TOWARDS A HYDROLOGICAL SOIL MAP OF SOUTH AFRICA (HYDROSOIL) – DEVELOPING A METHODOLOGY AND SHOWCASING ITS USES

Report to the
Water Research Commission

by

George van Zijl[1], Johan van Tol[2], Eddy Smit[2], Molebaleng Sehlapelo[1], Anru Kock[1], Willie Cloete[1], Jacques Faul[1], Jay Le Roux[2], Eddie Riddell[3], Altus Jacobs[1], Elouise Verwey[1], Vian Cooke[1], Willem de Clercq[4], Alen Manyevere[5], Simon Lorentz[6]

[1] Unit for Environmental Sciences and Management, North-West University
[2] Department of Soil, Crop and Climate, University of the Free State
[3] South African National Parks
[4] Stellenbosch University
[5] University of Fort Hare
[6] SRK Consulting

WRC Report No. 3145/1/24
ISBN 978-0-6392-0623-3

May 2024

WATER RESEARCH COMMISSION

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

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

This is the final report of WRC project no. C2020-2021-00455.

EXECUTIVE SUMMARY

BACKGROUND

On 20 February 2020, just before the outbreak of the Covid-19 pandemic, the Department of Water and Sanitation hosted a workshop titled 'Towards a detailed hydrological soil map for South Africa' in Pretoria. The workshop identified an urgent need for a detailed hydrological soil map for South Africa, and that this map should be parameterised with hydrological soil property data, which will require pedotransfer functions for ease of predictions. This project was duly initiated to commence the fulfilment of the needs expressed at the workshop shortly afterwards.

AIMS

The following were the aims of the project:

1. Develop the methodology by which a national hydrological soil map (HYDROSOIL) could be created with digital soil mapping methods.
2. Compile a legacy soil point and hydrological properties database through data rescue of largely paper-based data sources currently contained in academic theses, research reports and with corporate institutions.
3. Create a HYDROSOIL map for each of the priority areas within six economically and/or ecologically important catchments of South Africa, using the database compiled in Aim 2. The catchments used in this project includes: The Sabie-Sand, The Olifants, the Jukskei, the uMngeni, the Tsitsa and the Goukou.
4. Develop pedotransfer functions by which hydrological parameters could be calculated from readily measured soil properties, using the database compiled in Aim 2.
5. Combine the hydrological soil maps created in Aim 3 with the pedotransfer functions developed in Aim 4 to create hydrological soil property maps for the study areas.
6. Determine the value of the HYDROSOIL map by hydrological modelling of all six study areas with both the best existing soils data, as well as the HYDROSOIL map.
7. At each catchment a unique aspect regarding the HYDROSOIL map was investigated. This includes:
    a. Calibration of the SWAT hydrological model through optimizing model parameters based on the expected hydrological response of the hydrological response unit (Sabie-Sand).
    b. Determining a method to incorporate the land type field maps into creating a digital soil map (Olifants).
    c. Determining the effect of pixel size on the SWAT hydrological model outcome (Jukskei).
    d. Quantifying the effect of land use change on the hydrological regime (uMngeni).
    e. Modelling sediment yield using SWAT (Tsitsa).
    f. Using the JAMS hydrological model (Goukou).

METHODOLOGY

Existing legacy soil data was collected and added to a soil point database. The ideal structure and quality control measures for such a soil database were determined through a Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis of an existing national and international soil point database. Collected data included all available soil data, not only confined to the six catchments of concern. Thereafter the soil data was used to create hydrological soil maps for the different catchments using digital soil mapping approach powered by machine learning. The legacy data was found to be inadequate for this purpose and therefore a soil data collection effort was launched in five of the catchments. This entailed soil profile descriptions and classification at various locations within each catchment, as well as collection of disturbed and undisturbed soil samples for laboratory analysis. Using this data together with the legacy data collected, HYDROSOIL maps were created with digital soil mapping methods for all six catchments. The soil maps were evaluated with an independent soil profile dataset. The maps were then used as improved soil information in modelling the hydrological response of each catchment. Hydrological modelling was performed for each catchment, using the previously best available soil information and the newly created digital soil map. The value of the HYDROSOIL maps

were expressed as the improvement in the model accuracy. Additionally, using existing and newly collected data, pedotransfer functions were created using machine learning whereby hydrological soil properties could be predicted with readily available soil properties. Near infrared spectroscopy was also investigated to determine whether it could be used to predict soil hydrological properties.

RESULTS AND DISCUSSION

The HYDROSOIL maps of all six catchments were deemed to be adequately accurate. Using the improved soil information, five of the six catchments recorded an improvement in modelling accuracy, showing the value of the HYDROSOIL maps. At each catchment a unique aspect regarding the HYDROSOIL map was investigated. This includes:

- Calibration of the SWAT hydrological model through optimizing model parameters based on the expected hydrological response of the hydrological response unit (Sabie-Sand).
- Determining a method to incorporate the land type field maps into creating a digital soil map (Olifants).
- Determining the effect of pixel size on the SWAT hydrological model outcome (Jukskei).
- Quantifying the effect of land use change on the hydrological regime (uMngeni).
- Modelling sediment yield using SWAT (Tsitsa).
- Using the JAMS hydrological model (Goukou).

The hydrological soil property predictions (both the pedotransfer functions and near infrared spectroscopy calibrations) did not yield acceptable results, most likely due to having too little data to adequately describe the soil variation within the six catchments used in the project. However, the following was determined during the study:

1. Local data is required to accurately predict hydrological soil properties, as internationally developed models (pedotransfer functions or soil spectroscopic calibrations) will not be able to account for the local soil variability.
2. Predictions created for smaller areas are generally more accurate than predictions for larger areas, most probably due to having less soil variability.
3. A lot more data is required to make accurate predictions of hydrological soil properties.
4. Sampling strategies to collect the required data should focus on smaller areas to produce useful prediction models, and over time sufficient data will be collected for predictions at a regional or national scale.

CONCLUSIONS

The project was created to learn lessons to apply on the future journey towards a hydrological soil map for South Africa. In terms of mapping and modelling, the project aims were met, indicating a need and use for a national HYDROSOIL map. A national HYDROSOIL map will be a valuable national asset that should be pursued. The methodology and capacity exist in South Africa to create the HYDROSOIL map.

The following lessons were learnt from this project:

1. A database structure and quality control measures were created whereby collected soil data of the future could be gathered and stored.
2. A method was developed whereby highly clustered data could be used to create an accurate soil map, without losing less-represented soil types.
3. How to digitise the approximately 200 000-300 000 soil observations recorded on paper copy maps stored at the Agricultural Research Council and use these in digital soil mapping to create the HYDROSOIL map.
4. A strategic approach to obtaining hydrological soil property data was determined.

It was currently not possible to create useful soil hydraulic property predictions from the HYDROSOIL map. This was due to a lack of data. The current costs associated with collecting hydrological soil data are the reason for the lack of sufficient data, and efforts should continue to explore ways to collect such data in a more timely and cost effective manner. This essentially provides the motivation to continue with this research.

RECOMMENDATIONS

Based on the lessons learnt during this project**, it is recommended that an adequately parameterised, accurate hydrological soil map (HYDROSOIL) for South Africa and its border catchments can, and should be created.** The following steps should be taken to achieve this goal:

1. Improve the soil database to become a cloud-based soil data repository with automatic quality control.
2. Digitise the land type field observations to be used in digital soil mapping to create the national HYDROSOIL map, using the methods determined in this project.
3. Apply the digital soil mapping methods used and newly learnt in this report to create the HYDROSOIL map.
4. Characterise the hydrological properties of the soils of South Africa, using hydrological soil measurements, pedotransfer functions and near infrared spectroscopy. This should be done by collecting data from smaller areas and first make useful local predictions. When sufficient data has been collected for the entire country, national prediction models should be created.

CAPACITY BUILDING

This project contributed towards four PhD degrees, two MSc degrees and three Honours level BSc Degrees. Three peer reviewed journal articles have already been accepted for publication, and another four are in preparation for publication.

ACKNOWLEDGEMENTS

The funding of the project by the Water Research Commission and the contribution of the team members is acknowledged gratefully.

The following members of the Reference Group are acknowledged for their valuable contributions and guidance:

Reference Group	**Affiliation**
Mr Wandile Nomquphu (Chair)	Water Research Commission
Ms Penny Jaca (Coordinator)	Water Research Commission
Dr Wietsche Roets	Department of Water and Sanitation (DWS)
Ms Anneliza Collett	Department of Agriculture Forestry and Fisheries (DAFF)
Dr David Clark	University of KwaZulu-Natal
Prof. Cathy Clarke	Stellenbosch University
Expert Contributions	
Dr Andrew Watson	Stellenbosch University
Editorial Support	
Dr Emily Botts	Independent

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS AND ABBREVIATIONS

ACRU	Agriculture Catchments Research Unit
ANN	Artificial Neural Network
ARC-SWC	Agricultural Research Council – Soil Climate and Water
AWC	Available Water Capacity
Bd	Bulk density
BI	Brightness Index
CEC	Cation Exchange Capacity
CI	Colouration Index
cLHS	Conditioned Latin Hypercube Sampling
CN	Curve Number
CSIR	Council for Scientific and Industrial Research
CV	Coefficient of Variation
DALRRD	Department of Agriculture, Land Reform and Rural Development
DEM	Digital Elevation Map
DSM	Digital Soil Mapping
DUL	Drained Upper Limit
DWS	Department of Water and Sanitation
EC	Electrical Conductivity
EFTEON	Expanded Freshwater and Terrestrial Environmental Observation Network
ET	Evapotranspiration
GIS	Geographic Information System
GPS	Global Positioning System
GSD	Ground Sample Distance
HRU	Hydrological Response Unit
HWSD	Harmonized World Soil Database
ID	Identity
ISRIC	International Soil Reference and Information Centre
IUSS	International Union of Soil Sciences
JAMS	Jena Adaptable Modelling System
KGE	Kling-Gupta Efficiency
K-means	K-means clustering
Ksat	Saturated hydraulic conductivity
LL	Lower Limit
MAP	Mean Annual Precipitation
ME	Mean Error
MLR	Multiple Linear Regression
MNLR	Multinomial Logistic Regression
MPC	Macropore Conductivity
MRRTF	Multi-Resolution Ridge Top Flatness

MRVBF	Multi-Resolution Index of Valley Bottom Flatness
NDVI	Normalised Difference Vegetation Index
NIR	Near-infrared
NIRS	Near Infrared Spectroscopy
NSE	Nash-Sutcliffe Efficiency
OC	Organic Carbon
OSSL	Open Soil Spectral Library
PBIAS	Percentage Bias
PLSR	Partial Least Squares Regression
PTF	Pedotransfer Function
RF	Random Forest
RI	Redness Index
RMSE	Root Mean Square Error
ROS	Random Oversampling
RPD	Ratio of Performance to Deviation
RPIQ	Ratio of Performance to Interquartile Distance
RS	Random Sampling
RUS	Random Undersampling
SAFCOL	South African Forestry Company Limited
SANBI	South African National Biodiversity Institute
SAWS	South African Weather Service
SCS	Soil Conservation Service
SI	Saturation Index
SMOTE	Synthetic Minority Oversampling Technique
SQL	Standard Query Language
SRS	Stratified Random Sampling
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
SWAT	Soil and Water Assessment Tool
SWOT	Strengths, Weaknesses, Opportunities and Threats
TEPA	Total Evaluation Point Accuracy
TMU	Terrain Morphological Unit
TWI	Topographic Wetness Index
USGS	United State Geological Survey
WISE	World Inventory of Soil Emission
WoSIS	World Soil Information Services
WRB	World Reference Base for Soil Resources
WRC	Water Research Commission

This page was intentionally left blank

CHAPTER 1: BACKGROUND

1.1 INTRODUCTION

Understanding and simulating internal catchment hydrological processes are becoming increasingly important to quantify the impacts of climate and land-use change in areas with highly variable water regimes, like most of southern Africa. Soil is a first order control of hydrological processes, as it splits precipitation into overland flow and infiltration (Park et al., 2001, and further influences the flow paths (deep drainage, lateral flow, etc.) which the infiltrated water will take. Therefore, the spatial distribution of soil properties directly influences the hydrological functioning of a catchment. To this end, it has been shown that improved spatial soil information not only improves hydrological modelling results at the catchment outlet, but also reflects internal catchment processes more accurately (Bieger et al., 2017 Van Tol et al., 2020). Although it is agreed that soil information is a vital input for physically-based hydrological models (Worqlul et al., 2018), this information is seldom adequately available in similar spatial detail as remote-sensed land-use, topography and climate data.

Currently the only spatial soil information covering the whole of South Africa is the land type database (Land Type Survey Staff, 1972-2006). The land type data is, however, not a soil map, but a compilation of 7 070 polygons called land types, each which demarcates an area with a “homogeneous, unique combination of terrain type, soil pattern and macroclimatic zone”, covering the entire country at a scale of 1: 250 000 (Paterson et al., 2015). The soil pattern of a specific land type is given in a land type inventory as an estimated percentage of coverage of a soil form, on a particular Terrain Morphological Unit (TMU). As it is the only national soil data source, it has been converted to hydrological parameters (mostly for the ACRU model), however in a lumped format, with averaged values representing the entire land type (Schulze, 2007). Recently, the Water Research Commission (WRC) funded research to improve the spatial detail of hydrological parameters to TMU-scale for quinary catchments (WRC proposal 2019/2020-00205). Although this is certainly a step in the right direction, care must be taken when the land type data is re-used or re-interpreted for smaller areas (see Van Tol & Van Zijl, 2020. Some of the reasons that this might be error prone are:

- Often, very dissimilar soil forms will occur in large proportions on the same TMU, rendering it virtually impossible to assign a hydrological response to a specific TMU, let alone the land type within which it occurs.
- The scale of 1: 250 000 at which it is published makes it useful for hydrological interpretation and modelling at small scales of large areas, but limits the use for small areas, where the impact of drastic land-use change (e.g. open-cast mining) on hydrological process should be predicted.
- The land type survey was done largely to determine the agricultural potential of South African soils, which meant a shallow observation depth either limited by root restricting layers or 1.2 m, whichever came first. However, the nature of the soil/bedrock interface plays an important role in generation of lateral flow or recharge of aquifers. The description of the soil/bedrock interface (and the depth at which it occurs) is largely unaccounted for in the land type information.
- Although the land type survey is based on a large number of soil observations, these were made only next to accessible roads, and therefore the quality of land types in certain areas with poor road access are often questionable. This is particularly true of mountainous terrain, often of great hydrological importance.
- As the land type survey was done over thirty years by a very large number of surveyors, the methodology followed differs between areas. Testament to this are the large land types of the Northern Cape and the Free State, in comparison to the small land types of KwaZulu-Natal.
- Further complicating the standardisation matters are the fact that two different soil classification systems were used (MacVicar et al., 1977; Soil Classification Working Group, 1991) in the land type survey. Both systems are currently out of print, which creates a knowledge gap to the hydrologists of the future who would require soil information. The newest edition of the South African Soil Classification system (Soil Classification Working Group, 2018) has a strong emphasis on describing

natural soils, which necessitates a dedicated effort to reinterpret the existing land types for hydrological purposes.

Another serious limitation in the regional soil information of South Africa is the unavailability of soil hydraulic parameters as inputs into the models. Estimated texture classes and depths of horizons (but only up to 1.2 m) are the only hydraulic parameters that accompany land type inventories. Significant efforts have been made to convert these, in association with soil forms and series, to model input parameters (mostly for the ACRU model). Currently the WRC is funding a novel project which aims to create soil input parameters for the SWAT model, but again this is at 1:250 000 scale (i.e. one set of lumped parameters for an entire land type). The cost associated with collecting and measuring hydraulic properties such as conductivity, water retention characteristics and porosity is mostly blamed for the absence of these properties. Researchers consequently rely on pedotransfer functions (PTFs) to predict hydraulic properties from available measured properties, such as texture. Most of these PTFs are only accurate for the area (soils) where they were developed. There is a great need to create and improve PTFs for South African soils. Hydraulic properties have, however, been measured in numerous research and consultancy projects. It is timeous that this data is mined from research reports, articles and theses to create and improve PTFs of South African soils.

The importance and limitations of the available soil information for South Africa are highlighted in the foregoing rationale. If soil information is not readily and freely available at appropriate scale it will continue to be a knowledge gap that will restrict efficient water management in South Africa.

To initiate the bridging of the knowledge gap left by the unavailability of hydrological soil information, the Department of Water and Sanitation hosted a workshop titled ‘Towards a detailed hydrological soil map for South Africa’ on the 20th of February 2020. The workshop was attended by 15 researchers representing 11 different research and government institutions. At the workshop the attendees agreed that:

1. The currently used best source of soil spatial data, the land type survey, is inadequate to address the hydrological challenges facing us today and in the future.
2. There is a need to improve PTFs for South African soils.
3. A detailed hydrological soil map of South Africa is urgently needed, using a digital soil mapping approach.

Digital soil mapping (McBratney et al., 2003) is a collective word used to describe advanced soil survey techniques which embrace advances of technology such as satellite imagery, digital elevation models and machine learning to produce soil maps at a fraction of the price and time of conventional mapping methods (Van Zijl et al., 2013). In South Africa different approaches have been proposed for different methods, with the machine learning approach ideal for the mapping of large areas (> 100 km^2) with a large amount of soil point data available (Van Zijl, 2019; Du Plessis et al., 2020). Digital soil mapping has the added advantage that it can be used to map soils accurately even in areas where access is limited (Van Zijl et al., 2012), which is a great advantage in mountainous regions. For hydrological purposes digitally produced maps have been used to great effect to provide improved soil information for hydrological modelling from small third order streams (Van Tol et al., 2015, Van Zijl et al., 2016) to fairly large (~ 650 km^2) catchments (Van Tol et al., 2020).

Internationally, digital soil mapping with legacy soil data have been used to create national soil maps for entire countries, including Australia, the United States of America and Denmark amongst others. Legacy data refers to soil data which has been previously collected. Such datasets have immense value to map soils today, but are generally unavailable hidden away in paper-based copies of theses, research reports or held by commercial companies. However, collecting such data into a readily available digital database will ensure its usefulness into the future, negating the need to collect such data again for mapping and other purposes. With this project a methodology will be developed by which a national hydrological soil map (HYDROSOIL) using digital soil mapping techniques and legacy soil data could be created. The methodology was developed and tested in priority areas of six ecologically and/or economically important catchments in South Africa. For each catchment the hydrology was modelled using the existing soil data (land types) and HYDROSOIL map to quantify the value of the soil map in each instance.

1.2 PROJECT AIMS

The following were the aims of the project:

1. Develop the methodology by which a national hydrological soil map (HYDROSOIL) could be created with digital soil mapping methods.
2. Compile a legacy soil point and hydrological properties database through data rescue of largely paper-based data sources currently contained in academic theses, research reports and with corporate institutions.
3. Create a HYDROSOIL map for each of the priority areas within six economically and/or ecologically important catchments of South Africa, using the database compiled in Aim 2. The catchments used in this project includes: The Sabie-Sand, The Olifants, the Jukskei, the uMngeni, the Tsitsa and the Goukou.
4. Develop pedotransfer functions by which hydrological parameters could be calculated from readily measured soil properties, using the database compiled in Aim 2.
5. Combine the hydrological soil maps created in Aim 3 with the pedotransfer functions developed in Aim 4 to create hydrological soil property maps for the study areas.
6. Determine the value of the HYDROSOIL map by hydrological modelling of all six study areas with both the best existing soils data, as well as the HYDROSOIL map.
7. Apply the HYDROSOIL map at specific sites for each of the catchments. This includes:
    - Calibration of a hydrological model through optimizing model parameters based on the expected hydrological response of the hydrological response unit (Sabie-Sand).
    - Determining a method to incorporate the land type field maps into creating a digital soil map (Olifants).
    - Determining the effect of pixel size on the model outcome (Jukskei).
    - quantifying the effect of land use change on the hydrological regime (uMngeni).
    - Modelling sediment yield (Tsitsa).
    - Using a different hydrological model (Goukou).

1.3 SCOPE AND LIMITATIONS

The scope of this project was to create six hydrological soil maps, one for each of the selected catchments, and for these maps to be used to improve the hydrological modelling for those catchments. Additionally, pedotransfer functions to predict hydrological soil properties were to be created. Both of these aims were to be met using largely existing or legacy soil data, for which a soil database was created.

The lack of available data was a severe limitation to the project, which led to the collection of soil data for five of the six catchments. Although this mitigated the need for data for soil mapping, the collected data was still not sufficient to create useful pedotransfer functions. Additionally, the lack of a gauging weir for model validation limited the modelling results for the Tsitsa catchment, while the lack of spatial rainfall data limited the modelling outcomes for the Goukou catchment.

CHAPTER 2: A SOUTH AFRICAN SOIL POINT DATABASE

Chapter 2 describes the process of determining and populating a soil point database, based on existing data. The crux of the work was to determine an optimal database format and put quality control measures in place when adding data to the database. It is envisaged that the database should become representative of the country's soils as new data will be added in the future. This work was primarily done by Molebaleng Sehlapelo as part of her MSc dissertation and has been prepared as a journal article to be submitted to the *South African Journal of Plant and Soil* (SAJPS, 2023).

2.1 THE STRUCTURAL DESIGN OF A ROBUST SOIL DATABASE FOR SOUTH AFRICA SOIL POINT OBSERVATIONS

2.1.1 Abstract

A lack of publicly available soil data has emerged as a factor that hinders sustainable food production, environmental protection, and policy formulations. To tackle the paucity of available soil data, an initiative must be taken to bring together all available soil data into a unified soil database. However, the development of a soil database requires a robust database structure design that ensures that data obtained from different sources is quality controlled, adequately stored and readily accessible. A Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis was conducted on the international World Soil Information Services (WoSIS) and the Agricultural Research Council – Soil Climate and Water (ARC-SCW) soil databases to propose an ideal structure for a national soil point database. Additionally, quality control measures required when adding data to the database were proposed based on the criteria used and adopted from the World Inventory of Soil Emission Potentials (WISE) and the African Soil Profiles Database version 1.1. Incorporating the strengths and opportunities while negating the weaknesses and threats associated with the analysed soil databases resulted in the development of a comprehensible and user-friendly soil point database. The database consists of quality-controlled soil point data, retrieved from 25 distinct sources, resulting in a total of 539 soil profiles and 1 518 soil horizons. The database serves as a platform for recording legacy soil data collected from diverse sources, documented in different formats, and created for various purposes.

2.1.2 Introduction

Soil is a non-renewable natural resource, difficult to rehabilitate and costly to restore or cultivate after erosion, physical deterioration, or chemical contamination (Pozza & Field, 2020). The deterioration and pollution of soil is often caused by pressure on land and results in a decrease in crop production capacity (Jie et al., 2002). Pressures on land may lead to soil degradation that may further result in soil loss or a decrease in soil functions (Oldeman, 1992). Therefore, it is important to protect and restore soil quality to maximise these soil functions (Lal, 2015). A general understanding of the composition, characteristics, dynamics, and functions of soil plays a role in the effective management and sustainability thereof (Schoover & Crim, 2015). The availability of accurate information regarding soil dynamics and characteristics obtained through analysis and description of soil in the field is the basic prerequisite for achieving effective management and sustainability of soils (Jahn et al., 2006).

The availability of detailed soil information plays a role in the implementation of policy for soil protection (Breure et al., 2012). Due to the continuously increasing pressures on soils, such as soil erosion and compaction in South Africa, the Department of Agriculture, Land Reform and Rural Development (DALRRD) was assisted by the Agricultural Research Council (ARC) to draft a policy to promote the protection of soil and the proposed Preservation and Development of Agricultural Land Framework Act (Paterson et al., 2015). However, for this bill to pass into legislation, it was necessary to have detailed soil information covering the whole of South Africa (Paterson et al., 2015). This is because a soil database provides relevant information for national and regional planning, with a detailed risk assessment of the rate of soil degradation, land assessment and land-use planning mechanisms (Oldeman & Van Engelen, 1993). Thus, it is necessary to develop an authoritative national database of soil information for South Africa.

As the amount of data and data users continue to increase, a systematic approach for database development is required (Fernández & Rusinkiewicz, 1993). The development of a database is associated with a well laid-out structure, an accurate description of its contents and specifications on the quality of the data to be recorded. There are several considerations when developing a database, such as addressing user requirements and eliminating obsolete data. Consequently, ensuring data quality and standardisation is crucial for effective data interpretation (Oldeman & Van Engelen, 1993). There are several stages involved in designing a soil database. The first stage is gathering data from various sources (Shangguan et al., 2014). The second stage involves data harmonisation using data processing procedures (Shangguan et al., 2014). Stage three encompasses data standardisation, which entails processing data from many sources to make it identical to allow data comparability during data analysis (Quevauviller, 1998). The resulting database should present data that are easily comprehensible, accessible, and feasible (Grealish et al., 2004).

Understanding the different aspects involved in database design can be achieved by analysing different databases to compare the structure, information recorded, accessibility, data usage and standardisation of the data within them (Gupta, 2000). The legibility of the database can be assessed by examining the type of software used to store data, the format in which the data are stored and presented, and the ease of navigating through the database to find data (Shofiyati et al., 2011). The quality of the database depends on the accuracy of the attributes used to provide information, like geographical coordinates and units and methods of measurements (Hoffmann et al., 2020; Paterson et al. 2015). Data accessibility can be assessed by the ease of obtaining the database and retrieving the information recorded in the database (Pangos et al., 2011). Data usage can be evaluated by analysing how easy it is to understand the data so that it can be used in other software packages, to expand the database so that new data can be populated, and the ability to perform quality control measures on the recorded data (Shofiyati et al., 2011). Finally, data standardisation can be evaluated by examining the units and methods of measurement, and classification systems used, as well as the database's ability to integrate with other databases (Ribeiro et al., 2015).

This comparison can be achieved through a SWOT analysis, which is a theory that refers to the Strengths, Weaknesses, Opportunities, and Threat factors (Nikolaou & Evangelinos, 2010). These factors are used to identify the internal and external factors that enhance or interfere with the performance of an initiated plan or project. Strengths and weaknesses are characterised as internal factors, while the opportunities and threats are characterised as external factors (Leigh, 2010). The analyses of these factors are based on the estimation of their contributions to reaching a certain goal and the approximation of their controllability (Benzaghta et al., 2021).

During compilation of a soil database, the data are mainly categorised into geographic and attribute data. Geographic data provide information about the location, extent and topology of each soil profile while the attribute data provide information about the soil physical and chemical properties of each soil profile (Bouma et al., 1999). Databases created from well-established, standard soil survey practices are characterised into two main groups of soil information – primary and secondary data (Bouma et al., 1999). Primary soil data are established through sampling and remote sensing (Mulder et al., 2011). while secondary data are established through continuous pedotransfer functions (Reddy & Das, 2023). Furthermore, these are sub-categorised into topographical data, soil fertility data and hydrological data (Bouma et al., 1999). Soil fertility data represent the ability of soil to sustain growth and improve production through continuous provision of nutrients (Hartermink, 2007). Hydrological data represent the movement, retention, and loss of water in the soil used for understanding physical and chemical processes (Vereecken et al., 2015). Lastly, topographic data refers to digital information about the relief of a terrain (Carter, 1988).

Challenges of soil data quality are not limited to the concept of uncertainty, but quality is considered a function of completeness and consistency (Hortensius & Welling, 2008). Several methods can be utilised to establish the accuracy, compatibility, and traceability of measured soil data. These include implementing and maintaining a quality management system in the laboratories and application of methods used for validation

and standardisation. Furthermore, the use of reference materials (that have been certified), participation and organising evaluations in inter-laboratory experiments are also essential (Theocharopous et al., 2004). Additionally, a variety of soil properties, including positional accuracy, attribute accuracy, logical consistency, completeness, and lineage properties may be used to assess data quality (Theocharopous et al., 2004). Utilising well-documented procedures and standards to compile and process large-scale soil data is essential. The successful compilation of a systematic soil database requires reliable sources of data and standardised methods of acquiring, processing, and storing this information (Batjes, 1995). Once this database has been obtained it would be valid for many years and represents a strategic, once-off investment (Bouma et al., 1999).

There have been ongoing discussions about the lack of a freely available soil point database in South Africa. The need was first expressed in 2014 at the Soil Information workshop hosted in Stellenbosch. An initial overview of the available data was created (Paterson et al., 2015), and a project instigated to develop a Soil Information strategy for South Africa (Collett & Rozanov, 2018). This culminated in an additional workshop and discussion held in Pretoria in 2018 titled “Soil information for sustainable development, agricultural conservation and land use policies in South Africa”. In 2023, a soil policy formulation document was supported by the Soil Science Society of South Africa, which proposed the primary goal of creating a national soil database (Rozanov et al., 2023). Despite all this effort and the numerous expressions of the need for such a database, there is still no functional, freely available soil database to which soil data could be submitted, with the necessary quality control in place.

The aims of this research were (1) to develop a robust structural design for a soil point database to record soil observation points collected in South Africa, and (2) to determine quality control measures that can be applied during the population of the soil data into the database. Data was gathered from various sources to populate the database, allowing the structure and quality control measures to be tested.

2.1.3 Materials and methods

To develop a soil database structure, an international and a national database were subjected to the SWOT analysis, by assessing the data legibility, quality, accessibility, usage, and standardisation. Additionally, quality control measures used and adopted by Batjes (2008) and Leenaars (2013) during the creation of the WISE and African Soil Profiles Database version 1.1. were applied during the population of soil points into the database, to ensure harmonisation and standardisation of the database and to improve the quality thereof.

SWOT analysis

To develop a robust soil database structure, SWOT analyses of leading national and international databases were evaluated to observe the legibility, quality, accessibility, usage, and standardisation for the different databases. The two databases that were selected include the World Soil Information Services (WoSIS, ISRIC, 2023) and the Agricultural Research Council – Soil Climate and Water (ARC-SCW Soil Database. 2014) databases. These databases were chosen because each database was used to record data for different geographical spaces and purposes resulting in very distinct structures. Furthermore, the WoSIS and the ARC-SWC are the leading international and national custodians of soil databases (Ribeiro et al., 2015; Paterson et al., 2015).

Data quality control

Quantitative soil properties to be recorded in a comprehensible soil database must be quality controlled and this process was characterised in two stages (Figure 2.1). The first stage of quality control involved removal of all the values detected as outliers from the basic quality control criteria as adopted from the WISE 3 (Batjes, 2008) and African Soil Profiles Database version 1.1 (Leenaars, 2013), which included identifying and excluding unknown geographical coordinates or soil point coordinates not falling within the borders of the study area of collected soil point observations (South Africa) (Leenaars, 2013). Furthermore, quality control involved the identification and removal of soil properties falling outside certain ranges of values (Table 2.1).

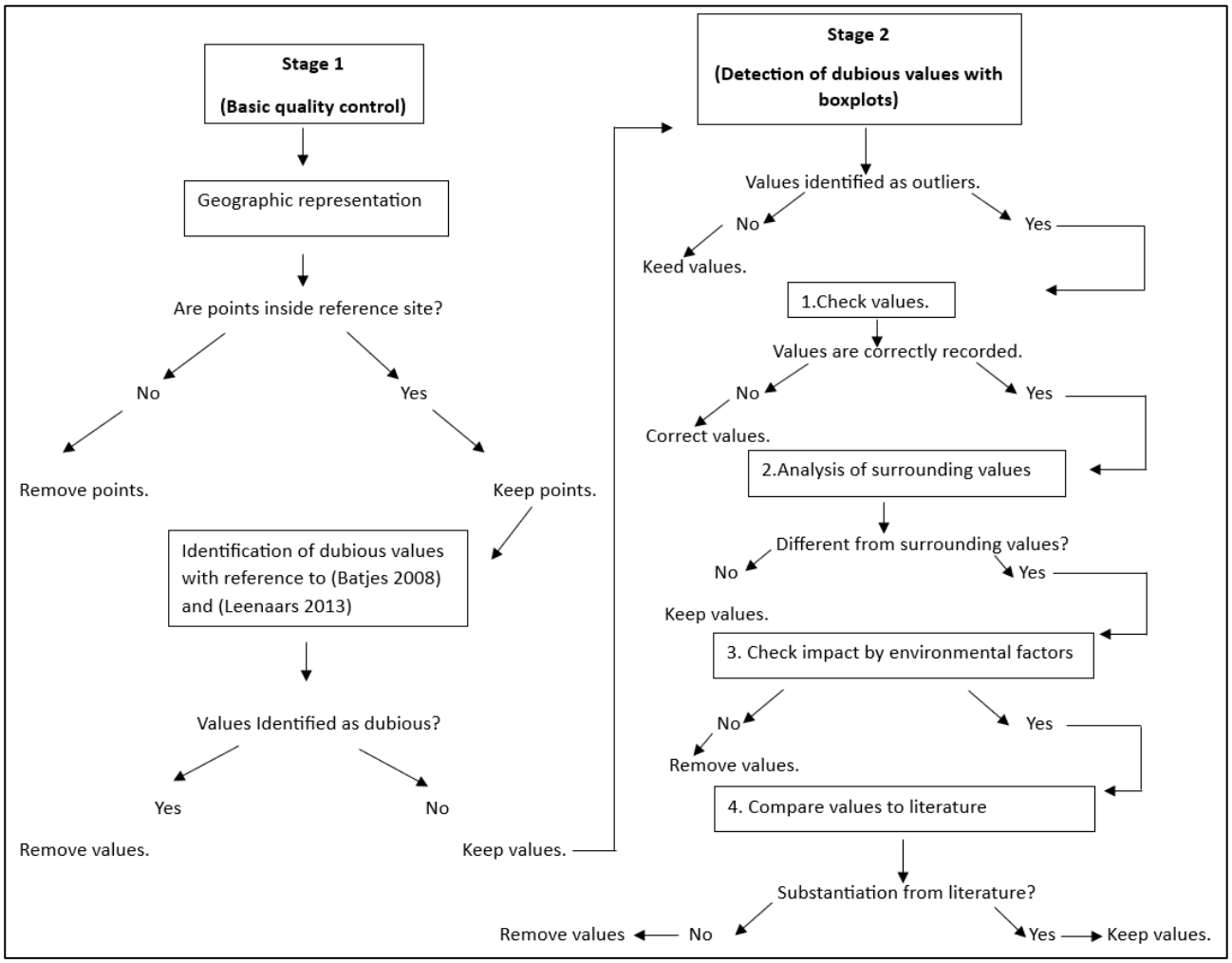

Figure 2.1: Stages involved in the quality control.

Table 2.1: Soil property values considered for basic quality control criteria adopted from Batjes (2008) and Leenaars (2013).

Soil property	Range of values	References
Sum of Sand, Silt, Clay Fractions (%)	> 90% and < 110%	Leenaars (2013)
Sand fractions (%)	> 0% and < 100%	Leenaars (2013)
Silt fractions (%)	> 0% and < 100%	Leenaars (2013)
Clay fractions (%)	> 0% and < 100%	Leenaars (2013)
Bulk Density (Pb)	> 0,1g/cm^3 and < 2,7 g/cm^3	Leenaars (2013); Batjes (2008)
Exchangeable Calcium (Ca)	> 0 cmol/kg and < 200 cmol/kg	Leenaars (2013)
Exchangeable Magnesium (Mg)	>0 cmol/kg and < 50 cmol/kg	Leenaars (2013)
Exchangeable Sodium (Na)	> 0 cmol/kg and < 200cmol/kg	Leenaars (2013)
Exchangeable Potassium (K)	> 0 cmol/kg and < 20 cmol/kg	Leenaars (2013)
Sum of Exchangeable Bases	> 150 cmol/kg	Leenaars (2013)
P-status	> 0 mg/kg and < 1 000 000 mg/kg	Leenaars (2013)
pH (H_2O)	> 2 and <12	Leenaars (2013); Batjes (2008)
pH (KCl)	> 2 and < 12	Leenaars (2013); Batjes (2008)
Cation Exchange Capacity (CEC)	> 1 cmolc/kg and < 150 cmolc/kg	Leenaars (2013)
Electrical Conductivity (EC)	< 0 mS/m	Leenaars (2013)

The second stage of quality control (Figure 2.1) involved the detection of outliers with boxplots. If any outliers were detected, four mandatory steps were followed before the exclusion of the values, because outliers detected within a dataset may not be erroneous anomalies with the potential to affect the quality of the dataset (Filzmoser & Gregorich, 2010). The first step was reviewing values, which involved rechecking the data source to ensure that the values were copied correctly. In the second step, a value detected as an outlier would be analysed and compared to the surrounding values to investigate whether there were any similarities, differences or trends. For any value that was detected as an outlier and was different from the surrounding values, the third step involved further analysis, where the geology and terrain were investigated as these environmental factors have the potential to alter the chemistry (chemical properties) of soil (Djodjic et al., 2021). The fourth step involved comparison of values with those observed in the literature.

Data population

The minimum requirement for inclusion of soil data into the soil database to ensure data legibility was the name of the provider and at least one soil attribute. Coordinates were not included as a prerequisite because data collected without coordinates could still be valuable for research, for instance in creating pedotransfer functions. Furthermore, minimum data requirements for soil profiles included information that aids soil classification, such as soil colour, texture or structure.

Soil data was collected from different sources and populated into the soil database. For data entry and collation, Microsoft Excel worksheets were utilised to ensure pragmatism and speed. Unique profile identities (IDs) were used when recording soil data to avoid duplication. These IDs referenced the original profile IDs used in different data sources. Datasets recorded in Portable Document Format (.pdf) documents were copied using the “copy” and “paste” functions, although this was time consuming and prone to human error. The soil database was composed from three academic transcripts, three research reports, two databases and 18 reports (15 irrigation suitability reports and three hydropedological survey reports). These resulted in a total of 567 soil profiles and 1 518 soil horizons.

Harmonisation and standardisation methods were carried out to incorporate all the soil data with different methods and expressed using a variety of units of measurements into the soil database (Table 2.2). Data harmonisation enables the integration of spatial and attribute data into a unified system, allowing for comparable representation of data from various sources (Sulaeman et al., 2013). Data standardisation involves assembling data into a common format to eliminate variation in classification, terminology, and measurements, for better data quality and interpretation (Hortensius & Nortcliff, 1991). Prior to population, each dataset was analysed to observe the units and methods of measurements. If the measuring units were different from the standard measurement units used in the soil database, a conversion process was carried out where possible to ensure standardisation (Table 2.2).

Table 2.2: Methods of data standardisation caried out.

Methods of standardisation
Calculation of coordinates in ArcMap 10.7.1 from shapefiles
Conversion of degrees, minutes, seconds to decimal degrees
Conversion of exchangeable cation values from mg/kg to cmolc/kg
Conversion of profile and horizon depth from cm to mm

2.1.4 Results and discussion

SWOT analysis of World Soil Information Services database

SWOT analysis of the WoSIS database (Table 2.3) was used to propose an optimal structure for a soil database.

- **Strengths**
  - *Legibility: Data can be imported into SQL (Standard Query Language) and statistical software.* All the soil information that makes up the WoSIS database is recorded in four different Tab Separated Value (.tsv) files. These files can be imported into SQL database or statistical software such as R where they may be joined using the profile_id. Importation of the dataset allows for handling and querying of the data as required by the user. Guidelines of this procedure are provided in the WoSIS Procedures Manual (Ribeiro et al., 2020).
  - *Data quality: Availability of information on database.* There are articles, reports and documentation about the database. The contents include definitions of soil properties and a set of attributes that can be used to express a description on a measurement. Downloads and links, contact details for enquiries and the source of the dataset can also be found on the International Soil Reference and Information Centre (ISRIC) World Soil Information page www.isric.org/data/data-download
  - *Accessibility: Database available online.* The WoSIS database is readily available online. The ISRIC World Soil Information page provides information on the standardised dataset as derived from WoSIS database, accessibility of the database and tutorials on how to use the R software to view and analyse the data.
  - *Standardisation of data: Methods described in columns*. The methods of measurement are readily described in the same table and column as the soil attribute, this results in a comprehensible and user-friendly database. There is also a specification of the soil classification used for each set of data.
  - *Data usage: Availability of link table.* Linking the different soil property tables allows easier navigation through the database, the user does not need to navigate back and forth through each file to use the data. Furthermore, the ease with which each .tsv file can be converted into an Excel spreadsheet improves the tabular presentation of soil attributes and the use of the database.
- **Weaknesses**

- *Legibility: Large amount of data.* There is a large volume of data recorded in many columns, which makes the data appear clustered resulting in poor data presentation and difficulty working through the database.
- *Standardisation of data: No specified units of measurements.* The units of measurements are not readily described in the same table and column as the soil attributes, searching for this information in other tables and/or documents can be time-consuming for the user.
- *Data usage: Data saved in .tsv files.* The users are limited due to the specialised data file formats used, which requires special skills such as R programming or conversion into Excel to access data.
- *Data spread over different files.* The data is spread over four soil attribute files, which makes it difficult to simultaneously use data from the same observation. This is overcome by linking the data in an SQL with the same identifier. Ideally, the user would want to have all the data in one file as not everyone is familiar with using SQL.
- *Unfamiliar column headings*. The headings are in codes and to overcome this, there is a separate file that provides the definitions. However, it is more effective to use a database that provides all the necessary information in a common file.

- **Opportunities**
  - *Data usage: Smaller geographical based databases.* Developing smaller databases based on the different geographical areas will result in a more comprehensible database because several columns will be omitted resulting in a single file for each database. This will result in easier navigation and lead to the use of familiar languages, codes and classification used for that area, resulting in a wider usage of the database.
- **Threats**
  - *Data usage: First time users could be deterred.* The use of SQL language, heading codes, .tsv file formats may be unfamiliar to first time users and the need to consult various documents for clarity may be time consuming and result in a limited number of users.
  - *Lack of quality control.* The accuracy of the database cannot be measured by first time users due to units of measurements being recorded in other files.

Table 2.3: Summary of the SWOT analysis of the World Soil Information Services (WoSIS) database.

Strengths (S)	**Weaknesses (W)**
.tsv files can easily be imported into an SQL database or statistical software such as R, after which they may be joined using the unique profile ID.	Large amount of data, incomprehensible and difficult to navigate.
Availability of information on the database.	No specified units of measurements in tables.
Database readily available online.	.tsv files are unfamiliar to most potential users.
Methods used are described in columns.	Data spread over different files.
Availability of link table to connect the profile to a given source.	Unfamiliar column headings.
Opportunities (O)	**Threats (T)**
Smaller geographical based datasets could improve usage considerably.	First time users could be deterred by unfamiliar file types and headings, and various documents to be consulted. Lack of quality control due to missing units of measurements.

SWOT analysis of the Agricultural Research Council – Soil Climate and Water database

SWOT analysis of the ARC-SCW database (Table 2.4) was used to propose an optimal structure for the soil database.

- **Strengths**
  - *Legibility: Database stored in Microsoft Access.* The ARC-SWC uses the database management system Access to store soil data. Access enables the user to perform queries

to acquire specific information from the database and to analyse the relationship between the recorded soil parameters. The program provides different options such as viewing or printing existing data, new registration of soil profiles, lab data requests or updating of existing data.

- *Database available in Excel.* The database can be opened and extracted into Excel, which is a program potential users of the database are familiar with. Excel presents the soil data in the database in a comprehensible manner, where a soil profile is recorded with accompanying horizons in sequential rows while the specific soil properties are recorded in designated columns. Furthermore, data in Excel can be analysed and manipulated as desired without changing the structure of the original database.
- *Data presented as .pdf printouts per soil profile.* When data from Access is requested, the option to obtain soil data by printing .pdf documents is available, which could be useful in certain situations, such as teaching and soil profile discussion. The .pdf documents present data in a table format, this results in easily comprehensible soil information. This soil data was clearly recorded, with each soil attribute in designated columns and clearly described as headings and the attribute features described as subheadings in the table.
  Data usage: Data entries and queries in Access. Access enables the user to perform queries to acquire specific information from the database and to analyse the relationship between the recorded soil parameters. Furthermore, the data entry program is easy to use as it provides drop down menus to enter the relevant information required for each soil profile. This information is automatically recorded in the Access database.
- *Availability of a user manual*. There is (Van Waveren & Bos, 1988) outlining the necessary steps for processes involved in data entries, queries and data acquisition. Additionally, the user manual provides information on the minimum data requirements for each soil profile.
- *Data quality: Units of measurements included.* There is a unit of measurement for each soil attribute presented on the .pdf printouts, resulting in easy comprehensibility and the performance of quality control to check for accuracy of the data recorded.
- *Availability of data validation tests.* The program provides various data validation tests that can be run on the recorded soil attributes to ensure data accuracy.
- *Standardisation of data: Necessary requirements for data entry provided.* The user manual provides guidance on the necessary information required before data entry. These include a soil classification, horizon definition, geographic location, and map sheets. This information improves the process of data standardisation during data unification.

- **Weaknesses**
  - *Accessibility: Database not readily available.* Contact via email had to be made to the provider to acquire soil data and information about the original structure of the database. This can sometimes be time-consuming because obtaining this information is dependent on the time it takes for the provider to reply to the request and other enquiries which may follow. For research purposes it can be motivated to obtain the data free of charge, however, for other purposes the data must be bought at a fee.
  - *Data quality: Description documents separated from database.* The description documents providing information about the soil attributes and the user manual providing information on how the recorded data can be used are provided. However, these are saved in separate documents, and the information in these documents is not available in the database. As a result, the user may have to open multiple documents and the database at the same time, which may cause confusion and may be time consuming.
  - *No analytical methods.* Although the main database gives descriptions of analytical methods, there is no description of these methods in the table.
  - *Data usage: Specialised skills required.* Although Access is an effective program to use for creating databases, not everyone is familiar with the functions involved in data queries. The effective use of these queries requires the user to partake in courses in Access for a clear understanding of the different functions. To illustrate this weakness, even the ARC-SCW has a limited number of employees with Access skills, which causes delays when requesting data when the skilled persons are not available.

- *Standardisation of data: Analytical methods not readily recorded in the table.* Methods used to measure the acquired soil properties and metadata such as analytical methods are not readily presented in the same table with the measured soil properties. This information is recorded in a separate document.

- **Opportunities**
  - *Accessibility: Data availability would increase usage.* Availability of this database would increase the opportunity for farmers, corporations, and researchers to readily have soil data that can help them with information required for productivity and sustainability, resulting in the wider usage of the database.
- **Threats**
  - *Accessibility: Database not used due to being unavailable.* The ARC-SCW is created for the provider's use and is not readily available to the public. As this database is not freely available online, it must be requested from the ARC-SCW at a cost for most purposes, which drastically reduces the usage of the database. Non-usage of such a database could render it obsolete as potential users will find different ways of obtaining the data required for their needs.
  - *Resignation of skilled Access users.* There are a limited number of skilled Access persons employed by the ARC-SCW. Therefore, should these persons resign, the lack of Access skills at the institution will render the database totally useless.

Table 2.4: Summary of the SWOT analysis of the Agricultural Research Council – Soil Climate and Water (ARC-SCW) database.

Strengths (S)	**Weaknesses (W)**
Database stored in Access. Database available in Excel. .pdf printouts for specific soil profiles available. Availability of a user manual. Units of measurements included. Availability of data validation tests. Necessary requirements for data entry provided.	Database not readily available. Description documents stored separate from database. No analytical methods presented in the table. Specialised skills required to extract and query data from Access.
Opportunities (O)	**Threats (T)**
Data availability would increase usage.	Database not used due to being unavailable. Resignation of skilled Access user will result in the database being unavailable.

Proposed structure of a robust soil point database

The SWOT analysis revealed various factors influencing the effectiveness of the WoSIS and ARC-SCW databases, highlighting both strengths and weaknesses. The legibility of a database was characterised by strengths such as the simplicity of importing data into a common database and other software, along with the use of table formats for data presentation, enhancing comprehensibility. However, weaknesses were identified, particularly in cases where recording large volumes of data into a single database led to cluttered information, hampering overall legibility. Concerning data quality, strengths were the availability of documents providing crucial information about the database, including information on units and measurement methods for each soil attribute, as well as the analytical methods employed. Conversely, weaknesses in data quality were associated with a lack of descriptive documents. The analysis also explored aspects related to data usage, noting strengths in a database's ease of navigation, reading, editing, analysis, importation, querying, quality control and data comparison. However, weaknesses were the complexity of the software used for data storage and recording, as well as the language, codes, and classifications employed, which could impact a database's usability for different users. In terms of accessibility, a strength was the ease of obtaining the database, while

a weakness was the lack thereof. Standardisation of the databases was promoted by the availability of units and measurement methods for each soil attribute, contributing to quality control.

From the SWOT analysis of the two databases, the strengths were incorporated and weaknesses negated to achieve a comprehensible and user-friendly structural design for the proposed soils database (Figure 2.2). The database was developed in an Excel spreadsheet with table headings displaying soil information with accompanying soil attributes, units and methods of measurement. Excel makes reading and editing easy. Six main soil attributes were recorded under the different table headings (Figure 2.2). The first table records the profile ID for soil profiles (Figure 2.2a). The second table contains landform and topography information of the soil profiles (Figure 2.2b). Soil morphological and physical properties of the soil profiles are shown under the third table (Figure 2.2b). The fourth table records the chemical properties of the soil profiles (Figure 2.2c, Figure 2.2d). The fifth table shows information on the hydrological properties and the sixth table was used to record geological properties (Figure 2.2e).

The table format optimised data presentation and comprehensibility, thus ensuring the legibility of the database. The database consists of six worksheets, one of which is used to describe all the attributes, units and methods of measurements, symbols, and a reference list of the source data. Each recorded attribute is accompanied by a unit of measurement displayed in the same column, thus improving the data quality. Soil attributes were recorded according to the method used for analysis. Therefore, there are options for different methods of measurements for a single soil attribute in the tables.

The database is made freely available upon request from the developer, to facilitate accessibility. Furthermore, the names of the soil attributes, units and methods of measurements used in the soil database are commonly used in soil sciences throughout South Africa. The soil database is used to record and store soil data for the geographical space of South Africa leading to the use of common languages, codes and classifications. Data recorded in the database can be easily imported and exported into other software using the “copy” and “paste” options, thereby promoting increased data usage.

(a)

	PROFILE ID						LANDFORM AND TOPOGRAPHY		
Profile No.	Profile cd.	DMS Coordinates		Name of provider	Lab name/code	Date	Slope%	Aspect	Elevation (m)
		X	Y						
1	1718	27,566666	-27,01667	ARC-ISCW	-	1921/02/02	1	South	1356
1	1718	27,566666	-27,01667	ARC-ISCW	C4362	1921/02/02	1	South	1356
1	1718	27,566666	-27,01667	ARC-ISCW	C4363	1921/02/02	1	South	1356
1	1718	27,566666	-27,01667	ARC-ISCW	C4364	1921/02/02	1	South	1356

(b)

SOIL MORPHOLOGICAL AND PHYSICAL DESCRIPTIVE PROPERTIES																			
Profile No.	Profile cd.	Soil Form	Soil Family	Master Horizon	Diagnostic Horizon	Depth (mm)	Auger Refusal	Soil Texture			Description								
								Clay %	Sand %	Silt %	Colour (dry)	Colour (wet)	Colour cd. (dry)	Colour cd. (wet)	Structure	Texture Class	Consistency	Transition	Other comments
1	1718	Glencoe		O		270													
1	1718	Glencoe		A1		550		5,70	92,30	0,60		yellowish brown		10YR4/4	massive	fine sand	friable		
1	1718	Glencoe		B21	Red Apedal	1000		8,80	85,90	4,80		yellowish brown		10YR5/6	single grain	loamy fine sand	friable	abrupt wavy	fine <2-6mm sesquioxide concre
1	1718	Glencoe		B22		1100	not reached	9,00	86,10	3,30						loamy fine sand			ferricrete

(c)

CHEMICAL PROPERTIES																							
Profile No.	Profile cd.	NH4OAc Cations (cmol(+)/kg soil)				Mehlich 3 Cations (mg/kg)				di-ammonium EDTA Micro Nutrients (mg/kg)					DTPA Micro Nutrients (mg/kg)					Mehlich 3 Micro Nutrients (mg/kg)				
		Na	K	Ca	Mg	Na	K	Ca	Mg	Zn	Mn	Cu	Co	B	Zn	Mn	Cu	Co	B	Zn	Mn	Cu	Co	B
1	1718																							
1	1718	0,00	0,10	4,40	0,10					0,09	0,80	0,65	0,14	0,47										
1	1718	0,00	0,10	0,20	0,10					0,44	0,10	0,69	0,03	0,32										
1	1718	0,00	0,00	1,90	0,10					0,18	0,30	0,78	0,15	0,32										

(d)

CHEMICAL PROPERTIES																
Profile No.	Profile cd.				Bray Test (mg/kg)			pH		CBD			CEC (cmol(+)/kg soil)	EC (mS/m)	WBM-Black Organic C %	Dry combustion OC%	Organic Carbon %
		MBM P(mg/kg)	Olsen P(mg/kg)	Mehlich 3 P(mg/kg)	1	2	KH2PO4 P-Soption (%)	H2O	KCl	Fe%	Al%	Mn%					
1	1718																
1	1718	0,90					27,04	5,10		0,37	0,04	0,00	1,70		0,30		
1	1718	0,40					63,04	4,80		0,54	0,07	0,00	2,30		0,20		
1	1718	0,70					87,49	5,80		2,11	0,22	0,00	3,00		0,20		

(e)

		HYDROLOGICAL PROPERTIES									GEOLOGICAL PROPERTIES	
Profile No.	Profile cd.	Water retention(%)	Bulk Density (g/cm3)	Bulk Density (mg/cm3)	AWR	Hydraulic conductivity (mm/ hr^{-1})					XRD MINERALOGICAL PROPERTIES	Parent material and/or Supergroup
						0mm	30mm	80mm	150mm	Average	Mineral symbols (um)	
1	1718											
1	1718	2,97			0,20						Qz, Fsl, Tcl, Kt,Tc, Mi, Ch, Vm	
1	1718	4,40			0,10						Qz, Fsl, Tcl, Kt,Tc, Mi, Ch	
1	1718	6,35			0,10						Qz, Tc, Fsl, Kt, Mi, Ch	

Figure 2.2: The complete soil database structure. Soil profile data is recorded in a single worksheet, and six main soil attributes recorded under different table headings: (a) Profile ID, landform and topography, (b) soil morphological and physical descriptive properties, (c) chemical properties, (d) chemical properties (cont.), and (e) hydrological and geological properties.

.

Data quality control

The soil point data collected included 567 soil profiles and 1 518 soil horizons from all nine provinces in South Africa (Figure 2.3) and characterised by a total of 58 soil forms. Between the different sources, data was collected for various geographical, topographical, morphological, physical, chemical, hydrological and geological soil properties. All the data populated in the soil database were subjected to quality control measures. Quantitative soil properties, including bulk density, soil texture, exchangeable cations, micronutrients, phosphorus status, phosphorus-sorption, pH, metal ions, electricity conductivity, organic carbon, water retention, relative soil saturation, hydraulic conductivity, and saturated hydraulic conductivity, were quality controlled. If a data entry was proven to have dubious values, and the rest of the soil attribute passes the quality control, only that soil attribute value is omitted. This is because the rest of the soil attribute values can be used for other analysis, including pedotransfer functions.

From basic quality control analysis, a total of seven outliers were detected from the geographical coordinates data, where the soil points were not plotted within the borders of South Africa. This was quite visible from the map (Figure 2.3). Furthermore, a total of seven soil point values were flagged as the points expressed values of Cation Exchange Capacity (CEC) exceeding 120 cmol/kg. The sources were rechecked to verify if the coordinates and CEC values were recorded properly. As this was the case, these soil attribute values were omitted from the database. A total of 12 soil horizons each with seven texture fractions resulted in the sum percentage of particle size exceeding 110%. As a result, 84 soil attribute values were identified as outliers and flagged to be investigated further in the second stage of quality control.

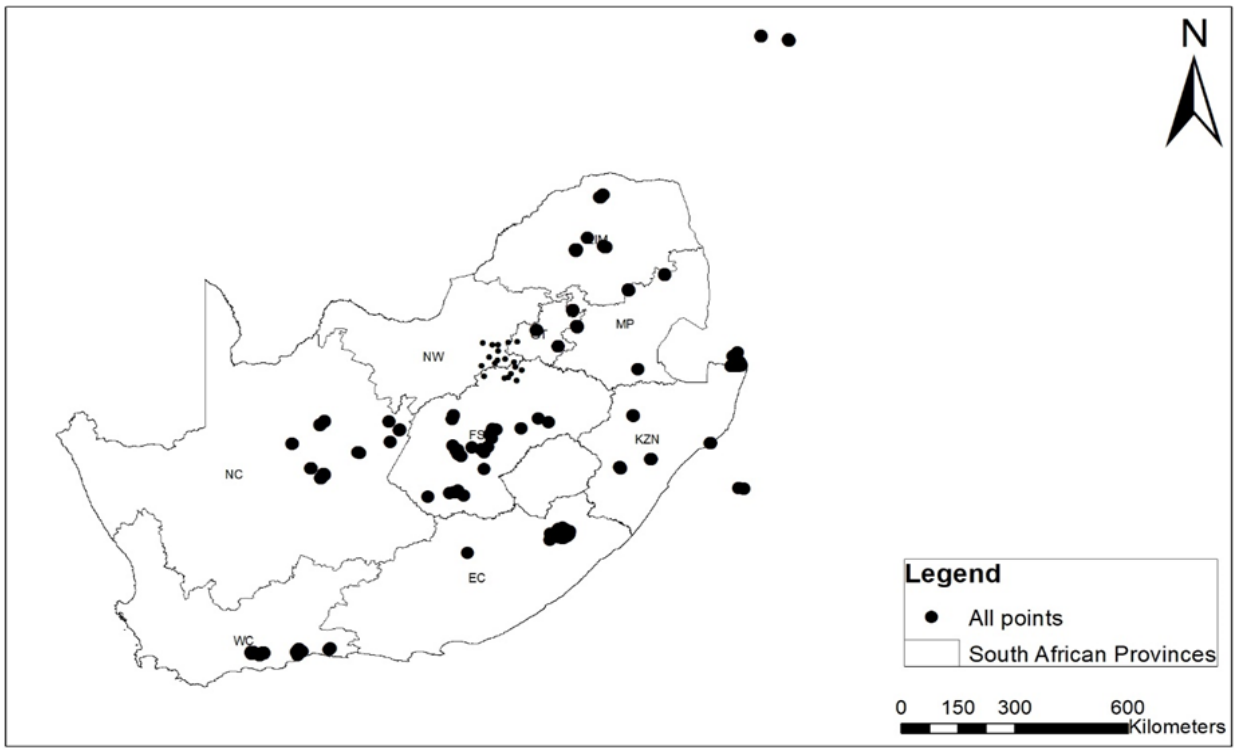

Figure 2.3: Map created for basic quality control analysis.

Boxplots were created for all quantitative soil properties during the second stage of quality control. No outliers were detected for bulk density (Figure 2.4a). However, a few values were detected as outliers for soil texture (Figure 2.4b, Figure 2.4c). A total of 45 texture fractions were detected as outliers based on the surrounding values, however, these values were not removed as the sum percentage of each

recorded soil horizon was well within the designated criteria for total sum percentage of texture fractions. The texture fractions of the 12 soil horizons flagged as outliers in the basic quality control stage were also detected as outliers in this stage of quality control. Therefore, these values were removed from the soil database.

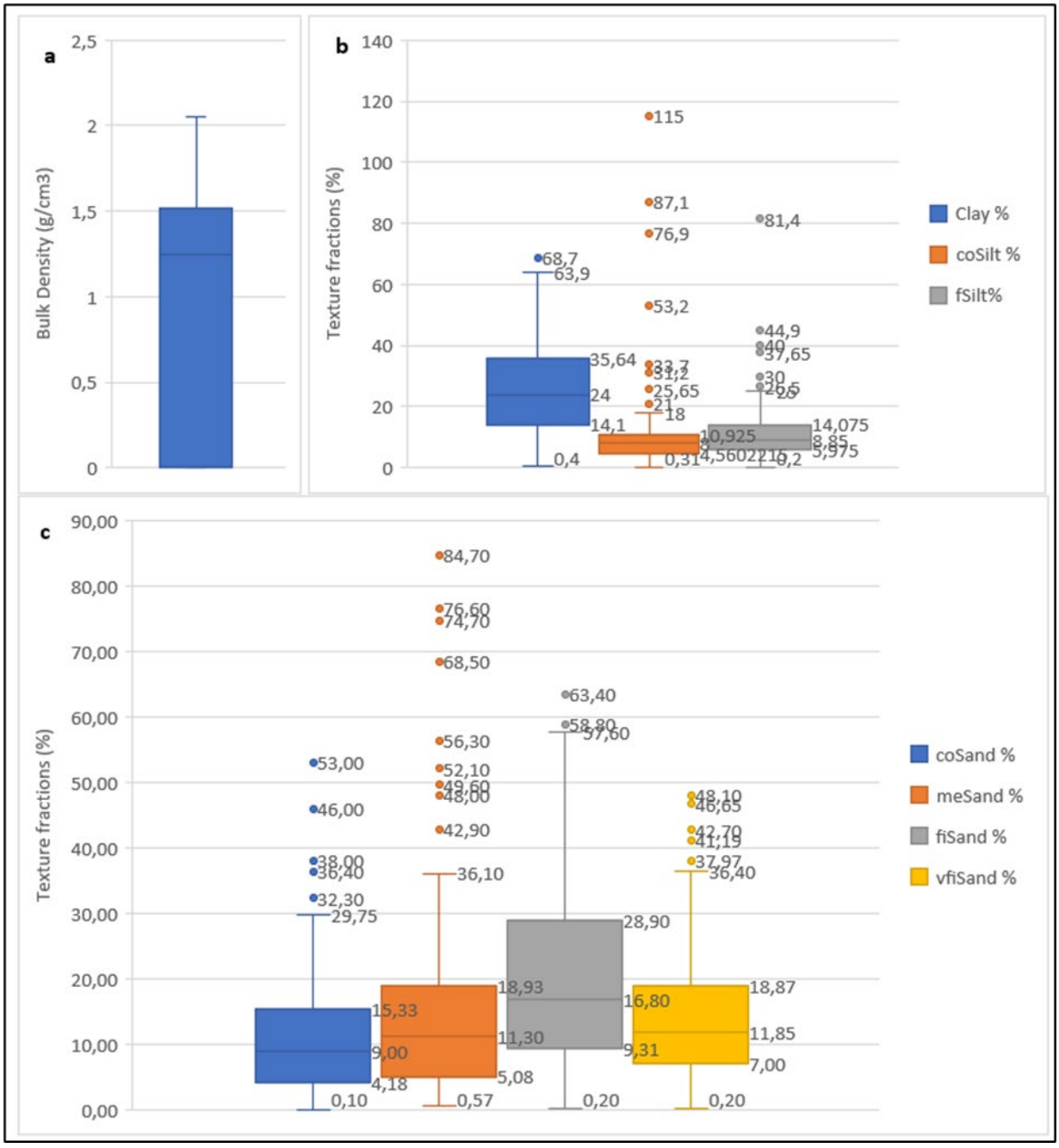

Figure 2.4: Boxplots for (a) bulk density, and (b, c) texture fractions percentages.

Boxplots for soil chemical and hydrological properties were also created and analysed, and if a soil point contained a soil property detected as an outlier, further analysis was carried out by following the four mandatory steps for quality control (Figure 2.1). Values detected as outliers for chemical and hydrological properties were not removed from the database as differed only one unit (or less) from the surrounding values. Likewise, geology and terrain values identified as outliers were identical to other values that were not identified as outliers. However, it could be that samples taken close together have

different values due to differences in soil forming factors. Therefore, detected outliers were not considered as true outliers as there was not enough substantiation for this output.

Post quality control, the total number of all the recorded soil profiles and soil horizons stayed the same. Only 98 soil attribute values were omitted from the soil database (Table 2.5).

Table 2.5: Summary of the number of different soil attributes values recorded before and after quality control

Soil attribute	Before quality control	After quality control	Removed attributes
Geographical coordinates	1166	1159	7
Clay%	623	611	12
coSand%, meSand%, fiSand%	206	194	12
vfiSand%, coSilt%	160	148	12
fiSilt%	162	150	12
Cation Exchange Capacity (CEC)	421	414	7

2.1.5 Conclusions

The aim of this study was to determine the optimal structure of the proposed soil point database, by examining the structure of an international (WoSIS) and a national (ARC-SCW) soil database. This aim was achieved through SWOT analysis of the structures of these soil databases, to evaluate the strengths, weaknesses, opportunities and threats thereof. The results revealed that a good soil database is dependent on the ease of data importation and exportation, soil data presentation and comprehensibility. The data quality of a database depends on the provision of information about the database, including definitions of the terminology, units and methods of measurements used in the database. Data usage is determined by how easy it is to read, edit, analyse, query, import data, navigate and perform quality control on the data. It is also dependent on the ease of understanding the software used to store and record data and the language, codes, and classification used in the database. A database should be easy to access from the source.

The proposed soil database drew from the findings of the SWOT analyses to optimise the legibility, data quality, data usage, accessibility, and standardisation of the database. It was developed as an Excel spreadsheet to promotes ease of usage, flexibility and interoperability. An attribute description created provides detailed information about the data recorded in the database, including definitions of terminology and symbols used, units and methods of measurements and the data source references.

The database was populated with data from various sources, which underwent an established two-stage quality control to assess outliers. The result is a soil profile database with a total of 567 soil profiles and 1 518 soil horizons which passed quality control. The intention is that the established quality control and standardisation measures will allow for data to be continuously added to this database such that it can become a freely available national asset.

CHAPTER 3: SABIE-SAND CATCHMENT

Chapter 3 presents the digital soil mapping and hydrological modelling of the Sabie-Sand catchment. This work formed the bulk of Eddy Smit's PhD thesis. The digital soil mapping entailed dealing with highly clustered data to produce an accurate soil map and has been published as a peer-reviewed paper in *Geoderma* (Section 3.1; Smit et al., 2023a). The digital soil mapping adds value to the hydrological modelling results to better mimic the hydrological processes at play within the catchment. These findings have been accepted as a peer-reviewed article in the journal *Vadose Zone* (Section 3.2; Smit et al., 2023b). The value of the HYDROSOIL map is further showcased by using the map to calibrate the hydrological model (Section 3.3). This paper has been submitted to the Hydrological Processes.

3.1 DOWNSCALING LEGACY SOIL INFORMATION FOR HYDROLOGICAL SOIL MAPPING USING MULTINOMIAL LOGISTIC REGRESSION

3.1.1 Abstract

In South Africa, there is a growing demand for large-scale, detailed hydrological soil maps for modelling and management purposes. However, legacy soil information often impedes the accurate creation of such maps by not being representative of the environmental complexity of large-scale catchments and containing imbalanced soil class distributions. The result is often the loss of minority soil classes, such as wetland and riparian soils, which are often of great hydrological importance. In this study, we propose a new downscaling approach to handle soil data within a large, low resolution legacy soil dataset to create an accurate hydrological soil map of the macro-scale (5 790 km^2) Sabie-Sand catchment using multinomial logistic regression (MNLR). The spatially localised legacy data was downscaled using k-means clustering and added to the broader legacy dataset. Five levels of legacy soil data were analysed in their representation of environmental covariates using QQ-plots and a Welsh's t-test and their mapping accuracy using confusion matrix and Kappa coefficient statistics. However, MNLR also requires balanced soil classes. The best performing legacy soil dataset was also compared to using all available soil information after both datasets had their soil class distributions fully balanced using Synthetic Minority Oversampling Technique (SMOTE). The 500ha/observation-SMOTE dataset resulted in the most accurate hydrological soil map with a validation point accuracy of 73% and a Kappa coefficient of 0.60, substantially outperforming the other downscaled soil maps as well as the SMOTE balanced dataset using all available soil information. This was due to the decreased variation between observations and catchment means, where the 500ha/observation dataset yielded the least variation between soil observation and catchment datasets as well as reducing the class imbalance within the legacy soil data. Downscaling spatially localised legacy soil data for environmental representation is an effective tool to improve digital soil mapping accuracy using MNLR.

3.1.2 Introduction

Digital soil mapping involves mathematical models for predicting soil properties using environmental covariates as predictors. The modelling procedure can be implemented using the framework of digital soil mapping (McBratney et al., 2003) to relate the environmental covariates with the target soil variable or class. Advances in digital soil mapping coincided with the need for detailed and accurate spatial soil information, which has widely been realised to provide appropriate solutions for conserving and managing agricultural and environmental resources. Soil information can improve modelling, policy making and scenario analysis at different spatial extents from catena to catchment and from national to global scales (Häring et al., 2012; Arrouays et al., 2014; Lamichhane et al., 2021).

Digital soil mapping has therefore developed into a cost-effective tool for creating large-scale, detailed soil maps. This is due to the expansion of highly-detailed remotely sensed covariate data, the drastic increase in desktop computing power and its availability to users as well as the availability of large swaths of legacy soil data (Häring et al., 2012). In South Africa, digital soil mapping has also been used for a wide range of agricultural and environmental applications including precision agriculture, land degradation studies, land capability studies, determining irrigation potential, as well as in Environmental Impact Assessments (EIAs) and town planning projects (Van Zijl, 2019).

Digital soil mapping has been increasingly used in the field of hydropedology, where hydrological soil maps created for modelling purposes have been shown to improve hydrological modelling accuracy (Van Tol et al., 2015; Van Tol et al., 2020; Harrison et al., 2022; Smit & Van Tol, 2022). The use of more detailed hydrological soil information within the ACRU (Agriculture Catchments Research Unit) model led to an increase in model efficiency of between 9% and 52% (Van Tol et al., 2015). Hydropedological soil information also improved modelling accuracy at three different catchment sizes compared to readily available soil information (Van Tol et al., 2020). The value of hydrological soil information may extend beyond the ability to accurately model long-term streamflow predictions. The argument is that hydrological soil information may serve as an effective ‘soft data’ tool, to better represent internal hydrological processes within a catchment (Smit & Van Tol, 2022).

In this light, the Water Research Commission has seen the potential of large-scale hydrological soil maps for the purposes of hydrological modelling and water resource management. It has authorised a project which builds towards a hydrological soil map of the country. Van Tol and Van Zijl (2022) provide rough steps to create a national hydrological soil map for South Africa. However, the creation of such large-scale hydrological soil maps remains reliant on using and interpreting large amounts of legacy soil data, from various soil surveys, which often use different classification systems at different spatial resolutions.

Digital soil mapping provides unique challenges in balancing legacy soil datasets because soils are never evenly distributed throughout a landscape, being a product of complex soil forming factors. Although machine learning models, such as Multinomial Logistic Regression (MNLR), are generally more accurate than simple models, the accuracy of these models are highly dependent upon the number of soil classes and the frequency of their distribution. The MNLR algorithm (Venables & Ripley, 2002) has been widely used for digital soil mapping purposes, specifically for large-scale mapping endeavours (Campling et al., 2002; Bailey et al., 2003; Hengl et al., 2007; Kempen et al., 2009; Debella-Gilo & Etzelmüller, 2009; Van Zijl, 2019). However, the performance of such algorithms is often poor when learning from imbalanced data, which is well documented in the field of categorical data modelling (López et al., 2013; Haixiang et al., 2017; Li et al., 2022).

Legacy soil datasets are often not representative of the environmental complexity of large-scale catchments, due to a limited number of soil observations and imbalanced soil class distributions. Most legacy soil datasets contain spatially localised sampling locations which were purposely selected according to the aim and purpose of the soil survey (Ma et al., 2019). For example, most agricultural soil surveys would refrain from sampling within specific low potential or prohibited positions in the landscape. When creating hydrological soil maps from legacy soil data, the minority soil classes are often of great importance (e.g. under-sampled wetlands and riparian soils). Predictive models based on these spatially-imbalanced legacy soil datasets, would fail to map these ecologically important soils.

When addressing the challenge of imbalanced datasets, three main approaches have been recognised. Firstly, there is the data-level approach where different resampling methods are used to create a balanced dataset. These methods include, under-sampling, where the majority soil class is reduced to

balance observations, over-sampling, which creates replications of the minority soil class, and synthetic data generation, where new artificial data of the minority observations are created to balance class distributions within datasets. The second approach is at the algorithm level, where algorithms are selected which are capable of handling imbalanced datasets, these include algorithms which incorporate cost sensitive learning and active learning (Chawla et al., 2002). The third approach is to apply a combination of both data-level and algorithm-level approaches to produce the most accurate results (He & Garcia, 2008; García & Herrera, 2009).

Within this paper we will be focussing on the data-level approach to create a balanced soil dataset. Fairly limited data balancing research has been conducted specific to the field of soil science (Heung et al., 2016; Sharififar et al., 2019a; Taghizadeh-Mehrjardi et al., 2019). Recently, Sharififar et al. (2019b) analysed the use of random oversampling (ROS) and random undersampling (RUS) methods to balance soil datasets using various machine learning models, where majority soil observations were down-sampled and minority soil observations up-sampled, preserving proportionality, to deal with the issue of imbalanced soil data with 452 profiles in an area of about 12 000 ha. They concluded that balancing soil datasets using a combined approach of both RUS and ROS significantly improved MNLR accuracy and decreased mapping uncertainty. However, these simplistic random resampling techniques potentially increase the likelihood of overfitting by discarding potentially useful observations, especially when large differences occur between the number of majority and minority soil classes (Zhu et al., 2017).

In an assessment of eight resampling strategies on five of the most-used machine learning algorithms on a national scale for Iran (1 648 195 km^2) using 7 664 soil observations, the researchers concluded the that highest increase in prediction accuracy was achieved using the Synthetic Minority Oversampling Technique (SMOTE) (Taghizadeh-Mehrjardi et al., 2019). As the name suggests, SMOTE generates synthetic examples of the minority soil classes. SMOTE uses the existing minority samples and interpolates between samples and their covariate attributes to generate new samples of the specific class (Chawla et al., 2002). Oversampling approaches of the minority soil class outperformed under-sampling techniques (Taghizadeh-Mehrjardi et al., 2019). This is because useful information, which was obtained by costly, time consuming and labour-intensive soil sampling, in the majority soil classes are ignored, leading to the degradation of classifier performance. ROS was also shown to provide poor performance results when legacy datasets contained large differences between majority and minority soil classes (Peri et al., 2018). ROS creates exact copies of the minority soil classes, which leads to a small decision region for the minority soil classes compared to the majority soil class and therefore the likelihood of overfitting increases substantially (Chawla et al., 2002; Zarinabad et al., 2017). More research is needed on the effects of different balancing techniques on different classifiers within the field of digital soil mapping (Sharififar et al., 2019b).

The main aim of this study was to create an accurate hydrological soil map of a macro-scale catchment in South Africa using MNLR and legacy soil information. This aim was achieved by providing an approach to handle large, spatially-localised legacy soil datasets for digital soil mapping purposes using MNLR. The first objective was to address the high spatial localisation and imbalances within the legacy soil data available. Localised data was first downscaled to improve the representativeness of environmental covariates in the broader legacy soil dataset prior to balancing soil classes. The second objective was to analyse the value of the downscaled legacy soil data. Mapping results were compared with and without balancing soil classes, as well as between the balanced legacy soil dataset and a dataset containing all available soil observations.

3.1.3 Materials and methods

The Sabie-Sand catchment

The 5 790 km^2 Sabie-Sand catchment is in the Mpumalanga province of South Africa (Figure 3.1) and forms part of the larger transboundary Incomati river basin. The catchment stretches from the Drakensberg escarpment in the west at an altitude of 2 200 meters above sea level (m.a.s.l.) and gradually flattens towards the east with an altitude of 150 m.a.s.l. before the Sabie River flows into Mozambique.

With a semi-arid warm and hot climate in the east of the catchment and a temperate warm climate in the west, a strong rainfall gradient exists, ranging from 1 600 mm in the west to 450 mm in the east. Rainfall occurs mainly in the austral summer (November through to March) and normally results from convective thunderstorms, although periodic high-intensity rainfall events do occur from cyclones that form over the Indian Ocean and track inland, where the orographic effect of the Drakensberg escarpment creates severe localised flooding (Kruger et al., 2002).

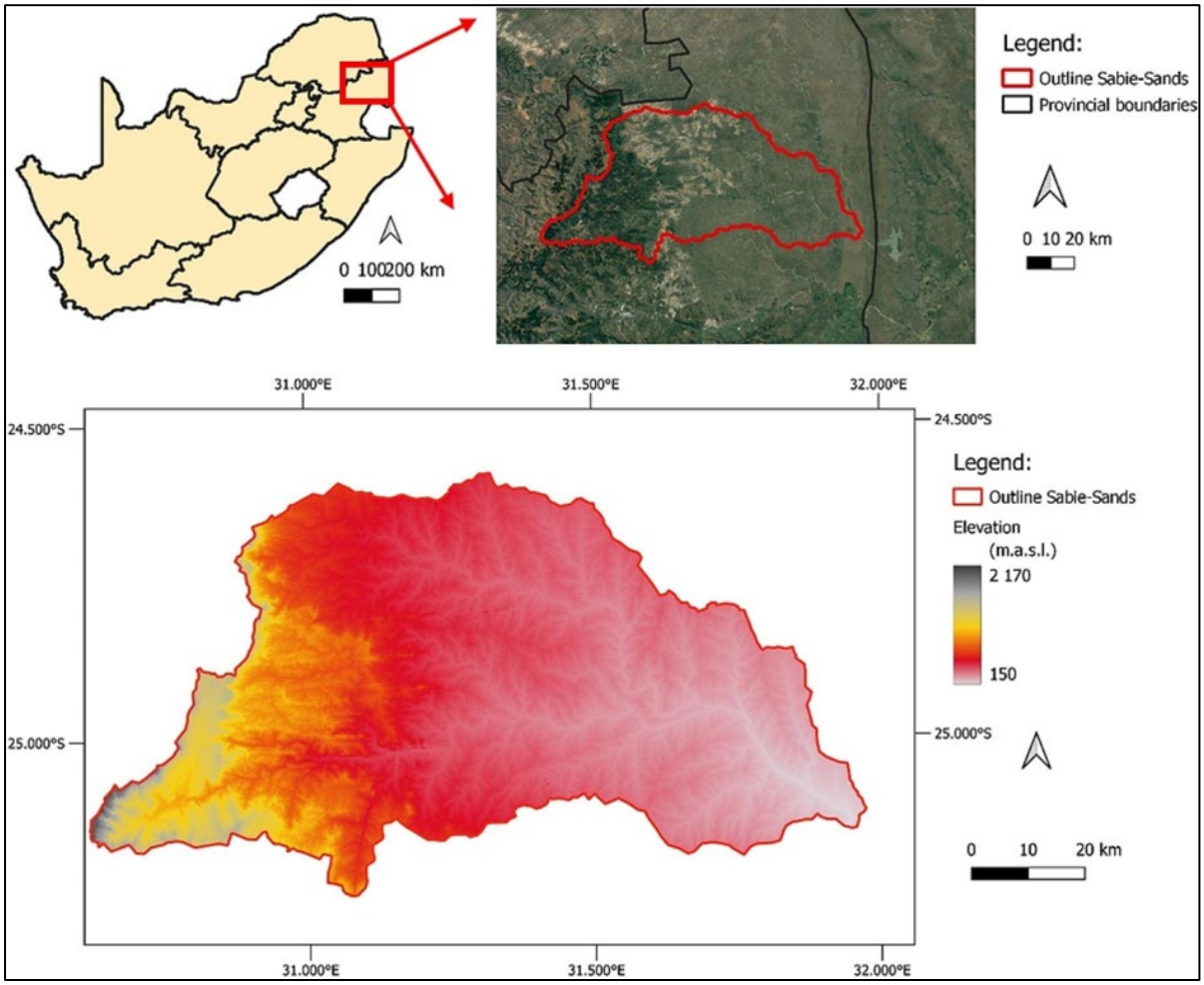

Figure 3.1: The location of the Sabie-Sand catchment.

The main bioregions of the catchment consist of savanna at lower altitudes and montane grasslands and montane forests in the mountainous regions, which have been heavily altered by commercial forestry plantations (Mucina & Rutherford, 2006). The area comprises various bedrock lithologies including, quartzites, granites, basalts, conglomerates, andesites, gneiss and shales (Council for Geoscience, 2007).

Soil data

The legacy soil data of the Sabie-Sand catchment in total amounts to 12 875 soil observations from various soil surveys. Among these are 380 observations as part of the national soil survey by the ARC (Land Type Survey Staff, 1976-2006), 108 soil observations by an in-field hydropedological survey and 118 observations from research and consultancy projects by various private and state-owned enterprises. However, the majority of soil observations within the catchment originate from a single forestry soil survey done by the South African Forestry Company Limited (SAFCOL) where 12 269 soil observations were made on 38 000 ha. The total soil observation density of the SAFCOL legacy data is 3.4 ha per soil observation compared to the 960 ha per soil observation for the remaining legacy soil information within the catchment. The spatial representation (Figure 3.2) of legacy soil observations within the catchment illustrates a clear bias in the mountainous (western) section of the catchment, visually illustrating the spatially localised nature of the SAFCOL soil dataset.

Due to the large size of the catchment area and large number of soil observations, a wide variety of soils occur within the catchment, driven by the differences in soil forming factors (parent material, climate, topography, organisms and time) between the afromontane regions and the semi-arid lowveld regions of the catchment. This spatial imbalance also results in an imbalanced number of the different soil types and therefore mapping classes within the catchment. Using the hydropedological groupings of South African soils (Van Tol & Le Roux, 2019) the soils of the Sabie-Sand were divided into six hydrological soil types (Table 3.1).

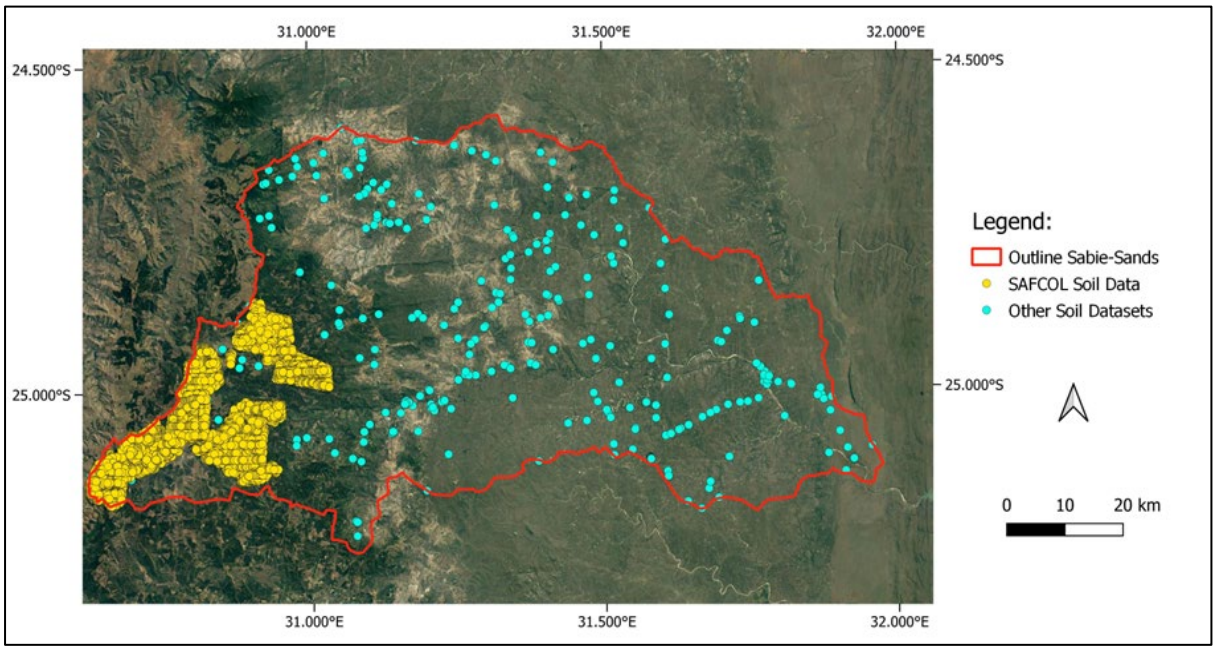

Figure 3.2: The spatial distribution of legacy soil datasets within the Sabie-Sand catchment.

Table 3.1: The defining characteristics of the hydrological mapping units of the Sabie-Sand catchment.

Hydrological mapping unit	Soil form	WRB Reference Groups	Number of obs.	Defining hydrological characteristic
Recharge deep	Hutton, Longtom, Kranskop	Acrisols, Nitisols, Fluvisols	11582	Deep soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying bedrock is the dominant flow direction.
Recharge shallow	Glenrosa, Nomanci	Leptosols	401	Shallow soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying fractured bedrock is the dominant flow direction.
Responsive saturated	Katspruit, Champagne	Gleysols	68	Soils with morphological evidence of long periods of saturation promoting the generation of overland flow due to saturation excess.
Responsive shallow	Mispah, Graskop	Leptosols	649	Shallow soils overlying relatively impermeable bedrock. Limited storage capacity results in the generation of overland flow after rainfall events.
A/B interflow	Estcourt, Sterkspruit	Solonetz	85	Duplex soils where the textural discontinuity facilitates build-up of water in the topsoil, with discharge in a predominantly lateral direction.
Soil/bedrock interflow	Fernwood, Cartref	Arenosols	173	Soils overlying relatively impermeable bedrock. Hydromorphic properties signify temporal build of water on the soil/bedrock interface and slow discharge in a predominantly lateral direction.

WRB = World Reference Base for Soil Resources

Recharge deep soils comprise approximately 90% of the total soil samples, recharge shallow soils 3%, responsive saturated soils comprise approximately 0.5%, responsive shallow soils 5%, A/B interflow soils comprise approximately 0.6% and soil/bedrock interflow soils 1%.

The mountainous soils of the Sabie-Sand are characterised by well-weathered soils being either deep apedal soils (Acrisols and Nitisols) or shallow apedal soils (Leptosols) on midslope and hillcrest positions depending on the parent material. Alluvial (Fluvisols) and saturated high clay (Gleysols) soils are also prominent in footslope and valley-bottom terrain positions depending on upslope environmental covariates.

The lowland soils are comparatively far shallower than the mountainous soils, gravellier, being less well-weathered and primarily controlled by differences in parent material. These soils show a more distinct toposequence with apedal soils (Acrisols and Nitisols) on hillcrest, albic soils (Arenosols) on midslope and footslope terrain positions. A small number of duplex soils (Solonetz) are also present on footslope and valley-bottom terrain positions. Apedal soils (Nitisols and Fluvisols) are also present on valley bottom terrain positions as floodplains form on the major river networks (IUSS Working Group WRB, 2015).

The hydropedological grouping of different soil types into six conceptual classes decreases the number of soil mapping classes but also creates larger variation between the number of soil observations per mapping class and thus further adds to the imbalance within the legacy soil dataset. This is especially

true when most of the soil observations are spatially localised as in the Sabie-Sand catchment, where the majority soil class (recharge deep, 90%) massively overshadows the smallest minority soil class (responsive saturated, 0.5%). Creating a national hydrological soil map would be highly reliant on these spatially-localised legacy soil datasets, especially in mountainous regions where the majority of soil observations consist of commercial forestry surveys. These detailed surveys would add a substantial amount of soil observations to the available soil mapping resources in South Africa. Therefore, developing a protocol for handling these localised datasets to improve digital soil mapping accuracy is imperative moving forward.

Multinomial logistic regression

MNLR forms a part of the broader family of generalised linear models and is applied when the target variable contains more than two categorical variables. This helps predict the probability of the occurrence of each unique soil mapping unit. If a given variable Y_i represents the observed soil mapping unit at a given observation location, with i=1,…, n and n is the number of soil mapping units within the study area (Kempen et al., 2009). In case n equals 2 and Y has outcomes Y_1 and Y_2. Both the counts of Y_1 and Y_2 therefore follow a binomial distribution. The probability of occurrence of Y_1 is $_1$ and that of Y_2 is $_2$. Logistic regression relates probability $_1$ to a set of predictors, in our case environmental covariates, using the logit link function:

$$logit(\pi_1) = ln\ \frac{\pi_1}{\pi_2} = ln\left(\frac{\pi_1}{1-\pi_1}\right) = x'\beta \tag{3.1}$$

where x is a vector of environmental covariates, and β is a vector of model coefficients that are typically estimated by maximum likelihood. Therefore, Equation 3.1 can also be rewritten as:

$$\frac{\pi_1}{1-\pi_2} = exp(x'\beta) = exp(n) \tag{3.2}$$

The quotient in Equation 3.2 is referred to as the *odds*. Equation 3.2 can then be reinterpreted as follows:

$$\pi_1 = \frac{exp(n)}{1+exp(n)} \tag{3.3}$$

The binomial logistic regression model is then generalised to the multinomial case. Where, there are n soil mapping units and also n variables Y_1,…, Y_n with corresponding probabilities of occurrence π_1,...π_n. Analogous to binomial logistic regression the odds π_1/π_n,…, π_{n-1}/π_n are modelled by means of exp(n_1),…, exp(n_{1-1}). From $\sum_{i=1}^{n}$ π_1 it then follows that:

$$\pi_i = \frac{exp(n_i)}{exp(n_1)+exp(n_1)+\ldots exp(n_n)} \tag{3.4}$$

where n_n= 0. This model ensures that all probabilities are in the interval [0,1] and that the probabilities sum to 1.

Covariate data and statistical analysis

A comprehensive environmental covariate dataset to describe the soil forming factors within the *scorpan* model (McBratney et al., 2003) was required for the MNLR algorithm to predictively map the different hydrological soil types of the Sabie-Sand catchment. These covariates were all resampled to a resolution of 30 m x 30 m, regardless of their original resolution.

Elevation was obtained from a 30 m x 30 m Shuttle Radar Topography Mission (SRTM) Digital Elevation Model (DEM; USGS, 2022). The covariates used to train the models included elevation, a 1: 250 000 geology map (Council for Geoscience, 2007), a broad land type map (Land Type Survey Staff, 1972-2006), planform curvature, profile curvature, vertical distance to channel network, topographical wetness index, climate covariates such as mean annual minimum temperature, mean annual maximum temperature, and mean annual precipitation (Schulze, 2007a). Slope, relative slope position, multiresolution valley-bottom flatness, multiresolution ridge-top flatness, LS-factor, longitudinal profile, flow accumulation index, cross-sectional curvature, convergence index and channel network distance were also incorporated. All topographic covariates were developed using the System for Automated Geoscientific Analysis (Conrad et al., 2015), from the DEM raster.

Additional spectral covariates were also developed for the Sabie-Sand catchment from Sentinel 2A satellite imagery to further differentiate between different soil types. These spectral covariates included brightness index, colouration index, redness index, saturation index and Normalised Difference Vegetation Index (NDVI) values for both the wet and dry season (Table 3.2; Bannari et al., 1995; Ray et al., 2004; Flynn et al., 2019a).

Table 3.2: Spectral bands, spectral covariates and their development (from Flynn et al., 2019a).

Bands	**Band origin (µm)**	**Symbol**
Blue	0.490	B
Green	0.560	G
Red	0.665	R
Near infrared (NIR)	0.842	NIR
Covariates	**Equation**	**Property**
Brightness index	$(R^2 + G^2 + B^2)/3^{0.5}$	Reflectance
Colouration index	$(R - G)/(R + G)$	Colour
Redness index	$R^2/(B * G^3)$	Hematite
Saturation index	$(R - B)/(R + B)$	Spectral slope
NDVI	$(NIR - R)/(NIR + R)$	Chlorophyll

NDVI = Normalised Difference Vegetation Index

Calibration and validation datasets

The SAFCOL data was used to create five levels of legacy soil information of which three were created using a downscaling approach. The downscaled legacy soil datasets were created using the base R software (R Core Team, 2022) in conjunction with the *prospectr* package for k-means clustering, trained on a comprehensive environmental covariate dataset, and the *nnet* package (Venables & Ripley, 2002) for running the MNLR algorithm.

The k-means clustering is a simple unsupervised non-linear clustering algorithm, where the algorithm seeks to partition the observations into a pre-specified number of k clusters (Hartigan & Wong, 1979). These clusters try to maximise the difference between clusters whilst also minimising the difference within clusters using the Euclidean distance between soil observation covariate data. K-means clustering is therefore an effective tool to downscale legacy soil data to a specific user-defined number of soil observations (number of clusters), while also retaining the most representative soil observations within the legacy soil dataset. When downscaling the legacy soil data, we therefore defined the number

of clusters (k) within R to be equal to the number of soil observations defined by the predetermined observation density (Stevens & Ramirez-Lopez, 2022).

The least detailed level of soil observation was using only the legacy soil information (Legacy) which excluded all of the SAFCOL soil data. The most detailed level of soil information included using all 12 875 available soil observations within the study area (All Observations), regardless of the balance or representativeness of soil observations. Observation density was then used to create the downscaled SAFCOL soil datasets at a further three levels. The third level of soil information included the legacy soil information as well as the SAFCOL data downscaled to the same observation density (960ha/observation). Therefore, our spatially localised dataset was downscaled to the same observation density as the soil observation density in the rest of the catchment. The fourth and fifth level of soil information downscaled the SAFCOL data to 500ha/observation and 100ha/observation in addition with the remaining legacy soil data. These observation densities were selected to establish if observation density affected the hydrological soil mapping accuracy and to potentially determine how many soil observations are required to create an accurate hydrological soil map of a macro-scale catchment.

All five levels of soil information were split into a calibration dataset (75%) and validation dataset (25%). However, a completely independent validation dataset (152 soil observations) was initially created consisting only of legacy soil information which excluded the SAFCOL soil data, which was removed from the different calibration datasets prior to creation. This prevented the differences in the downscaling procedures from affecting the number of observations within the validation dataset or its internal hydrological soil class distribution.

The MNLR algorithm in conjunction with the above-mentioned environmental covariates were then applied. The resulting hydrological soil maps were analysed using a confusion matrix as well as a Kappa coefficient statistic for both the calibration and validation datasets. How the different downscaled legacy soil datasets affected the representation of the larger covariate soil dataset was tested using a Welsh's t-test analysing the mean annual precipitation, slope, topographic wetness index and NDVI (dry season), with a p-value of 0.05 being used as a threshold for significant difference between the datasets. These four select covariates each represent different soil forming factors, giving an indication of the representativeness of legacy soil datasets. The Welsh's t-test compares the two means between datasets, the basic null hypothesis is that the means are equal. The QQ-plot was constructed by determining the (k/n + 1)-th quantiles of the large dataset, (where k = 1, ...n and n is the number of observations of the legacy soil dataset) and plotting those quantiles against the values of the observation's covariates of the legacy soil datasets, sorted from small to large. The closer the values of the QQ-plots to the identity line (x=y), the more representative the legacy soil dataset is of environmental covariates.

However, the MNLR algorithm requires a balanced soil mapping class dataset rather than one balanced by the respective environmental covariates. The most representative legacy soil dataset therefore still needed to be balanced by hydrological soil type and assessed.

As under-sampling has shown to decrease classifier performance by losing potentially useful information of the majority soil class and because the majority soil class has already been downscaled, only an oversampling approach was used to balance soil classes and compare soil mapping accuracy. However, because ROS causes overfitting when dealing with large differences between majority and minority soil classes, which still exists after downscaling, this approach would be nonsensical (Taghizadeh-Mehrjardi et al., 2019). Based on the available literature, the SMOTE technique to generate synthetic data of the minority soil classes was selected to balance hydropedological soil

classes. This technique was applied for both the most representative and best performing legacy soil dataset and all available soil observations using the *smotefamily* package (He et al., 2008). This allows us to compare if the downscaling of spatially localised soil observations for covariate representatives adds value to the digital soil mapping process, or if all legacy soil observations should simply be balanced using SMOTE for the best results.

The two SMOTE-balanced legacy soil datasets were then also applied within the MNLR algorithm to create a hydrological soil map of the Sabie-Sand catchment. The resulting hydrological soil maps were analysed using the same validation dataset and statistical measurements (confusion matrix and Kappa coefficient).

3.1.4 Results and discussion

Downscaling legacy soil data

QQ-plots illustrate select environmental covariates of the different legacy soil datasets and the corresponding covariates of the entire catchment (Figure 3.3). The QQ-plot of the mean annual precipitation (MAP) covariate illustrates the observation bias of the All Observations and 100ha/observation datasets for values ranging from 1 200-1 600 mm per annum, resulting from the localised SAFCOL data within the higher rainfall regions of the catchment.

The 500ha/observation, 960ha/observation and Legacy datasets are all more representative, with the 500ha/observation and 960ha/observation datasets yielding the most representative results for MAP values. The same trend can be observed when analysing the slope, where the 500ha/observation and 960ha/observation datasets yielded the most accurate representation of slope values across the catchment. Due to the bias of the legacy datasets towards the mountainous regions of the catchment where large slope values are present, the All Observations and 100ha/observation datasets are biased to higher slope values, whereas the Legacy soil dataset and 960ha/observation are slightly biased to low slope values common in the east of the catchment.

For topographical wetness index values, all datasets excluding the All Observations dataset yielded accurate representations of the specific environmental covariate for the landscape. The All Observations dataset remains biased to low topographic wetness index values, which is due to the fact that these observations were focussed on upslope terrain positions because proportionately limited SAFCOL observations were made in valley-bottom positions.

The 500ha/observation yielded the most accurate representation of catchment covariate values within the NDVI QQ-plot, which once again followed the same trend where both the All Observations and 100ha/observation datasets were biased to high NDVI values, indicative of the evergreen forestry activities in the mountainous areas in the catchment. The Legacy and 960ha/observation were slightly biased toward low NDVI values, indicative of the lack of vegetation in the savanna dry season.

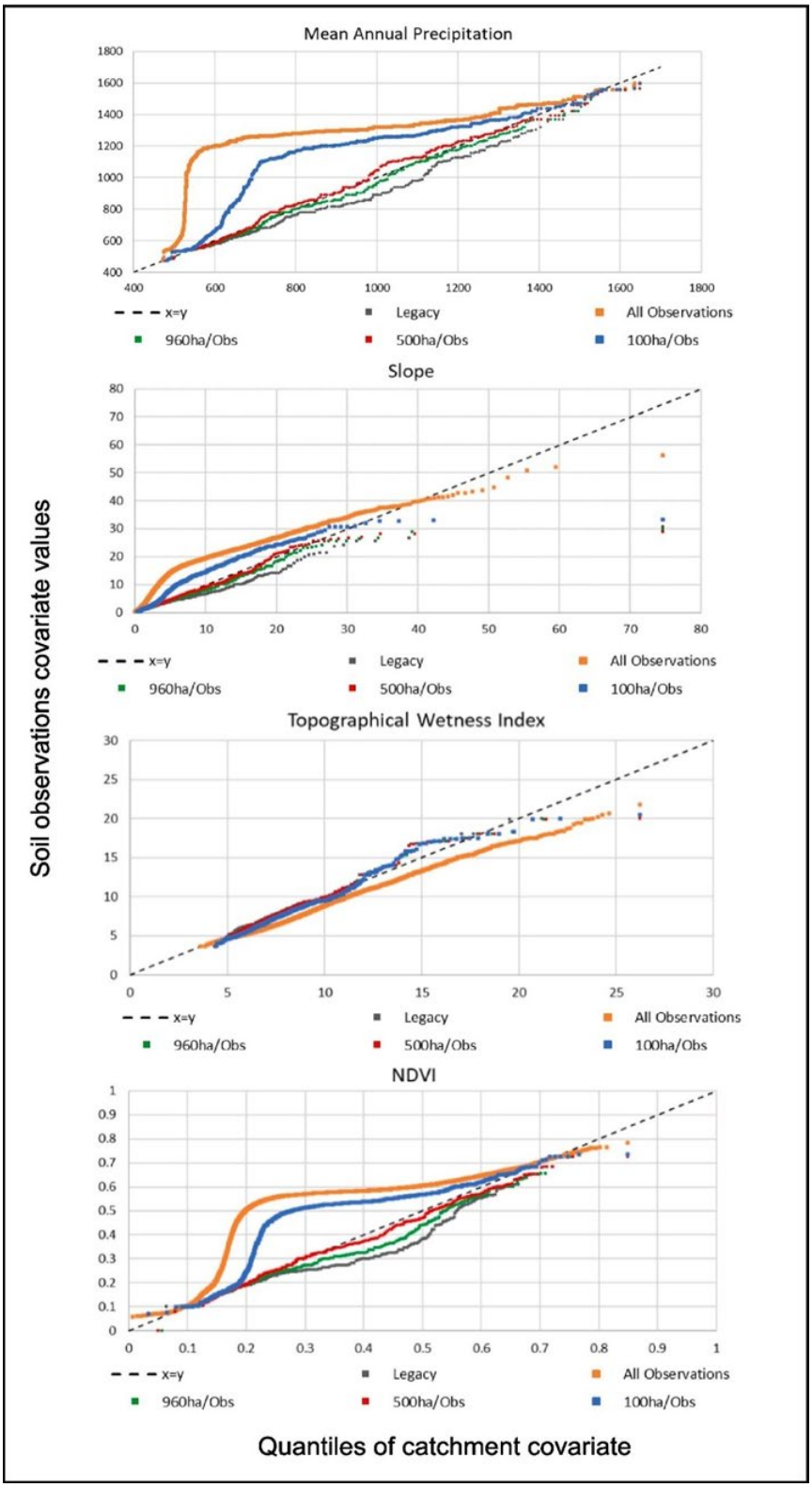

Figure 3.3: The QQ-plots of the different legacy soil datasets and select environmental covariates.

Table 3.3 illustrates the results of a Welsh's t-test between the MAP, slope, Topographic Wetness Index (TWI) and NDVI values (dry) of the different soil datasets and the corresponding catchment environmental covariate.

Table 3.3: The Welsh's t-test of select environmental covariates.

Covariate	Legacy dataset	Welsh's t-test			
		p-value	t-value	Mean soil	Mean covariate
MAP	Legacy	<0.002	-4.09	758.52	800.43
	All Observations	<0.002	315.42	1233.34	
	960ha/Observation	0.258	-1.11	788.36	
	500ha/Observation	0.369	0.90	810.25	
	100ha/Observation	<0.002	14.82	951.59	
Slope	Legacy	<0.002	-8.06	4.15	5.51
	All Observations	<0.002	101.54	11.84	
	960ha/Observation	<0.002	-3.94	4.75	
	500ha/Observation	0.034	-2.07	5.09	
	100ha/Observation	<0.002	8.34	7.35	
TWI	Legacy	0.008	2.63	8.07	7.82
	All Observations	<0.002	-61.48	6.75	
	960ha/Observation	0.094	1.68	7.98	
	500ha/Observation	0.141	1.48	7.96	
	100ha/Observation	<0.002	-3.25	7.57	
NDVI (dry)	Legacy	<0.002	-5.05	0.24	0.26
	All Observations	<0.002	176.75	0.48	
	960ha/Observation	0.003	-2.96	0.25	
	500ha/Observation	0.300	-1.03	0.26	
	100ha/Observation	<0.002	12.25	0.33	

MAP = Mean Annual Precipitation; TWI = Topographic Wetness Index; NDVI = Normalised Difference Vegetation Index

With all p-values below 0.05 the means of the Legacy, All Observations and the 100ha/observation soil datasets were significantly different from catchment environmental covariates. This is to be expected for catchments the size of the Sabie-Sand (5 790 km^2) where large ranges of environmental covariates exist, and legacy soil datasets are relatively small in comparison. However, the means of the 960ha/observation were not significantly different for mean annual precipitation and topographical wetness index (Table 3.3). The means of the 500ha/observation were not significantly different for all four catchment covariates.

The t-value, which measures the size of the difference of the observation data relative to the variation in our catchment data, where t-values closest to 0 indicate the lowest variation between the catchment covariates and soil observations covariates, improves as the soil data is downscaled due to the improved representation of catchment environmental covariates.

The QQ-plots and Welsh's t-test results illustrate the improved catchment covariate representation which can be achieved using a downscaling approach on localised legacy soil information. The

500ha/observation and 960ha/observation datasets provided a substantially improved representation of catchment environmental covariates compared to the Legacy, All Observations, and 100ha/observation datasets. In particular, the 500ha/observation dataset provided the most accurate representation of environmental covariates within the Sabie-Sand catchment.

Digital soil mapping

Table 3.4 illustrates the mapping accuracy of the five different legacy datasets in relation to their calibration and validation datasets. Focussing on the calibration datasets, the best performing was the Legacy soil dataset with a confusion matrix accuracy of 72% and Kappa coefficient of 0.52, whereas using All Observations resulted in the highest confusion matrix accuracy (88%) but the lowest Kappa coefficient (0.21). The three downscaled approaches provided modest calibration accuracy results with 960ha/observation and 500ha/observation yielding confusion matrix values and Kappa coefficient (Table 3.4). However, the 100ha/observation dataset yielded the east accurate calibration dataset within the different downscaling approaches. Although calibration accuracy should not be considered when assessing mapping accuracy, insight can be gained on how the models learned from the calibration legacy soil data.

Table 3.4: The statistical accuracy of the legacy datasets.

Legacy dataset	Dataset used	Point accuracy (%)	Kappa coefficient
Legacy	Calibration	72	0.52
	Validation	50	0.34
All Observations	Calibration	88	0.21
	Validation	46	0.28
960ha/observation	Calibration	50	0.48
	Validation	48	0.42
500ha/observation	Calibration	64	0.47
	Validation	62	0.46
100ha/observation	Calibration	63	0.27
	Validation	54	0.33

When analysing the validation results of the five levels of soil information, the 500ha/observation dataset yielded the most accurate hydrological soil map with a confusion matrix value of 62% and Kappa coefficient value of 0.46, compared to the two control datasets (Legacy and All Observations; Table 3.4). Both the 960ha/observation and 500ha/observation datasets represent a moderate strength of agreement with reality, outperforming both control datasets which represent a fair agreement with reality. The 100ha/observation only slightly outperformed the All Observations dataset also indicating a fair agreement with reality (Landis & Koch, 1977).

The 500ha/observation legacy soil dataset yielded the most accurate mapping results (Table 3.4) as well as being the most representative legacy soil dataset for the selected environmental covariates. This dataset was therefore selected for further balancing and comparison using the SMOTE technique to balance soil mapping classes.

The All Observations-SMOTE dataset yielded a validation point accuracy of 53% and a Kappa coefficient of 0.47 (Figure 3.4a), which is similar to results achieved by the 500ha/observation dataset prior to SMOTE balancing and represents a moderate strength of agreement with reality. Although the

All Observations-SMOTE hydrological soil map provided relatively accurate distributions of dominant hydrological soil types within the SAFCOL dataset such as recharge deep, recharge shallow, responsive shallow and responsive saturated, the ability to map the hydrological soils outside of these areas was poor. Especially the prediction of A/B interflow soils, where the MNLR algorithm vastly overestimated its presence within the catchment area to the detriment of recharge shallow and soil/bedrock interflow soils (Figure 3.4a) in the lowveld areas of the catchment.

This overprediction is most likely due to the specific nature of A/B interflow soils relative to the other hydrological soil types, where a limited range of soil forming factors result in these very specific soils compared to a far wider range resulting in soil/bedrock interflow and recharge shallow soils. Therefore, when SMOTE generated a substantial amount of synthetic data from the existing A/B interflow data, the resulting data was comparative to random oversampling with a small region of specific examples being created, which led to overfitting.

The 500ha/observation-SMOTE map (Figure 3.4b) yielded a validation point accuracy of 72% and a Kappa coefficient of 0.60, which was the most accurate hydrological soil map of the Sabie-Sand catchment and resulted in a substantial agreement with reality (Landis & Koch, 1977).

These results are comparable with the bulk of the other hydrological soil maps created in South Africa, such as with a point accuracy of 69% and Kappa coefficient value of 0.59 (Van Zijl et al., 2019); point accuracy of 69% and Kappa coefficient value of 0.59 (Van Zijl et al., 2012); point accuracy of 88% and Kappa coefficient value of 0.82 (Van Zijl et al., 2020), and a point accuracy of 74% and Kappa coefficient of 0.68 (Smit & Van Tol, 2022). The results are also comparable to other digital soil mapping projects globally, such as a point accuracy of 69% (MacMillan et al., 2010) and 76% (Zhu et al., 2008).

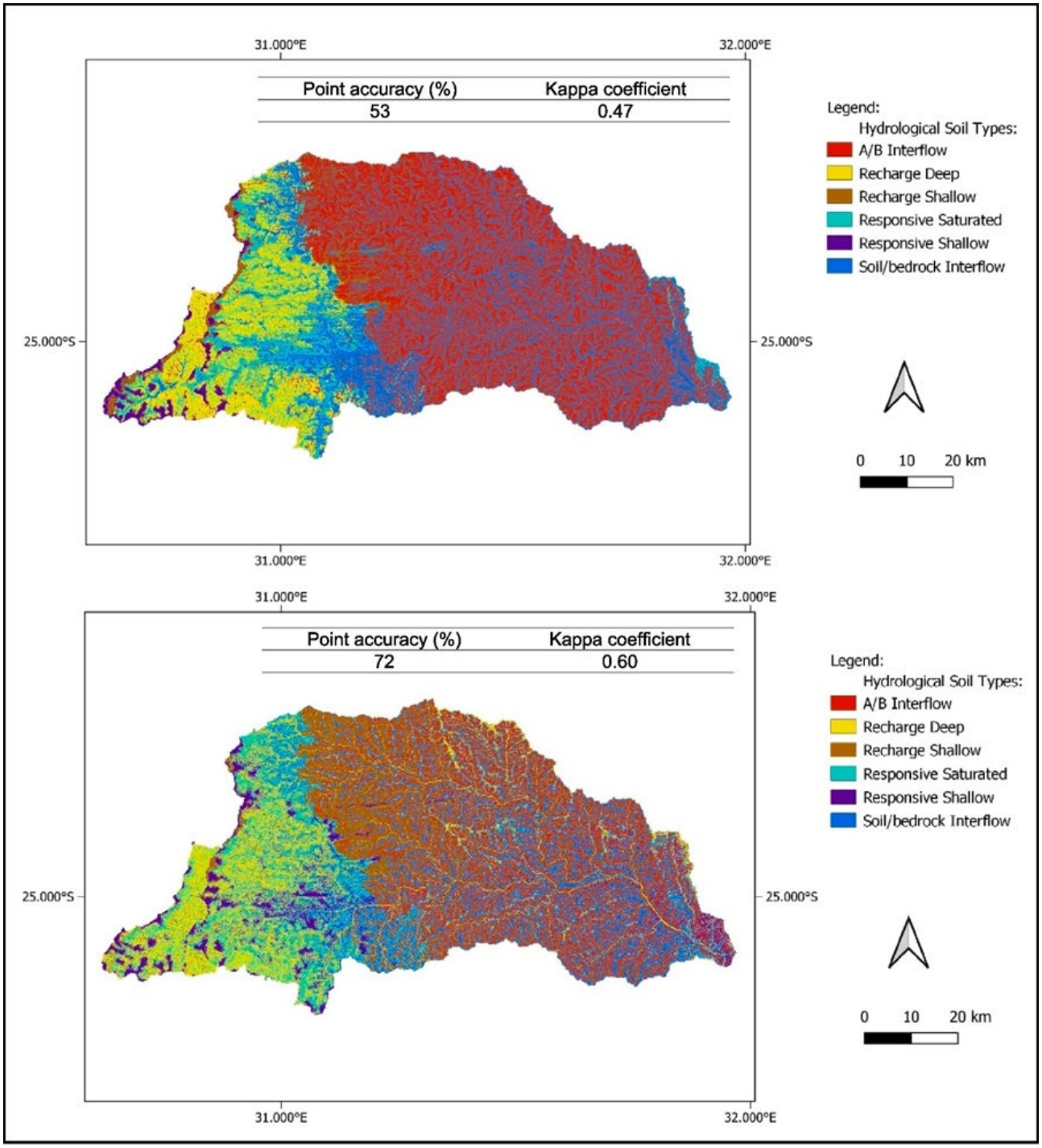

Figure 3.4: The hydrological soil types of (a) All Observations-SMOTE and (b) 500 ha/observation-SMOTE maps and their accompanying validation accuracy.

The confusion matrix for the All Observations-SMOTE hydrological soil map (Table 3.5) indicated that not all the soil classes were sufficiently mapped, with user's accuracy below 50% for recharge shallow, responsive shallow and soil/bedrock interflow classes. In general, the producer's mapping accuracy performed slightly better with only A/B interflow soils yielding an accuracy value below 50%.

Table 3.5: Confusion matrix of the All Observations-SMOTE hydrological soil map.

Mapping unit		User accuracy							
		A/B interflow	Recharge deep	Recharge shallow	Responsive saturated	Responsive shallow	Soil/bedrock interflow	Correct	%
Producer's accuracy	A/B interflow	23	14	6	1	4	10	23	40
	Recharge deep	2	50	4	1	1	3	50	82
	Recharge shallow	0	2	7	0	1	0	7	70
	Responsive saturated	0	1	2	5	0	0	5	62
	Responsive shallow	0	1	0	0	4	0	4	80
	Soil/bedrock interflow	0	5	0	0	0	5	5	50
	Correct	23	50	7	5	4	5	62	
	%	92	69	37	71	40	28		

The confusion matrix for the 500ha/observations-SMOTE hydrological soil map (Table 3.6) indicated that all the soil classes were sufficiently mapped, with all the user's and producer's accuracies above 50%. In general, mapping accuracy decreased with decreased validation observations as seen with the 57% user's accuracy of responsive saturated soils where only seven observations were present compared to the 84% user's accuracy of recharge deep soils where 73 observations were present.

Table 3.6: Confusion matrix of the 500ha/observations-SMOTE hydrological soil map.

Mapping unit		User accuracy							
		A/B interflow	Recharge deep	Recharge shallow	Responsive saturated	Responsive shallow	Soil/bedrock interflow	Correct	%
Producer's accuracy	A/B interflow	15	3	1	0	0	1	15	75
	Recharge deep	5	61	4	2	3	5	61	76
	Recharge shallow	3	2	14	1	1	2	14	61
	Responsive saturated	0	1	0	4	0	0	4	80
	Responsive shallow	2	1	0	0	6	0	6	67
	Soil/bedrock interflow	0	5	0	0	0	10	10	67
	Correct	25	73	19	7	10	18	72	
	%	15	61	14	4	6	10		

The 500ha/observation-SMOTE hydrological soil map (Figure 3.4b) contains a distinct toposequence in the mountainous and lowland areas within the catchment. In the mountainous regions, responsive shallow soils dominate the steepest of slopes. Recharge shallow soils also occur on the hillcrest terrain positions on the highest peaks where overland flow and relatively quick recharge conditions are the dominant hydrological processes. This could potentially be based on differences in parent material between soil classes. The less steep midslope and footslope terrain positions are dominated by recharge deep soils where vertical drainage of water through the soil profile is the dominant hydrological

response. This vertical drainage is most likely followed up by the lateral movement of water within the shallow groundwater aquifer.

The deep soils resulted from the processes of illuviation and colluviation, where the clayey material has been predominantly removed from midslope positions, with colluvial deposits forming at footslope terrain positions. The valley-bottom terrain positions are primarily dominated by responsive saturated soils typical of riparian and wetland areas, where saturation excess leads to the overland flow of water on top of the soil surface. These soils most likely originate due to the process of eluviation which results in the addition of clays from upslope terrain positions to these soils, expressed as the gleyed subsoils clay rich of these terrain positions.

The lowland hillcrest and midslope terrain positions are dominated by soil/bedrock interflow and recharge shallow soils where the lateral movement of water at the soil/bedrock interface (more prevalent in the east and southeast) and vertical drainage to the shallow aquifer (more prevalent in the northwest and west) are the dominant hydrological responses. A/B interflow soils occur on footslope positions where the dominant hydrological process is the lateral movement of water at the topsoil/subsoil interface where the textural discontinuity between soil horizons lead to the build-up and lateral movement of water through the soil profile. These soils were created by the eluviation of clays from the recharge shallow and soil/bedrock interflow upslope positions. Valley-bottom terrain positions are dominated by a mixture of recharge deep, due to sandy floodplains on the major river networks forming during periodic flooding caused by cyclone events, and responsive saturated soils on the valley bottom positions which are not exposed to the same periodic flooding.

The value of downscaling and balancing legacy soil data

In South Africa, a large volume of available legacy soil data exists, which remains largely untapped within commercial and semi-commercial sources that have not been freely available in the past (Paterson et al., 2015). However, these sources have become more frequently available in recent years by the improved cooperation between various public and private sector stakeholders as seen with the acquisition of the SAFCOL legacy soil dataset. The opportunity to add substantial amounts of additional spatially localised legacy soil data for use in digital soil mapping across South Africa should also coincide with research regarding how best to apply these datasets for digital soil mapping purposes.

The 500ha/observation downscaled dataset statistically improved the existing legacy soil dataset and provided the best representation of environmental covariates within the catchment, resulting in improved mapping accuracy prior to further balancing. Therefore, the downscaling of spatially localised legacy soil information to improve environmental covariate representation is an effective tool to improve the representation of legacy soil datasets. However, these results only consider environmental covariate representation and not necessarily the representativeness of the minority soil classes, which was why SMOTE was still required to improve the representation of the minority soil classes.

However, simply applying SMOTE balancing of mapping units using all available soil information still provided comparative results when evaluating point accuracy and Kappa coefficient values. When balancing soil classes using SMOTE was applied to our best representative legacy soil dataset, the resulting hydrological soil mapping accuracy was significantly improved compared to using only SMOTE balancing on all available soil information. These improved results using SMOTE are in accordance with results from others (Taghizadeh-Mehrjardi et al., 2019; Chawla et al., 2002; Tantithamthavorn et al., 2018). The value of downscaling spatially localised legacy soil information therefore must coincide with additional class balancing when using MNLR. These results also reaffirm the importance of balancing legacy soil information for mapping units across the entire catchment area because

imbalanced data affects the predictive ability of the MNLR algorithm. Our results confirm the ability of the SMOTE resampling technique to handle major class imbalances.

However, simply finding and adding additional highly imbalanced legacy soil data and balancing the minority soil classes using SMOTE did not improve mapping accuracy using MNLR compared to the best performing environmental covariate balanced map. Therefore, there is a limit to the capability of SMOTE for resampling and care should be taken when applying synthetic data generation techniques to balance legacy soil data. Emphasis should remain on using a representative legacy soil dataset for digital soil mapping purposes.

Sharififar et al. (2019b) did not encounter this problem because the researchers collected soil samples from a grid spacing of 500 m across their entire study area of 12 000 ha, meaning that environmental covariate representation of soil samples was guaranteed. However, as the size of the study area increases, grid sampling becomes too costly and time consuming, where legacy soil datasets are not necessarily representative of environmental covariates. Taghizadeh-Mehrjardi et al. (2019) used the national soil database of Iran consisting of 7 664 samples, which was created using stratified random sampling (~87%), grid sampling (~8%) and the conditioned latin hypercube sampling approach (~5%), with the assumption being that these samples are representative of the soils and environmental covariates of Iran. However, the readily available soil observations in the national soil database of South Africa (Land Type Survey Staff, 1976-2006) amounts to 2 500 modal profile observations, less than half of that of Iran. This study therefore provides a relevant protocol to use highly spatially localised legacy soil datasets to improve accuracy of digital soil mapping by downscaling and adding additional soil observations that improve overall representation and balance of environmental covariates data within the legacy soil dataset.

Therefore, downscaling highly spatially localised legacy soil data using k-means clustering for improving environmental covariate representation is an effective method to improve accuracy of digital soil mapping . Our best performing hydrological soil map thus more accurately represents the dominant hydrological processes throughout the Sabie-Sand catchment, than the readily available soil information in South Africa. An improved representation of internal catchment processes could hold the key to improved climate- and land-use change scenario analyses, and improved water resource management practices at catchment scale due to the fact that the major hydrological processes are better understood both spatially and temporally. This approach could potentially aid in the mapping of the hydrological soils for macro-scale catchments, by optimally using large, spatially localised, imbalanced legacy soil datasets, such as soil surveys within the forestry, mining, and agricultural sectors. The approach may be particularly applicable in South Africa where large amounts of spatially localised legacy soil information exists within large scale mapping projects, such as creating a hydrological soil map of South Africa.

Balancing spatially localised legacy soil datasets for use in large scale digital soil mapping extends beyond merely downscaling and upscaling soil observations of the majority and minority soil classes. The representativeness of soil observations within the catchment environmental covariates are measurable and are indicative of how accurate the resulting soil maps should be. Downscaling highly spatially localised legacy soil observations should strive to improve catchment representation and is also easily repeatable across soil datasets of various sizes and various resolutions.

A problem in dealing with spatially localised legacy soil observations is that a comparatively small validation dataset is used to validate mapping results. In this study only 152 observations were used which creates further uncertainty regarding the accuracy of minority soil classes within the different hydrological soil maps, as limited observations are available to validate the minority soil classes

(Sharififar et al., 2019b). Future research should focus on different balancing approaches. At the data level, different upscaling and downscaling approaches across different catchment sizes with different imbalanced legacy soil datasets preferably with larger validation datasets should be applied. Research should also be conducted at the algorithm level, where algorithms specifically developed to handle class imbalanced datasets, such as boosting algorithms, should also be considered as viable alternatives to well-established digital soil mapping algorithms. Recently, using the algorithm-level approach to deal with imbalance soil data revealed that a one-class support vector machine combined with multi-class classification yielded the most accurate soil map and adequately represented the minority soil class (Sharififar & Sarmadian, 2022). Lastly, a combination of these approaches should also be researched once the best approaches of each level have been firmly established in the field of soil science.

3.1.5 Conclusions

An accurate hydrological soil map of the macro-scale Sabie-Sand catchment was created using machine learning-based digital soil mapping and legacy soil information. The downscaling of spatially localised legacy soil datasets to improve the representation of environmental covariates was applied using k-means clustering, where the 500ha/observation dataset resulted in the best improved representation of catchment environmental covariates. However, the improved catchment representation does not necessarily result in improved mapping accuracy, especially when dealing with imbalanced soil mapping classes. Further balancing the imbalanced soil classes of the 500ha/observation dataset using SMOTE, significantly improved mapping accuracy compared to using SMOTE on all available soil information.

Therefore, downscaling spatially localised legacy soil information is an effective tool to improve legacy soil data covariate balance and representation which leads to improved accuracy of digital soil mapping. This approach is of value where large spatially localised datasets exist. Our main recommendation would be to further test the use of downscaling spatially localised legacy soil information to improve digital soil mapping accuracy across different catchment sizes with different legacy soil datasets.

3.2 EXAMINING THE VALUE OF HYDROPEDOLOGICAL INFORMATION ON HYDROLOGICAL MODELLING AT DIFFERENT SCALES IN THE SABIE CATCHMENT, SOUTH AFRICA

3.2.1 Abstract

Detailed soil information is increasingly sought after for catchment-scale hydrological modelling to better understand the soil-water interactions at a landscape level. In South Africa, 8% of the surface area is responsible for 50% of the mean annual runoff. Thus, understanding the soil-water dynamics in these catchments is imperative to future water resource management. In this study, the value of hydropedological information is tested by comparing a detailed hydropedological map based on in-field soil information to the best readily available soil information at five different catchment sizes (48 km^2, 56 km^2, 174 km^2, 674 km^2 and 2 421 km^2) using the SWAT+ model in the Sabie catchment, South Africa. The aim was to determine the value of hydropedological information at different scales as well as illustrating the value of hydropedology as 'soft data' to improve hydrological process representation. Better hydropedological information significantly improved long-term streamflow simulations at all catchment sizes, except for the largest catchment (2 421 km^2). It is assumed that the resulting improved streamflow simulations are a direct result of the improved hydrological process representation achieved by the hydropedological information. Here we argue that hydropedological information should form an important 'soft data' tool to better understand and simulate different hydrological processes.

3.2.2 Introduction

One of the modern challenges related to water resource management is understanding and representing soil-water interactions within a landscape-scale context (Kahmen et al., 2005; Smith, 2014; Zhang et al., 2015; Wei et al., 2016). With both soil and water being fundamental components of the hydrological processes within a catchment, understanding these processes is imperative to understanding how a catchment responds to different land-use management and climate change regimes (Bouma, 2016). Therefore, accurate soil information, especially soil hydraulic properties, are an important input parameter into physically-based hydrological models (Worqlul et al., 2018). The interdisciplinary field of hydropedology (Lin, 2003) includes the fields of hydrology, pedology and soil physics and enables the study of soil-water dynamics at various scales (Lin et al., 2005). Hydropedology has been particularly applicable in hydrological modelling as it provides a more detailed spatial understanding of soil-water interactions by improving the accuracy of internal catchment hydrological processes such as infiltration, runoff, lateral flow, percolation, return flow and evapotranspiration at different scales within greatly varying catchments (Bryant et al., 2006; Me et al., 2015).

By combining modern techniques for digital soil mapping (McBratney et al., 2003) with hydropedological insight (Lin et al., 2006; Van Tol et al., 2021a), soil scientists can now produce detailed large-scale hydrological soil datasets for modelling purposes (Julich et al.,2012; Van Tol et al., 2015; Wahren et al., 2016; Van Zijl et al., 2020; Van Tol & Van Zijl, 2022). Several studies have indicated that improved soil information does indeed improve hydrological modelling efficiency (Romanowicz et al., 2005; Bossa et al., 2012; Diek et al., 2014; Smit & Van Tol, 2022). A notable instance is the utilisation of the Soil Land Inference Model (SoLIM) to produce a more detailed soil map in a catchment with limited data in north-central Portugal (Wahren et al., 2016). The refined soil map yielded a 7% enhancement in prediction accuracy compared to traditional soil data, and simultaneously, it contributed to a reduction in parameter uncertainty.

Detailed hydropedological information has also been applied widely in South Africa. For example, hydrological soil information applied to three different catchment sizes (640 km^2, 550 km^2, 54 km^2) in an urbanised catchment improved modelling accuracy at all three catchment sizes when compared to readily available soil information (Van Tol et al., 2021a). Although long-term streamflow simulations

were similar using hydropedological information compared to readily available soil information, hydropedological information substantially improved the simulation of soil hydrological processes (Smit & Van Tol, 2022). These results also indicate that accurate streamflow simulations do not necessarily mean accurate internal hydrological processes. Hydropedological insight and measured hydraulic properties substantially improved lateral flow simulations in a mountainous catchment of South Africa (Harrison et al., 2022). However, others maintain that the statistically small modelling improvements do not necessarily justify the cost and time to gather the improved soil information, or that improved soil information does not necessarily lead to more accurate hydrological modelling (Geza & McCray, 2008; Chen et al., 2016).

The value of hydropedological information transcends its role in precisely modelling long-term streamflow predictions. There is an argument that hydropedological information can serve as a powerful 'soft data' tool, particularly in basins lacking reliable streamflow data (Seibert & McDonnell, 2002). The term 'soft data' in hydrological modelling refers to information that may not be directly measured but can be linked to hydrological processes (Winsemius et al., 2009; Van Tol et al., 2021b).

In this study, the primary objective was to evaluate the impact of hydropedological information on process-based hydrological modelling. This was achieved by statistically comparing long-term streamflow modelling accuracy using two levels of soil information, namely, hydropedological information and South Africa's most readily available soil information. The focus was on the direct contribution of soil information on modelling efficiency and therefore we did not calibrate the model through extensive automated calibration techniques to favour one model, but rather kept all inputs constant except for the soil information between modelling simulations and essentially treating the catchment as ungauged.

3.2.3 Materials and methods

The Sabie catchment

The 2 421 km^2 Sabie catchment is located in the Mpumalanga province of South Africa (Figure 3.5) and forms part of the larger transboundary Incomati River basin. The catchment stretches from the Drakensberg escarpment in the west at an altitude of 2 218 m.a.s.l. and gradually flattens towards the east with an altitude of 250 m.a.s.l. as the Sabie River continues to flow eastward into Mozambique. The ambient geology is predominantly crystalline (igneous and metamorphic) and comprises various ranges of bedrock lithologies, primarily quartzites, granites, andesites and gneiss formations (Council for Geoscience, 2007).

With a semi-arid warm and hot climate in the east of the catchment and a temperate warm climate in the west, a strong rainfall gradient exists ranging from 1 600 mm in the west to 550 mm in the east. Rainfall occurs mainly in the austral summer and normally results from convective thunderstorms, although periodic high-intensity rainfall events do occur from cyclones that form over the Indian Ocean and track inland, where the orographic effect of the Drakensberg escarpment creates severe localised flooding (Kruger et al., 2002). The main bioregions of the catchment consist of savanna at lower altitudes and montane grasslands and montane forests in the mountainous regions, which have been heavily altered by commercial forestry plantations (Mucina & Rutherford, 2006).

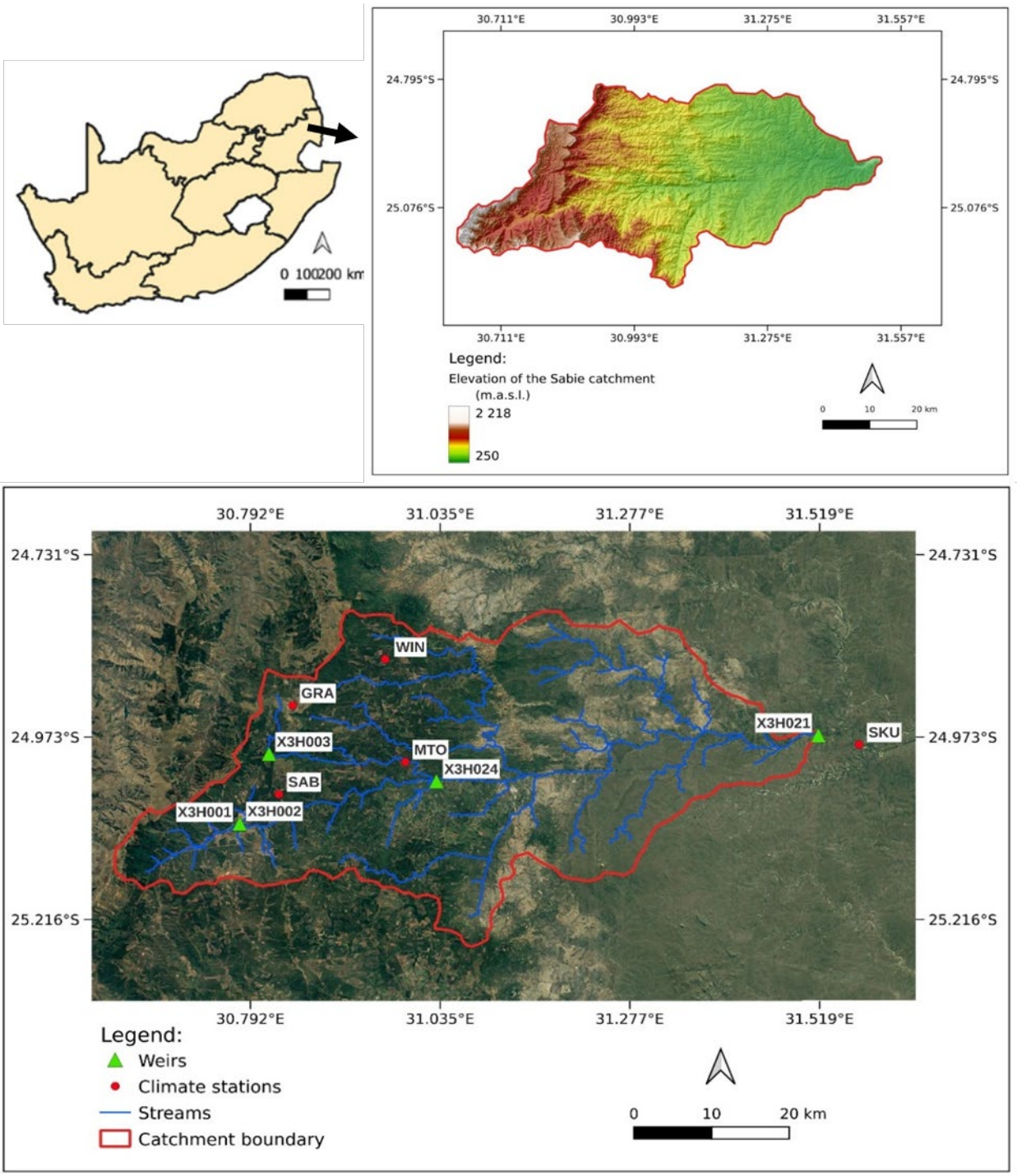

Figure 3.5: The Sabie catchment, including elevation, weirs and climate stations.

Model, inputs and setup

The SWAT (Soil and Water Assessment Tool) model is a process-based, semi-distributed catchment model which is widely used to simulate water quality and quantity predictions, and assess the impacts of physical changes such as land use and climate changes in catchments across the globe (Neitsch et al., 2011). SWAT+ is an enhanced iteration of the renowned SWAT model (Arnold et al., 1998; Bieger et al., 2017). The QSWAT+ (v. 2.3) plugin was used to set up the catchment. As an initial step, the model partitions the catchment into hydrological response units (HRUs), with each HRU representing a homogenous area in terms of soil, land use and slope. The model then calculates a range of water balance components for each individual HRU, including overland flow, infiltration, lateral flow,

percolation, return flow, evapotranspiration and discharge to the stream. The model was run from the start of January 2000 until the end of December 2019. The model warm-up period lasted for the first four years, followed by a 16-year validation period.

Daily rainfall data was obtained from four climate stations, namely, Sabie, Dunnottar at MTO Forestry, Rietspruit near God's Window and at Skukuza. Minimum and maximum temperatures were obtained from two climate stations, namely, Skukuza and Graskop. All data was received curtesy of the South African Weather Service. Daily solar radiation, relative humidity and wind speed were obtained from the Climate Forecast System Reanalysis which was done by the National Center for Environmental Prediction (Saha et al., 2015).

The DEM was obtained from a 30 m x 30 m SRTM (USGS, 2022). The land-cover data (Figure 3.6) were acquired from the 2013/14 South African National Land Cover Map (GeoTerra Image, 2015). For the land cover input, predefined SWAT values associated with various land-use classes were utilised. Additionally, dams identified in the land cover were integrated into the model setup, designated as 'reservoirs' and assigned default values.

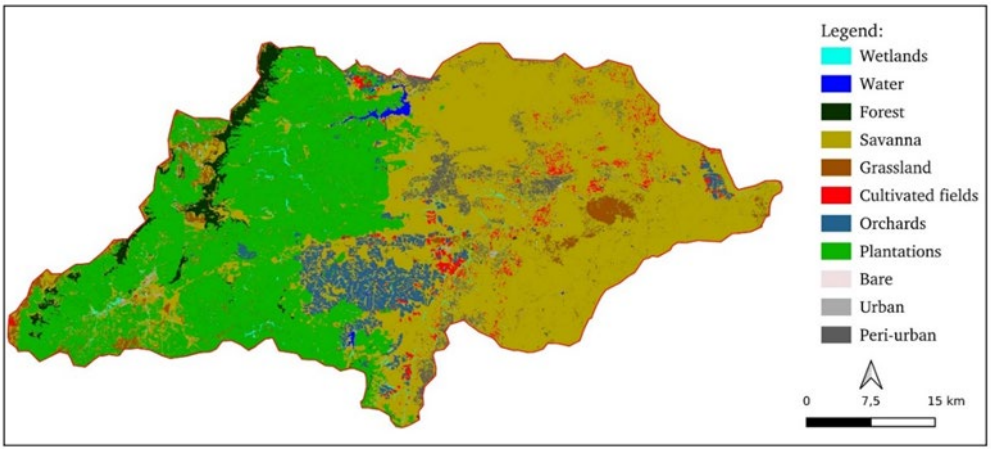

Figure 3.6: The land uses within the Sabie catchment as demarcated from the 2013/2014 National Land Cover Map.

Both model runs were left uncalibrated to ensure that differences between model runs were only due to differences in soil input information, so that direct comparisons between soil datasets could be made. This was done since any form of manual calibration would benefit either one or the other in predicting streamflow accuracy because of differences in how each model simulated different hydrological processes. This in essence meant that the catchment was treated as ungauged for the duration of the study.

Soil information

Soil properties govern the movement of water and air through the soil profile and have a major impact on the cycling of water, sediment and nutrients within each HRU. SWAT+ requires both the spatial soil mapping unit as well as physical properties for each individual soil horizon with the unit, such as depth to bottom of soil layer, bulk density, available water capacity, saturated hydraulic conductivity, organic carbon content, clay, silt, sand, and rock fragment contents, as well as moist soil albedo and the soil erodibility factor.

The Land Type database was developed between 1972 and 2002 and covers the entire country of South Africa at a 1:250 000 scale. A Land Type polygon is defined as "a homogeneous, unique

combination of terrain type, soil pattern and macroclimate zone." The Land Type survey identified 7 070 unique Land Type polygons based on some 400 000 soil observations (approximately 1 observation per 300 ha) (Paterson et al., 2015). The Land Type database has already been converted to a readily available spatial soil database specifically for use within the SWAT model (Le Roux et al., 2023). In the Sabie catchment, there are seven broad Land Type groups which could be divided into 42 different individual Land Types each with their own set of hydraulic properties (Figure 3.7; Table 3.7).

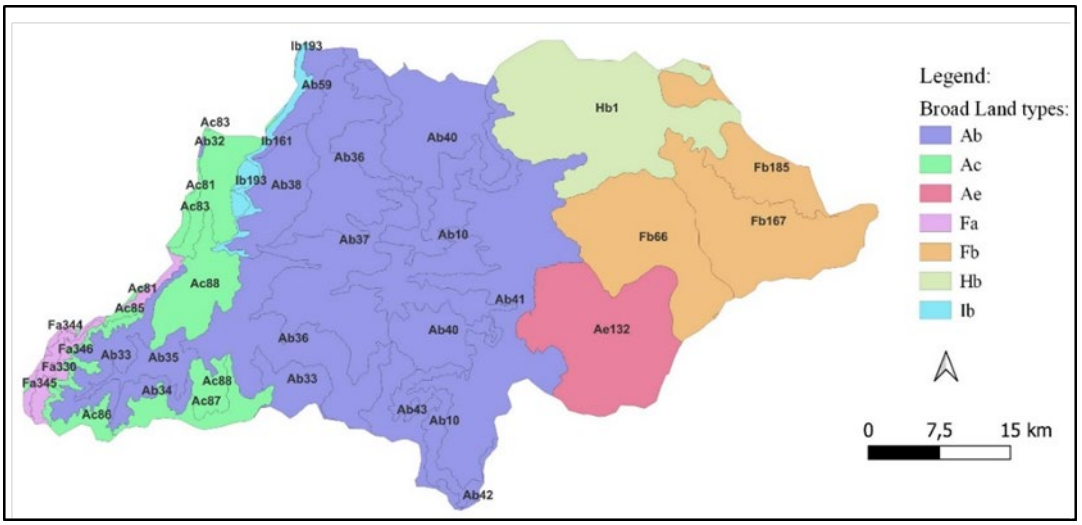

Figure 3.7: The Land Types present within the Sabie catchment (Land Type Survey Staff, 1972-2002).

Table 3.7: The main hydraulic properties of the Land Type mapping units (means are followed by minimum and maximum values in brackets).

Broad Land Types	Horizon	Hydro-group	Depth	Bd	AWC	Ksat	OC	Clay	Silt	Sand
			mm	g/cm^3	mm/mm	mm/h	%	%	%	%
Ae	A	A	270 (230,300)	1.45 (1.40,1.50)	0.085 (0.08,0.094)	145 (13,210)	0.7	8.3	13.3	78.3
	B		550 (450-670)	1.45 (1.40,1.50)	0.086 (0.68,0.09)	85 (13,210)	0.1	9.3	13.3	77.3
Ac	A	A	270 (230,300)	1.45 (1.40,1.50)	0.087 (0.085,0.089)	146 (13,210)	0.5	6.2	12.5	81.3
	B		500 (400-700)	1.45 (1.40,1.50)	0.097 (0.74,0.11)	135 (13,210)	0.1	6.6	12.5	80.9
Ab	A	A	290 (240,300)	1.45 (1.40,1.50)	0.089 (0.086,0.094)	147 (13,210)	0.8	9.2	13.3	77.5
	B		500 (450-650)	1.45 (1.40,1.50)	0.093 (0.68,0.10)	85 (13,210)	0.1	9.4	13.3	77.3
Fb	A	C	270 (270,270)	1.50 (1.40,1.60)	0.074 (0.068,0.080)	9 (4.3,13)	1.0	19.3	17.5	63.2
	B		550 (450-670)	1.45 (1.40,1.50)	0.056 (0.48,0.09)	9 (4.3,13)	0.2	19.3	17.5	63.2
Fa	A	C	280 (270,290)	1.50 (1.4,1.60)	0.08 (0.074,0.090)	35 (13,210)	0.9	15.3	16.3	68.5
	B		300 (300,300)	1.45 (1.40,1.50)	0.07 (0.68,0.09)	35 (13,61)	0.1	15.3	16.3	68.5
Hb	A	B	250 (210,300)	1.45 (1.40,1.50)	0.076 (0.068,0.090)	145 (13,210)	0.5	11.5	14.7	73.7
	B		300 (300-300)	1.45 (1.40,1.50)	0.076 (0.68,0.094)	50 (13,210)	0.1	11.5	14.7	73.7
Ib	A	D	290 (290,290)	1.45 (1.40,1.50)	0.088 (0.086,0.090)	110 (13,210)	0.4	11.0	15.0	74.0
	B		375 (300-450)	1.45 (1.40,1.50)	0.085 (0.68,0.09)	40 (13,61)	0.1	11.0	15.0	74.0

Bd = bulk density; AWC = Available Water Capacity; Ksat = saturated hydraulic conductivity; OC = organic carbon

The second soil dataset was the hydropedological dataset (HYDROSOIL) developed using modern techniques for digital soil mapping, an in-field hydropedological soil survey and legacy soil information. Details on the digital soil mapping approach are described in Section 3.1 (Smit et al., 2023a) but briefly, we developed an extensive environmental covariate dataset which included, geology, terrain variables such as planform curvature, profile curvature, etc., climate variables such as mean annual minimum temperature, mean annual maximum temperature, etc., and lastly spectral covariates such as brightness index, colouration index, redness index, saturation index and NDVI values for both the wet and dry seasons. A massive number of legacy soil observations (n = 12 875) were obtained from various legacy soil datasets which were reclassified in accordance with the hydropedological groupings of South African soils (Table 3.8) (Van Tol & Le Roux, 2019). A further 108 soil observations were made by hand auger during an in-field hydropedological survey which underwent the same reclassification.

Table 3.8: The characteristics of the hydrological mapping units of the Sabie catchment.

Hydrological mapping unit	Soil form	WRB Reference Groups	Defining hydrological characteristic
Recharge deep	Hutton, Longtom, Kranskop	Acrisols, Nitisols, Fluvisols	Deep soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying bedrock is the dominant flow direction.
Recharge shallow	Glenrosa, Nomanci	Leptosols	Shallow soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying fractured bedrock is the dominant flow direction.

Hydrological mapping unit	Soil form	WRB Reference Groups	Defining hydrological characteristic
Responsive saturated	Katspruit, Champagne	Gleysols	Soils with morphological evidence of long periods of saturation promoting the generation of overland flow due to saturation excess.
Responsive shallow	Mispah, Graskop	Leptosols	Shallow soils overlying relatively impermeable bedrock. Limited storage capacity results in the generation of overland flow after rainfall events.
A/B interflow	Estcourt, Sterkspruit	Solonetz	Duplex soils where the textural discontinuity facilitates build-up of water in the topsoil, with discharge in a predominantly lateral direction.
Soil/bedrock interflow	Fernwood, Cartref	Arenosols	Soils overlying relatively impermeable bedrock. Hydromorphic properties signify temporal build of water on the soil/bedrock interface and slow discharge in a predominantly lateral direction.

WRB = World Reference Base for Soil Resources

The hydropedological database was divided into training (75%) and evaluation (25%) datasets. We used the well-known k-means clustering algorithm to overcome the imbalance training data. The final soil map was then created in the R environment by running the multinomial logistic regression algorithm on the training data and using the validation data to test the accuracy of the hydropedological map, which had an evaluation point accuracy of 62% and a Cohen's Kappa statistic value of 0.46. These results indicated that the hydropedological map obtained moderate agreement with reality and was therefore deemed to be acceptable for use in the modelling exercise (Figure 3.8).

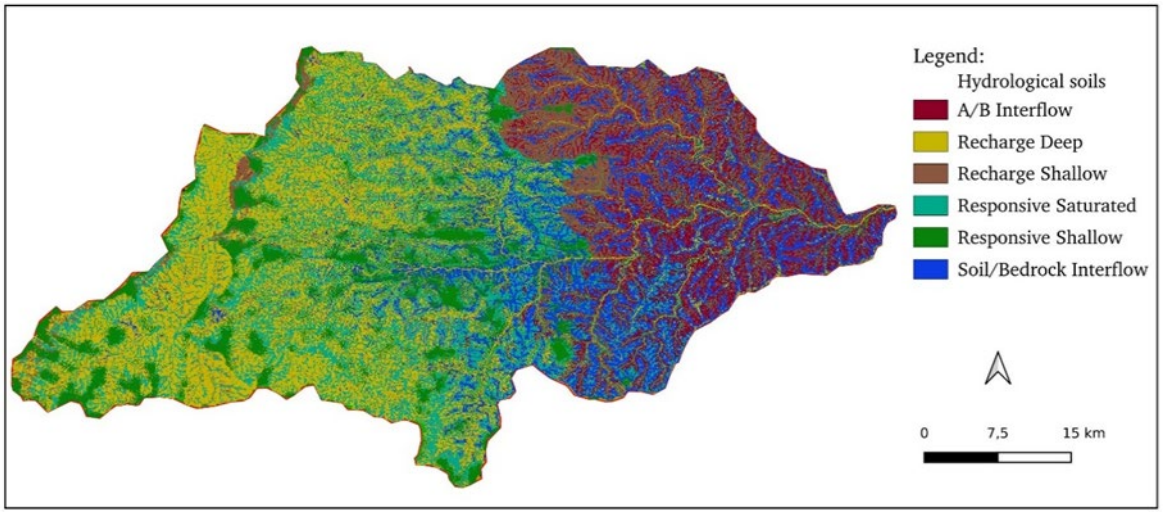

Figure 3.8: The hydropedological map of the Sabie catchment.

Undisturbed core samples were collected from 78 representative diagnostic horizons within the study area during the field survey. These core samples were used to determine bulk density, particle size distribution and the water retention characteristics. These results were combined with the already existing Land Type modal profile data, and then the required SWAT+ hydraulic parameters were obtained by averaging these properties for each hydropedological soil type (Table 3.9).

Table 3.9: The main hydraulic properties of the HYDROSOIL mapping units.

Hydrological soil types	Horizon	Hydro-group	Depth	Bd	AWC	Ksat	OC	Clay	Silt	Sand
			mm	g/cm^3	mm/mm	mm/h	%	%	%	%
A/B interflow	A	C	150	1.38	0.101	40.1	1.1	18	21	61
	B		1400	1.58	0.122	8.54	0.2	40	13	47
Recharge deep	A	A	200	1.41	0.107	11.58	5.2	31	27	42
	B		2400	1.46	0.119	7.8	1.5	34	29	37
	C		3000	1.51	0.119	5.7	0.2	33	32	35
Recharge shallow	A	D	400	1.4	0.107	27.27	5.5	32	12	56
Responsive saturated	A	C	200	1.36	0.104	32.01	4	25	18	57
	B		1800	1.51	0.121	11.74	1.8	38	14	48
Responsive shallow	A	D	300	1.34	0.104	40.5	5	25	15	60
Soil/bedrock interflow	A	A	200	1.4	0.100	40	1.5	17	20	62
	B		1600	1.38	0.113	78	0.4	19	5	76

Bd = bulk density; AWC = Available Water Capacity; Ksat = saturated hydraulic conductivity; OC = Organic Carbon

Two model runs were set up for the two levels of soil information. Only the soil information differed between setups as all other factors were constant for both simulation runs. However, the HYDROSOIL and Land Type soil datasets differed both spatially (Figure 3.7; Figure 3.8) and in their hydraulic properties (Table 3.7; Table 3.9).

Researchers have shown the soil conservation service curve number II (CN2) as the most sensitive parameter in SWAT streamflow simulations (Eckhardt, 2005; Mengistu et al., 2019) as it reflects the characteristics of the catchment prior to a rainfall event and largely determines surface runoff. It is dependent on the initial CN value assigned to the HRU by the model. CN2 is the runoff curve number for Moisture Condition II, calculated by the soil conservation service (SCS) runoff equation and adjusted soil moisture before a precipitation event. The CN2 value is therefore also directly affected by the initial soil hydrologic group of each soil mapping unit and will differ between model runs. Surface runoff is calculated using the following equations:

$$Qsurf = \frac{(P_i - I_a)^2}{(P_i - I_a + S)} \quad (3.5)$$

where Q_{surf} is the overland runoff or rainfall excess (mm H_2O), P_i is the precipitation depth for the day (mm H_2O), I_a is the initial abstraction lost from canopy interception, surface storage, and infiltration prior to runoff. The water retention parameter (S) (mm H_2O) is estimated by:

$$S = 25.4\left(\frac{1000}{CN} - 10\right) \quad (3.6)$$

Where CN is the curve number at a daily time step which is a function of soil permeability, land use and antecedent soil moisture content. The values are based on the soil hydrologic group of the soil mapping unit, land use and initial hydrologic condition, with the soil hydrologic group and land use being the most important variables within the equation. In addition, the value of each HRU is updated according to the antecedent soil moisture content for each daily timestep (Neitsch et al., 2011; Zhang et al., 2019a).

SWAT divides soil into four distinct soil hydrologic groups based on the infiltration characteristics of the soil, namely, A, soils with a low runoff potential, containing high infiltration rates and being well drained with a high rate of water transmission. B, soils with moderate infiltration rates with moderate rates of water transmission and being moderately well drained. C, soils with low infiltration rates often containing a layer that impedes the downward movement of water with low rates of water transmission. D, soils with a high runoff potential, with very slow infiltration rates with very slow rates of water transmission (Neitsch et al., 2011). These hydrologic groups largely determine the surface runoff potential of different soils as they directly affect the SCS curve designation given by the model. As these hydrologic groups differ spatially between datasets, the curve numbers and associated runoff characteristics will differ greatly between model runs.

Lateral flow is calculated by SWAT using a kinematic storage model, which simulates the movement of water in a two-dimensional cross section of a hillslope (Neitsch et al., 2011). Lateral flow therefore occurs when soil water exceeds field capacity with the underlying layer being impermeable or semi-permeable. The kinematic approximation method assumes that the flowpaths are parallel to the bedrock and that the hydraulic gradient equals the slope of the hill (Equation 3.7).

$$SW_{excess} = \frac{1000.H_0.\Theta_d.L_{hill}}{2} \quad (3.7)$$

Where SW_{excess} equals the drainable water volume within the saturated zone of the soil per unit area (mm), H_0 equals the saturated thickness of the hillslope outlet as a fraction of the total thickness (mm/mm), Θ_d equals the drainable porosity of the soil (mm/mm), and L_{hill} equals the length of the hillslope (m) (Neitsch et al., 2011). The drainage porosity of the soil equals the total porosity of the soil minus the soil porosity when the soil horizon is at field capacity. The increased spatial resolution of the HYDROSOIL map should result in an increased number of HRUs which would result in a more complex model structure compared to the Land Type dataset. More HRUs and differences in porosity between soil datasets will affect how the model simulates lateral flow values.

The differences in hydraulic properties between the two levels of soil information should also affect modelling accuracy. The increased soil depth, Available Water Capacity (AWC), clay content and decreased Ksat values of the HYDROSOIL map should result in more water being stored within the soil profile for longer periods, leading to more available water for root uptake, plant growth and evapotranspiration. More antecedent moisture within the soil should also lower CN2 values, which remains one of the most sensitive parameters within the SWAT model (Wahren et al., 2016, Mengistu et al., 2019).

Validation data and statistical comparison

Five weirs, which are managed by the Department of Water and Sanitation (DWS), were used to validate long-term streamflow simulations. These gauges, from smallest drainage area to largest were X3H003 which drains 48 km^2, X3H002 which drains 56 km^2, X3H001 which drains 174 km^2, X3H024 which drains 674 km^2 and X3H021 which drains the entire study area at 2 421 km^2. Daily streamflow was converted to monthly average values for comparison purposes.

For statistical comparison, four widely used statistical indicators were employed, namely coefficient of determination (R^2), percentage bias (PBIAS), Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE). Percentage bias (PBIAS) measures the average tendency of the simulated data to be larger or smaller than their observed counterparts. The optimal value of PBIAS is 0.0, with low

magnitude values indicating accurate model simulation. Positive values indicate model underestimation bias, while negative values indicate model overestimation (Moriasi et al., 2007).

3.2.4 Results and discussion

Streamflow simulations

The two model set-ups for the two levels of soil information had an identical number of sub-basins (119) and landscape units (616), because the same DEM was used to delineate these. The number of HRUs differed significantly where the HYDROSOIL model contained 11 883 HRUs compared to the 3 332 HRUs contained within the Land Type model. The large discrepancy between model HRUs is purely a result of the spatial differences between the soil input information. Even though the HYDROSOIL soil dataset contained less individual mapping units, the far greater level of detail (30 m x 30 m) of these mapping units still resulted in a significantly increased number of HRUs.

A KGE value surpassing -0.41 indicates a model prediction that aligns better with the mean observed values (Knoben et al., 2019). Refined evaluation criteria for hydrologic and water-quality models deem streamflow simulations to be satisfactory when $R^2 > 0.6$, NSE > 0.5, and PBIAS ≤ 15% (Moriasi et al., 2015).

Based on these criteria (Moriasi et al., 2015), the simulations of the HYDROSOIL model at gauging weirs at 48 km^2, 174 km^2 and 674 km^2 all yielded satisfactory results ($R^2 > 0.6$; Table 3.10). On the other hand, the Land Type model simulations provided satisfactory results at 174 km^2 and 674 km^2. All HYDROSOIL simulations achieved satisfactory KGE values above the -0.41 threshold (Knoben et al., 2019). However, the Land Type model did not meet the minimum KGE threshold at 48 km^2 and 56 km^2 (Table 3.10).

Both models produced disappointing PBIAS values, where only the Land Type model achieved PBIAS values below the 15% threshold (Moriasi et al., 2015) at 674 km^2 and 2 421 km^2. However, the HYDROSOIL model provided more accurate PBIAS values at 48 km^2, 56 km^2 and 174 km^2, although they did not meet the 15% criteria. Analysing NSE values, the HYDROSOIL model outperformed the Land Type model at each catchment scale, with only the HYDROSOIL model achieving an acceptable NSE value at 2 421 km^2 (Table 3.10).

Table 3.10: Statistical indicators of monthly streamflow simulations at five catchment levels.

Catchment	Soil data	R^2	PBIAS	NSE	KGE
X3H003 (48 km^2)	Land Type	0.46	68.27	-0.76	-0.43
	HYDROSOIL	0.66	53.92	0.03	0.41
X3H002 (56 km^2)	Land Type	0.42	37.34	-3.24	-0.55
	HYDROSOIL	0.57	43.67	-0.22	0.41
X3H001 (174 km^2)	Land Type	0.68	41.04	0.3	0.48
	HYDROSOIL	0.67	37.27	0.48	0.58
X3H024 (674 km^2)	Land Type	0.7	11.6	-0.41	0.09
	HYDROSOIL	0.71	20.85	0.54	0.67
X3H021 (2421 km^2)	Land Type	0.44	13.29	0.28	0.63
	HYDROSOIL	0.54	33.5	0.49	0.42

PBIAS = Percentage bias; NSE = Nash Sutcliffe Efficiency; KGE = Kling-Gupta Efficiency

Peak flows were overestimated by both models. However, the HYDROSOIL dataset yielded far lower peak flows than the Land Type dataset (Figure 3.9), which improved modelling accuracy at smaller scales (48 km^2, 56 km^2 and 174 km^2), but resulted in the underestimation of peak flows at the largest catchment scale (2 421 km^2). Baseflow simulations were also substantially underestimated by both models at all catchment levels but particularly at smaller catchment sizes (48 km^2, 56 km^2 and 174 km^2) where considerable baseflow contributions exist (Figure 3.9). The positive PBIAS values across all model simulations also equates to the general underestimation of total streamflow values which can also be attributed to the underestimation of baseflow values across all catchment levels. SWAT+ allows users to adjust groundwater parameters to mitigate or correct baseflow values.

Statistical comparison between the two models over the 16-year simulation period indicated a substantial difference between the two levels of soil information (Table 3.10). It could be inferred that the HYDROSOIL model outperformed the Land Type model, at four of the five catchment scales (48 km^2, 56 km^2, 174 km^2, 674 km^2) when comparing monthly observed streamflow values where the higher R^2, NSE and KGE values indicated improved model performance. The improved modelling accuracy is presumed to be a direct result of the improved simulation of real-world hydrological processes by improving the accuracy of soil information in the model. However, it does seem as if the importance of soil information decreases as the catchment size increases, which could be a result of the increased variability of soil properties, such as texture, organic matter content, and hydraulic conductivity, which may vary widely across catchments. The ineffectiveness of improved soil information at our largest catchment scale (2 421 km^2) is in accordance with both Chen et al. (2016), who found soil resolution was relatively insignificant when modelling streamflow and sediment yield in a 2 421 km^2 catchment, and Ayana et al. (2019) who concluded that improved soil data resolution only marginally improved streamflow simulations with an improved NSE of only 1% in a 16 000 km^2 catchment in Ethiopia. These results suggest that improved soil information does not necessarily improve modelling accuracy in large-scale catchments.

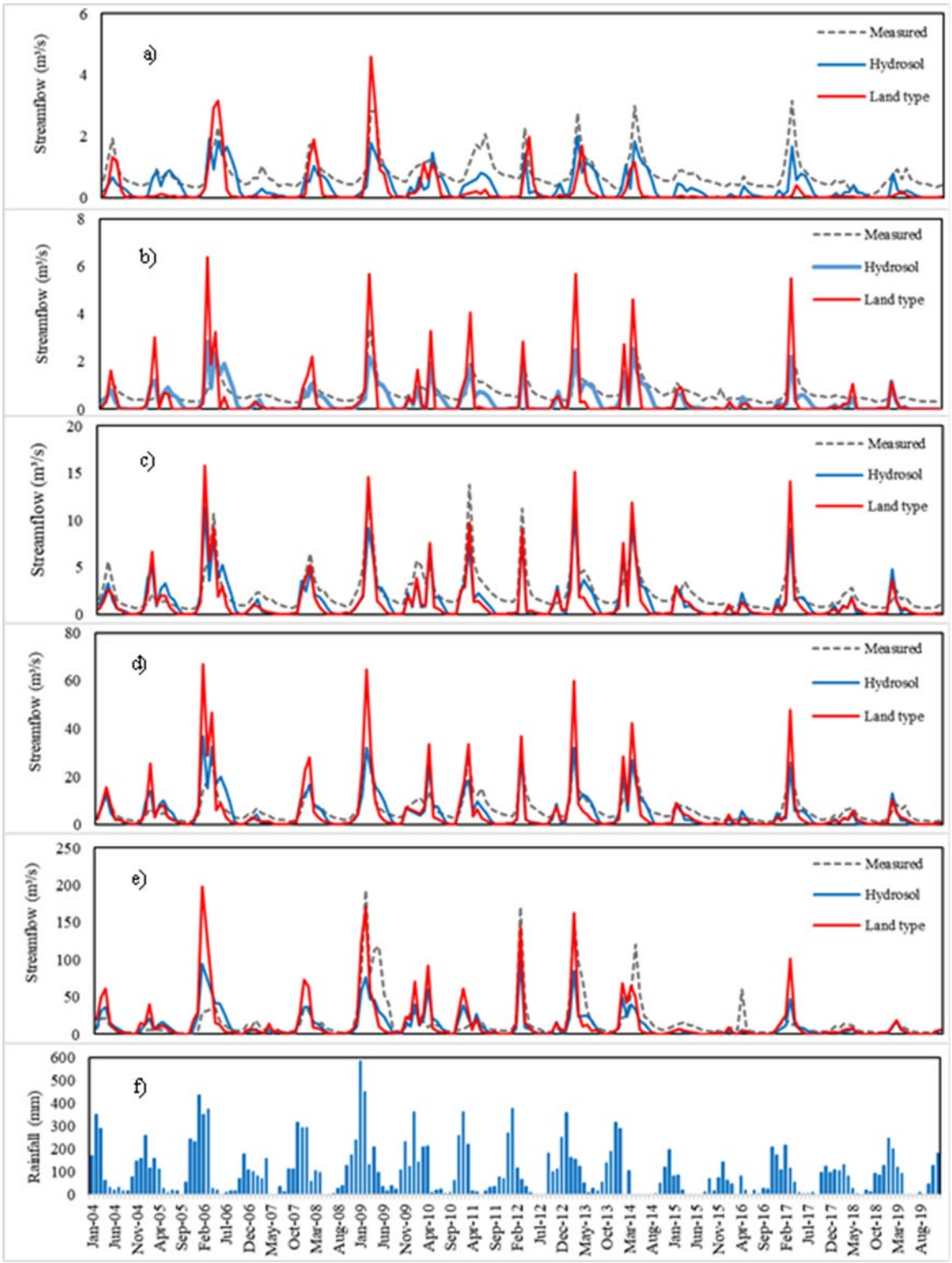

Figure 3.9: Monthly simulated streamflow for the Land Type and HYDROSOIL (Hydrosol) model runs compared to observed streamflow at (a) X3H003, (b) X3H002, (c) X3H001, (d) X3H024, (e) X3H021, together with (f) the average monthly rainfall during the validation period.

Hydrological processes

The differences in streamflow simulations and hydrological processes are a direct result of the differences between soil input datasets and how soil input data affects the simulation of these different hydrological processes.

The major hydrological processes differed substantially between model simulations (Table 3.11) at each of the five catchments. The HYDROSOIL simulations resulted in far lower average annual overland flow values than its Land Type counterpart (Table 3.11).

As surface runoff is directly impacted by the permeability of soils, land use and antecedent soil water conditions within each HRU (and land-use values remained constant between simulations), the difference in soil hydrological group and accompanying soil physical properties severely impact how surface runoff is simulated within the model (Neitsch et al., 2011; Zhang et al., 2019a). Recharge deep soils which are the dominant hydropedological soil within each of the five catchment scales contains the hydrologic soil group A designation, where low SCS curve numbers prohibit large overland flow values from being simulated, resulting in more infiltration within the soil profile.

Table 3.11: Average annual hydrological processes at each catchment scale.

Catchment	Soil data	Precipitation	Overland flow	Lateral flow	Perco	ET
		mm.year^{-1}				
X3H003 (48 km^2)	HYDROSOIL	1375	105	5	165	1006
	Land Type		232	8	126	949
X3H002 (56 km^2)	HYDROSOIL	1374	135	5	152	1034
	Land Type		323	24	63	964
X3H001 (174 km^2)	HYDROSOIL	1374	173	10	150	960
	Land Type		346	76	97	936
X3H024 (674 km^2)	HYDROSOIL	1285	136	7	138	958
	Land Type		296	37	74	932
X3H021 (2421 km^2)	HYDROSOIL	1109	106	4	99	842
	Land Type		239	17	43	813

Perco = Percolation; ET = Evapotranspiration

The HYDROSOIL dataset resulted in consistently lower lateral flow simulation at all five catchment scales (Table 3.11). The same factors that affect the soil runoff process affects lateral flow, where lower AWCs and shallower soil profiles of the Land Type dataset allows for more lateral flow to occur, because less water is needed to reach field capacity. These results are also in accordance with the hydrological soil types within the catchments. Lateral flow or interflow soils (A/B and soil/bedrock) are the least prevalent hydrological soil types within the Sabie River system, where X3H021 contains the only substantial amount of interflow soils at 25.5%.

Far higher percolation values were simulated by the HYDROSOIL model than the Land Type model at all five scales, decreasing as the catchment size increased from 165 mm per year at 48 km^2 to 99 mm per year at 2 421 km^2 (Table 3.11). This is presumably a result of differences in soil hydraulic properties but also in the decreased amount of precipitation within larger catchments. The SWAT model allows water to percolate if the soil water content exceeds field capacity for the specific soil layer and the underneath soil layer is still unsaturated and is therefore a function of the amount of soil water available

to percolate, the field capacity of soil layers as well as their saturated hydraulic conductivity. The variability of percolation is therefore largely affected by the spatial variability of various soil properties such as the depth of the soil profile, bulk density, saturated hydraulic conductivity and AWC of the soils, but also a product of SCS curve numbers where low curve numbers yield higher infiltration rates, allowing more water to enter the soil profile and potentially be available for percolation.

The HYDROSOIL dataset also simulated higher evapotranspiration compared to the Land Type dataset at all five catchment scales (Table 3.11). These results are comparable to other studies in the region such as Van Eekelen et al. (2015) with values of 1 143 mm for plantations, 1 087 mm for forest and woodlands and 690 mm per year for savanna and shrublands. Riddell et al. (2020) also found riparian savanna vegetation would record evapotranspiration values between 765 and 806 mm for one hydrological year within the region. Therefore, both models simulated reasonably accurate evapotranspiration values with high values where plantations and forests are the dominant land use, such as 48 km^2, 56 km^2 and 174 km^2, with decreasing values at the larger catchments which subsequently include more savanna and shrubland vegetation, such as 674 km^2 and 2 241 km^2. However, differences between evapotranspiration values are a direct result of differences in soil properties, where more water stored within the soil profile, due to deeper soils with large AWCs, results in more water being available for root uptake and evapotranspiration, as can be seen by the higher evapotranspiration values of the HYDROSOIL model compared to the Land Type model.

Figure 3.10 and Figure 3.11 illustrate the average annual surface runoff, lateral flow and percolation differences between each soil mapping unit between the two model simulations as well as the percentage spatial coverage of each mapping unit within each catchment. On average the HYDROSOIL soils simulated far lower average annual lateral flow, lower percolation rates and higher surface runoff values than their Land Type counterparts, except for recharge deep soils. Recharge deep soils are the dominant hydrological soil types within the HYDROSOIL map and are prevalent at all five catchment scales, the average annual surface runoff and lateral flow values at each catchment outlet therefore remained lower than the simulated values of the Land Type model.

Recharge deep soils contain the hydrologic soil group A designation, where low SCS curve numbers would prohibit large overland flow values to be simulated but rather result in higher infiltration rates. Recharge deep soils contained deeper soil profiles, with higher AWCs than the Land Type dataset, which means more water can infiltrate and be stored within the soil profile, without the profile reaching field capacity, affecting the simulation of different hydrological processes. These results are in accordance with other studies focusing on soil information in hydrological modelling (Wang & Melesse, 2006; Bouslihim et al., 2019).

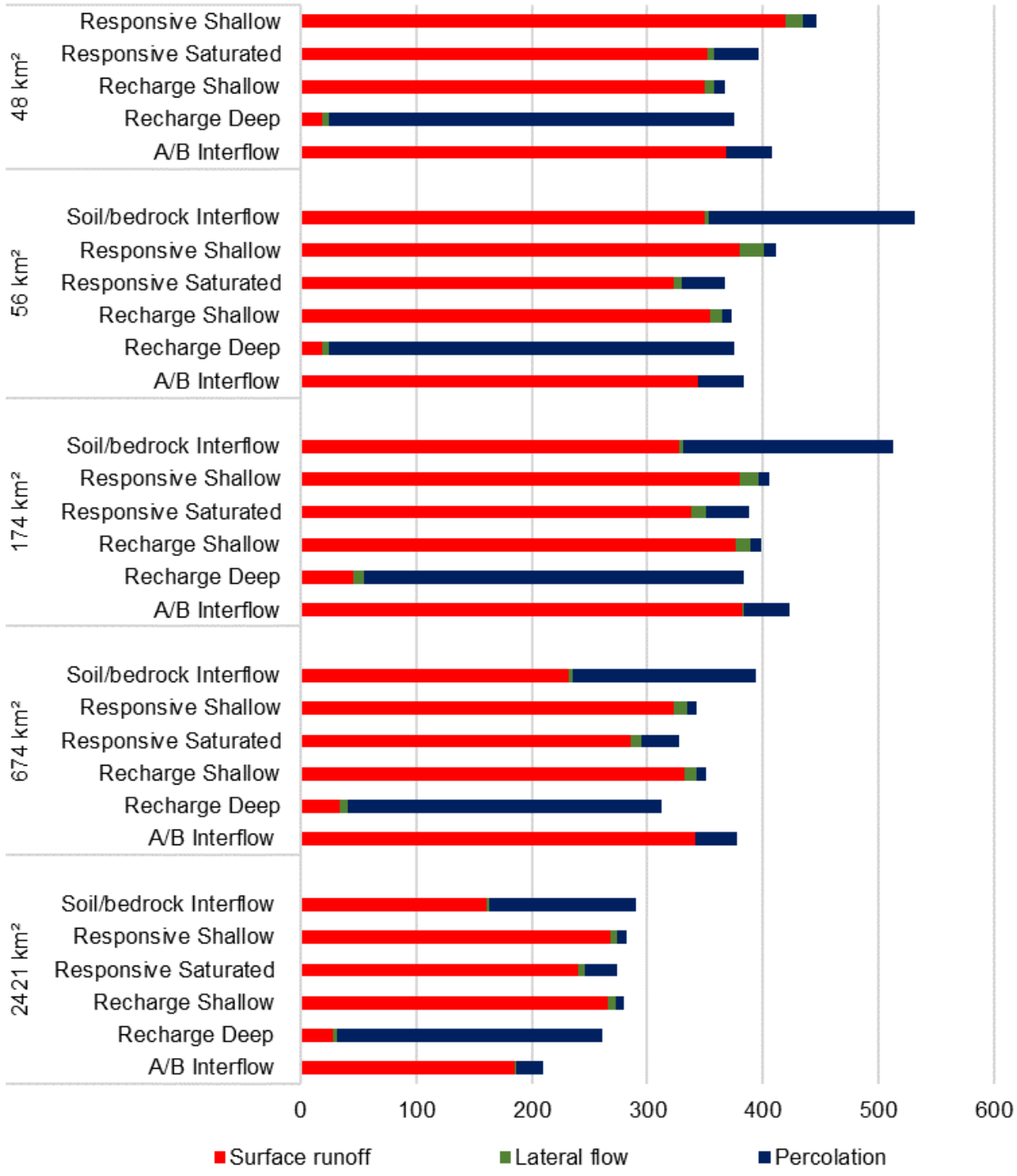

Figure 3.10: Average annual percolation, surface runoff and lateral flow values (mm) for the HYDROSOIL dataset as well as percentage of each mapping unit.

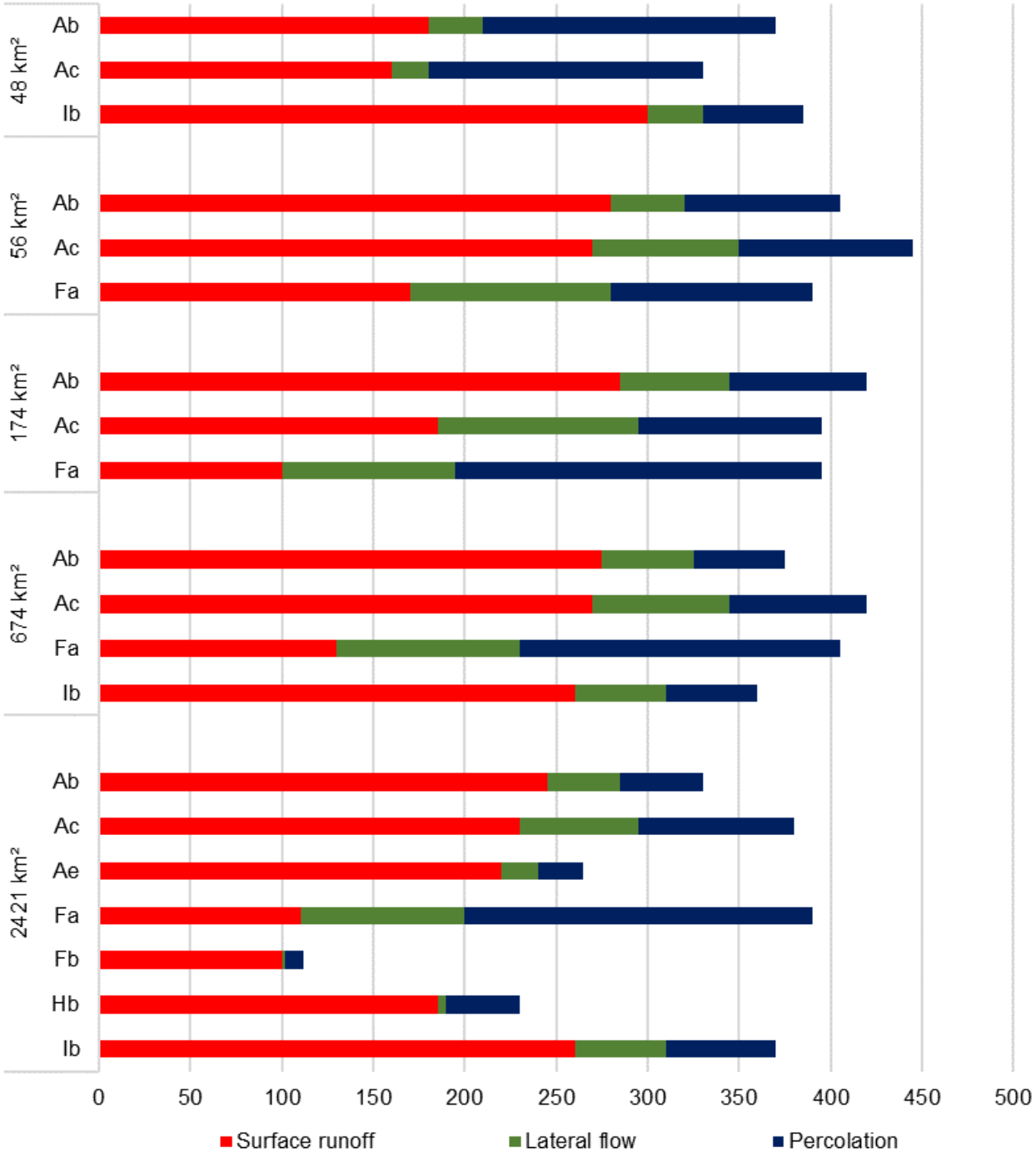

Figure 3.11: Average annual percolation, surface runoff and lateral flow values (mm) for the Land Type dataset as well as percentage of each mapping unit.

Differences between individual mapping units simulate different hydrological processes under the same hydrological conditions based on soil hydraulic properties (Figure 3.10; Figure 3.11). These results suggest that even though soils are mapped according to their hydropedological characteristics, these characteristics are not necessarily reflected within the modelling outputs. For example, due to the shallow depth and comparable soil hydraulic properties, both recharge shallow and responsive shallow soils simulate similar hydrological processes at each catchment scale. The same could be said for the A/B and soil/bedrock interflow soils which struggle to simulate large volumes of lateral flow compared to the other mapping units within the same catchments. These results suggest that additional calibration

of model parameters would be required to reflect the hydrological responses of different soils more adequately for different catchments. These results agree with Harrison et al. (2022), who required the calibrated lateral lag-time coefficient parameter within the SWAT+ model to improve the simulation lateral flow for each hydrological soil type within a mountainous research catchment in South Africa.

Differences in average annual lateral flow and percolation rates between mapping units under the same environmental conditions highlight the importance of soil hydraulic information. In particular these results suggest that Ksat and AWC values, which have been shown to be sensitive parameters within the model (Mengistu et al., 2019), severely affect how these two hydrological processes are simulated. Both are calculated when soil water exceeds the field capacity of the specific soil layer. However, higher Ksat and porosity values and steep slopes encourage water to drain laterally to the nearest stream channel, whereas lower Ksat and porosity values inhibit lateral flow to the channel and encourages the percolation of excess soil water to the underlying layer (Neitsch et al., 2011). The accurate representation of these hydraulic parameters will affect whether these processes are simulated accurately.

Large percolation values also correlate extremely well with the most dominant hydrological soil type across the Sabie River catchment, recharge soils (deep and shallow), as well as the large baseflow contributions seen within the measured streamflow data. The defining characteristic of these soils is the absence of any morphological indication of saturation. Vertical flow through and out of the soil profile into the underlying bedrock is the dominant flow direction. These soils also show no indication of permanent or periodic saturation within the soil profile, no indications of major runoff events at the soil surface and no indication of the lateral movement of water at the soil/bedrock or A/B interface (Van Tol & Le Roux, 2019). Hydropedologically speaking, 42.3% of the entire Sabie catchment should primarily be contributing recharge (percolation) to the shallow aquifer, with this value increasing in the mountainous catchments all the way up to 72.3% of the soils in 48 km^2.

The spatial disparity of average annual percolation values is evident (Figure 3.12), where the HYDROSOIL model simulated far higher percolation values than the Land Type model and at a far greater resolution.

The HYDROSOIL model primarily simulated high percolation values where recharge deep soils dominate in the mountainous sections of the catchment, where high precipitation and infiltration values also exist. Low percolation values were simulated on responsive shallow and recharge shallow soils as a result of their hydrologic soil group, position on the landscape and shallow soil profile. Percolation values in the east of the catchment show definite spatial variability along catenas with higher percolation rates associated with soil/bedrock interflow and recharge deep soils. The Land Type model did not show the same volume generation or spatial distribution of percolation across the catchment. Rather, percolation values were haphazardly spatially distributed as a function of the soil mapping units. These results are similar to Smit & Van Tol (2022), where large spatial and temporal differences were created between model simulations with differing soil input information. The average annual percolation values differed between the two levels of soil information within the Sabie catchment (Figure 3.13), where most of the catchment differed by average annual values greater than 50 mm, especially in the headwaters of the catchment (48 km^2, 56 km^2, 174 km^2, 674 km^2), becoming less pronounced in the drier eastern savanna segments of the catchment. Differences in soil input information also translates to differences in hydrological process simulations in hydrological models (Figure 3.13).

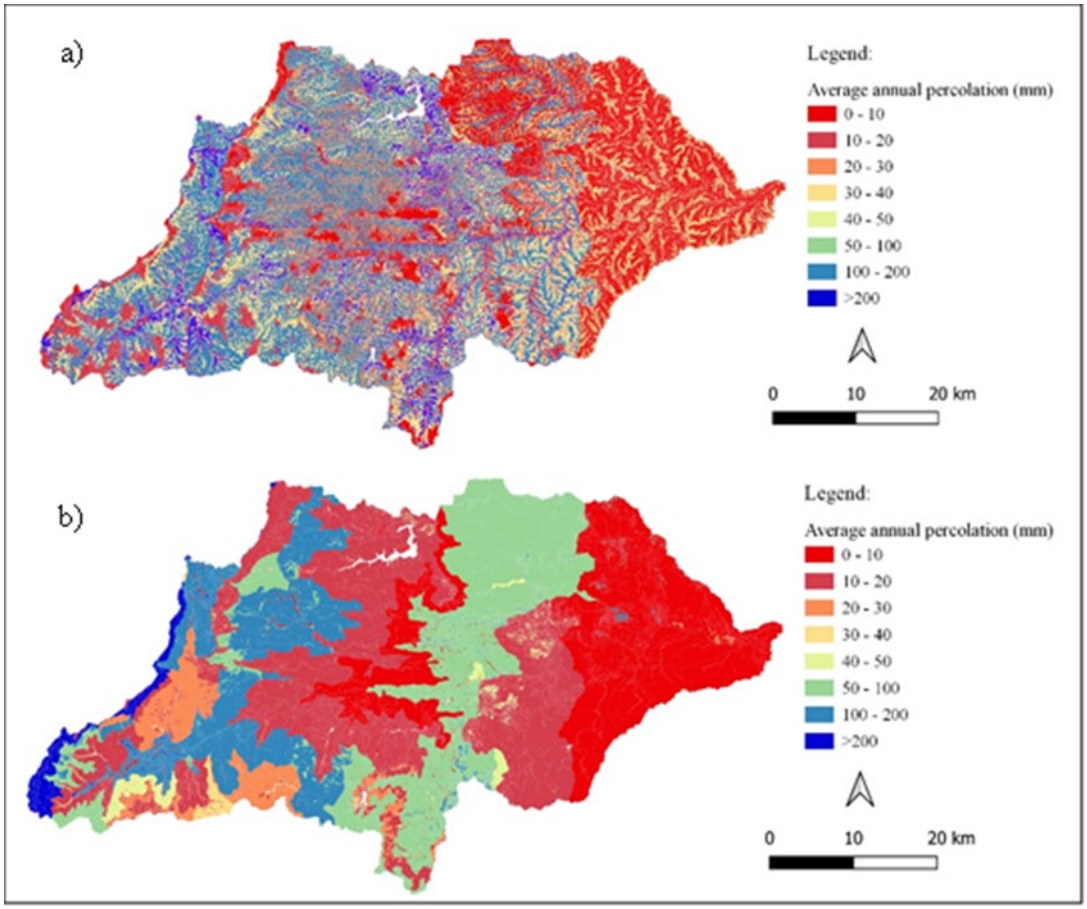

Figure 3.12: Average annual percolation values (mm) for the (a) HYDROSOIL and (b) Land Type dataset at the HRU level.

Our assumption remains that detailed hydropedological information, based on modern techniques for digital soil mapping and in-field measured soil physical properties represent a more accurate representation of real-world percolation rates within the Sabie catchment. The ability of the Land Type model to therefore simulate any form of land-use change or climate change scenario should be called into question as it is clear the internal hydrological process simulation, determined by the soil input data, is left wanting (Van Tol et al., 2021a; Smit & Van Tol, 2022). The argument remains that hydropedological information may serve as an effective 'soft data' tool to better represent internal hydrological processes within a catchment, leading to improved catchment management practices (Seibert & McDonnell, 2002; Smit & Van Tol, 2022), however further calibration is required to achieve this goal.

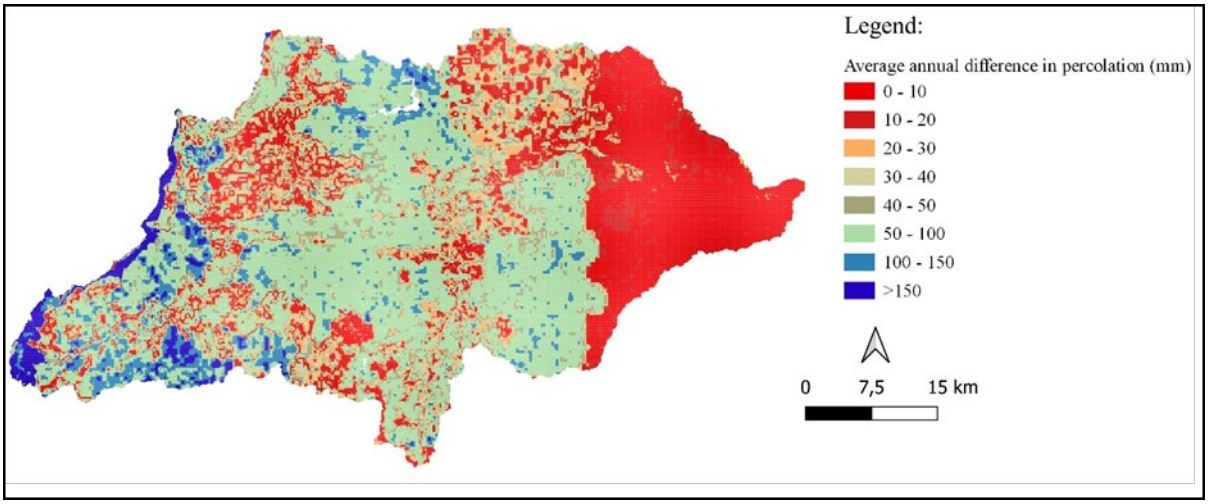

Figure 3.13: Gridded (100 m x 100 m) average annual percolation difference (mm) between the two levels of soil information.

The results of this study agree with other research which emphasises the importance of understanding the hydropedological information available within a catchment and its transferability for hydrological modelling purposes (Bouma et al., 2011; Sierra et al., 2018; Van Tol et al., 2020; Van Tol et al., 2021a). Soil information plays a crucial role in refining model predictions and should be used in supporting informed decision-making in hydrological modelling and water resource management (Bouslihim et al., 2019). It would be worth exploring if a multigauge calibration using the range of in-field measured soil properties can continue to improve modelling accuracy, especially at large scales where improved soil information diminishes in value. In terms of water resource management implications, this study does suggest that if large-scale applications of water quantity simulations are the primary objective, then the impact of hydropedological information is negligible, especially when comparing the modelling accuracy between the two levels of soil information at 2 421 km^2. However, detailed soil information improves the hydrological process representation and modelling accuracy at smaller scales. Modern water resource management plans are, however, concerned with impacts at the local sub-catchment level, where the improved detail and accuracy of hydropedological information is more applicable than coarse soil information. The value of hydropedological information should also be further investigated for use in ungauged basins as a means of improving modelling accuracy where long-term measurements are absent.

3.2.5 Conclusions

Detailed hydrological soil information, developed using digital soil mapping techniques, resulted in more accurate streamflow simulations at four of the five scales. The improved simulation accuracy at these scales was obtained without a calibration period, but rather by more accurately representing the internal hydrological processes of the catchment, based on hydropedological insight. This is especially promising for hydrological modelling in ungauged catchments, where hydropedology could form an important ‘soft data’ tool to aid modelling efforts where reliable streamflow measurements are absent.

The value of improved soil information decreases as the catchment size increases when analysing mean monthly streamflow simulations, which agrees with similar research findings globally. Future research should focus on determining the ideal level of soil information for hydrological modelling for different sized micro-, meso- and macro-scale catchments and focus on calibrating hydrological modelling using a range of in-field measured soil input parameters.

3.3 MODAL CALIBRATION USING HYDROPEDOLOGICAL INSIGHTS TO IMPROVE INTERNAL HYDROLOGICAL PROCESSES WITHIN SWAT+

3.3.1 Abstract

Soils affect hydrological processes by partitioning precipitation into different components of the water balance. Therefore, understanding soil-water dynamics at a catchment scale is imperative to future water resource management. This study investigates the value of hydropedological insights to calibrate a process-based model. Soil morphology was used as 'soft data' to assist in the calibration of the SWAT+ model at five different catchment sizes (48 km^2, 56 km^2, 174 km^2, 674 km^2 and 2 421 km^2) in the Sabie River catchment, South Africa. The aim is to calibrate the SWAT+ model to accurately simulate long-term monthly streamflow predictions, as well as to reflect internal soil hydrological processes, using hydropedology as a calibration tool in a multigauge system. Results indicated that calibration improved streamflow predictions where R^2 and Nash-Sutcliffe Efficiency (NSE) improved substantially, R^2 improved by 2 to 8% and NSE from negative correlations to values exceeding 0.5 at four of the five catchment scales compared to the uncalibrated model. Results confirm that soil mapping units can be calibrated individually within SWAT+ to improve the representation of hydrological processes. Particularly, the spatial linkage between hydropedology and hydrological processes, which is captured within the soil map of the catchment, can be adequately reflected within the model structure after calibration. This research should lead to an improved understanding of hydropedology as 'soft data' to improve hydrological modelling accuracy.

3.3.2 Introduction

Soils play a pivotal role in shaping hydrological processes within a landscape as they actively partition precipitation into various components of the water balance. This functionality stems from the soil's capacity to absorb, store and transmit water across diverse spatial and temporal scales (Park et al., 2001). These hydrological processes largely determine the volume, variability and residence times of water resources within a landscape, which in turn determines the agricultural potential, functionality of ecosystems and economic opportunities within different catchments (Wenninger et al., 2008). However, the logistical impracticality of measuring these hydrological processes at landscape scale means that these processes remain most practically quantified using hydrological models, which simplify and represent real-world hydrological systems (Gassman et al., 2007; Devia et al., 2015).

Soil and water are inextricably linked (Bouma et al., 2011). Soils provide valuable ecosystem services, such as food production, carbon and nutrient cycling, flood mitigation, and water filtration and purification (Lal et al., 2021). These services are intimately linked to the spatio-temporal variation of hydrological flowpaths, such as surface runoff, infiltration, lateral flow, evaporation and percolation. Water, on the other hand, plays a fundamental role in soil formation and results in soil properties such as soil colour, the formation and distribution of mottles, as well as soil texture and structure (Lal et al., 2021). These identifiable soil properties are a direct product of the dominant hydrological processes present during their formation and may be linked to different hydrological processes based on selected soil properties. The dynamic interplay between soil and water forms the cornerstone of the interdisciplinary field known as hydropedology (Lin, 2003). This field has proven instrumental in conceptualising diverse hydrological processes, particularly in regions lacking or having limited hydrometric measurements (Gassman et al., 2007; Devia et al., 2015).

However, one of the primary issues with modern hydrological models is their reliance on a calibration period, where results are primarily focused on accurately simulating a specific point observation, such as streamflow gauges (Beven & Freer, 2001). Modern hydrological models contain a level of complexity within their structure, which allows a high level of parameterisation and adaptability. This phenomenon

frequently leads to a situation where various model configurations produce comparable outputs, a concept known as equifinality (Beven & Freer, 2001; Beven, 2006). While these models and their associated approaches may offer statistically accurate simulations concerning point observations, there is a lingering question about their ability to truly capture the pertinent internal hydrological processes or to accurately simulate scenarios related to land-use or climate change beyond the environmental conditions for which they were initially calibrated (Kirchner, 2006).

While researchers concur on the importance of adequately incorporating internal catchment processes into the model structure and parameters, even at the cost of sacrificing some modelling accuracy (Arnold et al., 2015; Yen et al., 2014), the available approaches to enhance the representation of internal catchment processes remain somewhat constrained. To address this challenge and enhance the accuracy of internal hydrological processes, the incorporation of 'soft data' has been suggested. 'Soft data' is defined as information that may not be measured directly but can be linked to hydrological processes or phenomena (Beven, 2006).

Hydropedology combined with digital soil mapping has provided an intriguing source of 'soft data' for hydrological modelling. It allows the spatial capture of different soil hydrological processes observable at the pedon level and enables the accurate extrapolation of these processes to hillslope and catchment level (Lin, 2003). Therefore, enabling the capture and transfer of information related to different hydrological processes within different soil mapping units for hydrological modelling purposes. Several researchers have assessed hydropedological insights as input to process-based hydrological modelling, illustrating the improved accuracy that hydrological soil information achieves (Van Tol et al., 2020; Smit & Van Tol, 2022; Smit et al., 2023a). For instance, hydropedological characteristics improved modelling accuracy by more accurately reflecting the lateral flow dynamics within different afromontane catchments in South Africa (Harrison et al., 2022).

Although soils can be grouped according to different hydrological processes, the improved modelling performance achieved by hydrological soil information is primarily based on the improved representation of measured hydraulic properties and not the improved representation of hydrological processes (Smit et al.,2023a; Van Tol & Van Zijl, 2022).

In this paper, we aim to reflect internal hydrological processes more accurately by applying a calibration approach focussing on hydropedological insights as 'soft data'. The study area remained the same as Section 3.2. The aim was achieved by deriving the dominant hydrological responses of various soil types based on the soil morphology and then applying a calibration approach. Select parameters were calibrated to reflect an accurate prediction of long-term measured streamflow as well as dominant soil hydrological process of each hydrological soil type. This approach was evaluated by statistical comparison with measured stream flow and visual interpretation of water balance components.

3.3.3 Materials and methods

Hydropedological approach to calibration

The SWAT+ model requires a calibration period which allows the most sensitive parameters to be adjusted to improve hydrological modelling accuracy. Practical calibration guidelines have been well established and can be found in the SWAT manual (Arnold et al., 2011) as well as in a multitude of research papers, where calibration is either automated or conducted manually by selecting the most sensitive parameters and adjusting them accordingly (Abbaspour et al., 2007; Arnold et al., 2011; Moriasi et al., 2007; Ahl et al., 2008; Tuppad et al., 2011; Mengistu et al., 2019).

Different parameters have different levels of sensitivity within the model, which allows modellers the opportunity to calibrate different hydrological processes such as surface runoff, lateral flow, return flow, and evapotranspiration rates (Mengistu et al., 2019). The most sensitive parameters for calibration within the Sabie catchment were determined using the Latin Hypercube Sampling approach with 2 000 iterations within *R-SWAT* (Nguyen et al., 2022) with Kling-Gupta Efficiency as objective function.

Researchers have shown the SCS curve number II (CN2) as the most sensitive parameter in SWAT streamflow simulations (Eckhardt, 2005; Shen et al., 2012; Mengistu et al., 2019) as it reflects the characteristics of the catchment prior to a rainfall event and largely determines surface runoff. It is dependent on the initial CN value assigned to the HRU by the model. Surface runoff is calculated using the Equations 3.5 and 3.6 (Section 3.2.3).

Lateral flow is calculated by SWAT using a kinematic storage model, which simulates the movement of water in a two-dimensional cross-section of a hillslope (Neitsch et al., 2011). Lateral flow therefore occurs when soil water exceeds field capacity with the underlying layer being impermeable or semi-permeable. The kinematic approximation method assumes that the flowpaths are parallel to the bedrock and that the hydraulic gradient equals the slope of the hill (Equation 3.8).

$$SW_{excess} = \frac{1000.H_0.\Theta_d.L_{hill}}{2} \quad (3.8)$$

Where SW_{excess} equals the drainable water volume within the saturated zone of the soil per unit area (mm), H_0 equals the saturated thickness of the hillslope outlet as a fraction of the total thickness (mm/mm), Θ_d equals the drainable porosity of the soil (mm/mm), and L_{hill} equals the length of the hillslope (m) (Neitsch et al., 2011). The drainage porosity of the soil equals the total porosity of the soil minus the soil porosity when the soil horizon is at field capacity.

$$Q_{lat} = 24.H_0.K_{sat}.slp \quad (3.9)$$

Where Q_{lat} is the water discharge from the hillslope outlet, H_0 equals the saturated thickness of the hillslope outlet as a fraction of the total thickness (mm/mm), K_{sat} is the saturated hydraulic conductivity for the specific soil layer and slp is the slope value as the increase in elevation per distance unit for the specific hillslope.

SWAT+ calculates percolation using a storage routing methodology, where water percolates if the soil water content exceeds the field capacity of the specific soil layer and the layer below is not yet saturated. The equation used to calculate the amount of percolation that occurs is:

$$w_{perc} = SW_{excess}.\left(1 - exp\left[\frac{-\triangle t}{TT_{perc}}\right]\right) \quad (3.10)$$

Where w_{perc} is the amount of water percolating from the specific soil layer, SW_{excess} is the drainable soil water available, $\triangle t$ is the time step (hours) and TT_{perc} is the travel time for percolation (hours). The travel time for percolation is defined in the model using:

$$TT_{perc} = \frac{SAT - FC}{K_{sat}} \quad (3.11)$$

Where TT_{perc} is the travel time for percolation (hours), SAT is the amount of water within the specific soil layer when completely saturated, FC is the soil water content at field capacity and K_{sat} is the saturated hydraulic conductivity for the specific soil layer.

Therefore, K_{sat} largely determines if lateral flow or percolation is simulated by the model, where the lower the K_{sat} value within the soil layer, the lower the lateral flow value and higher the percolation ratio simulated.

Manipulating and calibrating major hydrological processes such as surface runoff, lateral flow and percolation remains imperative to accurately modelling water resources throughout a landscape (Brouziyne et al., 2017; Wagner et al., 2022). The uncalibrated SWAT+ model for the Sabie-Sand catchment (Smit et al., 2023a) illustrated that the hydrological soil types within the catchment did not necessarily accurately reflect the dominant hydrological process associated with each soil mapping unit, such as interflow soils not simulating sufficient lateral flow volumes and recharge shallow soils primarily contributing surface runoff volumes, with limited percolation contributions (Figure 3.10).

SWAT+ incorporates parameters that allow the calibration of these processes, where PERCO and LATQ_CO parameters are linear coefficients that can be applied to the hillslope storage equation to limit lateral flow and percolation values (Wagner et al., 2022). Therefore, these parameters could potentially be manually calibrated to link soil morphology to dominant hydrological processes and accurately reflect these processes for each mapping unit as the current soil hydraulic properties fail to correctly simulate these processes.

Practically, calibration was performed through the calibration of the most sensitive parameters identified by the sensitivity analyses and literature, performing several final manual iterations to fine-tune results. The calibration procedure needed to accomplish two goals. Firstly, the model needed to be calibrated to reflect accurate streamflow values by more accurately representing baseflow and peak flow volumes and secondly, each hydrological soil type needed to reflect their dominant hydrological response identified by their inherent hydrological soil mapping unit. Therefore, CN2, PERCO and LATQ_CO were calibrated for each mapping unit to ensure lateral flow dominates in interflow soils (A/B and soil/bedrock), percolation dominates in recharge soils (deep and shallow), and surface runoff dominates in responsive soils (saturated and shallow). Hydropedology is therefore applied as a source of ‘soft data’ informing the calibration procedure to better represent hydrological processes spatially.

Validation

Five weirs, which are managed by the DWS, were used to validate long-term streamflow simulations. These gauges, from smallest drainage area to largest were X3H003 which drains 48 km^2, X3H002 which drains 56 km^2, X3H001 which drains 174 km^2, X3H024 which drains 674 km^2 and X3H021 which drains the entire study area at 2 421 km^2. Monthly streamflow averages were used for statistical analysis.

Four commonly employed statistical indicators – coefficient of determination (R^2), percentage bias (PBIAS), Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) – were used for statistical comparison. Percentage bias (PBIAS) specifically evaluates the average tendency of simulated data to either exceed or fall short of their observed counterparts.

3.3.4 Results and discussion

Sensitivity analyses

Table 3.12 illustrates the most commonly used parameters for calibration, their description and their relative sensitivity within the SWAT+ model for the Sabie River catchment. The t-stat and p-value are two statistical measurements which assess the sensitive rank of each parameter. The t-stat represents

a range of sensitivity, while the p-value identifies the significance of sensitivity. The higher absolute value of t-stat and lower p-value (< 0.05) indicates a sensitive parameter.

Table 3.12: The most commonly used parameters for calibration, their description and their relative sensitivity within the SWAT+ model for the Sabie River catchment.

Parameter	Description	t-stat	p-value
CN2	Initial SCS runoff curve number for moisture condition II	-5.47	0.00
ALPHA_BF	Baseflow alpha factor (l/days)	-5.16	0.00
LATQ_CO	Lateral flow coefficient	-5.26	0.00
CH_K2	Effective hydraulic conductivity in main channel alluvium (mm/h)	1.48	0.14
SOL_AWC	Available water capacity of the soil layer (mm H_2O /mm soil)	1.46	0.15
SOL_K	Soil hydraulic conductivity of the soil layer (mm H_2O /hour)	-1.45	0.15
SURLAG	Surface runoff lag coefficient	-1.03	0.30
EPCO	Plant uptake compensation factor	0.97	0.34
PERCO	Percolation coefficient	0.63	0.43
RCHRG_DP	Aquifer percolation coefficient for water to percolate from the shallow to the deep aquifer.	-0.34	0.64
ESCO	Soil evaporation compensation factor	-0.34	0.74
REVAP	Threshold depth of water in the shallow aquifer for 'revap' to occur (mm H_2O)	0.65	0.83
ESCO	Soil evaporation compensation factor	-0.34	0.84
REVAP	Threshold depth of water in the shallow aquifer for 'revap' to occur (mm H_2O)	0.65	0.9

Firstly, parameters SURLAG, ESCO, EPCO, ALPHA_BF and RCHRG_DP were calibrated to better reflect the peak flow and baseflow characteristics of each of the five catchments, specifically by lower evapotranspiration values, and increasing peak flow and baseflow values (Table 3.13). This study altered the CN2, PERCO and LATQ_CO parameters to calibrate each hydrological soil type (Table 3.14).

In general, CN values were increased for all soil mapping units to improve surface runoff dynamics in the catchments and improve peak flow estimations. The soil hydrologic group for recharge shallow soils were altered from group D, as suggested by Neitsch et al. (2011), for shallow soils, to group A which would drastically lower the associated soil curve numbers, facilitating increased infiltration rates and therefore decreasing surface runoff values. The PERCO parameter was also calibrated accordingly, where values were kept at default for recharge deep and recharge shallow soils but decreased for interflow soils (A/B and soil/bedrock), which would inhibit percolation and increase lateral flow for these soils. PERCO was also slightly decreased for responsive soils (saturated and shallow) to slightly inhibit percolation, resulting in higher soil water contents and potentially more surface runoff.

Table 3.13: Calibrated model parameters, the methods of change used and the final calibrated values.

Parameter	Catchment	Method of change	Min value	Max value	Fitted value
CN2	All	Relative	-15	15	6
SURLAG	X3H021, X3H024	Replace	0	20	10
ESCO	All	Replace	0	1	0.98
EPCO	All	Replace	0	1	0.45
ALPHA_BF	X3H003	Replace	0.005	0.48	0.004
	X3H002	Replace	0.005	0.48	0.0015
	X3H001	Replace	0.005	0.48	0.005
	X3H024, X3H021	Replace	0.005	0.48	0.04
RCHRG_DP	X3H003, X3H002, X3H001	Replace	0.001	0.05	0.01
	X3H021, X3H024	Replace	0.001	0.05	0.03
CH_K	All	Replace	0.5	150	34

The LATQ_CO parameter was decreased for recharge (deep and shallow) to potentially limit lateral flow volumes, meaning more water is available to be stored within the soil profile, leading to more antecedent moisture content and potentially increased percolation values. LATQ_CO was not adjusted for interflow soils (A/B and soil/bedrock) or responsive saturated soils as these soils are major contributors to lateral flow. The adjustments of selected SWAT+ parameters should improve the representation of internal hydrological processes by linking them to soil morphology.

Table 3.14: Manually calibrated parameters applied to improve the representation of soil hydrological processes.

Hydrological soil type	Default soil hydrologic group	Calibrated soil hydrologic group	Parameter	Default value	Calibrated value
A/B interflow	C	C	CN	70-85	76-89
			PERCO	1	0.5
			LATQ_CO	1	1
Recharge deep	A	A	CN	32-67	35-70
			PERCO	1	1
			LATQ_CO	1	0.8
Recharge shallow	D	A	CN	79-89	35-70
			PERCO	1	1
			LATQ_CO	1	0.8
Responsive saturated	C	C	CN	70-85	75-88
			PERCO	1	0.5
			LATQ_CO	1	1
Responsive shallow	D	D	CN	79-89	82-91
			PERCO	1	0.5
			LATQ_CO	1	1
Soil/bedrock interflow	A	A	CN	32-67	35-70
			PERCO	1	0.5
			LATQ_CO	1	1

Streamflow predictions

The SWAT+ model had the exact same number of sub-basins (119) and landscape units (616) prior to calibration, during calibration as well as during validation as used in Section 3.2 (Smit et al., 2023a). A KGE value surpassing -0.41 indicates a model prediction that aligns better with the mean observed values (Knoben et al., 2019). Refined evaluation criteria for hydrologic and water-quality models deem streamflow simulations satisfactory when $R^2 > 0.6$, NSE > 0.5, and PBIAS ≤ 15% (Moriasi et al., 2015).

The simulations of the HYDROSOIL dataset during and after calibration and validation yielded satisfactory results at four of the five catchments, namely, 48 km^2, 56 km^2, 174 km^2, and 674 km^2 (Table 3.15). These results are an improvement on the uncalibrated model which achieved satisfactory R^2 at only three of the five catchments, namely, 48 km^2, 174 km^2, and 674 km^2 (Smit et al., 2023a).

Table 3.15: Statistical indicators of streamflow prediction accuracy during calibration and validation for all five catchment scales.

Catchment	Model period	R^2	PBIAS	NSE	KGE
X3H003 (48 km^2)	Uncalibrated	0.66	53.92	0.03	0.41
	Calibration	0.70	-9.91	0.80	0.73
	Validation	0.79	2.32	0.68	0.86
X3H002 (56 km^2)	Uncalibrated	0.57	43.67	-0.22	0.41
	Calibration	0.63	-4.52	0.84	0.74
	Validation	0.62	15.34	0.67	0.65
X3H001 (174 km^2)	Uncalibrated	0.67	37.27	0.48	0.58
	Calibration	0.72	-25.45	0.36	0.53
	Validation	0.79	-20.00	0.61	0.60
X3H024 (674 km^2)	Uncalibrated	0.71	20.85	0.54	0.67
	Calibration	0.79	-21.51	0.56	0.58
	Validation	0.83	-15.87	0.76	0.74
X3H021 (2421 km^2)	Uncalibrated	0.54	33.5	0.49	0.42
	Calibration	0.42	-27.54	0.42	0.51
	Validation	0.48	-25.25	0.34	0.64

PBIAS = Percentage bias; NSE = Nash Sutcliffe Efficiency; KGE = Kling-Gupta Efficiency

Both during calibration and validation disappointing PBIAS values were achieved, where no catchment greater than 250 km^2 achieved PBIAS values below the 15% threshold (Table 3.15; Moriasi et al. 2015). However, PBIAS values were decreased by between 6-8% during calibration and validation compared to the uncalibrated model (Smit et al., 2023a). In general, the negative PBIAS values during both the calibration and validation signifies a general overestimation of total streamflow within the Sabie River catchment, which may be attributed to the overestimation of peak flows within the catchments which could potentially be attributed to rainfall uncertainty within the Sabie River system.

When analysing NSE values during the calibration and validation periods, only the calibration period at 174 km^2 and calibration and validation periods at 2 421 km^2 fell below the accepted 0.5 threshold. The best NSE values were obtained during calibration of both 48 km^2 and 56 km^2. The best NSE values during the validation period were observed at 48 km^2 and 674 km^2. Both calibration and validation NSE

values significantly improved from the uncalibrated hydrological model at all five scales, especially at the 48 km^2, 56 km^2 and 174 km^2 catchment sizes.

All KGE values met the satisfactory threshold values of -0.41 (Knoben et al., 2019). The best calibration KGE values were obtained at 48 km^2 and 56 km^2, whereas the best validation KGE values were observed at 48 km^2 and 674 km^2. Both calibration and validation KGE values significantly improved from the uncalibrated hydrological model at all five scales, especially at 48 km^3, 56 km^2, 174 km^2, which indicates the value of calibrated representative hydrological models at smaller scales, which is a notion supported by Smit et al. (2023a) and Van Tol et al. (2020).

Peak flows simulations were improved by increasing the SURLAG parameter from 2 421 km^2 and 674 km^2, which lags surface runoff at the two largest catchments to be more realistic of real-world conditions. This is due to the fact that as catchments increase in size, the surface runoff lag time becomes substantial. However, peak flows were slightly overestimated at the four smallest catchment sizes (48 km^2, 56 km^2, 174 km^2, 674 km^2), with the largest overestimation of peak flow values occurring at catchment 2 421 km^2 (Figure 3.14).

Peak flows were underestimated at 2 421 km^2, which is the largest of the five catchment scales. What is readily observable is that baseflow simulations were greatly improved during calibration, where the uncalibrated model substantially underestimated baseflow contributions, particularly at smaller catchment sizes (48 km^2, 56 km^2 and 174 km^2) where considerable baseflow contributions exist (Figure 3.14). This is primarily a result of calibrating the selected groundwater parameters, such as ALPHA_BF, and RCHRG_DP, where both parameters were substantially decreased to improve the representation of baseflow values within the five catchments. Baseflow underestimation is still, however, prevalent at 48 km^2 and 56 km^2, which suggests that additional contributions from the deep aquifer potentially augment baseflow volumes provided by the shallow aquifer and are in accordance with other studies in the region (Saravia Okello et al., 2018) as well as the broader region of South Africa (Van Tol et al., 2020).

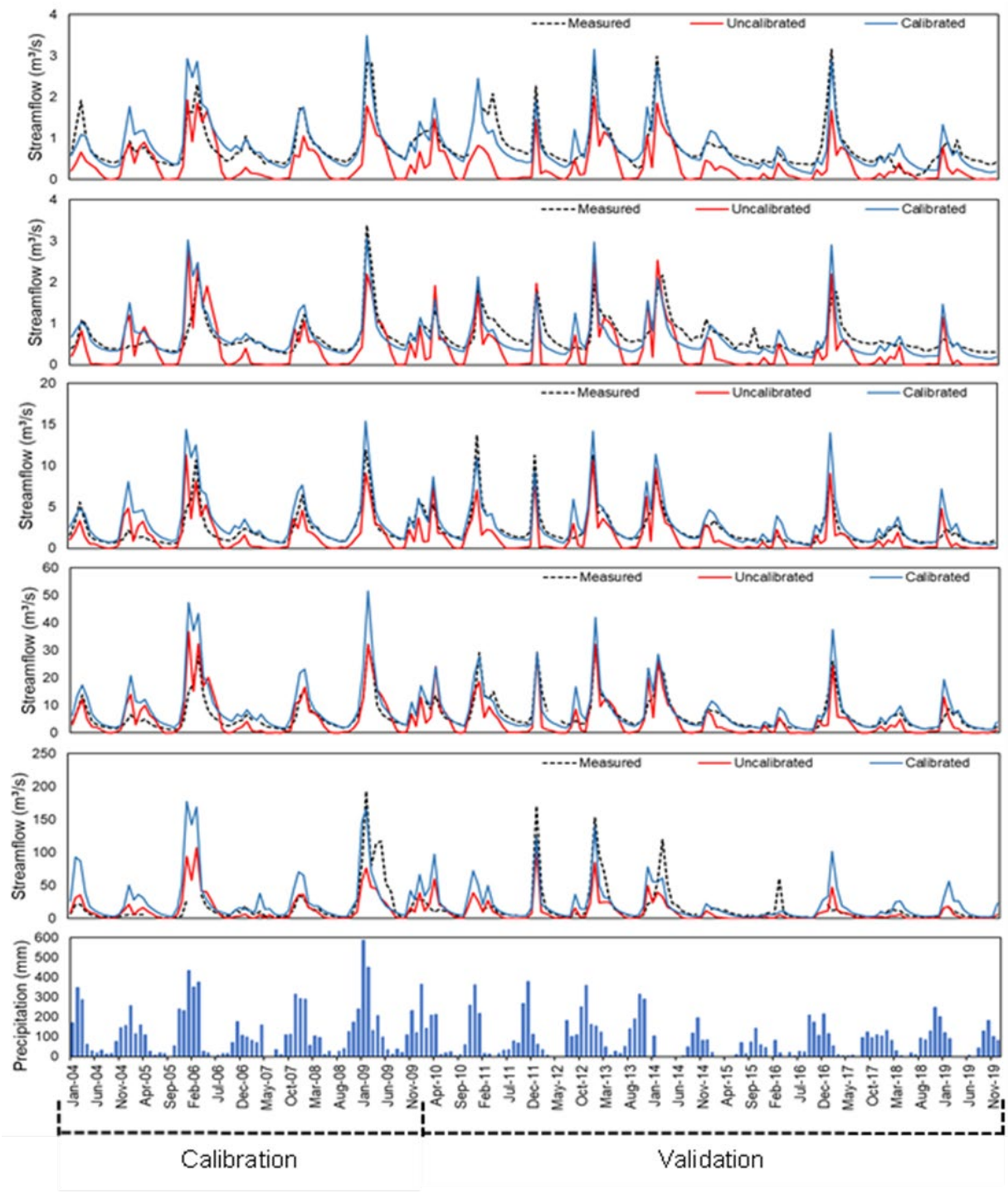

Figure 3.14: Monthly simulated streamflow for the HYDROSOIL model runs compared to observed streamflow at (a) X3H003, (b) X3H002, (c) X3H001, (d) X3H024, (e) X3H021 together with (f) the average monthly rainfall during the validation period.

Hydrological processes

The differences in streamflow simulations are a direct product of the calibration of selected parameters, which affected how the major hydrological processes were simulated. The major hydrological processes differed between the uncalibrated and calibrated model simulations (Table 3.16) at each of the five catchment scales.

In general, the calibrated model simulated far higher overland flow compared to the uncalibrated model (Table 3.16). The same was also true for average annual lateral flow values, where the calibrated model simulated substantially higher values at each of the five catchments (Table 3.16). However, calibrated average annual percolation values only slightly increased at 48 km^2 and 174 km^2. Whereas calibrated percolation rates slightly decreased at 56 km^2, 674 km^2 and 2 241 km^2. The calibrated model also simulated less evapotranspiration compared to the uncalibrated model.

Table 3.16: Average annual hydrological processes at each catchment scale.

Catchment	Model run	Precipitation	Surface runoff	Lateral flow	Percolation	ET
		mm.year^{-1}				
X3H003 (48 km^2)	Uncalibrated	1375	105	5	165	1006
	Calibrated		270	77	184	984
X3H002 (56 km^2)	Uncalibrated	1374	135	5	152	1034
	Calibrated		214	93	145	955
X3H001 (174 km^2)	Uncalibrated	1374	173	10	150	960
	Calibrated		230	130	167	944
X3H024 (674 km^2)	Uncalibrated	1285	136	7	138	958
	Calibrated		209	121	133	932
X3H021 (2421 km^2)	Uncalibrated	1109	106	4	99	842
	Calibrated		170	75	96	802

ET = Evapotranspiration

As surface runoff is directly impacted by the permeability of soils, associated land use and antecedent soil water conditions within each HRU (where land-use values remained constant between simulations), the difference in selected calibrated parameters, especially CN2 values, affected how surface runoff was simulated within the two models (Neitsch et al., 2011; Zhang et al., 2019a). On average, the CN2 value of each hydrological soil group increased after calibration, which allowed more surface runoff to be simulated.

However, ESCO also increased for the entire basin, less water was allowed to be removed by evaporation from lower levels in the soil, allowing more water to be available to either be stored within the soil, percolate, or flow laterally. EPCO was also decreased for the basin during calibration, which allowed less variation of the original depth distribution from which plants could meet their transpiration demands and therefore decreased evapotranspiration values for the entire basin, further increasing the amount of water available to either be stored within the soil profile, flow laterally out of the soil profile or percolate to the underlying soil or shallow aquifer (Neitsch et al., 2011). These changes within the model resulted in lower evapotranspiration rates after calibration which are still in accordance with other studies in the region where evapotranspiration rates of between 690 mm to 1 143 mm were recorded, depending on vegetation types, where savanna and shrubland vegetation types resulted in lower evapotranspiration rates than plantations and indigenous forests (Van Eekelen et al., 2015; Riddell et al., 2020). As 48.5% of the catchment consists of savanna and only 40% consists of plantations and forest (Mucina & Rutherford, 2006), these evapotranspiration values are reasonable.

A substantial amount of additional water was therefore made available for different hydrological processes within the Sabie River system which were partitioned at their appropriate hydrological response at the soil level using the above-mentioned calibration approach (Table 3.17; Figure 3.15).

After calibration, recharge deep soils illustrated increases in both surface runoff and lateral flow, however, the dominant hydrological flow path remains recharge to the groundwater with recharge deep soils contributing 180 mm of recharge annually (Table 3.17).

Table 3.17: Average annual surface runoff, lateral flow and percolation and change (mm) for each hydrological soil type between the calibrated and uncalibrated SWAT+ models (2 421 km^2).

Soil	Coverage	Model	Surface runoff	Lateral flow	Percolation
	%		mm.year^{-1}		
A/B interflow	9.9	Uncalibrated	185.9	0.4	23.7
		Calibrated	268.0	42.8	3.0
Recharge deep	29.3	Uncalibrated	27.8	4.0	229.8
		Calibrated	54.2	34.9	179.9
Recharge shallow	13	Uncalibrated	265.4	7.0	6.9
		Calibrated	51.6	39.3	315.1
Responsive saturated	19.8	Uncalibrated	170.0	5.6	28.3
		Calibrated	189.7	122.4	15.5
Responsive shallow	12.5	Uncalibrated	268.0	6.1	7.7
		Calibrated	304.8	60.1	6.9
Soil/bedrock interflow	15.6	Uncalibrated	161.0	2.4	127.1
		Calibrated	201.6	148.8	9.9

Recharge shallow soils differed substantially after calibration where far less surface runoff was simulated, slightly higher average annual lateral flow values and substantially higher percolation values were also simulated, with an average annual increase of 308 mm of recharge being simulated (Figure 3.15). These increases can be attributed by changing the soil hydrologic group from D to group A, therefore drastically lowering CN values, increasing infiltration and limiting lateral flow. The dominant hydrological response of recharge shallow soils now reflect their hydropedological characteristics. The vertical flow through and out of the profile into the underlying fractured bedrock is the dominant flow direction, without any morphological evidence of temporary or permanent periods of saturation (Van Tol et al., 2015).

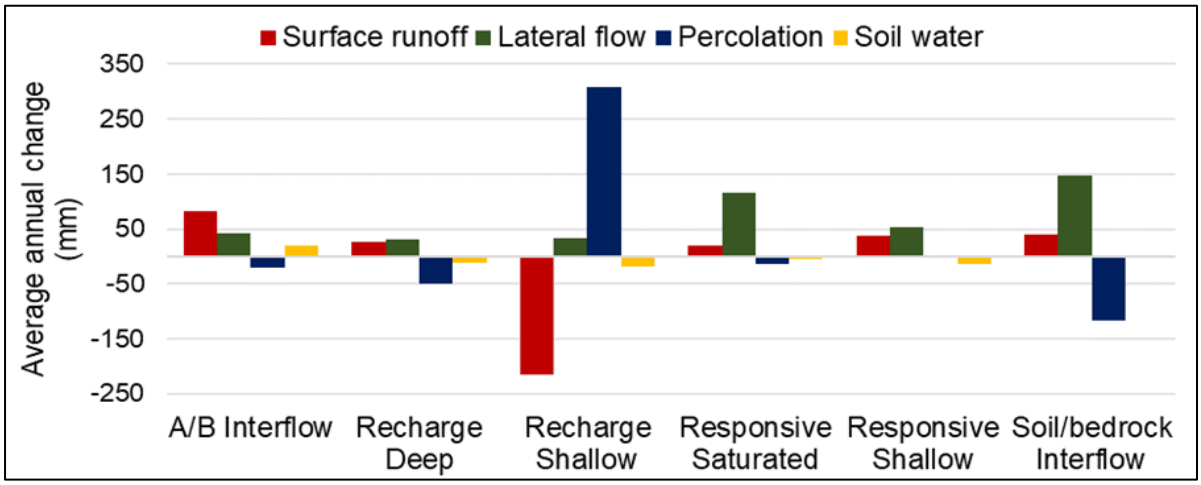

Figure 3.15: Average annual change in hydrological processes (mm) at the soil level from uncalibrated to calibrated model runs for the entire catchment.

The same can be said for both responsive saturated and responsive shallow soils where surface runoff values slightly increased after calibration (Table 3.17). These soils contributed limited lateral flow and percolation volumes, primarily only contributing surface runoff. Therefore, their hydropedology characteristics are accurately reflected within the model.

Soil/bedrock interflow soils showed a marked increase in average annual lateral flow volumes after calibration with an increase of 132 mm (Table 3.17), improving the hydropedological representation of these soils. The increase in lateral flow is in large part due to the PERCO parameter being adjusted, limiting percolation, therefore, increasing lateral flow volumes, indicated by the 117 mm decrease in percolation annually. Surface runoff is still the dominant hydrological process with an average annual contribution of 172 mm, which is to be expected (Van Tol et al., 2015).

A/B interflow also showed improved lateral flow simulations, with decreases in percolation volumes. Surface runoff remains the primary hydrological response with an average annual surface runoff contribution of 268 mm (Table 3.17). However, it remains uncertain if A/B interflow soils are adequately reflected within the model as the SWAT+ model does not output hydrological processes for each soil horizon but rather aggregates these processes, which is a current shortcoming of the SWAT+ model.

The spatial distribution of hydrological processes also changed between uncalibrated and calibrated models (Figure 3.16). More surface runoff is simulated in the drier savanna sections of the catchment after calibration as a result of increases in curve numbers between model simulations. The calibrated recharge shallow soils also now illustrate significantly less surface runoff, especially at the most mountainous sections of the catchment.

The additional water provided after calibration for different hydrological processes significantly impacted the spatial distribution of lateral flow within the catchment, where calibration allowed significant increases in lateral flow to be simulated. Low lateral flow values in the uncalibrated model could be attributed to the low saturated hydraulic conductivity values within the HYDROSOIL datasets which mostly inhibited lateral flow, allowing percolation to be the dominant hydrological process within the soil. Most lateral flow was simulated in the west of the catchment where steeper slopes and significantly more rainfall occurs (Figure 3.16). Percolation also changed spatially after calibration where the spatial distribution of percolation aligns with recharge deep and recharge shallow soils.

Average annual soil water contents also showed a significant spatial change after calibration, where increases in average annual soil water for interflow soils (A/B and soil/bedrock) increases soil water content in the savanna regions of the catchment in the east (Figure 3.16). Decreases in soil water content for recharge (deep and shallow) and responsive shallow soils also affected in the spatial distribution of soil water in the mountainous regions in the west of the catchment, where recharge deep soils stored the most soil water due to the soils deep profile and relative position in the landscape.

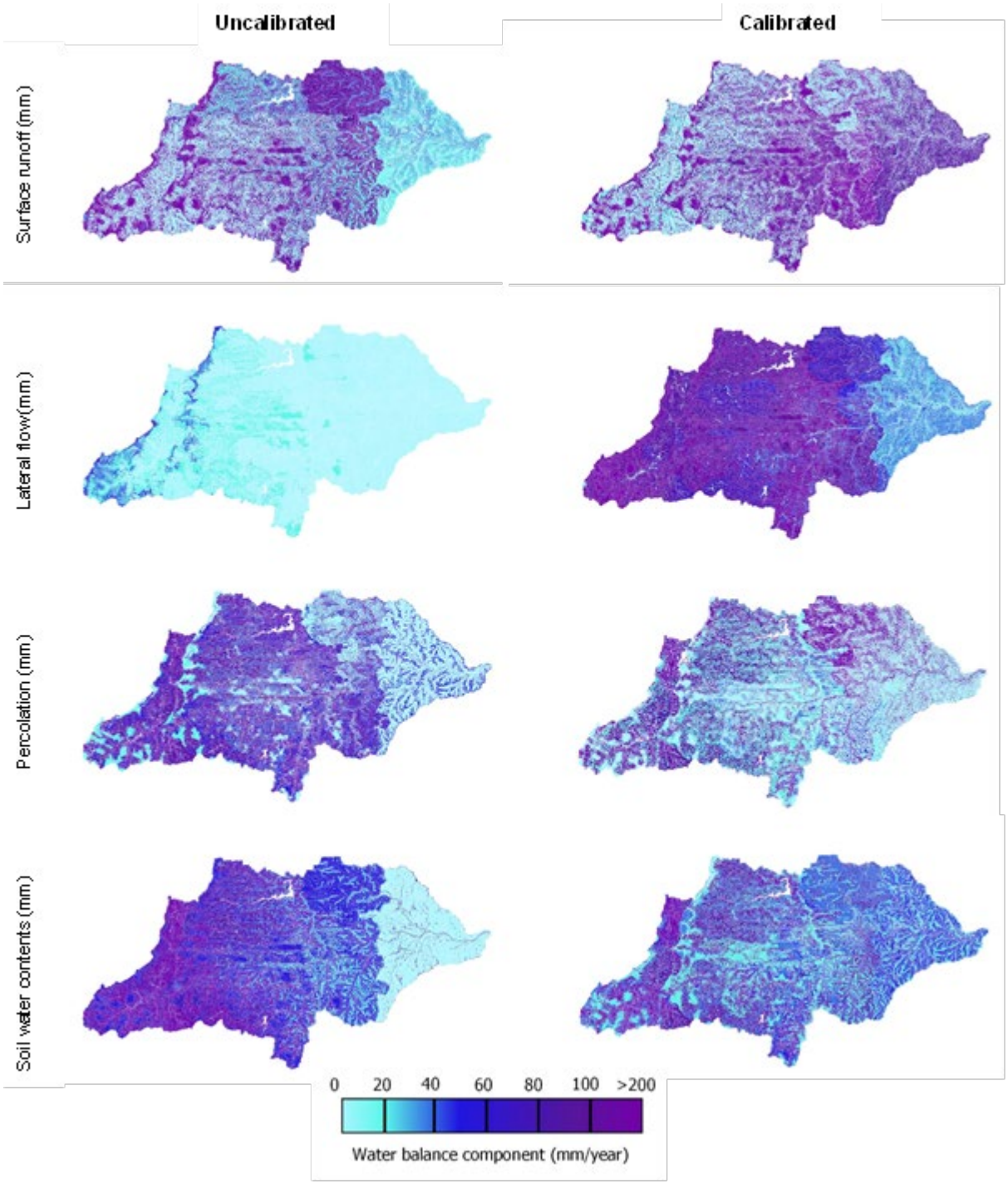

Figure 3.16: Average annual hydrological processes (mm) at the HRU-level for the uncalibrated and calibrated hydrological models.

These results are in accordance with other studies focusing on hydrological soil information in hydrological modelling (Van Tol et al., 2013; Sierra et al., 2018; Smit & Van Tol, 2022; Harrison et al., 2022), They emphasise the potential of calibrating hydrological models using hydropedology as an information carrier to improve representation of hydrological process. This aligns with the study by Van Tol et al. (2021a) which illustrated the ability of an accurate hydrological soil map to act as data carrier for hydrological modelling purposes. Additionally, Bouma et al. (2022) cited the need to adopt soil classification as data source and information carrier to provide solutions to the Sustainable Development Goals inextricably linked to soil function (Lal et al., 2021; Bouma et al., 2021).

This research shows that accurate hydrological soil information, based on hydropedology, can carry information that may serve as an effective 'soft data' to better represent internal hydrological processes within a catchment, leading to improved representation of internal hydrological processes (Seibert & McDonnell, 2002; Van Tol et al., 2021a; Smit & Van Tol, 2022). Further calibration of SWAT+ parameters, in particular PERCO, LATQ_CO and CN2, allowed the linkage between hydropedology and different dominant hydrological processes, which is captured within the HYDROSOIL dataset, to be better represented within the SWAT+ model.

These results also agree with other studies focusing on the importance of hydropedology within physically-based hydrological models (Bouma et al., 2011; Van Tol et al., 2021b). Representing the spatial distribution of dominant hydrological processes remains imperative to hydrological modelling for decision-making and policy purposes (Bossa et al., 2012; Wahren et al., 2016; Bouslihim et al., 2019). Modern water resource management plans are concerned with impacts at the local level, where the improved detail and accuracy of hydropedological information has been shown as more applicable than less detailed soil information (Harrison et al., 2022; Smit et al., 2023a).

3.3.5 Conclusions

Statistical analyses indicated substantial modelling improvement during both calibration and validation compared to the uncalibrated model. Further calibration of SWAT+ parameters, in particular PERCO, LATQ_CO and CN2, allowed the linkage between hydropedology and different dominant hydrological processes, which implies that hydropedology could be considered as a viable source of 'soft data' within the SWAT+ model. Accurate hydrological soil maps should form an integral part of modern process-based hydrological modelling as they can act as important data sources and information carriers relating the variability of different hydrological processes across a landscape.

This is particularly promising for hydrological modelling in ungauged basins as well as the regionalization of hydrological soil information, where hydropedology could form an important 'soft data' tool to aid different modelling approaches. Future research should focus on testing the calibration of hydrological soil types in different hydrological conditions, determining the applications of calibrated hydrological soil information for ungauged basins, as well as its impact on modelling land-use and climate change scenario analyses.

CHAPTER 4: OLIFANTS CATCHMENT

Chapter 4 describes the digital soil mapping and hydrological modelling of the Olifants River catchment. It centres on how Land Type field data could be incorporated into digital soil mapping in South Africa (Section 4.1). This work forms part of the PhD thesis of Molebaleng Sehlapelo, and has been prepared as a peer-review paper to be submitted to the *South African Journal of Plant and Soil* (SAJPS,2023). describes how the soil map was then used to model the hydrology of the Olifants River catchment (Section 4.2). Eddy Smit prepared this part of the manuscript.

4.1 INVESTIGATING THE ACCURACY OF DIGITISED LAND TYPE FIELD DATA IN DIGITAL SOIL MAPPING

4.1.1 Abstract

The acquisition of soil data in South Africa, as a developing country, has always been limited due to the lack of resources. However, these limitations were overthrown by the introduction of digital soil mapping with machine learning algorithms. Over the years, soil data collection has been recorded manually on topo-cadastral maps. This data was perceived to be geographically inaccurate due to the unavailability of Global Positioning Systems (GPS) at the time of data collection. However, the introduction of geo-referencing and digitising of scanned maps increased the availability of recorded soil data in digital format. The aim is to investigate whether using digitised soil point data in digital soil mapping affects the accuracy of soil maps. The Olifants catchment was chosen as the study area where soil point data and covariate data were collected to create hydropedological maps. The validation point accuracy and Kappa coefficient values were used to investigate the accuracy of the maps. The results indicated that adding Land Type field data to the collected soil point data decreased the accuracy of the maps. However, adding buffers to negate the assumed inaccurate geographical position of the Land Type field data increased the accuracy. Furthermore, the addition of a 100 m buffer resulted in the most accurate map yielding a validation point accuracy of 73.3% and Kappa value of 0.86. Therefore, Land Type field data can be used in digital soil mapping, however, it is necessary to negate the uncertainties associated with the geographical positions of the digitised Land Type field data.

4.1.2 Introduction

Soil is the most complex and diverse natural resource in the world. It is a vital ecosystem that is non-renewable, which serves important environmental, economic and social functions (Blum et al., 2006. Soil has become a limited resource due to rapid human population growth and intensified agricultural practices aiming for higher crop yields. Consequently, soil faces significant pressure because of the diverse services it provides (Kopittke et al., 2019). Some of these services include food security, raw materials, infrastructure support, water resources, carbon storage and land degradation neutrality (Padarian et al., 2015). Comprehending the impacts of the pressures on soil is essential for sustainable soil management. However, the ability to monitor and manage soil relies heavily on the accessibility of precise spatial information about the soil (Zhang et al., 2017). Soil properties, agricultural performance and yield can exhibit significant variations over short distances due to the diverse nature of soils (Iqbal et al., 2005. Therefore, recording variation in soil properties on a larger scale is essential in understanding their impact on agricultural and environmental processes (Cook et al., 2008).

Due to the variation of soil properties, the acquisition of data becomes cumbersome and necessitates significant financial and technical efforts. Moreover, the process of measuring, recording and mapping these variations is labour-intensive and time-consuming (Paterson et al., 2015). Many developing countries lack detailed information on their soil resources (Dewitte et al., 2013). The lack of data results in a significant gap in our understanding of soils physical and biological properties, making it the sole

missing layer of information on a global to local scale (Nachtergaele et al., 2010). There is an increasing recognition of the critical importance of soil knowledge and raising awareness of its global values to the public, policy makers and land managers (Hengl et al., 2017).

There is a trend where fewer field-based soil data are being collected, and older soil data are favoured over new field soil data (Pangos, 2011). Therefore, data collected in the past must be preserved because they serves as a foundation for many of the research that is conducted today. Their scientific value in analysing historical changes in soils over time can be used now and in future (Taghizadeh-Mehrjardi et al., 2019). Therefore, it is imperative that important soil data, which is now only available on papers and maps, be better saved and digitised before they are lost (Pangos, 2011).

In South Africa, the lack of soil information was resolved through the introduction of soil surveys that were used to gather information on soil resources (Zeraatpisheh et al., 2020). The national Land Type Survey in South Africa started in the early 1970s based on field surveys which used 1:50 000 topo-cadastral maps as base maps for collecting point information on soil properties at different locations (Paterson et al., 2015). The primary focus was to map and record the distribution of soils and their functions. The introduction of digital computing in the early 1990s allowed soil data to be transferred to a digital format. The 1:50 000 topo-cadastral soil maps could be digitised and edge mapped to represent the coverage of South Africa (Paterson et al., 2015). Although the method of soil surveys was effective for mapping small fields, it could not be used to create maps for larger areas (Van Zijl et al., 2014b). Due to the growing need for soil data and the unavailability of soil maps, modelling techniques were developed to spatially predict soil properties. Such includes the application of digital soil mapping with machine learning algorithms (Padarian et al., 2020).

Soil data acquisition in South Africa improved due to the introduction of digital soil mapping, as the tool utilises representative and spatially distributed soil data (Van Zijl, 2019). Digital soil mapping exploits soil point data and environmental covariates that are put through machine learning methods to derive the relationship between the soil forming factors and soil properties (Minasny & McBratney, 2015). It depends on the accurate evaluation of the correlation between covariates and a set of observations, which is influenced by the selection of the covariates to be used to represent the relationship between soil and environment. In digital soil mapping, selecting the appropriate covariates (scorpan factors) is often the key to creating soil maps that can clearly indicate the soil knowledge (Peng et al., 2020). The scorpan factors are derived from remote sensing, proximal sensing or easily measured soil properties (Flynn et al., 2019a). These refer to numerical descriptions of the connections between soil and factors that are spatially referenced (McBratney et al., 2003). There are seven factors that make up the scorpan model which is written as: $S_c = f\,(s.c.o.r.p.a.n)$ or $S_a = f(s,c,o,p,a,n)$, where S_c is soil classes and is S_a is soil attributes (McBratney et al., 2003) . The acronym stands for each factor, *s* (soil), *c* (climate), *o* (organisms), *r* (topography), *p* (parent material), *a* (time) and *n* (spatial position) (Grunwald, 2009).

This study investigated whether Land Type field data can be used to create accurate maps in digital soil mapping. Since the Land Type field data was collected without the use of GPS devices to verify the coordinates, it is assumed that the geographical position of the recorded soil points is inaccurate. Therefore, it is important to understand whether this inaccuracy affects the accuracy of the created maps in digital soil mapping.

4.1.3 Materials and methods

The Olifants catchment

The study area for mapping was the upper Olifants catchment, which falls within two provinces (Gauteng and Mpumalanga) of north-eastern South Africa (Figure 4.1). The upper Olifants catchment area is known for mining, agricultural and power generation activities, which are greatly dependent on a range of goods and services obtained from local aquatic ecosystems (Dabrowski & De Klerk, 2013). The catchment is characterised by ground and surface water pollution due to the anthropogenic stressors in the catchment, including extensive coal mining resulting in acidic water (Hobbs et al., 2008).

The catchment receives rainfall during the summer months (October to April), with an annual rainfall ranging between 500 and 800 mm in most parts of the catchment. The rainfall is characterised by high variability of semi-arid climate and a temperature ranging between -4 and 45℃ (Olabanji et al., 2020). The geology of the catchment consists of all three major rock forms (igneous, metamorphic and sedimentary). The oldest rock formation is exposed in the eastern lowveld of the catchment, which is the Archean Granite and Gneiss Basalt Complex. Additionally, the rock formation consists of phyllites, banded ironstone, quartzites, conglomerate and limestone that have gone through metamorphism. There is also a group of igneous rocks embedded in the same rock formation, including amphibolites, greenstone lavas, and chlorite-schists (Thomas, 2015). The catchment is characterised by a variety of soil types, and the major soil types are moderately deep sandy to sandy-clay loams (Idowu et al. 2010).

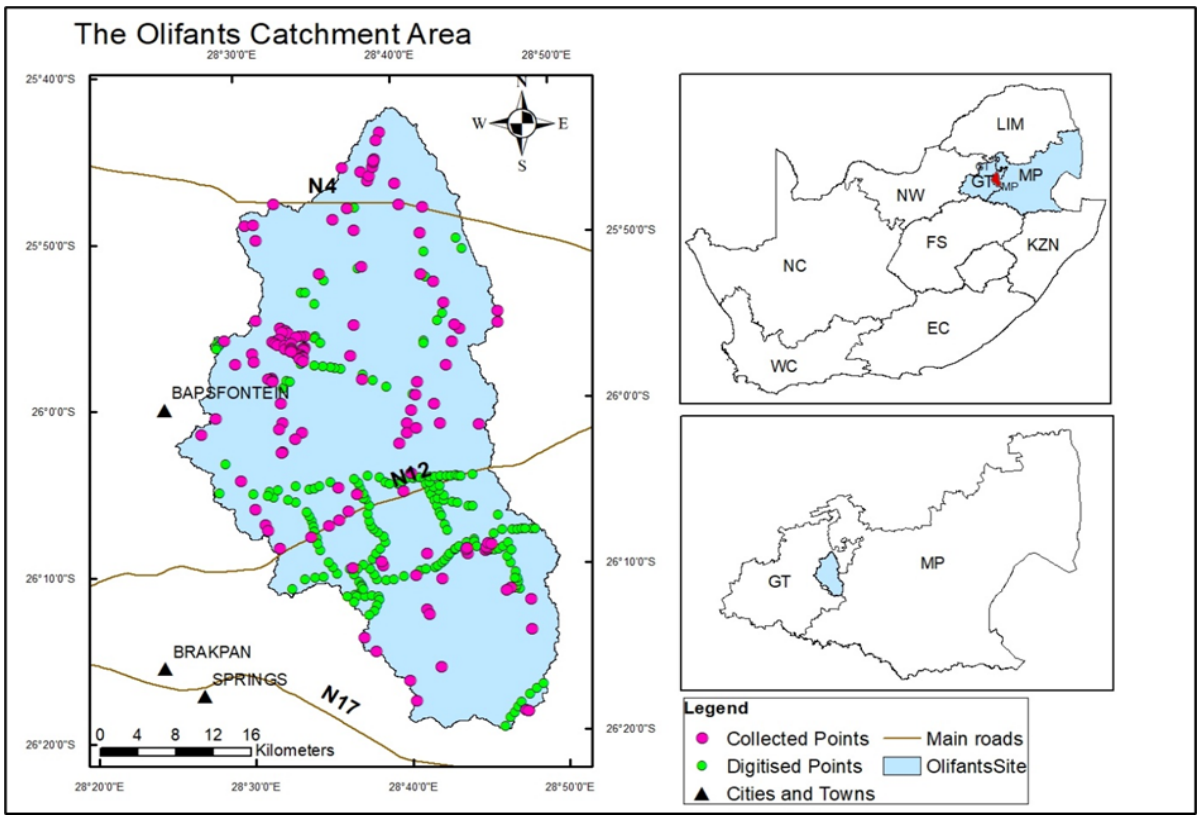

Figure 4.1: The Olifants catchment and soil point data used in the study.

Soil point data

Two different sets of soil data were collected for this study. Firstly, Land Type field data was collected from the ARC. The data was recorded on 1:50 000 topographic sheets with estimated geographical

positions, recorded as a dot and soil form on the map with a pencil or pen by the surveyor. The pen or pencil was assumed to have a thickness of 5 mm. Therefore, 5 mm on the topo-cadastral sheet is 50 m on the ground, meaning that the recorded point is already in a 50 m displacement. The maps were scanned, georeferenced and digitised in ArcMap 10.7.1 to obtain geographical coordinates of the recorded points. This resulted in a total of 193 digitised soil points.
The second set of soil data was collected from legacy soil point data obtained by other institutions during fieldwork. Soil point data recently (2022) collected by North-West University students was also used for this study. This resulted in a total of 136 collected soil points with coordinates obtained from GPS devices used during fieldwork. These soil observations were classified to soil-form level according to the South African Soil Classification System (Soil Classification Working Group, 1991), and divided into conceptual hydropedological properties from the soil descriptions (Table 4.1; Van Tol & Le Roux, 2019).

Table 4.1: Soil forms in the Olifants catchment divided into hydropedological classes according to Van Tol and Le Roux (2019).

Recharge		**Interflow**		**Saturated responsive**	**Stagnating**
Deep	**Shallow**	**A/B horizon**	**Soil/bedrock**		
Bonheim	Glenrosa	Cartref	Avalon	Katspruit	Dresden
Carolina	Mayo	Constantia	Bainsvlei	Rensburg	
Clovelly	Mispah	Estcourt	Fernwood	Willowbrook	
Glen	Rustenburg	Kransfontein	Glencoe		
Graffin		Kroonstad	Pinedene		
Hutton		Longlands	Sepane		
Nkonkoni			Tukulu		
Oakleaf			Westleigh		
Shortlands					
Swartlands					
Tongwane					
Valsrivier					

Covariate data

One or more covariates were selected to represent the different scorpan factors – *s* (soil), *c* (climate), *o* (organisms), *r* (topography), *p* (parent material), *a* (time) and *n* (spatial position). Satellite images were collected from Sentinel Hub for the wet (16/04/2021) and dry (23/09/2021) seasons. Four sets of bands were obtained (blue-band 2, green-band 3, red-band 4, and NIR-band 8) from the satellite images to calculate the required indices (see Table 3.2) through mathematical manipulation carried out in SAGA-GIS 2.2.5.

Furthermore, a DEM was obtained from SRTM with a 90 m resolution (USGS, 2015). Topographic covariates (Table 4.2) were derived from the DEM using the basic analysis tool in SAGA-GIS 2.2.5. Similarly, the Multiresolution Index of Valley Bottom Flatness (MRVBF) was derived from DEM using the Morphometry analysis tool. However, not all the covariates were used in creating the maps. Data on the geology, land types, temperature and rainfall of the study area was also obtained. This resulted in a total of 28 covariates, which were resampled to have the same grid extent resolution of 30 m.

Table 4.2: Topographic covariates derived from the Digital Elevation Model.

Topographic covariates
Analytical Hill shading
Slope
Aspect
Plan Curvature
Profile Curvature
Convergence Index
Closed Depression
Total Catchment Area
Topographic Wetness Index
Slope Length and Steepness factor (LS-Factor)
Channel Network Base Level
Channel Network Distances
Valley Depth
Relative Slope Position
MRVBF
MRRTF

MRVBF = Multiresolution Index of Valley Bottom Flatness;
MRRTF = Multiresolution Ridge Top Flatness

Map creation

The HYDROSOIL maps were generated from various soil point datasets. Three different maps were created using three different sampling methods (Conditioned Latin Hypercube, K-means clustering and Stratified Random Sampling), with the multinomial regression algorithm in R studio using all the covariate layers generated. These sampling methods were used to divide the observation points into training (75%) and validation datasets (25%). Three types of soil point data were used to create the different maps (collected soil point data, Land Type field data and Land Type field data with different sized buffers). Different hydropedological properties were used as mapping units for the created maps. The most accurate map created using the three-sampling method was used as a baseline map. The baseline map was generated using the collected soil point data to compare or observe the changes in accuracy following the addition of Land Type field data.

The accuracy of the maps was investigated using the Kappa coefficient and validation point accuracy to measure whether the map was an acceptable representation of reality. The validation dataset (25%) was used to calculate the Kappa coefficient and validation point accuracy values during validation testing. The validation points were added to the map units of the generated maps, and each validation point was examined to confirm alignment with the corresponding map unit. Total point accuracy refers to the total number of validation observations correctly predicted, and the Kappa coefficient is the reflection of reality by the map, with values close to 0 indicating a random designation of mapping units and values close to 1 indicating an accurate representation of reality (Van Zijl , 2019).

The next map was created using both the collected soil point data and the Land Type field data, using the sampling method that resulted in the most accurate map in the previous step. This map was created for the purpose of investigating the impact that Land Type field data could have on the accuracy of the

baseline map. The accuracy of this map was investigated using the same validation dataset as used for the baseline map, the validation dataset was kept the same for statistical purposes.

The last four maps were created using both the collected soil point data and Land Type field data. A 50, 100, 200 and 500 m buffer was added to the Land Type field data using the *Shapes Buffer* tool in SAGA-GIS 2.2.5. The different sized buffers were added to Land Type field data for the purpose of negating the uncertainties associated with the geographical locations or the removal of the geographical displacement of the points recorded on the topo-cadastral maps. This was done to investigate which buffer is most effective for the Olifants catchment.

4.1.4 Results and discussion

The accuracy of the three maps generated through Conditioned Latin Hypercube (Figure 4.2a), K-means clustering (Figure 4.2b), and Stratified Random Sampling (Figure 4.2c) were assessed based on an accuracy matrix for the hydropedological classes used as mapping units (Table 4.3).

The map created with the Conditioned Latin Hypercube sampling method and collected soil point data demonstrated the highest validation point accuracy of 50% and a Kappa coefficient of 0.47, surpassing the accuracy achieved by maps generated with alternative sampling methods (Table 4.4). Therefore, the map was used as the baseline map to compare how the accuracy of the maps changed with the inclusion of land type field point data. The same sampling method was used to create all the maps due to the high accuracy. The map accuracy was assessed using the same validation dataset.

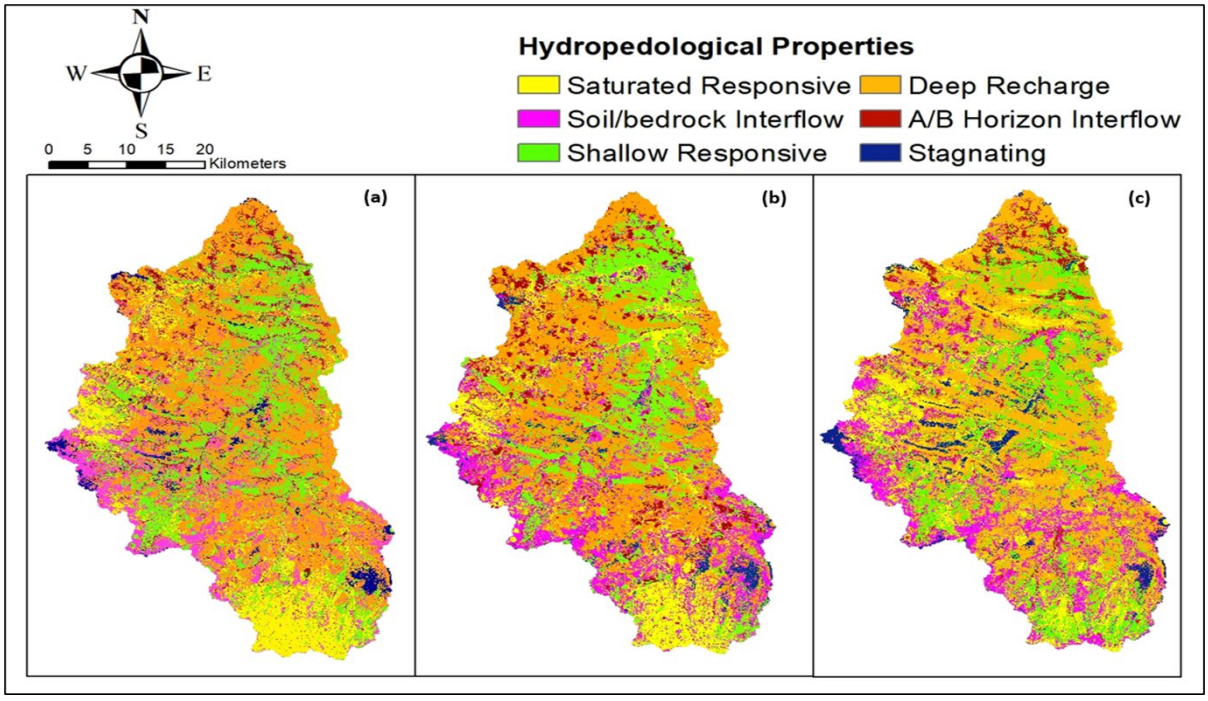

Figure 4.2: Hydropedological maps created using different sampling methods including (a) Conditioned Latin Hypercube, (b) K-means Clustering and (c) Stratified Random Sampling.

Table 4.3: Accuracy matrix for hydropedological soil class map created with the Conditioned Latin Hypercube sampling method.

		Map units									
		Deep recharge	**Shallow recharge**	**Interflow A/B**	**Interflow Soil/bedrock**	**Stagnating**	**Shallow responsive**	**Saturated responsive**	**Total**	**Correct**	**%**
Observations	**Deep recharge**	15	5			2	2	2	26	15	57.7
	Shallow recharge	1							1	0	0.0
	Interflow A/B	9				1			18	8	44.4
	Interflow Soil/bedrock	2		8	3		1		9	3	33.3
	Stagnating	1		5	2	5			9	5	55.6
	Shallow responsive	2		1		1	3		6	3	50.0
	Saturated responsive							7	7	7	100.0
	Total	30	5	12	5	9	6	9	76		
	Correct	15	0	8	3	5	3	7		41	
	%	50.0	0.0	66.7	60.0	55.6	50.0	77.8			50.0
	Kappa	0.47									

Table 4.4: Kappa coefficient and validation point accuracy (%) of maps created using different sampling methods.

Sampling method	Kappa coefficient	Validation point accuracy (%)
Conditioned Latin Hypercube	0.47	50
K-means Clustering	0.36	33.3
Stratified Random Sampling	0.29	39.4

The next five maps were created using the Conditioned Latin Hypercube sampling method, collected soil point data and digitised Land Type field data. The first map with no buffers (Figure 4.3a) resulted in a Kappa coefficient of 0.45 and validation point accuracy of 63.6% (Table 4.5). These results indicate a slight decrease in Kappa value from the baseline map, although there is an increase in the validation point accuracy.

Adding a 50 m buffer (Figure 4.3b) resulted in an increased Kappa coefficient of 0.76 and a validation point accuracy of 70%, when compared to the map created without any buffers. Increasing the buffer size from 50 m to 100 m (Figure 4.3c) again positively influenced the accuracy of the map, with a Kappa coefficient of 0.86 and validation point accuracy of 73.3%. A 200 m buffer (Figure 4.3d) yielded a Kappa coefficient of 0.59 and validation point accuracy of 66.7% – a reduction in both the Kappa coefficient and validation point accuracy when compared to the previous buffers. Nevertheless, these values remain higher than those obtained for the map created with digitised Land Type field data without any added buffers. The 500 m buffer (Figure 4.3e) yielded a Kappa coefficient of 0.70 and validation point accuracy of 80%. This is highest validation point accuracy compared to all the created maps.

Table 4.5: Kappa coefficient and validation point accuracy (%) of maps created using collected soil point data and Land Type field data with different sized buffers.

Buffer	Kappa coefficient	Validation point accuracy (%)
No buffer	0.45	63.3
50 m	0.76	70.0
100 m	0.86	73.3
200 m	0.59	66.7
500 m	0.70	80.0

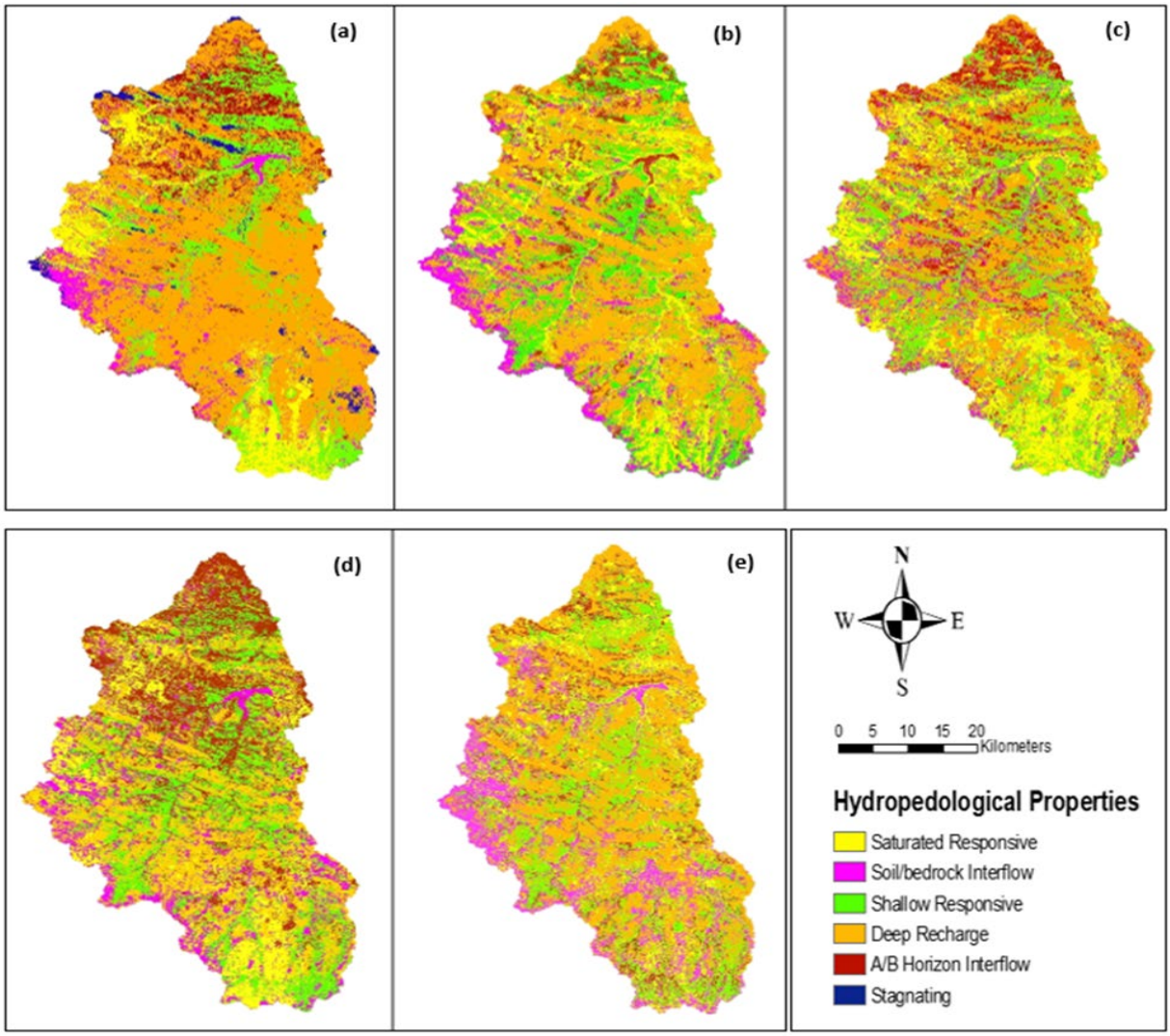

Figure 4.3: Hydropedological maps created using collected soil point data and Land Type field data with (a) no buffer, (b) 50 m buffer, (c) 100 m buffer, (d) 200 m buffer and (e) 500 m buffer.

The findings suggest that incorporating various-sized buffers to the Land Type field data led to a notable increase in the Kappa coefficient, signifying enhanced map accuracy (Figure 4.4). Despite a dip in the Kappa coefficient with the addition of a 200 m buffer, the map's accuracy remained higher than that of the map created with land type field data without any buffer. Notably, the inclusion of a 100 m buffer with the land type field data resulted in the highest Kappa coefficient, signifying the most accurate map.

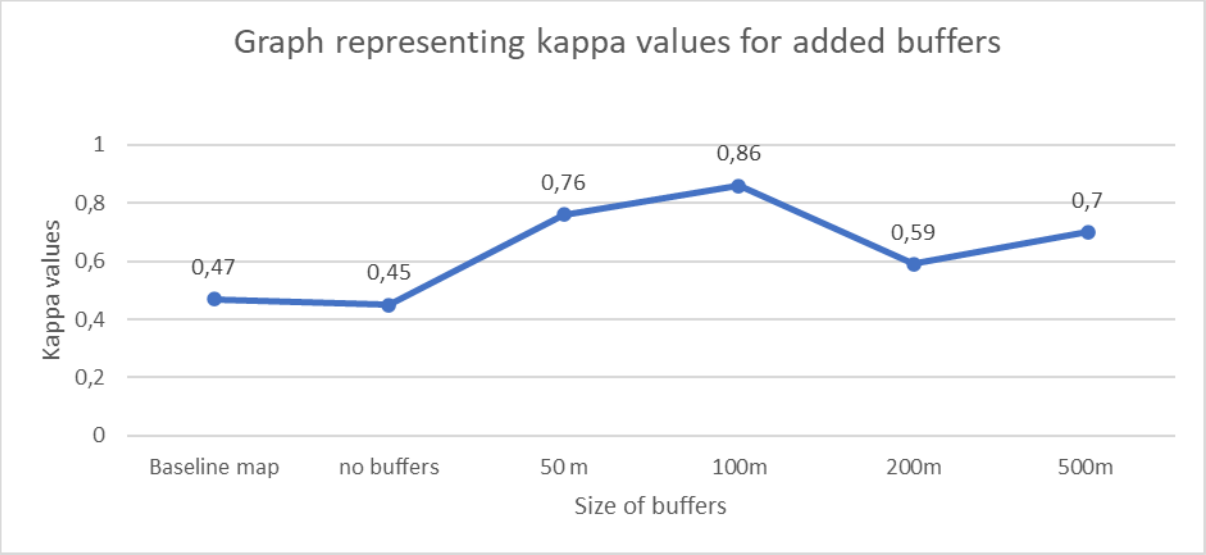

Figure 4.4: The accuracy of the created hydropedological maps with different buffer sizes as measured using the Kappa coefficient.

4.1.5 Conclusions

In the past, data collection predominantly relied on paper and maps before the introduction of electronic storage options. Unfortunately, some valuable soil data was either lost or consigned to archives, leading to a scarcity of available soil data or duplication thereof. This challenge is intensified by the labour-intensive and time-consuming nature of collecting and analysing new soil data. The introduction of data digitisation and georeferencing enabled the transition from traditional topographic maps to digital records. However, the coordinates for these soil points were inaccurate as they were estimated in the absence of GPS devices. The question arose whether such soil data, marked by geographical uncertainties, could be effectively utilised in digital soil mapping. The findings revealed that adding the Land Type field data to collected soil point data slightly reduced the accuracy of the resulting map. However, introducing buffers to mitigate geographical uncertainties significantly improved map accuracy. Incorporating a 100 m buffer to the digitised Land Type field data yielded the highest accuracy for the Olifants catchment. In conclusion, digitised land type field data can indeed be used in digital soil mapping. However, to address uncertainties associated with geographical positions, it is essential to incorporate buffers. Further investigations are essential to determine the most effective buffer size for the specific study area.

4.2 COMPARING HYDROSOIL AND LAND TYPE SOIL INFORMATION IN THE UPPER OLIFANTS CATCHMENT USING SWAT+

4.2.1 Introduction

As was done for the Sabie catchment (Section 3.2), this section again compares the created HYDROSOIL information to readily available Land Type data in the upper Olifants catchment using SWAT+. The aim was to compare long-term daily streamflow data between the two models in a highly anthropogenically altered, dolomite-dominated catchment at two catchment scales.

4.2.2 Materials and methods

The upper Olifants catchment

The 330 km^2 upper Olifants catchment is located in the Gauteng province of South Africa (Figure 4.5). The catchment elevation ranges from 1 683 m.a.s.l. and gradually flattens towards the north-east with an altitude of 1 450 m.a.s.l., as the Olifants River continues to flow toward Mozambique. Dolomite and quartzite are the primary geology present in the catchment area (Council for Geoscience, 2007).

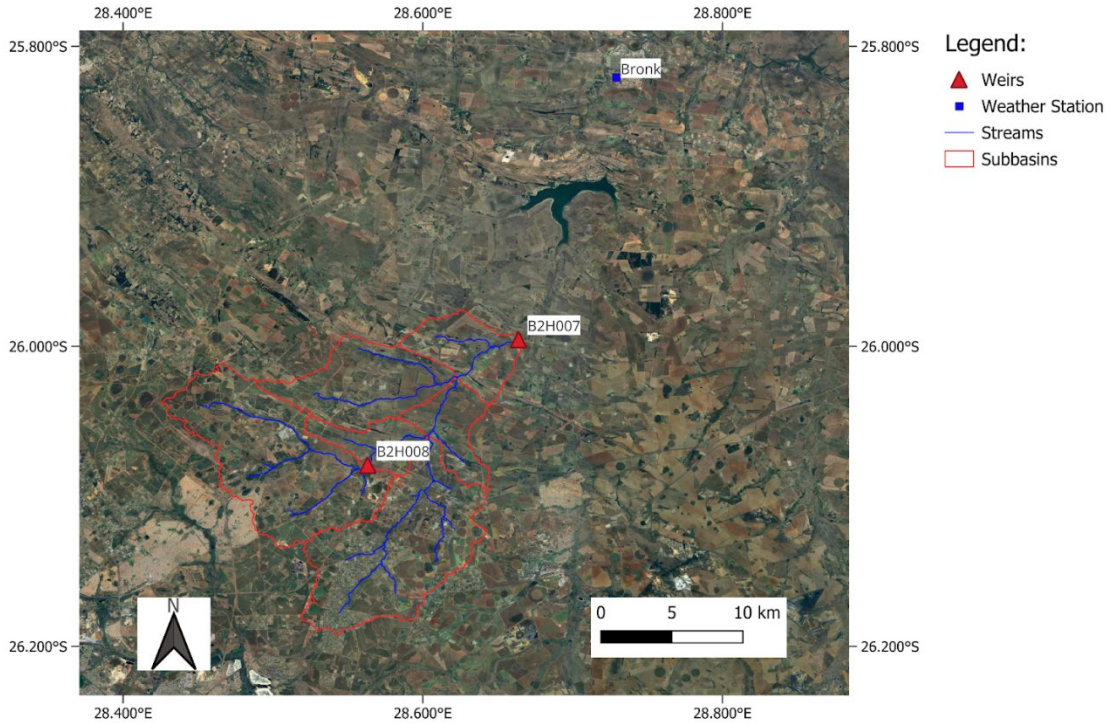

Figure 4.5: The upper Olifants River catchment, with weirs, streams, subbasins and climate station.

Model, inputs and setup

The same process as in the Sabie catchment (Section 3.2) was used. The QSWAT+ (v. 2.3) plugin was used to set up the catchment. The model warm-up period lasted for the first four years, followed by a five-year daily validation period.

Daily rainfall, maximum and minimum temperature data was obtained from the Bronkhorstspruit (Bronk) climate station. All data was received courtesy of the South African Weather Service. Daily solar radiation, relative humidity and wind speed were obtained from the Climate Forecast System Reanalysis which was done by the National Center for Environmental Prediction (Saha et al., 2015).

The DEM was obtained from a 30 m x 30 m SRTM (USGS, 2022; Figure 4.6). The land cover data (Figure 4.7), was acquired from the 2013/14 South African National Land-Cover Map (GeoTerra Image, 2015). For the land cover input, predefined SWAT values associated with various land-use classes were utilised. Additionally, dams identified in the land cover were integrated into the model setup, designated as 'reservoirs' and assigned default values.

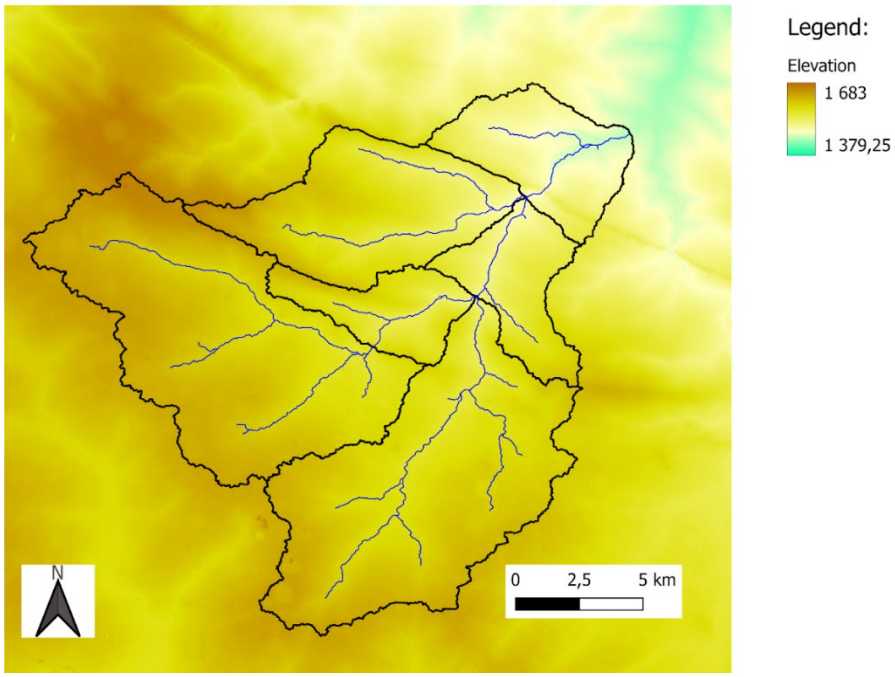

Figure 4.6: The elevation of the upper Olifants River catchment.

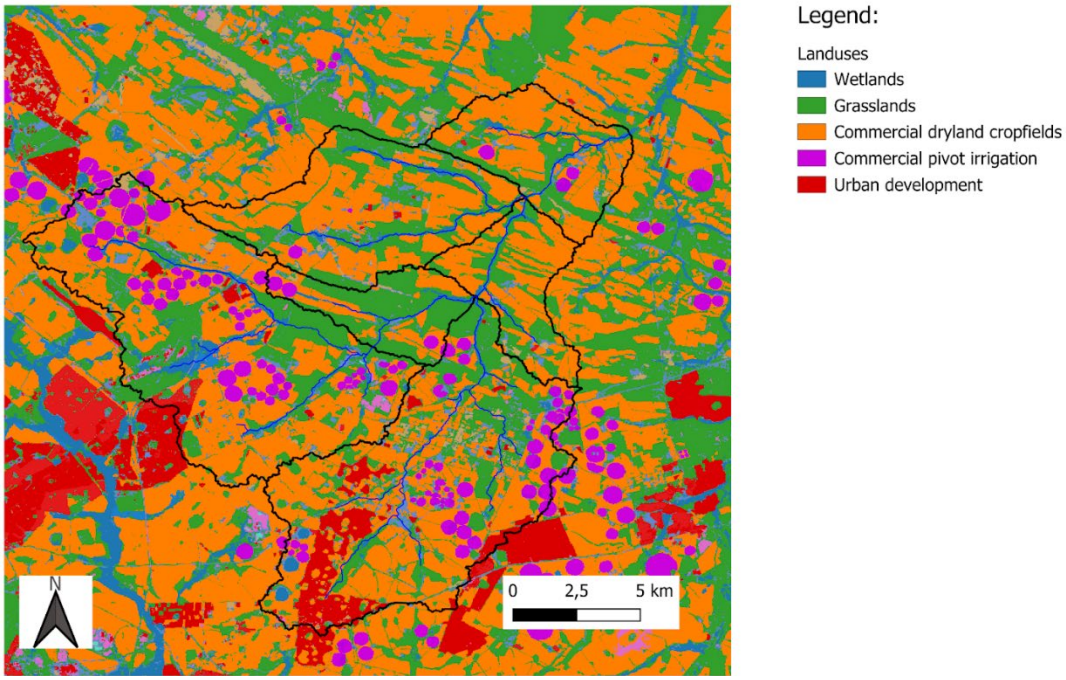

Figure 4.7: The land uses within the upper Olifants catchment as demarcated from the 2013/2014 South African National Land Cover Map.

Soil information

Maps using the HYDROSOIL soil information and the Land Types were compared for the upper Olifants catchment (Figure 4.8). The Land Type database had already been converted to a readily available spatial soil database specifically for use within the SWAT model (Le Roux et al., 2023). In the upper Olifants catchment, there are five Land Type groups each with their own set of hydraulic properties (Table 4.6).

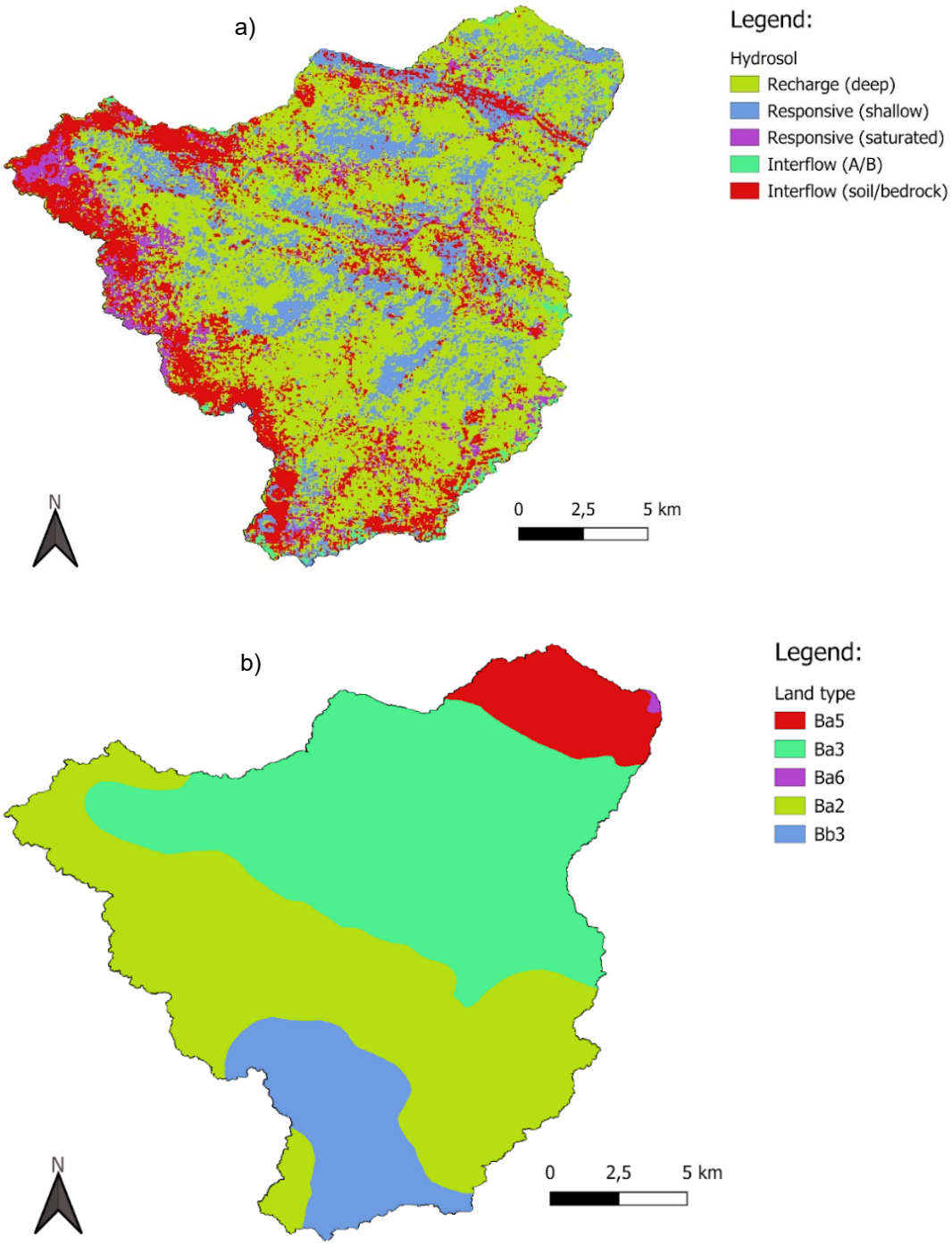

Figure 4.8: a) The HYDROSOIL map and b) Land Types present within the Olifants catchment.

Table 4.6: The main hydraulic properties of the Land Type mapping units.

Land Type	Horizon	Hydro-group	Depth	Bd	AWC	Ksat	OC	Clay	Silt	Sand
			mm	g/cm^3	mm/mm	mm/h	%	%	%	%
Ba5	A	B	300	1.5	0.084	13	1	18.6	17.5	63.9
	B		870	1.5	0.076	210	0	18.6	17.5	63.9
Ba3	A	B	300	1.5	0.089	4.3	1.25	20.8	17.5	61.7
	B		790	1.6	0.082	210	0.25	20.8	17.5	61.7
Ba6	A	B	300	1.5	0.084	4.3	1	19.1	17.5	63.4
	B		870	1.5	0.076	210	0	19.1	17.5	63.4
Ba2	A	B	300	1.5	0.086	4.3	1.0	21.2	17.5	61.3
	B		1000	1.5	0.074	210	0	21.2	17.5	61.3
Bb3	A	B	300	1.5	0.08	4.3	1	17.5	61.4	17.5
	B		990	1.5	0.067	210	0.1	21.1	17.5	61.4

Bd = Bulk density; AWC = Available Water Capacity; Ksat = saturated hydraulic conductivity; OC = Organic Carbon.

The second spatial soil dataset was the hydropedological dataset (HYDROSOIL), which was created in Section 4.1 using a 100 m buffer zone. A Kappa coefficient of 0.86 indicates a strong agreement between the soil map and actual observations (Section 4.1). The general morphological descriptions of the HYDROSOIL are given in Table 4.7.

Table 4.7: The characteristics of the hydrological mapping units of the upper Olifants catchment.

Hydrological mapping unit	Soil form	WRB Reference Groups	Defining hydrological characteristic
Recharge deep	Hutton, Longtom, Kranskop	Acrisols, Nitisols, Fluvisols	Deep soils without any morphological indication of saturation. Vertical flow through and out of the profile into the underlying bedrock is the dominant flow direction.
Responsive saturated	Katspruit, Champagne	Gleysols	Soils with morphological evidence of long periods of saturation promoting the generation of overland flow due to saturation excess.
Responsive shallow	Mispah, Graskop	Leptosols	Shallow soils overlying relatively impermeable bedrock. Limited storage capacity results in the generation of overland flow after rainfall events.
A/B interflow	Estcourt, Sterkspruit	Solonetz	Duplex soils where the textural discontinuity facilitates build-up of water in the topsoil, with discharge in a predominantly lateral direction.
Soil/bedrock interflow	Fernwood, Cartref	Arenosols	Soils overlying relatively impermeable bedrock. Hydromorphic properties signify temporal build of water on the soil/bedrock interface and slow discharge in a predominantly lateral direction.

WRB = World Reference Base for Soil Resources

Undisturbed core samples were collected from 12 representative diagnostic horizons within the Olifants catchment during the field survey. These core samples were used to determine bulk density, particle size distribution and the water retention characteristics. The results were combined with the already existing Land Type modal profile data, and then the required SWAT+ hydraulic parameters were obtained by averaging these properties for each hydropedological soil type (Table 4.8).

Table 4.8: The main hydraulic properties of the HYDROSOIL mapping units.

Hydrological soil types	Horizon	Hydro-group	Depth	Bd	AWC	Ksat	OC	Clay	Silt	Sand
			mm	g/cm^3	mm/mm	mm/h	%	%	%	%
Recharge (deep)	A	A	250	1.38	0.078	26.6	1.4	10.7	4.2	85.0
	B		1800	1.42	0.095	4.827	0.8	23.4	14.4	62.2
Responsive (shallow)	A	D	300	1.45	0.085	18.3	2.0	12.8	9.5	77.7
Responsive (saturated)	A	B	200	1.48	0.123	1.5	3.2	51.6	17.9	30.6
	B		1500	1.53	0.122	1.479	0.9	49.6	17.9	32.5
Interflow (A/B)	A	D	200	1.42	0.072	49.8	0.6	7.6	0.7	91.7
	B		1200	1.39	0.079	8.476	0.2	18.1	3.9	78.0
Interflow (soil/bedrock)	A	C	250	1.41	0.096	8.0	1.6	18.9	16.0	65.1
	B		1400	1.41	0.091	2.622	0.7	29.7	9.2	61.1

Bd = bulk density; AWC = Available Water Capacity; Ksat = saturated hydraulic conductivity; OC = Organic Carbon.

Two model runs were set up for the two levels of soil information. Only the soil information differed between setups as all other factors were constant for both simulation runs. However, the HYDROSOIL and Land Type soil datasets differed both spatially and in their hydraulic properties (Table 4.6, Table 4.8). These differences would therefore affect the simulations due to how different hydrological processes are simulated by the model. As the hydrologic groups differ spatially between datasets, the curve numbers and associated runoff characteristics will differ greatly between model runs.

The differences in hydraulic properties between the two levels of soil information should also affect modelling accuracy. The increased soil depth, AWCs and clay content and decreased Ksat values of the HYDROSOIL map should result in more water being stored within the soil profile for longer periods, leading to more available water for root uptake, plant growth and evapotranspiration. More antecedent moisture within the soil should also lower CN2 values, which remains one of the most sensitive parameters within the SWAT model (Wahren et al., 2016, Mengistu et al., 2019).

Accounting for streamflow reduction

The Botleng aquifer is a dolomite aquifer system present in the Olifants catchment (Pietersen et al., 2012). This aquifer is used for large-scale agricultural irrigation and domestic use within the Delmas Local Municipality. It has been estimated that 10 Ml of potable water is abstracted from three major wellfields per day which falls just outside of the catchment (Pieterson et al., 2012). Preliminary modelling results suggested that substantially decreased streamflow is measured compared to that expected from the climate of the catchment. The dolomite aquifer drives streamflow within the catchment, where the majority of streamflow arises from springs within the dolomite areas, a substantial amount of streamflow reduction should be accounted for. To account for this within the SWAT+ model, the coefficient which determines the amount of groundwater lost from the system due to deep recharge and in this case, abstraction was increased to 0.8 for both models. Therefore, the fraction of root zone percolation that reaches the deep aquifer for both models was set to 0.8, mimicking the water lost from the system.

Validation data and statistical comparison

Two weirs were used to validate long-term streamflow simulations which are managed by the DWS. These gauges were B2H008 (100 km^2) and B2H007 (330 km^2) where total daily streamflow was used for comparison purposes.

For statistical comparison, four widely used statistical indicators were employed, namely coefficient of determination (R^2), percentage bias (PBIAS), Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE). A KGE value surpassing -0.41 indicates a model prediction that aligns better with the mean observed values (Knoben et al., 2019). Streamflow simulations satisfactory when $R^2 > 0.6$, NSE > 0.5, and PBIAS ≤ 15% (Moriasi et al., 2015).

4.2.3 Results and discussion

Streamflow simulations

The model set-ups for the two levels of soil information had an identical number of subbasins (6) and landscape units (42), because the same DEM was used to delineate these. The number of HRUs differed significantly where the HYDROSOIL model contained 966 HRUs compared to the 386 HRUs contained within the Land Type model. The large discrepancy between model HRUs is purely a result of the spatial differences between the soil input information.

The simulations of the HYDROSOIL and Land Type models yielded satisfactory results at both gauges based on R^2 values (Table 4.9). All HYDROSOIL simulations achieved satisfactory KGE values (Knoben et al., 2019). However, the Land Type model did not meet the minimum KGE threshold at 100 km^2 with a value of -1.57 and yielded a weak KGE value of 0.22 at 330 km^2.

Both models produced disappointing PBIAS values at 100 km^2, where neither model achieved PBIAS values below the 15% threshold (Moriasi et al., 2015). However, the HYDROSOIL model provided more accurate PBIAS values (27.17) compared to the Land Type model (-43.2). At 330 km^2 the HYDROSOIL model yielded satisfactory PBIAS results (-0.45) compared to the unsatisfactory value (-50.29) for the Land Type model. Analysing NSE values, the HYDROSOIL model substantially outperformed the Land Type model at each catchment scale, with only the HYDROSOIL model achieving acceptable NSE values.

Table 4.9: Statistical indicators of monthly streamflow simulations at two catchment levels.

Catchment	Soil data	R^2	PBIAS	NSE	KGE
B2H008 (100 km^2)	Land Type	0.76	-43.2	-7.18	-1.57
	HYDROSOIL	0.78	27.17	0,69	0,69
B2H007 (330 km^2)	Land Type	0.87	-50.29	-1.69	0.22
	HYDROSOIL	0.74	-0.45	0,74	0.74

PBIAS = Percentage bias; NSE = Nash Sutcliffe Efficiency; KGE = Kling-Gupta Efficiency

Peak flows were drastically overestimated by the Land Type model, whereas the HYDROSOIL dataset yielded far lower peak flows than the Land Type dataset (Figure 4.9), which improved modelling accuracy. Baseflow simulations were also substantially overestimated by the Land Type model at all catchment levels. The negative PBIAS values across all model simulations also equates to the general overestimation of total streamflow values by the Land Type model which can also be attributed to the

overestimation of baseflow values across all catchment levels. SWAT+ allows users to adjust groundwater parameters to mitigate or correct baseflow values.

Statistical comparison between the two models over the five-year simulation period indicated a substantial difference between the two levels of soil information (Table 4.9). It could be inferred that the HYDROSOIL model outperformed the Land Type model, at both catchment scales (100 km^2 and 300 km^2) when comparing daily observed streamflow values. The improved modelling accuracy is presumed to be a direct result of the improved simulation of real-world hydrological processes by improving the accuracy of soil information in the model.

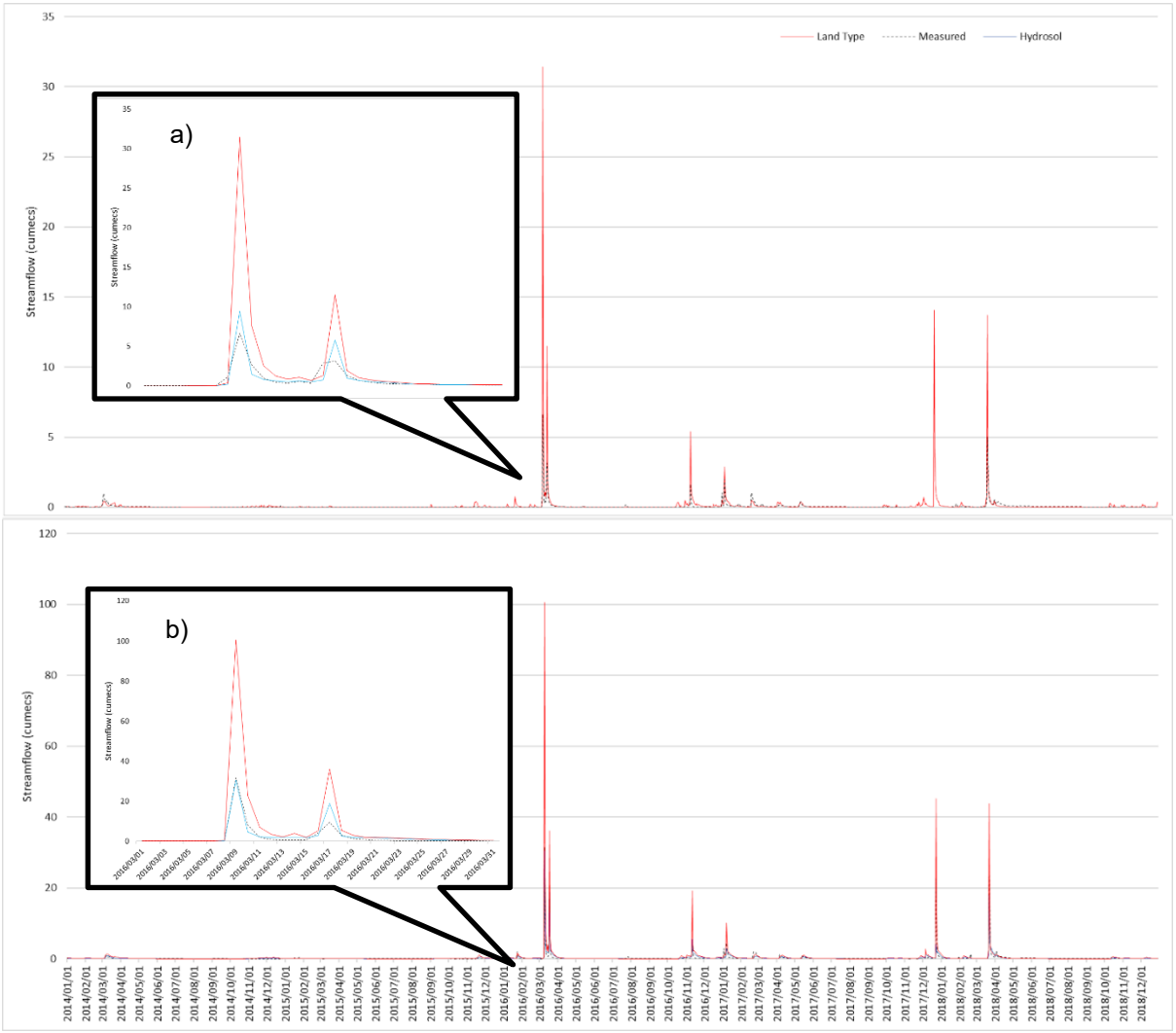

Figure 4.9: Daily simulated streamflow for the Land Type and HYDROSOIL (Hydrosol) model runs compared to observed streamflow at a) B2H008 and b) B2H007.

Hydrological processes

These differences in streamflow simulations and hydrological processes are a direct result of the differences between soil input datasets and how soil input data affects the simulation of these different hydrological processes. The major hydrological processes differed substantially between model simulations (Table 4.10) for the upper Olifants catchment (B2H007).

The upper Olifants catchment is significantly influenced by anthropogenic changes. This is very apparent when analysing the streamflow of each model as a fraction of precipitation where the HYDROSOIL streamflow equates to 20% of precipitation and Land Type streamflow equates to 16%. The low precipitation conversion rate can be explained by the substantial abstraction from agricultural and domestic practices.

Baseflow as a fraction of total flow also differs between the models, where the HYDROSOIL baseflow equates to 86% of total flow whereas the Land Type baseflow equates to 60% of total flow. The upper Olifants sits at the edge of the Botleng dolomite aquifer which is used for large-scale agricultural irrigation and domestic use. The unconfined dolomite aquifer is also therefore responsible for a substantial portion of the total flow within the catchment as can be seen by the high baseflow rates for both models. Surface runoff as a fraction of total flow also differs between both models, where the HYDROSOIL surface runoff equates to 14% of precipitation compared to the 40% of the Land Type surface runoff. This difference can be explained by the differences between soil hydro-group and soil saturated hydraulic conductivity.

Percolation and deep percolation for the HYDROSOIL model equates to 26% and 21% of precipitation, respectively. Percolation and deep percolation for the Land Type model equates to 19% and 15% of precipitation, respectively. These results can also be related to differences between soil information. Finally, evapotranspiration as a fraction of precipitation equals 72% for the HYDROSOIL model and 76% for the Land Type model.

Table 4.10: Average annual hydrological processes for the upper Olifants catchment.

Hydrological processes	HYDROSOIL	Land Type
Streamflow as a fraction of precipitation	0,2	0,16
Baseflow as a fraction of total flow	0,86	0,6
Surface Runoff as a fraction of total flow	0,14	0,4
Percolation as a fraction of precipitation	0,26	0,19
Deep recharge as a fraction of precipitation	0,21	0,15
Evapotranspiration as a fraction of precipitation	0,72	0,76

The most substantial difference between the two model simulations is the large discrepancy between surface runoff as a fraction of precipitation (Table 4.10). As all model inputs remained constant except for the soil information, it is the difference in soil information which results in these differences in surface runoff.

The HYDROSOIL model simulates substantially less surface runoff than the Land Type model, where surface runoff is primarily concentrated in the urban and wetland areas of the catchment (Figure 4.10a). This is not the case for the Land Type model where the majority of surface runoff is simulated under pivot irrigation and within urban areas (Figure 4.10b). However, the rest of the catchment also simulates substantially more surface runoff compared to the HYDROSOIL model. The higher surface runoff value is also reflected in the substantially higher peak flows generated by the Land Type model (Figure 4.9).

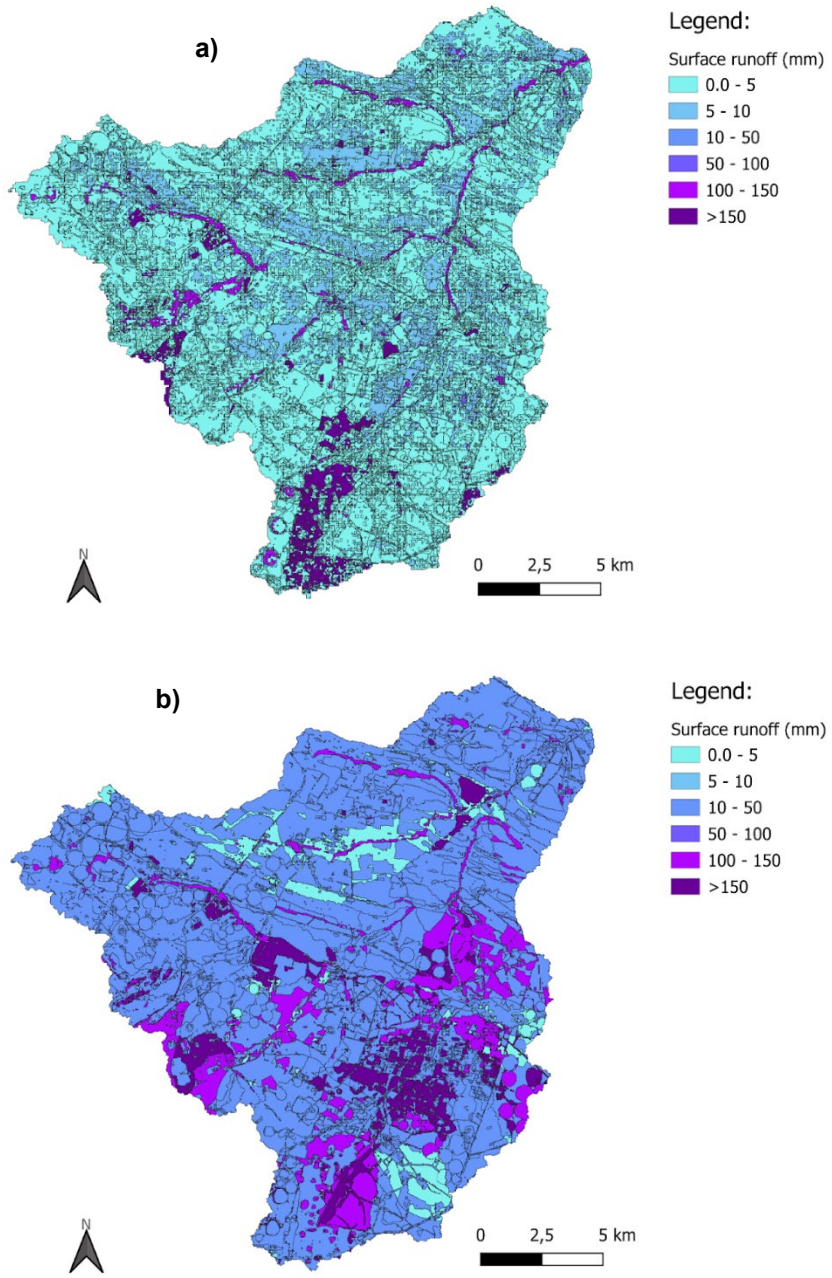

Figure 4.10: Average surface runoff (mm) for a) the HYDROSOIL and b) the Land Type dataset.

The most prevalent HYDROSOIL soil in the catchment, recharge (deep), is designated as a hydro-group A soil, which lowers the CN number and therefore leads to more infiltration and less runoff. The soils of the Land Type model are all designated as hydro-group B soils, which should indicate moderate rates of infiltration. However, the majority of these soils also contain low saturated hydraulic conductivity values (less than 5 mm/h), especially compared to the HYDROSOIL dataset which were measured in the field (more than 15 mm/h). The lower the saturated hydraulic conductivity value of the soil the less permeable the soil and the less water is allowed to infiltrate, flow laterally or percolate within the soil.

The results for the Olifants catchment agree with other research that emphasises the importance of understanding the hydropedological information available within a catchment and its transferability for

hydrological modelling purposes (Bouma et al., 2011; Sierra et al., 2018; Van Tol et al., 2021; Smit & Van Tol., 2022).

4.2.4 Conclusions

Detailed hydrological soil information for the upper Olifants catchment, developed using digital soil mapping techniques, resulted in more accurate streamflow simulations at both catchment scales in a highly anthropogenically-altered catchment. These results also illustrate the importance of soil information for simulating hydrological processes even in systems which are groundwater dominated.

CHAPTER 5: JUKSKEI RIVER CATCHMENT

Chapter 5 revisits a modelling study conducted in the Jukskei catchment (Van Tol et al., 2020), which evaluated the impact of two different levels of soil information on streamflow predictions and water balance components. Building on this model, the chapter tests the hypothesis that there is a ceiling in the benefit that can be obtained from improving soil information.

5.1 EXPLORING THE OPTIMAL LEVEL OF SPATIAL DETAIL IN SOIL INFORMATION FOR HYDROLOGICAL MODELLING

5.1.1 Introduction

The advances in digital soil mapping have paved the way for more detailed soil information. Several studies have found that using more detailed soil information improves accuracy in modelling results and reduces parameter uncertainty during calibration (Julich et al., 2012; Thompson et al., 2012; Van Tol et al., 2015; Van Zijl et al., 2016; Wahren et al., 2016; Van Zijl et al., 2020). With continuous advances in remote sensing and more detailed, readily-available ancillary data, one can expect that finer and finer scale digital soil maps will be created. The question then is, when is enough, enough? Can we achieve the same simulation accuracies and process representations without needing more detailed soil information? Some argue that small improvements in modelling accuracy do not justify the cost and time to gather more soil information (Geza & McCray, 2008). Although digital soil mapping techniques largely reduce the costs of data accumulation, it is a reasonable argument in relation to computing efficiency and realistic representation of processes. Ultimately, the modelling results should inform decision-making, and realistic-sized management units should drive the detail of modelling input data.

A previous study for the Jukskei catchment compared models created using advanced digital soil mapping techniques with soil information derived from the Land Type database. In general, the digital soil mapping resulted in more accurate simulations of streamflow than the Land Type data when compared with measured values (Van Tol et al., 2020). The improved simulation accuracy was obtained without calibration of the model, which is promising for hydrological modelling in ungauged areas where long-term streamflow monitoring for calibration is absent. The ideal level of detail (or scale) of soil information compared to catchment size remained an important question. However, the SWAT model is sensitive to soil inputs, and the spatial representation of dominant hydrological processes is captured more accurately with more detailed soil information (Van Tol et al., 2020). Therefore, a reasonable effort should be made to improve soil information to realistically reflect hydrological processes to enhance land-use planning, especially in areas dedicated to urbanisation.

This chapter tests the hypothesis that there is a ceiling in the benefit that can be obtained from improving soil information. It was tested in the Jukskei catchment using three levels of detail obtained from a digital soil mapping exercise. The digital soil mapping data was also compared against the new SWAT spatial layers and attribute data for South African soils (Le Roux et al., 2023).

5.1.2 Materials and methods

The Jukskei catchment

The Jukskei catchment spans approximately 630 km^2 and is situated between Johannesburg, the largest city in South Africa, and the capital city, Pretoria (Figure 5.1). Located in Gauteng province, this region accommodates a quarter of the country's population and significantly contributes to the majority of the gross domestic product. Given its economic importance, the area faces substantial development pressure driven by urbanisation.

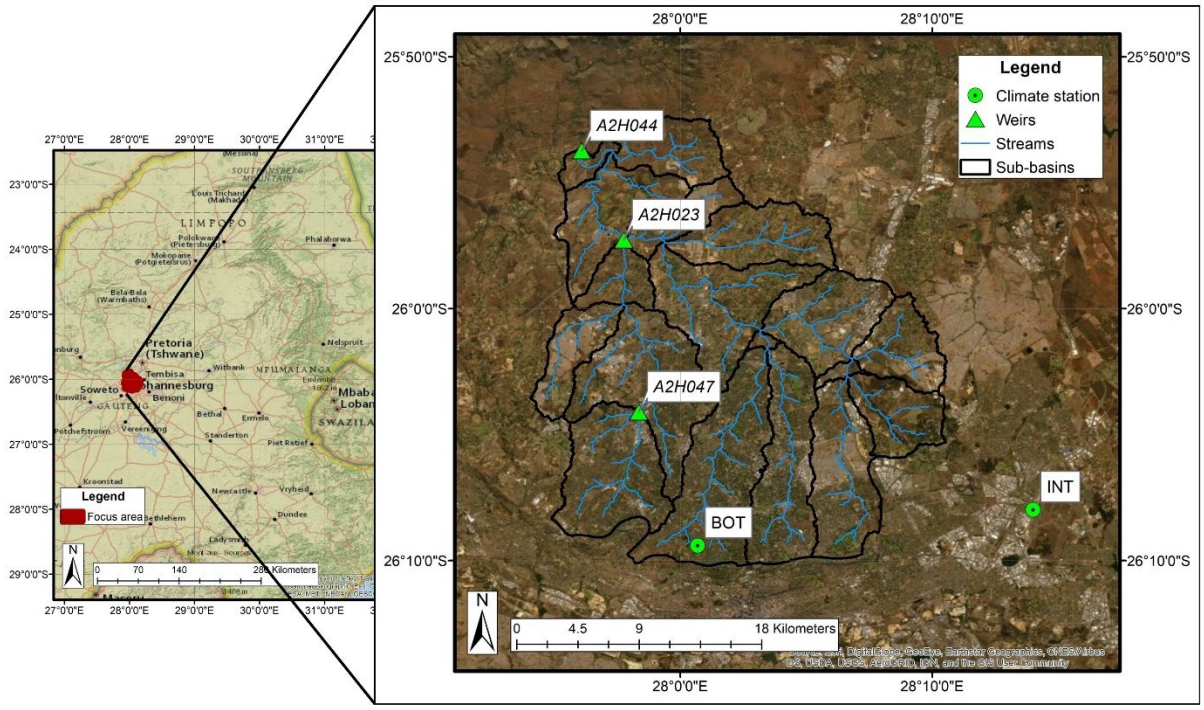

Figure 5.1: The Jukskei catchment with sub-basins, weirs and climate stations (Van Tol et al., 2020).

The Jukskei River drains the catchment in a northerly direction. The geological composition of the study site consists of granite and gneiss from the Lanseria Gneiss of the Johannesburg Dome Granite (Dippenaar & van Rooy, 2014), featuring dominant Reference Groups soils such as Leptosols, Plinthosols, Cambisols, Stagnosols, and Fluvisols (IUSS, 2015. The vegetation type is Egoli Granite Grassland, forming part of the Mesic Highveld Grassland Bioregion (SANBI, 2012). Unfortunately, more than two-thirds of this vegetation unit has undergone transformation due to urbanisation. Positioned between 1 245 and 1 709 m.a.s.l. on the Highveld of South Africa, the catchment exhibits hilly terrain, with the majority of hillslopes having an average slope of less than 5%. The climate is characterised by convectional thunderstorms during summer months (October to April), with an average annual rainfall of approximately 700 mm. Summer days are hot, reaching an average maximum temperature of around 25°C, while winter nights are cold, with an average minimum temperature of approximately 4°C.

Mapping

To create a soil map, firstly environmental covariates were collected for the entire Halfway House Granites area, as well as the Hospital Hill Subgroup and Swazian Erathem geological formations, to use as ancillary variables in the mapping process. These layers included wet and dry season Landsat 8 images (USGS, 2018) taken on 10 April 2004 (wet season) and 31 July 2004 (dry season) respectively, and the 30 m SRTM DEM (USGS, 2018). Covariate layers were resampled to have the same grid extent at a resolution of 30 m. Secondary covariate layers were derived from the Landsat 8 and DEM layers in SAGA-GIS. Terrain derivative secondary covariate layers included: slope, profile curvature, planform curvature, aspect, topographic wetness index, flow accumulation, altitude above channel network, relative slope position and multi-resolution index of valley bottom flatness. From the wet and dry season satellite images the NDVI was derived. The geological map was rasterised and resampled to fit the grid extent of the other covariate layers.

A soil observation database was created for the Jukskei catchment by combining various databases from different projects done within the greater Johannesburg area. Seventy observations were collected from Van Zijl and Bouwer (2012), 142 from Van Zijl et al. (2019), 61 from Van Zijl et al. (2020) and 113

from Bouwer et al. (2020). These observations were made by soil auger, profile pits and spot observations in areas where the soil was not deemed to be disturbed. Soils were described per horizon with soil texture, structure, colour, redox morphology, stone content and transitions being noted. Observations were classified to soil family level according to the South African Soil Classification System (Soil Classification Working Group, 1991). Samples for soil physical measurements were taken of selected soil horizons. Additionally, 48 virtual soil profiles were placed within wetlands as delineated on the wetland map of Johannesburg. Therefore, the final database consisted of 434 soil observation points (Figure 5.2). Hydropedological soil forms (Van Tol & Le Roux, 2019) were derived from the soil form classifications (Table 5.2).

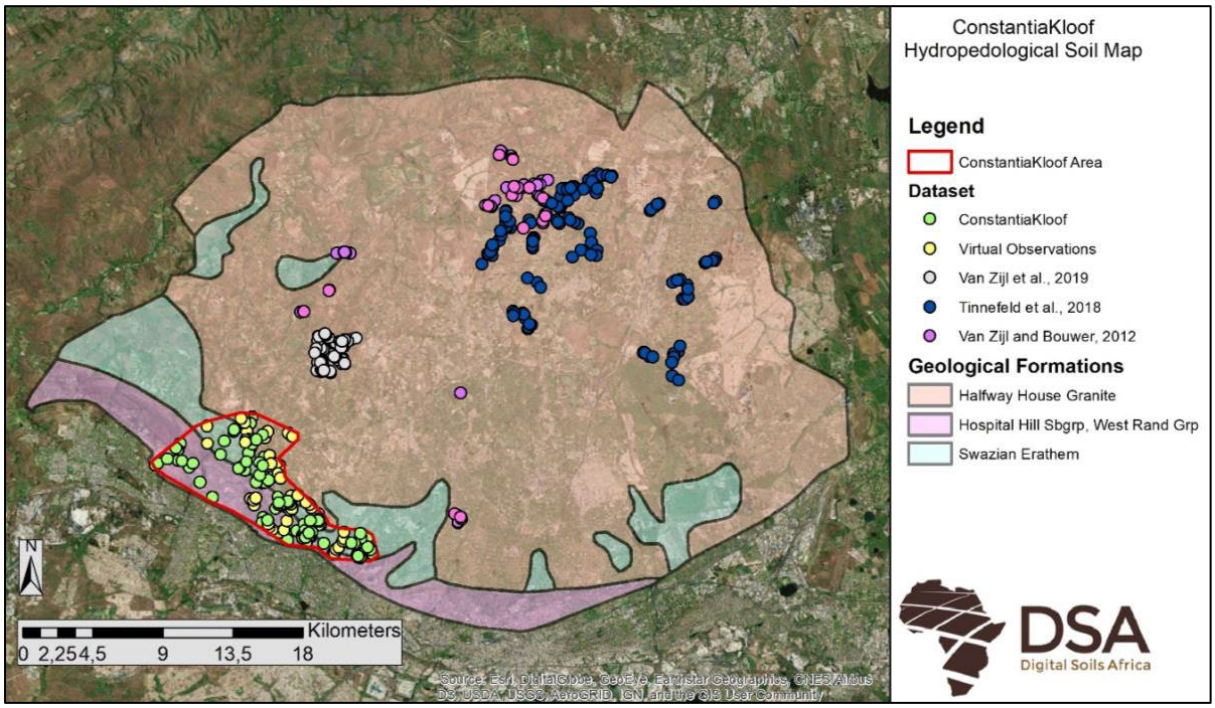

Figure 5.2: The soil observations of the three Halfway House Granites soil observation databases.

The combined soil database was divided into a training and validation dataset, by stratified random sampling, using soil form as a stratifier, with 25% of the observations points of each hydropedological soil form being included in the validation dataset. The virtual soil profiles were all included into the training dataset. The training dataset consisted of 334 observations and the validation dataset 100 observations.

Using the training dataset, environmental covariates and the multinomial logistic regression method (Kempen et al., 2009), a hydropedological soil map was created at a 30 m resolution. To decrease the number of mapping units, the map was simplified by resampling it to a larger pixel size of 100 m and 200 m. It is important to note that the application of a filter was first tried, but linear features, such as the saturated responsive soils found along streams were eliminated using a filter. All three maps (30 m, 100 m, and 200 m) were validated with the validation dataset. A one-pixel buffer was observed (Van Zijl et al., 2012).

Total validation point accuracy, user's and producer's accuracy, and the Kappa coefficient were determined for the validation dataset, to measure whether the map was an acceptable representation of reality. Total validation point accuracy is the total number of observations correctly mapped, expressed as a percentage of the total number of validation observations. The user's accuracy reflects the accuracy of the map from the user's perspective. It is the number of validation observations correctly

mapped within a specific map unit, expressed as a percentage of the total number of observations found on that specific map unit. The producer's accuracy reflects the accuracy of a map from the producer's perspective. It is the number of validation observations, within a specific class, correctly mapped, expressed as a percentage of the total number of observations within that specific class. The Kappa coefficient represents how well the map reflects reality, when compared to a random designation of mapping units. Kappa coefficient values range between 0 and 1, with values close to 0 indicating that the map is equal to a random designation and values close to 1 indicating that the map represents reality significantly better than a random designation would.

The SWAT model, model inputs and setup

The hydrological modelling utilised SWAT+ (version 2.2.3). SWAT+ represents an updated iteration of the widely recognised Soil and Water Assessment Tool (SWAT; Arnold et al., 1998; Bieger et al., 2017). SWAT is a semi-distributed catchment-scale model renowned for its process-based approach, extensively employed for simulating water quality and quantity to predict and assess the impacts of factors such as land use, climate change, soil erosion and pollution.

The modelling period spanned from January 2000 to December 2013, with the initial three years designated as a warm-up period, followed by 11 years for validation. Notably, the study's objective, of evaluating the direct contribution of improved soil information to modelling efficiency, precluded the inclusion of a calibration period.

Topography and land use

Elevation data were acquired from a 30 m SRTM DEM (USGS) (Figure 5.3a). Current land use information was extracted from the 2013-2014 South African National Land Cover Map dataset (GeoTerra Image, 2015). To align with SWAT modelling requirements, the land-cover classifications were re-grouped into specific SWAT land uses, each characterised by pre-defined parameters (Figure 5.3b).

Climate information

Daily rainfall, as well as minimum and maximum temperatures, were sourced from two climate stations: the Johannesburg Botanical Gardens and OR Tambo Airport. These climate stations are part of the South African Weather Service. Additionally, daily solar radiation, relative humidity and wind speed data were retrieved from the National Center for Environmental Prediction (Saha et al., 2015). This comprehensive set of meteorological information was employed to calculate daily potential evapotranspiration, utilising the Penman-Monteith approach.

Soil information

To implement SWAT, a comprehensive soil dataset is essential, serving as a spatial layer with detailed information on soil horizons. Key attributes, including depth, particle size distribution, saturated hydraulic conductivity, bulk density, carbon content, and available water capacity (AWC), are required for each layer. The AWC is synonymous with the more commonly known plant available water.

This study incorporates four levels of soil information. The first utilises the spatial layer and associated soil attribute data derived from a project designed to provide SWAT-ready data for South Africa (Le Roux et al., 2023). Due to its origin in the Land Type database (Land Type Survey Staff, 1972-2002), this layer is referred to as Land Type throughout this document. The Land Type database, covering the entire country at a 1:250 000 scale, categorises Land Types based on relatively homogeneous soil-

forming factors, such as climate, geology, and topography. There are only two Land Types in the Jukskei catchment (Bb1 and Bb2; Figure 5.4).

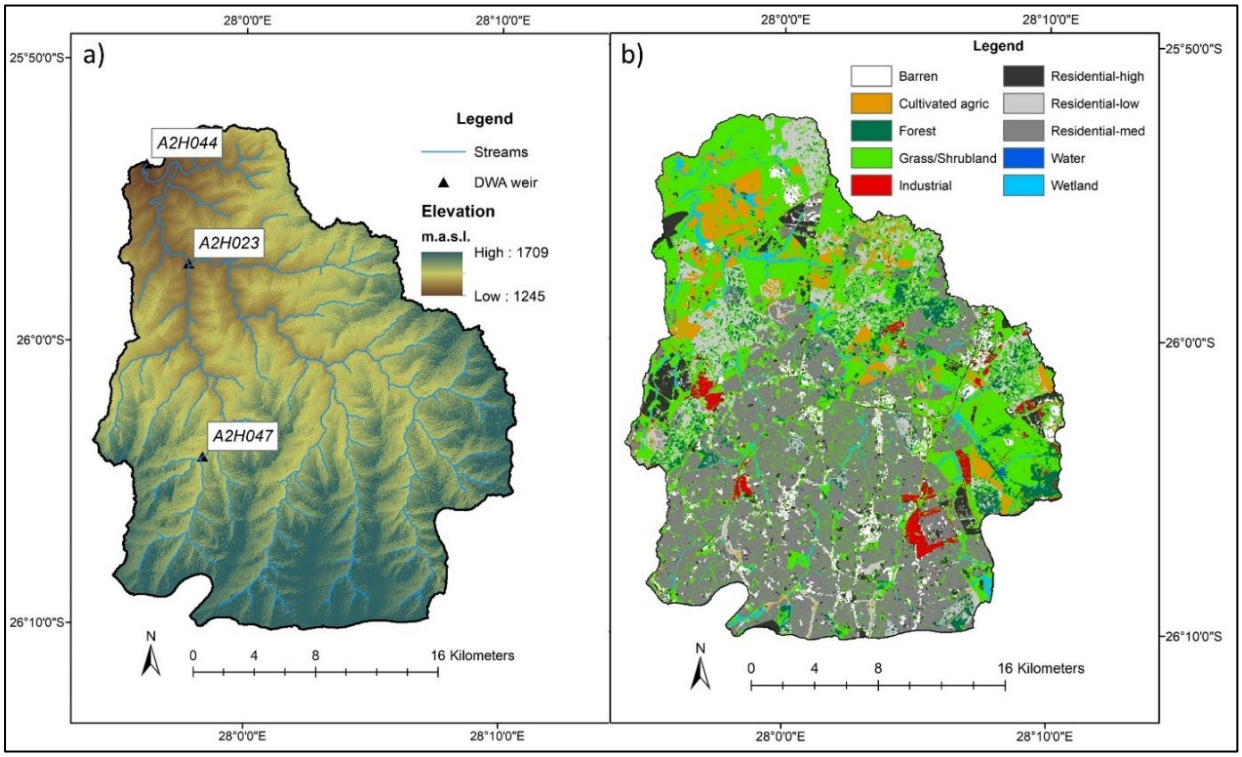

Figure 5.3: a) Elevation of the Jukskei catchment with streams and weirs, b) dominant land-use in the Jukskei catchment as obtained from the South African National Land Cover 2013-14.

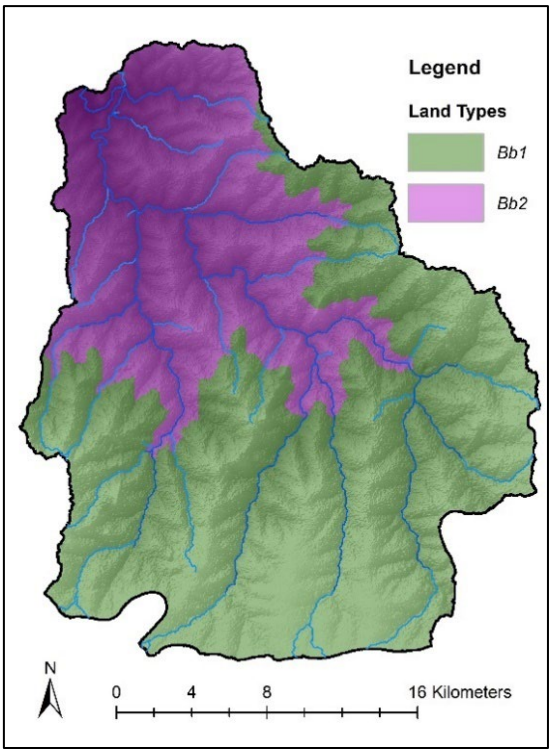

Figure 5.4: The Land Type information for the Jukskei catchment (from Le Roux et al., 2023).

In contrast to the Land Type database, digital soil maps exhibit a higher diversity of soils (six vs. two) and offer a more detailed spatial distribution of these soils. The HYDROSOIL dataset primarily features

the hydropedological classes interflow (soil/bedrock) and recharge (deep). To assess the impact of varying spatial detail, three HYDROSOIL maps, (Detail, Medium, Coarse) were utilised. Although these maps share the same map units (Table 5.1), they differ in spatial detail, resulting in varying numbers of HRUs. The Land Type dataset, with only two soil types, yielded 4 826 HRUs. In contrast, the Coarse, Medium, and Detail datasets generated 11 400, 13 844, and 33 196 HRUs, respectively.

Validation data and statistical comparison

Streamflow data were collected at three weirs managed by the DWS within the catchment (Figure 5.3). This study specifically focuses on the entire Jukskei catchment, covering 630 km^2, drained by the A2H044 outlet. Daily streamflow measurements were transformed into monthly average values for the sake of comparison.

For statistical assessments, three widely recognised indices were employed: the coefficient of determination (R^2), the Root Mean Square Error (RMSE), the NSE, and the KGE. Beyond these statistical measures, a comparative analysis of water balance components across different model runs was conducted to evaluate the influence of soil information on the modelling outcomes.

Table 5.1: Selected hydraulic properties of the soil horizons in different soil information datasets (Land Type from Le Roux et al., 2023 and HYDROSOIL from Van Tol et al., 2020).

	Soil Group	Hydro-Group	Horizon	Depth	Db	AWC	Ks	OC	Clay	Silt	Sand
				mm	g.c m^{-3}	mm. mm^{-1}	$mm.h^{-1}$			%	
Land Type	Bb1	B	A	300	1.4	0.09	13	1.0	15.0	15.0	70.0
			B	660	1.5	0.09	210	0.0	15.0	15.0	70.0
	Bb2	B	A	300	1.4	0.09	13	1.0	15.0	15.0	70.0
			B	660	1.5	0.09	210	0.0	15.0	15.0	70.0
HYDROSOIL	Recharge (deep)	A	A	300	1.4	0.09	218.5	1.2	21.6	11.1	67.6
			B	1200	1.3	0.09	172.0	0.8	29.7	13.2	57.2
			C	1500	1.4	0.08	56.9	0.4	27.1	15.7	57.6
	Recharge (shallow)	A	A	300	1.4	0.12	218.5	1.6	21.6	11.1	67.6
	Interflow (A/B)	C	A	300	1.4	0.06	112.5	1.8	21.6	11.1	67.6
			E	600	1.3	0.09	87.5	0.6	29.1	14.7	56.6
			B	1200	1.4	0.08	2.0	0.5	46.2	14.2	39.7
	Interflow (soil/bedrock)	B	A	300	1.4	0.13	218.5	1.8	21.6	11.1	67.6
			B	800	1.3	0.07	172.0	0.8	29.1	14.7	56.6
			C	1000	1.5	0.06	15.0	0.4	46.2	14.2	39.7
			R	1500	1.8	0.06	0.1	0.0	46.2	14.2	39.7
	Responsive (wet)	D	A	300	1.4	0.06	10.2	2.1	21.6	11.1	67.6
			G	1000	1.2	0.07	5.0	0.9	52.8	19.6	27.6
			G2	1300	1.6	0.06	0.1	0.4	52.8	19.6	27.6
	Responsive (shallow)	C	A	300	1.4	0.13	10.2	1.8	21.6	11.1	67.6
			R	500	1.8	0.07	1.0	0.0	46.2	14.2	39.7

Db = bulk density; AWC = Available Water Capacity; Ks = saturated hydraulic conductivity; OC = Organic Carbon.

5.1.3 Results and discussion

Digital soil map results

As could be expected, the 30 m pixel map achieved a higher accuracy than the resampled maps (Figure 5.5). Its validation point accuracy was 65%, and it had a Kappa value of 0.53, while the 100 m and 200 m pixel maps only achieved validation point accuracies of 52% (both) and kappa values of 0.34 and 0.35 respectively. The 30 m pixel map is therefore deemed to have a moderate agreement with reality, while the other two only being deemed to have a fair agreement with reality.

Table 5.2: Hydropedological soil types used in the HYDROSOIL data, their dominant characteristics and reference groups.

Hydropedological soil type[1]	Soil forms[2]	Reference Groups[3]	Defining characteristic
Recharge (deep)	Clovelly, Constantia, Griffen, Hutton, Shortlands	Acrisols, Nitisols	Soil profiles showing no signs of wetness in the profile. Fast vertical drainage through and out of the profile is dominant.
Recharge (shallow)	Mispah, Glenrosa, Mayo	Leptosols	Shallow soils with chromic colours in the topsoil. Underlying bedrock is permeable and drainage out of profile dominant.
Interflow (A/B)	Kroonstad, Longlands, Sterkspruit, Wasbank	Stagnosols, Planosols, Plinthosols	Hydromorphic properties between top and subsoil signify periodic saturation. Typically duplex soils with textural discontinuity between top and subsoil, resulting in a perched water table at A/B horizon interface and interflow.
Interflow (soil/bedrock)	Avalon, Bainsvlei, Bloemdal, Dresden, Fernwood, Glencoe, Pinedene, Tukulu, Westleigh	Acrisols, Stagnosols, Arenosols, Plinthosols, Stagnosols	Hydromorphic properties at the soil/bedrock interface indicate saturation due to relatively impermeable bedrock. Perched water table at bedrock interface will result in interflow at soil/bedrock interface.
Responsive (wet)	Katspruit, Rensburg	Gleysols	Gleyed subsoils indicate long periods of saturation, typical of wetland soils. Soils will respond quickly to rain event and promote overland flow due to saturation excess.
Responsive (shallow)	Mispah, Glenrosa	Leptosols	Shallow soils with bleached colours in the topsoil indicate that underlying bedrock is slowly permeable. Small storage capacity of the soil will quickly be exceeded following rainstorms and promote overland flow generation.

[1]Van Tol & Le Roux, 2019; [2]Soil Classification Working Group, 2018 [3]IUSS WRB, 2015.

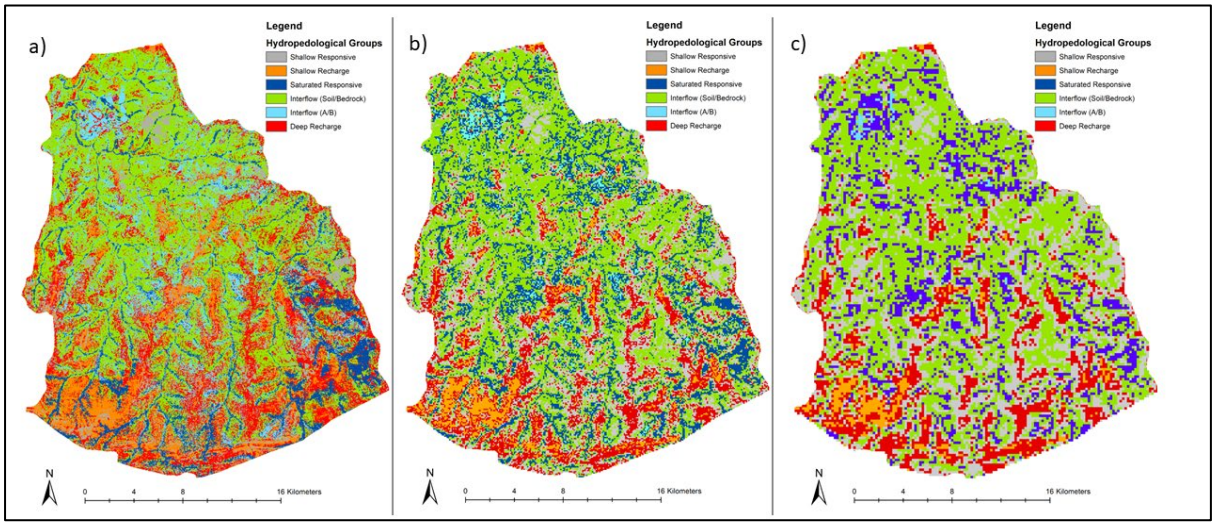

Figure 5.5: Digital Soil Mapping derived input data a) DSM_detail, b) DSM_Medium, resampled to 100 m grid and c) DSM_coarse, resampled to a 200 m grid.

Modelling results

The HYDROSOIL map yielded somewhat improved streamflow simulations compared to the Land Type dataset (Figure 5.66; Table 5.3). Despite this enhancement, the NSE values remain below the generally accepted threshold of 0.5 for both model runs (Moriasi et al., 2007). However, the KGE values surpass the 0.5 threshold, indicating an acceptable level of performance. It is noteworthy that both models exhibit a tendency to underestimate streamflow, as evidenced by positive PBIAS values.

The underestimation is more pronounced in the case of the HYDROSOIL dataset, likely attributed to its tendency to underestimate both peak flows and baseflows. Both Land Type and HYDROSOIL model runs reveal a substantial underestimation of baseflow (Figure 5.66). Interestingly, a visual inspection suggests that the HYDROSOIL dataset outperforms the Land Type dataset in predicting baseflow, although improvements are still needed.

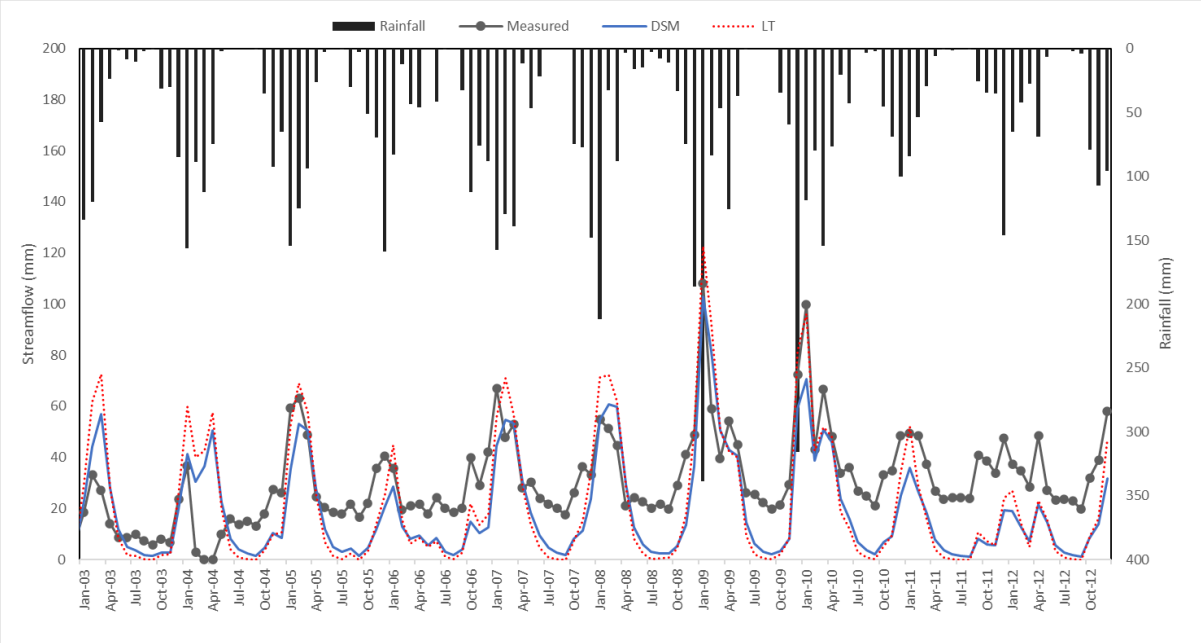

Figure 5.6: Simulated streamflow for the Land Type (LT) and detailed HYDROSOIL (DSM) model runs compared to observed streamflow.

The results demonstrate marked progress in comparison to earlier simulations (Van Tol et al. 2020). The earlier KGE values were 0 for the Land Type dataset and 0.28 for the HYDROSOIL map, indicating significant deficiencies (Van Tol et al., 2020). In contrast, these updated simulations exhibit noticeable improvements for both Land Type and HYDROSOIL scenarios (Table 5.3). It is crucial to note that these enhancements may stem from updates to the model, modifications to default parameters, and alterations to the overall model structure, rather than solely relying on changes in the soil dataset.

Table 5.3: Statistical streamflow prediction accuracies when using HYDROSOIL as input compared to Land Type.

Soil level	R^2	RMSE	PBIAS	NSE	KGE
HYDROSOIL	0.64	16.61	39.64	0.39	0.54
Land Type	0.65	17.55	32.56	0.32	0.43

RMSE = Root Mean Square Error; PBIAS = Percentage bias; NSE = Nash Sutcliffe Efficiency; KGE = Kling-Gupta Efficiency

While the simulations fall short of the ideal NSE threshold, the acceptable KGE values and notable improvements over previous studies underscore the positive impact of model updates and structural modifications. Further refinement, especially in addressing underestimation issues, remains a pertinent focus for future enhancements in streamflow simulations.

Examining the impact of different levels of spatial detail on water balance components revealed interesting findings, particularly when comparing the Detail and Medium simulation outputs. Despite reducing the number of HRUs by over 20 000, the water balance components showed only minor changes, maintaining consistent representation of processes with similar volumes of water (Table 5.4; Figure 5.7).

Table 5.4: Water balance component estimates (mm) when using various scale HYDROSOIL inputs and the Land Type dataset. Differences are expressed as % change from the Detail model run.

	Detail	Medium		Coarse		Land Type	
Component			% change		% change		% change
Rainfall	664.2	664.2		664.0		663.9	
Streamflow	210.7	211.5	0.4	233.4	10.8	234.6	11.3
Overland flow	60.1	59.9	-0.3	58.9	-2.1	56.7	-5.7
Lateral flow	150.6	151.6	0.6	174.5	15.9	177.9	18.1
Percolation	17.3	15.1	-12.7	16.5	-4.7	38.2	120.1
ET	430.7	431.6	0.2	415.0	-3.6	392.3	-8.9
Transpiration	125.4	127.3	1.5	96.7	-22.8	99.0	-21.0
Evaporation	295.7	294.6	-0.4	311.3	5.3	284.3	-3.9
ET0	1760.5	1760.5		1760.6		1760.6	
Profile water	106.0	105.2	-0.7	115.0	8.6	40.0	-62.3
Topsoil water	17.5	18.0	2.8	16.8	-4.2	10.9	-38.0

ET = Evapotranspiration

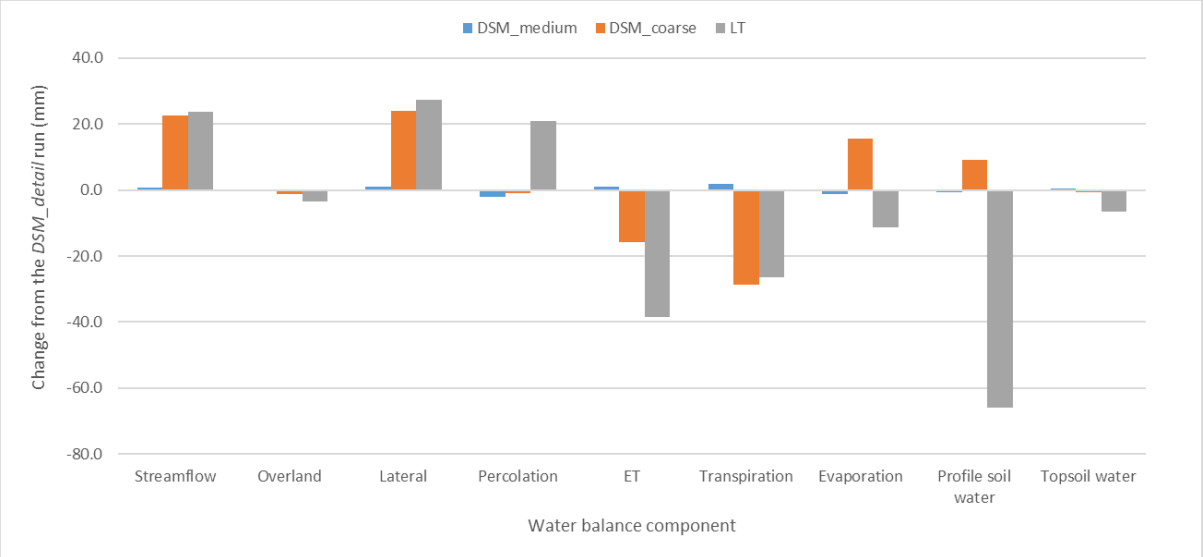

Figure 5.7: Absolute change (mm) between various water balance components from different model runs.

In contrast, a more significant difference was observed between the Detail and Coarse assessments of the water balance (Table 5.4). Simulations using the Coarse dataset exhibited comparable trends to the Land Type dataset, demonstrating increased streamflow due to heightened lateral flows. Transpiration also decreased in a similar manner as in the Land Type inputs. Interestingly, evaporation and profile soil water content increased with the Coarse dataset, whereas they declined when using the Land Type dataset. The most substantial difference between the HYDROSOIL and Land Type datasets was observed in the average profile soil water. Shallower soils in the Land Type dataset led to increased lateral flow, higher percolation, reduced evapotranspiration and altered soil water storage dynamics.

At least for this specific catchment, there exists a critical threshold where soil data becomes too coarse to provide meaningful insights (Figure 5.7; Table 5.4). It appears that, at a grid resolution exceeding 100 m, the spatial processes lose accuracy. Potential explanations include the inadequate coverage of significant soil types or the diminishing spatial connectivity at this coarse resolution. These nuances should be the focus of subsequent investigations, guiding future advancements in understanding the intricate relationship between spatial detail and the accurate representation of water balance components.

5.1.4 Conclusions

This study revisited a hydrological modelling assessment conducted in the Jukskei catchment, Gauteng province, to scrutinise the impact of varying levels of spatial detail in soil information on streamflow predictions and water balance components. Building on the work of Van Tol et al. (2020), three levels of soil detail were obtained from a digital soil mapping exercise and compared against the Land Type dataset. The investigation sought to address the crucial question of whether finer-scale digital soil maps, driven by advancements in remote sensing and ancillary data availability, offer substantial benefits in terms of modelling accuracy compared to coarser representations.

The results revealed that, despite reducing the number of HRUs by more than 20 000 units, the differences in water balance components between the Detail and Medium simulations were negligible. This suggests that, within certain limits, the spatial processes and volumes were adequately captured even with a reduction in spatial detail. However, significant disparities emerged when comparing the Detail and Coarse simulations. Notably, the Coarse dataset exhibited similarities to the Land Type dataset in terms of streamflow trends and transpiration declines, suggesting that there might be a threshold beyond which coarser spatial resolutions compromise the accuracy of hydrological simulations.

The comparison of HYDROSOIL and Land Type datasets, along with the observed trends in water balance components, hints at a critical threshold of detail in soil information for robust hydrological modelling. Beyond a grid resolution of 100 m, spatial processes may not be accurately reflected, raising questions about the effectiveness of soil information at coarser resolutions.

These results contribute to the ongoing discourse on balancing the benefits of detailed soil information against computational efficiency and realistic representation of hydrological processes. While the simulations show improvements over previous studies and highlight the sensitivity of the SWAT model to soil inputs, the study suggests that there might be diminishing returns in terms of modelling accuracy with excessively detailed soil data. This prompts further investigation into the specific conditions under which coarser resolutions become less effective, potentially due to the inadequate representation of soil types or diminishing spatial connectivity.

In conclusion, a nuanced approach is needed to determine the optimal level of spatial detail in soil information for hydrological modelling. Future research should delve into the intricate relationship between grid resolution, spatial processes and modelling accuracy to provide valuable insights for land-use planning, especially in regions undergoing rapid urbanisation.

CHAPTER 6: UMNGENI, TSITSA AND GOUKOU CATCHMENTS

Chapter 6 describes the general methodology for the digital soil mapping in the uMngeni, Tsitsa and Goukou catchments. They are grouped in this chapter to avoid repetition, as the methodology is so similar.

The value of the HYDROSOIL in the uMngeni catchment (Section 6.2) is shown through simulating land-use change and its effect on the water balance when all the grasslands are converted into forests. Willie Cloete was responsible for creating the HYDROSOIL, while Johan van Tol did the hydrological modelling.

The Tsitsa catchment, situated within the Eastern Cape province is severely eroded (Du Plessis et al., 2020). The HYDROSOIL was used for hydrological as well as sediment modelling, to showcase its use. Willie Cloete was responsible for the mapping while Jay le Roux conducted the modelling.

The creation of the HYDROSOIL and hydrological modelling of the Goukou River is showcased through the use of the JAMS model, to show that different hydrological models could also use the HYDROSOIL. The soil mapping was done by Molebaleng Sehlapelo, while Willem de Clerq and Andrew Watson did the modelling.

6.1 METHODOLOGY

6.1.1 Digital Soil Mapping methodology

Sentinel 2A satellite images were retrieved from the Sentinel Hub at a resolution of 10 m for each catchment. The satellite images were collected for the dry season and wet season at 0% cloud coverage. The spectral bands (including red, green, blue and near-infrared) were used to calculate spectral indices: normalised difference vegetation index (NDVI), colouration Index (CI), redness Index (RI), saturation Index (SI) and brightness Index (BI) (see Table 3.2). The 30 m SRTM DEM (USGS) was used to derive various topographic variables, including slope, plan curvature, profile curvature, convergence index, closed depression, total catchment area, topographic wetness index, LS-factor, channel network base level, channel network distance, valley depth, relative slope position, Multi-Resolution Index of Valley Bottom Flatness (MRVBF and Multi-Resolution Index of Ridge Top Flatness (MRRTF). Climatology data were obtained from the *South African Atlas of Climatology and Agrohydrology* (Schulze, 2007) including annual median rainfall, maximum temperature (January) and minimum temperature (June). Additionally, geological maps were also used as covariates (Council for Geoscience, 2007). Land type data (Land Type Survey Staff, 1972-2002) were also used as a covariate and retrieved from the land type database (Land Type Survey Staff, 1972-2002).

Soil point data was collected, which included legacy soil data, recently sampled soil data and digitised soil data. The legacy soil data was collected from different institutions, scientists and entities, while recently sampled data was collected during 2022. Each soil observation had a hydropedological soil form assigned to it based on its classification (Van Tol & Le Roux, 2019).

The soil observation data was split into training and validation datasets in the ratio of 75:25, using all three of the Stratified Random Sampling (SRS), K-means clustering (K-means) and Conditioned Latin Hypercube sampling algorithms. This allowed for the algorithms to be compared.

Using the training datasets, a HYDROSOIL map was created for each training dataset with the multinomial logistic regression algorithm. The maps were tested using the independent validation

datasets. The validation point accuracy and the Kappa coefficient was calculated for each map. The Kappa coefficient indicates the maps' representation of the reality, above a random representation, with values ≤ 0 indicating no agreement, 0.01-0.20 as none to slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial and 0.81-1.00 as almost perfect agreement.

6.1.2 Hydrological modelling

The next step is the application of the HYDROSOIL maps in hydrological modelling, which allows appraisal of their performance. More specifically, performance of the HYDROSOIL maps is determined by comparing the streamflow and sediment yield results from hydrological modelling with the HYDROSOIL maps, as well as a soil map (Le Roux et al., 2022) derived from the Land Type Database of South Africa (ARC, 2012). Comparing the streamflow results and accuracies of the two datasets (HYDROSOIL data versus Land Type data) allows appraisal of the performance of the HYDROSOIL data.

Sections 6.2 and 6.3 make use of the Soil and Water Assessment Tool (SWAT) model, while Section 6.4 uses the Jena Adaptable Modelling System (JAMS) model. The SWAT model is described here.

SWAT is a catchment-scale and continuous time model operating on a daily time-step to simulate water, sediment and chemical fluxes in large catchments with varying climatic conditions, soil properties, stream channel characteristics, land use and management practices (Srinivasan et al., 1998; Arnold et al., 2012). SWAT considers most hydrological and sedimentological aspects into one simulation package, including factors controlling runoff on hillslopes and streamflow in river channels, as well as sediment generation, channel transport and deposition into sinks (Gassman et al., 2007).

The SWAT model has graphical user interface applications that streamline access to databases and facilitate the preparation of input datasets including topography, drainage network, land cover, soil, climate and land management. SWAT is routinely coupled within GIS platforms which offer unprecedented flexibility in the representation and organisation of spatial data (Chen & Mackay, 2004). Although SWAT and its baseline input datasets were developed for use in the United States of America, the model has gained international acceptance and has been applied to support various large catchment (10-10 000 km^2) modelling studies across the world (e.g. Mishra et al., 2007; Wang et al., 2009; Srinivasan et al., 2010; Gassman et al., 2014).

SWAT+ represents a comprehensive revision of the well-established SWAT (Bieger et al., 2017; Arnold et al., 1998). Widely employed for water quality and quantity simulations, SWAT is instrumental in predicting and evaluating the impacts of land use changes, climate variations, soil erosion, and pollution. Neitsch et al. (2009) provide an in-depth description of the SWAT model, while Bieger et al. (2017) outlines the modifications and updates introduced in the SWAT+ version.

Operating as a process-based semi-distributed catchment-scale model, SWAT+ begins by subdividing the catchment into Hydrological Response Units (HRUs), each representing a homogeneous area in terms of soils, land use, and slope. The model then computes water balance components, encompassing overland flow, infiltration, lateral flow, percolation, evapotranspiration, and discharge to the stream from each HRU. The hydrologic component is based on the water balance equation in the soil profile integrating several processes, including surface runoff volume using the infiltration method (Green & Ampt, 1911) or the curve number method (USDA, 1972).

6.2 DIGITAL SOIL MAPPING AND MODELLING OF THE UMNGENI CATCHMENT

6.2.1 Introduction

Advancements in computational capabilities have empowered spatially distributed hydrological models to handle landscape heterogeneity intricacies. Models like SWAT and SWAT+, seamlessly integrated with GIS interfaces such as ArcMap and QGIS, utilise topography, land use, and soil data layers to delineate HRUs (Arnold et al., 1998; Bieger et al., 2017). While remote sensing has improved topographical and land-use data globally, the availability of detailed soil information in many developing countries remains a challenge. Despite evidence suggesting that more realistic hydraulic properties enhance modelling accuracy and reduce parameter calibration uncertainty, comprehensive soil data is often lacking (Romanowicz et al., 2005; Bossa et al., 2012; Diek et al., 2014; Van Tol et al., 2015; Wahren et al., 2016; Gagkas et al., 2021; Van Tol et al., 2021).

The scarcity of suitable soil information can be attributed to the fact that soil maps are not typically designed for hydrological modelling purposes (Zhu & Mackay, 2001). Additionally, the costs and time associated with quantifying spatial variations in crucial soil hydraulic properties further hinder data availability. In South Africa, the Land Type database, a 30-year initiative primarily for agricultural purposes, offers countrywide soil coverage, albeit at a limited scale of 1:250 000. Despite efforts to convert these Land Types into hydrological modelling inputs, using lumped average soil parameters (Pike & Schulze, 1995; Schulze et al., 2007; Le Roux et al., 2023), it has inherent limitations (Van Tol & Van Zijl, 2020). One of the objectives of this report is to evaluate whether these limitations have significant impacts on regional scale modelling.

Recent developments in digital soil mapping have facilitated the generation of detailed soil information at an appropriate scale and format for hydrological modelling studies at relatively low costs (McBratney et al., 2003; Zhu & Mackay, 2001; Thompson et al., 2012; Van Tol et al., 2015; Van Zijl et al., 2016; Wahren et al., 2016; Van Zijl et al., 2020). Digital soil mapping allows for the remapping of legacy soil data, like the Land Type database, at finer scales with improved accuracy, addressing some of the limitations associated with existing datasets. Notably, advancements in machine learning, expert knowledge and the disaggregation of Land Types into detailed soil polygons have been instrumental in mapping soils for hydrological purposes in South Africa (Van Zijl, 2019; Van Zijl et al., 2016; Van Tol et al., 2020).

This section revisits hydrological modelling with SWAT+ in the uMngeni catchment using distinct soil datasets (Van Tol & Van Zijl, 2022). The objectives are to create improved HYDROSOIL datasets using various methods to determine the most effective approach. Additionally, the impact of different soil input datasets will be assessed with a newer version of the SWAT model, incorporating two distinct soil datasets, namely, a HYDROSOIL map and Land Type soil inputs (Le Roux et al., 2023). Furthermore, the simulated impact of potential land-use changes on hydrological processes is evaluated when different soil inputs are employed.

6.2.2 Materials and methods

The uMngeni catchment

The study focused on three quaternary catchments located in the KwaZulu-Natal midlands: U20A (upper uMngeni River), U20B (Lions River), and U20D (Karkloof River) (Figure 6.1). These catchment areas encompass 299 (U20A), 358 (U20B) and 339 km^2 (U20D). The average annual precipitation ranges from 1 250 mm per annum in U20D to 850 mm per annum in the drier central areas of U20B, with most of the rainfall occurring between October and March (Schulze & Lynch, 2007). Summer and

winter mean daily air temperatures are approximately 19°C and 11°C, respectively (Schulze & Maharaj, 2007). The natural vegetation includes Midlands Grassland, Drakensberg Foothill Moist Grassland, and Southern Mistbelt Forests (SANBI, 2018). The predominant current land uses are commercial forestry and crop production (Figure 6.2).

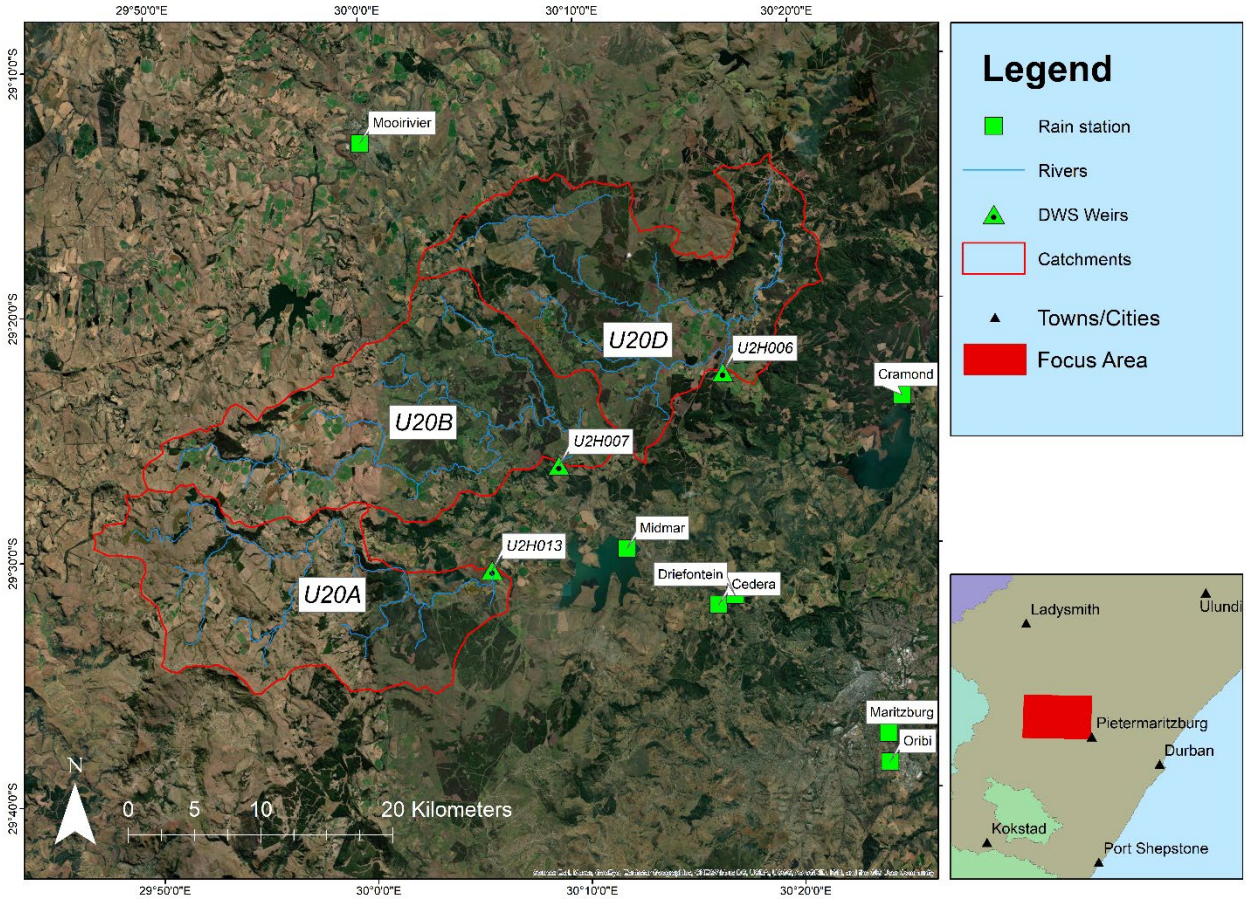

Figure 6.1:The uMngeni catchment, represented by catchments U20A, U20B and U20D, together with the location of rainfall stations and weirs draining the catchments (Van Tol & Van Zijl, 2022).

Digital soil mapping

For a description of the digital soil mapping methodology, please see Section 6.1.

Model, simulations and input data

The simulations used the SWAT+ model (v 2.2.3) (Section 6.1.2; Arnold et al., 1998; Bieger et al., 2017).

Land-cover data were acquired from the 2013/14 South African National Land Cover map (GeoTerra Image, 2015; Figure 6.2). Pre-defined SWAT values for various land-use classes served as input data for land cover. Dams identified from the land cover were included in the model setup as 'reservoirs' with estimated parameters, limited to relatively large dams (> 1 ha), amounting to 3 (U20A), 2 (U20B), and 3 (U20D). Smaller ponds and farm dams were assigned default SWAT+ parameters for a 'water' land use class in the model.

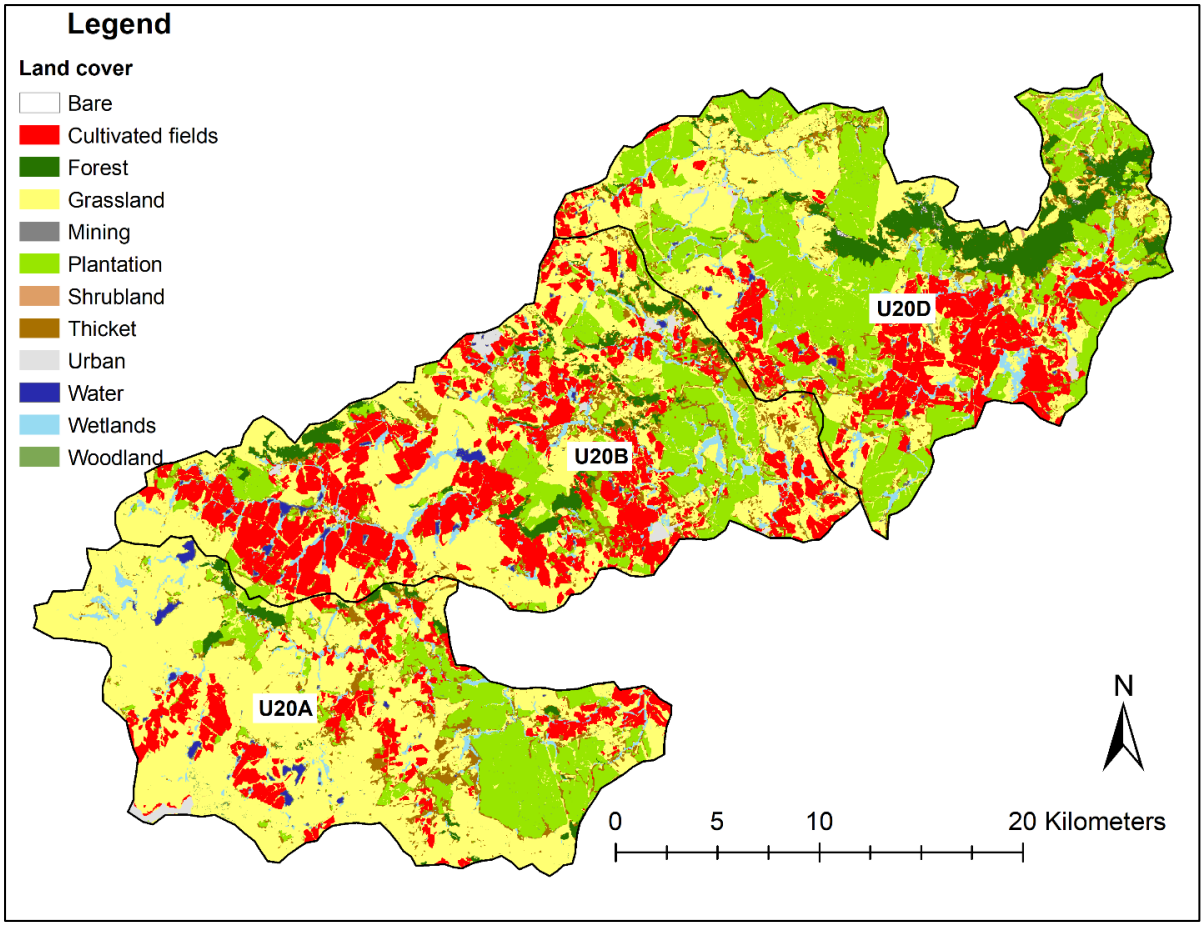

Figure 6.2: Land cover of the uMngeni catchment, simplified from the 2013/14 South African National Land Cover dataset (GeoTerra Image, 2015; adopted from Van Tol & Van Zijl, 2022).

The soil inputs were the HYDROSOIL map with the best accuracy (see results, Section 6.1) and the Land Type data (Figure 6.3) converted to SWAT ready input data (Le Roux et al., 2023). There are 91 Land Types in the three catchments each with their own input attributes (Table 6.1). The hydraulic input parameters were adopted from Van Tol & Van Zijl (2022) with alterations to accommodate two soil types namely responsive (shallow) and Stagnant which did not form part of the previous modelling exercise (Table 6.1).

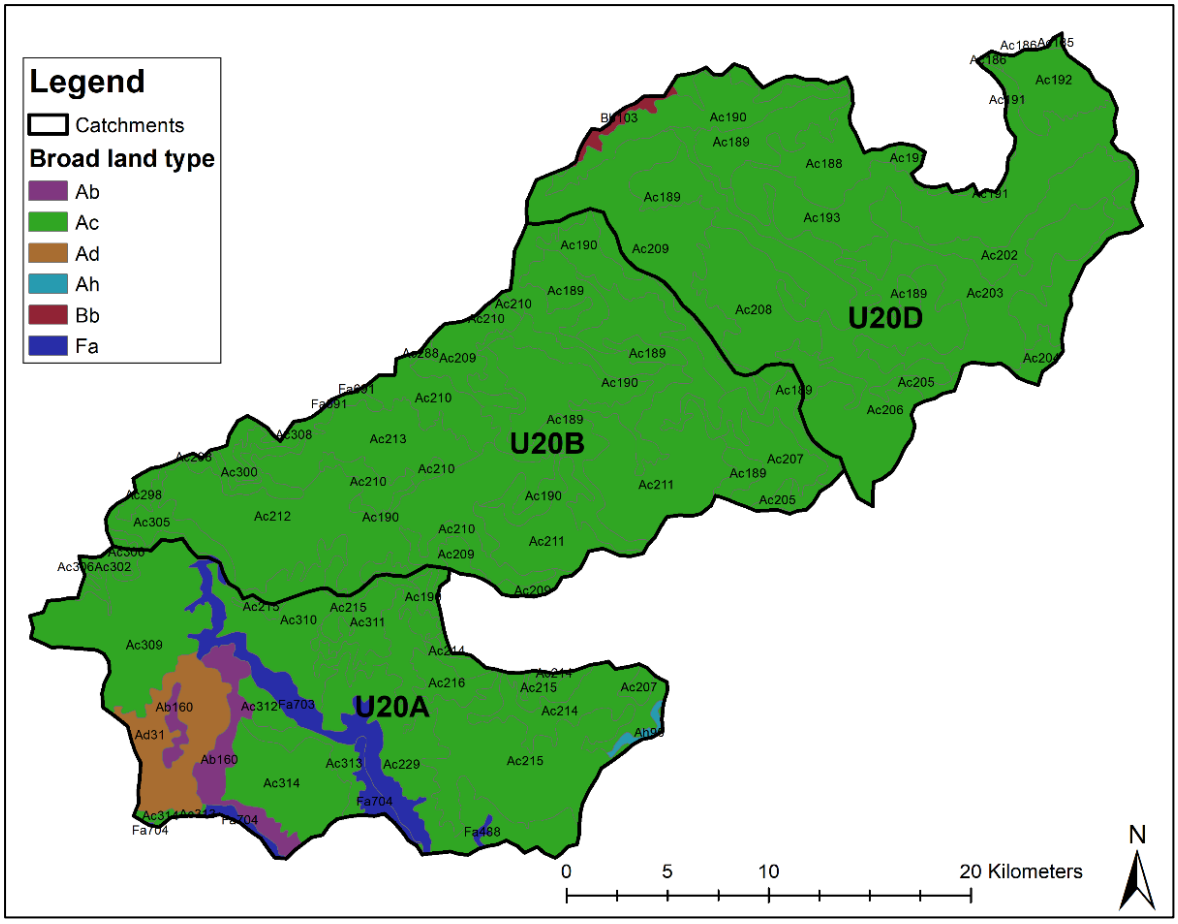

Figure 6.3: Land types present in the uMngeni catchments (Land Type Survey Staff, 1972-2002).

Table 6.1: Summary of hydraulic input parameters for the Land Type soil dataset (Le Roux et al., 2023; Van Tol & Van Zijl, 2022)

	Master horizon	Depth	Bulk density	AWC	Clay	Silt	Sand	OC	Ks
		mm	g.cm^{-3}	mm.mm^{-1}	%				mm.h^{-1}
U20D	A	300 (300, 300)	1.52 (1.49, 1.59)	0.092 (0.071, 0.105)	39.1	30.9	30.0	5.0	24.0
	B	580 (400, 710)	1.53 (1.51, 1.57)	0.091 (0.059, 0.122)	48.4	29.3	22.0	1.5	6.0
U20B	A	300 (290, 300)	1.54 (1.49, 1.60)	0.092 (0.084, 0.105)	39.1	30.9	30.0	5.0	24.0
	B	520 (320, 790)	1.55 (1.52, 1.58)	0.087 (0.066, 0.122)	48.4	29.3	22.0	1.5	6.0
U20A	A	300 (290, 300)	1.55 (1.49, 1.60)	0.102 (0.084, 0.105)	39.1	30.9	30.0	5.0	24.0
	B	530 (430, 790)	1.55 (1.52, 1.58)	0.114 (0.066, 0.122)	48.4	29.3	22.0	1.5	6.0

AWC = Available Water Capacity; OC = Organic Carbon; Ks = Saturated hydraulic conductivity

Table 6.2: Hydraulic input parameters for the HYDROSOIL and Land Type model runs.

Hydro-pedological & SWAT soil group	Master horizon	Depth	OC	Clay	Silt	Sand	Bulk density	AWC	Ks
		(mm)	%				$g.cm^{-3}$	$mm.mm^{-1}$	$mm.h^{-1}$
Recharge (deep) A	A	300	6.77	33.83	36.50	22.35	1.15	0.16	17.94
	B	1500	1.14	42.29	30.83	25.18	1.50	0.17	7.71
	C[4]	3000	0.24	35.30	34.98	28.68	1.50	0.18	3.75
Recharge (shallow) A	A	300	6.77	33.83	36.50	22.35	1.15	0.16	17.94
	C	700	0.24	35.30	34.98	28.68	1.50	0.18	3.75
Interflow (deep) B	A	300	6.77	33.83	36.50	22.35	1.15	0.16	17.94
	B	1500	1.14	42.29	30.83	25.18	1.50	0.17	7.71
	B2[4]	2000	0.35	49.25	39.35	10.50	1.50	0.19	3.79
Responsive (wet) C	O	300	9.36	34.00	48.50	12.70	1.00	0.18	37.71
	G	2000	0.35	49.25	39.35	10.50	1.50	0.19	3.79
Shallow (responsive) C	A	300	6.77	33.83	36.50	22.35	1.15	0.16	17.94
	C	500	0.24	35.30	34.98	28.68	1.50	0.18	3.75
Stagnant A	A	300	6.77	33.83	36.50	22.35	1.15	0.16	17.94
	B	1000	1.14	42.29	30.83	25.18	1.50	0.17	7.71
	C	1500	0.24	35.30	34.98	28.68	1.50	0.18	3.75

OC = Organic Carbon; AWC = Available Water Capacity; Ks = Saturated hydraulic conductivity

Streamflow data were recorded at DWS weirs U2H013, U2H007, and U2H006 (Figure 6.1). Daily rainfall records were sourced from seven rainfall stations provided by the South African Weather Service and DWS (Figure 6.1). The average annual rainfall during the simulation period (2000-2013) at these stations was 675 mm. In instances of malfunctioning rainfall stations, the average daily rainfall recorded at the remaining stations was used to fill in the days without data. Daily minimum and maximum temperatures, along with relative humidity, were obtained from weather stations. Solar radiation and wind speed data were sourced from the National Center for Environmental Prediction (Saha et al., 2010). Daily potential evapotranspiration was calculated using the Penman-Monteith approach (Monteith, 1965).

The model ran individually on the three catchments from January 1998 to December 2013, incorporating two levels of soil input data. For each of the catchments a scenario of change was also included, where all grasslands are converted to forestry. That is, the ‘before’ scenario relied on land cover data of 2014 (Figure 6.2) and the ‘after’ scenario used the same land cover but used pine forestry input parameters for grasslands. This resulted in 12 model runs. The initial two years served as a warm-up period, followed by 14 years of validation, as no model calibration period was included since the focus was not on optimisation.

Statistical analysis

The comparison between simulated monthly streamflow and measured flow at the three stream gauges (Figure 6.1) involved the use of five widely recognised indices: the coefficient of determination (R^2), the Root Mean Square Error (RMSE), percentage bias (PBIAS), the Nash-Sutcliffe Efficiency (NSE), and the Kling-Gupta Efficiency (KGE). PBIAS serves to quantify the degree of overestimation or underestimation in the simulations relative to observed values (Gupta et al., 1999). NSE assesses the magnitude of variance between simulated and observed values (Nash & Sutcliffe, 1970). A value greater than 0.5 generally indicates satisfactory model performance when comparing monthly data (Moriasi et al., 2015). Higher KGE values (Gupta et al., 2009) signify better model performance, and

values smaller than -0.41 suggest that the means of the observations provide a better fit than the model (Knoben et al., 2019).

6.2.3 Results and discussion

HYDROSOIL maps

The HYDROSOIL maps were generated using the multinomial logistic regression algorithm with the different sampling techniques – random sampling (Figure 6.4), K-means (Figure 6.5) and Conditioned Latin Hypercube (Figure 6.6).

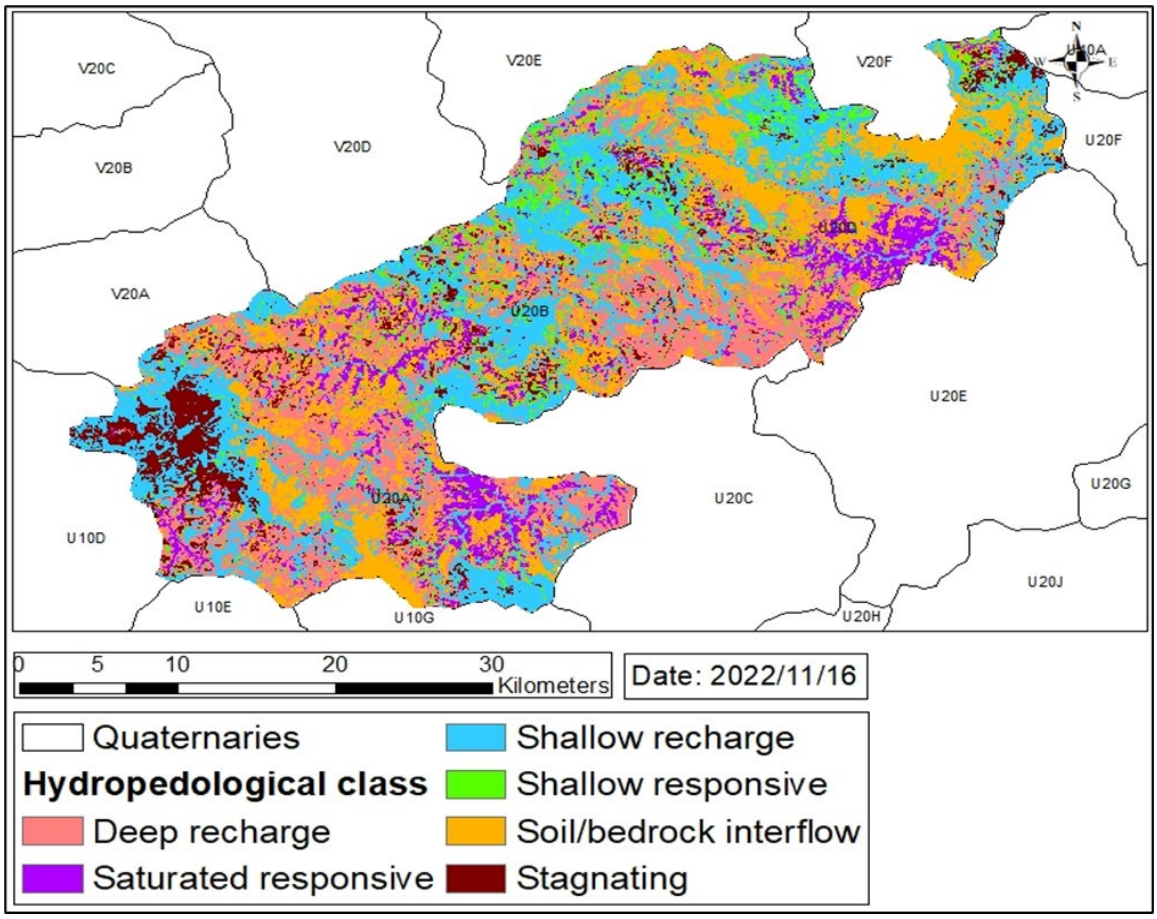

Figure 6.4: Hydrological soil map (HYDROSOIL) of the uMngeni catchment area created using the multinomial logistic regression algorithm with random sampling.

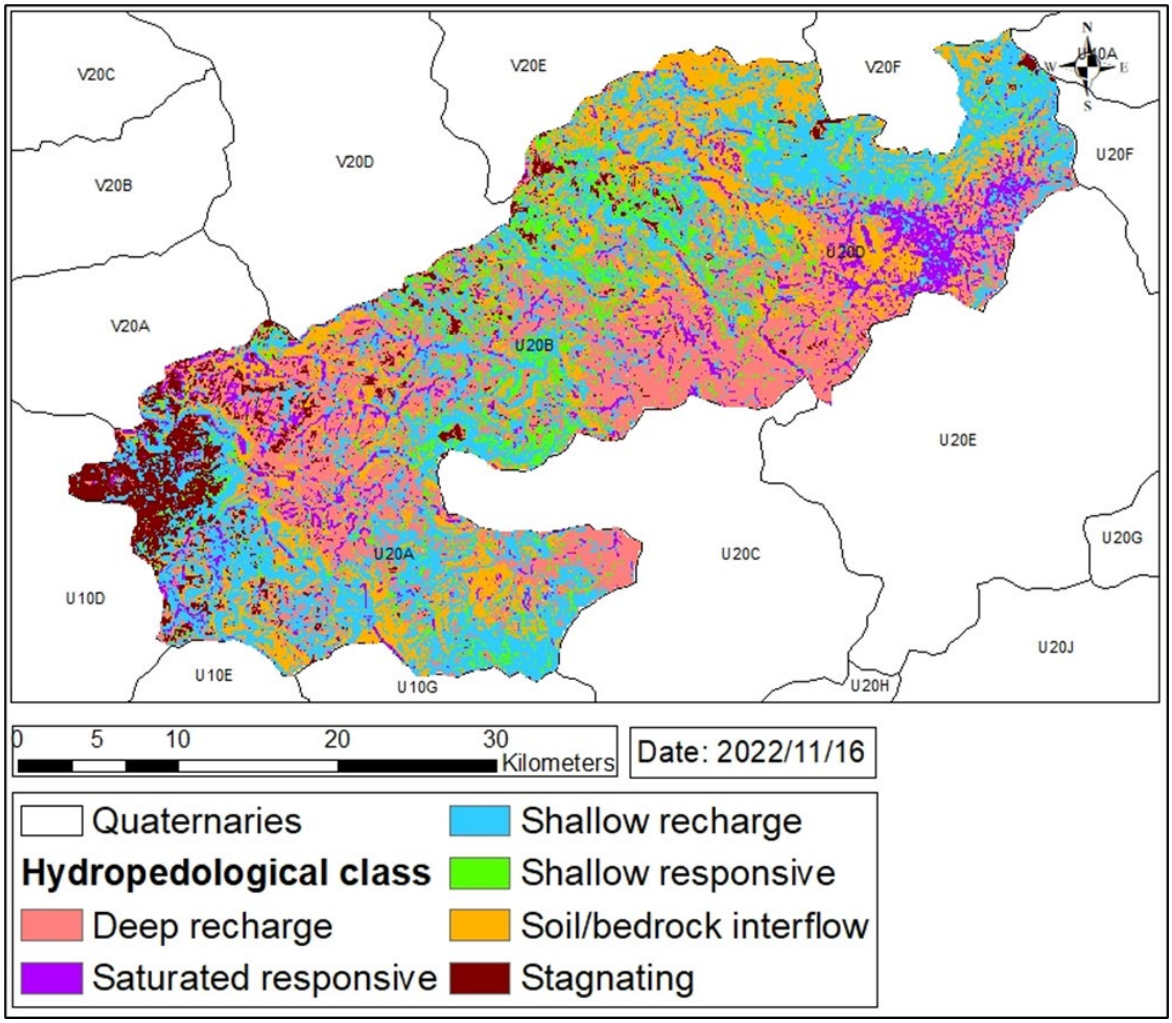

Figure 6.5: Hydrological soil map (HYDROSOIL) of the uMngeni catchment area created using the multinomial logistic regression algorithm with K-means clustering.

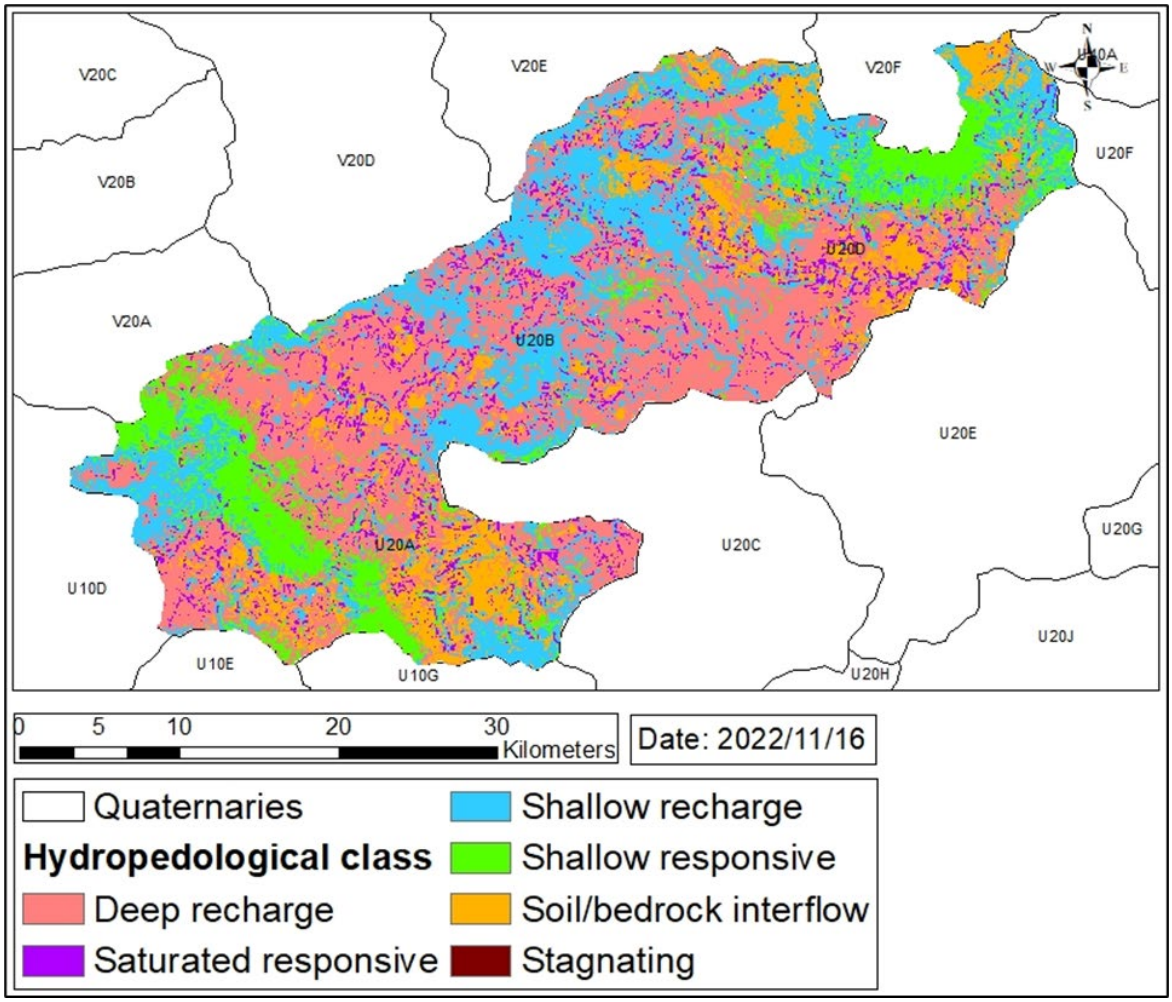

Figure 6.6: Hydrological soil map (HYDROSOIL) of the uMngeni catchment area created using the multinomial logistic regression algorithm with Conditioned Latin Hypercube sampling.

The HYDROSOIL map generated by multinomial logistic regression with random sampling resulted in a Kappa coefficient that showed almost perfect agreement (0.92), while the total evaluation point accuracy was very high (95.8%; Table 6.3). Although the Kappa and total evaluation point accuracy is high, only four of the six hydropedological classes were represented for validation.

Table 6.3: Confusion matrix of the HYDROSOIL map generated by multinomial logistic regression with stratified random sampling.

		Map units				Total	Correct	%
		Deep recharge	Shallow recharge	Shallow responsive	Saturated responsive			
Observations	Deep recharge	15	1			16	15	93.8
	Shallow recharge		4			4	4	100.0
	Shallow responsive			1		1	1	100.0
	Saturated responsive				3	3	3	100.0
Total		15	5	1	3	24		
Correct		15	4	1	3		23	
%		100.0	80.0	100.0	100.00			95.8

The HYDROSOIL map generated by multinomial logistic regression with K-means clustering resulted in a Kappa coefficient that showed substantial agreement (0.78), while the total evaluation point accuracy was also very high (87.5%; Table 6.4).

Table 6.4: Confusion matrix of the HYDROSOIL map generated by multinomial logistic regression method with K-means clustering.

		Map units				Total	Correct	%
		Deep recharge	**Shallow recharge**	**Shallow responsive**	**Saturated responsive**			
Observations	Deep recharge	13	2		1	16	13	81.3
	Shallow recharge		4			4	4	100.0
	Shallow responsive			1		1	1	100.0
	Saturated responsive				3	3	3	100.0
Total		13	6	1	4	24		
Correct		13	4	1	3		21	
%		100.0	66.7	100.0	75.0			87.5

Table 6.5 shows the confusion matrix of The HYDROSOIL map generated by multinomial logistic regression with Conditioned Latin Hypercube sampling resulted in high Kappa and total evaluation point accuracy, although only five of the six hydropedological classes were represented for validation. The Kappa coefficient showed substantial agreement (0.75), while the total evaluation point accuracy was also very high (87.5%).

Table 6.5: Confusion matrix of the map generated by multinomial logistic regression method with Conditioned Latin Hypercube sampling.

		Map units					Total	Correct	%
		Deep recharge	**Shallow recharge**	**Shallow responsive**	**Saturated responsive**	**Interflow (soil/bedrock)**			
Observations	Deep recharge	16	1				17	16	94.1
	Shallow recharge		1		1	1	3	1	33.3
	Shallow responsive			1			1	1	100.0
	Saturated responsive				3		3	3	100.0
	Interflow (soil/bedrock)						0	0	
Total		16	2	1	4	1	24		
Correct		16	1	1	3	0		21	
%		100.0	50.0	100.0	75.0	0.0			87.5

From the total evaluation point accuracy and Kappa coefficient, all the HYDROSOIL maps created using the three different sampling techniques were judged to be acceptable for hydrological modelling. All HYDROSOIL maps showed significant higher accuracy compared to the previous map (73%; Van Tol & Van Zijl, 2022). However, the previous hydropedological soil map was focused on the disaggregation of Land Type data and only four hydrological classes were used (Van Tol & Van Zijl, 2022). Future focus should be on the validation of all hydropedological classes, which might decrease the accuracy of the maps for all three types of sampling. The HYDROSOIL map created using random sampling was

judged to be the most accurate, as this map showed the highest total evaluation point accuracy (95.8%) and Kappa coefficient (0.92; Table 6.3), and therefore was used for hydrological modelling.

Modelling results for catchment U20A

In the assessment of U20A, both HYDROSOIL and Land Type simulations exhibited suboptimal performance, as evidenced by low R^2 and NSE values (Figure 6.7). Although HYDROSOIL simulations showed slight improvement compared to those with the Land Type dataset, they still fell short of being considered 'satisfactory', especially considering the 0.5 threshold for NSE (Moriasi et al., 2007). Analysing PBIAS, overestimations noted by Van Tol and Van Zijl (2022) in both HYDROSOIL and Land Type datasets were substantially mitigated with the new model setup, but performance remained inferior to previous simulations. Notably, the model exhibited significant underestimation of baseflow, a deficiency that could potentially be rectified through calibration.

It is worth highlighting that, after relatively dry years, observed streamflow failed to respond to rainfall, whereas simulated streamflow notably increased. This phenomenon is likely attributable to the necessity of filling numerous small farm dams before generating streamflow. The underestimation of baseflow could be attributed to extremely low percolation volumes (Table 6.6; Figure 6.8), which are unrealistic for catchments dominated by freely-drained soils. Excessive lateral flows suggest a need for future work to emphasise percolation over lateral flows.

Interestingly, there was minimal disparity between simulated water balance components using HYDROSOIL and Land Type datasets under natural land-use conditions (Figure 6.8a). Differences primarily arose in soil water content due to variations in assigned soil depths and storage parameters. However, under the change scenario (Figure 6.8b), substantial differences in streamflow and lateral flow generation emerged. This discrepancy underscores that internal catchment processes are simulated differently by distinct soil datasets, emphasising the critical importance of accurately representing these processes in scenarios of change (Yen et al., 2014; Arnold et al., 2015).

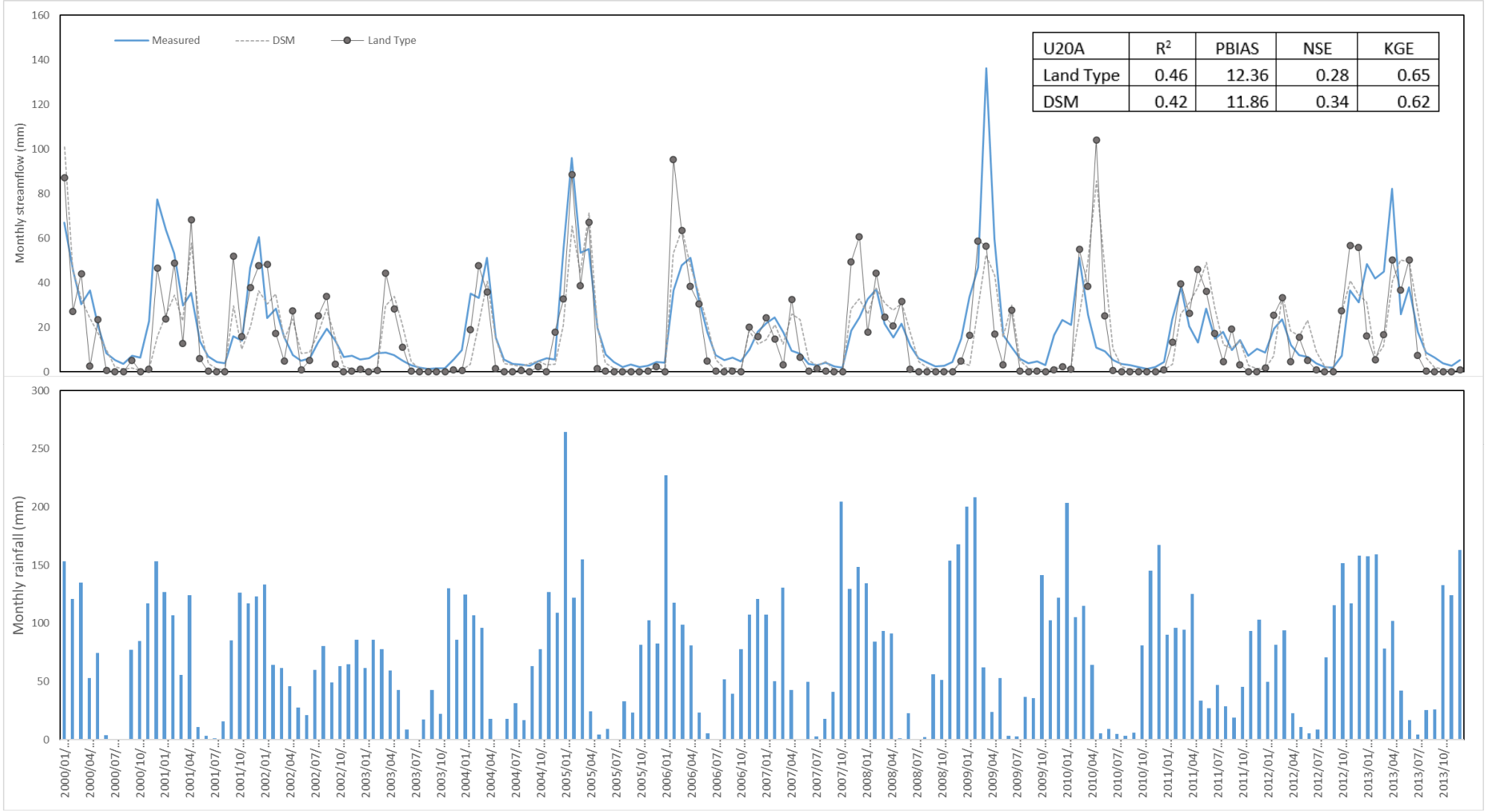

Figure 6.7: Streamflow simulations and accuracies for HYDROSOIL (DSM) and Land Type datasets in catchment U20A.

Table 6.6: Water balance components (mm) for 'before' and 'after' afforestation scenarios using two different soil inputs in U20A.

	HYDROSOIL			Land Type		
	Natural	**Forestry**	**%Change**	**Natural**	**Forestry**	**%Change**
Rainfall	729.3	729.3		729.4	729.3	
Overland flow	7.1	5.3	-26.2	6.2	4.7	-24.4
Lateral flow	194.7	32.4	-83.4	194.8	65.2	-66.5
Water yield	201.9	37.7	-81.3	201.0	69.9	-65.2
Percolation	4.0	0.5	-86.1	0.8	0.3	-57.7
ET	531.8	703.7	32.3	530.5	663.4	25.1
Transpiration	291.4	544.0	86.7	258.1	504.7	95.5
Evaporation	199.7	90.9	-54.5	232.7	91.2	-60.8
Profile soil water	240.6	74.3	-69.1	49.8	26.8	-46.2
Topsoil water	31.5	18.5	-41.3	14.3	8.6	-39.8
ET0	1332.2	1332.2		1332.2	1332.2	

ET = Evapotranspiration

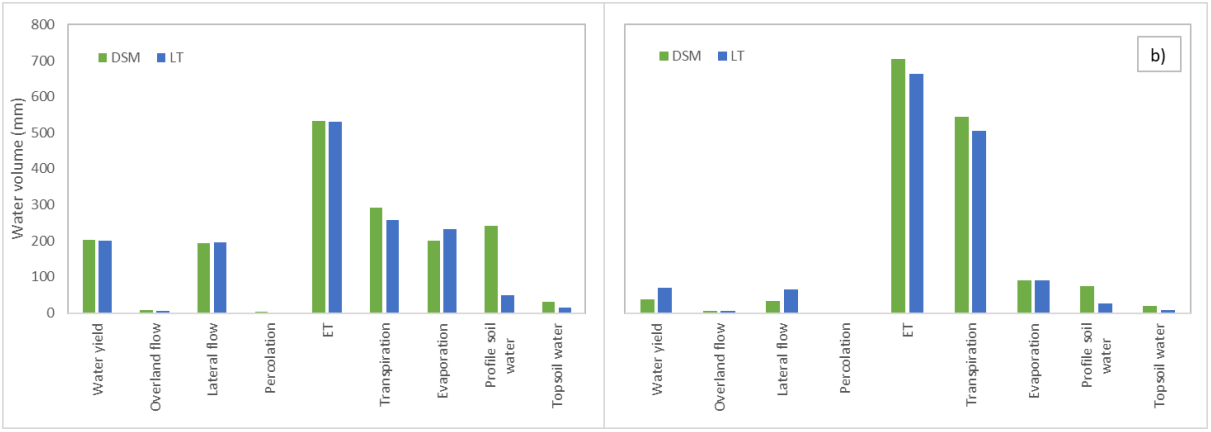

Figure 6.8: Simulated water balance components for U20A using the HYDROSOIL (DSM) and Land Type (LT) soil inputs for a) before scenario and b) scenario where all grasslands were converted to forestry.

Modelling results for catchment U20B

In the assessment of U20B, the HYDROSOIL dataset demonstrated superior performance over the Land Type dataset when considering all statistical indices (Figure 6.9). The only exception was observed in PBIAS, where the Land Type dataset exhibited better performance. Despite the overall modest results, the improvement in soil information led to more accurate streamflow predictions, particularly noteworthy given that the model underwent no calibration. Similar to U20A, the notable underestimation of baseflow persists as a significant concern in the model configuration.

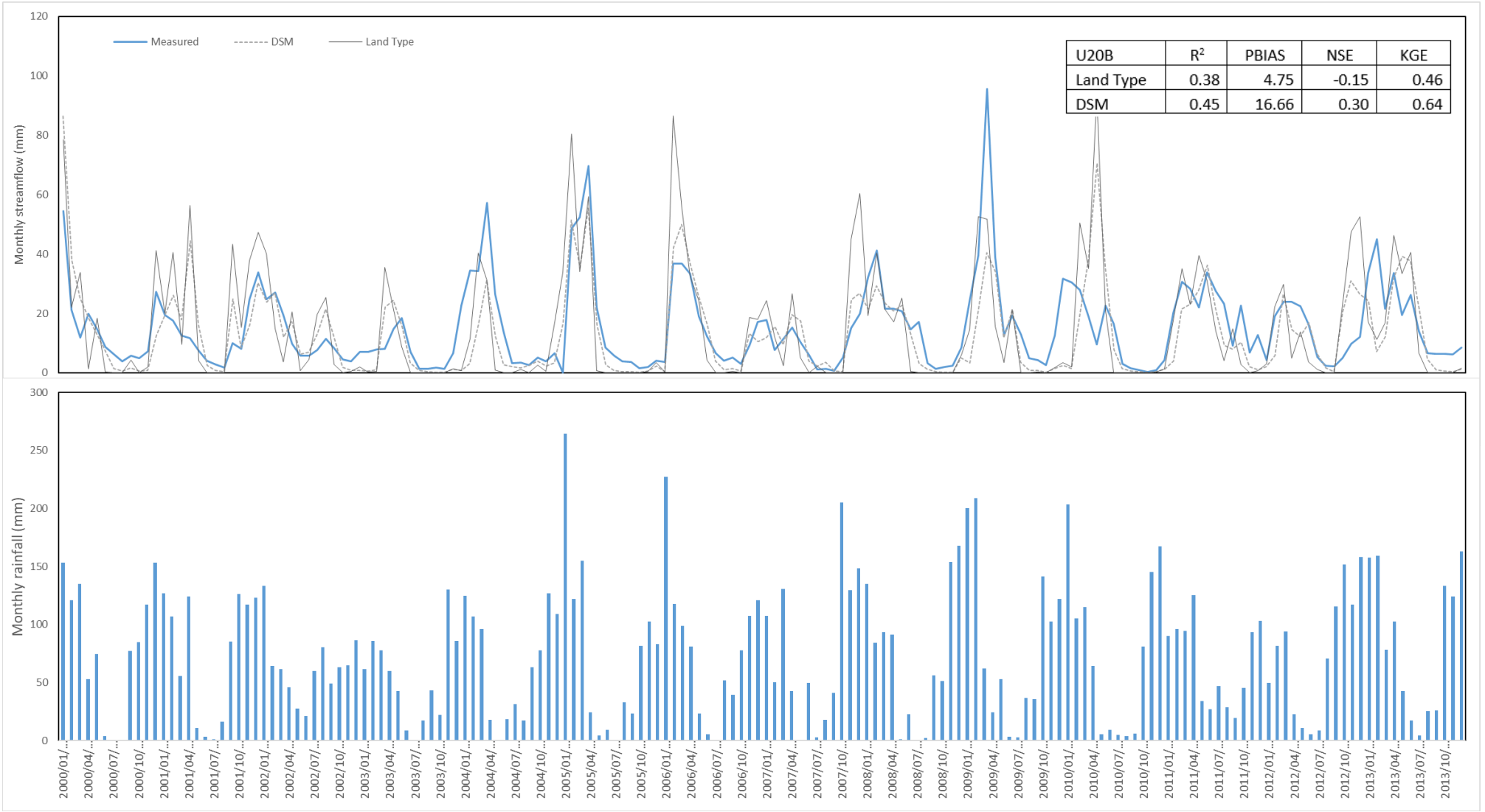

Figure 6.9: Streamflow simulations and accuracies for HYDROSOIL (DSM) and Land Type datasets in catchment U20B.

The deficient baseflow estimation could be linked to the low percolation values (Table 6.7; Figure 6.10. Allowing more water to recharge groundwater stores, which are gradually released into streams, has the potential to increase baseflow while reducing lateral flows. Intriguingly, discrepancies in water balance components between different soil input datasets are more pronounced in U20B than in U20A. Variations in soil water content remain a prominent distinction between simulations using different input datasets. The HYDROSOIL soil dataset leads to higher transpiration due to deeper soils capable of storing more water. Notably, under 'forestry' simulations, evapotranspiration accounts for over 90% of the water balance.

Table 6.7: Water balance components (mm) for 'before' and 'after' afforestation scenarios using two different soil inputs in U20B.

	HYDROSOIL			Land Type		
	Natural	**Forestry**	**%Change**	**Natural**	**Forestry**	**%Change**
Rainfall	725.1	725.0		725.0	725.0	
Overland flow	11.7	9.9	-15.3	9.3	8.8	-5.8
Lateral flow	145.7	40.3	-72.3	169.3	78.1	-53.9
Water yield	157.3	50.2	-68.1	178.7	86.9	-51.4
Percolation	3.2	1.0	-70.0	0.5	0.3	-36.3
ET	573.5	686.1	19.6	549.3	642.6	17.0
Transpiration	300.1	463.7	54.5	263.9	426.0	61.4
Evaporation	234.6	165.4	-29.5	247.3	160.2	-35.2
Profile soil water	203.4	104.8	-48.5	54.1	35.0	-35.3
Topsoil water	29.1	19.7	-32.4	13.6	9.6	-29.7
ET0	1349.9	1350.0		1350.0	1350.0	

ET = Evapotranspiration

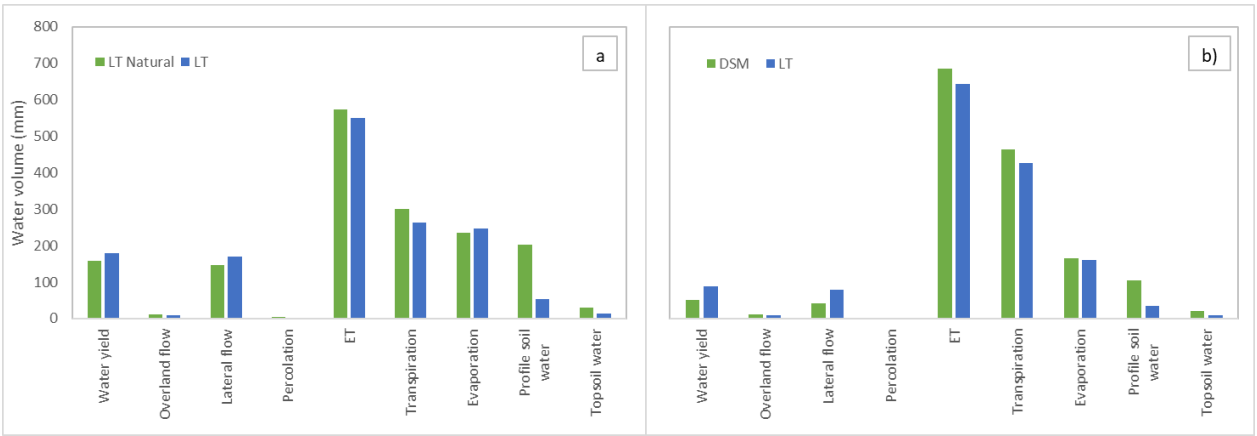

Figure 6.10: Simulated water balance components for U20B using the HYDROSOIL (DSM) and Land Type (LT) soil inputs for a) before scenario and b) scenario where all grasslands were converted to forestry.

Modelling results for catchment U20D

Catchment U20D exhibited the poorest simulations among the three catchments (Figure 6.11), with both HYDROSOIL and Land Type datasets yielding NSE and KGE values below 0.5. Surprisingly, the

Land Type dataset performed notably better than the HYDROSOIL dataset across all statistical indices. The primary reason for this suboptimal performance lies in the consistent underestimation of both baseflows and peak flows, evident in PBIAS values below -100. Remarkably, even with the existing land cover, evapotranspiration accounts for more than 85% of the water balance (Table 6.8), a remarkably high proportion for semi-arid to sub-humid regions.

Two plausible explanations for this excessive evapotranspiration simulation exist. Firstly, the inadequacy of rainfall stations, positioned outside the catchment in an area with diverse topography, may not accurately represent local rainfall variations (Figure 6.1). U20D, with the most natural forests and plantations among the three catchments (Figure 6.2), experiences the lowest recorded rainfall, suggesting potential inaccuracies in rainfall inputs. Secondly, an overestimation of abstraction by forests and plantations, influenced by default parameters derived from the northern hemisphere, where temperature, not water, typically limits growth, may contribute to excessive evapotranspiration. Adjustments to plant parameters might be necessary to better accommodate semi-arid plants adapted to limited water availability.

Comparing before-and-after scenarios in U20D with HYDROSOIL and Land Type datasets (Figure 6.12), both exhibit similar magnitudes of change. Despite inadequate percolation and excessive lateral flows, the main cause of simulation errors is a notable water balance discrepancy between simulated and actual streamflow, attributable to underestimation of both peak and baseflows.

The disappointment lies in the fact that detailed soil information did not lead to improved simulations. However, this underscores the importance of comprehending the catchment’s water balance. While enhanced soil information may not eliminate all modelling uncertainties, it contributes to better reflecting processes, instilling confidence in modelers to calibrate parameters with the assurance that the model configuration is accurate (Van Tol et al., 2021).

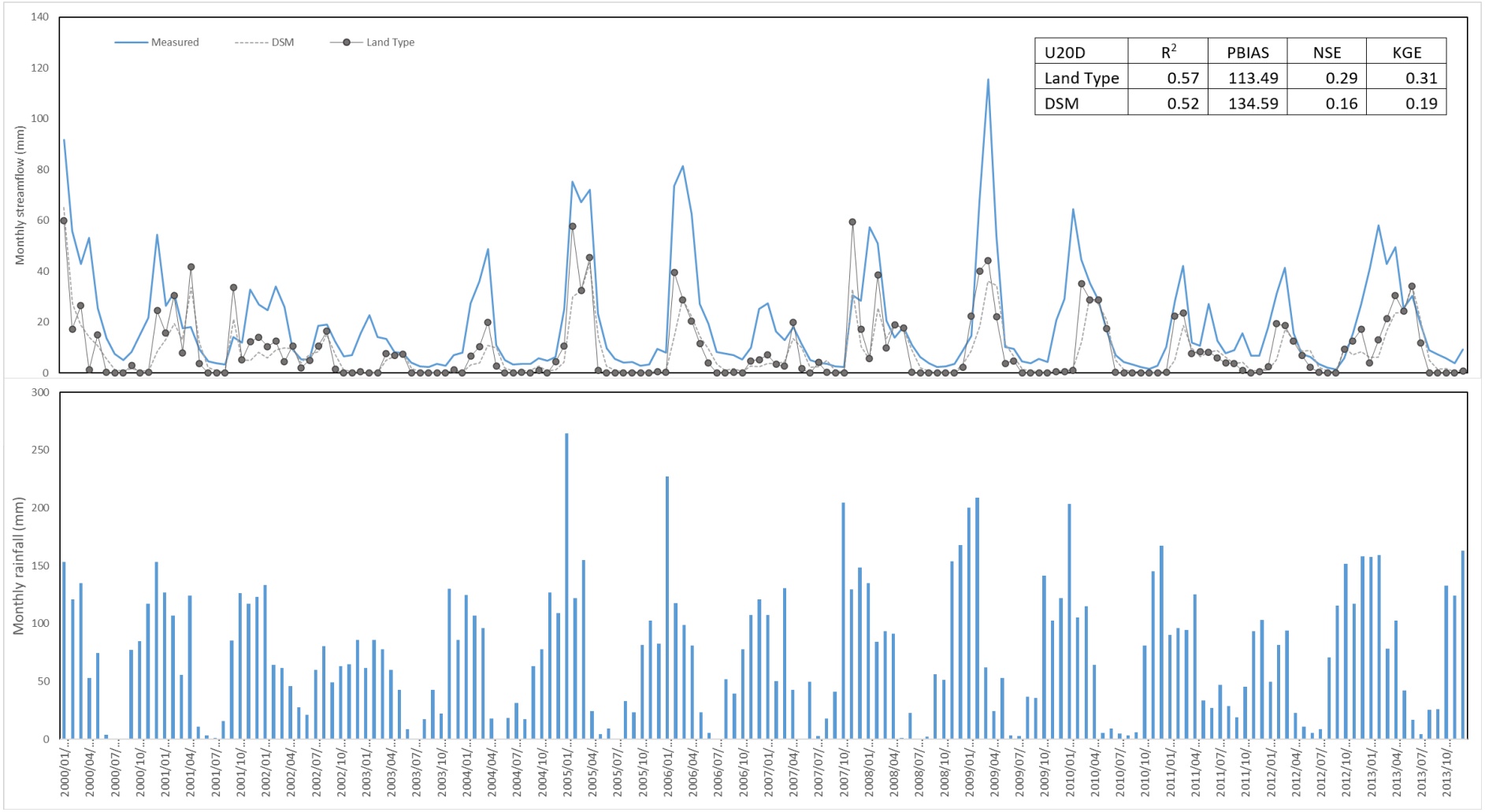

Figure 6.11: Streamflow simulations and accuracies for HYDROSOIL (DSM) and Land Type datasets in catchment U20D.

Table 6.8: Water balance components (mm) for 'before' and 'after' afforestation scenarios using two different soil inputs in U20D.

	HYDROSOIL			Land Type		
	Natural	**Forestry**	**%Change**	**Natural**	**Forestry**	**%Change**
Rainfall	657.9	657.9		657.9	657.8	
Overland flow	6.5	5.1	-22.1	4.0	3.8	-5.5
Lateral flow	93.3	28.2	-69.8	107.8	41.6	-61.4
Water yield	99.8	33.2	-66.7	111.8	45.3	-59.4
Percolation	2.6	0.9	-66.7	0.4	0.2	-50.5
ET	560.7	631.7	12.6	548.0	615.3	12.3
Transpiration	366.7	475.1	29.5	349.1	463.5	32.7
Evaporation	155.6	107.1	-31.1	161.1	103.0	-36.0
Profile soil water	157.3	77.9	-50.5	50.4	30.4	-39.8
Topsoil water	24.1	16.3	-32.6	12.3	8.7	-29.3
ET0	1455.3	1455.4		1455.5	1455.5	

ET = Evapotranspiration

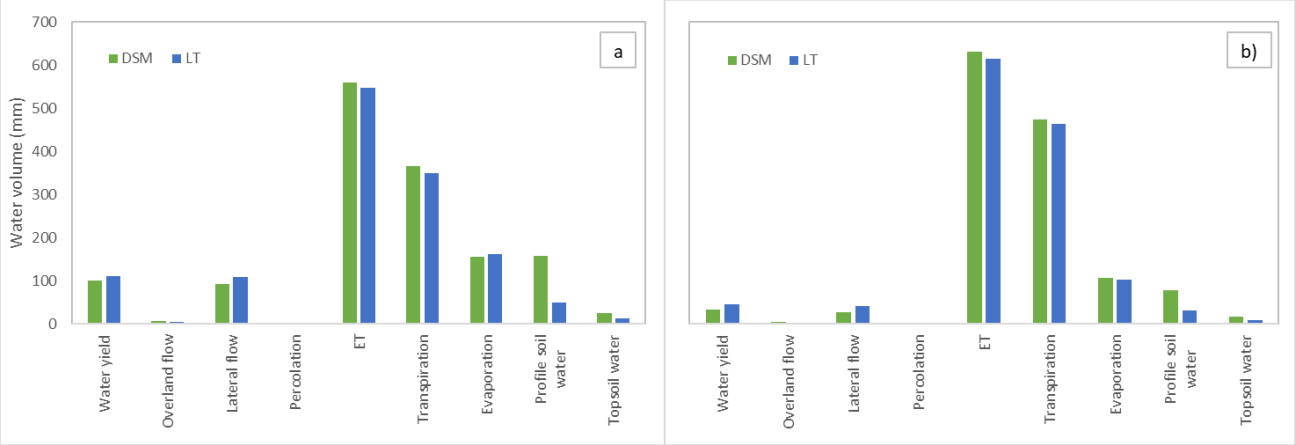

Figure 6.12: Simulated water balance components for U20D using the HYDROSOIL (DSM) and Land Type (LT) soil inputs for a) before scenario and b) scenario where all grasslands were converted to forestry.

6.2.4 Conclusions

Digital soil mapping proved to be useful in mapping the uMngeni catchment. It was, however, evident that the type of sampling that is done to compile the datasets as well as the method used for the prediction, affects the accuracy of the maps. The stratified random sampling was the better sampling method compared to the K-means and the Conditioned Latin Hypercube sampling, resulting in an overall accuracy of 95.8% and Kappa coefficient of 0.92. Therefore, using stratified random sampling generated the most acceptable HYDROSOIL map to be used for hydrological modelling. However, future focus should be on the validation of all hydropedological classes, which might decrease the accuracy of the maps regarding all three types of sampling.

In terms of the modelling, the HYDROSOIL dataset consistently outperformed the Land Type soil dataset in two of the three catchments. U20D presented a departure from this trend, likely due to inaccuracies in rainfall data. While U20A and U20B showed promising results, it is evident that calibration will be imperative to attain

satisfactory model performance. The persistent underestimation of baseflows suggests that more water needs to reach the groundwater aquifers and not contribute to evapotranspiration or lateral flows. Yielding realistic lateral flows and percolation from various hydropedological soil types is an important activity for studies.

The results underscore that despite the varying success in improving streamflow comparisons, the detailed soil information does enhance the representation of internal catchment processes. This insight is crucial as it provides confidence to modellers that, even if streamflow outcomes are not dramatically improved, the model captures the underlying processes accurately. The amplification of simulation differences when considering scenarios of change emphasises the critical role of precise soil information in anticipating the impacts of land-use modifications and environmental variations. In essence, while calibration remains a necessity for improved model accuracy, the results show that detailed soil information contributes to a more robust understanding of catchment dynamics, paving the way for more reliable hydrological simulations.

6.3 APPLICATION OF HYDROSOIL INPUT DATA IN THE TSITSA CATCHMENT

6.3.1 Introduction

The combination of models and remote sensing techniques within a GIS framework is commonly utilised to assess hydrological processes such as streamflow, water erosion, sediment yield dynamics and nutrient inputs/outputs (e.g. Guzha et al., 2018). One of the biggest challenges in hydrological modelling in developing countries is to obtain appropriate input data, especially soil data. Soil data preparation and model set-up is a laborious task, especially due to the lack of appropriate and representative data (Glenday et al., 2021). The application of inappropriate soil data at a catchment scale could lead to errors and uncertainty in hydrological simulations. This challenge can be addressed by creating hydrological soil property maps by means of digital soil mapping techniques and the application of pedotransfer functions to generate the required hydraulic parameters.

The aim of this study is to set-up and run the SWAT model in the Tsitsa catchment with the HYDROSOIL map and the Land Type database. SWAT has been applied in South Africa to support various large catchment modelling studies (Glenday et al., 2021). HYDROSOIL maps will not only assist users to set up and run the SWAT model in South Africa with appropriate soil data, but will also assist in the standardisation of hydrological modelling efforts in South Africa.

6.3.2 Materials and methods

The Tsitsa catchment

The Tsitsa catchment is located in the Eastern Cape province of South Africa. The SWAT model was applied in quaternary catchment T35E, which is nested in the Tsitsa catchment with a drainage area of 49 007 ha (Figure 6.13). The Tsitsa River drains the Drakensberg escarpment (approximately 2 600 m a.s.l.) and flows east into the Mzimvubu River (at approximately 200 m a.s.l.) after a flow length of approximately 200 km. The climate is sub-humid with mean annual rainfall ranging from 625 mm in the lower plains to 1 327 mm in the mountains (ARC, 2012). The catchment falls mainly within the Grassland biome, with narrow bands of Bushveld along the river networks in the lower part of the catchment, as well as pockets of Afromontane Forest in fire protected ravines (Mucina & Rutherford, 2006). The main land use is extensive grazing with areas of pine and gum plantations, and maize cultivation in the upper catchment.

The geology consists of a succession of sedimentary layers of the Triassic age, including Adelaide mudrock succeeded by mudstones of the Tarkastad, Molteno and Elliot Formations (Council for Geoscience, 2007). Mudstones are overlain by sandstone and siltstone of the Clarens Formation and capped by Drakensberg basaltic lava of the Jurassic age. Karoo dolerite sills and dykes are present in the sedimentary formations, leading to more resistant base level controls.

Although soils in the catchment vary significantly, those from the mudstone parent material in the central part of the catchment are associated with duplex soils that are highly erodible with widespread gully erosion. Duplex soils have a marked increase in clay content from the topsoil to subsoil and an abrupt transition with respect to texture, structure and consistency (ARC, 2012). Soil forms that often have duplex properties include Katspruit, Kroonstad, Sterkspruit, Estcourt, and to a lesser extent Valsrivier, Swartland and Bonheim. These soils limit intrinsic permeability since water does not move readily into the subsurface matrix, which often leads to increased subsurface flow (Van Tol et al., 2013) causing tunnel and subsequent gully erosion. In the Tsitsa catchment, duplex soils often have prismacutanic subsoils that can easily be identified by the large structured prisms that are exposed on gully sidewalls or where the topsoil is completely eroded. Importantly, the subsurface matrix of duplex soils is often dispersive as a result of high sodium absorption (Van Zijl et al., 2014).

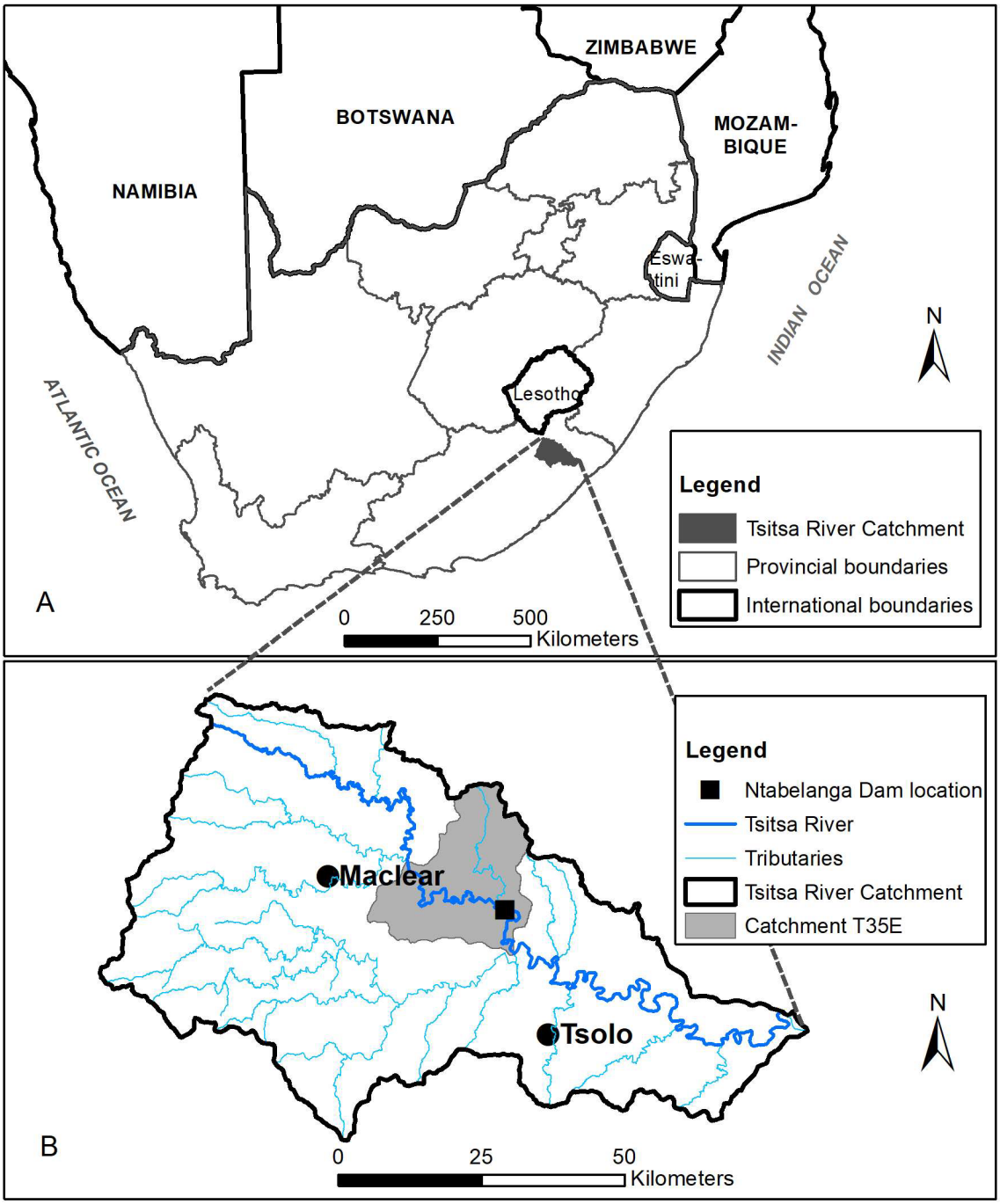

Figure 6.13: Location of the Tsitsa catchment T35E in the Eastern Cape province, South Africa.

Digital soil mapping

For a description of the digital soil mapping methods, please see Section 6.1.

Hydrological modelling

ArcSWAT-2012 was used for simulations, which is a graphical user interface for SWAT and ArcGIS® software extension (Srinivasan et al., 1998). Topographic, land use, soil, climate and hydrological data were utilised to configure and parameterise the Tsitsa catchment T35E (Table 6.9).

Table 6.9: Summary of topographic, land cover, soil, climate and hydrological input data used to parameterise the Tsitsa catchment T35E.

	Input data	
DEM GSD (m)	STRM DEM (USGS, 2015) 30	
Land cover data GSD (m)	SANLC (GeoTerra Image, 2015) 30	
Soil data	Land Type database (ARC, 2012) South African Atlas of Climatology and Agrohydrology (Schulze, 2007)	HYDROSOIL
Usable scale	1:250 000	1:100 000
HRUs	610	616
Slope class (%) Thresholds (%)	0-5; 5-10; 10-20; 20-40; >40 Land use 10; Soil 10; Slope 10	
Climate data Number of stations Timeframe Simulation period (years)	ARC Agroclimatology database (2012) 2 2005-2012 8 (including 3-year warm up)	

GSD = ground sample distance

First, using the 30 m SRTM DEM at 90 m resolution (USGS, 2015), topographic and drainage networks of the catchment were partitioned into sub-catchments that are comparative in size and representing all relevant river tributaries. A total of 47 sub-catchments were delineated in T35E (Figure 6.14).

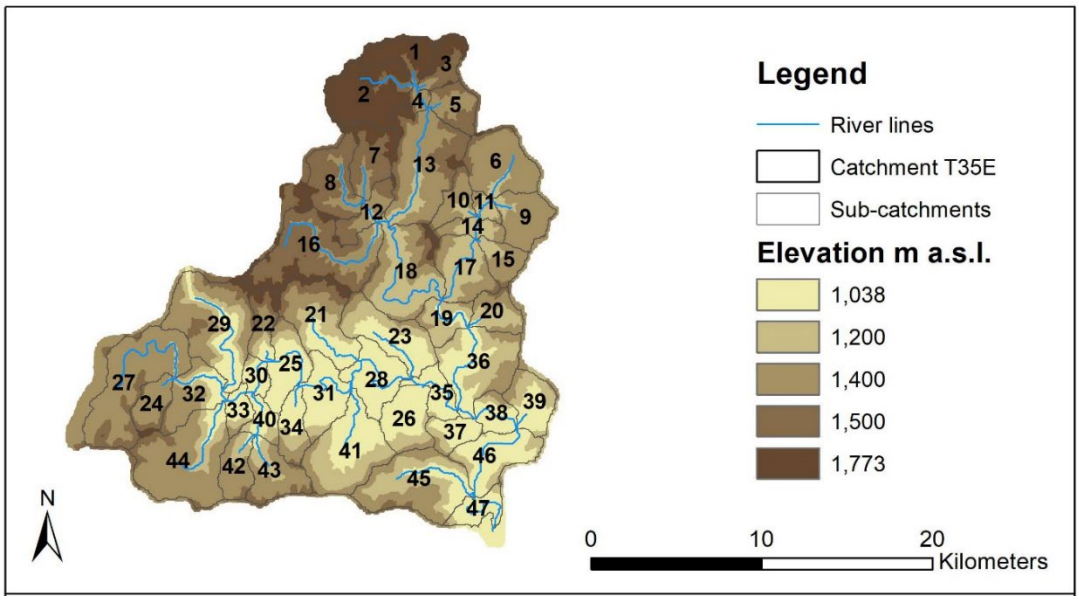

Figure 6.14: Tsitsa catchment T35E illustrating the 47 delineated sub-catchments and streams.

Land cover data were derived from the South African National Land Cover map (GeoTerra Image, 2015) creating 12 land cover classes for T35E (Figure 6.15). These land cover classes were linked to the land cover types in the SWAT database. Next, soil texture and hydraulic parameter values were assigned to the Land Types and the HYDROSOIL classes of T35E.

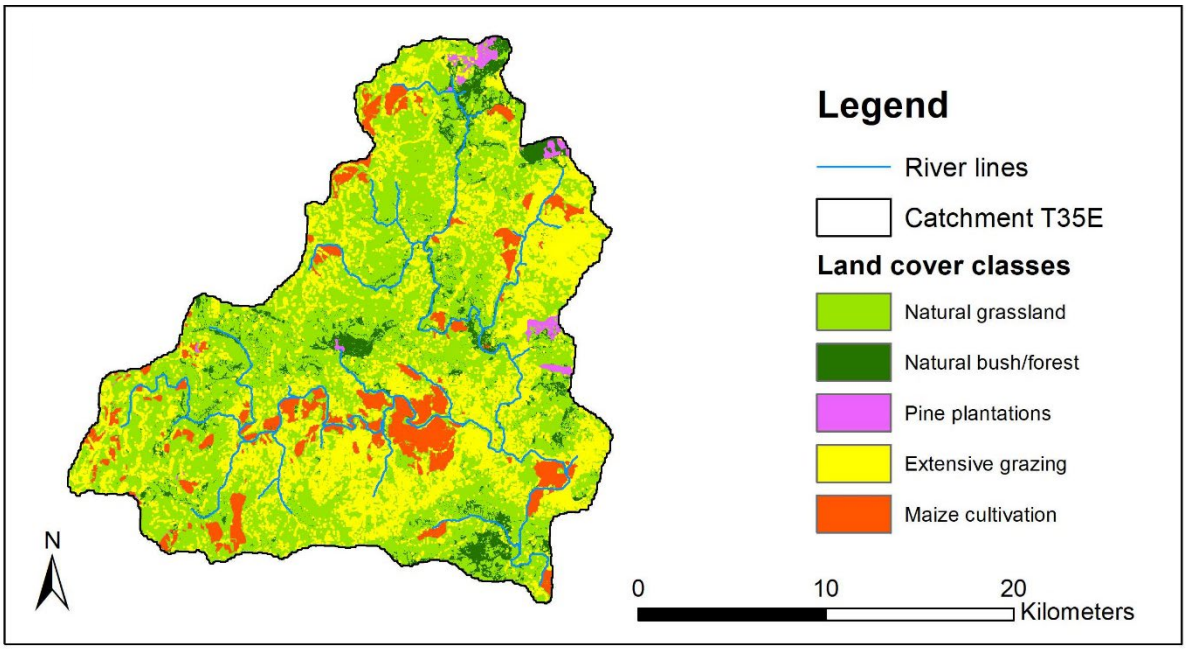

Figure 6.15: Simplified land cover map of the Tsitsa catchment T35E showing the extent of five most prominent land cover classes.

Pedotransfer functions (Van Tol et al., 2013; Van Zijl et al., 2016; Van Tol et al., 2020) were used to generate the required hydraulic parameters, including available water capacity and saturated hydraulic conductivity. Soil parameter values (Table 6.10) were assigned to the Land Types in catchment T35E (Figure 6.16).

Table 6.10: Description and reasoning used to assign soil parameter values to each soil component of the Land Type data of the Tsitsa catchment T35E.

Soil parameter	Reasoning
Number of layers in the soil	Two soil layers/horizons were incorporated into each soil component of the Land Type database.
Depth from soil surface to bottom of layer (mm)	Depth descriptions/classes in the Land Type database and Schulze (2007) were used to assign depth to each Land Type.
Maximum rooting depth of soil profile (mm)	As above.
Soil Hydrologic Group (A, B, C, D) in terms of runoff potential, Soil Group A = low, B = moderately low, C = moderately high, D = high.	Soil hydrological groups were based on the broad soil patterns given in the Land Type database as follows: A for deep and freely drained apedal soils with humic topsoils as well as podzols; B for apedal soils with plinthic subsoils or deep alluvial soils; C for shallow soils or planosols comprising sandier topsoil abruptly overlying more clayey subsoil; D for rock outcrops.
Available water capacity of the soil layer (mm H_2O/mm soil)	For each Land Type, Schulze (2007) calculated plant available water content as the difference between water content at field capacity and permanent wilting point.
Saturated hydraulic conductivity (mm/hr)	Values were derived from the Rosetta Model (Schaap et al., 2001) based on the soil texture classes for each Land Type.
Bulk density (Mg/m^3 or g/cm^3)	Bulk density (BD) was estimated using porosity (PO) data in Schulze (2007) for each Land Type: PO = 1-BD/2.65.
Soil albedo (non-dimensional value between 0 and 1)	Albedo values were assigned to broad soil patterns in the Land Type database ranging between 0.25 for light-coloured sands to 0.7 for dark clays.
Clay content with diameter of < 0.002 mm (% soil weight)	Clay content in the A-horizon was assigned using the average topsoil clay classes given to each Land Type. Clay content in the B-horizon was assigned to each Land Type by adjusting the clay values of the A-horizon to clay-factors (Le Roux et al., 2022).

Soil parameter	Reasoning
Silt content with diameter of 0.05-0.002 mm (% soil weight)	Due to the lack of data, silt content for A and B horizons were assigned values between 10-22.5%, increasing with increase in clay as follows (Le Roux et al., 2022): percentage of Land Type with <= 6% clay = 10% silt; 6.1-15% clay = 15% silt; 15.1-25% clay = 17.5% silt; 25.1-35% clay = 20% silt; 35.1-55% clay = 22.5% silt.
Sand content with diameter of > 2 mm (% soil weight)	Sand content for A and B horizons were assigned as follows: Sand = 100% – (%clay + %silt).
Rock fragment content (% soil weight)	Rock content was based on the agricultural restriction/rock (MB) classes in the Land Type database as follows: MB0=0% (no rock); MB1=20%; MB2=50%; MB3=20%; MB4=100% (no soil).
Organic carbon content (% soil weight)	A soil organic carbon map of South Africa (Schulze & Schütte, 2020) were used to assign average carbon values for A and B horizons per Land Type.
USLE K factor in SI units t/ha per unit 'erosivity'	Erodibility units (Le Roux et al., 2008) were assigned to each Land Type.

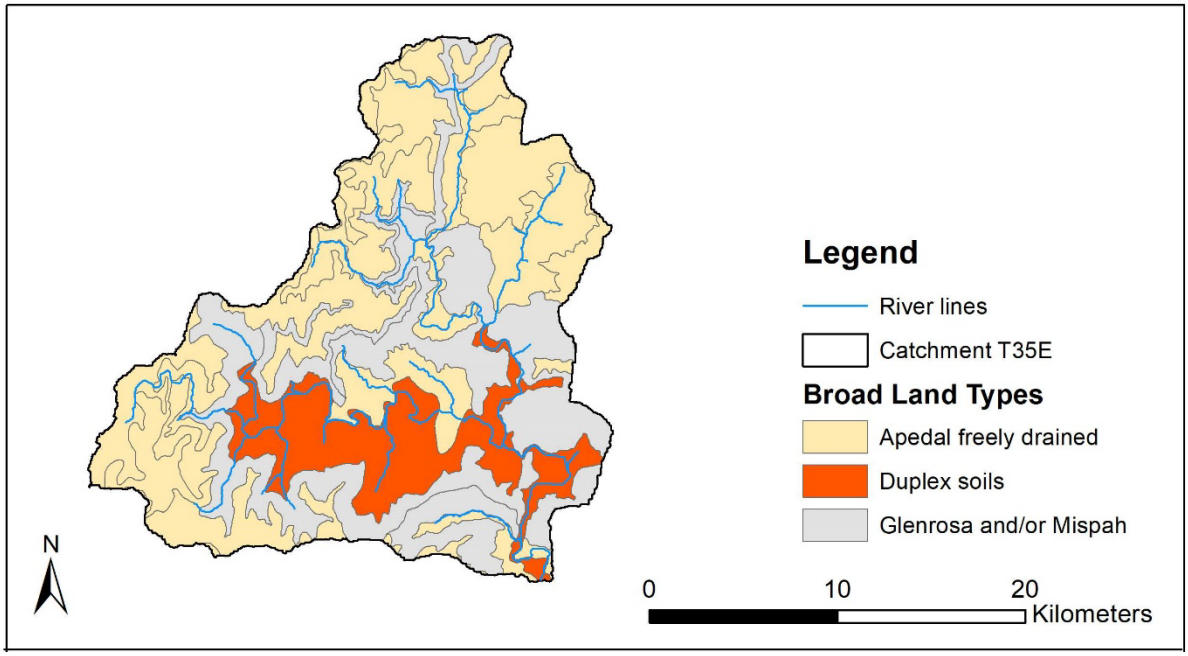

Figure 6.16: Land Type map of the Tsitsa catchment T35E.

Next, textural and soil hydraulic parameter values were assigned to each HYDROSOIL component. Although similar reasoning was followed in the assignment of the required parameter values to both the Land Type and HYDROSOIL data models, the HYDROSOIL data model used additional soil analytical/sample data and hydrological pedotransfer functions based on the spatial distribution and hydropedological grouping of soils (Van Tol et al., 2013; Van Tol & Le Roux, 2019). Soil parameter values (Table 6.11) were assigned to the HYDROSOIL units of catchment T35E (Figure 6.17).

Table 6.11: Description and reasoning used to assign soil parameter values to each HYDROSOIL component of Tsitsa catchment T35E.

Soil parameter	Reasoning
Number of layers in the soil	Two soil layers/horizons were incorporated into each soil component of the HYDROSOIL map.
Depth from soil surface to bottom of layer (mm)	The minimum, maximum and mean depth descriptions/classes in the Land Type database and Schulze (2007) were used to assign depth to each HYDROSOIL unit as follows: Shallow recharge = 290 mm; Deep recharge = 1 180 mm; A/B horizon interflow = 844 mm; Soil/ bedrock interflow = 930 mm; Shallow responsive = 300 mm; Saturated responsive = 1 180 mm.
Maximum rooting depth of soil profile (mm)	As above.

Soil parameter	Reasoning
Soil Hydrologic Group (A, B, C, D) in terms of runoff potential, Soil Group A = low, B = moderately low, C = moderately high, D = high.	Soil hydrological groups were based on Van Tol et al. (2013) as follows: Shallow recharge = B; Deep recharge = A; A/B horizon interflow = C; Soil/ bedrock interflow = C; Shallow responsive = D; Saturated responsive = D.
Available water capacity of the soil layer (mm H_2O/mm soil)	Schulze (2007) calculated plant available water content for each Land Type as the difference between water content at field capacity and permanent wilting point.
Saturated hydraulic conductivity (mm/hr)	Average values for each HYDROSOIL class were calculated using profile sample data.
Bulk density (Mg/m^3 or g/cm^3)	Average values for each HYDROSOIL class were calculated using profile sample data.
Soil albedo (non-dimensional value between 0 and 1)	Albedo values were assigned to broad soil patterns in the Land Type database ranging between 0.25 for light coloured sands to 0.7 for dark clays.
Clay content with diameter of < 0.002 mm (% soil weight)	Average clay values for each HYDROSOIL class were calculated using profile sample data. Clay content in the B-horizon was assigned to each Land Type by adjusting the clay values of the A-horizon to clay-factors (Le Roux et al., 2022).
Silt content with diameter of 0.05-0.002 mm (% soil weight)	Average silt values for each HYDROSOIL class were calculated using profile sample data. Due to the lack of data, silt content for B horizons were assigned values between 10-22.5%, increasing with increase in clay as follows (Le Roux et al., 2022): percentage of HYDROSOIL unit with <= 6% clay = 10% silt; 6.1-15% clay = 15% silt; 15.1-25% clay = 17.5% silt; 25.1-35% clay = 20% silt; 35.1-55% clay = 22.5% silt.
Sand content with diameter of > 2 mm (% soil weight)	Sand content for A and B horizons were assigned as follows: Sand = 100% – (%clay + %silt).
Rock fragment content (% soil weight)	Rock content was based on the agricultural restriction/rock (MB) classes in the Land Type database (2012) as follows: MB0=0% (no rock); MB1=20%; MB2=50%; MB3=20%; MB4=100% (no soil).
Organic carbon content (% soil weight)	Average values for each HYDROSOIL class were calculated using profile sample data.
USLE K factor in SI units t/ha per unit 'erosivity'	Erodibility units (Le Roux et al., 2008) were assigned to each Land Type.

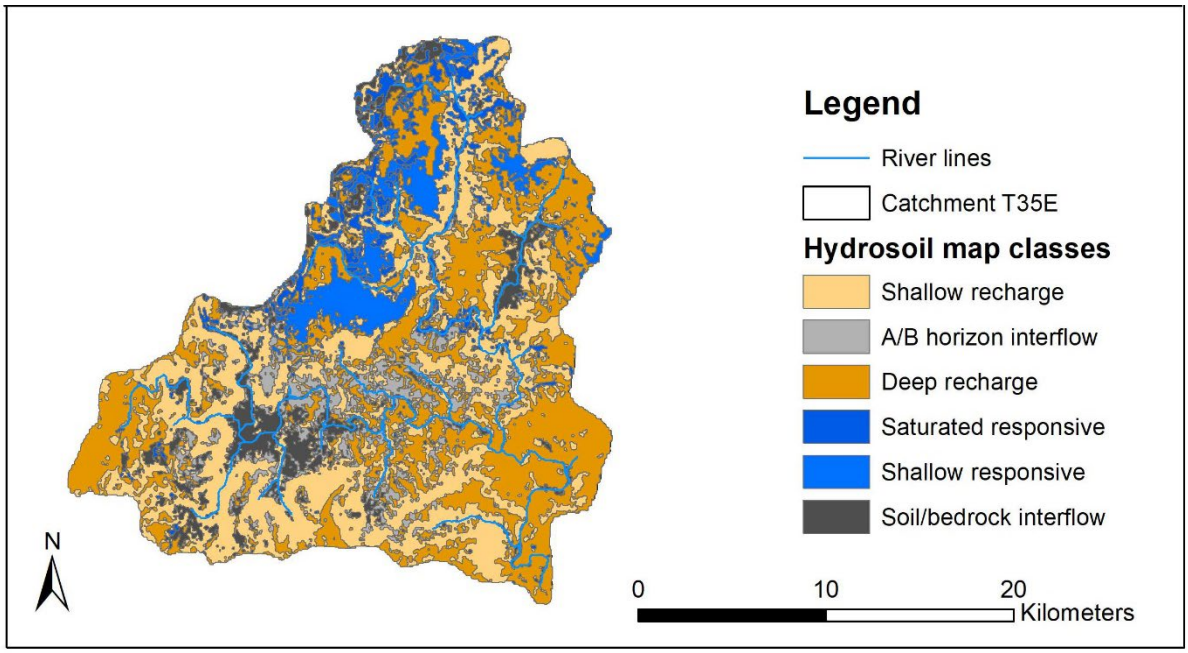

Figure 6.17: HYDROSOIL map of the Tsitsa catchment T35E.

The overlay of land cover and soil maps created 610 and 616 HRUs for the Land Type and HYDROSOIL maps, respectively.

SWAT also requires climate parameters including precipitation, temperature, solar radiation, relative humidity and wind speed. Daily precipitation and temperature data were acquired from two meteorological stations of the ARC Agroclimatology Database (2012) over a six-year period (Figure 6.18a). In addition, Weather Generator input files consist of weather statistics including precipitation, temperature, solar radiation, relative humidity and wind speed. Weather Generator files are needed by SWAT to generate representative daily climate data for simulated catchments in two instances: when the user specifies that simulated weather will be used or when measured data is missing. Weather Generator files were created by acquiring and interpreting climate data from the two weather stations. Using the SWAT Weather Database (Essenfelder, 2016), the Weather Generator files were prepared covering the period 2001-2020.

Hydrological parameters included flow contributions from the Tsitsa River inlet (Figure 6.18b). No reservoirs were present. The Penman-Monteith equations were used to calculate potential (and actual) evapotranspiration for each catchment, considering soil moisture and crop development (Aouissi et al., 2016).

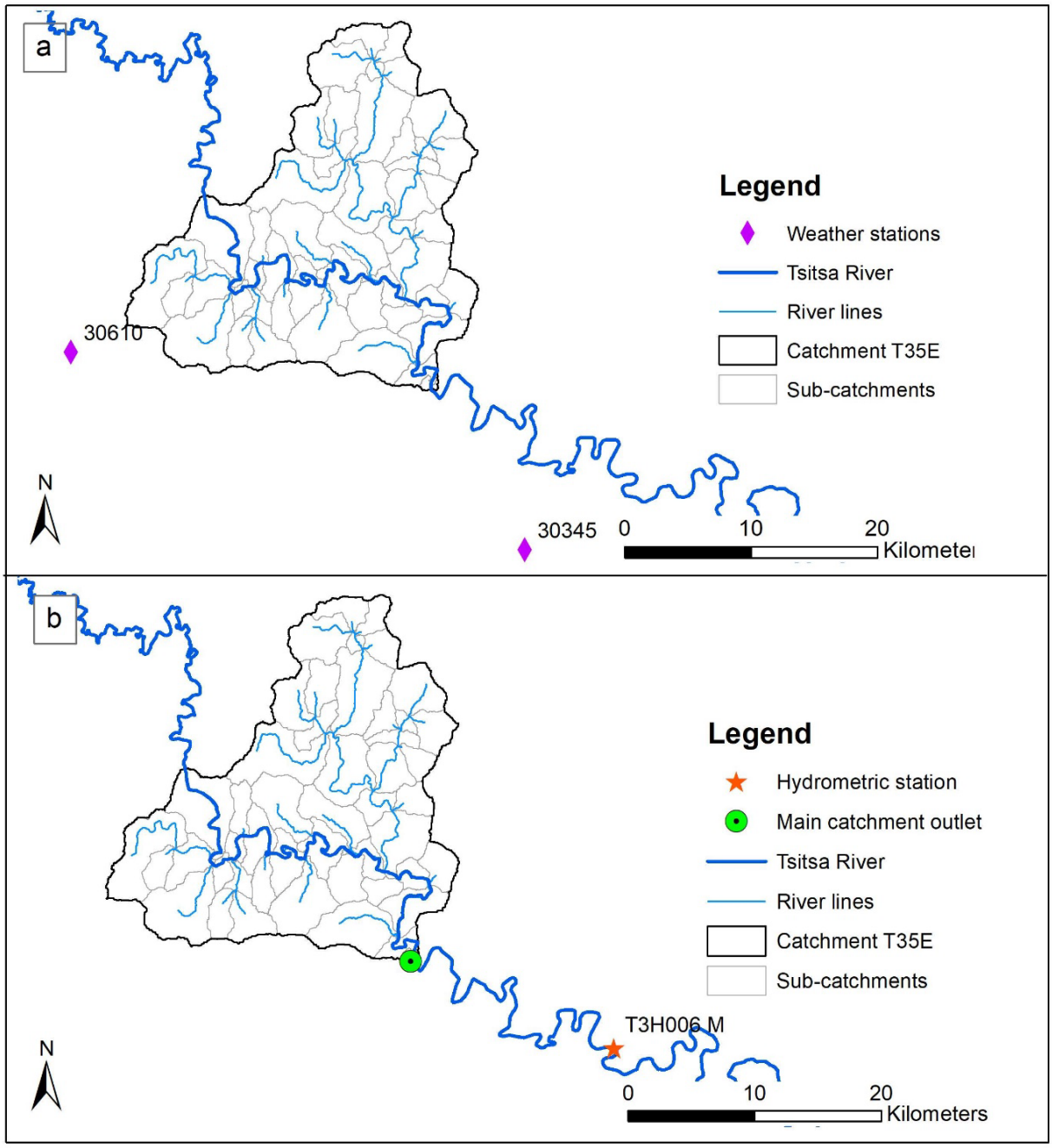

Figure 6.18: Tsitsa catchment T35E illustrating the a) weather station, and b) hydrometric, as well as the main Tsitsa River inlet and outlet locations.

Management practices include tillage, nutrient applications, irrigation schedules and harvest. These practices affect the water balance and sediment/nutrient load generation through the impacts of the plant growth cycle on evapotranspiration. Due to the lack of data on management practices, however, parameter values were assigned to represent each management practice according to values provided in the SWAT database.

Model simulations and validation

Finally, model simulations were conducted over five years, preceded by a three-year warm-up period to get the hydrological cycle fully operational. Catchment T35E was therefore simulated twice using the same weather data, over the same timeframes. The reason for duplicating the application of SWAT is to compare the streamflow and sediment yield results of the HYDROSOIL versus the Land Type database, which allows appraisal of the performance of the HYDROSOIL data. Model performances of streamflow were determined by the Nash-Sutcliffe Efficiency (NSE), as well as the coefficient of determination (R^2). A percent deviation method (Dv) (Martinec & Rango, 1989) was used as a measure of goodness-of-fit between simulated and measured streamflow data at the main catchment outlets. It is noteworthy here that the closest hydrometric station (weir) is more than 20 km downstream of the main catchment outlet.

6.3.3 Results and discussion

Digital soil mapping

The HYDROSOIL maps were generated using the multinomial logistic regression algorithm with the different sampling techniques – random sampling (Figure 6.19), K-means (Figure 6.20) and Conditioned Latin Hypercube (Figure 6.21).

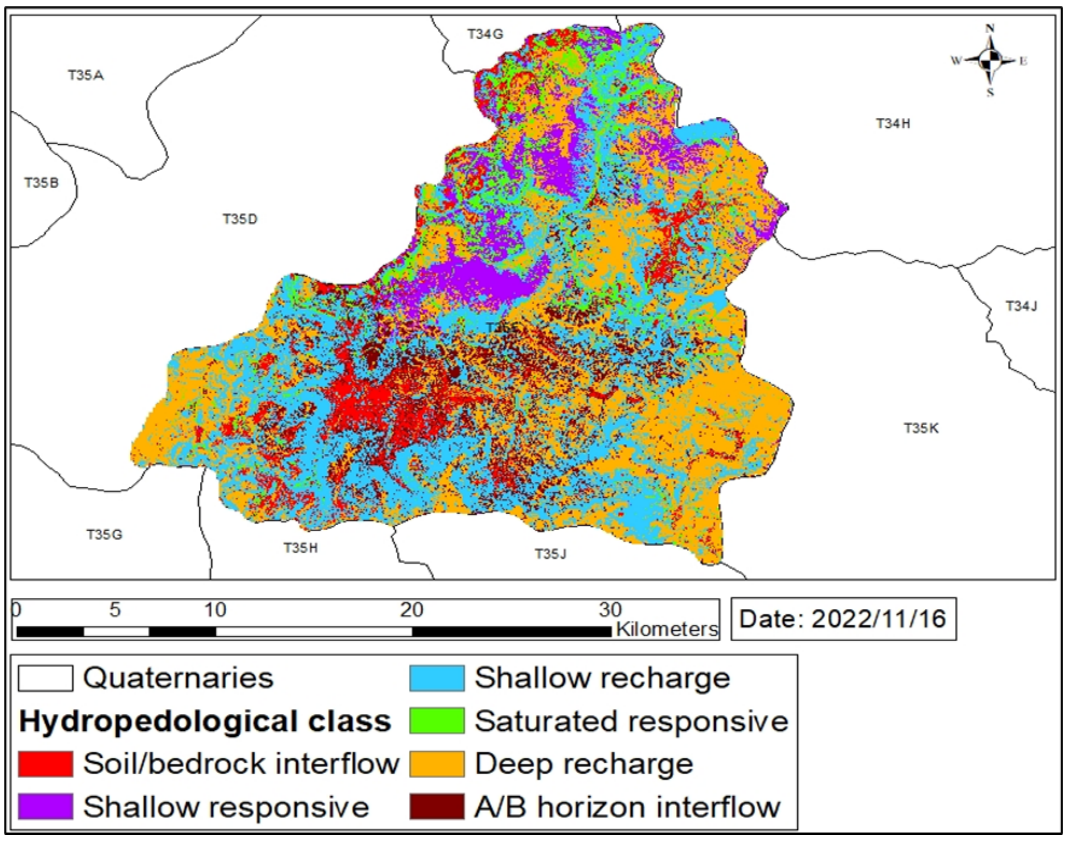

Figure 6.19: Hydrological soil map (HYDROSOIL) of the Tsitsa catchment created using the multinomial logistic regression algorithm with random sampling.

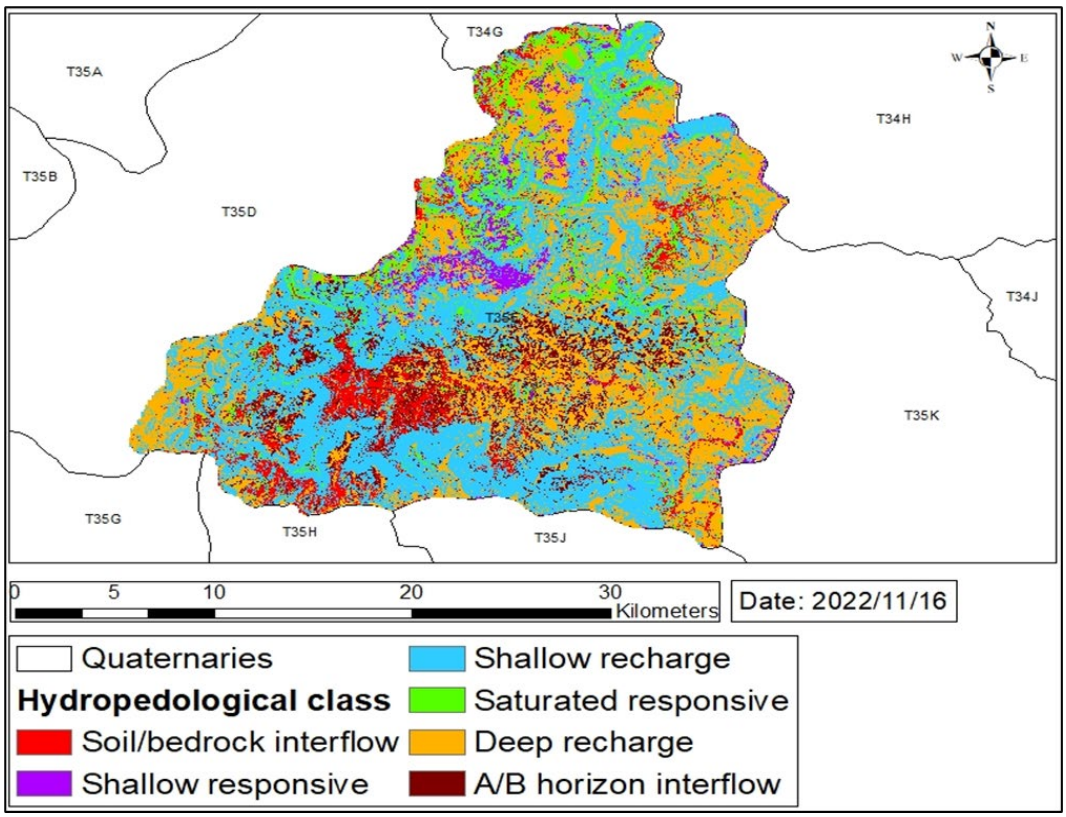

Figure 6.20: Hydrological soil map (HYDROSOIL) of the Tsitsa catchment created using the multinomial logistic regression algorithm with K-means clustering.

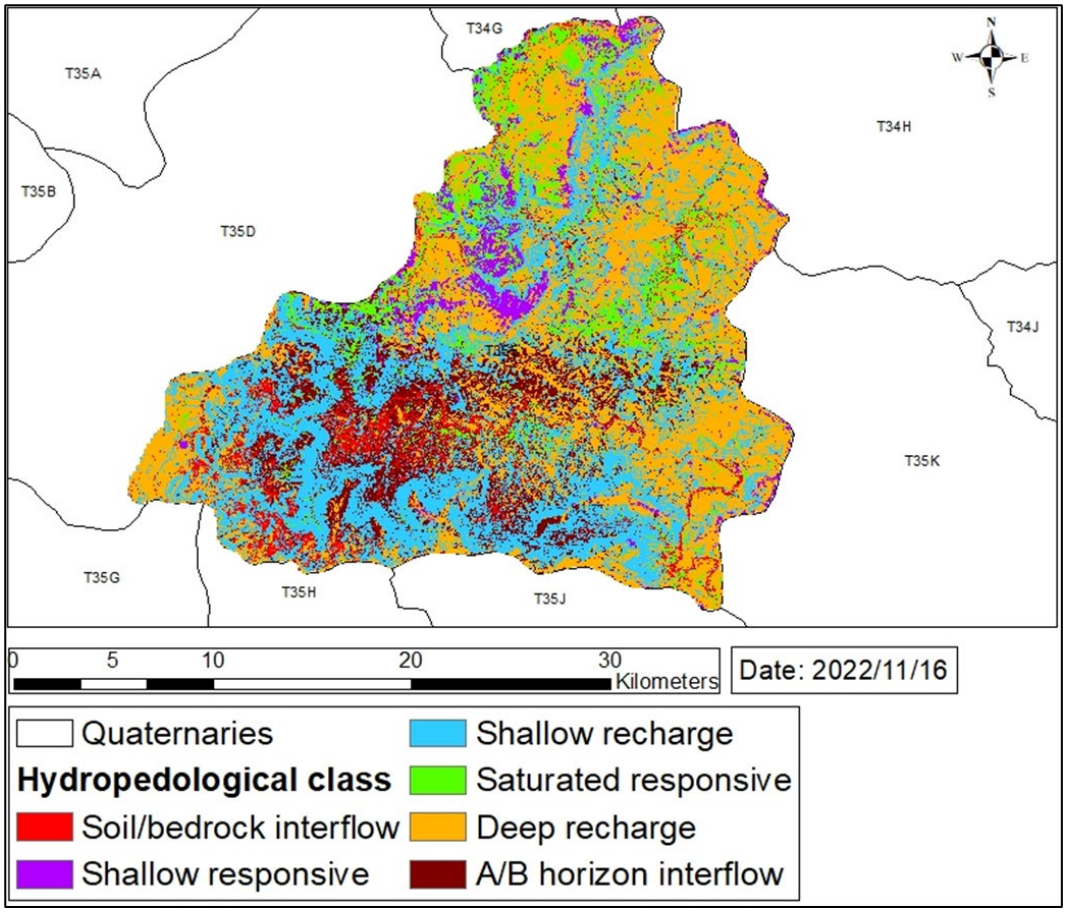

Figure 6.21 Hydrological soil map (HYDROSOIL) of the Tsitsa catchment created using the multinomial logistic regression algorithm with Conditioned Latin Hypercube sampling.

The HYDROSOIL map generated by multinomial logistic regression with random sampling resulted in a Kappa coefficient that showed substantial agreement (0.66), while the total evaluation point accuracy was also high (74.2%; Table 6.12). The HYDROSOIL map generated by multinomial logistic regression with K-means clustering resulted in a Kappa coefficient that showed moderate agreement (0.58), while total evaluation point accuracy was also moderate (68.5%; Table 6.13).

Table 6.12: Confusion matrix of the HYDROSOIL map generated by multinomial logistic regression method with stratified random sampling.

		Map units								
		Deep recharge	Shallow recharge	Interflow A/B	Interflow Soil/bedrock	Shallow responsive	Saturated responsive	Total	Correct	%
Observations	Deep recharge	22	4	3	2			31	22	71.0
	Shallow recharge	3	17	2	1			23	17	73.9
	Interflow A/B	2	1	15				18	15	83.3
	Interflow Soil/bedrock		3		3			6	3	50.0
	Shallow responsive	1	1			4		6	4	66.7
	Saturated responsive						5	5	5	100.0
	Total	28	26	20	6	4	5	89		
	Correct	22	17	15	3	4	5		66	
	%	78.6	65.4	75.0	50.0	100.0	100.0			74.2

Table 6.13: Confusion matrix of the HYDROSOIL map generated by multinomial logistic regression method with K-means clustering.

		Map units								
		Deep recharge	Shallow recharge	Interflow A/B	Interflow Soil/bedrock	Shallow responsive	Saturated responsive	Total	Correct	%
Observations	Deep recharge	22	4	3	2			31	22	71.0
	Shallow recharge	5	15		3			23	15	65.2
	Interflow A/B	3	2	13				18	13	72.2
	Interflow Soil/bedrock	1	1		3		1	6	3	50.0
	Shallow responsive					3		3	3	100.0
	Saturated responsive	2	1				5	8	5	62.5
	Total	33	23	16	8	3	6	89		
	Correct	22	15	13	3	3	5		61	
	%	66.7	65.2	81.3	37.5	100.0	83.3			68.5

The HYDROSOIL map generated by multinomial logistic regression with Conditioned Latin Hypercube sampling resulted in high the Kappa (0.64) and total evaluation point accuracy (72.8%; Table 6.14).

Table 6.14: Confusion matrix of the HYDROSOIL map generated by multinomial logistic regression method with Conditioned Latin Hypercube sampling.

		Map units								
		Deep recharge	Shallow recharge	Interflow A/B	Interflow Soil/bedrock	Shallow responsive	Saturated responsive	Total	Correct	%
Observations	Deep recharge	22	3		2		1	28	22	78.6
	Shallow recharge	2	18	3	1		0	24	18	75.0
	Interflow A/B	5	3	15				23	15	65.2
	Interflow Soil/bedrock	1	1	1	3			6	3	50.0
	Shallow responsive	1				4		5	4	80.0
	Saturated responsive		1				5	6	5	83.3
	Total	31	26	19	6	4	6	92		
	Correct	22	18	15	3	4	5		67	
	%	71.0	69.2	78.9	50.0	100.0	83.3			72.8

From the total evaluation point accuracy and Kappa coefficient, the HYDROSOIL maps that were created using random sampling (Kappa = 0.66, point accuracy = 74.2%) and Conditioned Latin Hypercube sampling (Kappa = 0.64, point accuracy = 72.8%) were judged to be most acceptable for usage in hydrological modelling. However, the HYDROSOIL map that was created using K-means clustering was also judged to be acceptable even if with a Kappa coefficient of 0.58 and point accuracy of 68.5%.

The three HYDROSOIL maps created were an improvement on the soil association with depth maps (Kappa < 0.50 and point accuracy < 68%) created by Du Plessis et al. (2020). The HYDROSOIL map that was created using random sampling was judged to be the most accurate map and therefore should be used for hydrological modelling.

Streamflow simulation results

Graphically, streamflow simulations with Land Type and HYDROSOIL data appear similar, with occasional steep peaks that can be associated with wetter months (Figure 6.22). For simulation with Land Type data, monthly streamflow at the main catchment outlet ranges between 0.1 m^3/s in September 2010 to 39.5 m^3/s in January 2011, with an average of 9.0 m^3/s during the simulation period (2008-2012). For simulation with HYDROSOIL data, streamflow at the main catchment outlet ranges between 0.2 m^3/s in August 2010 to 42.3 m^3/s in in January 2011 with an average of 11.3 m^3/s. Streamflow simulated by the HYDROSOIL data model is 20% higher than the Land Type data model.

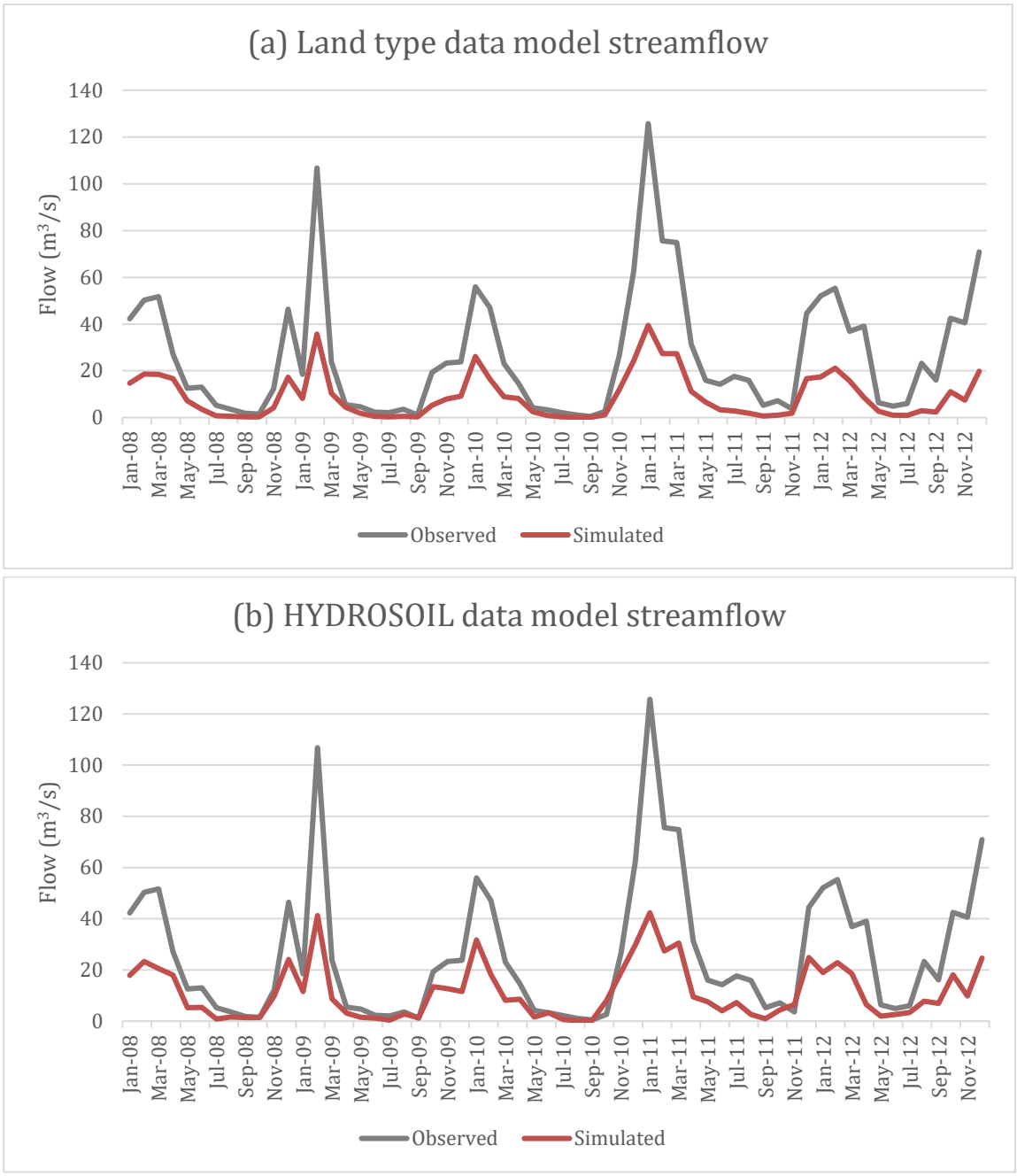

Figure 6.22: Comparison of observed monthly streamflow (in m^3/s) with the (a) Land Type and (b) HYDROSOIL data models in the Tsitsa catchment T35E (2008-2012).

The HYDROSOIL data model was slightly superior compared to the Land Type data model during validation, as shown by the higher NSE, R^2 and Dv values (Table 6.15). The HYDROSOIL data model underpredicted streamflow by 131.4% as determined by Dv, the goodness-of-fit expressed by NSE was -3.62% and R^2 was 95%. The Land Type data model under-predicted streamflow by 191.54% as determined by Dv, the goodness-of-fit expressed by NSE was -5.81% and R^2 was 97%. It is noteworthy here that both models would probably be more accurate if the hydrometric station (weir) was not located more than 20 km downstream of the catchment outlet. Underpredictions would have been considerably less due to additional flow contributions from the Tsitsa River and its tributaries upstream of the weir. The main reason for the slightly superior performance of the HYDROSOIL data model is due to differences between soil datasets.

Table 6.15: Performance metrics (R^2, NSE and Dv in %) obtained from monthly streamflow validation for National and HYDROSOIL data models for the Tsitsa catchment T35E.

Data model	Land Type	HYDROSOIL
R^2 accuracy (%)	0.97	0.95
NSE accuracy (%)	-5.81	-3.62
Dv underprediction (%)	-191.54	-131.69

NSE = Nash-Sutcliffe Efficiency; Dv = percent deviation

Sediment yield results

SWAT simulations with Land Type and HYDROSOIL data show similar trends in sediment load estimations, with occasional steep peaks that can be associated with wetter months (Figure 6.23). For simulation with Land type data, sediment load at the main catchment outlet ranges between 0.0 t in September 2010 to 4 746.0 t in January 2011 with an annual average load of 958.6 t/yr and a total load of 57 518.3 t during the simulation period (2008-2012). For simulation with HYDROSOIL data, sediment load at the main catchment outlet ranges between 3.8 t in August 2010 to 6 976.0 t in January 2010 with an average load of 1 889.6 t/yr and a total load of 113 374.3 t during the simulation period (2008-2012). The sediment load simulated by the HYDROSOIL data model is 50% higher than the Land Type data model.

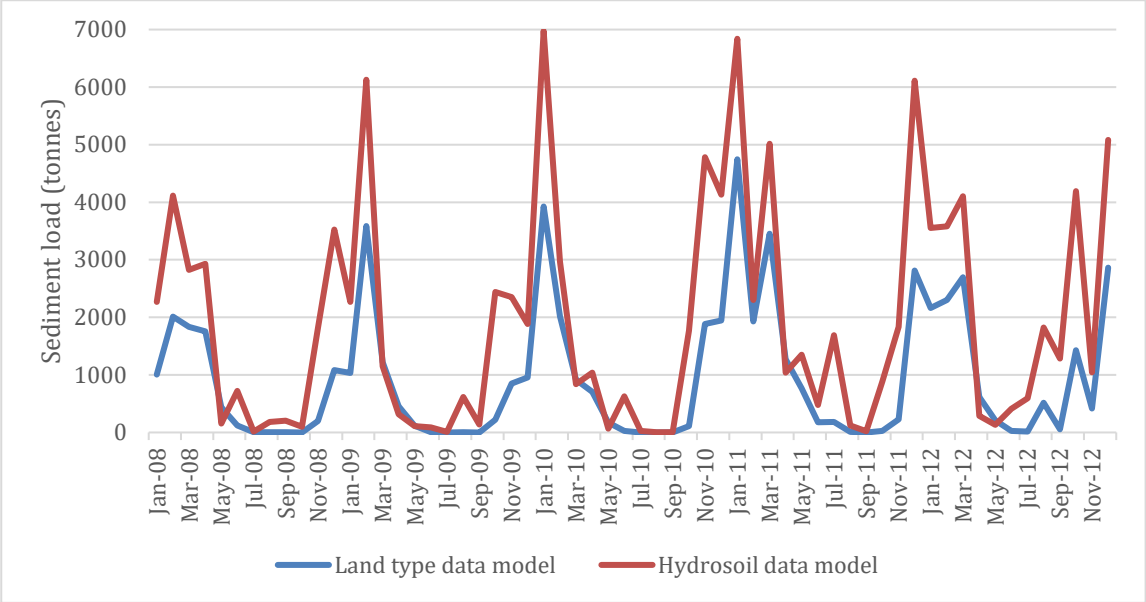

Figure 6.23: Comparison of total annual sediment load (in metric t) for SWAT simulations with the Land Type and HYDROSOIL data models in the Tsitsa catchment T35E (2008-2012).

Both simulations show that the sediment load is mainly high during the summer rainfall season (extending from October to April) (Figure 6.24). Low rainfall months (extending from May to August) have low sediment loads due to low or no rainfall during winter in the catchment.

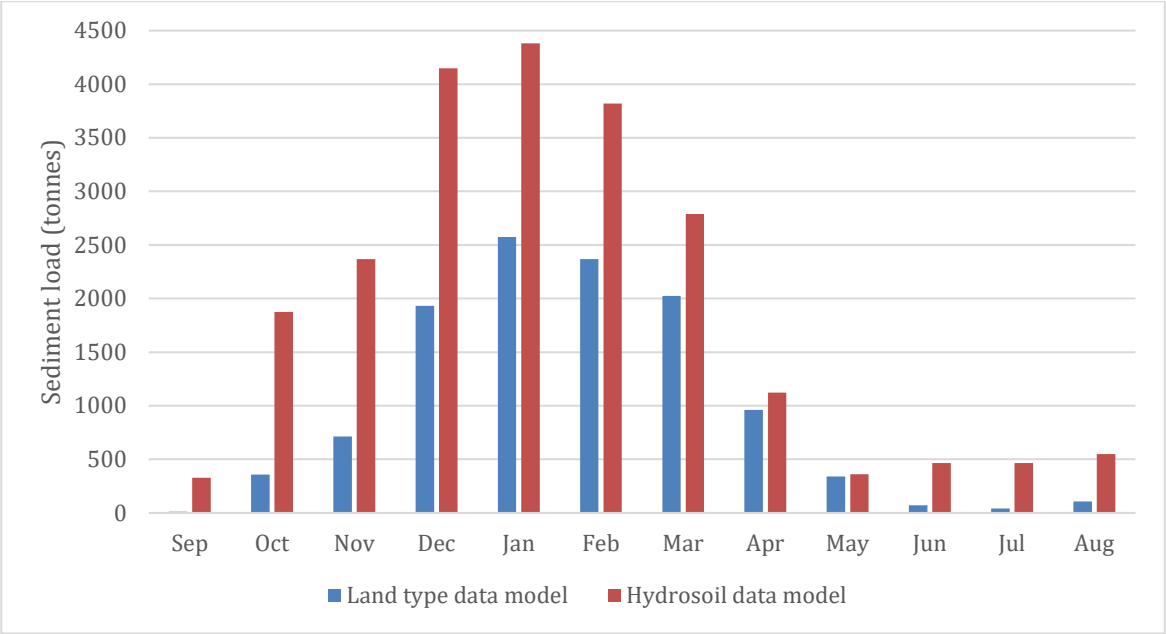

Figure 6.24: Comparison of monthly average sediment load (in metric t) for SWAT simulations with the Land Type and HYDROSOIL data models in the Tsitsa catchment T35E (2008-2012).

Although the average sediment yield of the Land Type and HYDROSOIL data models are similar (5.7 and 4.6 t/ha/yr respectively), the models identified different sediment source areas (Figure 6.25). The Land Type data model identifies the central and lower half of the catchment as important sediment source areas (5-20 t/ha/yr), whereas the HYDROSOIL data model identifies high sediment yield values in the upper catchment areas. The largest difference occurs in sub-catchment 16, where the Land Type data model simulates very low sediment yield (0.3-0.5 t/ha/yr), whereas the HYDROSOIL model simulates very high sediment yield (20-25 t/ha/yr). The spatial differences in sediment yield between the data models are attributed to soil parameter variances since the topography, land cover and climate parameters in both data models are similar.

6.3.4 Discussion of data model differences

The spatial differences in sediment yield between the data models are attributed to soil distribution patterns and parameter variances, especially the soil hydrological parameters. Although similar reasoning was followed in the assignment of the required parameter values to both the Land Type and HYDROSOIL data models (see Table 6.10; Table 6.11), the HYDROSOIL data model used soil analytical/sample data and hydrological pedotransfer functions based on the spatial distribution and hydropedological grouping of soils (Van Tol et al., 2013; Van Tol & Le Roux, 2019). The Land Type map illustrates the upper catchment areas consist of mainly freely drained apedal soils (hydrological group A and B) where infiltration rates are relatively high and runoff and subsequent erosion is low. In contrast, the HYDROSOIL map illustrates that large parts of the upper catchment consist of shallow responsive soils (hydrological group C and D) where infiltration rates are relatively low and runoff and subsequent erosion is high. In these areas, streamflow simulated by the HYDROSOIL data model is approximately > 20% higher than the Land Type data model. Higher streamflow accounts for relatively high sediment yield in these sub-catchments. As a result, the Land Type data model identifies the central and lower half of the catchment as important sediment source areas, whereas the HYDROSOIL data model identifies high sediment yield values in the upper catchment areas (Figure 6.25).

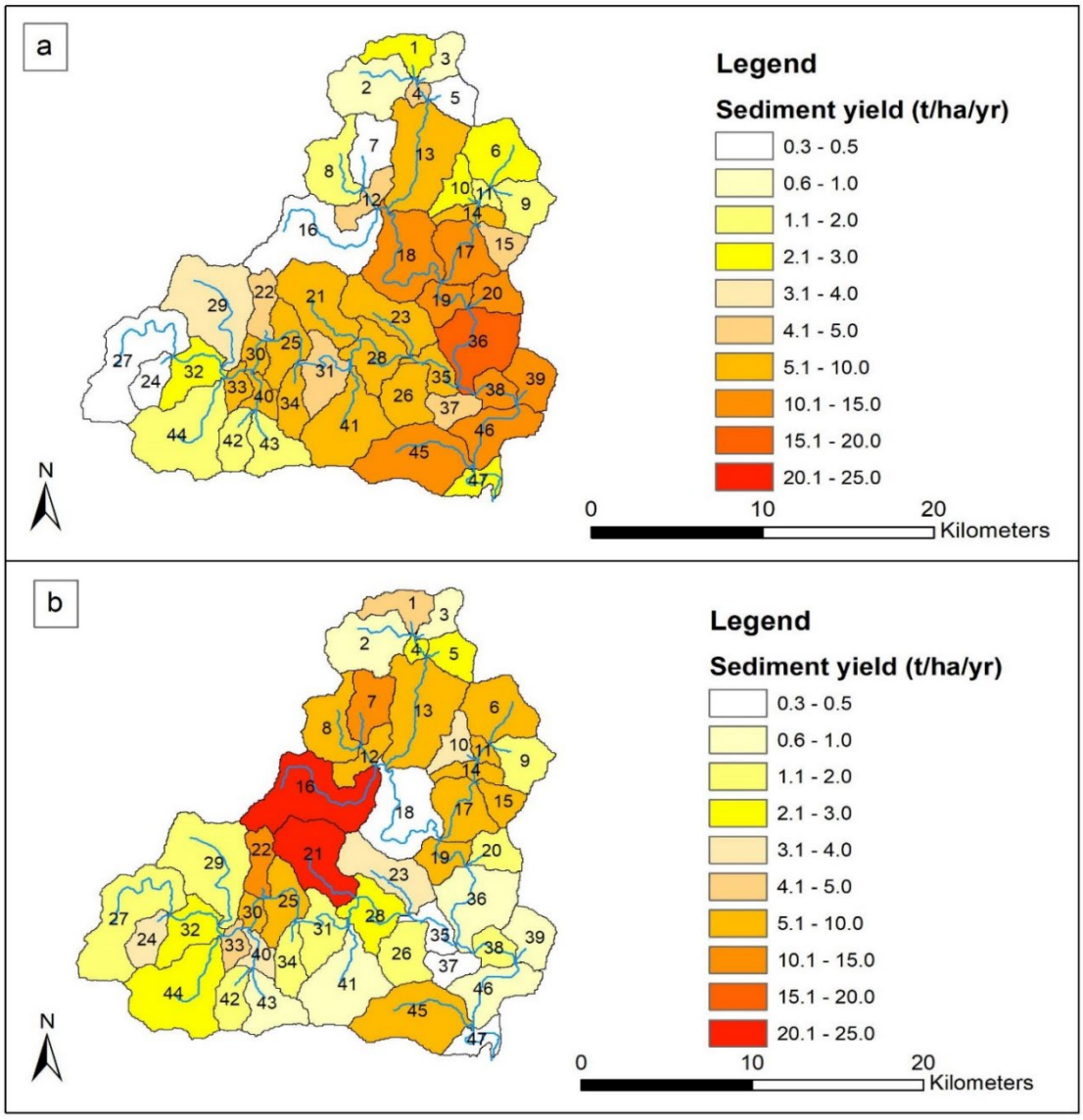

Figure 6.25: Spatial comparison of average annual sediment yield (in t/ha/yr) simulated by the SWAT model with the (a) Land Type and (b) HYDROSOIL data models in Tsitsa catchment T35E.

Overall, the sediment load simulated by the HYDROSOIL data model is approximately 50% higher than the Land Type data model (113 374.3 t and 57 518.3 t respectively). However, results should not be interpreted as absolute values due to the absence of measured sediment data for validation. ArcSWAT utilises the Modified USLE equation (Williams & Brendt, 1977) that excludes gully erosion processes. The USLE is an empirical model that was developed from runoff plots (22 m long and 2 to 3 m wide) and rainfall simulation experiments (Wischmeier & Smith, 1978). Plots capture soil loss from rill-interrill erosion but not from gully erosion which occurs at larger scales (> 0.03 km^2) (Kirkby et al., 2003). Therefore, the SWAT model underestimates sediment yield in sub-catchments where gullies are prominent.

Nevertheless, in terms of streamflow, the HYDROSOIL data model was superior compared to the Land Type data model during validation, especially in terms of Dv values. The Land Type data model underpredicted streamflow by 191.54%, whereas the HYDROSOIL data model underpredicted streamflow by 131.4%. As mentioned above, underpredictions would have been less if the hydrometric station (weir) was not located more than 20 km downstream of the catchment outlet, due to additional flow contributions from the Tsitsa River and its tributaries upstream of the weir. Despite these limitations, the HYDROSOIL data model was superior compared to the Land Type data model during validation and appears to be more efficient than Land Type data in modelling high sediment yield values in the severely eroded catchment.

6.3.5 Conclusions

Digital soil mapping proved to be useful in mapping the Tsitsa catchment. It was however evident that the type of sampling that is done to compile the datasets as well as the method used for the prediction, affects the accuracy of the maps. The random sampling was the better sampling method compared to the K-means and the Conditioned Latin Hypercube sampling, resulting in an overall accuracy of 74.2% and Kappa coefficient of 0.66. Therefore, it was clear that using random sampling generated the most acceptable HYDROSOIL map to be used for hydrological modelling.

In terms of streamflow, the HYDROSOIL data model was superior compared to the Land Type data model during validation, especially in terms of Dv values. The Land Type data model under-predicted streamflow by 191.54%, whereas the HYDROSOIL data model under-predicted streamflow by 131.4%. Underpredictions would have been substantially less if the hydrometric station (weir) was not located more than 20 km downstream of the catchment outlet. Overall, streamflow simulated by the HYDROSOIL data model is approximately > 20% higher than the Land Type data model, whereas sediment load simulated by the HYDROSOIL data model is approximately 50% higher than the Land Type data model. The HYDROSOIL data model appears to be more efficient than Land type data in modelling high sediment yield values in the severely eroded catchment. HYDROSOIL is an important step forward in the application of hydrological models to assist agricultural water management.

Further refinement will be possible given additional research. Firstly, it is recommended to expand the HYDROSOIL map in the Tsitsa catchment to ensure that flow monitoring points spatially overlay with catchment outlet points for calibration and validation of model simulations with measured data. Furthermore, stream channel processes and hydrological structures need to be characterised, allowing deposition of excess sediment depending on the carrying capacity and/or sediment storages where connectivity is reduced (Chen & Mackay, 2004). Ancillary information regarding management practices in the catchment should also be incorporated, including tillage operations, nutrient applications, irrigation scheduling and harvesting operations.

6.4 THE MAPPING AND HYDROLOGICAL MODELLING OF THE GOUKOU RIVER CATCHMENT

6.4.1 Introduction

This section focusses on the creation of the HYDROSOIL and hydrological modelling of the Goukou River catchment, in the southern Cape region. A recent ecological state report for the Goukou River catchment indicates that the river condition has rapidly deteriorated from the source to the sea (CSIR, 2011; Nzonda, 2016). The Goukou catchment is also has reasonably pristine wetland areas (Nzonda, 2016; Royal HaskoningDHV, 2018). Therefore understanding the hydrological processes in this catchment is very important to enable adequate mapping of this river system to avoid further deterioration. The use of the HYDROSOIL for this catchment is showcased through the use of the JAMS model, to show that different hydrological models could also use the HYDROSOIL map.

Two distinct model runs utilising the JAMS hydrological framework were compared. The first model run used published soils information based on the Land Type surveys. The second run introduced the HYDROSOIL into the model without redefining the HRUs.

The primary objective was to conduct a statistical comparison of the outcomes from these two runs. This would shed light on the potential impact that incorporating new soils information may have on hydrological modelling and prediction. By systematically comparing the results, we seek to discern any significant differences and draw insights into how the updated soils information influences the overall hydrological behaviour predicted by the model.

6.4.2 Materials and methods

The Goukou catchment

The Goukou catchment is situated in the southern Cape of South Africa. The river flows from the Langeberg Mountains in the north, southwards to Stilbaai at the coast. The river is only approximately 64 km long. The catchment comprises five quaternary catchments (Figure 6.26). The last 19 km of the river is an estuary of high ecological importance (CSIR, 2011). The catchment stretches from sea level, where it enters the sea at Stilbaai, to 1 458 m above sea level in the mountains north at Riversdale (Figure 6.27). The Goukou River reaches sea level about 18 km inland, where it meets the estuary. The land use includes mostly forestry, formal irrigated grazing, irrigated vineyards, wheat and other rain-fed crop production systems.

The Goukou system is spread out over three distinct Land Types (Figure 6.28). The coastal belt system is an area of small mountains and recent deep sands which is a mixture of river deposits and aeolian sands. Further inland a system with deep carved valleys exists. In the upper reaches of the Goukou a mountain landscape exists, with distinct occurrences of large wetlands in an undulating landscape. The three Land Types also have large differences in soil occurrence, and large differences related to the runoff pathways, whether it be surface, subsurface or deep drainage transvers of water in the system (Table 6.16). Land Type Ib168 in the north, Dc32 in the middle section and Fc17 in the south or coastal zone are the more prominent Land Types (Figure 6.28).

Table 6.16: The dominant soil parameters per macro positions in the Goukou catchment (Land Type Survey Staff, 1972-2002).

Land Type	Soil depth	Soil type	Clay A %	Clay deep %	Depth restriction	Texture class
Ib168	0-300	Rock and Cartref	0-6	12	Rock	me/coSa-LmSa
Dc32	450	Va	15-35	55	Vr, Ca	Cl
Fc17	1200	Mispah	0-10	6	ka	meSa

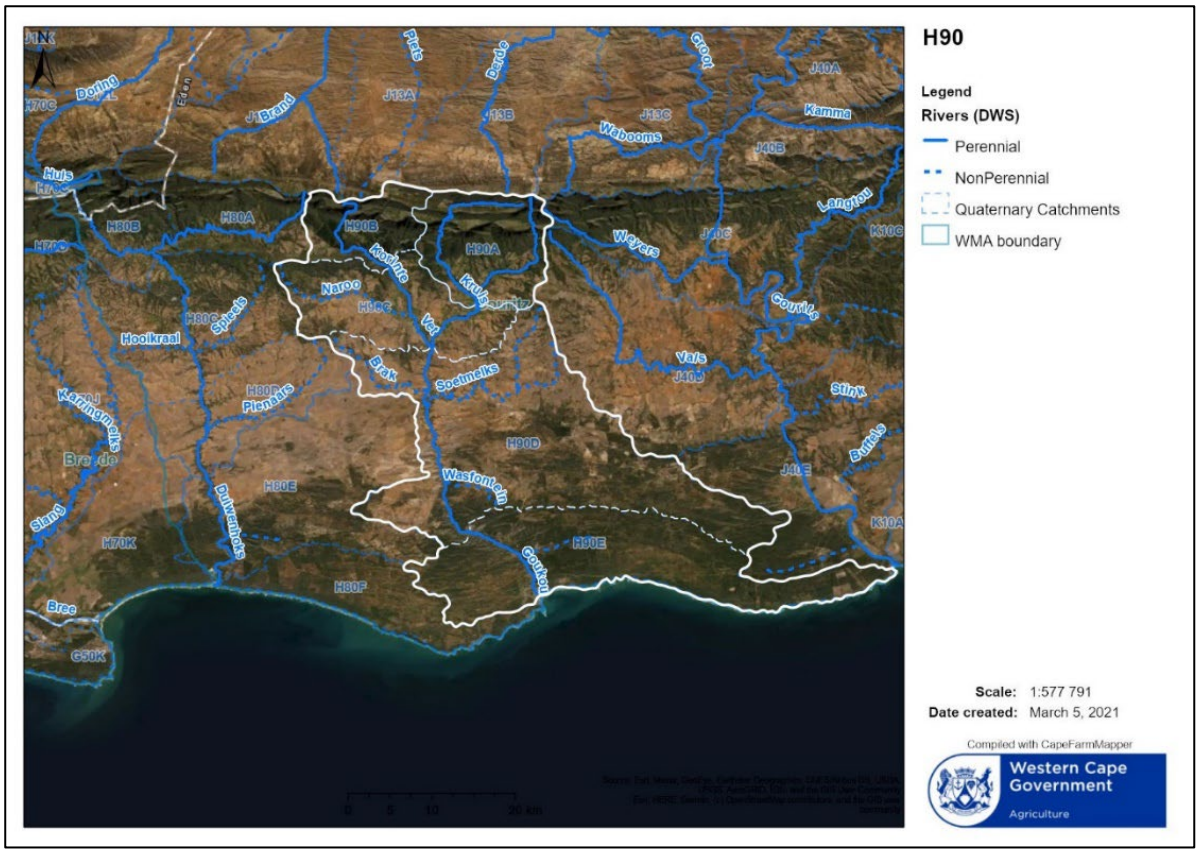

Figure 6.26: The Goukou catchment (H90, solid white line) and its five quaternary catchments (broken white lines).

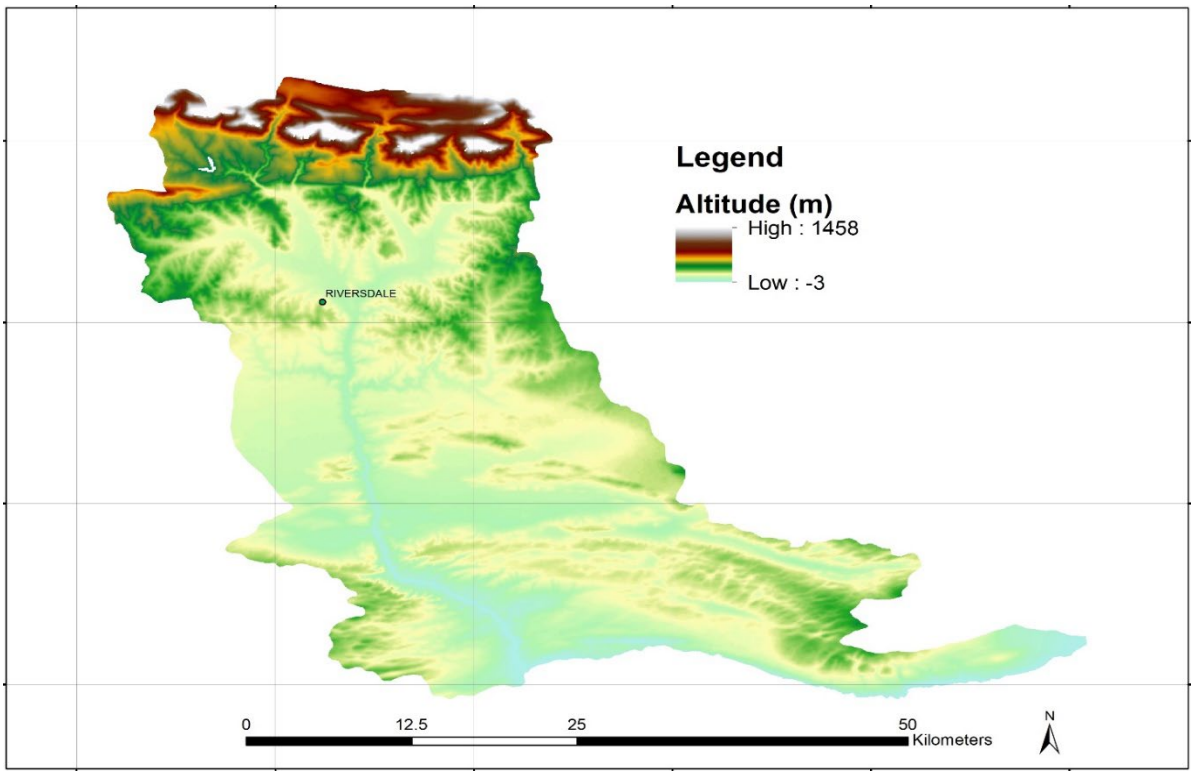

Figure 6.27: The altitude of the Goukou catchment.

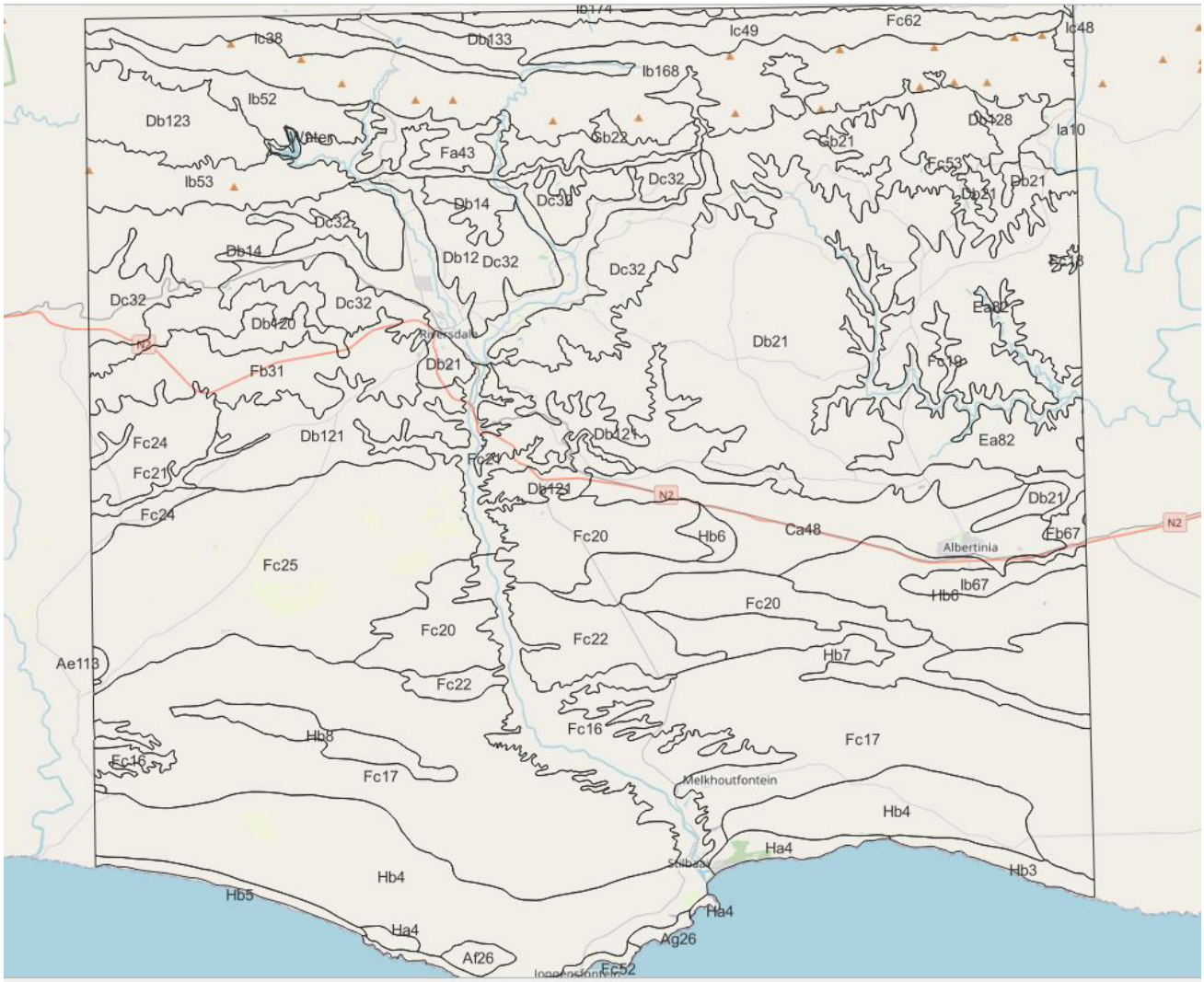

Figure 6.28: The Land Types of the Goukou catchment (Land Type Survey Staff, 1972-2022).

HRU mapping with the SWAT divided the catchment into three sub-catchments (Figure 6.29). The delineation is based on second-order streams. The upper Goukou, including the mountain region, is therefore represented in sub-catchment 1, while most of the coastal system is represented in sub-catchments 2 and 3. It is important to see from this result, the fact that the Goukou catchment, HRU 1, makes a contribution directly to the estuary while HRU 3 has a multitude of small streams connecting to the estuary. For our purposes, the JAMS modelling focussed mainly on sub-catchment 1.

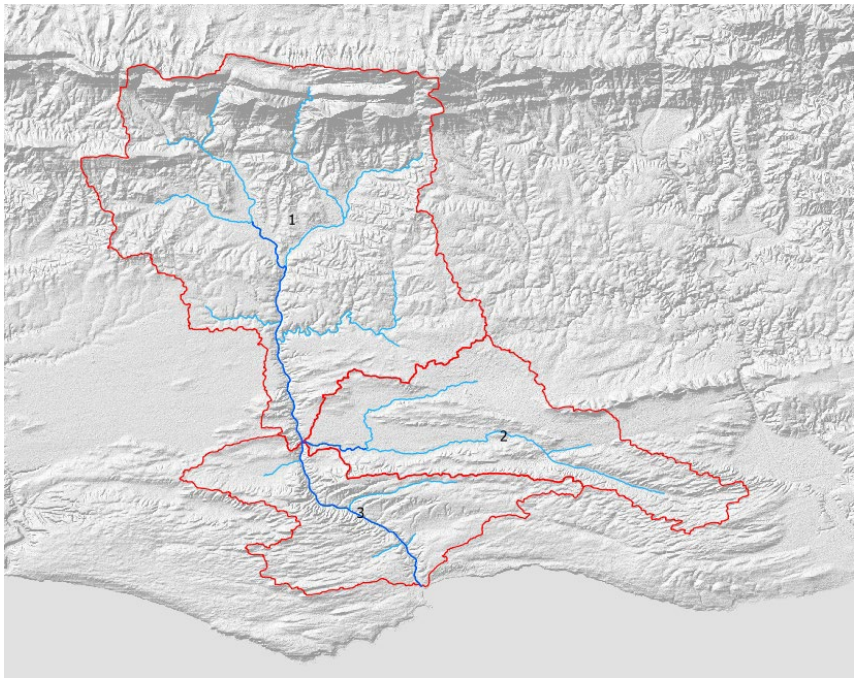

Figure 6.29: Hydrological response units as mapped with the SWAT in QGIS.

Digital soil mapping

For the digital soil mapping methodology, please see Section 6.1.

Soil parameterisation

Soil parameterisation for the Goukou JAMS hydrological model began by acquiring crucial data on the depth of various soil horizons and their corresponding texture composition, represented by the percentages of soil, silt and clay. This essential information is readily available within the Land Type soil dataset. It is also available in the ACRU support data for modelling in South Africa, and contributes to the Harmonised World Soil Database (HWSD).

The texture data proves particularly valuable in characterising the soil water retention curves. To accomplish this, the input data in the form of soil texture percentages were fed into the Rosetta component within Hydrus 1D. Rosetta, operating within this software framework, facilitates the exploration of pedotransfer functions related to soils under three distinct hypothetical pressure scenarios: 0 mbar, 60 mbar and 15 000 mbar.

To categorise the soil texture, the classification system based on specific pore volumes was adhered to. This comprehensive approach enables a robust soil parameterisation file that captures the intricate dynamics of the soil-water relationship, laying the foundation for a more accurate and nuanced representation within the hydrological model. This contribution serves to enhance the understanding of the critical role played by soil characteristics in shaping hydrological processes.

Terrain and morphon mapping

Terrain and morphon mapping were done for the Goukou catchment by Stellenbosch University researchers to support the hydrological modelling and the fate of water in the system (Figure 6.30). These results were available for further modelling efforts. What is important to see is that the Goukou River system is carved into the landscape and that considerable landmasses occur above the river system. This makes coastal aquifer recharge along the coastal system from inland water sources quite difficult.

A morphon map of the same region was generated for this project (Figure 6.31). The map indicates slope classes (with soil info), which can be linked to overland flow, subsurface flow and zones where deep drainage occurs, as a basis for HRU development in hydrological modelling. The occurrence in soils also shows a marked difference between the soils towards the north (green) compared to the soils of the south (pink). The green soils are generally soils with lower infiltration capacities and therefore larger overland flow that generally causes floods to occur. The soils of the south (pink) on the other hand have higher infiltration capacity and more water is stored in this system.

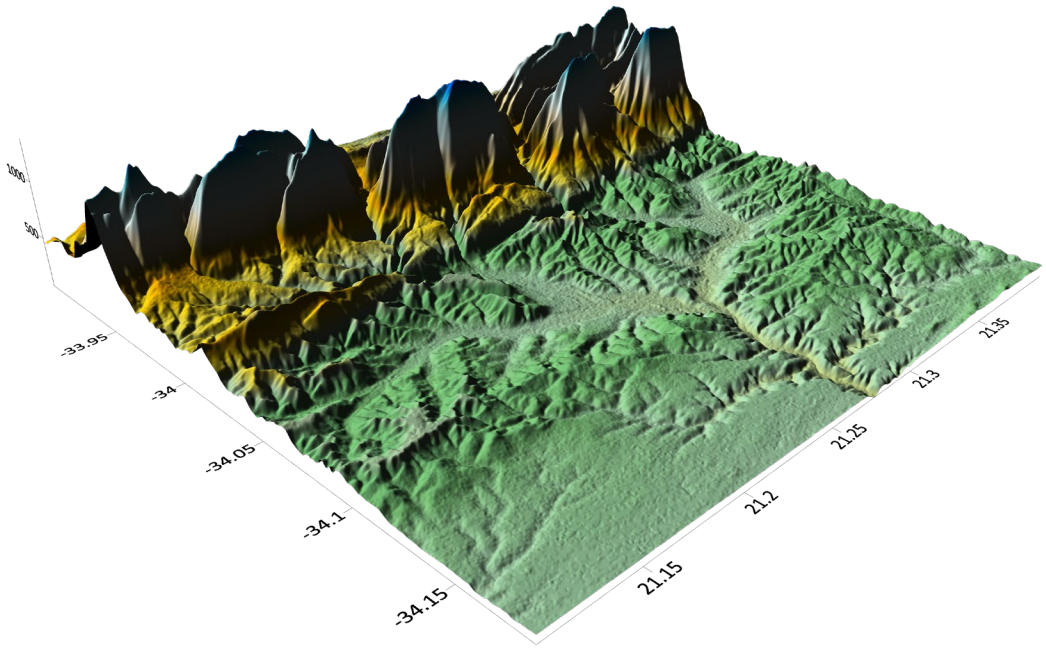

Figure 6.30: The exaggerated topography of the Riversdale region based on the 30 m digital elevation model.

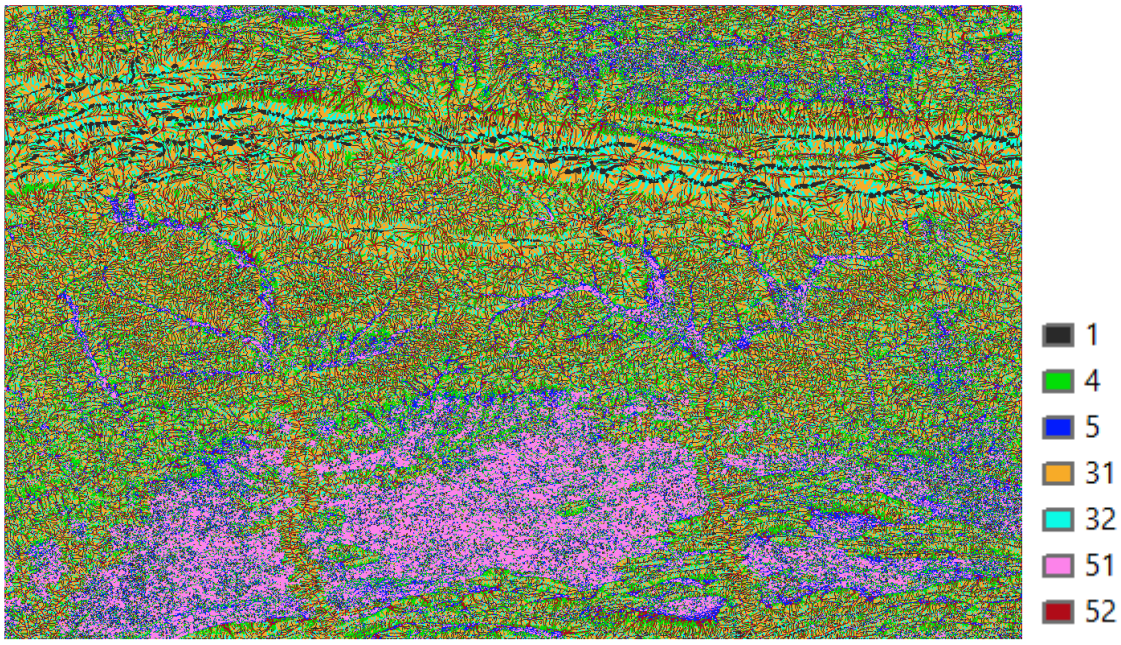

Figure 6.31: Soils and terrain map of the Hessequa region. Terrain classes are indicated, linked to the Land Types. This was used as the basis for soils mapping and vegetation distribution for the Land Type model (De Clercq et al. 2023).

Soil property designation

The HYDROSOIL raster was overlaid on the HRU map and soils information was transferred to the HRU map to be used in the JAMS modelling. The soils information was subjected to Hydrus 1D, where the information was developed into a format acceptable for JAMS modelling. The relevant hydrological soils information was derived from the HSWD database and used in the first simulation (Table 6.17). Similar relevant hydrological soils information used in the HYDROSOIL simulation (Table 6.18).

Table 6.17: Hydrological properties derived for the Land Type simulation.

MU_global	Depth (dm)	Air cap (mm)	Field cap SUM (mm)
60364	10	5.87	28.02
60644	100	63.35	268.88
60672	100	177.4	170
60688	100	66.73	264.63
60696	100	221.24	113.84
60702	100	104.46	230.56
60710	10	5.87	28.02
60726	100	104.46	230.56
60754	30	19.14	81.6
60756	100	66.73	264.63
60766	100	65.24	269.33
60780	100	221.24	113.84
60786	100	221.24	113.84

MU_global =

Table 6.18: Hydrological properties derived for the HYDROSOIL simulation

ID	Depth (dm)	Air Cap (mm)	Field cap SUM (mm)
A/B interflow	100	125.9	216.7
Shallow responsive	100	82.3	243.7
Deep recharge	100	211.1	121.9
Saturated responsive	100	275.9	57.3
Stagnating	100	181.2	148.7
Soil/bedrock interflow	100	161.2	166.8

The model employed 2 234 HRUs and used six soil IDs (Table 6.19). This table also outlines the specific definitions of each HRU concerning area, elevation, slope, aspect, position in the landscape, watershed, subbasin, land use and soil.

Table 6.19: A snipped HRU parameter file used in the JAMS model run

# hru.par created Fri, 15 Oct 2021, 17:32:42 by GRASS-HRU														
ID	area	elevation	slope	aspect	x	y	watershed	subbasin	landuseID	soilID	hgeoID	to_poly	to_reach	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	
999999	9999999	10000	90	360	9999999	9999999	999999	999999	9999	9999	9999	999999	999999	
n/a	m2	m	deg	deg	m	m	n/a	n/a	n/a	n/a	n/a	n/a	n/a	
1	575100	555	3.882	285	521111.7132	6244737.665	4	978	11	6	5	0	978	
2	664200	567	5.225	267	521066.7132	6245187.665	4	978	8	6	5	1	0	
3	980100	594	8.379	247	521516.7132	6245187.665	4	978	8	6	5	9	0	
4	891000	645	15.593	255	523361.7132	6244647.665	4	978	8	6	5	3	0	
5	1093500	849	23.52	220	513011.7132	6244287.665	4	974	5	3	5	7	0	
6	623700	895	26.915	304	511931.7132	6244017.665	4	974	8	3	5	70	0	
7	445500	855	24.156	301	512381.7132	6244197.665	4	974	5	3	5	6	0	
8	494100	694	9.205	274	527411.7132	6244017.665	5	996	5	2	5	23	0	
9	866700	556	2.2	207	522416.7132	6244287.665	4	978	8	6	5	0	978	
10	1069200	633	4.341	209	525476.7132	6243927.665	4	978	8	5	5	21	0	

Hydrological modelling

The JAMS/J2000 model was applied for the Goukou catchment (Figure 6.32). It involved a calibration process against the Duiwenhoks system model. Due to limited monitoring of the Goukou, we opted to calibrate its flows against the neighbouring Duiwenhoks system.

The decision to employ JAMS/J2000 for both catchment stems from ongoing research in the Western Cape on station density, hydrogeological property variability and transmission loss. In this application, a standard JAMS/J2000 model was employed, accounting for canal abstractions in the headwaters (DWS station H9H006) for Goukou and dam releases for Duiwenhoks (DWS station H8R001). A time series data substitution was used to mitigate the impact of anthropogenic activities (De Clercq et al., 2023; Watson et al., 2022).

Given the limited gauged portion of the Goukou, parameter sets from Duiwenhoks were adapted for the Goukou model to establish upper and lower limits in simulated streamflow. However, the breakdown of the flow components (surface runoff, interflow and baseflow) for Goukou was deemed unrealistic, necessitating further exploration with an additional parameter set. This iterative approach underscores the commitment to refining the model and ensuring its applicability to the unique characteristics of each catchment.

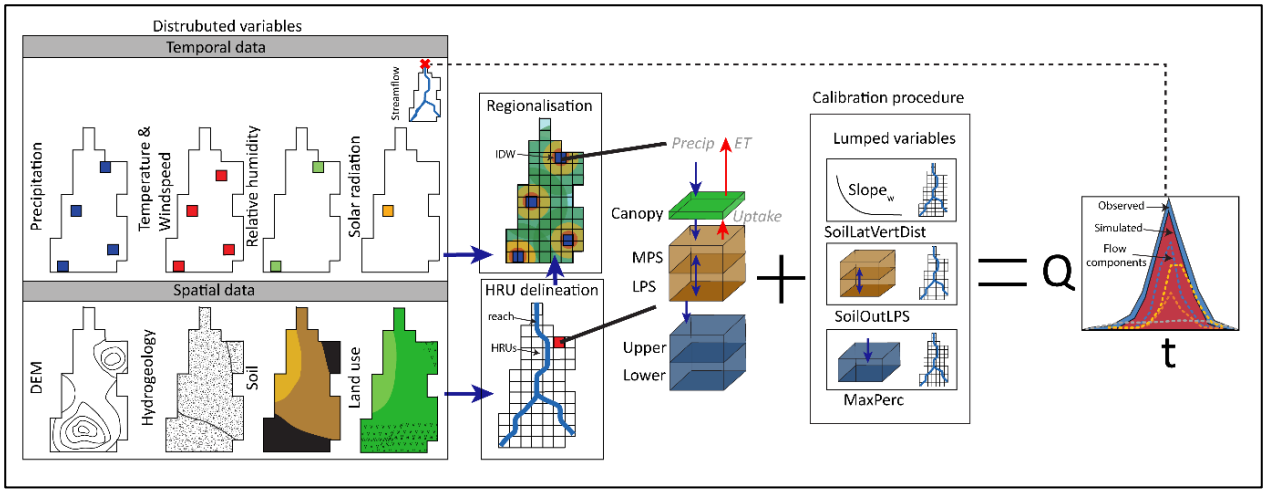

Figure 6.32: Layout of the modelling procedure, input data and parameters, calibration, and estimation of different flow components.

6.4.3 Results and discussion

HYDROSOIL

While there are striking similarities between the three HYDROSOIL maps for the different sampling techniques (Figure 6.33; Figure 6.34; Figure 6.35), their accuracies did differ somewhat (Table 6.20; Table 6.21; Table 6.22). The map created using the K-means clustering performed the best, with a validation point accuracy of

65.6% and a Kappa coefficient of 0.6, indicating a moderate agreement with reality. The other two maps were found not to be sufficiently accurate. Both these maps only achieved a slight representation of reality. Therefore, the K-means clustered map was used for the hydrological modelling.

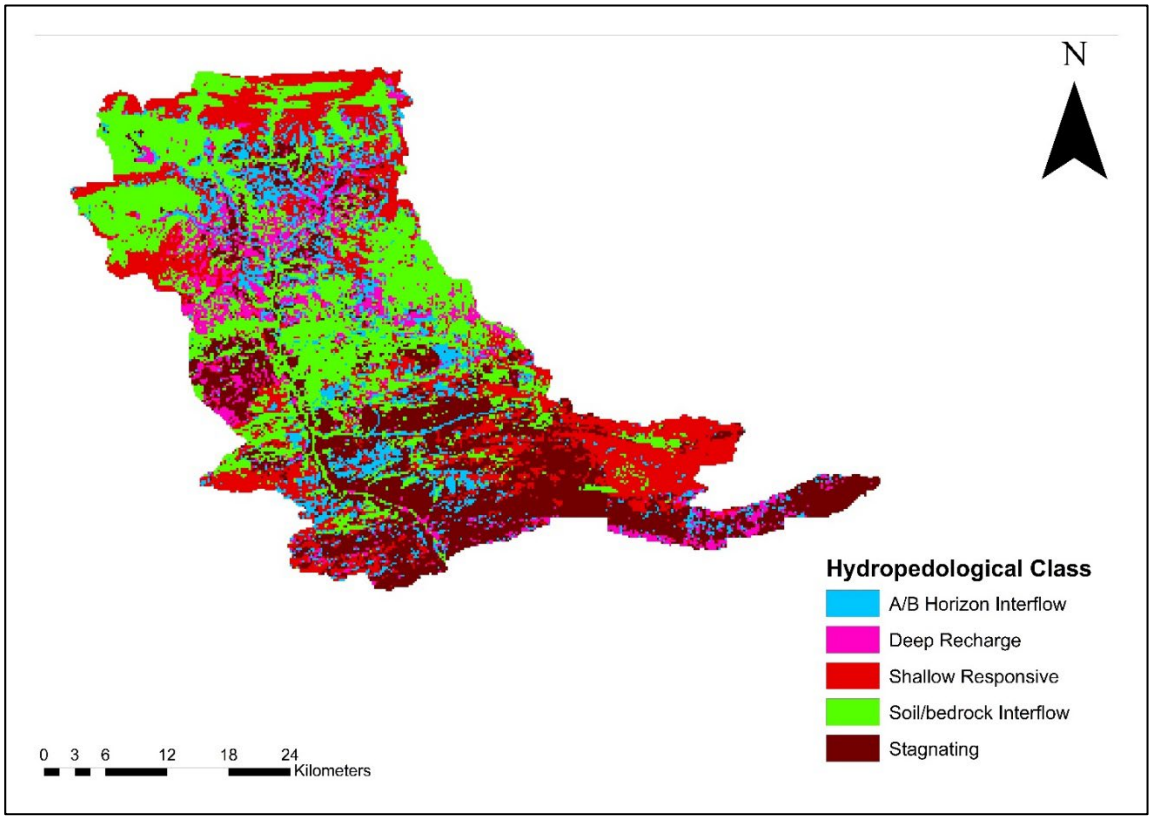

Figure 6.33: The HYDROSOIL map of the Goukou catchment created using the stratified random sampling method.

Table 6.20: The confusion matrix for the HYDROSOIL map created using stratified random sampling method.

		Map units									
		Deep recharge	Shallow recharge	Interflow A/B	Interflow soil/bedrock	Stagnating	Shallow responsive	Saturated responsive	Total	Correct	%
Observations	Deep recharge	10		2	1		1		14	10	71.4
	Shallow recharge								0	0	
	Interflow A/B	17		13	1	1	1	2	35	13	37.1
	Interflow soil/bedrock			2	2				4	2	50.0
	Stagnating				1	2			3	2	66.7
	Shallow responsive	3		4	4		3		14	3	21.4
	Saturated responsive								0	0	
	Total	30		21	9	3	5	2	70		
	Correct	10		13	2	2	3	0		30	
	%	33.3		61.9	22.2	66.7	60.0	0.0			42.9

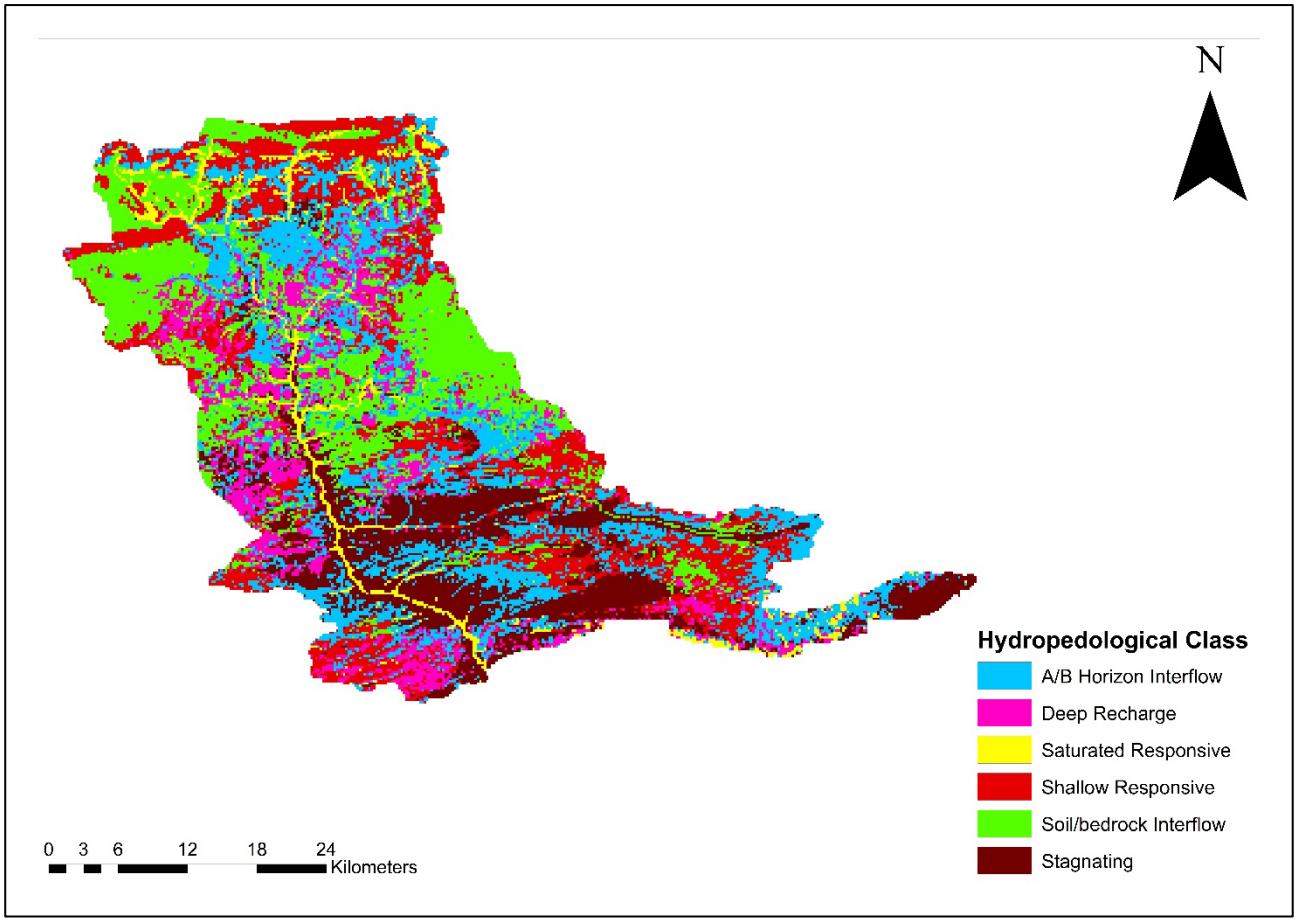

Figure 6.34: The HYDROSOIL of the Goukou catchment created using the K-means clustering method.

Table 6.21: The confusion matrix for the HYDROSOIL created using K-means clustering.

		Map units									
		Deep recharge	Shallow recharge	Interflow A/B	Interflow soil/bedrock	Stagnating	Shallow responsive	Saturated responsive	Total	Correct	%
Observations	Deep recharge	16				2	2	2	22	16	72.7
	Shallow recharge								0	0	
	Interflow A/B	9		8		2			18	8	44.4
	Interflow soil/bedrock	2		3	3		1		9	3	33.3
	Stagnating	1			2	5			8	5	62.5
	Shallow responsive	2				1	3			3	
	Saturated responsive							7	7	7	100.0
	Total	30	0	11	5	9	6	9	64		
	Correct	16	0	8	3	5	3	7		42	
	%	53.3		72.7	60.0	55.6	50.0	77.8			65.6

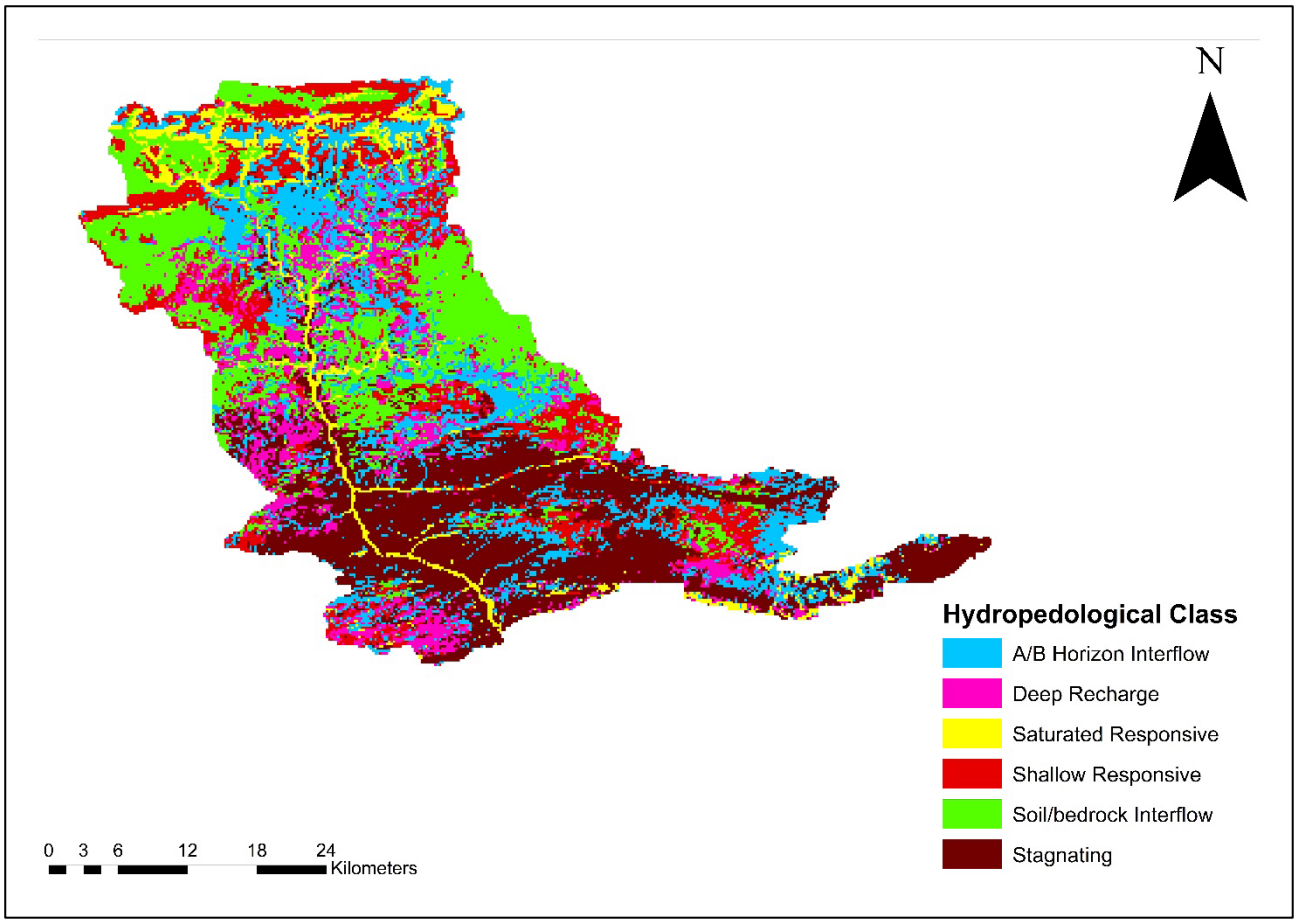

Figure 6.35: The HYDROSOIL of the Goukou catchment created using the Conditioned Latin Hypercube sampling method.

Table 6.22: The confusion matrix for the HYDROSOIL created using the Conditioned Latin Hypercube sampling method.

		Map units									
		Deep recharge	Shallow recharge	Interflow A/B	Interflow soil/bedrock	Stagnating	Shallow responsive	Saturated responsive	Total	Correct	%
Observations	Deep recharge	9		9	1		1	9	29	9	31.0
	Shallow recharge								0	0	
	Interflow A/B	10		10		1			24	10	47.6
	Interflow soil/bedrock	3		2	3		1		9	3	33.3
	Stagnating	1			2	2			5	2	40.0
	Shallow responsive	3				1	3			3	
	Saturated responsive							2	0	0	
	Total	26	0	21	6	4	5	11	64		
	Correct	9	0	10	3	2	3	0		27	
	%	34.6		47.6	50.0	50.0	60.0	0.0			42.2

JAMS hydrological modelling

Figure 6.36 presents the modelled hydrographs for both the Duiwenhoks and Goukou basins using the Land Type soil data. Despite the effective calibration in JAMS modelling, there remains a lingering question about its reliability. A recent study conducted at Stellenbosch University has suggested that, for most small catchments, the modelling timestep cannot exceed 16 days (Du Plessis, 2023). Beyond this threshold, a noteworthy overestimation in flow prediction becomes apparent. This observation underscores the importance of critically assessing the modelling methodology and its temporal resolution, particularly in the context of small catchments.

Comparing observed streamflow and simulated streamflow using both the Land Type data and the HYDROSOIL data, shows minimal distinction between the two simulated flows (Figure 6.37). The Land Type data exhibits an NSE of 0.31, like that of the HYDROSOIL data of 0.29. The Bias value for the Land Type data of -0.2 is also very similar to the HYDROSOIL value of -0.21. This indicates that the model in its current set-up is insensitive to soil information and the impact of the soils are masked by the role of rainfall distribution.

Unfortunately, the challenges associated with lumping variables together, especially rainfall distribution became evident in these simulations. The size of the catchment, and its placement in the headwaters of the catchment increases the model parameter uncertainty, which limited the accuracy of the model. The improved soil information could, however, benefit the modelling through using it for Penman modelling. Additionally a soil sensitivity analysis is required to determine the significance of the differences between the different soil maps. However, the sensitivity of the model to changing the rainfall distribution input should also be assessed as it seems that it masks any soil input changes.

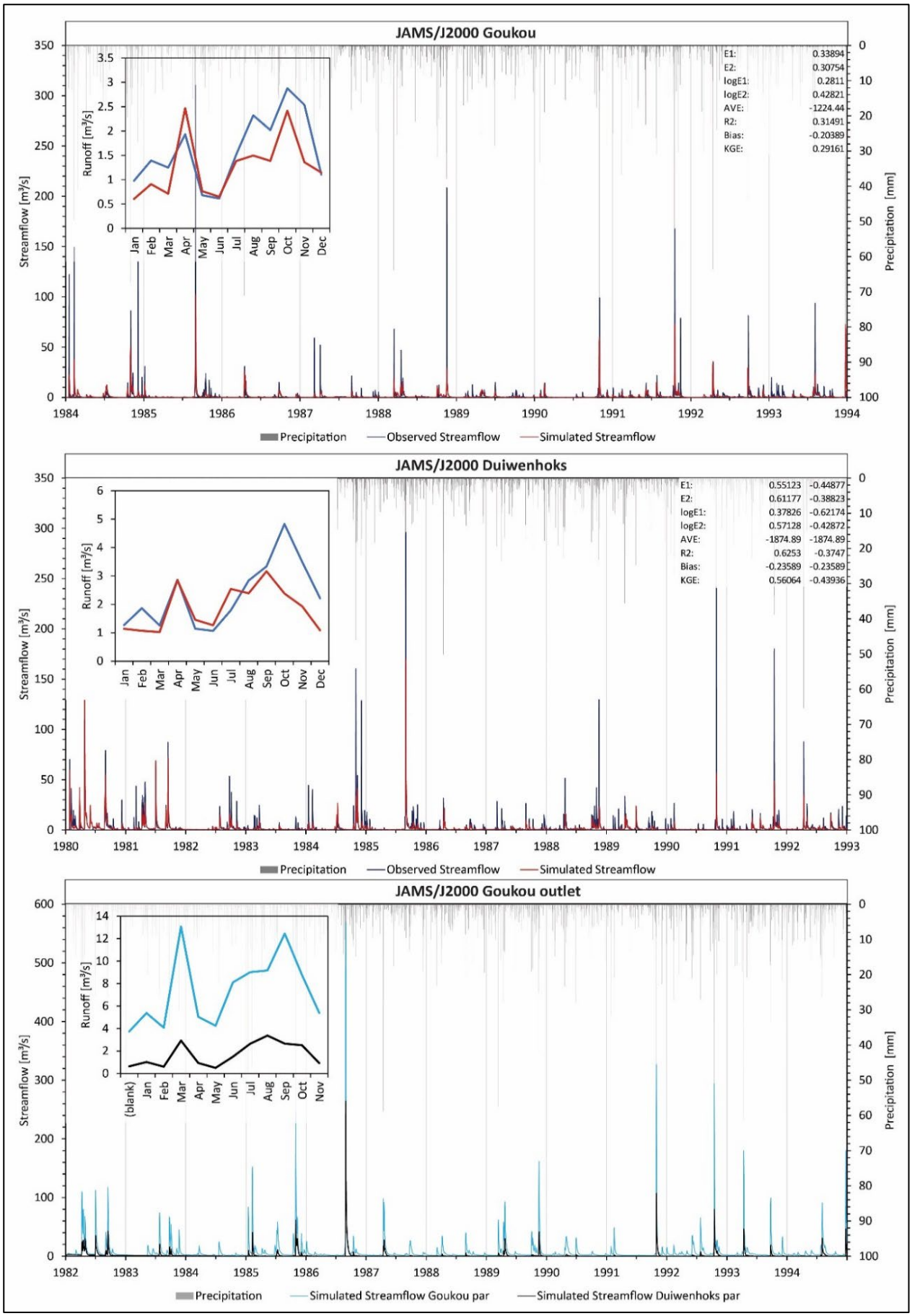

Figure 6.36: An overview of the modelled results produced by JAMS for the Goukou system, indicating peak flows modelled on a daily timestep and compared to the Duiwenhoks basin model (De Clercq et al., 2023).

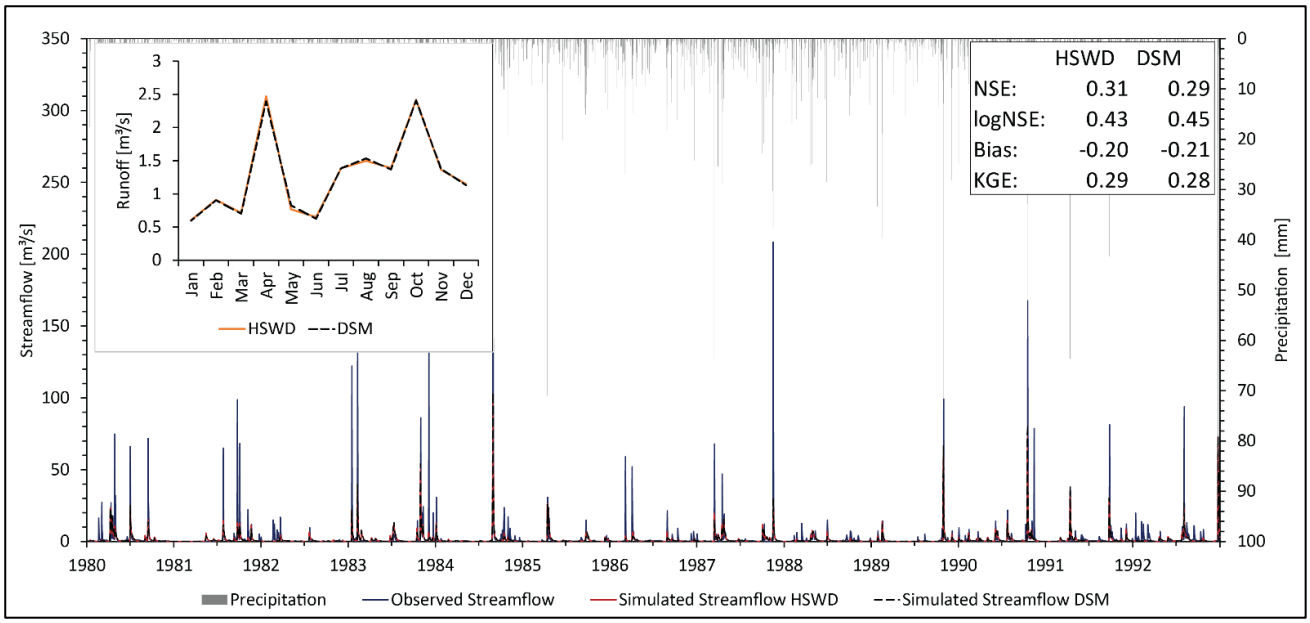

Figure 6.37: An overview of the modelled results produced by JAMS for the Goukou system, indicating peak flows modelled on a monthly timestep, for both the Land Type (HSWD) and HYDROSOIL (DSM) maps.

6.4.4 Conclusions

Using digital soil mapping, an acceptable HYDROSOIL for the Goukou catchment could be developed, using K-means clustering to split the available soil data into a training and validation dataset. However, using this soil map to model the hydrological response using the JAMs model within the catchment made an insignificant difference to the modelling accuracy. It is therefore concluded that the rainfall distribution with the current model set-up within the Goukou catchment masks any modelling effects which the soils could have. A soil sensitivity analysis is required to determine the significance thereof. Conversely, a sensitivity analysis of the model outcomes to changing the rainfall input should be done as higher priority as it has a larger effect on the model output.

Understanding the hydrological regime within the mountainous parts of the country remains an important challenge, as this is where most of the rain falls. Therefore more studies should be conducted in such terrains, despite the difficult circumstances and lack of data often encountered in these areas. This is true for the Goukou catchment, but also for most of catchments with mountainous parts, including the Drakensberg.

CHAPTER 7: HYDRAULIC PEDOTRANSFER FUNCTIONS

In addition to mapping soil more accurately, to understand and model the hydrological response of an area, the soils need to be more accurately parameterised as well (Van Tol & Van Zijl, 2022). However, hydrological soil measurements are cumbersome and expensive. Therefore, pedotransfer functions (PTF's) should be developed, whereby soil properties difficult and expensive to measure can be predicted using soil properties easy and inexpensive to measure.

This chapter will describe two separate studies where the development of PTFs was attempted. The creation of PTF's using legacy soil data was presented at the Kirkham Conference in 2022 held at Skukuza, South Africa by Anru Kock (Section 7.1). The development of PTF's using collected data formed part of the 4th year project of Altus Jacobs, Elouise Verwey and Vian Cooke (Section 7.2). It was decided not to combine the legacy data with the newly acquired data, as the legacy data is mostly measured in the field, while the newly acquired data was measured from undisturbed samples in a laboratory. It seems these measurements are not comparable.

7.1 CREATING PEDOTRANSFER FUNCTIONS TO DETERMINE IMPORTANT SOIL HYDRAULIC PROPERTIES

7.1.1 Introduction

Soil and water play an essential role on the surface and within the subsurface of the Earth. Surface soil regulates and controls the water balance via infiltration, evapotranspiration, surface runoff, groundwater recharge and therefore has a substantial effect on regional and global land surface water (Zhang & Schaap, 2019). Soil and water parameters are also important for hydrological modelling of catchments (Abbaspour et al., 2019). Models require soil and water data for effective use and implementation. Soil datasets that contain the required soil property data for certain areas of interest are not always available to be included into the models. Measured soil properties vary from dataset to dataset and the amount of data as well as the type of data that is captured may differ due to the needs, aim and cost constraints of a respective study. Certain soil properties are almost always measured from soil samples for, example soil texture, pH, cation exchange capacity (CEC). Other soil hydraulic properties are more time consuming or more expensive to measure, for example, saturated hydraulic conductivity (Ks), soil water content at field capacity (θ_33) and soil water content at wilting point (θ_1500) are not readily included.

PTFs aim to solve this problem by estimating soil properties, especially soil hydraulic properties, using readily available soil data. For decades PTFs have been created and used for many important soil hydraulic properties including, Ks, θ_33, θ_1500 and Available Water Capacity (AWC), which is the difference between θ_33 and θ_1500. The first equations to relate land characteristics and soil properties were in 1987 (Bouma & Van Lanen, 1987; Li et al., 2007). Modern PTFs have been widely created around the world and mainly use soil texture, bulk density (BD) and organic carbon (OC) as predictors for soil hydraulic properties (Li et al., 2007).

A relatively small number of PTFs for some soil properties have been created and used for South African soils. These include: water conducting microporosity by (Van Tol et al., 2012) and liquid limit, plastic limit, linear shrinkage and plasticity index used in engineering and land evaluation (Van Tol et al., 2016a). This study will focus on creating PTFs for Ks, θ_33, θ_1500 and BD for South African soils.

7.1.2 Materials and methods

Legacy soil data used for pedotransfer functions

A soil dataset with 221 soil samples measuring various soil properties (Table 7.1) was used to create the PTFs (Van Tol, 2022). It is important to note that not all soil samples contain the same amount of soil data for each soil property, some clusters of soil samples have no values for certain soil properties. This was addressed during the development of the respective PTFs.

Table 7.1: Summary of predictor variables from the soil dataset (Van Tol, 2022).

Soil property	*n*	Unit	Max.	Min.	Average	Standard deviation
Saturated hydraulic conductivity (Ks)	220	$mm.h^{-}$	12000	0	473.45	1361.35
K (3)	147	$mm.h^{-}$	56	0.02	7.9	11.96
Macropore conductivity (MPC)	144	$mm.h^{-}$	1768.7	0.0153	155.4	283.9
Bulk density (BD)	169	$Mg.m^{-3}$	1.9852	0.4841	1.35	0.27
Organic carbon (OC)	123	%	9.36	0.04	1.48	1.79
Cation Exchange Capacity (CEC)	147	cmol	37.5	1.65	9.6	7.08
Sand	221	%	93.302	5	54.06	20.60
Silt	221	%	54.4	0.6	17.03	10.27
Clay	221	%	68.7	3.7	28.12	14.42
Drained upper limit (DUL)	74	$mm.mm^{-}$	0.5199	0.0454	0.26	0.1
Lower limit (LL)	63	$mm.mm^{-}$	0.47	0.0089	0.15	0.10

***n* = number of observations, Max = Maximum unit value, Min = Minimum unit value**

Development of pedotransfer functions

To create the PTFs for the respective soil hydraulic properties, multiple regressions were used to train the models with soil properties as predictors (Table 7.2). All PTFs were created using the R programming language (R Core Team, 2022). The soil dataset (Van Tol, 2022) was first reduced to 149 soil observations after removing samples that contained Ks values that were error sum. The dataset was then randomly divided into two sets of data, one training set containing 75% of the soil observations and a second set that contain the other 25% of soil observations. This step was repeated multiple times for each soil hydraulic property to maintain the maximum number of samples with complete data. The training set of data was used to train the multiple linear regression models and the test set was used for an independent validation. For Ks five multiple regression models were created, three for BD, two for Drained Upper Limit (DUL) and three for Lower Limit (LL) using different combinations of available soil property data. One cubist model was created for Ks with all soil properties to evaluate the performance of machine learning algorithms compared to standard multiple linear regression models.

Table 7.2: All pedotransfer functions (PTFs) developed with the soil properties as predictors.

Soil property PTFs	Method	Predictors	n_t
Ks-A	MLR	CEC, OC	112
Ks-B	MLR	Sand, Silt, Clay	112
Ks-C	MLR	Sand, Silt, Clay, CEC	112
Ks-D	MLR	Sand, Silt, Clay, OC	112
Ks-E	MLR	Sand, BD, OC	112
Ks-N	Cubist	BD, OC, Sand, Silt, Clay	112
BD-F	MLR	Sand, Silt, Clay	63
BD-G	MLR	Sand, Silt, Clay, CEC	63
BD-H	MLR	Sand, Silt, Clay, OC	63
BD-P	Cubist	OC, CEC, Sand, Silt, Clay	63
θ_{33}-I	MLR	Ks, BD	51
θ_{33}-J	MLR	Ks, Sand, Silt, Clay, BD	51

Soil property PTFs	Method	Predictors	n_t
θ_{33}-Q	Cubist	Ks, BD, Sand, Silt, Clay	51
θ_{1500}-K	MLR	Ks, Sand, Silt, Clay, BD	36
θ_{1500}-L	MLR	BD,	36
θ_{1500}-M	MLR	Sand, Silt, Clay, BD,	36
θ_{1500}-R	Cubist	Ks, BD, Sand, Silt, Clay,	36

n_t = number of samples used for training; MLR = multiple linear regression; CEC = Cation Exchange Capacity; OC = Organic Carbon; BD = bulk density.

Validation of pedotransfer functions

Validation is necessary to determine the performance of the models. A validation set, or in this case the 25% test set, is used to accomplish this. Using each model trained (Table 7.2), the respective predictor values are used to make a prediction on the response variable of the specific PTF. The predicted values from the respective PTFs are then compared against the actual values in the soil dataset using statistical performance parameters. These parameters include Mean Error (ME), Root Mean Square Error (RMSE), coefficient of determination (R^2), ratio of performance to deviation (RPD) and coefficient of variation (CV) (Wadoux et al., 2021). All statistical calculations except for CV were calculated using the *eval* function in the *soilspec* package in R studio, CV was manually calculated using the R programming language.

Comparing pedotransfer functions for saturated hydraulic conductivity

From all five PTFs created for Ks, one was selected that produced the best results and that also used the same predictors as found in the literature (Weynants et al., 2009). The following equation was used:

$$Ks\ (cm.d^{-1}) = exp\ (1.9582 + 0.0308 * Sand(\%) - 0.6142 * BD(g.cm^{-3}) - 0.01566 * OC(g.Kg^{-1}) \qquad (7.1)$$

Units for each soil property differ from that of the units used in this study (Table 7.1). This was corrected by first converting the soil property values in this study to match that of the above equation and then converting the Ks results back to $mm.h^{-}$ to be able to compare the values from both PTFs.

7.1.3 Results

The statistical performance parameter values for each PTF is given in Table 7.3 and the best performing PTF is compared to that of Weynants et al. (2009) in Table 7.4.

Table 7.3: Performance indicators for validation of the pedotransfer functions (PTFs).

PTF	ME	RMSE	R^2	RPD	CV	n_v
Ks-A (CEC, OC)	-2,54	161,16	0,59	1,58	105.77	37
Ks-B (SSC)	1,12	221,22	0,15	1,1	143.57	37
Ks-C (SSC, CEC)	3,41	191,27	0,36	1,27	122.32	37
Ks-D (SSC, OC)	2,37	171,64	0,53	1,49	109.12	37
Ks-E (Sand, BD, OC)	58,7	217,61	0,49	1,44	66.02	37
Ks-N (Cubist)	-1,78	152,1	0,6	1,6	100.6	37
BD-F(SSC)	0,11	0,21	0,4	1,32	14.19	21
BD-G (SSC, CEC)	0,11	0,21	0,43	1,36	14.2	21
BD-H (SSC, OC)	0,04	0,11	0,87	2,88	7.81	21

PTF	ME	RMSE	R^2	RPD	CV	n_v
BD-P (Cubist)	-0.01	0.08	0.92	3.7	5.88	21
θ_{33}-I (ALL)	0,02	0,07	0.04	1.05	28.92	17
θ_{33}-J (Ks, BD)	0,02	0,08	-0,32	0,9	32.43	17
θ_{33}-Q (Cubist)	0.01	0.06	0.25	1.19	25.67	17
θ_{1500}-K (ALL)	0	0.06	-0.39	0.89	28.92	12
θ_{1500}-L (BD, θ_{33})	0	0.06	-0.29	0.92	44.88	12
θ_{1500}-M (SCC, BD, θ_{33})	0	0,07	-0.6	0.82	52.95	12
θ_{1500}-R (Cubist)	0	0.06	-1.51	-0.66	45.44	12

ME = Mean Error; RMSE = Root Mean Square Error; R^2 = coefficient of determination; RPD = ratio of performance to deviation; CV = coefficient of variation; n_v = number of samples used for validation.

Table 7.4: Performance indicators for comparing pedotransfer functions (PTFs) to Weynants et al. 2009.

PTF	ME	RMSE	R^2	RPD	CV	n_v
Ks-S (Sand, BD, OC)	58.68	217.62	0.49	1.44	66.02	17
Ks-Wey (SSC)	-170.67	384.74	-0.61	0.81	383.57	17

ME = Mean Error; RMSE = Root Mean Square Error; R^2 = coefficient of determination; RPD = ratio of performance to deviation; CV = coefficient of variation; n_v = number of samples used for validation.

Scatter plots were used to indicate the individual performance of each created PTF for the various hydraulic soil properties by plotting the observed values to the predicted values obtained from the PTFs (Figure 7.1; Figure 7.2; Figure 7.3; Figure 7.4). With the most accurate PTF, the data points should be centred around the 1:1 line, indicated as a solid line. The best Ks PTF created in this study was compared with the PTF for Ks as created by Weynants et al. (2009) (Figure 7.5).

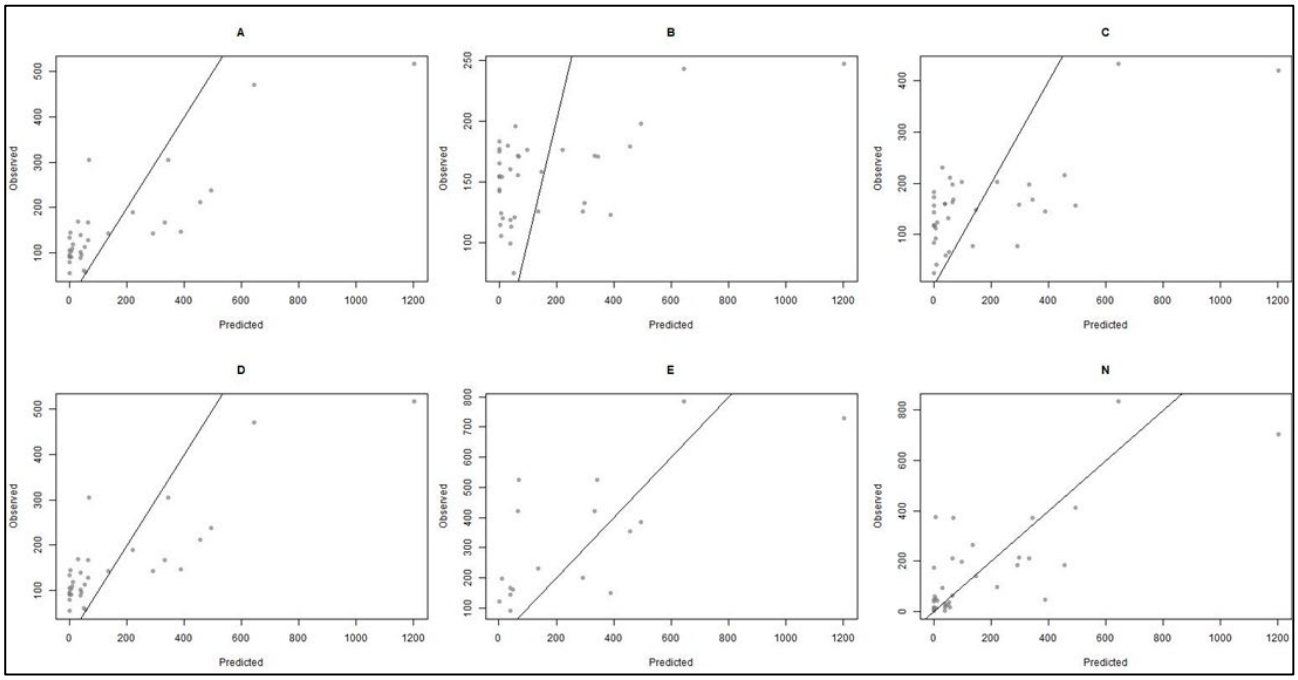

Figure 7.1: Scatter plots for validating PTFs for Ks, with a 1:1 line. Plots A-E and N correspond to each PTF for Ks in Table 7.2.

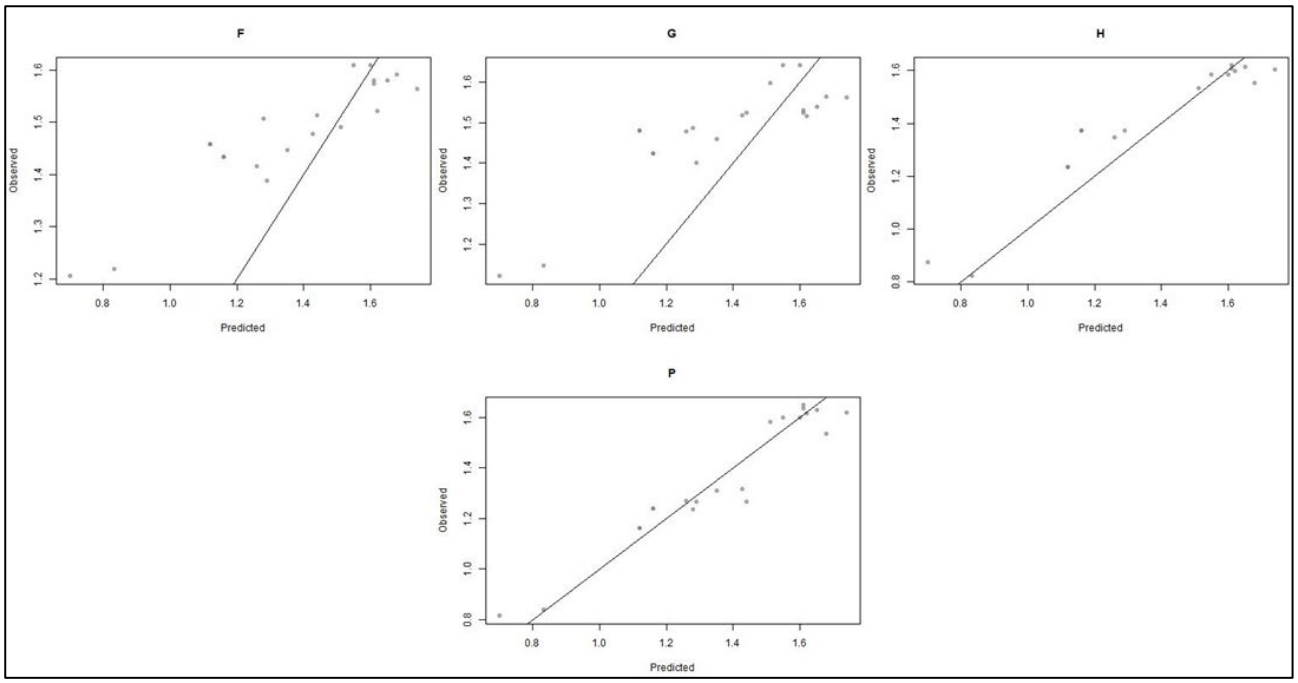

Figure 7.2: Scatter plots for validating PTFs for bulk density, with a 1:1 line. Plots F-H and P corresponds to each PTF for BD in Table 7.2.

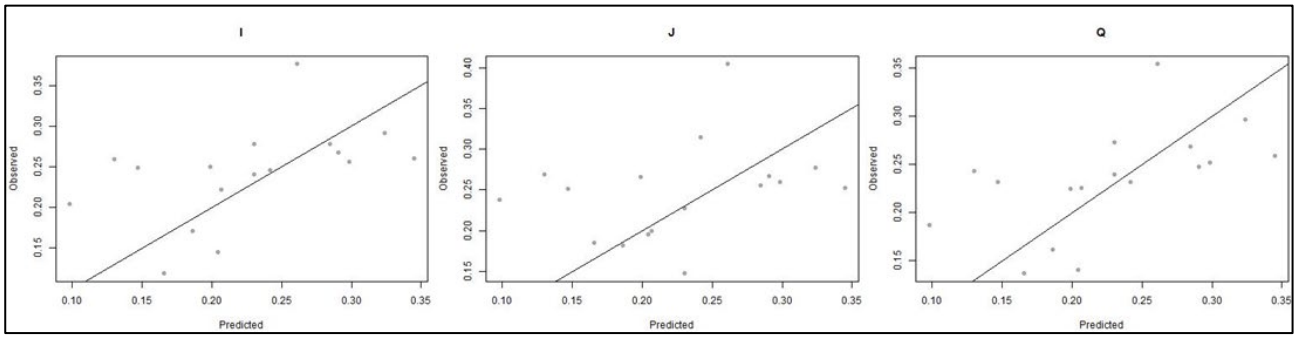

Figure 7.3: Scatter plots for validating PTFs for θ_33, with a 1:1 line. Plots I and J corresponds to each PTF for θ_33 in Table 7.2.

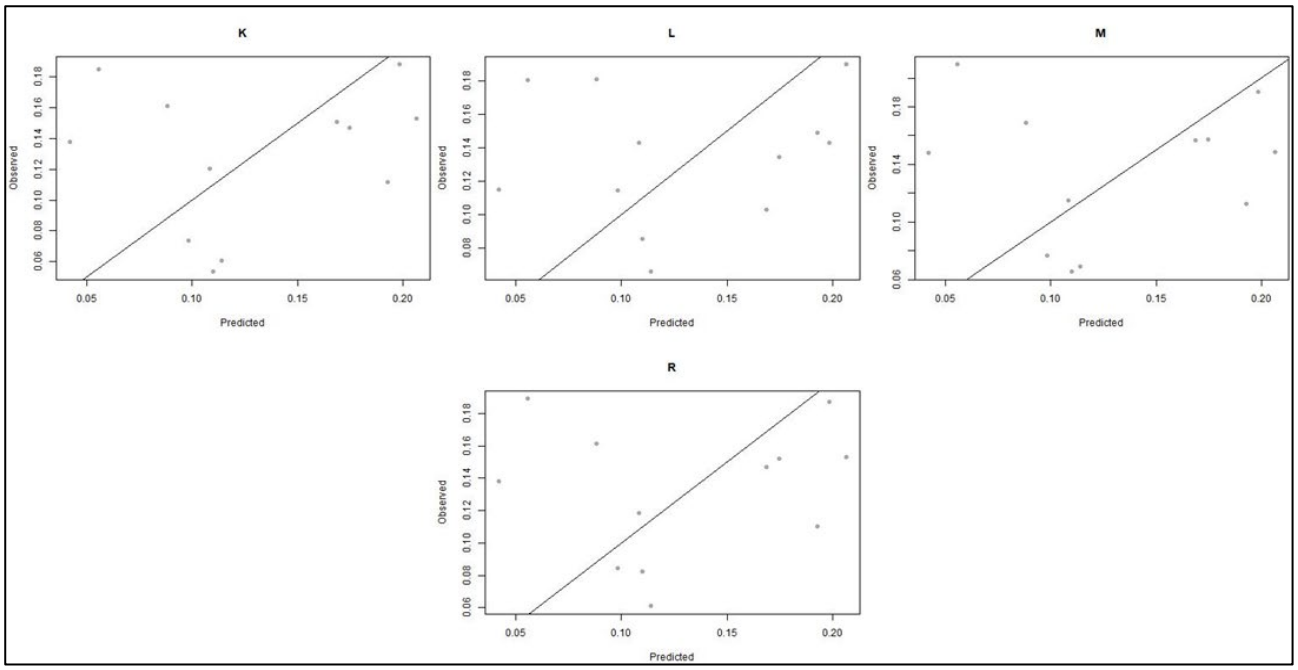

Figure 7.4: Scatter plots for validating PTFs for θ_1500, with a 1:1 line. Plots I and J corresponds to each PTF for θ (1500) in Table 7.2.

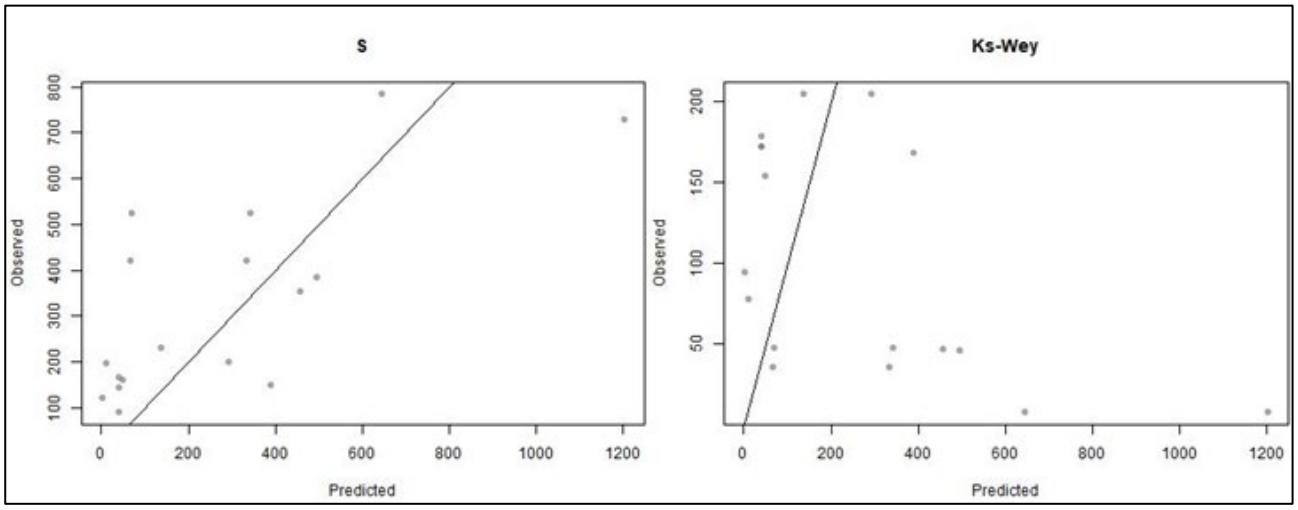

Figure 7.5: Scatter plots for comparing Ks-S and Ks-Wey PTFs, with a 1:1 line

7.1.4 Discussion

Saturated hydraulic conductivity

Saturated hydraulic conductivity, which is one of the important soil properties, had variable results for PTF's. Pedotransfer functions for Ks created using cubist and all available soil properties as predictors had the best overall performance (ME = -1,78, RMSE = 152,1, R^2 = 0,6, RPD = 1,6 and CV = 100.6). The cubist PTF had the highest R^2 and RPD with the lowest RMSE values compared to the others. The Ks-A PTF with multiple linear regression follows in performance (ME = -2.54, RMSE = 161.16, R^2 = 0,59, RPD = 1,58 and CV = 105.77) with relatively low RMSE, high R^2 and RPD. The Ks-E PTF created with Sand, BD and OC had slightly weaker statistical performance except for having the lowest CV of 66.02, which makes it a relatively good PTF compared to the others for Ks with higher CV values. Another Ks PTF (Ks-D) that used OC as a predictor also showed comparable performance with slightly weaker performance statistics (ME = 2.37, RMSE = 171.64, R^2 = 0,53, RPD = 1,49 and CV = 109.12). PDFs using only sand, silt and clay or CEC had the worst performance (Table 7.3). These results are substantiated with the scatter plots (Figure 7.1). The cubist Ks-N PTF had most observations on the 1:1 line which indicated good performance for predicting lower Ks values. Other PTFs show the tendency to underpredict the Ks values with observations plotting above the 1:1 line. Predictions for Ks values greater than 200 $mm\,.\,h^{-1}$ tend to be spread out from the 1:1 indicating weak model performance greater Ks values.

Bulk Density

Three PTFs for BD were created using standard multiple linear regression and one with cubist. Two of the BD PTFs had the best performance with BD-P created using cubist having the best performance (ME = -0.01, RMSE = 0.08, R^2 = 0.92, RPD = 3.7 and CV = 5.88). This makes the PTF BD-P an excellent PTF for predicting BD. With multiple linear regression, PTF BD-H had the second-best performance with ME = 0.04, RMSE = 0.11, R^2 = 0.87, RPD = 2.88 and a CV of 7.81 which also makes it a good BD PTF. Scatter plots for BD-P and BD-H show that observations are close to and on the 1:1 line (Figure 7.2). The performance from these two PTFs are more impressive seeing that only 63 soil samples were used to develop the PTFs. Scatter plots for BD-F and BD-G have observations further from the 1:1 line which is in line with the statistical lower R2 values for these PTFs.

Drained upper limit Θ_{33}

PTF development for θ_{33} had less soil samples to work with – only 51 samples available for model development and 17 for model validation. This resulted in less optimal conditions for calibration of the θ_{33} PTFs. Low statistical performances were observed (Table 7.3) for all the θ_{33} PTFs, with the θ_{33}-Q PTF having the best results. Scatter plots (Figure 7.3) support these findings. Some predictions for all three θ_{33} PTFs are close to the 1:1 with low ME and RMSE values. None of these PTFs are useful and more calibration and samples are needed to develop better performing PTFs for θ_{33}.

Lower limit Θ_{1500}

Just like the θ_{33} PTFs, results for the θ_{1500} PTFs show no potential with regard to statistical performance (Table 7.3). Only 36 soil samples had available data to be used for model development and 12 samples were used to validate the models. The best performing model was θ_{1500}-L which used BD and θ_{33} as predictors (ME = 0, RMSE = 0.06, R^2 = -0.29, RPD = 0.92 and CV = 44.88). Scatter plots for θ_{1500} (Figure 7.4) show almost no correlation between the predicted and observed values for all three PTFs. For θ_{1500}, multiple linear regression created the relatively best performing PTFs, whereas the cubist PTF θ_{1500}-R had lower than expected performance when compared to the performance of cubist PTFs for the other soil properties.

Comparing pedotransfer for saturated hydraulic conductivity

The PTF for Ks-S from this study, which used sand, BD and OC (Equation 7.1) was compared against the performance of another PTF with the same predictors from Weynants et al. (2009), with the equation:

$$Ks - S = 965.44 + 2.367 * Sand - 612.45 * Bulk\ density + 36 * organic\ carbon \quad (7.2)$$

Both PTFs were used to make Ks predictions on 17 soil observations which contained enough soil data to make the comparison. Statistical performance parameters (Table 7.4) show that Ks-S had better predictions than the PTF Ks-Wey used from Weynants et al. (2009). Lower ME of 58.68, RMSE of 217.62 were observed compared to a ME of -170.67 and RMSE of 384.74 for Ks-Wey. Scatterplots (Figure 7.5) confirm that Ks-S is better developed to make predictions closer to the 1:1 especially for Ks > 500 $mm.h^{-}$.

7.1.5 Conclusions and recommendations

This study was able to successfully produce PTFs for all four soil hydraulic properties Ks, BD, θ_{33} and θ_{1500} from legacy soil data. PTFs for Ks using multiple linear regression are suboptimal for practical use. The Ks PTF developed with cubist showed more potential as being useful as a PTF in predicting Ks from BD, OC, CEC, sand, silt and clay. This, however, is only true if enough soil data is available to develop the PTF. Bulk density PTFs were the best performing PTF from all the soil properties chosen in this study and produced PTF's that are accurate and reliable enough to be practically used. PTFs for θ_{33} and θ_{1500} were not as successful as for Ks and BD are not usable for any predictions.

To improve the accuracy of PTFs, a larger soil dataset with more complete soil observational and soil chemical data must be used to develop the PTFs, especially for Ks, θ_{33} and θ_{1500}. More OC data will lead to better PTF development for Ks as seen from the performance of PTFs which used OC as predictors.

7.2 CREATING AN HYDRAULIC PEDOTRANSFER FUNCTION FOR SOUTH AFRICAN SOILS

7.2.1 Introduction

Problem statement

In modern agriculture, precision farming is gradually becoming the 'normal' way of farming. The challenge comes with the bottomless need for data to make the best decisions. To acquire data for certain soil characteristics, such as hydraulic conductivity, can be very time consuming and expensive. To bridge this problem, hydraulic PTFs are an alternative method to obtain such data, namely.

Mathematical models called hydraulic PTFs link the properties of soil to hydraulic conductivity and soil water retention (Rawls et al., 1982). They are widely used in agriculture for estimating soil water availability and movement, which are critical for crop growth and yield (Wösten et al., 1999). PTFs have been developed using various techniques, including regression and machine learning methods, and have been applied in different agricultural systems and regions worldwide (Nemes & Schaap, 2006; Schaap et al., 2001). However, the accuracy and transferability of PTFs depend on the quality and quantity of input data, the model structure and complexity, and the validation and calibration procedures (Schaap et al., 2001; McBratney et al., 2002).

Despite the significant progress in the development of PTFs globally, there is a lack of hydraulic PTFs specifically tailored for South African soils. The unique characteristics of South African soils, influenced by diverse climates, vegetation types and parent materials, necessitate the development of locally calibrated PTFs. These functions are crucial for predicting soil hydraulic properties, which are fundamental inputs for hydrological and land surface models. Recent studies have demonstrated the value of PTFs in predicting soil properties such as Atterberg limits, soil moisture content at field capacity and permanent wilting point (Van Tol et al., 2016a). However, these studies also highlight the challenges associated with developing accurate and

reliable PTFs, including the need for large and diverse soil datasets (Miti et al., 2023). Furthermore, a recent study on European soils has shown that incorporating prediction uncertainty into PTFs can significantly improve their accuracy (Szabó et al., 2021). However, such advancements have not yet been applied to South African soils, and the question is how an accurate and reliable PTF can be developed using locally available soil data.

Hydraulic conductivity and data collection methods

Hydraulic conductivity (K) is an important property that governs the water transmission capacity of soils, influencing critical processes such as water infiltration, drainage and plant water availability (Hillel, 2003; Rawls et al., 2003). It represents the ability of a porous material to transfer water and is governed by Darcy's law, which states that the hydraulic gradient and hydraulic conductivity are directly proportional to the fluid flow through a medium (Batezini & Balbo, 2015). The amount of organic matter, soil structure and texture all have a significant impact on the K value, with two critical parameters used to describe the conductivity: Ksat (saturated hydraulic conductivity) and Kunsat (unsaturated hydraulic conductivity), corresponding to fully and partially saturated conditions, respectively (Hillel, 2003; Rawls et al., 1982; Schaap et al., 2001).

Both laboratory and field-based techniques are utilised for measuring K, with Ksat ranging from 1×10^{-6} m/s (for clayey soils) to 1×10^{-3} m/s (for sandy soils) (Hillel, 2003; Rawls et al. 1982). An extensive database of European soils' hydraulic properties (Wösten et al., 1999), including hydraulic conductivity, provides valuable information that was collected from different sources and subjected to quality checks to ensure consistency and reliability. The data revealed significant variations in hydraulic conductivity across different European soils, with sandy soils exhibiting higher values compared to clayey soils. A comprehensive overview offers laboratory-based procedures for determining the hydraulic conductivity and diffusivity of soil (Klute & Dirksen, 1986), which are essential parameters for understanding water flow in soils. Field-based techniques include the double ring, the single ring, the tension, and the disk infiltrometer. These methods are used to gauge the soils in-situ hydraulic conductivity.

Two hydraulic conductivity measurement techniques are the constant head method and the falling head method. The constant head method measures the hydraulic conductivity of highly permeable soils, while falling head method measures hydraulic conductivity for low to moderately permeable soils and both are widely used in soil science research (Hillel, 2003). These two methods are used in the field along with the double and single ring. It describes how the water going into the soil is facilitated, by measuring the time for a drop in distance (falling head), or having the head be constant, and the volume of water changed measured at the water source (constant head).

Lefranc's test, which involves measuring the amount of time it takes for a water column inside a tube to drop to a specific height, is the foundation for the falling head method, a dependable laboratory technique for determining the hydraulic conductivity of soils (Mualem,1976; Wösten et al., 1999; Pedescoll et al., 2011). The falling head method can distinguish between the hydraulic conductivities typically seen for non-clogged and clogged systems, which range between 200 and 300 m/day and less than 50 m/day, respectively, it has been used to assess on-site hydraulic conductivity in a variety of environmental facilities. The falling head method will be used in this study, as most of the soils tested are moderately permeable.

Hydraulic pedotransfer functions

PTFs are a useful and practical tool for estimating hydraulic conductivity (Nemes & Schaap, 2006). Using empirical or semi-empirical methods, these functions are created based on more easily quantifiable soil properties such as bulk density, texture and organic matter content (Schaap et al. 2001). PTFs are a valuable alternative to direct measurements of hydraulic properties because they can be time-consuming and expensive (Rawls et al., 2003). However, the accuracy of PTFs can be limited by specific soil types and conditions (Schaap et al., 1998).

For predicting soil water retention parameters and hydraulic conductivity, several PTFs have been created using various databases and variables (Schaap et al., 2001). A PTF for estimating the soil water retention parameters was developed using a database of 633 soils based on organic matter content and soil texture (Rawls & Brakensiek, 1985). This PTF was found to be accurate for a variety of textures and organic matter levels. Another PTF for estimating hydraulic conductivity and soil water retention using organic matter content, soil texture and bulk density was developed from a database of 542 soils (Saxton & Rawls, 2006). This PTF was also found to be accurate for many different types of soil. A PTF was developed specifically for European soils, focusing on water retention and unsaturated hydraulic conductivity, using a large dataset of 12 000 soils (Wösten et al., 1999). This PTF includes organic matter content, soil texture and bulk density and was found to be accurate for predicting the hydraulic properties of a wide range of European soils.

It is critical to note that PTFs have limitations, and appropriate PTFs should be developed for specific soil types and regions (Schaap et al., 2001). PTFs should be used with caution due to their limitations (Vereecken et al., 2010). A review of the development, validation and application of PTFs in soil hydrology noted that PTFs are an essential tool for estimating soil hydraulic properties in regions with scarce data (Tóth et al., 2015). They also emphasised the importance of validating PTFs with independent datasets to improve their accuracy and reliability (Hutson, 1983).

There are several PTFs that have been developed for South African soils among which is the model developed by Van Tol et al. (2016), which calculates hydraulic conductivity and soil water retention curves using easily measurable soil characteristics like texture, bulk density and organic matter content. Another PTF to predict the hydraulic properties of soil in South Africa (Myeni et al., 2021) uses bulk density, soil texture and organic matter content as indicators. Both models were validated using soil data from South African regions.

Machine learning

Machine learning has emerged as a valuable tool in soil science, with applications ranging from soil classification and mapping to modelling. For predicting soil characteristics like organic carbon content, pH, Cation Exchange Capacity (CEC), and bulk density, algorithms like multiple linear regression, convolutional neural networks, random forests, artificial neural networks, support vector machines, decision trees, Cubist, memory-based learning, partial least square regression, principal component analysis and multivariate adaptive regression splines, have been used (Wang et al., 2023; Aydin et al., 2023; Bondi et al., 2018). PTFs can be developed using machine learning that are useful for estimating soil hydraulic properties (Szabo et al., 2021).

When trained on large datasets of direct measurements and inferred properties, empirical studies have shown that these methods are effective at predicting soil properties with high accuracy (Benke et al., 2020). A recent study conducted in Sri Lanka employed various machine learning algorithms to create PTFs for tropical Sri Lankan soils (Gunarathna et al., 2019). The findings showed that random forest was the most reliable algorithm for creating PTFs in this situation. A large number of additional machine learning algorithms used in soil science.

Ensemble learning combines several base models to produce an optimal predictive model that has been employed to enhance the accuracy of PTFs for soil hydrology (Lamorski et al., 2008). Specifically, two ensemble methods, bagging and additive regression, were utilised in a study to enhance the accuracy of a single regression model. The results indicated that ensemble methods are widely used in improving the performance of data-driven regression models (Cisty et al., 2012). Partial least squares regression, Cubist and random forests were identified as model calibration methods that are accurate and reliable (Dangal et al., 2019). This technique will also be used in this study.

Evaluation of pedotransfer functions

Statistical measures like the Root Mean Square Error (RMSE), coefficient of determination (R^2), and residual plots are frequently used to assess PTFs (Romano & Palladino, 2002; Myeni et al., 2021). These statistical measurements are used to evaluate the accuracy and dependability of the PTFs in predicting the hydraulic properties of the soil.

In soil science, the RMSE statistic is frequently used to assess the precision of models that forecast soil characteristics like hydraulic conductivity and water retention. It is calculated as the square root of the mean of the squared differences between the predicted and observed values and measures the difference between the predicted and observed values (Chai & Draxler, 2014). RMSE has been used in several studies to assess how well PTFs predict soil hydraulic properties. (e.g. Wösten et al., 1999; Klopp et al., 2020). The smaller the RMSE value, the better the fit of the model to the data.

The statistical measure known as the coefficient of determination (R^2) illustrates how closely the regression line resembles the actual data points. It has a range of 0 to 1, with 0 denoting that the model does not fit the data at all and 1 denoting that it does so flawlessly (Nagelkerke, 1991). Higher R^2 values indicate better model performance.

Residual plots are graphical representations of the differences between the observed and predicted values that are used to assess the goodness-of-fit of a regression model. They can be used to spot patterns in the data that the model did not account for, such as non-linear relationships or outliers. (Kutner et al., 2005). The most common way to evaluate the regression model's goodness-of-fit is using residual plots to look for a random scatter of points around the zero line, indicating that the model has captured all the relevant information in the data, while deviations from this pattern may indicate that the model is inadequate (Montgomery et al., 2012).

In a study in the Yellow River Delta region of China, a dataset of 100 soil samples was collected from the coastal salt-affected mud farmland, where support vector machine (SVM), multiple linear regression (MLR), and artificial neural network (ANN) models were used to create PTFs for calculating saturated hydraulic conductivity (Ks) based on readily observable soil characteristics. The coefficient of determination (R^2), RMSE, and mean absolute (MAE) error were used to assess the performance of these models. The results indicated that the SVM model outperformed the ANN and MLR models when it comes to predicting Ks, with an R^2 value of 0.89, RMSE of 0.10, and MAE of 0.08, while the MLR model had an R^2 value of 0.77, RMSE of 0.16, and MAE of 0.13, and the ANN model had an R^2 value of 0.85, RMSE of 0.12, and MAE of 0.10 (Yao et al., 2015).

Similarly, another study conducted in South Africa (Myeni et al., 2021) employed SVM, ANN, and MLR models to develop PTFs for estimating soil moisture content based on easily measurable soil physico-chemical properties at field capacity and permanent wilting point. The same statistical indices were used to assess these models' performance. The results of this study also showed that the SVM model outperformed the ANN and MLR models in terms of estimating soil moisture levels, with an R^2 value of 0.91, RMSE of 0.03, and MAE of 0.02, while the MLR model had an R^2 value of 0.82, RMSE of 0.05, and MAE of 0.04, and the ANN model had an R^2 value of 0.87, RMSE of 0.04, and MAE of 0.03 (Myeni et al., 2021).

Research aim, objectives and hypothesis

The aim of this study is to create suitable PTFs for five specific regions in South Africa as well as a general PTF which can be used to determine the saturated hydraulic conductivity (Ksat) from readily measured soil properties.

To reach the aim, the following objectives must be met:

1. Create a soil profile database with values for the readily available soil properties as well as Ksat, representative of South Africa.

2. Using the dataset created in (1), determine PTFs to predict the Ksat with the readily obtainable soil properties.
3. Evaluate the effectiveness of the PTFs.
4. Test the created PTF's against a PTF obtained from the literature.

The hypothesis tested through this project is that by using South African soil data, a local PTF could be created to predict Ksat from readily available soil data, which is adequately accurate and has practical implications for soil science and agriculture. Furthermore, the PTF's created from local data will be better suited to predict Ksat in South African soils than PTF's obtained from literature created with data from elsewhere.

The development of an hydraulic PTF for South African soils is timely and relevant. It will fill a significant gap in the knowledge and provide a valuable tool for researchers and practitioners working in fields such as hydrology, agriculture and environmental management in South Africa. This study will involve collecting a large and diverse dataset of soil properties from across South Africa. Various statistical and machine learning techniques will be used to develop the PTF. The function will then be validated using independent data sets. The study will also explore the incorporation of prediction uncertainty into the PTF, following recent advancements in PTF development.

7.2.2 Materials and methods

A total of 214 samples of disturbed and undisturbed soil were taken from five catchments in South Africa (Table 7.5; Figure 7.6). At each sampling location, the soil was described per soil horizon and classified according to the Soil Classification Working Group (SCWG, 2018). Samples were generally taken from the topsoil, although some subsoil samples were also taken. The disturbed samples were sent to EcoAnalytica and were analysed for:

1. Seven-fraction texture with hydrometer method.
2. Total carbon with dry combustion method.
3. Organic carbon using the Walkley Black method.

The undisturbed core samples were used to determine the soil water retention curve using the pressure plate method, scanned with SI-ware Near Infrared Spectrometer, dried for bulk density, and used to measure Ksat in a laboratory – using the falling head infiltration method. The laboratory work was done at the North West University.

Table 7.5: The size and significance of each catchment where samples were collected.

Site	Unique significance	Size (km^2)
Sabie	Environmental significance, EFTEON site	5043
Olifants	Coal mining	4698
uMngeni	Sugar cane farming	466
Tsitsa	Soil erosion	494
Goukou	EFTEON site	1613

EFTEON = Expanded Freshwater and Terrestrial Environmental Observation Network

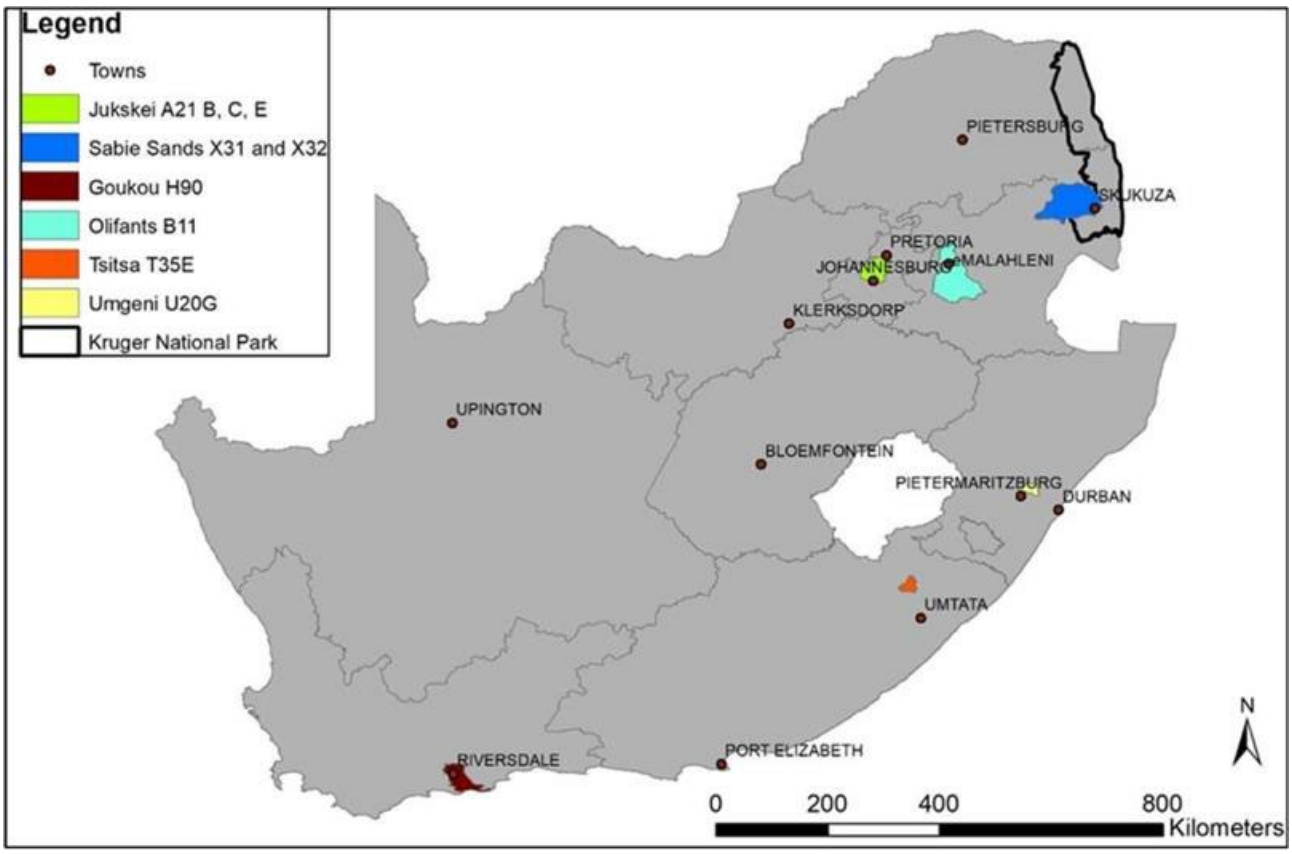

Figure 7.6: Location of the study sites across South Africa.

The sites (Figure 7.6) were selected to encompass a diverse range of climatic conditions, offering a comprehensive dataset. The Sabie-Sand study site, for instance, registered an average temperature of 20°C and received 540 mm of rainfall in 2022 (SAWS, 2023). At the Olifants study site, 760 mm of rainfall was recorded in 2022, accompanied by a daily average temperature of 16.3°C. Similarly, the Jukskei site experienced a daily temperature average of 15.9°C with 794 mm of precipitation. uMngeni reported an average temperature of 22°C and received 966 mm of precipitation. Umtata, in proximity to the Tsitsa site, typically receives 111.7 mm during the rainy season and maintains an annual average temperature of 17.11°. Lastly, the Goukou site near Riversdale received 407 mm of rainfall with an average temperature of 16.7°C (SAWS, 2023).

The falling head method was used in the laboratory to determine hydraulic conductivity (Figure 7.7). This involved the following steps:

1. Samples were wetted in a tub until 100% saturation from the bottom of the sample to the top.
2. The samples were carefully removed and placed in a pot containing coarse sand to introduce a suction gradient. Fine cloths were situated between the samples and the sand to keep the soil core intact.
3. The single ring infiltrometers were then placed into the sample to a depth of three centimetres.
4. Water then got added to a height of approximately nine centimetres and the infiltration time was taken to a certain hight, depending on the tempo of infiltration.
5. This process was repeated until the infiltration time varied only within a five percent margin to ensure accuracy.
6. The average of the last three measurements were then used to calculate the saturated hydraulic conductivity of each sample.

Figure 7.7: Measuring saturated hydraulic conductivity using the falling head method in the laboratory.

The following formulas were used to calculate the saturated hydraulic conductivity:

$$K = (L * A * \ln(h1/h2)) / (t * (h1 - h2)) \quad (7.3)$$

Here, K represents hydraulic conductivity, L is the length of the specimen, A is the cross-sectional area of the standpipe, h1 is the initial head, h2 is the final head, and t is the time interval.

In addition to saturated hydraulic conductivity, five other soil properties were calculated to be used as inputs for the PTFs: pH in a 1:2.5 solution; bulk density through drying and weighing a sample of known volume; seven-fraction texture with the hydrometer method; total carbon with total dry combustion method; and organic carbon with the Walkley-Black method.

The data were portioned into training and validation datasets using the stratified random sampling method. This method is instrumental in evaluating the effectiveness of PTFs in predicting soil properties. It entails dividing the available dataset into several subgroups based on catchments (e.g. Sabie, Goukou, Olifants, etc.), and then randomly selecting a certain percentage of samples from each sub-group for testing, while the remainder are utilised for model development. This approach ensures that the testing dataset is representative of the overall dataset, allowing for evaluation of PTFs across a diverse range of soil types.

One pedotransfer function from literature was used to test the effectiveness of locally calibrated models. The mathematical equation for the PTF (Saxton & Rawls, 2006) is as follows:

$$Ksat = (7.755+0.0352*silt+0.93*topsoil-0.967*dbd^2 -0.000484*clay^2 -0.000322*silt^2 +0.001/silt-0.0748/LECO-0.643*ln\_(silt)-0.01398*dbd*clay-0.1673*dbd*LECO+0.02986* topsoil*clay-0.03305*topsoil*silt) \quad (7.4)$$

Where silt is the percentage of silt, topsoil is either 1 for topsoil or 0 for subsoil, dbd for dry bulk density, clay for the percentage of clay and LECO for the percentage of organic carbon.

Model validation is a crucial step in both model calibration and validation. It serves to assess the models generated during calibration, and compare the predicted values of soil properties with the actual values obtained through laboratory measurements. Coefficient of determination (R^2), RMSE, ratio of performance to deviation (RPD) and scatter plots were used.

7.2.3 Results

Soil property database

The descriptive statistics show that the soil properties are very diverse and cover a range of values (Table 7.6).

Table 7.6: Basic descriptive statistics for the soil property database for all five catchments. The database is the combination of all five catchments data.

Soil property	Median	Mean	σ	Min	Max
Ksat (mm/h^{-1})	31.0	33.5	194.9	0.2	833.4
Dry bulk density (g/cm^3)	1.2	1.1	0.3	0.3	1.8
Total organic carbon (%)	1.8	1.7	3.1	0.3	28.4
Very coarse sand (%)	5.4	4.4	8.7	0.0	37.1
Coarse sand (%)	7.0	6.2	7.0	0.0	34.2
Medium sand (%)	18.8	14.9	11.1	0.1	52.3
Fine sand (%)	18.5	16.9	13.0	0.3	69.0
Very fine sand (%)	6.8	6.7	9.2	0.7	98.1
Silt (%)	14.4	10.5	13.2	0.2	58.0
Clay (%)	14.7	12.3	11.5	0.3	56.3

σ = standard deviation; Min = minimum; Max = maximum; Ksat = saturated hydraulic conductivity.

Model creation

In total six Cubist models were created to predict Ksat from the predictor soil properties, namely the dry bulk density, total organic carbon and the seven fractions of texture. Regional specific models were created for the Sabie, Olifants, uMngeni, Tsitsa and Goukou areas respectively. One national model was developed using the combined data from the five regions. These models were calibrated with a training dataset and evaluated with the validation dataset. The performance was determined with statistical analysis and scatter plots (Table 7.7; Figure 7.8). A PDF from literature was used to predict Ksat with the same validation data used for the national model (Figure 7.9). The same statistical analysis was done for this model.

Table 7.7: Results for each pedotransfer function created.

Model	R^2	RMSE	RPD
National	0.58	154.96	1.33
National_Lit	0.07	401.96	0.51
Sabie	0.79	58.42	1.23
uMngeni	0.55	310.66	0.83
Tsitsa	1	7.78	1.65
Goukou	0.85	82.18	2.8
Olifants	0.07	61.98	0.71

R^2 = coefficient of determination; RMSE = Root Mean Square Error; RPD = ratio of performance to deviation.

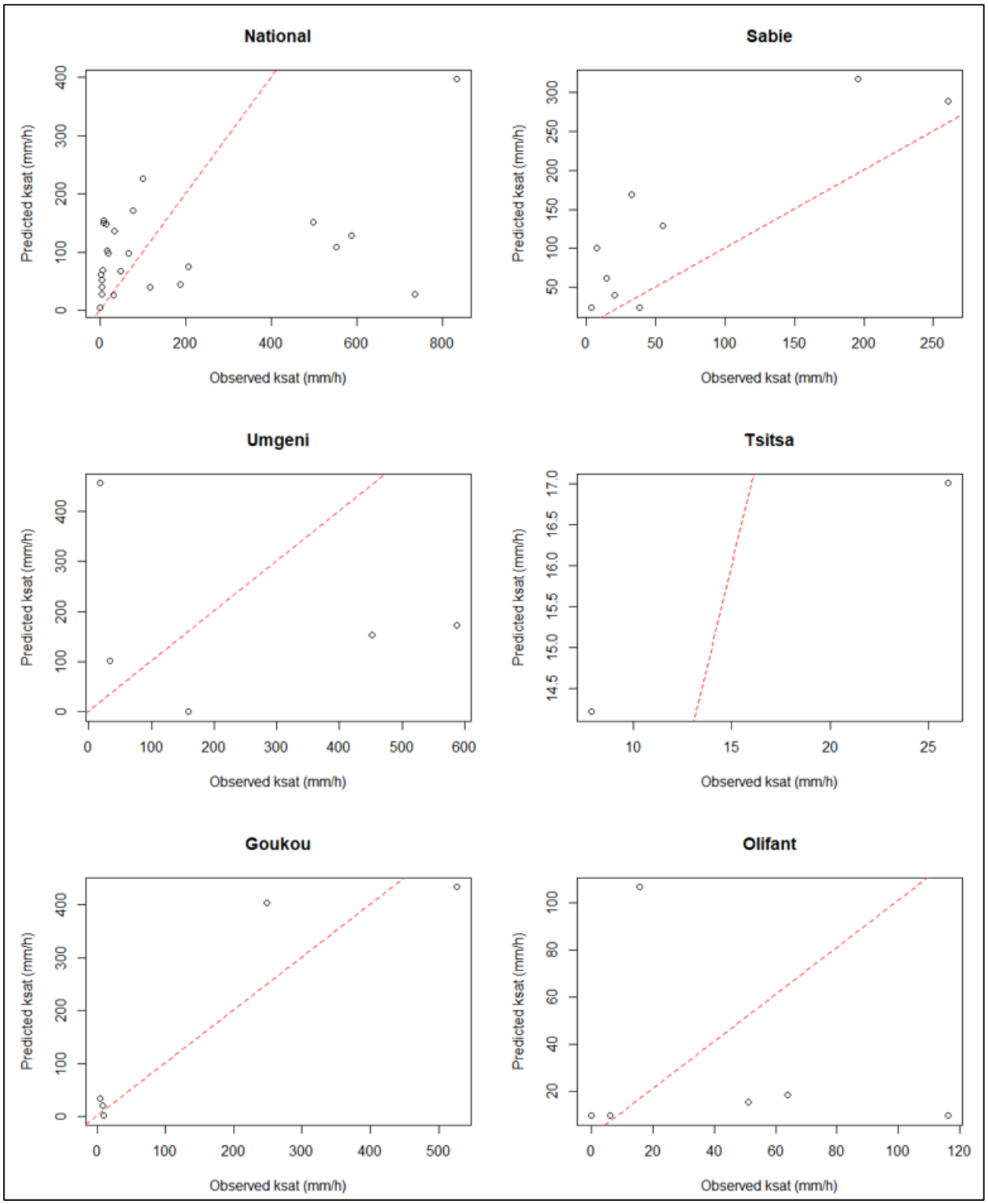

Figure 7.8: Scatter plots show the regression plots for the national and regional validation model for predicting saturated hydraulic conductivity. The red line represents 1:1.

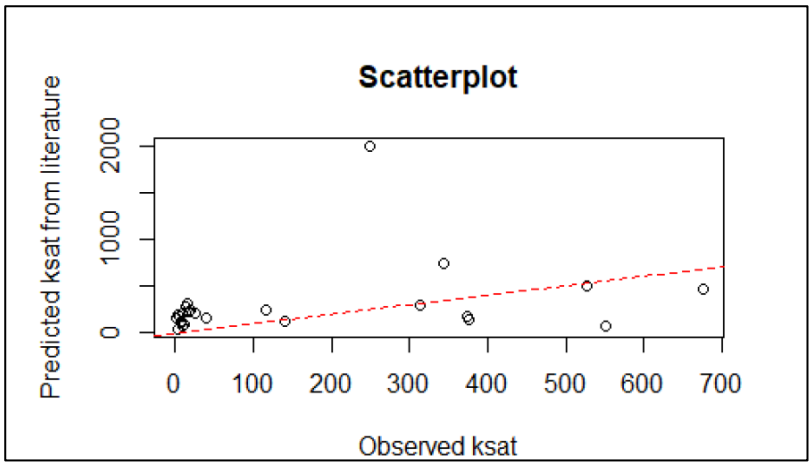

Figure 7.9: Scatter plot showing the predicted Ksat using a pedotransfer function (Saxton & Rawls, 2006). The red line represents the 1:1 line.

7.2.4 Discussion

Evaluating the national model

Results show that the Cubist model is not able to predict the saturated hydraulic conductivity to a satisfactory level (RMSE = 154.96, R^2 = 0.58, RPD = 1.33; Table 7.7). The RMSE is very high, as the average value for Ksat from the database was 127.40. With a RPD below two the model is considered to not be reliable at predicting Ksat (Dangal et al., 2019). Thus, the PTF for the national model did not meet requirements. Looking at the spread of the data the reason for the inaccuracy becomes apparent. Box and whiskers plots (Figure 7.10) show that the soil properties cover a large range of values, with several extreme outliers for nearly all the soil properties. While covering a large range of values for each soil property is a good thing, as it allows for a robust model to be created, it also means one needs more data to be able to create an accurate model. The number of samples used in this study was inadequate to allow for an accurate PTF to be created.

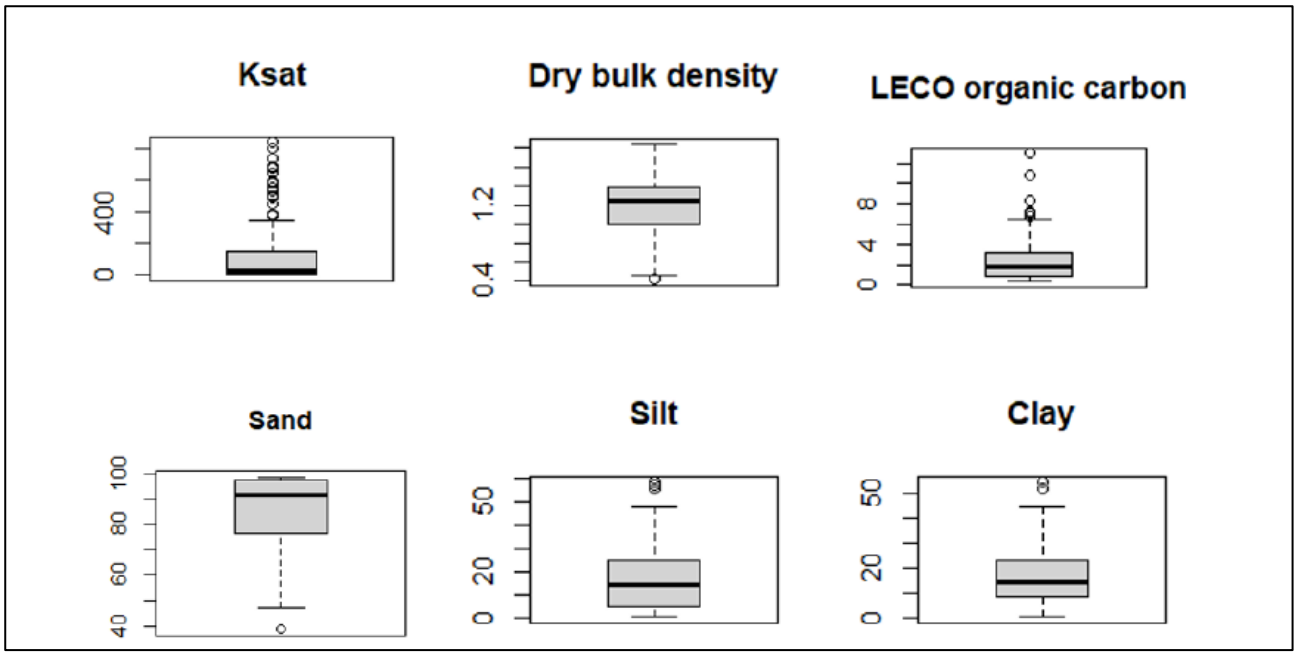

Figure 7.10: Box and whisker plots showing the distribution of the various variables used in the study.

Evaluating the regional models

When the accuracy and reliability of the regional models are evaluated the number of samples in each catchment needs to be considered, as many of the catchments had very few soil samples collected leading to inaccuracy when developing the model.

When considering the RPD, the Goukou model performed the best and was the most reliable at predicting Ksat (Table 7.7). Goukou had a total of 23 samples. The Tsitsa model performed well, but the biggest problem with the Tsitsa model was the amount of soil samples collected. At only 11, the four validation samples were inadequate to be able to assess the accuracy of the model correctly. Sabie achieved a R^2 value of 0,79, RMSE of 58,42 and a RPD of 1,23. The Sabie catchment had a total sample count of 47, the most data points for a region in this study. The model for uMngeni had a R^2 value of 0,55, RMSE of 310,66 and a RPD of 0,83. The uMngeni catchment had 22 soil samples. The local model that performed the worst was (Table 7.7). The Olifants catchment had 28 samples.

All the catchment models are inadequate for practical use, as the RMSE is too large to accurately predict Ksat, or too few samples were collected for an adequate evaluation of accuracy. However, despite the poor results, nearly all the catchments had a lower RMSE than the national PTF, except for the uMngeni catchment. This indicates that PTFs are probably better defined for local areas than for larger areas.

Evaluating the model obtained from literature

Saxton and Rawls (2006) developed a PTF for predicting Ksat using the same predictor variables used in this study. The model was notably less accurate than the nationally calibrated model. The model from Saxton and Rawls (2006) achieved a R^2 of 0,07, RMSE of 401,96 and a RPD of 0,51, compared to the national model with a R^2 of 0,58, RMSE of 154,96 and RPD of 1,33. This emphasises the need for the development of PTF's with local data.

7.2.5 Conclusions

With the national model achieving an R^2 value of 0,58, RMSE of 154.96 and RPD of 1,33 the potential for a national model was evident. Although these results may not be sufficient for practical use, it does indicate that there is potential to develop a PTF for Ksat in South Africa. However more data is needed to represent the diversity of soils to create accurate and robust PTFs.

The importance of locally calibrated models was shown when the PTF created was compared to one from literature (Saxton & Rawls, 2006). Applying the South African data to the existing PTF produced results much worse than the locally produced PTFs. Therefore the hypothesis that locally created will be more accurate and practical than PTFs developed elsewhere can be accepted. The importance of local data to create a PTF was further emphasised by the catchment specific PTFs created. Generally, the results were more accurate, indicating that within South Africa, regional-specific PTFs need to be created.

Going forwards, the creation of hydrological PTFs shows potential, provided sufficient data can be collected to create robust models. The need for soil data and the difficulty obtaining it shows the need for a PTF. In this way Ksat values could be predicted from more easily obtainable soil properties. One way to collect the data needed is to use a region-by-region approach. This would concentrate data gathering into smaller areas, potentially allowing for sufficient data to be collected to represent the soil diversity in the specific area. This should lead to robust accurate PTFs. Over time, when sufficient data has been collected from smaller regional areas, then the development of a national PTF using machine learning could be plausible.

CHAPTER 8: NEAR INFRARED SPECTROSCOPY TO MEASURE SOIL PROPERTIES

Chapter 8 describes how a Near Infrared Spectroscopy (NIRS) can be calibrated to measure soil moisture content and similar hydrological soil properties. NIRS can be a non-destructive, quick and safe way to measure soil properties, provided appropriate calibration algorithms exist. This chapter formed the bulk of Jacques Faul's MSc dissertation.

8.1 CREATING NEAR INFRARED SPECTROSCOPY CALIBRATION ALGORITHMS TO MEASURE SELECTED HYDROLOGICAL SOIL PROPERTIES

8.1.1 Introduction

It has been suggested that the potential cause of the next world war could be a struggle over freshwater resources (Singh et al., 2012; Tignino, 2010), given that it is arguably the most precious resource known to humankind. It has also been said that water is the cornerstone of all life (Skalko, 2013). The importance of water on our planet cannot be overstated, as it forms the foundation of the biochemical function in all living organisms (Chaplin, 2001). Simply put, without water, life on earth would simply cease to exist.

Only about 2.5% of water on Earth is freshwater usable by humans, approximately 1.7% of global freshwater reserves are found below the surface as groundwater, while the rest is confined to glaciers and the atmosphere (Kikkas & Kulik, 2018). The availability of freshwater has declined over recent years due to the rapid growth of global population, the unrelenting progression of urbanisation, and the surge in demand for goods and services that often involve clean, freshwater to manufacture and deliver (Hanjra & Qureshi, 2010). The situation in South Africa is worse than the global average, as the country is seen as being water scarce (Viljoen & Van der Walt, 2018). South Africa is also faced with the growing issue of land availability accompanied by a growing population. In a country that is plagued by erosion and improper land-use practices, the availability of usable land for agriculture only decreases, making effective water management extremely crucial (Phinzi et al., 2020).

Soil is a first-order control of the hydrological cycle (Yamanaka et al., 2007). Understanding its role in the hydrological cycle is thus of utmost importance when it comes to the management of water resources. A critical factor to consider then, is the dynamic, interactive relationship between water and soil, also known as hydropedology (Van Tol et al., 2013). Hydropedology is an important field of study, especially in countries where agriculture under irrigation plays a key role in the economy, like in South Africa (Rapanyane & Ngoepe, 2019). Understanding how water interacts with soil can aid in effectively managing irrigation schedules and prevent issues like waterlogging and soil erosion (Yerro & Ceccato, 2023). The quality and success of land management decisions also rely greatly on the accuracy of soil measurements upon which they are based (Packer et al., 2019). Because South Africa is a developing country with a struggling economy (Rapanyane & Ngoepe, 2019), the expense associated with measuring soil properties is an important issue that needs to be addressed, especially soil moisture (Paterson et al., 2015). Soil characteristics tend to vary on a fine scale (Paterson et al., 2015), making the measurements of soil characteristics even more difficult, demanding sufficient accuracy when they are being taken.

In addition to South Africa's land, soil and water issues, there is also the lack of existing soil databases to monitor the health and development of soils (Paterson et al., 2015). Understanding soil properties and identifying certain soil characteristics, especially in agriculture, can drastically increase productivity by helping people utilise the soil more effectively. Not only can it increase the productivity of agriculture and crop production but may increase the efficiency at which water is managed as well. Conducting soil measurements, especially water related soil measurements, can prove to be challenging due to the

expensive costs, inaccuracy in readings, as well as tedious processes they require (Afzali et al., 2021; Placidi et al., 2021).

Water content in soils has traditionally been measured using various methods, like tensiometers, electrical conductivity using moisture probes, neutron moisture meters and gravimetric measurements. Methods like moisture probes and tensiometers are stationary, while methods like the neutron moisture meter and remote sensing are sometimes expensive. Gravimetric and volumetric soil moisture measurements are furthermore not always reliable in accuracy (Afzali et al., 2021; Placidi et al., 2021). This naturally, leaves room to find a better, more cost-effective solution. When observing newer technologies and methods, spectroscopy shows real promise. Near Infrared Spectroscopy (NIRS) stands out in spectroscopy since it provides portable and seamlessly accurate measurements without disturbing the soil in any way (Knox et al., 2015). NIRS can also provide real-time data to accommodate fluctuating moisture levels. This method is similar to remote sensing, other than that it is portable, less prone to noise, more economical and more rapid than remote sensing solutions (Knadel et al., 2017).

For the scanner to identify what it scans, however, requires a calibration using an algorithm which acts as a lexicon by which the scanner identifies features in the spectra it retrieves. The issue, however, is that there are no freely and readily available NIRS algorithms for predicting soil water content in South Africa. Most of these algorithms either demand membership or payment to access, and algorithms from other countries are not suitable due to the diverse and unique spectral signatures of South African soils.

The hypothesis tested in this study is that NIRS can accurately determine soil water content in a variety of South African soils if efficient calibration algorithms are available.

The aim of the study is to create NIRS calibration models for soil water content and bulk density prediction for a wide variety of soils in South Africa. To test the hypothesis and meet the aim of the study, the following objectives must be met:

1. To establish regional NIRS soil water content and bulk density calibration algorithms for soils from five catchments in South Africa.
2. To determine at which scale NIRS calibration algorithms perform best by establishing water content and bulk density prediction calibration algorithms for each individual catchment.
3. To compare created algorithms of drained upper limit, lower limit and bulk density against freely available international calibration algorithms for the same properties.

8.1.2 Materials and methods

Study sites

Undisturbed soil samples were collected from five diverse catchments spread through South Africa (Figure 8.1). These catchments include the Goukou catchment in the Western Cape, the uMngeni catchment in KwaZulu-Natal, the Tsitsa catchment in the Eastern Cape, the Sabie catchment in Mpumalanga and Limpopo, and the Olifants catchment in Gauteng and Mpumalanga. The samples used for this study was therefore collected from six different provinces and should include a very diverse set of soil properties.

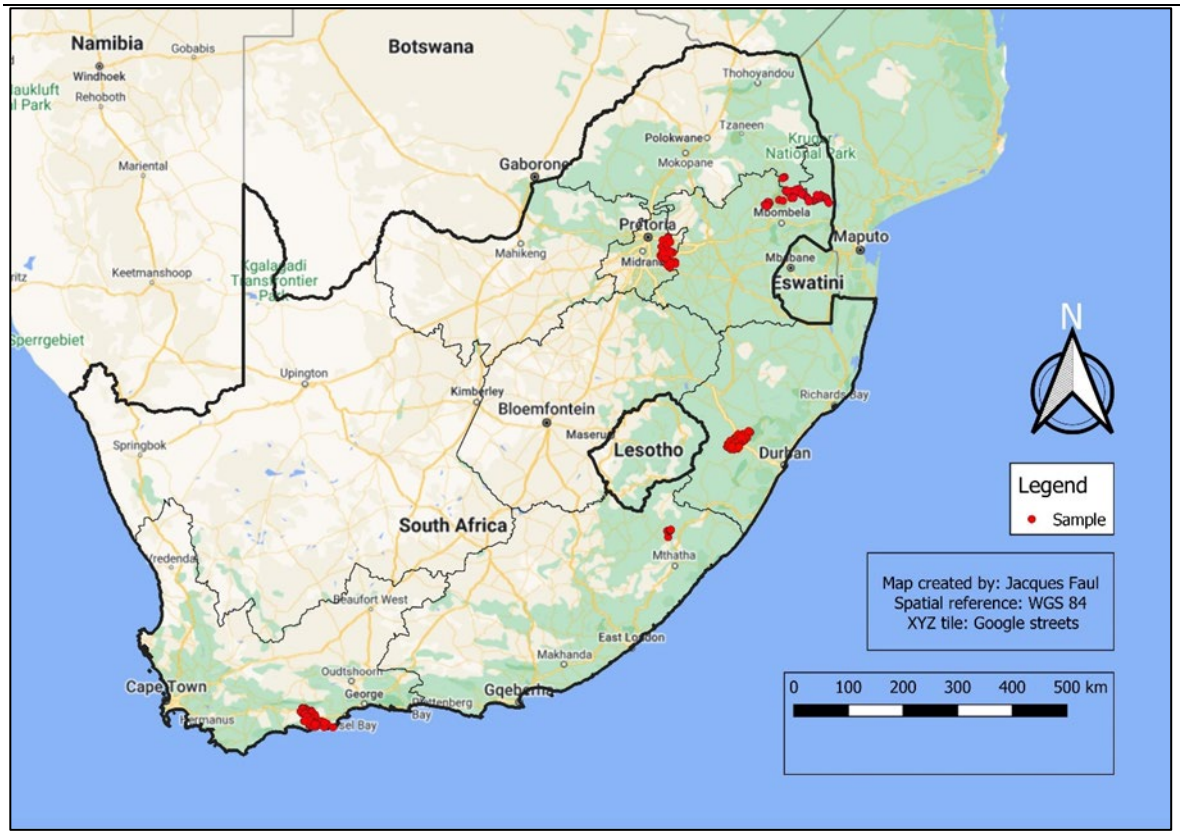

Figure 8.1: The locations of the samples collected for this study.

Methodology

A total of 213 undisturbed soil core samples with known volume were collected from the surface layer of the soil (depth of 0-20 cm) at various locations within the five mentioned catchments. Sample locations for each catchment were determined using conditioned Latin hypercube sampling (Minasny & McBratney, 2006), using covariates for each catchment, giving a number of sample points that are all representative of the spatial variability of the catchment.

After the samples were collected, they were moved to a laboratory where they were wetted until saturated, weighed and scanned with the handheld Neospectra NIRS by placing the scanner on top of the sample and scanning for a total time of 14 seconds. The handheld Neospectra NIRS has a spectral resolution of 16 nm and the spectral range was from 1 250-2 500 nm.

The samples were then placed in a pressure pot that subjected them to different pressures to force the water from the core sample. The different pressures used were from 33 kPa (drained upper limit), 100 kPa, 500 kPa and 1 500 kPa (lower limit). Once the soil water reached a constant level at each pressure, the sample was weighed and then scanned again with the Neospectra spectrometer.

Each sample will thus have a spectral measurement at five different water levels, giving a total of 1 065 individual scans. Afterwards the samples were oven dried and weighed again to determine the dry bulk density (*pb*). The bulk density can be calculated using:

$$Bulk\ density(\rho b) = \frac{Mass\ of\ the\ dry\ sample\ (g)}{Volume\ of\ the\ dry\ sample\ (cm^3)} \quad (8.1)]$$

The saturated and completely dry weights of the samples were used to calculate the gravimetric water content at each measurement from which the volumetric water content (θv) was calculated using the previous *pb* equation.

$$Volumetric\ water\ content\ (\theta v) = gravimetric\ moisture\ content * bulk\ density \quad (8.2)$$

For the calibration process, approximately 75% of the 1 065 spectra were used in the creation of the calibration model, where the remaining 25% were used as training data in the creation of the algorithm. The fuzzy K-means clustering sample selection algorithm was used to split the data into a training and validation set at a 75:25 ratio on the spectra. Pre-processing methods applied included Standard Normal Variate (SNV), multivariate scatter correction, Standardisation, Savitzky-Golay, and outlier removal (Table 8.1).

Table 8.1: Summary and comparison of the pre-processing methods used.

Pre-processing	Description
No pre-processing	Uses the spectra without any modifications.
Savitzky-Golay	Applies a smoothing filter to the spectra to reduce noise (Mouazen & Al-Asadi, 2018).
Savitzky-Golay + removal of outliers	Applies a smoothing filter to the spectra and then removes any remaining multivariate outliers (Mouazen & Al-Asadi, 2018; Wadoux et al., 2021).
Standard Normal Variate	Standardises the spectra by mean centring and dividing by the standard deviation (Zhang et al., 2019b).
Multiplicative Scatter Correction	Mean centres the spectra and then scales it to have a uniform standard deviation (Zhang et al., 2019b).
Standardisation	Centres the spectra by subtracting the mean and then scales it to have a unit standard deviation (Wadoux et al., 2021).

The calibration algorithms were created using the training dataset with different algorithms, including Partial Least Squares Regression (PLSR), Random Forest (RF) and Cubist, with the pre-processing methods for the water content prediction (Table 8.2). A total of 18 calibrations were created.

Table 8.2: Combinations of models and pre-processing methods used for the volumetric water content prediction on the regional dataset.

Model	Pre-processing
Partial Least Squares Regression Cubist Random Forest	No pre-processing
	Savitzky-Golay
	Savitzky-Golay + Removed outliers
	Standard Normal Variate
	Multiplicative Scatter Correction
	Standardisation

For the bulk density algorithms, the entire spectral dataset of 1 065 spectra were again split into a 75:25 calibration and validation ratio using fuzzy K-means clustering. All three models (PLSR, Random Forest, and Cubist) were used firstly with no pre-processing, and then with Savitzky-Golay and standardisation or outlier removal, depending on which results were best. Giving a total of nine calibrations.

For the catchment-specific calibrations, only the best performing model and pre-processing method were used. This was done for each catchment for volumetric water content and dry bulk density. Similar to the previous calibrations, fuzzy K-means cluster was used on the spectra to split the data into a 75:25 training and validation sets. A total of 213 samples scanned at five different water contents (Table 8.3).

Table 8.3: Number of samples for each catchment at five different water contents.

Catchment	Number of samples	Number of spectra (samples * 5)
Goukou	35	175
Olifants	39	195
Sabie	77	385
Tsitsa	28	140
uMngeni	34	170

The entire dataset for each catchment was used and split into a 75:25 training and validation datasets using fuzzy K-means clustering. This resulted in five catchment specific calibrations for both volumetric water content and dry bulk density.

To compare the created algorithms against freely available algorithms, Open Soil Spectral Library (OSSL) algorithms from the Soil Spectroscopy for Global Good project were used. For the comparison, algorithms were created for the drained upper limit, lower limit and bulk density using all of the data at the lower limit. The OSSL estimation service does not have a prediction option for volumetric water content, but only for drained upper limit, lower limit and bulk density.

It is important to note that the OSSL models were created using samples that were dried and sieved. This study, however, did not scan the samples at completely dried moisture level, due to the fact that water content below the lower limit is not necessarily important for agricultural purposes. For the comparison, we used the driest moisture content available (which is the lower limit) data to compare with the OSSL models. It is, however, not a like-for-like comparison.

The OSSL models were created using a combination of the Cubist machine learning algorithm and the standard normal variate pre-processing method. For this reason, Cubist was used to create models for the drained upper limit, lower limit and bulk density to ensure that the comparison is valid. The data was again split into a 75:25 training and validation dataset using fuzzy K-means clustering. Cubist with no pre-processing was used since standard normal variate did not provide any noticeable improvement on calibration results with Cubist algorithms for predicting volumetric water content using all the data. The validation dataset of each created algorithm is exported as a .csv file, which is then uploaded to the OSSL estimation service for them to make predictions for the same parameters, which was then loaded into R to compare the calibrations.

Validation was conducted on the independent validation dataset. The following statistical measurements were used to evaluate the usefulness of the derived calibration algorithms: R^2, displaying the correlation between the predicted and measured values; the mean error (ME) indicating any bias, the RMSE which indicates the magnitude of error uncertainty; the concordance correlation coefficient (rhoC) which measures the agreement between the predicted and measured values; the ratio of performance to deviation (RPD), which is the ratio between the standard deviation of a variable and the standard error of that variable; ratio of performance to interquartile distance (RPIQ) which is the interquartile distance of the prediction errors.

Additionally, the texture of the samples was determined using the hydrometer method for particle size analysis, and the total percentage carbon was determined using the dry combustion method (LECO).

8.1.3 1Results and discussion

Soil water content database

Some measured soil properties indicate that the samples are very diverse and represent a wide range of soils (Figure 8.2). For the lower limit (Figure 8.2a) and drained upper limit (Figure 8.2b), the data ranges from a very low 7 and 10%, to a very high 63 and 71%, respectively. Looking at the bulk density (Figure 8.2c) the samples cover almost the entire range of bulk density classes (Hazelton & Murphy, 2007), from very low (< 1.0) to high (1.6-1.8), lacking only the very high class of bulk density which is higher than 1.9, which is considered to be soils that are very compacted (Hazelton & Murphy, 2007). The clay content of the samples (Figure 8.2d) covers the entire range of all texture grades from sand to heavy clays (Hazelton & Murphy, 2007). The range of organic carbon in the samples (Figure 8.2c) also cover the entire range of organic carbon levels from extremely low (< 0.4% carbon) to organic soil material (> 8.7%; Hazelton & Murphy, 2007) The box and whiskers plots indicate that the samples represent a very good range of diverse soil conditions.

Most catchments represent a significant amount of variability in soil water content parameters (Figure 8.3; Figure 8.4), and bulk density (Figure 8.5) apart from the Tsitsa catchment. The uMngeni catchment has the most variability with both the drained upper limit and the lower limit ranging significantly, as well as the bulk density ranging from almost 0.3 to 1.3 g/cm^3. The extremely low bulk density found within the uMngeni catchment can be attributed to abundant humic soils in the area. However, it may also be possible that the upper litter layer of organic matter was sampled.

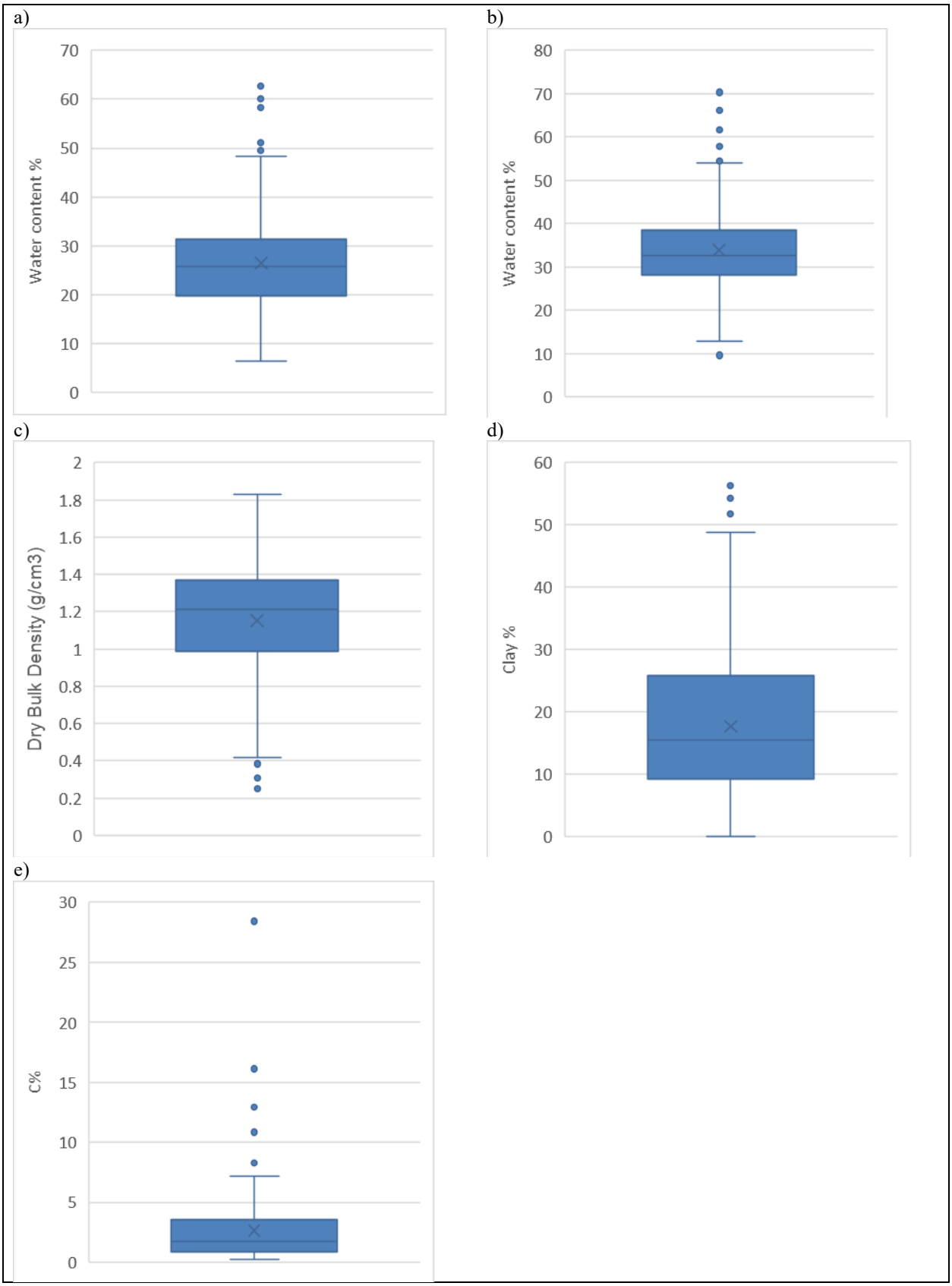

Figure 8.2: Box and whiskers plots for selected soil properties: a) Drained Upper Limit (DUL), b) Lower Limit, c) the dry bulk density (DBD) in g/cm3, d) the clay percentage of all samples, and e) the carbon percentage of all samples.

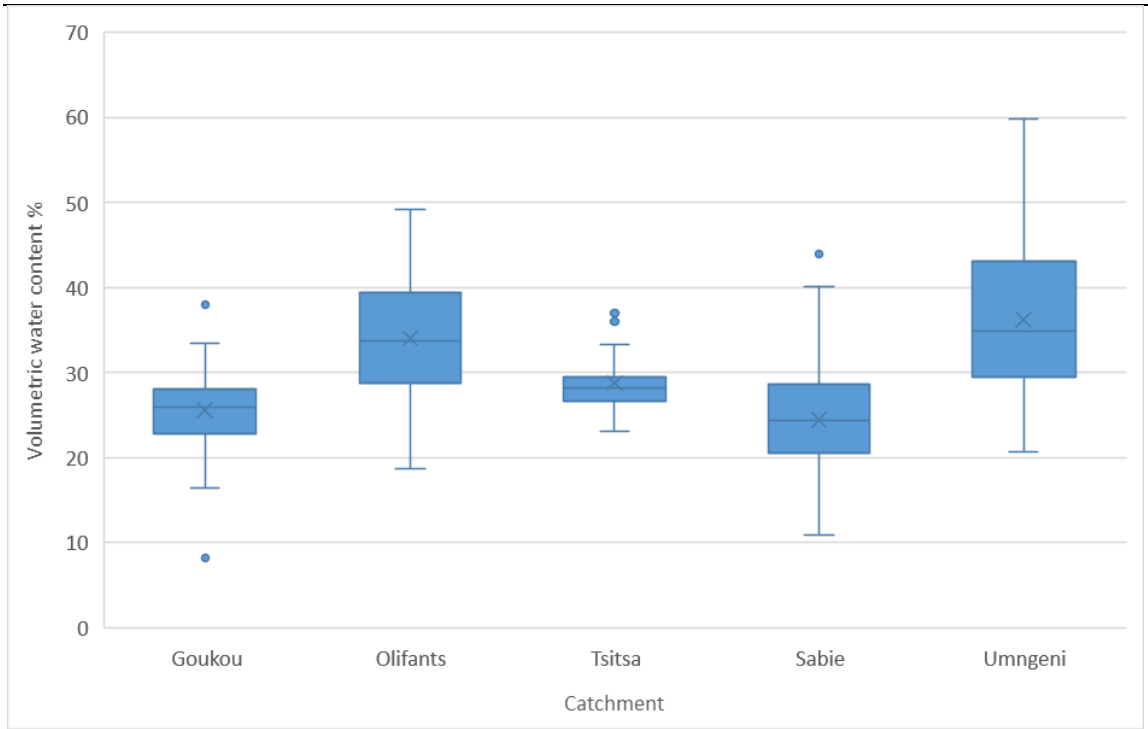

Figure 8.3: Box and whiskers plot of the volumetric water content (%) for each catchment at the drained upper limit.

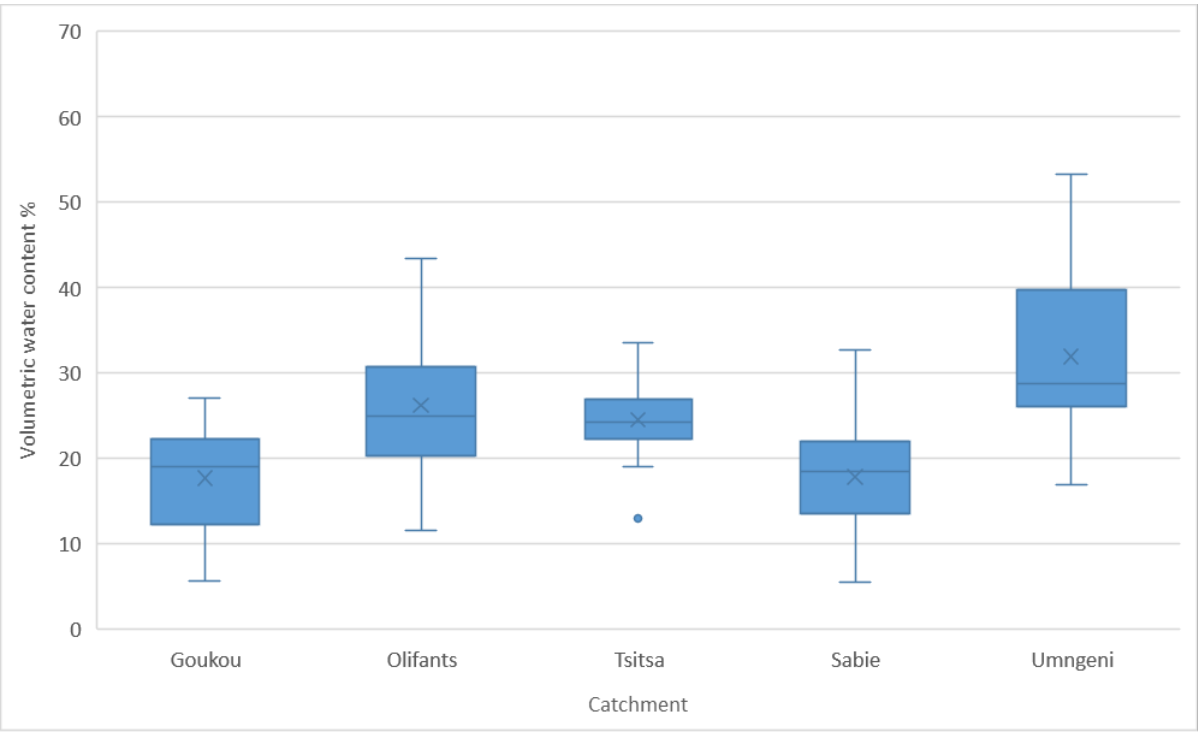

Figure 8.4: Box and whiskers plot of the volumetric water content (%) for each catchment at the lower limit.

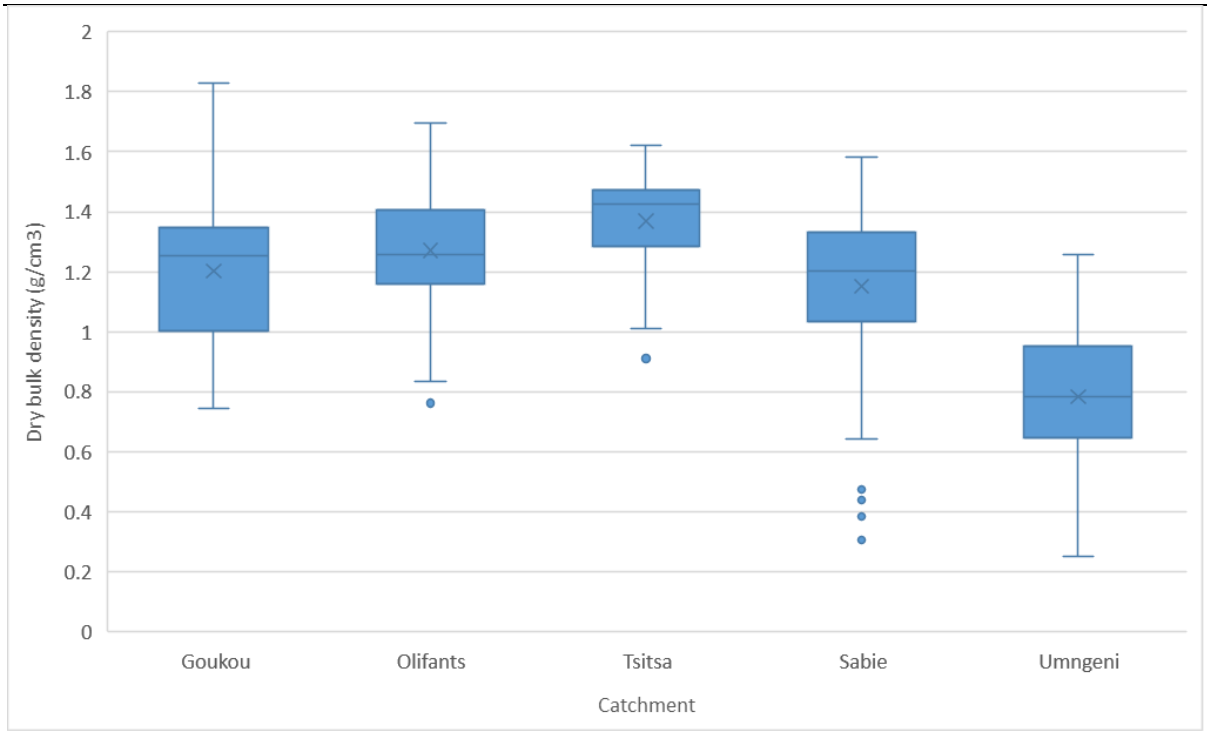

Figure 8.5: Box and whiskers plot of the dry bulk density (g/cm3) for each catchment.

Volumetric water content

For the calibration results, most of the volumetric water content models calibrated well with rhoC values above 0.6, indicating that the models performed satisfactorily in terms of the calibration data (Table 8.4). This also means that the models are capturing both the systematic bias and random error in the predictions, resulting in a reliable and consistent relationship between the predicted and observed values. The RPIQ values of the PLSR calibrations did not perform as well, and are all below two, while the Cubist and RF RPIQ values are all above two which is satisfactory. The RMSE values of all the calibrations are below 8.6%, which is moderately accurate, but leaves desire for improvement.

For the validation, Cubist and RF performed well with Savitzky-Golay and had relatively similar results, with RF having a slightly lower RMSE of 6.96% in comparison to Cubist with an RMSE of 7.01% (Table 8.4). Moreover, RF had a slightly higher R^2, rhoC, and RPD compared to Cubist, but with similar RPIQ values. RF did, however, have a higher bias of 0.75, while Cubist only has 0.38. While the performance of Cubist and RF are largely similar, RF shows signs of overfitting due to the calibration performing well, but the validation performing poorly, which may be due to the relatively small dataset size. RF typically performs better with larger datasets, due to its usage of decision trees that are prone to overfitting when there is not enough data to train the algorithm (Cosenza et al., 2020). The combination of Cubist and Savitzky-Golay shows promise, since Cubist performs better on smaller datasets (Katuwal et al., 2020), while Savitzky-Golay's derivative and smoothing capabilities work well with Cubist to ensure a robust calibration algorithm (Zimmermann & Kohler, 2013). Similarly, Clingensmith and Grunwald (2022) concluded that PLSR underperformed in predicting soil properties in vis-NIR when compared to RF and Cubist due to data complexity and non-linear relationships.

The results are promising but can be improved. Liang et al. (2012) obtained substantially better results in predicting soil water content using NIRS, with an RMSE of 1.2% and an R^2 of 0.99. The study, however, created calibration algorithms on a much smaller area with much less spatial variation, accompanied by a spectrometer with a higher spectral resolution of only 6.8 nm, which may significantly increase the accuracy of results. Bullock et al. (2004) produced similar results to this study in predicting the water content in five

soil horizons with an RMSE of 6.4% and a R^2 of 0.95. They also utilised a spectrometer with a very high resolution of only 2 nm with a limited sample variation of five soil horizons. Considering information from these studies, the results are quite satisfactory, considering the spatial and property variation of the samples are very high, and the spectral resolution of the scanner used is only 16 nm. Calibrations on a smaller area with more samples should improve results.

Dry bulk density

The models for bulk density calibrated rather well, with rhoC values ranging from 0.66 to 0.95 highlighting that there is a relatively good agreement between predicted and actual values (Table 8.5). The RMSE values are all below 0.2 g/cm^3 which is acceptable but requires further refinement for increased accuracy. Apart from the PLSR and Cubist models without pre-processing, the models all have RPIQ values above two, indicating good predictive accuracy and robustness.

For the validation, the Cubist and RF models performed the best, yielding similar results with both algorithms having an RMSE of 0.16 g/cm^3 (Table 8.5). The best Cubist model combination (Cubist and Savitzky-Golay with removed outliers) yielded an R^2 value of 0.71, a rhoC of 0.82, an RPD of 1.86, and an RPIQ of 2.31 against the best RF model's R^2 of 0.7, rhoC 0.81, RPD of 1.84, and RPIQ 2.29. All three of the models had a bias value of 0.01. Cubist paired with Savitzky-Golay again performed well, in this case the removal of outliers further improved results.

Katuwal et al. (2020) predicted soil bulk density using vis-NIRS using a spectrometer with a resolution of 0.5 nm and provided promising results with an RMSE of 0.04 g/cm^3 and an R^2 of 0.94. The study used samples from different datasets, giving a total of 2 462 samples from across Denmark. Considering that the number of samples were much higher than this study (1 065 data points), and the spectral resolution was much higher, the results can be improved. Davari et al. (2021) achieved poorer results for bulk density calibration with an RMSE of 0.15 g/cm^3 and an R^2 of 0.26. The study utilised vis-NIRS on 220 soil samples within a 2 000 m^2 area, and concluded that a larger number of samples could possibly improve calibration results.

Table 8.4: Results of the volumetric water content algorithms.

Algorithm and pre-processing method	Calibration						Validation					
	ME	RMSE	R^2	rhoC	RPD	RPIQ	ME	RMSE	R^2	rhoC	RPD	RPIQ
PLSR (without pre-processing)	0	8.54	0.49	0.66	1.4	1.7	0.7	8.12	0.38	0.65	1.27	1.65
PLSR (Savitzky-Golay)	0	8.21	0.53	0.69	1.46	1.78	1.78	8.6	0.31	0.59	1.21	1.45
PLSR (Savitzky-Golay + outliers)	0	8.22	0.52	0.68	1.44	1.74	0.68	8.94	0.34	0.6	1.23	1.46
PLSR (Standard Normal Variate)	0	8.71	0.46	0.63	1.36	1.66	1.52	8.99	0.32	0.56	1.21	1.43
PLSR (Multiplicative Scatter Correction)	0	8.66	0.45	0.62	1.35	1.66	1.03	9.13	0.37	0.56	1.26	1.41
PLSR (Standardisation)	0	8.09	0.54	0.7	1.47	1.79	0.2	8.1	0.42	0.67	1.32	1.63
Cubist (without pre-processing)	0.85	7.23	0.62	0.74	1.63	1.93	1.13	7.96	0.49	0.65	1.41	1.81
Cubist (Savitzky-Golay)	0.96	6.70	0.68	0.77	1.76	2.18	0.38	7.01	0.6	0.72	1.57	1.81
Cubist (Savitzky-Golay + outliers)	0.82	6.31	0.71	0.8	1.87	2.21	-0.51	7.7	0.53	0.65	1.46	1.78
Cubist (Standard Normal Variate)	0.55	6.87	0.66	0.75	1.71	2.12	0.32	7.85	0.5	0.62	1.42	1.5
Cubist (Multiplicative Scatter Correction)	0.57	6.71	0.68	0.77	1.77	2.17	0.25	7.77	0.48	0.61	1.39	1.63
Cubist (Standardisation)	0.77	7.05	0.63	0.75	1.65	1.98	0.75	8.18	0.49	0.64	1.41	1.74
RF (without pre-processing)	0.08	3.67	0.9	0.94	3.17	3.84	0.46	8.86	0.41	0.57	1.31	1.57
RF (Savitzky-Golay	0.02	3.14	0.93	0.96	3.75	4.68	0.75	6.96	0.61	0.75	1.61	1.8
RF (Savitzky-Golay + outliers)	0.01	3.11	0.93	0.96	3.75	4.42	0.5	7.07	0.63	0.75	1.64	2.04
RF (Standard Normal Variate)	0.03	3.86	0.9	0.93	3.1	3.91	1.62	8.68	0.3	0.47	1.2	1.4
RF (Multiplicative Scatter Correction)	0.03	3.82	0.89	0.93	3.06	3.83	1.38	9.08	0.35	0.48	1.24	1.33
RF (Standardisation)	0.04	3.76	0.9	0.94	3.17	3.94	0.25	8.28	0.39	0.57	1.29	1.53

PLSR = Partial Least Squares Regression; RF = Random Forest; ME = Mean Error; RMSE = Root Mean Square Error; R^2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

Table 8.5: Results of the dry bulk density algorithms.

Algorithm and pre-processing method	Calibration						Validation					
	ME	RMSE %	R^2	rhoC	RPD	RPIQ	ME	RMSE %	R^2	rhoC	RPD	RPIQ
PLSR (without pre-processing)	0	0.2	0.54	0.71	1.48	1.97	-0.01	0.2	0.54	0.69	1.47	1.8
PLSR (Savitzky-Golay)	0	0.19	0.6	0.75	1.57	2.09	0	0.21	0.44	0.66	1.34	1.69
PLSR (Standardisation)	0	0.19	0.59	0.74	1.56	2.06	-0.01	0.19	0.58	0.74	1.54	2
Cubist (without pre-processing)	0.02	0.17	0.7	0.66	0.76	1.71	0.01	0.19	0.64	0.59	0.7	1.57
Cubist (Savitzky-Golay)	0.01	0.11	0.88	0.85	0.91	2.62	0.01	0.16	0.68	0.67	0.79	1.74
Cubist (Savitzky-Golay + outliers)	0.01	0.11	0.88	0.85	0.91	2.6	0.01	0.16	0.72	0.71	0.82	1.86
RF (without pre-processing)	0	0.1	0.94	0.89	0.93	3.01	0	0.23	0.39	0.37	0.5	1.27
RF (Savitzky-Golay)	0	0.07	0.96	0.95	0.97	4.48	0.01	0.16	0.66	0.66	0.78	1.71
RF (Savitzky-Golay + outliers)	0	0.07	0.96	0.95	0.97	4.4	0.01	0.16	0.71	0.7	0.81	1.84

PLSR = Partial Least Squares Regression; RF = Random Forest; ME = Mean Error; RMSE = Root Mean Square Error; R^2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

Catchment specific calibrations

Site-specific calibration algorithms were created for each catchment to compare with the regional algorithms that used all the samples. Cubist paired with Savitzky-Golay was used, since it showed consistent and robust results previously. It is important to note that this section only focuses on the validation statistics for model evaluation.

It is evident that the results are noticeably better for catchments than for the regional scale, apart from the uMngeni catchment (Table 8.6). The results are satisfactory. A rhoC value of 0.58 for the Tsitsa catchment indicates that there is a relatively fair agreement between the predicted values and the observed values but is, however, not as good as the rhoC values of the other catchments. Apart from the Tsitsa algorithm, all of the rhoC values are above 0.69, which indicates a good agreement between the predicted and measured values. Of all the catchments, the Tsitsa catchment had the lowest number of data points of 140, which might explain the poor R^2 value of 0.32 and the lowest rhoC value of 0.58. The most robust catchment algorithms were the Sabie and the Olifants catchments. The Sabie achieved a rhoC value of 0.74 and a satisfactory RPIQ value of 2.06, while the Olifants achieved a high rhoC value of 0.81 and an RPIQ value of 1.94. This is because the Sabie catchment had the most data points of 385, and the Olifants the second most of 195.

The variability of the soil properties within the different scale areas (Figure 8.6; Figure 8.7; Figure 8.8) helps to explain the findings, as the catchments with the least spatial variability performs better, which is the case for the Tsitsa, Sabie and Olifants catchments. Catchments with significant variability like the uMngeni performed poorly, as expected. Others have concluded that local calibrations prove superior to regional and international calibrations due to less spatial variation and samples being more concentrated to better capture the spatial variation that may be present (Canal Filho et al., 2023; Koirala et al., 2022).

Table 8.6: Results of the catchment specific calibrations for drained upper limit using Cubist.

Catchment	ME	RMSE (%)	R^2	rhoC	RPD	RPIQ
Goukou	1.06	5.35	0.58	0.75	1.55	1.33
Olifants	0.59	5.07	0.7	0.81	1.85	1.94
Sabie	1.3	5.52	0.58	0.74	1.56	2.06
Tsitsa	0.22	3.83	0.32	0.58	1.23	1.4
uMngeni	0.6	8.63	0.54	0.69	1.49	1.86

ME = Mean Error; RMSE = Root Mean Square Error; R^2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

For the catchment-specific dry bulk density calibrations, again using Cubist paired with Savitzky-Golay pre-processing, the Sabie catchment yielded the best results (Table 8.7). The predictions for the Goukou, Olifants and uMngeni catchment are relatively poor, whereas the predictions for the Sabie and the Tsitsa are significantly better with RMSE values below 0.1 g/cm^3 (Katuwal et al., 2020). The results of the catchment calibrations naturally leave the desire for improvement, which can be achieved by increasing the number of samples in the area to better represent the spatial variation and increase algorithm performance. Although these models might not be suitable for precision measurements, they would still be useful for agricultural water management, irrigation, as well as creating real-time moisture maps to monitor moisture changes in the soil.

Table 8.7: Results of the catchment specific calibrations for dry bulk density (g/cm3) using Cubist.

Algorithm	ME	RMSE (g/cm3)	R^2	rhoC	RPD	RPIQ
Goukou	0.04	0.14	0.64	0.76	1.68	1.72
Olifants	0.02	0.14	0.64	0.74	1.69	1.78
Sabie	0	0.08	0.92	0.96	3.52	3.75
Tsitsa	0	0.1	0.46	0.62	1.39	1.2
uMngeni	0.02	0.17	0.45	0.54	1.37	2.46

ME = Mean Error; RMSE = Root Mean Square Error; R^2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

Comparing created algorithms against freely available algorithms

It is evident that the algorithms created here are noticeably more accurate than the OSSL models for water content (Figure 8.6; Table 8.8). It is important to note that the OSSL models were calibrated with dry samples, while the created algorithms were created with the data for the lower limit, as this was the dryest soils we had. Thus, the calibrations from this project should be more accurate than that of the OSSL. The OSSL algorithm for the drained upper limit indicated poor results with an RMSE of 17.12%, a mean error of 12.38, and very low RPD and RPIQ of 0.59 and 0.51 respectively. The created algorithm for the drained upper limit performed better, with an RMSE of 8.89%, a mean error of 1.98, and an RPD and RPIQ of 1.13 and 0.99 respectively. Although the created algorithms results are better, they are still too poor to be practically applied in the field, which may be due to the spatial variation of all the samples used in the calibration.

A similar outcome was found with the lower limit calibrations (Figure 8.7; Table 8.9). The OSSL model for the lower limit performed very poorly and the created algorithm performed better (Table 8.9). Even though the results from the created algorithm are better, they still lack the accuracy and reliability that is desired. This is disappointing, as lower limit data was used to create the calibration algorithm.

It was expected that the OSSL models would not accurately predict the South African soil values, as local soils were not included in their calibration algorithms. Soils were collected from across the globe and scanned under different measuring protocols, which may significantly hinder large-scale soil prediction models (Zhou et al., 2022). As for the created algorithms, the data points used still vary greatly due to large variation, where local calibrations would probably perform better.

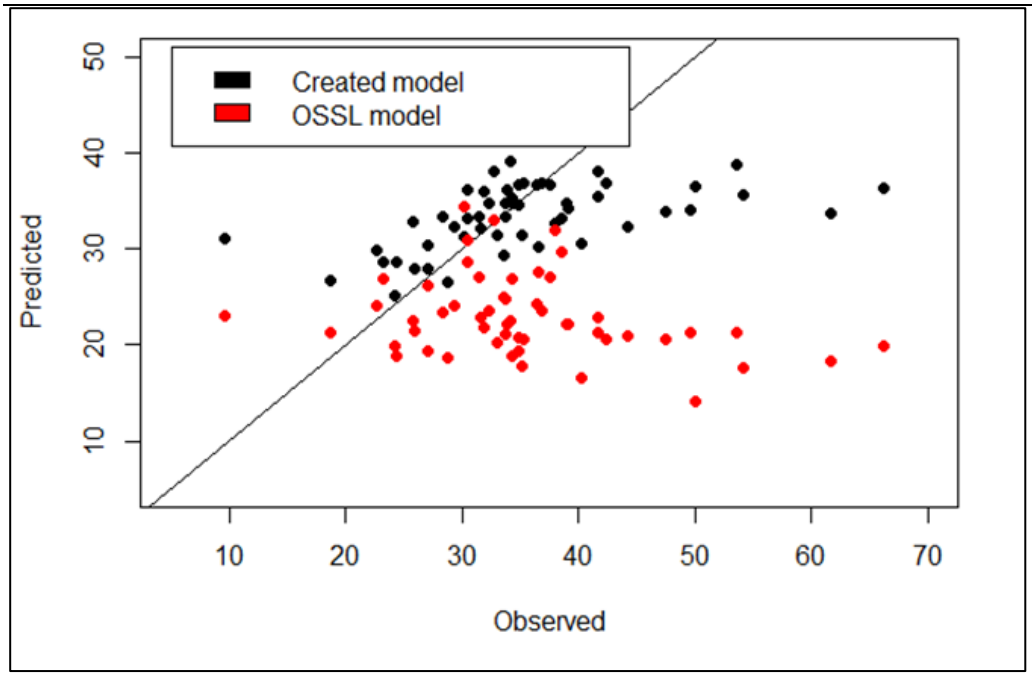

Figure 8.6: Comparison of the created Cubist model of the drained upper limit water content (%) against the OSSL model

Table 8.8: Validation results of the created Cubist model of the drained upper limit water content (%) against the OSSL model.

Model	ME	RMSE	R^2	rhoC	RPD	RPIQ
OSSL model	-12.38	17.12%	-1.95	-0.09	0.59	0.51
Created model (without pre-processing)	-1.98	8.89%	0.2	0.31	1.13	0.99

OSSL = Open Soil Spectral Library; ME = Mean Error; RMSE = Root Mean Square Error; R2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

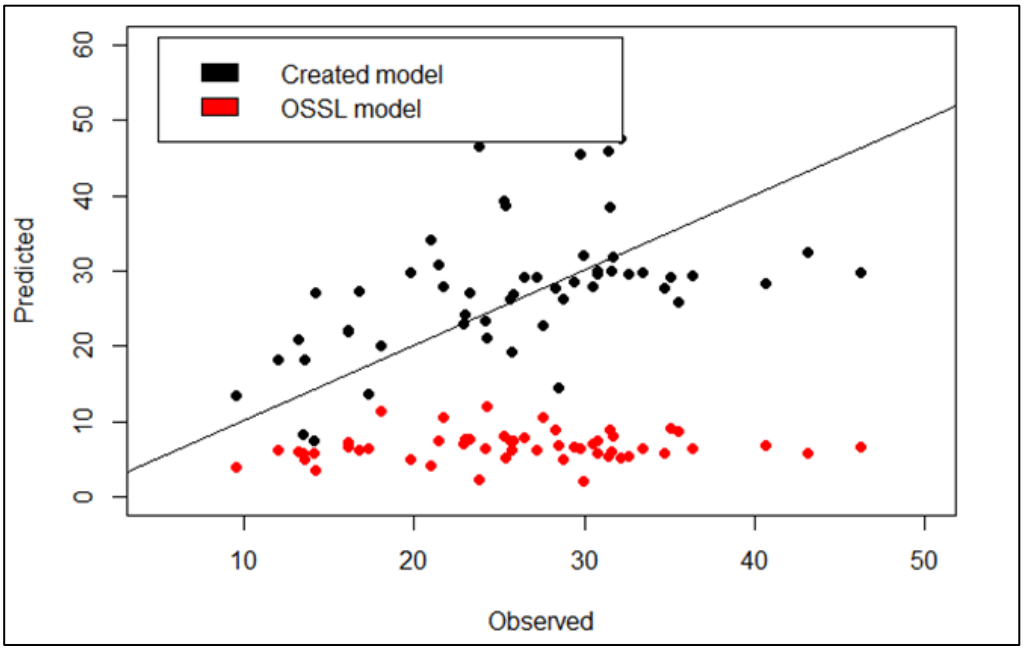

Figure 8.7: Comparison of the created Cubist model of the lower limit water content (%) against the OSSL model.

Table 8.9: Validation results of the created Cubist model of the lower limit water content (%) against the OSSL model.

Model	ME	RMSE	R^2	rhoC	RPD	RPIQ
OSSL model	-19.24	20.85%	-5.72	0.01	039	0.5
Created model (without pre-processing)	-1.57	8.35%	-0.08	0.5	0.97	1.24

OSSL = Open Soil Spectral Library; ME = Mean Error; RMSE = Root Mean Square Error; R2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

The created algorithms were more accurate in determining the dry bulk density of the soil (Figure 8.8; Table 8.10), where the RMSE of the created model is noticeably better at 0.22 g/cm^3 than the RMSE of the OSSL model at 0.38 g/cm^3. These results are unfortunately poor in comparison to other studies (Katuwal et al., 2020), which requires an RMSE of < 0.1 g/cm^3 to be used effectively in the field.

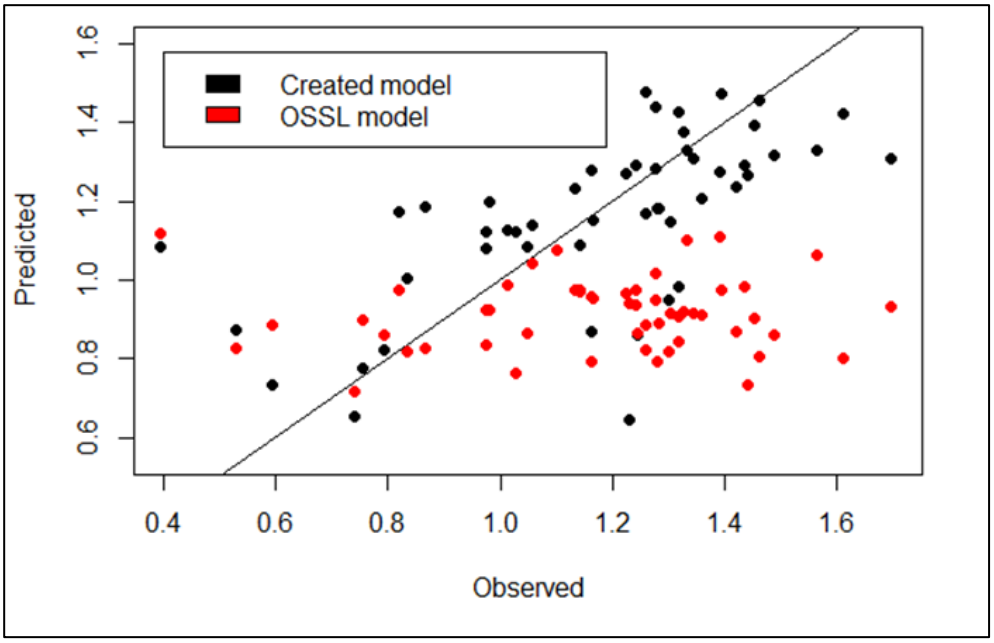

Figure 8.8: Comparison of the created Cubist model of the dry bulk density (g/cm3) against the OSSL model.

Table 8.10: Validation results of the created Cubist model of the dry bulk density (g/cm^3) against the OSSL model.

Model	ME	RMSE (g/cm^3)	R^2	rhoC	RPD	RPIQ
OSSL model	-0.26	0.38	-1.02	0.01	0.71	0.8
Created model (Without pre-processing)	-0.01	0.22	0.33	0.58	1.23	1.38

OSSL = Open Soil Spectral Library; ME = Mean Error; RMSE = Root Mean Square Error; R2 = correlation coefficient; rhoC = Lin's concordance coefficient; RPD = Ratio of Performance Deviation; RPIQ = Ratio of Performance to Interquartile Distance.

Various studies (Chen et al., 2021; Koirala et al., 2022; Mouazen & Al-Asadi, 2018) also concluded that increasing the number of samples for calibration can improve results, and that the development of local calibrations can feed regional and national calibrations, and that regional calibrations can furthermore aid in ultimately improving international calibrations. But for the comparison of international models, they served the purpose of highlighting that local calibrations are required to create useful calibration algorithms, as shown for the Sabie River catchment.

8.1.4 Conclusions and recommendations

The first objective, to create calibration algorithms for regional water content, was met by creating 18 algorithms to predict volumetric water content. The most accurate algorithm was Random Forest with Savitzky-Golay pre-processing, probably because Random Forest performs very well with large datasets.

The second objective, to determine at which scale NIRS calibration algorithms perform best, was achieved by creating calibration algorithms for each catchment using Cubist and Savitzky-Golay pre-processing. The results showed that the catchment calibrations performed significantly better than the regional algorithms.

For the third objective, the process was repeated for bulk density, where nine algorithms to predict dry bulk density were created. The best performing algorithm was Cubist with Savitzky-Golay pre-processing. Catchment calibrations for dry bulk density were also created using Cubist and Savitzky-Golay pre-processing. Again, the results were noticeably better than the regional calibration results, again supporting the theory that local calibrations are necessary for enhanced accuracy and reliability.

The fourth and final objective was to compare created algorithms to freely available international algorithms. Although differences in sampling meant the comparison was not exactly equivalent, the results indicated that the created algorithms were significantly more accurate than the OSSL algorithms. This was expected since the OSSL algorithms utilise data from across the world that differs from South African soils, and that were collected using different methods.

The hypothesis that NIRS can effectively predict soil water content in a variety of soils is thus accepted, with the condition that local calibrations are to be made for more accurate predictions. Because soil characteristics vary on such a fine scale, especially South African soils that are so diverse, local calibrations for NIRS are highly recommended. Although this might require more calibrations in the long term, the accuracy and reliability of algorithms will certainly gain from it, which will result in better water management decisions that will promote a more sustainable future where water scarcity is mitigated for future generations. It is also recommended to improve handheld NIRS devices to have a better spectral resolution, as this may show an improvement in results. Furthermore, it would be advisable to increase the number of samples in areas to better capture spatial variation when creating calibration algorithms, as this may also add significant value to calibrations. The creation of local calibrations may aid in the compilation of regional calibrations, where the continued compilation of data may add to accuracy and reliability of international algorithms which would be the ultimate goal.

As for the application of these models, specifically those created to compare with the OSSL models, the models perform too poorly for high-accuracy measurements. The catchment models, however, are accurate enough to be practically implemented for the use of irrigation management in agriculture and should provide sufficient accuracy.

NIRS certainly shows promise as a non-invasive, rapid and cost-effective method of soil moisture determination, but for its effective and accurate application, the collection of sufficient data on local scale is imperative.

CHAPTER 9: CONCLUSIONS AND RECOMMENDATIONS

9.1 CONCLUSIONS

The HYDROSOIL, a detailed hydrological soil map created by digital soil mapping methods, does usually improve the hydrological modelling of a catchment. An increase in soil information led to an increase in model accuracy in five of the catchments. In addition to the improved spatial distributions of soil, the increase in model accuracy is also dependent on correctly parameterising the soil mapping units. The uses of the HYDROSOIL were showcased by:

1. Using the HYDROSOIL map to calibrate a hydrological model by optimising model parameters based on the expected hydrological response in the Sabie catchment.
2. Determining the effect of pixel size on the model outcome in the Jukskei catchment.
3. Quantifying the effect of land-use change on the hydrological regime within the uMngeni catchment.
4. Modelling sediment yield in the Tsitsa catchment.
5. Using a different hydrological model in the Goukou catchment.

The Goukou catchment was the only catchment where the improved soil information had no effect on the accuracy of the hydrological modelling. Here, it was shown that accurate rainfall data is the primary determinant of model outcome. Improving the soil data will be secondary to ensuring accurate rainfall data is used.

Efforts to enable better parameterisation of soils included creating pedotransfer functions to predict hydrological soil properties using more easily measured soil properties, and creating calibration algorithms for Near Infrared Spectroscopy to measure hydrological soil properties. Neither of these efforts yielded acceptable results, probably due to too little data being collected to adequately represent the large variety of soils within the study sites. This strengthens the argument to further pursue efforts to easily determine the hydrological properties of soils, since the lack of data was caused by the difficulty and cost associated with collecting such data. Easier and cheaper methods are required to parameterise soils in the future.

Both efforts led to the same conclusions, 1) that local data is required to accurately predict hydrological soil properties, as internationally developed models will not be able to account for the local soil variability; 2) that predictions created for smaller areas are generally more accurate than predictions for larger areas, most probably due to having less soil variability; 3) That a lot more data is required to make accurate predictions of hydrological soil properties and 4) Sampling strategies to collect the required data should focus on smaller areas to produce useful prediction models, and over time sufficient data will be collected for predictions at a regional or national scale.

The title of this project is "*Towards a hydrological soil map of South Africa (HYDROSOIL) – developing a methodology and showcasing its uses".* The 'towards' indicates that this project should not be the final product but that the work started here should be continued until an accurate hydrological soil map has been created for South Africa. However, this project was created to learn lessons to apply on the future journey towards a hydrological soil map for South Africa. The following lessons were learnt:

1. A database structure and quality control measures were created whereby collected soil data of the future could be gathered and stored.
2. A method was developed whereby highly clustered data could be used to create an accurate soil map, without losing less-represented soil types.
3. How to digitise the approximately 200 000-300 000 soil observations recorded on paper copy maps stored at the Agricultural Research Council and use these in digital soil mapping to create the HYDROSOIL map.
4. A strategic approach to obtaining hydrological soil property data was determined.

9.2 RECOMMENDATIONS

Based on the lessons learnt during this project**, it is recommended that an adequately parameterised, accurate hydrological soil map (HYDROSOIL) for South Africa and its border catchment can and should be created.** The following steps should be taken to achieve this goal:

1. Improve the soil database to become a cloud-based soil data repository with automatic quality control.
2. Digitise the land type field observations to be used in digital soil mapping to create the national HYDROSOIL map, using the methods determined in this project.
3. Apply the digital soil mapping methods used and newly learnt in this report to create the HYDROSOIL map.
4. Characterise the hydrological properties of the soils of South Africa, using hydrological soil measurements, pedotransfer functions and near infrared spectroscopy. This should be done by collecting data from smaller areas and first make useful local predictions. When sufficient data has been collected for the entire country, national prediction models should be created.

REFERENCES

Abbaspour, K.C., Vaghefi, S.A., Yang, H. & Srinivasan, R. (2019). Global soil, landuse, evapotranspiration, historical and future weather databases for SWAT Applications. Scientific Data, 6(1), 1-11.

Abbaspour, K.C., Vejdani, M., Haghighat, S. (2007). SWAT-CUP calibration and uncertainty programs for SWAT. In Oxley, L. and Kulasiri, D. Eds., MODSIM 2007 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand, 1596-1602.1

Afzali, H., Tasumi, M. & Nishiwaki, A. (2021). Use of hand-held NIR sensor to estimate water status of leaves and soils. Journal of Rainwater Catchment Systems, 26(2), 1-6.

Ahl, R.S., Woods, S.W. & Zuuring, H.R. (2008). Hydrologic calibration and validation of SWAT in a snow-dominated rocky mountain watershed, Montana, USA. Journal of the American Water Resources Association, 44(6), 1411-1430.

Aouissi, J., Benabdallah, S., Lili Chabaâne, Z. & Cudennec, C. (2016). Evaluation of potential evapotranspiration assessment methods for hydrological modelling with SWAT – Application in data-scarce rural Tunisia. Agricultural Water Management, 174, 39-51. doi:10.1016/j.agwat.2016.03.004

ARC (2012). Land Types of South Africa: Maps (69 sheets) and Memoirs (39 books). Agricultural Research Council – Soil, Climate and Water, Pretoria.

ARC (2012). Agroclimatology Database, unpublished, Agricultural Research Council – Soil, Climate and Water, 600 Belvedere, Pretoria 0083, South Africa.

ARC-SWC Soil Database (2014). Agricultural Research Council – Soil Water and Climate. Pretoria.

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B. & Neitsch, S.L. (2012). Soil and Water Assessment Tool “SWAT” Input/Output Documentation. Version 2012. Texas Water Resources Institute, TR-439, College Station, 650. https://swat.tamu.edu/media/69296/swat-io-documentation-2012.pdf

Arnold, J.G., Youssef, M.A., Yen, H., White, M.J., Sheshukov, A.Y., Sadeghi, A.M., Moriasi, D.N., Steiner, J.L., Amatya, D.M., Skaggs, R.W., Haney, E.B., Jong, J., Arabi, M. & Gowda, P.H. (2015). Hydrological processes and model representation: Impact of soft data on calibration. Transactions of the ASABE, 58, 1637-1660. doi:10.13031/trans.58.10726.

Arnold, J.G., Kiniry, J.R., Srinivasan, R., Williams, J.R., Haney, E.B., Neitsch, S.L. (2011). Soil and Water Assessment Tool Input/Output File Documentation-Version 2009. Texas Water Resources Institute Technical Report 365.

Arnold, J.G., Srinivasan, R., Muttiah, R.S. & Williams, J.R. (1998). Large area hydrologic modelling and assessment, part I: Model development. Journal of the American Water Resources Association, 34(1), 73-89. doi:10.1111/j.1752-1688.1998.tb05961.x

Arrouays, D., McKenzie, N., de Forges, A.R., Hempel, J. & McBratney, A.B. (2014). GlobalSoilMap: Basis of the Global Spatial Soil Information System (1st ed.). CRC Press, Balkema, Leiden.

Ayana, E., Dile, Y., Narasimhan, B. & Srinivasan, R. (2019). Dividends in flow prediction improvement using high-resolution soil database. Journal of Hydrology: Regional Studies, 21(2019), 159-175.

Aydın, Y., Işıkdağ, Ü., Bekdaş, G., Nigdeli, S.M. & Geem, Z.W. (2023). Use of machine learning techniques in soil classification. Sustainability, 15(3), 2374. doi:10.3390/su15032374.

Bailey, N., Clements, T., Lee, J.T. & Thompson, S. (2003). Modelling soil series data to facilitate targeted habitat restoration: a polytomous logistic regression approach. Journal of Environmental Management, 67, 395-407.

Bannari, A., Morin, D., Bonn, F. & Huete, A.R. (1995). A review of vegetation indices. Remote Sensing Reviews, 13, 95-120. doi:10.1080/02757259509532298.

Batezini, R. & Balbo, J.T. (2015). Study on the hydraulic conductivity by constant and falling head methods for pervious concrete. Revista IBRACON de Estruturas e Materiais, 8(3), 248-259.

Batjes, N.H. (1995). A homogenized soil data file for global environmental research: a subset of FAO, ISRIC, and NRCS profiles (Version 1.0). Working Paper and Preprint 95/10b, International Soil Reference and Information Centre.

Batjes NH. (2008). ISRIC-WISE Harmonized Global Soil Profile Dataset (Ver. 3.1). Report 2002/02, ISRIC-World Soil Information, Wageningen.

Benke, K.K., Norng, S., Robinson, N.J., Chia, K., Rees, D.B. & Hopley, J. (2020). Development of pedotransfer functions by machine learning for prediction of soil electrical conductivity and organic carbon content. Geoderma, 366, 114210.

Benzaghta, M.A., Elwalda, A., Mousa, M.M., Erkan, I. & Rahman, M. (2021). SWOT analysis applications: An integrative literature review. Journal of Global Business Insights, 6(1), 55-73. doi:10.5038/2640-6489.6.1.1148.

Beven, K. & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology, 249, 11-29.

Beven, K.A. (2006). Manifesto for the equifinality thesis. Journal of Hydrology, 320, 18-36.

Bieger, K., Arnold, J.G., Rathjens, H., White, M.J., Bosch, D.D., Allen, P.M. & Srinivasan, R. (2017). Introduction to SWAT+, a completely restructured version of the Soil and Water Assessment Tool. Journal of the American Water Resources Association, 53(1), 115-130.

Blum, W.E.H., Warkentin, B.P. & Frossard, E. (2006). Soil, human society and the environment. Geological Society Publications, 266, 1-8. https://doi.org/10.1144/GSL.SP.2006.266.01.01

Bondi, G., Creamer, R., Ferrari, A., Fenton, O. & Wall, D. (2018). Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation. Geoderma, 318, 137-147. doi:10.1016/j.geoderma.2017.11.035.

Bossa, A.Y., Diekkrüger, B., Igué, A.M. & Gaiser, T. (2012). Analyzing the effects of different soil databases on modelling of hydrological processes and sediment yield in Benin (West Africa). Geoderma, 173, 61-74. doi:10.1016/j.geoderma.2012.01.012.

Bouma, J. (2016). Hydropedology and the societal challenge of realizing the 2015 United Nations Sustainable Development Goals. Vadose Zone Journal, 15. doi:10.2136/vzj2016.09.0080.

Bouma, J., Bonfante, A., Basile, A., van Tol, J., Hack-ten Broeke, M.J.D., Mulder, M., Heinen, M., Rossiter, D.G., Poggio, L. & Hirmas, D.R. (2022). How can pedology and soil classification contribute towards sustainable development as a data source and information carrier? Geoderma, 424, 115988. doi:10.1016/j.geoderma.2022.115988.

Bouma, J., Droogers, P., Sonneveld, M. P. W., Ritsema, C. J., Hunink, J. E., Immerzeel, W. W. & Kauffman, S. (2011). Hydropedological insights when considering catchment classification. Hydrology and Earth System Sciences, 15, 1909-1919. doi:10.5194/hess-15-1909-2011.

Bouma, J., Pinto-Correia, T. & Veerman, C. (2021). Assessing the role of soils when developing sustainable agricultural production systems focused on achieving the UN-SDGs and the EU Green Deal. Soil Systems, 5, 56. doi:10.3390/soilsystems5030056.

Bouma, J., Stoorvogel, J., van Alphen, B.J. & Booltink, H.W.G. (1999). Pedology, precision agriculture and the changing paradigm of agricultural research. Soil Science Society of America Journal, 63, 1763-1768.

Bouma, J. & van Lanen, H. A. J. (1987). Transfer functions and threshold values: from soil characteristics to land qualities. In K. J. Beek, P. A. Burrough & D. E. MacCormack (Eds.), Proceedings of the international workshop on Quantified land evaluation procedures : held in Washington, DC, 27 April-2 May 1986 (pp. 106-110)

Bouslihim, Y., Rochdi, A., El Amrani Paaza, N. & Liuzzo, L. (2019). Understanding the effects of soil data quality on SWAT model performance and hydrological processes in Tamedroust Watershed (Morocco). Journal of African Earth Sciences, 160, Article 103616. doi:10.1016/j.jafrearsci.2019.103616.

Bouwer, D., Van Zijl, G.M., Van Tol, J.J., Le Roux, P.A.L., Hydropedological Report of Constantia Kloof. Report created for the Johannesburg Roads Agency.

Breure, A.M., De Deyn, G.B., Dominati, E., Englin, T., Hedlund K., Van Orshoven, J. & Posthuma, L. (2012). Ecosystem services: a useful concept for soil policy making. Current Opinion in Environmental Sustainability, 4(5), 578-585.

Bryant, R.B., Gburek, W.J., Veith, T.L. & Hively, W.D. (2006). Perspectives on the potential for hydropedology to improve watershed modelling of phosphorus loss. Geoderma, 131, 299-307.

Bullock, P., Li, X. & Leonardi, L. (2004). Near-infrared spectroscopy for soil water determination in small soil volumes. Canadian Journal of Soil Science, 84(3), 333-338.

Campling, P., Gobin, A. & Feyen, J. (2002). Logistic modelling to spatially predict the probability of soil drainage classes. Soil Science Society of America Journal, 66, 1390-1401.

Canal Filho, R., Molin, J.P., Wei, M.C. & Silva, E.R. (2023). Soil attributes mapping with online near-infrared spectroscopy requires spatio-temporal local calibrations. AgriEngineering, 5(3), 1163-1177.

Carter, J.R. (1988). Digital representations of topographic surfaces. Photogrammetric Engineering and Remote Sensing, 54(11), 1577-1580.

Chai, T. & Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Geoscientific Model Development Discussions, 7(1), 1525-1534.

Chaplin, M. (2001). Water: its importance to life. Biochemistry and Molecular Biology Education, 29(2), 54-59.

Chawla, N.V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

Chen, E. & Mackay, D.S. (2004). Effects of distribution-based parameter aggregation on a spatially distributed agricultural nonpoint source pollution model. Journal of Hydrology, 295, 211-224.

Chen, L., Wang, G., Zhong, Y. & Shen, Z. (2016). Evaluating the impacts of soil data on hydrological and nonpoint source pollution prediction. Science of The Total Environment, 563-564, 19-28. doi:10.1016/j.scitotenv.2016.04.107.

Chen, Y., Li, L., Whiting, M., Chen, F., Sun, Z., Song, K. & Wang, Q. (2021). Convolutional neural network model for soil moisture prediction and its transferability analysis based on laboratory vis-nir spectral data. International Journal of Applied Earth Observation and Geoinformation. 104:1-9.

Cisty, M., Bezak, J. & Skalova, J. (2012). Pedotransfer Functions Development by means of the Ensemble Data-Driven Methodology. Proceedings of the Eighth International Conference on Engineering Computational Technology.

Clingensmith, C. & Grunwald, S. (2022). Predicting soil properties and interpreting Vis-NIR models from across Continental United States. Sensors, 22(9), 3187. doi:10.3390/s22093187.

Collet, A. & Rozanov, A. (2018). SA-EU Dialogue on Soil Information Report on the study tour to Europe (Italy; Germany; France).

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V. & Böhner, J. (2015). System for automated geoscientific analysis (SAGA). In: Geoscientific Model Development. doi:10.5194/gmd-8-1991-2015.

Cook, S.E., Jarvis, A. & Gonzalez, J.P. (2008). A new global demand for digital soil information. In: Hartemink, A.E., McBratney, A., Mendonça-Santos, M.d. (eds) Digital soil mapping with limited data. Springer, Dordrecht. doi:10.1007/978-1-4020-8592-5_3.

Cosenza, D., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J., Næsset, E., Gobakken, T., Soares, P. & Tomé, M. (2020). Comparison of linear regression, K-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. Forestry: An International Journal of Forest Research, 94(2), 311-323.

Council for Geoscience. (2007). Geological Data 1:250 000. Pretoria, South Africa: Council for Geoscience.

Dabrowski, JM. & De Klerk, LP. (2013). An assessment of the impact of different land use activities on water quality in the upper Olifants River catchment. Water SA, 39(2), 231-244. http://dx.doi.org/10.4314/wsa.v39i2.6Dangal, S., Sanderman, J., Wills, S. & Ramirez-Lopez, L. (2019). Accurate and precise prediction of soil properties from a large mid-infrared spectral library. Soil Systems, 3(1), 11. doi:10.3390/soilsystems3010011.

Davari, M., Karimi, S., Bahrami, H., Taher Hossaini, S. & Fahmideh, S. (2021). Simultaneous prediction of several soil properties related to engineering uses based on laboratory vis-nir reflectance spectroscopy. CATENA, 197, 1-12.

De Clercq, W.P., De Witt, M., Watson, A., Helness, H. & Daman, S. (2023). Evidence-based assessment of NWRM for sustainable water management “EviBAN”. WRC Report No. 3084/1/23.

Debella-Gilo, M. & Etzelmüller, B. (2009). Spatial prediction of soil classes using digital terrain analysis and multinomial logistic regression modelling integrated in GIS: examples from Vestfold County, Norway. CATENA, 77, 8-18.

Devia, G.K., Ganasri, B.P. & Dwarakish, G.S. (2015). A review on hydrological models. Aquatic Procedia, 4, 1001-1007.

Dewitte, O., Jones, A., Spaargaren, O., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Micheli, E., Montanarella, L., Thiombiano, L., Van Ranst, E., Yemefack, M. & Zougmore, R. (2013). Harmonisation of the soil map of Africa at the continental scale. Geoderma, 211-212, 138-153. doi:10.1016/j.geoderma.2013.07.007.

Diek, S., Temme, A.J.A.M. & Teuling, A.J. (2014). The effect of spatial soil variation on the hydrology of a semi-arid Rocky Mountain catchment. Geoderma, 235, 113-126. doi:10.1016/j.geoderma.2014.06.028.

Dippenaar, M.A. & Van Rooy, J.L. (2014). Review of engineering, hydrogeological and vadose zone hydrological aspects of the Lanseria Gneiss, Goudplaats-Hout River Gneiss and Nelspruit Suite Granite (South Africa). Journal of African Earth Sciences,91, 12-31.

Djodjic, F., Bieroza M. & Bergström L. (2021). Land use, geology and soil properties control nutrient concentrations in headwater streams. Science of the Total Environment 772.145108.

Du Plessis, C., van Zijl, G., van Tol, J. & Manyevere, A. (2020). Machine learning digital soil mapping to inform gully erosion mitigation measures in the Eastern Cape, South Africa. Geoderma, 368. doi:10.1016/j.geoderma.2020.114287.

Eckhardt, K. (2005). How to construct recursive digital filters for base-flow separation. Hydrological Processes, 19(2), 507-515.

Essenfelder, A.H. (2016). SWAT Weather Database: A Quick Guide. Version: v.0.16.06. doi:10.13140/RG.2.1.4329.1927.

Fernández, R.N. & Rusinkiewicz, M. (1993). A conceptual design of a soil database for a geographical information system. International Journal of Geographic Information Systems, 7(6), 525-539.

Filzmoser, P., Garrett RG. & Reimannn C. (2005). Multivariate outlier detection in exploration geochemistry. Computer & Geosciences, 31, 579-587.

Flynn, T., de Clercq, W., Rozanov, A. & Clarke, C. (2019a). High-resolution digital soil mapping of multiple soil properties: an alternative to the traditional field survey? South African Journal of Plant and Soil, 36(4), 237-247.

Flynn, T., Van Zijl, G.M., Van Tol, J.J., Botha, C., Rozanov, A., Warr, B. & Clarke, C. (2019b). Comparing algorithms to disaggregate complex soil polygons in contrasting environments. Geoderma, 352, 171-180. doi:10.1016/j.geoderma.2019.06.013.

Gagkas, Z., Lilly, A. & Baggaley, N.J. (2021). Digital soil maps can perform as well as large-scale conventional soil maps for the prediction of catchment baseflows. Geoderma, 400, 115230. doi:10.1016/j.geoderma.2021.115230 .

García, S. & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: proposals and taxonomy. Evolutionary Computation, 17, 275-306. doi:10.1162/evco.2009.17.3.275.

Gassman, P.W., Reyes, M.R., Green, C.H. & Arnold, J.G. (2007). The Soil and Water Assessment Tool: historical development, applications, and future research directions. Transactions of the ASABE, 50(4), 1211-1250. doi:10.13031/2013.23637

Gassman, P.W., Sadeghi, A.M. & Srinivasan, R. (2014). Applications of the SWAT model special section: overview and insights. Journal of Environmental Quality, 43, 1-8. doi:10.2134/jeq2013.11.0466

GeoTerra Image (2015). 2013-2014 South African National Land Cover Dataset; Report Created for Department of Environmental Sciences; DEA/CARDNO SCPF002: Implementation of Land Use Maps for South Africa; Department of Environmental Affairs: Pretoria, South Africa.

Geza, M. & McCray, J.E. (2008). Effects of soil data resolution on SWAT model stream flow and water quality predictions. Journal of Environmental Management, 88, 393-406. doi:10.1016/j.jenvman.2007.03.016

Glenday, J., Gokool, S., Gwapedza, D., Holden, P., Rebelo, A., Tanner, J., Jumbi, F. & Metho, P. (2021). Critical catchment model inter-comparison and model use guidance development. WRC report K5/2927. Water Research Commission: Pretoria, South Africa

Grealish, G., King, P., Omar, S. & Roy, W. (2004). Geographic information system and database for the soil survey for the State of Kuwait-design and outputs. Kuwait Journal of Science and Engineering, 31(1), 135-148

Green, W.H. & Ampt, G.A. (1911). Studies on soil physics, I. Flow of air and water through soils. Journal of Agricultural Science, 4, 11-24. doi:10.1017/S0021859600001441

Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modelling approaches. Geoderma, 152(3-4), 195-207. doi:10.1016/j.geoderma.2009.06.003

Gunarathna, M.H.J.P., Sakai, K., Nakandakari, T., Momii, K. & Kumari, M.K.N. (2019). Machine learning approaches to develop pedotransfer functions for tropical Sri Lankan soils. Water, 11(9), 1940. doi:10.3390/w11091940

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, 377, 80-91. doi:10.1016/j.jhydrol.2009.08.003.

Gupta, R. (2000). SWOT analysis of geographic information: The case of India. Current Science, 79(4), 489-498. doi:10.1019/j.currentsci.2000.07.001

Guzha, A.C., Rufino, M.C., Okoth, S., Jacobs, S. & Nóbrega, R.L.B. (2018). Impacts of land use and land cover change on surface runoff, discharge, and low flows: Evidence from East Africa. Journal of Hydrology: Regional Studies, 15, 49-67.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73, 220-239.

Hanjra, M. & Qureshi, M. (2010). Global water crisis and future food security in an era of climate change. Food Policy, 35(5), 365-377.

Häring, T., Dietz, E., Osenstetter, S., Koschitzki, T. & Schröder, B. (2012). Spatial disaggregation of complex soil map units: A decision-tree based approach in Bavarian forest soils. Geoderma, 185-186, 37-47.

Harrison, R.L., Van Tol, J.J. & Toucher, M.L. (2022). Using hydropedological characteristics to improve modelling accuracy in Afromontane catchments. Journal of Hydrology: Regional Studies, 39, 100986. doi:10.1016/j.ejrh.2021.100986

Hartemink, A.E. (2007). Soil fertility decline: definitions and assessment. ISRIC-World Soil Information, 1618-1621.

Hartigan, J.A. & Wong, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm. Applied Statistics, Royal Statistical Society, 100-108.

Hazelton, P. & Murphy, B. (2007). Interpreting soil test results: What do all the numbers mean? Melbourne: CSIRO Publishing.

He, H., Bai, Y., Garcia, E. & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. Proceedings of IJCNN 2008 (IEEE World Congress on Computational Intelligence), IEEE International Joint Conference, 1322-1328.

He, H. & Garcia, E. A. (2008). Learning from imbalanced data. IEEE Transactions on Knowledge & Data Engineering, 21, 1263-1284.

Hengl, T., De Jesus, J.M., Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S. & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PloS ONE, 12(2), 1-40.

Hengl, T., Toomanian, N., Reuter, H.I. & Malakouti, M.J. (2007). Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. Geoderma, 140, 417-427.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E. & Schmidt, M. G. (2016). An overview and comparison of machine learning techniques for classification purposes in digital soil mapping. Geoderma, 265, 62-77.

Hillel, D. (2003). Principles and processes of environmental soil physics: The state and the transport of matter and energy in the soil/plant atmosphere. Oxford: Academic.

Hobbs, P., Oelofse, S.H.H. & Rascher, J. (2008). Management of environmental impacts from coal mining in the upper Olifants river catchment as a function of age and scale. International Journal of Water Resources Development, 24, 417-431. https://doi.org/10.1080/07900620802127366

Hoffmann, C., Schulz, S., Eberhardt, E., Grosse, M., Stein, S., Specka, X., Svoboda, N. & Heinrich, U. (2020). Data standards for soil-and agricultural research. Report number: BonaRes Series 2019/6. doi:10.20387/BonaRes-ARM4-66M2

Hortensius D. & Norcliff S. (1991). International standardization of soil quality measurement procedures for the purpose of soil protection. Soil Use and Management, 7(3), 163-166.

Hortensius, D. & Welling, R. (2008). International standardization of soil quality measurements. Communications in Soil Sciences and Plant Analysis, 27(3-4), 387-402. doi:10.1080/00103629609369563

Hutson, J.L. (1983). Estimation of hydrological properties of South African soils. University of KwaZulu-Natal. https://researchspace.ukzn.ac.za/xmlui/handle/10413/11019

Idowu O.A. Lorentz S.A. Annandale J.G. Aken M. McCartney M.P. Thornton-Dibb SLC. Westhuizen A. (2010). Comparative assessment of widespread irrigation with low quality mine-water in undisturbed and rehabilitated mine-lands in the upper Olifants using the ACRU200 model. *Water SA* 36 (5): 543-552. https://hdl.handle.net/10520/EJC116745

IUSS Working Group WRB (2015). World Reference Base for Soil Resources 2014, Update 2015 International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. World Soil Resources Reports No. 106. FAO, Rome.

Iqbal J., Thomasson JA., Jenkins JN., Owens PR. & Whisler, FD. (2005). Spatial variability analysis of soil physical properties of alluvial soils. Soil science society of American journal 69(4): 1338-1350. https://doi.org/10.2136/sssaj2004.0154

Jahn, R., Blume, H.P., Asio, V.B., Spaargaren, O. & Schad, P. (2006). Guidelines for soil description. 4th edition. Rome, Italy: FAO.

Jie, C., Jing-zang, C., Man-zhi, T. & Zi-tong, G. (2002). Soil degradation: A global problem endangering sustainable development. Journal of Geographical Sciences, 12, 243-252.

Julich, S., Breuer, L. & Frede, H.-G. (2012). Integrating heterogeneous landscape characteristics into watershed-scale modelling. Advances in Geosciences, 31, 31-38. doi:10.5194/adgeo-31-31-2012

Kahmen, A., Perner, J. & Buchmann, N. (2005). Diversity-dependent productivity in semi-natural grasslands following climate perturbations. Functional Ecology, 19, 594-601.

Katuwal, S., Knadel, M., Norgaard, T., Moldrup, P., Greve, M. & de Jonge, L. (2020). Predicting the dry bulk density of soils across Denmark: Comparison of single-parameter, multi-parameter, and vis-NIR based models. Geoderma, 361, 1-10. doi:10.1016/j.geoderma.2019.114078.

Kempen, B., Brus, D.J., Heuvelink, G.B.M. & Stoorvogel, J.J. (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. Geoderma, 151, 311-326.

Kikkas, K. & Kulik, S. (2018). Modelling the effect of human activity on freshwater extraction from the earth's reserves. IOP Conference Series: Earth and Environmental Science, 180(2018), 12-17.

Kirchner, J.W. (2006). Getting the right answer for the right reason: Linking measurements, analysis, and models to advance the science of hydrology. Water Resources Research, 42, W03S04. doi:10.1029/2005WR004362.

Kirkby, M.J., Bull, L.J., Poesen, J., Nachtergaele, J. & Vandekerckhove, L. (2003). Observed and modelled distributions of channel and gully heads—with examples from SE Spain and Belgium. Catena, 50(2-4), 415-434.

Klopp, H., Arriaga, F., Daigh, A. & Bleam, W. (2020). Analysis of pedotransfer functions to predict the effects of salinity and sodicity on saturated hydraulic conductivity of soils. Geoderma, 362, 114078. doi:10.1016/j.geoderma.2019.114078 .

Klute, A. & Dirksen, C. (1986). Hydraulic conductivity and diffusivity: Laboratory methods. In: Klute, A. (Ed.), Methods of Soil Analysis. Part 1: Physical and Mineralogical Methods, 2nd Edition, Agronomy Monograph No. 9, ASA, Madison, 687-734. doi:10.2136/sssabookser5.1.2ed.c28

Knadel, M., Gislum, R., Hermansen, C., Peng, Y., Moldrup, P., de Jonge, L. & Greve, M. 2017. Comparing predictive ability of laser-induced breakdown spectroscopy to visible near-infrared spectroscopy for soil property determination. *Biosystems Engineering*. 156:157-172.

Knoben, W.J.M., Freer, J.E. & Woods, R.A. (2019). Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrology and Earth System Sciences, 23, 4323-4331. doi:10.5194/hess-23-4323-2019

Knox, N., Grunwald, S., McDowell, M., Bruland, G., Myers, D. & Harris, W. (2015). Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy. Geoderma, 239-240, 229-239.

Koirala, B., Zahiri, Z. & Scheunders, P. (2022). A robust supervised method for estimating soil moisture content from spectral reflectance. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-13. doi:10.1109/TGRS.2021.3109277

Kopittke, P.M., Menzies, N.W., Wang, P., McKenna, B.A. & Lombi, E. (2019). Soil and intensification of agriculture for global food security. Environmental International, 136, 1-8. doi:10.1016/j.envint.2019.105078

Kruger, A.C., Makamo, L.B. & Shongwe, S. (2002). An analysis of Skukuza climate data. Koedoe, 45(1), 87-92. doi:10.4102/koedoe.v45i1.16

Kutner, M.H., Nachtsheim, C.J., Neter, J. & Li, W. (2005). Applied linear statistical models (5th ed.). New York: McGraw-Hill/Irwin

Lal, R. (2015). Restoring soil quality to mitigate soil degradation. Sustainability, 7(5), 5875-5895. doi:10.3390/su7055875

Lal, R., Bouma, J., Brevik, E., Dawson, L., Field, D.J., Glaser, B., et al. (2021). Soils and Sustainable Development Goals of the United Nations: An IUSS Perspective. Geoderma Regional, 25, e000398. doi:10.1016/j.geodrs.2021.e000398.

Lamichhane, S., Kumar, L. & Adhikari, K. (2021). Updating the national soil map of Nepal through digital soil mapping. Geoderma, 394, 115041.

Lamorski, K., Pachepsky, Y., Sławiński, C. & Walczak, R. (2008). Using support vector machines to develop pedotransfer functions for water retention of soils in Poland. Soil Science Society of America Journal, 72(5), 1243-1247. doi:10.2136/sssaj2007.0280n

Land Type Survey Staff (1972-2002). Land Types of South Africa: Digital Map (1:250,000 Scale) and Soil Inventory Datasets. ARC-Institute for Soil, Climate and Water: Pretoria, South Africa.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Leigh D. (2010). SWOT Analysis. In: Handbook of Improving Performance in the Workplace, vol. 2, Selecting and implementing performance interventions. Pepperdine University. pp 115-140.

Le Roux, J.J., Morgenthal, T.L., Malherbe, J., Sumner, P.D. & Pretorius, D.J. (2008). Water erosion prediction at a national scale for South Africa. Water SA, 34(3), 305-314.

Le Roux, J.J., Mararakaney, N., Mudaly, L., Weepener, M. & van der Laan, M. (2023). Development of a South African national database to run the SWAT model in a GIS. WRC Report No. 3053/1/22. Water Research Commission, Pretoria.

Le Roux, J.J., Mararakanye, N., Mudaly, L., Weepener, H.L. & van der Laan, M. (2022). Development of a South African national input database to run the SWAT model in a GIS. WRC report in press. Water Research Commission: Pretoria, South Africa.

Leenars, J.G.B. (2013). African Soil Profiles Database Version 1.1. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). ISRIC Report 2013/03, 1-160.

Li, Y., Adams, N. & Bellotti, T. (2022). A relabelling approach to handling the class imbalance problem for logistic regression. Journal of Computational and Graphical Statistics, 31(1), 241-253. doi:10.1080/10618600.2021.1978470

Li, Y., Chen, D., White, R.E., Zhu, A. & Zhang, J. (2007). Estimating soil hydraulic properties of Fengqiu County soils in the North China Plain using pedo-transfer functions. Geoderma, 138(3-4), 261-271. doi:10.1016/j.geoderma.2007.04.003

Liang, X., Li, X. & Lei, T. (2012) Paper delivered at the International conference on Systems and Informatics (ICSAI 2012), Yantai. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6223659 Date of access: 6 Oct. 2023.

Lin, H., O'Geen, A. T., Zhang, R. & Horwath, W. R. (2005). Spatial variability of soil hydraulic properties in a paddy field. Geoderma, 124(3-4), 331-344. doi:10.1016/j.geoderma.2004.05.008

Lin, H., Thompson, J. A. & Green, R. E. (2006). Spatial scaling of soil hydraulic properties using geostatistics and neural networks. Soil Science Society of America Journal, 70(5), 1705-1715. doi:10.2136/sssaj2005.0369

Lin, H.S. (2003). Hydropedology: Bridging disciplines, scales, and data. Vadose Zone Journal, 2, 1-11. doi:10.2136/vzj2003.1000

Lin, H.S., Kogelman, W., Walker, C. & Bruns, M.A. (2006). Soil moisture patterns in a forested catchment: A hydropedological perspective. Geoderma, 131, 345-368. doi:10.1016/j.geoderma.2005.03.013

López, V., Fernandez, A., García, S., Palade, V. & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences, 250, 113-141.

Ma, Y.X., Minasny, B., Malone, B.P. & McBratney, A.B. (2019). Pedology and digital soil mapping (DSM). European Journal of Soil Science, 70, 216-235.

MacMillan, R.A., Moon, D.E., Coupé, R.A. & Phillips, N. (2010). Predictive ecosystem mapping (PEM) for 8.2 million ha of forestland, British Columbia, Canada. In J.L. Boettinger, D.W. Howell, A.C. Moore, A.E. Hartemink & S. Kienast Brown (Eds.), Digital Soil Mapping: Bridging Research, Environmental Application, and Operation. Springer.

MacVicar, C.N., de Villiers, J.M., Loxton, R.F., Verster, E., Lambrechts, J.J.N., Merryweather, F.R., Le Roux, J., van Rooyen, T.H., Harmse, H.J. von M. (1977). Soil classification: a binomial system for South Africa. Pretoria: Department of Agriculture Technical Services.

Martinec, J. & Rango, A. (1989). Merits of statistical criteria for the performance of hydrological models. Water Resources Bulletin, 25, 421-432.

McBratney, A.B., Minasny, B., Cattle, S.R. & Vervoort, W.R. (2002). From pedotransfer functions to soil inference systems. Geoderma, 109(1-2), 41-73. doi:10.1016/s0016-7061(02)00139-8.

McBratney, A.B., Mendoça Santos, M.L. & Minasny, B. (2003). On digital soil mapping. Geoderma, 117, 3-52. doi:10.1016/S0016-7061(03)00223-4

Me, W., Abell, J.M. & Hamilton, D.P. (2015). Effects of hydrologic conditions on SWAT model performance and parameter sensitivity for a small, mixed land use catchment in New Zealand. Hydrology and Earth System Sciences, 19, 4127-4147. doi:10.5194/hess-19-4127-2015

Mengistu, A.G., van Rensburg, L.D. & Woyessa, Y.E. (2019). Techniques for calibration and validation of SWAT model in data scarce arid and semi-arid catchments in South Africa. Journal of Hydrology: Regional Studies, 25, 100621.

Minasny, B. & McBratney, A. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32(9), 1378-1388.

Minasny, B. & McBratney, A.B. (2015). Digital soil mapping: A brief history and some lessons. Geoderma, 1-11. doi:10.1016/j.geoderma.2013.12.014

Mishra, A., Froebrich, J. & Gassman, P.W. (2007). Evaluation of the SWAT model for assessing sediment control structures in a small watershed in India. Transactions of the ASABE, 50(2), 469-478.

Miti, C., Mbanyele, V., Mtangadura, T., Magwero, N., Namaona, W., Njira, K., Sandram, I., Lubinga, P.N., Chisanga, C.B., Nalivata, P.C., Chimungu, J.G., Nezomba, H., Phiri, E. & Lark, R.M. (2023). The appraisal of pedotransfer functions with legacy data; an example from southern Africa. Geoderma, 439, 116661. doi:10.1016/j.geoderma.2023.116661

Monteith, J.L. (1965). Evaporation and environment. In G.E. Fogg (Ed.), Symposium of the Society for Experimental Biology, The State and Movement of Water in Living Organisms (Vol. 19, pp. 205-234). Academic Press, Inc., New York.

Montgomery, D.C., Peck, E.A. & Vining, G.G. (2012). Introduction to Linear Regression Analysis. John Wiley & Sons.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D. & Veith, T.L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900. doi:10.13031/2013.23153

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. Transactions of the ASABE, 58(6), 1763-1785. doi:10.13031/trans.58.10715.

Mouazen, A. & Al-Asadi, R. (2018). Influence of soil moisture content on assessment of bulk density with combined frequency domain reflectometry and visible and near infrared spectroscopy under semi field conditions. Soil and Tillage Research, 176, 95-103.

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media. Water Resources Research, 12(3), 513-522. doi:10.1029/WR012i003p00513.

Mucina, L. & Rutherford, M.C. (2006). The vegetation of South Africa, Lesotho and Swaziland. Strelitzia 19, South African National Biodiversity Institute, Pretoria.

Mulder, V.L., De Bruin, S., Schaepman, M.E. & Mayr, T.R. (2011). The use of remote sensing in soil and terrain mapping – A review. Geoderma, 162(1-2), 1-9. doi:10.1016/j.geoderma.2010.12.018.

Myeni, L., Mdlambuzi, T., Paterson, D.G., De Nysschen, G. & Moeletsi, M.E. (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), p.2639. doi:10.3390/w13192639.

Nachtergaele, F., Van Velthuizen, H., Verelst, L., Batjes, N., Dijkshorn, K., Van Engelen, V., Fische, G., Jones, A., Montanarella, L., Petri, M., Prieller, S., Shi, X., Teixeira, E. & Wiber, D. (2010). The harmonised world soil data. World Congress of Soil Science, Soil Solutions for a Changing World, 34-37.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. Biometrika, 78(3), 691-692. doi:10.1093/biomet/78.3.691.

Nash, J.E. & Sutcliffe, J.V. (1970). River flow forecasting through conceptual models, Part I: A discussion of principles. Journal of Hydrology, 10, 282-290. doi:10.1016/0022-1694(70)90255-690255-6).

Neitsch, S.L., Williams, J., Arnold, J. & Kiniry, J. (2009). Soil and Water Assessment Tool Theoretical Documentation Version 2009. Texas Water Resources Institute: College Station, TX, USA.

Neitsch, S.L., Williams, J., Arnold, J. & Kiniry, J. (2011). Soil and Water Assessment Tool Theoretical Documentation Version 2009. Texas Water Resources Institute: College Station, TX, USA.

Nemes, A. & Schaap, M.G. (2006). Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of Hydrology, 329(1-2), 85-104.

Nguyen, T.V., Dietrich, J., Dang, D.T., Tran, D.A., Doan, B.V., Sarrazin, F.J., Abbaspour, K. & Srinivasan, R. (2022). An interactive graphical interface tool for parameter calibration, sensitivity analysis, uncertainty analysis, and visualization for the Soil and Water Assessment Tool. Environmental Modelling & Software, 156, 105497. doi:10.1016/j.envsoft.2022.105497

Nikolaou, I.E. & Evangelinos, K.I. (2010). A SWOT analysis of environmental management practices in Greek Mining and Mineral Industry. Resources Policy, 35, 226-234.

Nzonda, G. (2016). Characterising historical land cover change, and understanding trends in the Goukou catchment, Western Cape, South Africa. (Thesis – MSc). University of KwaZulu-Natal.

Olabanji M.F., Ndarana T., Davis N. & Archer E. (2020). Climate change impact on water availability in the olifants catchment (South Africa) with potential adaptation strategies. Physics and Chemistry of the Earth, Parts A/B/C, 120 (5), 1-10. http://dx.doi.org/10.1016/j.pce.2020.102939

Oldeman, L.R. (1992). Global Extent of Soil Degradation. In Bi-Annual Report 1991-1992/ISRIC, pp. 19-36.

Packer, I., Chapman, G. & Lawrie, J. (2019). On-ground extension of soil information to improve land management. Soil Use and Management. 35(1):75-84.

Padarian, J., Minasny, B. & McBratney, A.B. (2015). Using Google's cloud-based platform for digital soil mapping. Computer & Geosciences, 83, 80-88. doi: 10.1016/j.cageo.2015.06.023

Padarian, J., Minasny, B. & McBratney, A.B. (2020). Machine learning and soil sciences: a review aided by machine learning tools. SOIL, 6, 35-52. doi: 10.5194/soil-6-35-2020

Pangos, P., Jones, A., Basco, C. & Kumar, S. (2011). European digital archive on soil maps (EuDASM): preserving important soil data for public free access. International Journal of Digital Earth, 4(5), 434-443.

Park, S.J., McSweeney, K. & Lowery, B. (2001). Identification of the spatial distribution of soils using a process-based terrain characterization. Geoderma, 103, 249-272.

Paterson, G., Turner, D., Wiese, L., Van Zijl, G., Clarke, C. & Van Tol, J. (2015). Spatial soil information in South Africa: Situational analysis, limitations and challenges. South African Journal of Science, 111(56), 1-7. http://dx.doi.org/10.17159/sajs.2015/20140178

Pedescoll, A., Samsó, R., Romero, E., Puigagut, J. & García, J. (2011). Reliability, repeatability and accuracy of the falling head method for hydraulic conductivity measurements under laboratory conditions. Ecological Engineering, 37(5), 754-757.

Peng, L., Cheng-zhi, Q., A-xing, Z., Zhi-wei, H., Nai-qing, F. & Yi-jie, W. (2020). A case-based method of selecting covariates for digital soil mapping. Journal of Integrative Agriculture, 19(8), 2127-2136.

Phinzi, K., Ngetar, N. & Ebhuoma, O. (2020). Soil erosion risk assessment in the Umzintlava catchment (T32E), Eastern Cape, South Africa, using RUSLE and random forest algorithm. *South African Geographical Journal.* 103(2):139-162

Pietersen, K., Beekman, H.E. & Holland, M. (2012). South African groundwater governance case study. Report prepared for the World Bank in partnership with the South African Department of Water Affairs and the Water Research Commission. WRC report no. KV 273/11. Pretoria: Water Research Commission; 2011.

Pike, A. & Schulze, R. (1995). AUTOSOILS: A program to convert ISCW soils attributes to variables usable in hydrological models. Department of Agricultural Engineering, University of Natal, Pietermaritzburg, South Africa.

Placidi, P., Morbidelli, R., Fortunati, D., Papini, N., Gobbi, F. & Scorzoni, A. (2021). Monitoring soil and ambient parameters in the IoT precision agriculture scenario: An original modelling approach dedicated to low-cost soil water content sensors. Sensors, 21, 1-28.

Pozza, L.E. & Field, D.J. (2020). The science of soil security and food security. Soil Security, 1, 100002. doi:10.1016/j.soisec.2020.100002.

Quevauviller, P. (1998). Operationally defined extraction procedures for soil and sediment analysis I. Standardization. Trends in Analytical Chemistry, 17(5), 289-297.

R Core Team (2022). R: A language and environment for statistical computing. https://www.R-project.org.

Rapanyane, M. & Ngoepe, C. (2019). The impact of illicit financial flows on the South African political economy under Jacob Zuma, 2009-2018. Journal of Public Affairs, 20(2), 1-7.

Rawls & Brakensiek (1985). A PTF for predicting soil water retention parameters from soil texture and organic matter content. doi:10.2136/sssaj1985.03615995004900050015x

Rawls, W.J., Brakensiek, D.L. & Saxton, K.E. (1982). Estimation of soil water properties. Transactions of the ASAE, 25(5), 1316-1320.

Rawls, W.J., Pachepsky, Y.A., Ritchie, J.C., Sobecki, T.M. & Bloodworth, H. (2003). Effect of soil organic carbon on soil water retention. Geoderma, 116(1-2), 61-76.

Ray, S.S., Singh, J.P., Das, G. & Panigrahy, S. (2004). Use of high-resolution remote sensing data for generating site-specific soil management plans. In: The International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences.

Reddy, N.N. & Das, B.S. (2023). Digital soil mapping of key secondary soil properties using pedotransfer functions and Indian legacy soil data. Geoderma, 429, 1-15. doi:10.1016/j.geoderma.2023.115682.

Ribeiro, E., Batjes, N.H. & van Oostrum, A.J.M. (2020). World Soil Information Service (WoSIS) – Towards the standardization and harmonization of world soil profile data. Procedure manual 2020, Report 2020/01, ISRIC-World Soil Information. doi:10.17027/isric-wdc-2020-01.

Ribeiro, E., Batjes, N.H., Leenaars, J.G.B., van Oostrum A.J.M. & Mendes de Jesus, J. (2015). Towards the standardization and harmonization of world soil data: Procedures manual ISRIC World Soil Information Services (WoSIS version 2.0), Report 2015/03, ISRIC-World Soil Information.

Riddell, E. S., Nel, J., Van Tol, J., Fundisi, D., Jumbi, F. & Van Niekerk, A. (2020). Groundwater-surface water interactions in an ephemeral savanna catchment, Kruger National Park. Koedoe, 62(2), a1583. doi:10.4102/koedoe.v62i2.1583

Romano, N. & Palladino, M. (2002). Prediction of soil water retention using soil physical data and terrain attributes. Journal of Hydrology, 265(1-4), 56-75. doi:10.1016/s0022-1694(02)00094-x00094-x

Romanowicz, A.A., Vanclooster, M., Rounsevelb, M., La Junesseb, I. (2005). Sensitivity of the SWAT model to the soil and land use data parametrization: A case study in the Thyle catchment, Belgium. Ecological Modelling, 187, 27-39. doi:10.1016/j.ecolmodel.2005.01.025

Royal HaskoningDHV (2018). Goukou Estuarine Management Plan. Document title and version: Goukou River Estuarine Management Plan Western Cape Estuary Management Framework and Implementation Strategy.

Rozanov, A., Collett, A., Mamphol, R. & Paterson, G. (2023). Soil Information and soil security in South Africa. Stellenbosch University.

Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D. & Co-authors (2015). The NCEP Climate Forecast System Reanalysis. Bulletin of the American Meteorological Society, 91, 1015-1057.

SANBI (2012). The Vegetation Map of South Africa, Lesotho and Swaziland. South African National Biodiversity Institute. Online. http://bgis.sanbi.org/SpatialDataset/Detail/18 Version, 2012.

SANBI (2018). The Vegetation Map of South Africa, Lesotho, and Swaziland. South African National Biodiversity Institute. Online. http://bgis.sanbi.org/Projects/Detail/186 Version 2018.

Saraiva Okello, A.M.L., Masih, I., Uhlenbrook, S., Jewitt, G.P.W. & Van der Zaag, P. (2018). Improved Process Representation in the Simulation of the Hydrology of a Meso-Scale Semi-Arid Catchment. Water, 10, 1549. doi:10.3390/w10111549.

SAWS Home – WeatherSA Portal. (2024). Available from: http://www.weathersa.co.za/

Saxton, K.E. & Rawls, W.J. (2006). A PTF for estimating soil water retention and hydraulic conductivity using soil texture, bulk density, and organic matter content. Soil Science Society of America Journal, 70(5), 1569-1578. doi:10.2136/sssaj2005.0117.

Schaap, M.G. & Leij, F.J. (1998). Database-related accuracy and uncertainty of pedotransfer functions. Soil Science, 163(10), 765-779.doi: 10.1097/00010694-199810000-00001.

Schaap, M.G., Leij, F.J. & van Genuchten, M.T. (2001). ROSETTA: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. Journal of Hydrology, 251(3-4), 163-176. doi:10.1016/S0022-1694(01)00466-800466-8

Schoonover, J.E. & Crim, J.F. (2015). An introduction to soil concepts and the role of soils in watershed management. Journal of Contemporary Water Research & Education, 154(1), 21-47. doi:10.1111/j.1936-704X.2015.03186.x.

Schulze, R.E. & Schütte, S. (2020). Mapping soil organic carbon at a terrain unit resolution across South Africa. Geoderma, 373. doi:10.1016/j.geoderma.2020.114447.

Schulze, R.E. (2007). South African Atlas of Climatology and Agrohydrology. WRC Report 1489/1/06, Water Research Commission: Pretoria, South Africa.

Schulze, R.E. (2007). Soils: Agrohydrological information needs, information sources and decision support. In: South African Atlas of Climatology and Agrohydrology. WRC Report 1489/1/06, Section 41, Water Research Commission: Pretoria, South Africa.

Schulze, R.E. & Lynch, S.D. (2007). Annual Precipitation. In: South African Atlas of Climatology and Agrohydrology. WRC Report 1489/1/06, Section 62, Water Research Commission: Pretoria, South Africa.

Schulze, R.E. & Maharaj, M. (2007). Mean Annual Temperature. In: South African Atlas of Climatology and Agrohydrology. WRC Report 1489/1/06, Section 72, Water Research Commission: Pretoria, South Africa.

Seibert, J. & McDonnell, J.J. (2002). On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. Water Resources Research, 38, 23.1-23.14. doi:10.1029/2001WR000978

Shangguan, W., Dai, Y., Duan, Q., Liu, B. & Yuan, H. (2014). A global soil data set for earth system modelling. Journal of Advances in Modelling Earth Systems, 6, 249-263.

Sharififar, A. & Sarmadian, F. (2022). Coping with the imbalanced data problem in digital mapping of soil classes. European Journal of Soil Science, 74(3).

Sharififar, A., Sarmadian, F., Malone, B.P. & Minasny, B. (2019a). Addressing the issue of digital mapping of soil classes with imbalanced class observations. Geoderma, 350, 84-92.

Sharififar, A., Sarmadian, F. & Minasny, B. (2019b). Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling techniques. Computers and Electronics in Agriculture, 159, 110-118.

Shofiyati, R. Bachri, S. & Sarwani, M. (2011). Soil Database Management Software Development for Optimizing Land Resource Information Utilization to Support National Food Security. Journal of Geographic Information System, 3, 211-216.

Sierra, A. L. M., Roqueñí-Gutiérrez, N. & Loredo-Pérez, J. (2018). Methodology for the generation of hydropedological parameters associated with edaphic GIS coverage and databases for hydrological modelling. Proceedings, 2, 1411. doi:10.3390/proceedings2231411

Singh, C., Shashtri, S., Rina, K. & Mukherjee, S. (2012). Chemometric analysis to infer hydro-geochemical processes in a semi-arid region of India. Arabian Journal of Geosciences, 6(8), 2915-2932.

Skalko, J. (2013). If food and water are proportionate means, why not oxygen?. The National Catholic Bioethics Quarterly, 13(3), 453-467.

Smit, I.E. & Van Tol, J.J. (2022). Impacts of soil information on process-based hydrological modelling in the upper Goukou catchment, South Africa. Water, 14(3), 407. doi:10.3390/w14030407.

Smit, I.E., Van Zijl, G.M., Riddell, E.S. & Van Tol, J.J. (2023a). Examining the value of hydropedological information on hydrological modelling at different scales in the Sabie catchment, South Africa. Vadose Zone Journal, 00, 1-18. doi:10.1002/vzj2.20280.

Smit, I.E., Van Zijl, G.M., Riddell, E.S. & Van Tol, J.J. (2023b). Downscaling legacy soil information for hydrological soil mapping using multinomial logistic regression. Geoderma, 436, 116568. doi:10.1016/j.geoderma.2023.116568.

Smith, M. (2014). The impact of soil erosion on water quality in rivers: A review. Science of the Total Environment, 468-469, 306-317. South African Journal of Plant and soil. Volume 40, Issue 4-5. (2023). Taylor & Francis Oline. England & Wales No 3099067. https://www.tandfonline.com/toc/tjps20/current

Soil Classification Working Group. (1991). Soil classification: a taxonomic system for South Africa. Pretoria: Department of Agricultural Development.

Soil Classification Working Group. (2018) Soil Classification: A natural and anthropogenic system for South Africa. ARC-Institute for Soil, Climate and Water, Pretoria.

Srinivasan, R., Ramanarayanan, T.S., Arnold, J.G. & Bednarz, S.T. (1998). Large area hydrologic modelling and assessment part II: Model application. Journal of the American Water Resources Association, 34(1), 91-101.

Srinivasan, R., Zhang, X. & Arnold, J. (2010). SWAT ungauged: Hydrological budget and crop yield predictions in the upper Mississippi river basin. Transactions of the ASABE, 53(5), 1533-1546.

Suleaman, Y., Minasny, B., McBratney, AB., Sarwani, M. & Sutandi, A. 2013. Harmonizing legacy soil data for digital soil mapping in Indonesia. Geoderma, 192, 77-85.

Szabó, B., Weynants, M. & Weber, T.K.D. (2021). Updated European hydraulic pedotransfer functions with communicated uncertainties in the predicted variables (euptfv2). Geoscientific Model Development, 14(1), 151-175. doi:10.5194/gmd-14-151-2021.

Taghizadeh-Mehrjardi, R., Schmidt, K., Eftekhari, K., Behrens, T., Jamshidi, M., Davatgar, N., Toomanian, N. & Scholten, T. (2019). Synthetic resampling strategies and machine learning for digital soil mapping in Iran. European Journal of Soil Science, 71(3), 352-368. doi:10.1111/ejss.12893.

Tantithamthavorn, C., Hassan, A.E. & Matsumoto, K. (2018). The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. IEEE Transactions on Software Engineering, 46(11).

Theocharopoulos, S.P., Mitsios, I.K. & Arvanitoyannis, I. (2004). Traceability of environmental soil measurements. TrAC Trends in Analytical Chemistry, 23(3), 273-251. doi:10.1016/S0165-9936(04)00317-600317-6)

Thomas, A. (2015). Modelling of spatially distributed surface runoff and infiltration in the Olifants river catchment/water management area using GIS. International Journal of Advanced Remote Sensing and GIS, 4(1), 828-862. http://dx.doi.org/10.23953/cloud.ijarsg.81

Thompson, J.A., Roecker, S., Gunwald, S. & Owens, P.R. (2012). Digital soil mapping: Interactions with and applications for hydropedology. In: Hydropedology: Synergistic Integration of Soil Science and Hydrology. Elsevier: Amsterdam, Netherlands, 665-709.

Tignino, M. (2010). Water, international peace, and security. International Review of the Red Cross. 92(879):647-674.

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G. and Tóth, G. (2015), New generation of hydraulic pedotransfer functions for Europe. Eur J Soil Sci, 66: 226-238. https://doi.org/10.1111/ejss.12192

Tuppad, P., Douglas-Mankin, K.R., Srinivasan, R. & Arnold, J.G. (2011). Soil and Water Assessment Tool (SWAT) hydrologic/water quality model: Extended capability and wider adoption. Transactions of the ASABE, 54, 1677-1684.

USDA SCS (1972). National Engineering Handbook, Section 4 Hydrology. USDA Agricultural Conservation Service: USA.

USGS (2022). Landsat images. United States Geological Survey. http://landsat.usgs.gov

USGS (2018). Landsat images. United States Geological Survey. http://landsat.usgs.gov

USGS (2015). NASA Shuttle Radar Topography Mission (SRTM) Version 3.0 (SRTM Plus) Product Release. Land Process Distributed Active Archive Center, National Aeronautics and Space Administration. https://lpdaac.usgs.gov/about/news archive/nasa shuttle radar topography mission SRTM version 30 SRTM plus product release

Van Eekelen, M.W., Bastiaanssen, W.G.M., Jarmain, C., Jackson, B., Ferreira, F., Van der Zaag, P., et al. (2015). A novel approach to estimate direct and indirect water withdrawals from satellite measurements: A case study from the Incomati basin. Agriculture, Ecosystems & Environment, 200, 126-142. doi:10.1016/j.agee.2014.10.023.

Van Tol, J.J., Dzvene, A.R., Le Roux, P.A.L. & Schall, R. (2016a). Pedotransfer functions to predict Atterberg limits for South African soils using measured and morphological properties. Soil Use and Management, 32(4), 635-643. doi:10.1111/sum.12303.

Van Tol, J., Le Roux, P., Lorentz, S. & Hensley, M. (2013). Hydropedological classification of South African hillslopes. Vadose Zone Journal, 12(4), 1-10.

Van Tol, J., Van Zijl, G. & Julich, S. (2020). Importance of detailed soil information for hydrological modelling in an urbanized environment. Hydrology, 7, 34. doi: 10.3390/hydrology7020034.

Van Tol, J.J., Bieger, K. & Arnold, J.G. (2021). A hydropedological approach to simulate streamflow and soil water contents with SWAT+. Hydrological Processes, 35(6), Article ID: e14242. doi:10.1002/hyp.14242.

Van Tol, J.J., Le Roux, P.A.L. (2019). Hydropedological grouping of South African soil forms. South African Journal of Plant and Soil, 36, 233-235. doi:10.1080/02571862.2018.1537012.

Van Tol, J.J., Le Roux, P.A.L. & Hensley, M. (2012). Pedotransfer functions to determine water conducting macroporosity in South African soils. Water Science and Technology, 65(3), 550-557.

Van Tol, J.J. & Van Zijl, G.M. (2022). South Africa needs a hydrological soil map: a case study from the upper uMngeni catchment. Water SA, 48(4), 335-347. doi:10.17159/wsa/2022.v48.i4.3977.

Van Tol, J.J. & Van Zijl, G.M. (2020) Regional soil information for hydrological modelling in South Africa. Water Wheel, March April 2020, 43-45.

Van Tol, J.J., Van Zijl, G.M., Riddell, E.S. & Fundisi, D. (2015). Application of hydropedological insights in hydrological modelling of the Stevenson-Hamilton Research Supersite, Kruger National Park, South Africa. Water SA, 41(4), 525-533. doi:10.4314/wsa.v41i4.12.

Van Waveren, E,J. & Bos, A.B. 1988. ISRIC soil information system-user manual-technical manual. Wageningen: International Soil Reference and Information Centre.

Van Zijl, G.M., Ellis, F. & Rozanov, A. (2014a). Understanding the combined effect of soil properties on gully erosion using quantile regression. South African Journal of Plant and Soil, 31(3), 163-172. doi:10.1080/02571862.2014.944228.

Van Zijl, G.M., Le Roux, P.A.L. & Smith, H.J.C. (2012). Rapid soil mapping under restrictive conditions in Tete, Mozambique. In B. Minasny, B.P. Malone & A.B. McBratney (Eds.), Digital Soil Assessments and Beyond, CRC Press, Balkema, pp. 335-339.

Van Zijl, G.M., Van Tol, J.J. & Riddell, E.S. (2016). Digital soil mapping for hydrological modelling. In: Zhang G.L., Brus D., Liu F., Song X.D. & Lagacherie P. (Eds.), Digital Soil Mapping Across Paradigms, Scales and Boundaries, Springer Environmental Science and Engineering. Springer, Singapore.

Van Zijl, G.M. (2019). Digital soil mapping approaches to address real-world problems in southern Africa. Geoderma, 337, 1301-1308. doi:10.1016/j.geoderma.2018.07.052

Van Zijl, G.M., Bouwer, D., Van Tol, J.J. & Le Roux, P.A.L. (2014b). Functional digital soil mapping: A case study from Namarroi, Mozambique. Geoderma, 219-220, 155-161. doi:10.1016/j.geoderma.2013.12.014

Van Zijl, G.M., Van Tol, J.J., Bouwer, D., Lorentz, S.A. & Le Roux, P.A.L. (2020). Combining historical remote sensing, digital soil mapping, and hydrological modelling to produce solutions for infrastructure damage in Cosmo City, South Africa. Remote Sensing, 12, 433. doi:10.3390/rs12030433.

Van Zijl, G.M., Le Roux, P.A.L. Turner, D.P., (2013). Disaggregation of land types Ea34 and Ca11 by terrain analysis, expert knowledge and GIS methods. South African Journal of Plant and Soil, 30(3), 123-129. http://dx.doi.org/10.1080/02571862.2013.806679

Van Zijl, G.M.; Van Tol, J.J.; Tinnefeld, M.; Le Roux, P.A.L. (2019). A hillslope based digital soil mapping approach, for hydropedological assessments. *Geoderma* **2019**, doi:10.1016/j.geoderma.2019.113888.

Van Zijl, G.M.; Bouwer, D. (2012) *Soil Observation Dataset from the Halfway House Granites*; University of the Free State dataset; University of the Free State, Bloemfontein, South Africa, 2012.

Venables, B. & Ripley, B.D. (2002). Modern Applied Statistics with S. Fourth edition. Springer.

Vereecken, A., Huisman, J.A., Hendricks Franssen, H.J., Brüggemann, N., Bogena, H.R., Kollet, S., Javaux, M., van der Kruk, J. & Vanderborght, J. (2015). Soil hydrology: Recent methodological advances, challenges, and perspectives. Water Resources Research, 51, 2616-2633. doi:10.1002/2014WR016852.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M.G. and van Genuchten, M.T. (2010), Using Pedotransfer Functions to Estimate the van Genuchten-Mualem Soil Hydraulic Properties: A Review. Vadose Zone Journal, 9: 795-820. doi:10.2136/vzj2010.0045

Viljoen, G. & van der Walt, K. (2018). South Africa's water crisis – an interdisciplinary approach. Tydskrif vir Geesteswetenskappe, 58(3), 483-500.

Wadoux, A., Malone, B., Minasny, B., Fajardo, M. & McBratney, A. (2021). Soil Spectral Inference with R: Analysing Digital Soil Spectra Using the R Programming Environment. Springer International Publishing AG, Cham. doi:10.1007/978-3-030-64896-1

Wahren, F.T., Julich, S., Nunes, J.P., Gonzalez-Pelayo, O., Hawtree, D., Feger, K.H. & Keizer, J.J. (2016). Combining digital soil mapping and hydrological modelling data in a data scarce watershed in north-central Portugal. Geoderma, 264, 350-362. doi:10.1016/j.geoderma.2015.08.023.

Wang, X., Yang, W. & Melesse, A.M. (2009). Using hydrologic equivalent wetland concept within SWAT to estimate streamflow in watersheds with numerous wetlands. Transactions of the ASABE, 51(1), 55-72. doi:10.13031/2013.24227.

Wang, X., Yang, Y., Jianglong, L. & He, H. (2023). Past, present and future of the applications of machine learning in soil science and hydrology. Soil and Water Research, 18(2), 67-80. doi:10.17221/94/2022-SWR.

Wang, X. & Melesse, A.M. (2006). Effects of STATSGO and SSURGO as inputs on SWAT model's snowmelt simulation. Journal of the American Water Resources Association, 42, 1217-1236.

Wei, X., Zhang, H., Wang, L., Zhang, X., Li, P., Shi, J. & Yan, C. (2016). Modelling the effects of soil and water conservation measures on water quality in the Loess Plateau of China. Water, 8(2), 54.

Wenninger, J., Uhkenbrook, S., Lorentz, S.A. & Leibungut, C. (2008). Identification of runoff generation processes using combined hydrometric, tracer, and geophysical methods in a headwater catchment in South Africa. Hydrological Sciences Journal, 53, 65-80.

Weynants, M., Vereecken, H. & Javaux, M. (2009). Revisiting Vereecken Pedotransfer Functions: Introducing a Closed-Form Hydraulic Model. Vadose Zone Journal, 8(1), 86-95. doi:10.2136/vzj2008.0062.

Wischmeier WH, Smith DD. 1978. Predicting Rainfall Erosion Losses, a Guide to Conservation Planning. USDA Agriculture Handbook No. 537. USDA: Washington DC, USA.

Worqlul, A. W., Ayana, E. K., Yen, H., Jeong, J., MacAlister, C., Taylor, R., Gerik, T. J. & Steenhuis, T. S. (2018). Evaluating hydrologic responses to soil characteristics using SWAT model in paired-watersheds in the Upper Blue Nile Basin. Catena, 163, 332-341. doi:10.1016/j.catena.2017.12.040.

Winsemius, H.C., Schaefli, B., Montanari, A. & Savenije, H.H.G. (2009). On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. Water Resources Research, 45, W12422. doi:10.1029/2009WR007706.

Wischmeier, W.H. & Smith, D.D. (1978). Predicting Rainfall Erosion Losses: A Guide to Conservation Planning. USDA Agriculture Handbook No. 537. USDA: Washington DC, USA.

WoSIS Soil Profile Database. 2023. ISRIC-World Soil Information. The Netherlands. https://www.isric.org/explore/wosis.

Wösten, J.H.M., Lilly, A., Nemes, A. & Le Bas, C. (1999). Development and use of a database of hydraulic properties of European soils. Geoderma, 90(3-4), 169-185. doi: 0.1016/S0016-7061(98)00132-300132-3

Yamanaka, T., Kaihotsu, I., Oyunbaatar, D. & Ganbold, T. (2007). Summertime soil hydrological cycle and surface energy balance on the Mongolian steppe. Journal of Arid Environments, 69(1), 65-79.

Yao, R.J., Yang, J.S., Wu, D.H., Li, F.R., Gao, P. & Wang, X.P. (2015). Evaluation of pedotransfer functions for estimating saturated hydraulic conductivity in coastal salt-affected mud farmland. Journal of Soils and Sediments, 15(4), 902-916.

Yen, H., Bailey, R.T., Arabi, M., Ahmadi, M., White, M.J. & Arnold, J.G. (2014). The role of interior watershed processes in improving parameter estimation and performance of watershed models. Journal of Environmental Quality, 43, 1601-1613.

Yerro, A. & Ceccato, F. (2023). Soil-water-structure interactions. *Geotechnics*. 3(2):301-305.

Zarinabad, N., Wilson, M., Gill, S.K., Manias, K.A., Davies, N.P. & Peet, A.C. (2017). Multiclass imbalance learning: Improving classification of pediatric brain tumors from magnetic resonance spectroscopy. Magnetic Resonance in Medicine, 77, 2114-2124.

Zeraatpisheh, M., Jafari, A., Bodaghabadi, M.B., Ayoubi, S., Taghizadeh-Mehrjardi, R., Toomanian, N., Kerry, R. & Xu, M. (2020). Conventional and digital soil mapping in Iran: Past present and future. Catena, 188: 1-15.

Zhang, D., Lin, Q., Chen, X. & Chai, T. (2019a). Improved curve number estimation in SWAT by reflecting the effect of rainfall intensity on runoff generation. Water, 11(1), 163. doi:10.3390/w11010163.

Zhang, F., Zhang, Y., Ruan, J. & Liu, H. (2019b). An overview of soil moisture measurement methods. Paper delivered at the 2nd International conference on intelligent systems research and mechatronics engineering (ISRME 2019), Taiyuan. https://webofproceedings.org/proceedings_series/ESR/ISRME%202019/ISRME19075.pdf

Zhang, Y. & Schaap, M.G. (2019). Estimation of saturated hydraulic conductivity with pedotransfer functions: A review. Journal of Hydrology, 575(June), 1011-1030. doi:10.1016/j.jhydrol.2019.05.058

Zhang, G., Liu, F. & Song, X. (2017). Recent progress and future prospect of digital soil mapping: A review. Journal of Integrative Agriculture, 12(12), 2871-2885.

Zhang, H., Wei, X., Shi, J., Cao, Y., Zhang, Y., Xue, W. & Yan, C. (2015). The effects of soil and water conservation measures on soil and water loss on the Loess Plateau in China. Catena, 128, 166-177.

Zhou, Y., Chen, S., Hu, B., Ji, W., Li, S., Hong, Y., Xu, H., Wang, N., Xue, J., Zhang, X., Xiao, Y. & Shi, Z. (2022). Global soil salinity prediction by open soil vis-NIR spectral library. Remote Sensing, 14(21), 1-13. doi:10.3390/rs14215627.

Zhu, A. X. & Mackay, D. S. (2001). Effects of spatial detail of soil information on watershed modelling. Journal of Hydrology, 248, 54-77.

Zhu, A.-X., Yang, L., Li, B., Qin, C., English, E., Burt, J. E. & Zhou, C. (2008). Purposive sampling for digital soil mapping for areas with limited data. In A.E. Hartemink, A.B. McBratney & M.D.L. Mendonça-Santos (Eds.), Digital Soil Mapping with Limited Data. Springer, Dordrecht.

Zhu, B., Baesens, B. & Vanden Broucke, S.K.L.M. (2017). An empirical comparison of techniques for the class imbalance problem in churn prediction. Information Sciences, 408, 84-99.

Zimmermann, B. & Kohler, A. (2013). Optimizing Savitzky-Golay parameters for improving spectral resolution and quantification in infrared spectroscopy. Applied Spectroscopy, 67(8), 892-902.