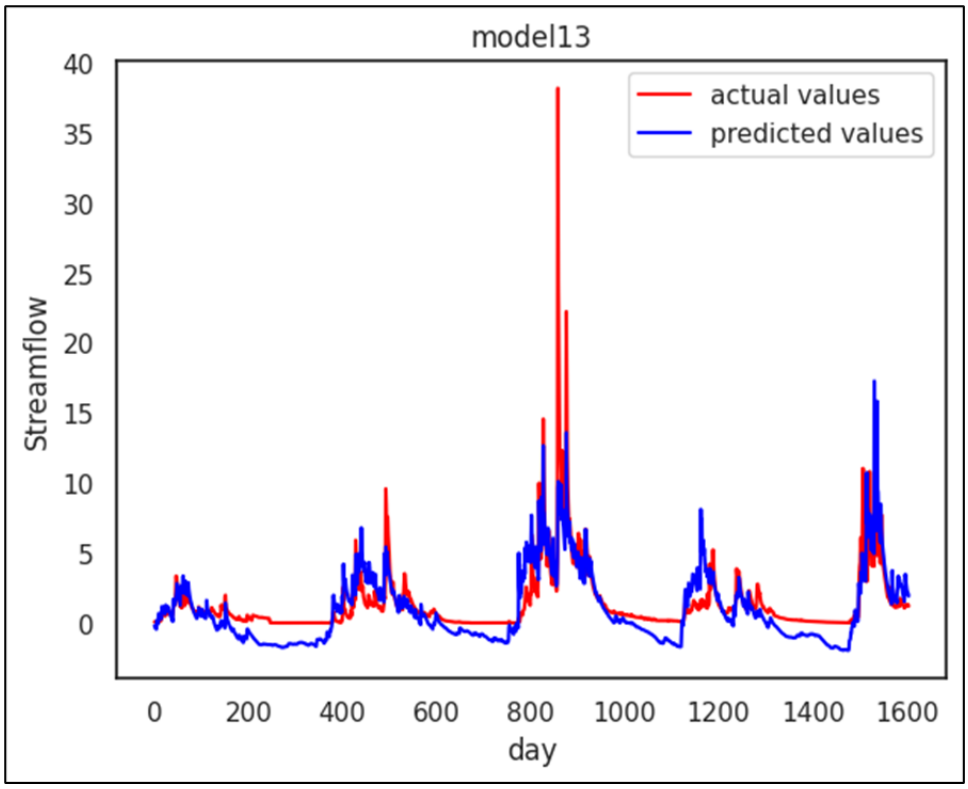


Figure 26: Negative streamflow predictions in Catchment A. The dotted line indicates 0 m^3s^{-1}.

Look-back window experiment

The look-back window (LBW) is an important hyperparameter in these models that specifies how many previous timesteps are considered by the LSTM/GRU network to predict the value at the next timestep. The LBW is connected to the climate and hydrological connectivity of a catchment and various LBW sizes have been used for LSTM models in different climates. Several studies in the northern hemisphere, including the United States, Canada and the United Kingdom, used LBWs between 270-360 days for LSTM and hybrid CNN-LSTM models (Kratzert et al., 2019c, Lees et al., 2021, Anderson and Radić, 2022), where snowmelt is a major driver of streamflow in many catchments. A LBW of only 15 days was sufficient for LSTM models in a subtropical wet climate zone of China Fan et al. (2020), and a LBW between 20-30 days worked well for a semi-arid catchment in Morocco (Nifa et al., 2023). Less information is available regarding the optimal LBW size for GRU networks, and no information is available regarding a LBW size in southern Africa.

The experiment studied the impact of LBW size on model performance. Based on literature, LBW sizes of three, five, 10, 20, 30, 60, 90 and 120 days, were identified for testing (Kratzert et al., 2019a, Fan et al., 2020, Nifa et al., 2023). The input variables included rainfall, streamflow, minimum and maximum temperatures (Combination 2) and were sourced from the ARC weather station data. For each of the two catchments, 20 GRU and 20 LSTM networks were trained for each LBW size.

It appears that for both catchments, the NSE values vary slightly with the change in LBW size, but there isn't a clear trend indicating that a larger or smaller LBW consistently leads to better model performance. LBWs of 10 to 120 days provided more accurate predictions than LBWs of three to five days. This agrees well with Nifa et al., 2023, Fan et al. (2020), that found LBWs ranging between 15-30 days to be optimal. Longer LBWs decreased computational efficiency considerably, resulting in both longer training and longer prediction times. LBWs of 10 to 30 days were generally deemed suitable for accurate predictions across both catchments and model types, and for computational efficiency. There is, however, potential for further optimization and adjustment.

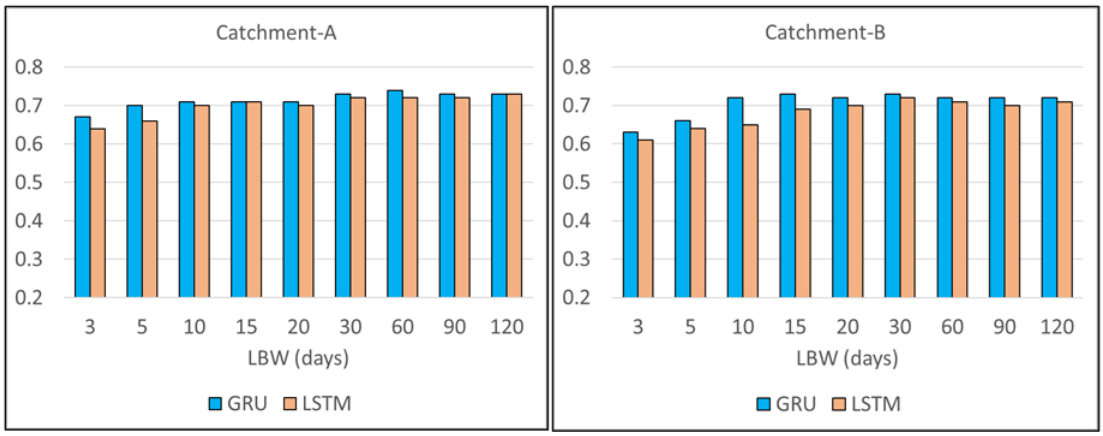


Figure 27: Nash-Sutcliffe Efficiency (NSE) values for different look-back window sizes.

The data and code used to develop the DL models for streamflow prediction is available through the WRO (https://data.waterresearchobservatory.org/metadata-form/deep-learning-for-streamflow-prediction-project-data) and can be used to replicate the experiments or to build new models for different catchments.