land

MDPI

Article

Machine Learning Techniques for Estimating Hydraulic Properties of the Topsoil across the Zambezi River Basin

Mulenga Kalumba [1,2,*], Edwin Nyirenda [3], Imasiku Nyambe [4], Stefaan Dondeyne [5] and Jos Van Orshoven [1]

[1] Department of Earth and Environmental Sciences, University of Leuven, Celestijnenlaan 200E, 3001 Leuven, Belgium; jos.vanorshoven@kuleuven.be
[2] Department of Agricultural Engineering, The University of Zambia, Lusaka P.O. Box 32379, Zambia
[3] Department of Civil and Environmental Engineering, School of Engineering, The University of Zambia, Lusaka P.O. Box 32379, Zambia; edwin.nyirenda@unza.zm
[4] Department of Geology, School of Mines, The University of Zambia, Lusaka P.O. Box 32379, Zambia; inyambe@unza.zm
[5] Department of Geography, Ghent University, Krijgslaan 281 S8, 9000 Gent, Belgium; stefaan.dondeyne@ugent.be
* Correspondence: mulenga.kalumba@unza.zm

Abstract: It is critical to produce more crop per drop in an environment where water availability is decreasing and competition for water is increasing. In order to build such agricultural production systems, well parameterized crop growth models are essential. While in most crop growth modeling research, focus is on gathering model inputs such as climate data, less emphasis is paid to collecting the critical soil hydraulic properties (SHPs) data needed to operate crop growth models. Collection of SHPs data for the Zambezi River Basin (ZRB) is extremely labor-intensive and expensive, thus alternate technologies such as digital soil mapping (DSM) must be explored. We evaluated five types of DSM models to establish the best spatially explicit estimates of the soil water content at pF0.0 (saturation), pF2.0 (field capacity), and pF4.2 (wilting point), and of the saturated hydraulic conductivity (Ksat) across the ZRB by using estimates of locally calibrated pedotransfer functions of 1481 locations for training and testing the DSM models, as well as a reference dataset of measurements from 174 locations for validating the DSM models. We produced coverages of environmental covariates from various source datasets, including climate variables, soil and land use maps, parent materials and lithologic units, derivatives of a digital elevation model (DEM), and Landsat imagery with a spatial resolution of 90 m. The five types of models included multiple linear regression and four machine learning techniques: artificial neural network, gradient boosted regression trees, random forest, and support vector machine. Where the residuals of the initial DSM models were spatially autocorrelated, the models were extended/complemented with residual kriging (RK). Spatial autocorrelation in the model residuals was observed for all five models of each of the three water contents, but not for Ksat. On average for the water content, the R^2 ranged from 0.40 to 0.80 in training and test datasets before adding kriged model residuals and ranged from 0.80 to 0.95 after adding model residuals. Overall, the best prediction method consisted of random forest as the deterministic model, complemented with RK, whereby soil texture followed by climate and topographic elevation variables were the most important covariates. The resulting maps are a ready-to-use resource for hydrologists and crop modelers to aliment and calibrate their hydrological and crop growth models.

Keywords: digital soil mapping; multilinear regression; residual kriging; saturated hydraulic conductivity; spatial autocorrelation; water retention

Citation: Kalumba, M.; Nyirenda, E.; Nyambe, I.; Dondeyne, S.; Van Orshoven, J. Machine Learning Techniques for Estimating Hydraulic Properties of the Topsoil across the Zambezi River Basin. *Land* **2022**, *11*, 591. https://doi.org/10.3390/land11040591

Academic Editors: Jianzhi Dong, Yonggen Zhang, Zhongwang Wei, Sara Bonetti and Wei Shangguan

Received: 7 March 2022
Accepted: 14 April 2022
Published: 18 April 2022

1. Introduction

In a context where water availability is declining and/or competition for water is increasing, it is imperative to produce 'more crop per drop' [1]. To this end, rainfed and irrigated crop production systems must be designed which are adapted to the local soil,

climate, and socio-economic conditions. Integrated and well parameterized crop growth models are key to designing such crop production systems. However, today, in most crop growth and hydrological modelling studies, a lot of attention is devoted to collecting model inputs such as climate data, whereas less attention is given to collecting and/or modelling of the crucial soil hydraulic properties (SHPs) data with sufficient spatial resolutions needed to operate the models at a large scale. These SHPs data are mainly needed by crop growth and hydrological models because they play a vital role in the soil-atmosphere-vegetation interactions, which together with other water balance components, such as rainfall, evapotranspiration, runoff, ground, and surface water flow, influence the overall hydrological cycle, and are key in making water available for crops. Therefore, the SHPs, such as the soil water content at pF0.0 (saturation), pF2.0 (field capacity), and pF4.2 (wilting point), and the saturated hydraulic conductivity (Ksat) are fundamental for predicting water and energy exchange processes at the transition zone between solid earth and atmosphere. To be useful, the data must be sufficiently accurate [2]. Apart from the fact that the spatial resolution of SHPs datasets available for vast areas such as the Zambezi River Basin (ZRB) is insufficient, the extensive collection of physical soil data such as SHPs for large territories as the ZRB is extremely labor-intensive and costly, necessitating the investigation of scientifically valid alternative approaches such as digital soil mapping (DSM).

The ZRB is the fourth largest basin in Africa, with abundant water and land resources, and where more than 40 million people live [3,4]. Given the population growth and economic development, the demand for water and land resources increases. In particular, hydropower generation and irrigated agriculture demand large quantities of water [3,5]. To analyze the water–energy–food nexus, and to develop a decision analytical framework in support of policy and decision makers for the ZRB [6], the hydrological model TOPKAPI [7–9] and the FAO crop growth model AquaCrop [10] have been used. Process-based hydrological models such as TOPKAPI and the AquaCrop crop growth model require soil hydraulic properties as input data, including the saturated hydraulic conductivity and the water retention characteristics. However, SHPs data with higher spatial resolutions are not available on a large scale, such as the ZRB, which spans eight African countries, and measuring them is both expensive and time intensive. Hence, measured data on these soil properties are normally estimated on a point scale from basic soil data by means of pedotransfer functions [11]. However, to generate spatial coverages of these soil hydraulic properties across a large basin such as the ZRB, digital soil mapping techniques can be used.

According to [12], DSM techniques target the generation of coverages of soil properties at a given spatial resolution, based on quantitative relationships between spatially explicit environmental data (predictor variables or covariates) and the properties of interest, observed or measured in the field or in the laboratory. It can be seen as a process whereby insights from both conventional soil surveys and geostatistical approaches are combined resulting in a hybrid approach (Figure 1).

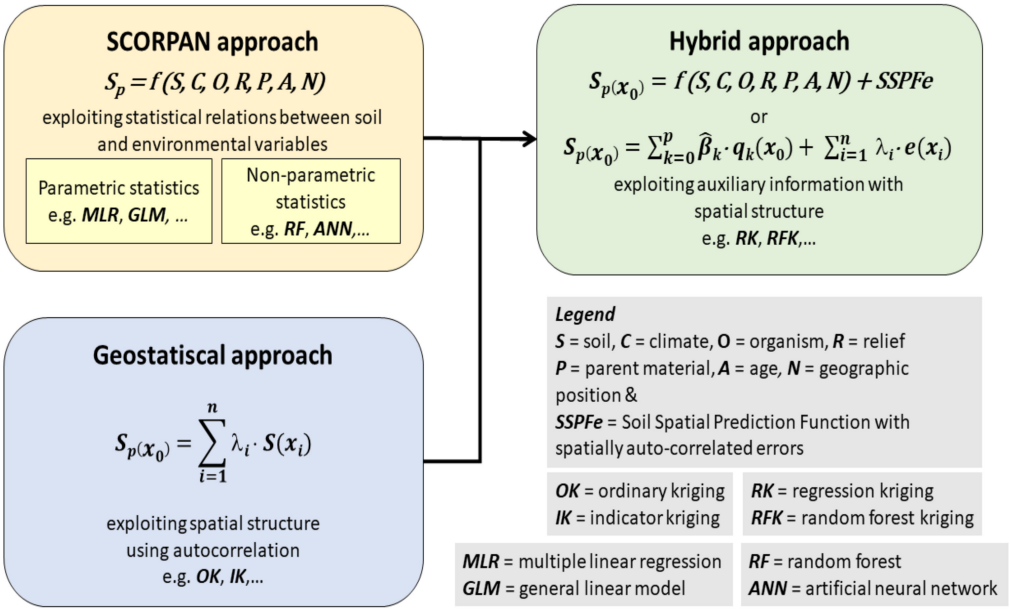

Figure 1. Integration of the SCORPAN model with geostatistical approaches for digital soil mapping (adapted from: [13]).

Whereas conventional soil survey approaches implicitly account for spatial variation and auto-correlation of Jenny's soil forming factors [14]. Ref. [12] further elaborated and reformulated such approaches into DSM models of the "SCORPAN-SSPFe" type. "SCORPAN" stands for a priori knowledge on the soils (S), besides Jenny's soil forming factors climate (C), organisms (O), relief (R), parent material (P) and age (A) and supplemented with information on the geographic position (N), and "SSPFe" stands for "soil spatial prediction function with spatially auto-correlated errors". The SCORPAN (Figure 1) term also refers to the collection of environmental covariates (predictors in the deterministic models), which are nowadays widely available as geodatasets, and include legacy soil maps, climatic data, land cover data often derived from remote sensing data, digital elevation models (DEMs) and their derivatives, and geological maps [15].

The four most fundamental components in any DSM approach are the response variable or dependent variable, a model, independent variables (the soil environmental covariates) representing the SCORPAN factors, and the SSPFe term. A model must be chosen to build a map of soil attributes or soil classes scattered throughout a landscape after picking an ideal set of SCORPAN variables. DSM distinguishes three basic categories of modeling techniques: (i) geostatistical techniques [16,17] e.g., ordinary kriging, (ii) non-geostatistical techniques [12,18] e.g., multiple linear regression (MLR) and machine learning (ML) approaches, which include: artificial neural network (ANN), gradient boosted regression trees (BRT), random forest (RF), and support vector machine (SVM), and lastly, (iii) hybrid methods that combine the benefits of the two preceding techniques. Figure 1 depicts the various approaches, and mentions several techniques, one of which is the geostatistical ordinary kriging (OK), also known as the best linear unbiased predictor [19]. OK has been modified to more accurate hybrid algorithms that account for secondary information, the most flexible of which is residual kriging (RK) [20]. The trend component of the RK model is typically derived by global linear regression of the soil environmental variables on the target or dependent variable [18]. Any prediction approach can theoretically be cast in RK if the linked model residuals are spatially auto correlated. Therefore, the term "residual kriging" is preferred, even if the same technique has been called "simple kriging with variable local means" by [21] or "regression kriging" by [20].

ML techniques have been explored and applied successfully for digital quantitative mapping of soil properties, such as soil organic carbon content and soil texture both over

large and small areas [12,18,22–25], but have rarely been applied for mapping soil hydraulic properties across areas as vast as the Zambezi River Basin. However, ML-based evaluations in water resources have been applied over complex terrain regions such as the ZRB, since hydraulic/hydrological estimates can be associated with significant errors due to variability and uncertainty introduced by orographic effects [26–28].

Using DSM techniques that are based on machine learning such as artificial neural network (ANN), gradient boosted regression trees (BRT), random forest (RF) and multinomial logistic regression, the International Soil Reference and Information Centre (ISRIC) is making great efforts to provide global gridded soil data at 250 m × 250 m resolution through its SoilGrids web interface [29]. However, although SoilGrids provides data on water content at saturation (pF0.0) and the available water capacity (AWC), it does not provide data for other soil hydraulic properties, such as saturated hydraulic conductivity, water content at field capacity (pF2.0), or at the wilting point (pF4.2), which are needed by most hydrological and crop-growth models. Furthermore, the accuracy of the SoilGrids data is moderate, with R^2 ranging from about 0.30 to 0.80 [29], hence it is advisable to improve on this data for specific projects. Most DSM projects based on machine learning models have been conducted in regions in the northern hemisphere and Australia, for which availability of field measured data is abundant [12,24,30,31], but only few applications of DSM have been reported for southern countries, such as Zambia in the ZRB. Apart from studies by [29] for the SoilGrids Database, [32] for projects concerning the Edgeroi district in north-western Australia across a 1500 km^2 area on a spatial resolution of 90 m × 90 m, and [33] for the 5775 km^2 Balaton catchment area in Hungary on a spatial resolution of 100 m × 100 m, studies implementing SCORPAN approaches in DSM to predict soil hydraulic properties over a vast area such as the ZRB are absent. To fill this gap, the aim of this study was to identify the best performing ML algorithm for digital mapping of soil hydraulic properties for the whole ZRB at a spatial resolution of 90 m × 90 m. The purpose was to produce data layers for Ksat, the water contents at pF0.0, pF2.0, and pF4.2, and the AWC of the topsoil (0–30 cm) that can be used in hydrologic and crop growth models. The specific objectives were therefore to:

a. Evaluate the performance of five alternative deterministic DSM models (1) multiple linear regression, (2) artificial neural network, (3) gradient boosted regression trees, (4) random forest and (5) support vector machine using easily available environmental covariates.
b. Verify whether the performance can be improved by accounting for the spatial autocorrelation among residuals from all five deterministic models, and by conducting residual kriging.
c. Establish the most appropriate technique after comparing the prediction performance of all five deterministic models each with and without related residual kriging.

2. Materials and Methods

2.1. Study Area

The Zambezi River Basin is a 1.6 million km^2 large basin in southern Africa which spans eight riparian countries (Figure 2). The seasonal climatic variations are determined by the movement of the Inter-Tropical Convergence Zone (ITCZ), with the major rainy season lasting from December to March. Annual rainfall ranges from less than 300 mm in the south to more than 1200 mm on the northern plateaus, and to more than 2300 mm on the highlands in Malawi and Tanzania. Mount Mulanje, in Malawi, measuring 3002 m above sea level, is the highest point in the basin [3]. The western and central parts of the ZRB are dominated by plains and plateaus [34]. The eastern parts are dominated by mountains, escarpments, and valleys, which are part of the East African Rift valley system [35,36]. The plateaus and plains of the western part of the basin are dominated by formations of the Karoo system. The Karoo formations are continental sediments, which include shales, mudrocks, sandstones, and basalt outflows [37]. In the western plains, these formations are covered by the Kalahari sands, which are Quaternary aeolian deposits [35]. Native

semi-deciduous woodland occupies more than half of the land area, followed by native grassland used for animal husbandry, semi-deciduous shrubland, agricultural crop land, water bodies, urban areas, and wetlands, in that order [38].

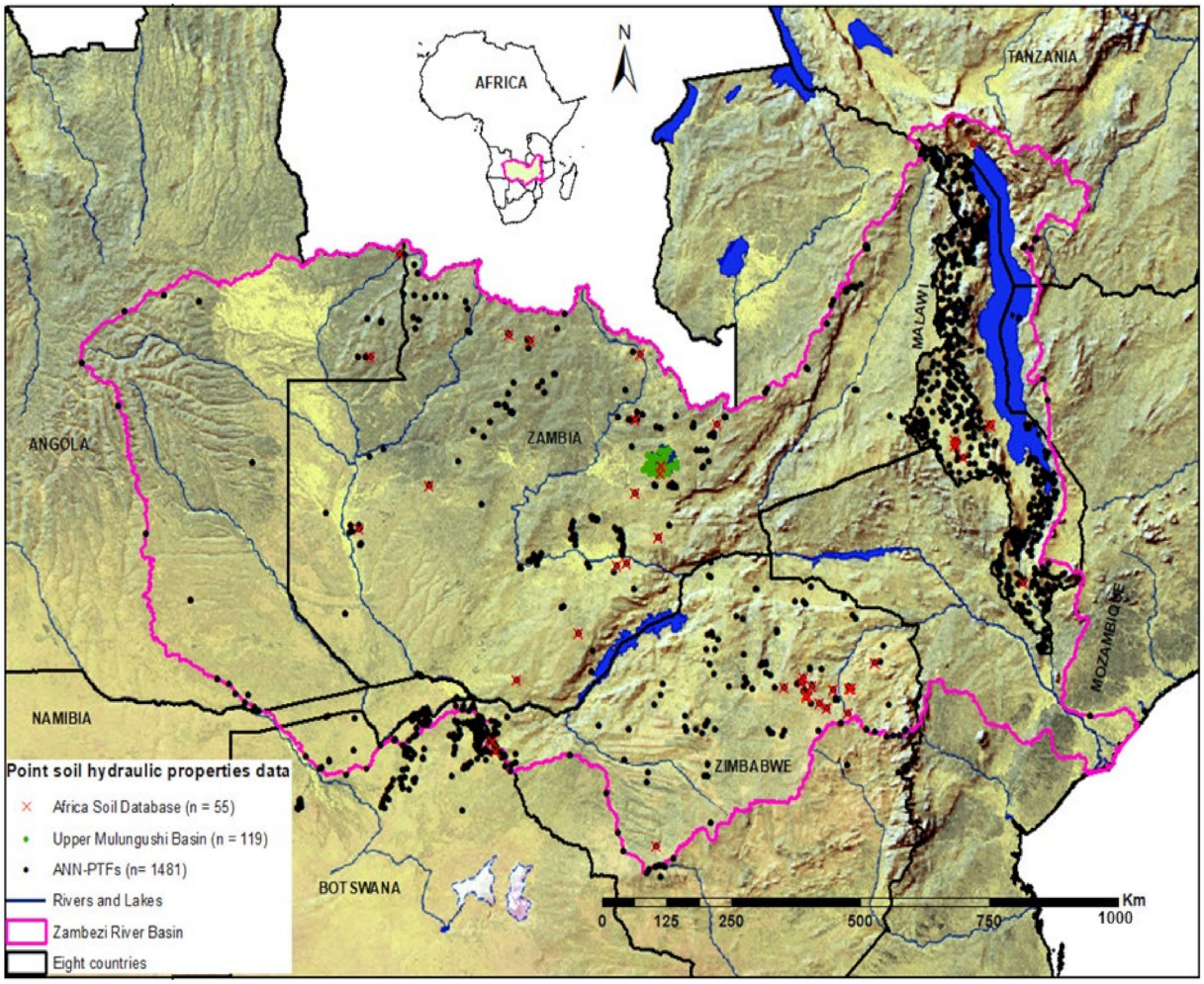

Figure 2. The Zambezi River Basin and location of (i) the soil profiles in the African Soil Profile database (n = 55 in red), for which soil hydraulic data are available, (ii) own soil sampling points in the Upper-Mulungushi subbasin in central Zambia (n = 119 in green), and (iii) of the soil profiles in the African Soil Profile database (n = 1481 in black), for which soil hydraulic data were estimated using ANN-PTFs [11].

2.2. Soil Hydraulic Data

In order to train and test the different DSM models, dataset #1 containing estimates of soil hydraulic data for (n = 1481) georeferenced points (Figure 2) throughout the ZRB and available in the Africa Soil Profiles Database, obtained from a set of locally calibrated artificial neural network-based pedotransfer functions (ANN-PTFs) [11]. The calibration of these PTFs was performed based on dataset #2 and #3. In 2018, a soil sampling and measurement campaign (Figure 2) was conducted in the 2000 km^2 Upper Mulungushi sub-basin (UMB) of the ZRB, resulting in a dataset containing measured soil hydraulic properties for depths of 30–40, 60–70, and 100–110 cm at (n = 119) georeferenced point locations, further termed as the UMB-dataset or dataset #2 (Figure 2). From the same 119 locations and at the same depth layers, we also took undisturbed core samples with Kopecky rings (100 cm^3) for measuring saturated soil hydraulic conductivity, and the water contents at three matric potentials: pF0.0, pF2.0, and pF4.2. The Ksat was measured in the laboratory by placing the Kopecky rings with the undisturbed soil samples in a constant

head permeameter apparatus. The resulting data served the validation of the outcome of the DSM for water contents at pF0.0, pF2.0, pF4.2, the AWC, and Ksat. In addition, a Legacy dataset (dataset #3) containing (n = 55) georeferenced points with measured water contents at pF2.0 and pF4.2 was extracted from the Africa Soil Profiles Database [39,40] (Figure 2). These were also used to validate the developed DSM maps for the water content at pF2.0, pF4.2, and for the AWC.

2.3. Environmental Covariates

In DSM studies, it is important to have a good number of all of the SCORPAN factors for optimal models to be developed, hence, in this study, 67 independent variables (Table 1) or the soil environmental covariates were considered to represent the five SCORPAN factors of soil, climate, organisms, relief, and parent materials [12,14]. According to [15,41], in DSM techniques, the use of many environmental covariates is highly encouraged. These variables are available as spatial coverages, each with its own spatial resolution ranging from 20 to 1000 m, over the whole ZRB. All of the covariate layers were projected in the same cartographic reference system (WGS84/UTM zone 35S), and resampled to a 90 m spatial resolution using the nearest neighbor approach for categorical covariates and the bilinear approach for continuous covariates [42]. The soil map, lithology map, land use map, and the landforms map provided the categorical environmental covariates. The number of distinct mapping units correspond to the number of classes that were transformed to as many dummy (indicator, binary) variables. Each dummy variable received a value equal to one (1) when a given class level was present, and zero (0) otherwise. Table 1 gives an overview of the 67 covariates that comprised the factors included in the SCORPAN concept that were used in this study.

2.3.1. Soil (S)

The soil map based on the Harmonized Soil Map of Africa [43] was complemented with the more detailed soil and terrain maps (SOTER) for Southern Africa [44] and the SOTER of Malawi [45]. The legend of the map is in accordance with the second edition of the international soil classification system, "World Reference Base for soil resources" (WRB) [46]. Data on silt, sand, and clay fractions, bulk density, organic carbon content, pH, and the depth to bedrock, at a spatial resolution of 250 m, and depth layers of 0–30 cm were downloaded from the SoilGrids database (https://soilgrids.org/ accessed on 18 January 2019) of ISRIC-World Soil Information [29].

2.3.2. Climate (C)

Climate variables, such as those informing about temperature and precipitation regimes, which are typically employed in DSM techniques, are frequently derived from long term meteorological station observations. Advances in remote sensing techniques, on the other hand, can overcome limited access to consistent datasets in specific places. Products derived from time series of surface moisture, temperature, and evapotranspiration extracted from remote sensing imagery can also be utilized as climate data [47]. In this study, 19 data layers of bioclimatic variables were obtained from the WorldClim 2.0 dataset, which gathers climate variables worldwide at a spatial resolution of 1 km [48]. Among other climate variables, this dataset provides the average monthly minimum, mean, and maximum temperature, precipitation, and potential evapotranspiration derived from observations over the period from 1970 to 2000.

2.3.3. Organisms (O)

This SCORPAN factor encompasses mainly vegetation and human activities. To represent these, the S2 prototype land cover map of Africa for the year 2016 obtained from http://2016africalandcover20m.esrin.esa.int/ (accessed on 18 January 2019) was used. It has a 20 m spatial resolution and features 10 classes (Table 2). Vegetation types over large areas are commonly mapped from remotely sensed spectral data, hence Enhanced Vegeta-

tion Index (EVI) images derived from MODIS-imagery at 250 m resolution as continuous covariates were downloaded from the Land Processes Distributed Active Archive Centre website [49]. Moreover, the continuous covariate tree canopy area fraction for the year 2000 was obtained at a 100 m spatial resolution [50].

2.3.4. Relief (R)

To account for relief as one of the influential factors for soil hydraulic properties, a number of attributes were derived from the SRTM-DEM obtained from http://srtm.csi.cgiar.org (accessed on 18 January 2019) at a spatial resolution of 90 m. Various combinations of terrain attributes can characterize geomorphic surfaces and describe processes related to soil development. The DEM was (pre-)processed using the SAGA software version 2.1.4 [51] by first filling artificial depressions ('sinks') using the Planchon/Darboux algorithm [52], and subsequently smoothened by applying a Gaussian filter [51]. Continuous DEM derivatives, such as the slope, aspect, profile curvature, plan curvature, slope-length factor (LS-factor), topographic wetness index (TWI), convergence index, and topographic roughness index, were derived using the basic terrain analysis tools provided by the SAGA software. Finally, using the landform classification tool of SAGA, a landform map providing 16 binary covariates, each corresponding to one landform class (Table 2), was generated.

2.3.5. Parent Material (P)

To incorporate parent material among the covariates, we used a lithology map with 20 legend classes (Table 2), obtained from http://geoportal.rcmrd.org/layers/servir%3Aafrica_surfacelithology (accessed on 18 January 2019), at a spatial resolution of 100 m. A selection of the environmental covariates used are displayed in Figure 3. The legend of the categorical covariates displayed in Figure 3 is in Table 2.

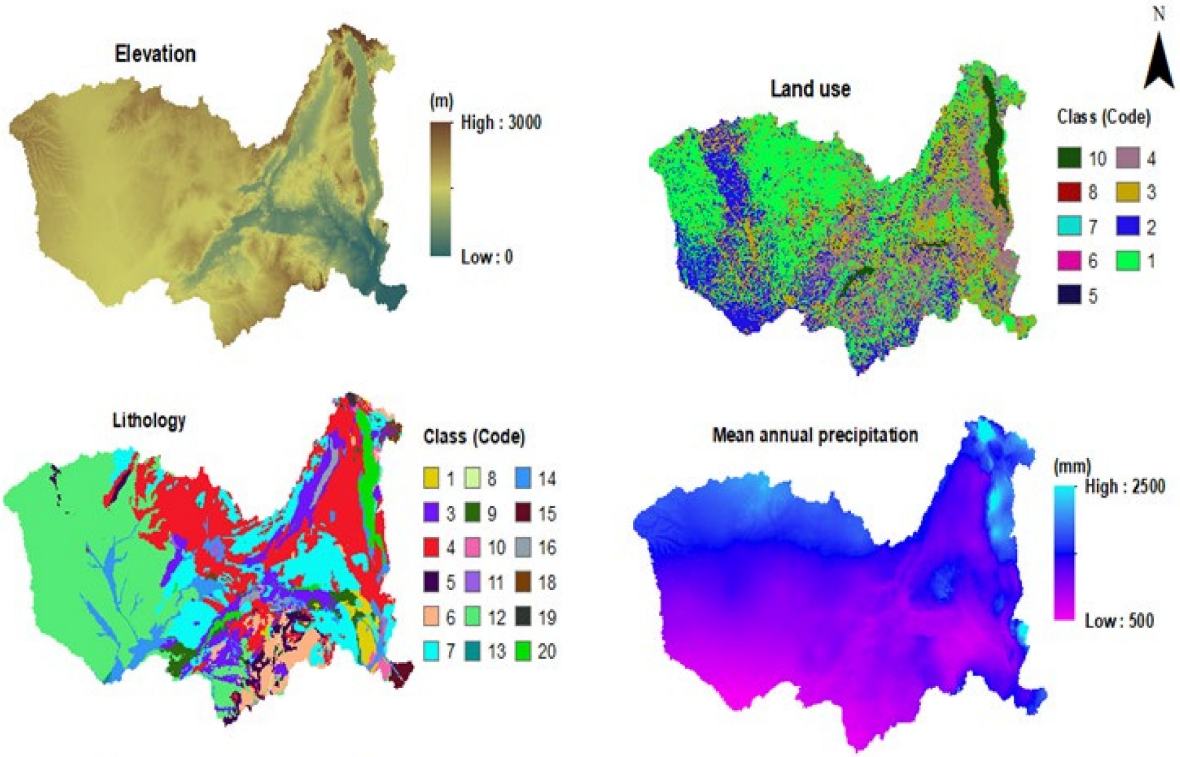

Figure 3. *Cont.*

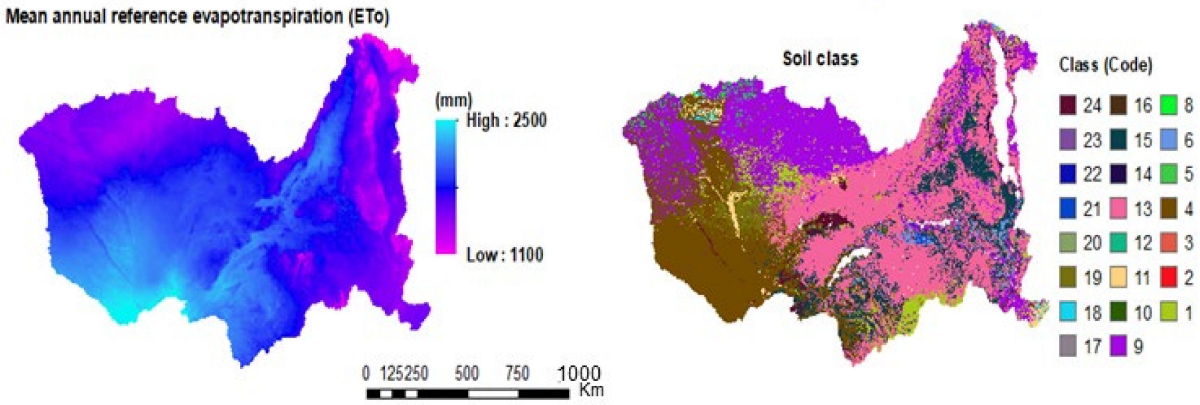

Figure 3. Six out of the total 67 examined SCORPAN covariates: elevation, land use, lithology, mean annual precipitation and evapotranspiration, and the soil class map. Legend of the categorical maps (soil, land use, and lithology) is presented in Table 2.

Table 1. Soil environmental covariates and their abbreviations, type, spatial resolution, units and range of values listed as "SCORPAN factors".

Abbreviation	Covariate	Type of Data	Spatial Resolu-tion	Units	Range of Values (ZRB)	Source
	Soil factor					
SOL	Soil class map	Categorical	250 m	RSG	24 classes	Soil Map of Africa; SOTER SAF & Malawi [43–45]
SND	Sand	Continuous	"	%	18–95	SoilGrids [53]
CLY	Clay	"	"	"	3–72	"
SLT	Silt	"	"	"	1–47	"
BLD	Bulk Density	"	"	$kg \cdot m^{-3}$	881–1849	"
OC	Organic Carbon	"	"	"	30–300	"
PH	Soil pH(H_2O)	"	"	–	4.5–8.5	"
BED	Depth to Bedrock	Continuous	"	m	0–2	"
	Climate factor					
PET	Potential Evapotranspiration	"	"	"	1000–2500	Global Aridity Index and PET Database [54]
BIO1	Annual Mean Temperature	"	"	°C	10–26	WorldClim database [48]
BIO2	Mean Diurnal Range	"	"	"	7–22	"
BIO3	Isothermality	"	"	%	52–74	"
BIO4	Temperature Seasonality	"	"	–	146–386	"
BIO5	Max Temperature of Warmest Month	"	"	°C	17–37	"
BIO6	Min Temperature of Coldest Month	"	"	"	2–17	"
BIO7	Temperature Annual Range	"	"	"	14–30	"
BIO8	Mean Temperature of Wettest Quarter	"	"	"	12–29	"
BIO9	Mean Temperature of Driest Quarter	"	"	"	8–24	"
BIO10	Mean Temperature of Warmest Quarter	"	"	"	12–30	"
BIO11	Mean Temperature of Coldest Quarter	"	"	"	7–23	"
BIO12	Annual precipitation	"	1000 m	mm	400–2200	"
BIO13	Precipitation of Wettest Month	"	"	mm	150–650	"

Table 1. *Cont.*

Abbreviation	Covariate	Type of Data	Spatial Resolution	Units	Range of Values (ZRB)	Source
BIO14	Precipitation of Driest Month	"	"	"	0–40	"
BIO15	Precipitation Seasonality	"	"	"	75–136	"
BIO16	Precipitation of Wettest Quarter	"	"	"	200–1340	"
BIO17	Precipitation of Driest Quarter	"	"	"	0–126	"
BIO18	Precipitation of Driest Quarter	"	"	"	74–760	"
BIO19	Precipitation of Coldest Quarter	"	"	"	0–160	"
Organism, landcover factor						
LAN	Land use map	Categorical	20 m	–	10 classes	Land Cover data of Africa [55]
EX1	Enhanced Vegetation Index (EVI) for Jan. & Feb.	Continuous	250 m	"	−1–1	MODIS Enhanced Vegetation Index [56]
EX2	EVI for March & April	"	"	"	"	"
EX3	EVI for May & June	"	"	"	"	"
EX4	EVI for July & August	"	"	"	"	"
EX5	EVI, September & October	"	"	"	"	"
Organism, landcover factor						
EX6	EVI, November & December	"	"	"	"	"
FORTC	Forest Tree Cover	"	90 m	%	0–100	Hansen tree cover data of 2000 [57]
Relief, topography factor						
ELE	Elevation	Continuous	90 m	m	0–2500	SRTM void filled data [57,58]
TRI	Terrain Ruggedness Index	"	"	"	0–32	Derived from SRTM data
VTR	Vector Terrain Ruggedness	"	"	–	−0.2–0.60	"
LSF	LS-factor	"	"	"	0–11	"
SLP	Slope	"	"	Radians	0–0.51	"
CRD	Local Downslope Curvature	"	"	"	−1.15–0.50	"
UPCUR	Upslope Curvature	"	"	"	−0.15–0.5	"
DNCUR	Downslope Curvature	"	"	"	−0.55–0.28	"
MRN	Melton Ruggedness Number	"	"	–	0–10	"
SPI	Stream Power Index	"	"	"	0–20000	"
TWI	Topographic Wetness Index	"	"	–	4–16	"
FOR	Landforms map	Categorical	"	"	16 classes	"
VBF	Valley Bottom Flatness	Continuous	"	"	0–11	"
CRU	Local Upslope Curvature	"	"	Radians	−0.65–0.62	"
TPI	Topographic Position Index	"	"	–	−8–10	"
POS	Positive Openness	"	"	"	1.2–1.6	"
NEG	Negative Openness	"	"	"	1.3–1.6	"
DVM	Deviation from Mean Value	"	"	"	−210–218	"
GECUR	General Curvature	"	"	Radians	−0.5–0.6	"
PRCUR	Profile Curvature	"	"	"	−0.1–0.4	"
PLCUR	Plan Curvature	"	"	"	−0.4–0.5	"
TACUR	Tangential Curvature	"	"	"	−0.14–0.2	"
VDP	Valley Depth	"	"	m	4–1026	"
LOCUR	Local Curvature	"	"	Radians	−0.55–0.68	"

Table 1. *Cont.*

Abbreviation	Covariate	Type of Data	Spatial Resolution	Units	Range of Values (ZRB)	Source
CSCUR	Cross Sectional Curvature	"	"	"	−0.03–0.05	"
CONVI	Convergence Index	"	"	"	−81–75	"
CNBL	Channel Network Base Level	"	"	m	17–1722	"
LNCUR	Longitudinal Curvature	"	"	Radians	−0.05–0.06	"
ASP	Aspect	"	"	"	0–7	"
	Parent material factor					
LIT	Lithology map	Categorical	100 m	–	20 classes	Africa Surface Lithology [59]

Covariates with the same type of data, spatial resolution, units, range of values and source (").

Table 2. Definition of classes for categorical variables.

Soil Class/RSG	Code	Land Use	Code	Lithology	Code	Landforms	Code
Acrisols	1	Tree Cover	1	Calcareous rocks	1	Very steep slope, high convexity	1
Alisols	2	Shrubs	2	Karst rocks	2	Very steep slope, high convexity	2
Andosols	3	Grasslands	3	Calcareous sedimentary rocks	3	Very steep slope, low convexity	3
Arenosols	4	Croplands	4	Meta-sedimentary rocks	4	Very steep slope, low convexity	4
Calcisols	5	Vegetation/Wetlands	5	Alkaline intrusive volcanic rocks	5	Steep slope, high convexity	5
Cambisols	6	Sparse vegetation	6	Silicic rocks	6	Steep slope, high convexity	6
Chernozems	7	Bare areas	7	Meta-igneous rocks	7	Steep slope, low convexity	7
Durisols	8	Built up areas	8	Ultramafic rocks	8	Steep slope, low convexity	8
Ferralsols	9	/	/	Extrusive volcanic rocks	9	Moderate slope, high convexity	9
Fluvisols	10	Open water bodies	10	Colluvium sediments	10	Moderate slope, high convexity	10
Gleysols	11			Water saturated and Organic sediments	11	Moderate slope, low convexity	11
Histosols	12			Aeolian sediments	12	Moderate slope, low convexity	12
Leptosols	13			Alluvium-(Fan deposits)	13	Gentle slope, high convexity	13
Lixisols	14			Alluvium-(Fluvial deposits)	14	Gentle slope, high convexity	14
Luvisols	15			Alluvium-(Beach & coastal deposits)	15	Gentle slope, low convexity	15
Nitisols	16			Alluvium-(Saline deposits)	16	Gentle slope, low convexity	16
Phaeozems	17			/	/		
Planosols	18			Alluvium-(other)	18		
Podzols	19			Volcanic-(Ash mudflow)	19		
Regosols	20			Water bodies	20		
Solonchaks	21						
Solonets	22						
Umbrisols	23						
Vertisols	24						

2.4. Covariate Selection

All the soil environmental covariates were in the *GeoTiff* raster format, and all data pre-processing, model training, testing, and validation were performed by means of R-software, version 3.5.0 [60]. The 67 environmental covariate layers were first reprojected into the WGS84/UTM zone 35S coordinate system, and then resampled to a 90 m spatial resolution. Using the *raster package* and *stack function* of the R-software, a RasterStack dataset consisting of all 67 covariates was created. A RegressionMatrix for each dependent or response variable was created by overlaying the georeferenced point dataset carrying the response variables, that is, the hydraulic soil properties estimated by means of the ANN-PTFs for the soil samples in dataset #1 (n = 1481, displayed in Figure 2), with the RasterStack dataset, and using the *extract function* of the R-software. Therefore, this RegressionMatrix consisted of the value of a response variable (n = 1481) for a depth of 30 cm on the one hand, and the corresponding pixel values of all the 67 environmental covariates used in this study on the other hand. From the RegressionMatrix, for each response variable at the 30 cm depth, potential covariates were selected by using the backward-stepwise selection of the Akaike Information Criterion (AIC) *R package* [61], and by testing 67 possible models until the AIC stopped decreasing. As a result, for each response variable, we selected the model with the potential covariates that had the lowest AIC number.

2.5. Deterministic SCORPAN Models, Training, and Testing

Once the potential covariates were selected for each of the dependent variables, the RegressionMatrix was randomly split into a training (70%, n = 1037, dataset #1a), and a test dataset (30%, n = 444, dataset #1b). With these datasets, and for each response variable at the 30 cm depth, five deterministic models, namely multiple linear regression (MLR) in R, artificial neural network (ANN) (using the *neuralnet* R package), Gradient boosted regression trees (BRT) (using the *gbm* R package), random forest (RF) (using the *randomForest* R package) and support vector machine (SVM) (using the *e1071* R package), were trained and tested. To avoid over-fitting, we used the default meta-parameters in the models, such that no model tuning was performed, apart from adjusting the number of trees (*n.trees*) in the BRT model from 100 to 500 to match the same default number of trees as in the RF model. The default meta-parameters included the number of *neutrons* and *hidden layers*, such that one neutron with one hidden layer were the default meta-parameters for the ANN; the *n.tree*, *shrinkage*, and *interaction depth* for the BRT; the *mtry* and *n.tree* for the RF; and the *gamma* and *cost function* for the SVM as some of the key meta-parameters that were used in default mode. The performance of all the models was evaluated using the coefficient of determination (R^2), the mean absolute error (MAE), and the root mean squared error (RMSE). Models that had high R^2, low MAE, and low RMSE were well performing. Furthermore, the potential covariates that were selected in the model training and testing were ranked according to their order of importance relative to the particular response variable using the *relative importance function* of the random forest model.

2.6. Spatial Autocorrelation and Model Validation

In DSM approaches, apart from the SCORPAN term, there is also the SSPFe (soil spatial prediction function with spatially auto-correlated errors) term to consider (Figure 1, [12]). It has been suggested and demonstrated [18] that adding model residuals to the deterministic component improves model performance [12]. Therefore, after training and testing each of the deterministic models, the model residuals were obtained by subtracting the model estimations from the reference value of each response variable. We then checked for spatial autocorrelation of the residuals by drawing semi-variograms using the *gstat* R package [62]. In case spatial autocorrelation was observed in the model residuals, a raster coverage in *GeoTiff* format of the model residuals was generated using ordinary kriging (OK) [20,63–65]. The kriged map of the model residuals was then added to the deterministic estimations of the training and testing models, and their performance was again evaluated using R^2, MAE and RMSE.

The trained and tested models built from the potential covariates were subsequently used to estimate coverages of the dependent variable for the whole ZRB using R's *predict function* and the *RasterStack* dataset consisting of all the 67 covariates, whereby only the potential covariates selected were used in the prediction. If there was no spatial autocorrelation, we predicted for the whole ZRB straight away, and then subsequently validated this hydraulic soil property map. However, if spatial autocorrelation was observed in the model residuals, then the kriged model residuals were added to the deterministic estimations for the whole ZRB. After this addition, maps in *raster* and *GeoTiff* formats of each response variable for 30 cm depth were created. Finally, to validate this hydraulic soil property map, it was overlaid with the validation point datasets consisting of the measured hydraulic soil properties data of each response variable in dataset #2 (n = 119) obtained from the Upper-Mulungushi subbasin, and dataset #3 (n = 55) retrieved from the African Soil Profile database (Figure 2). Finally, the R^2, MAE, and RMSE were computed for the 55 + 119 (n = 174) data points.

3. Results

3.1. Selected Covariates

As an illustration, the results of a random forest model in Figure 4 present the twenty highest ranked covariates out of the 67 candidate predictor variables (Table 1). Variables are ranked in terms of importance on the y-axis (with variables of highest importance at the top), with a mean decrease in accuracy (%IncMSE) should that particular variable be removed from the random forest model on the x-axis. The mean decrease in accuracy (%IncMSE) shows how much the model accuracy would decrease if we were to leave out a particular variable. For the water content at pF0.0, pF2.0, and pF4.2, sand content (SND), silt content (SLT), climate variables such as average annual rainfall (BIO12), and mean annual reference evapotranspiration (PET) and temperature variables such as the minimum temperature of the coldest month (BIO6) were among the highest ranked covariates. Clay content (CLY), soil organic carbon content (OC), and elevation (ELE), followed by climate variables such as the mean temperature of the coldest quarter (BIO11) were the most important covariates for the AWC, while for Ksat, sand content, the reference soil group arenosols (SOL with code 4) (Table 2), and climate variables, such as the average annual rainfall and the mean temperature of the coldest quarter, were among the most important covariates (Table 1).

3.2. Spatial Autocorrelation

We also observed spatial autocorrelation of the residuals for each of the five models that were developed for the water content at pF0.0, pF4.2, and the AWC. For the water content at pF2.0, spatial autocorrelation of the model residuals was only observed with the RF model. Furthermore, there was no spatial autocorrelation of the model residuals for all five models for Ksat. As an illustration, Figure 5, shows the semivariograms of the model residuals for all five models for the water content at pF0.0 in the topsoil (0–30 cm). For the ANN model, model residuals exhibited spatial autocorrelation up to a range of about 9.8 km. The range in the variograms of the BRT, MLR, RF, and SVM model residuals was about 7.6, 10.6, 7.1, and 10.3 km, respectively (Figure 5).

3.3. Model Performance Evaluation

The evaluation of the performance of the five DSM models each without and complemented with residual kriging was based on comparing the R^2, MAE, and RMSE for the training (dataset #1a), test (dataset #1b), and validation data sets (datasets #2 and #3) (Figures 6–8). Overall, the lowest RMSE, the lowest MAE, and the highest R^2 were observed in the training and test data sets after adding the kriged model residuals to the deterministic models.

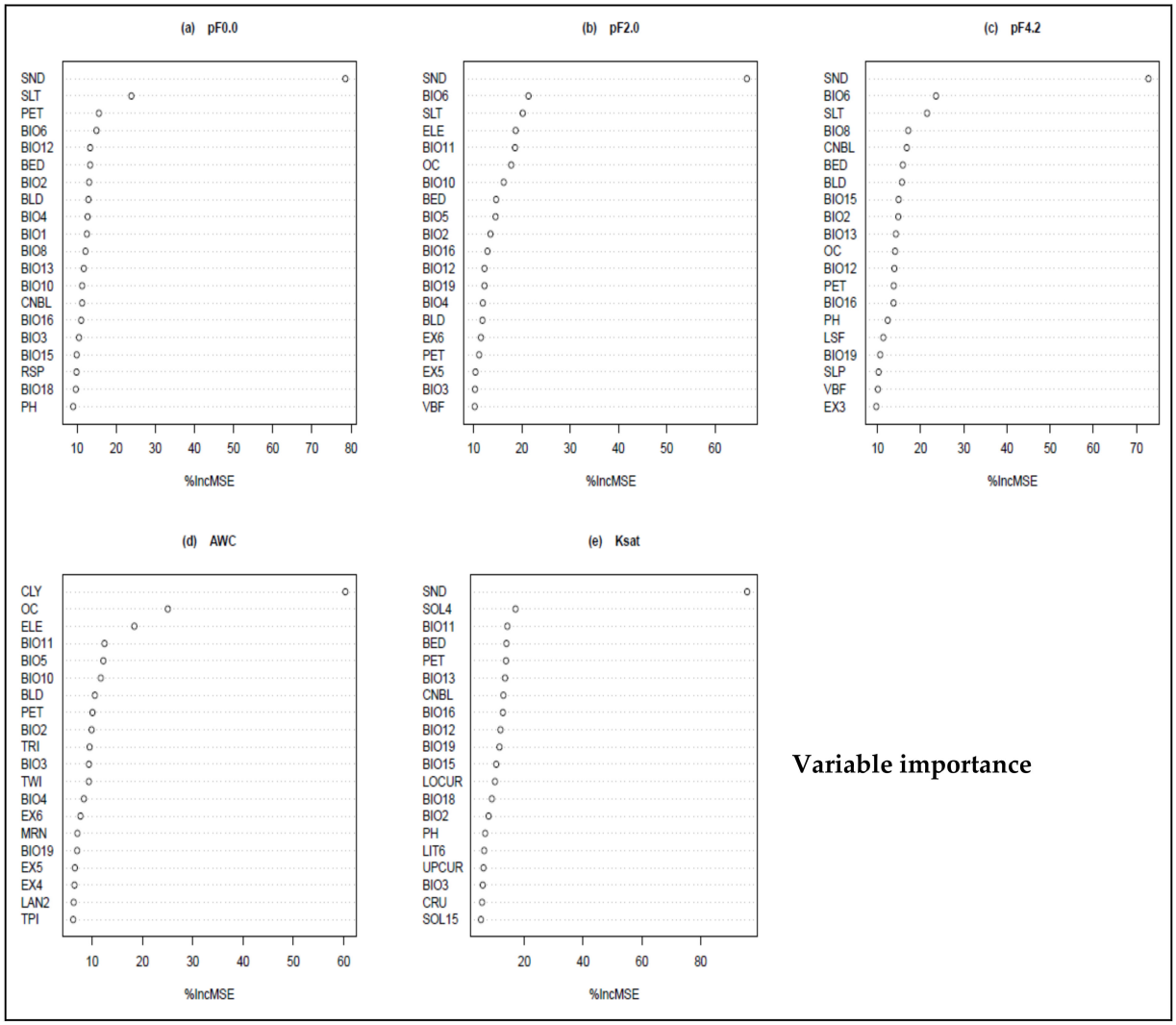

Figure 4. The 20 most important covariates for each of the 5 response variable expressed in terms of %IncMSE, as derived from a random forest model. Abbreviation of the covariates are given in Table 1. Panel (**a**) shows the covariates for response variable pF0.0, panel (**b**) for pF2.0, panel (**c**) for pF4.2, panel (**d**) for AWC and panel (**e**) for Ksat.

For the water content at pF0.0 and pF2.0, the R^2 ranged from 0.60 to 0.80 for all five models in both training and test data sets before adding the model residuals. After adding the model residuals, the R^2 went up to around 0.95 for all five models in training and test data for the water content at pF0.0 (Figure 6). For pF2.0, the R^2 after adding the random forest residuals went up to around 0.95. There was no spatial autocorrelation at pF2.0 for the other four models, while for the water content at pF4.2 and the AWC, the R^2 ranged from 0.40 to 0.60 for all five models in training and test data sets before adding the model residuals, while after adding the model residuals, the R^2 increased to about 0.70 to 0.85 in training and test data. No spatial autocorrelation was present for Ksat for any of the five models. The R^2 ranged from 0.40 to 0.75 for all five models in training and test data sets. Overall, for all of the response variables, and looking at the validation data sets, the R^2 was rather erratic, although it stands out for the RF model (Figure 6). Figures 6 and 8 display the RMSE and MAE showing high, therefore worse, values for the validation data sets and lower, hence better, values for the training and test data sets with model residuals. The lowest values for the RMSE and MAE in the training, test, and validation datasets for the RF model depict a better performance of this model than the other four models.

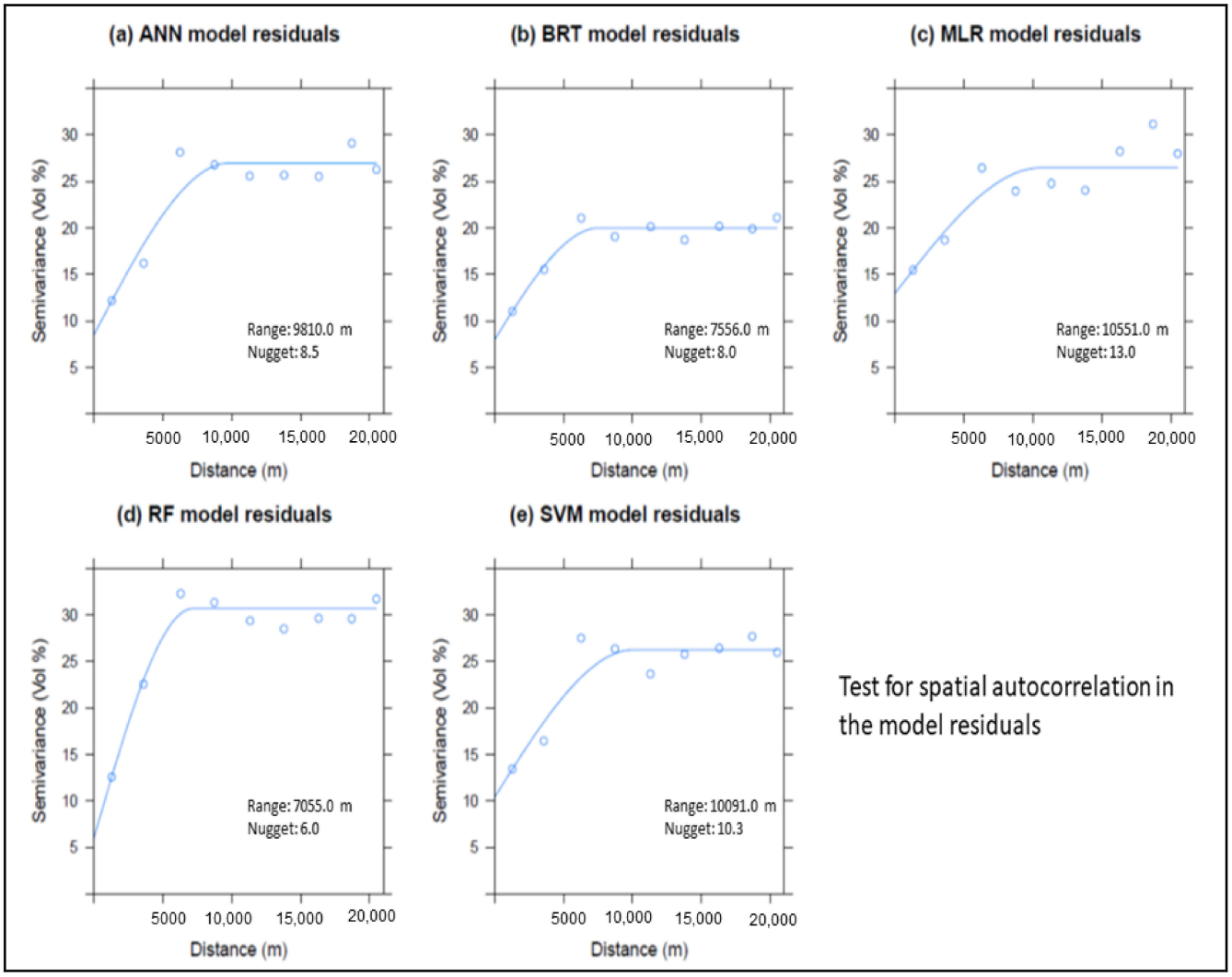

Figure 5. Spatial autocorrelation observed in the model residuals for the water content at pF0.0 in the topsoil (0–30 cm) depth layers. (**a**) ANN, artificial neural network; (**b**) BRT, gradient boosted regression trees; (**c**) MLR, multiple linear regression; (**d**) RF, random forest; (**e**) SVM, support vector machine.

3.4. Digital Soil Maps

Since the random forest method complemented with residual kriging proved to have the best predictive power for all considered hydraulic properties, we used this approach to elaborate maps of estimates of the water content at pF0.0, pF2.0, pF4.2, and AWC at a depth of 30 cm and with a spatial resolution of 90 m (Figure 9) for the whole ZRB. Since the saturated hydraulic conductivity residuals of the random forest model were not spatially autocorrelated, the estimates of the saturated hydraulic conductivity were mapped based solely on the random forest deterministic model. The resulting digital maps show that water content at pF0.0, pF2.0, pF4.2, and the AWC have lower values in the southwestern part of the basin, and higher values in the north central and southeastern part of the ZRB. On the contrary, the saturated hydraulic conductivity has higher values in the southwestern part of the basin, and lower values in the north central and southeastern part of the basin (Figure 9).

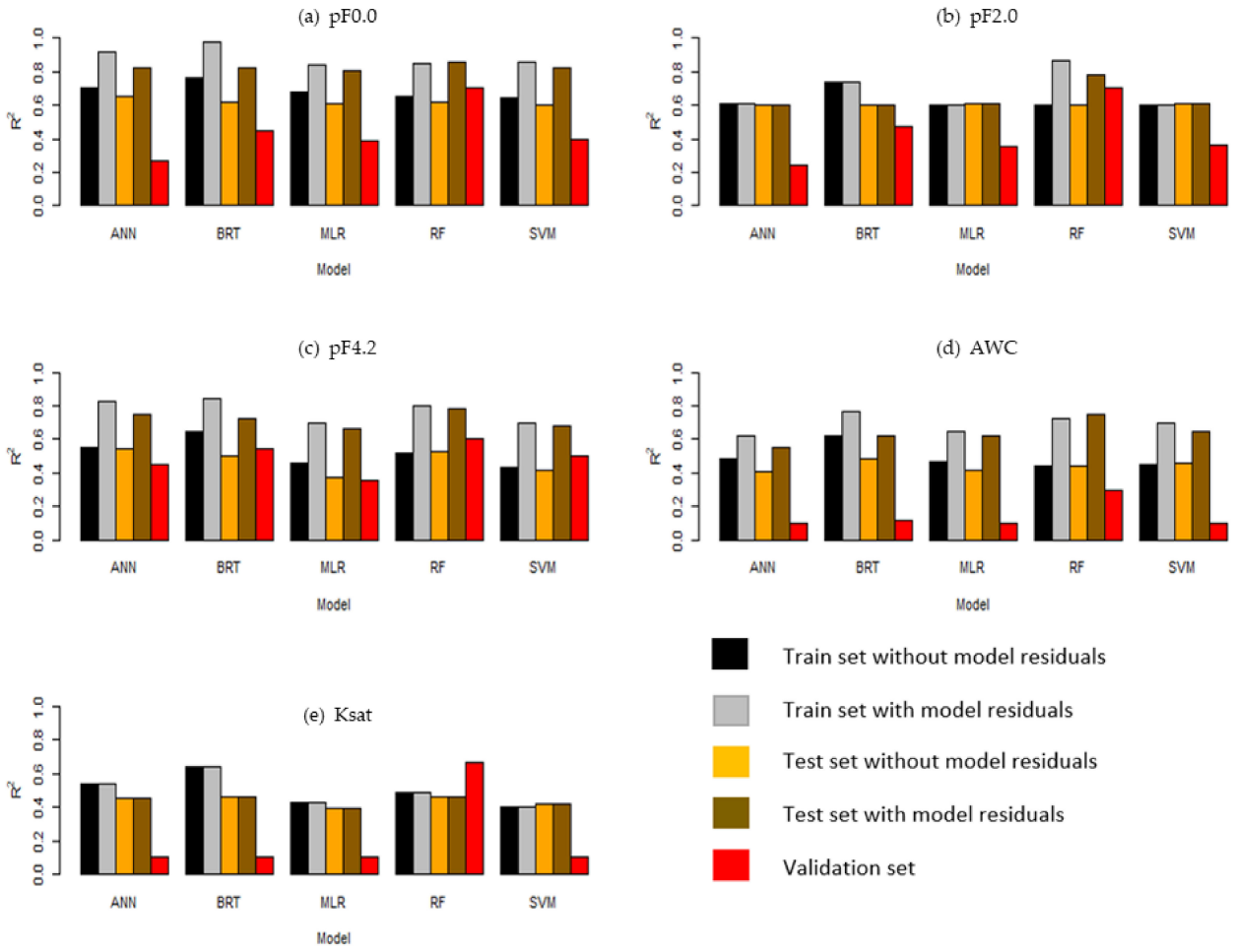

Figure 6. Coefficients of determination (R^2) for the five DSM models applied to the training sets without residuals (black), training sets with residuals (grey), test sets without residuals (orange), test sets with residuals (brown), and the validation data sets (red). ANN, artificial neural network; BRT, gradient boosted regression trees; MLR, multiple linear regression; RF, random forest; SVM, support vector machine. Panel (**a**) shows the response variable pF0.0, panel (**b**) for pF2.0, panel (**c**) for pF4.2, panel (**d**) for AWC and panel (**e**) for Ksat at a depth of 30 cm.

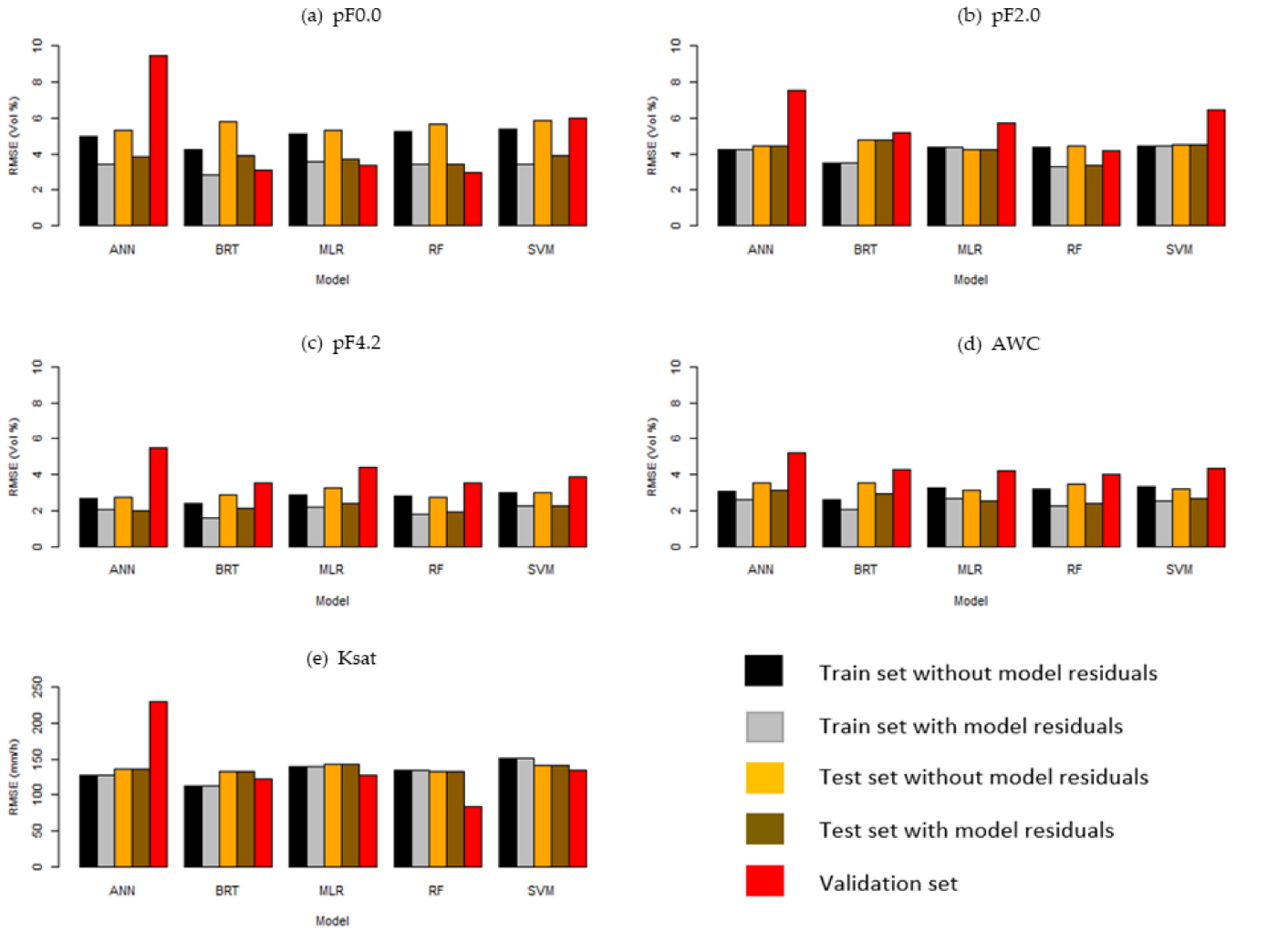

Figure 7. Root mean squared error (RMSE) for the five DSM models applied to the training sets without residuals (black), training sets with residuals (grey), test sets without residuals (orange), test sets with residuals (brown), and the validation data sets (red). ANN, artificial neural network; BRT, gradient boosted regression trees; MLR, multiple linear regression; RF, random forest; SVM, support vector machine. Panel (**a**) shows the response variable pF0.0, panel (**b**) for pF2.0, panel (**c**) for pF4.2, panel (**d**) for AWC and panel (**e**) for Ksat at a depth of 30 cm.

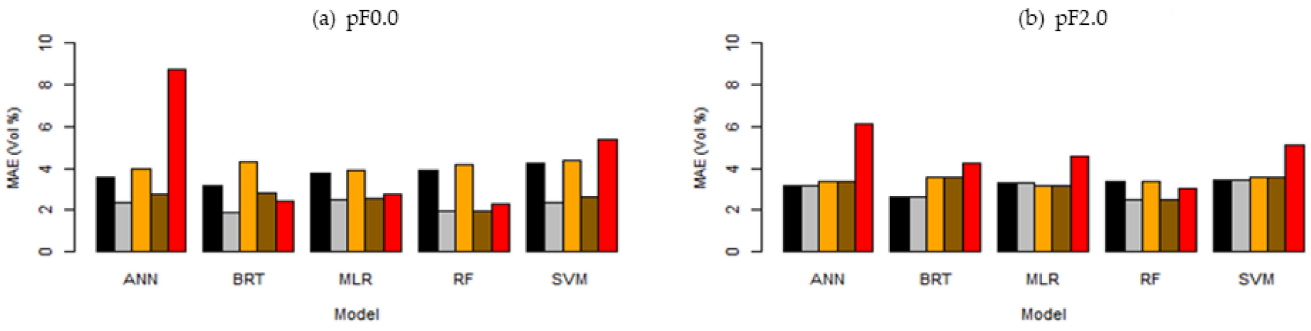

Figure 8. *Cont.*

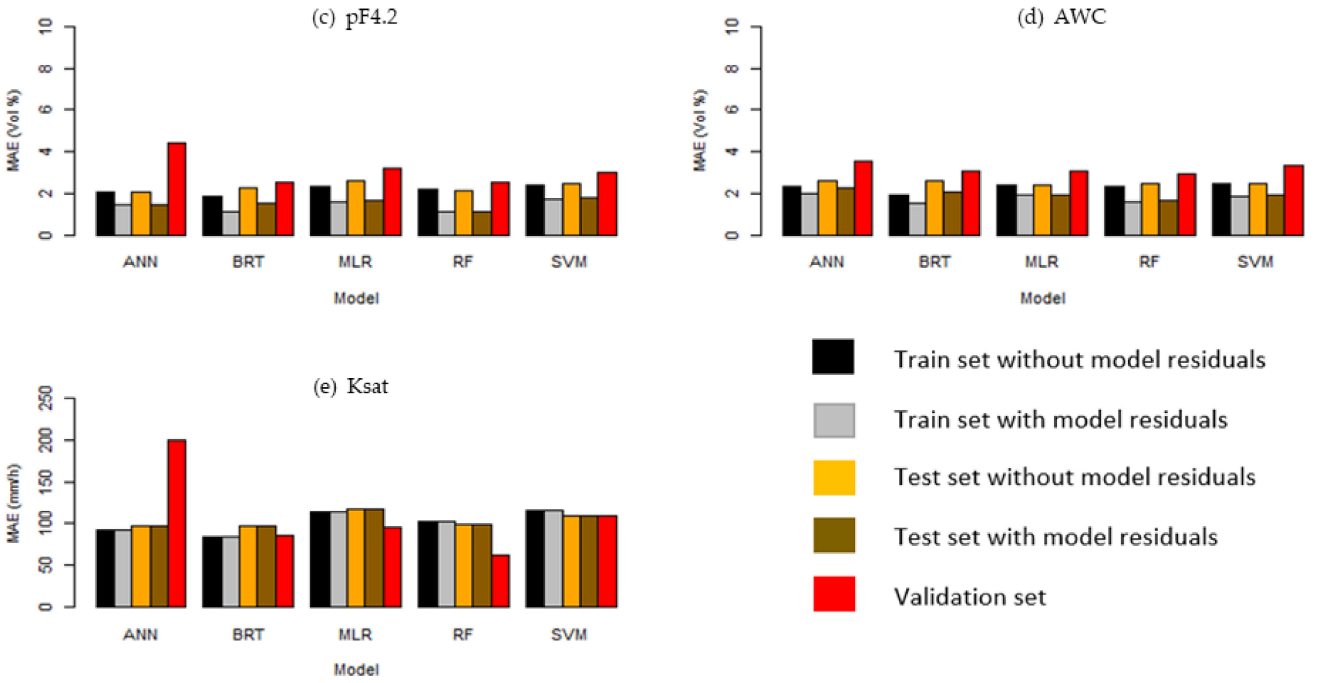

Figure 8. Mean absolute error (MAE) for the five DSM models applied to the training sets without residuals (black), training sets with residuals (grey), test sets without residuals (orange), test sets with residuals (brown), and the validation data sets (red). ANN, artificial neural network; BRT, gradient boosted regression trees; MLR, multiple linear regression; RF, random forest; SVM, support vector machine. Panel (**a**) shows the response variable pF0.0, panel (**b**) for pF2.0, panel (**c**) for pF4.2, panel (**d**) for AWC and panel (**e**) for Ksat at a depth of 30 cm.

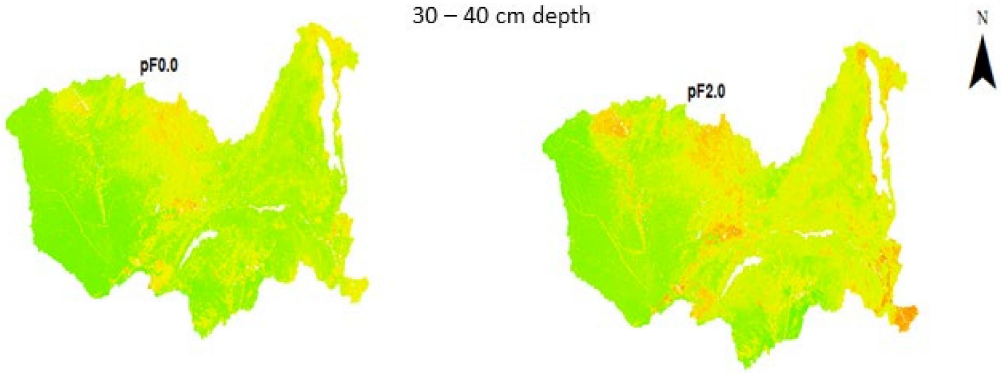

Figure 9. *Cont.*

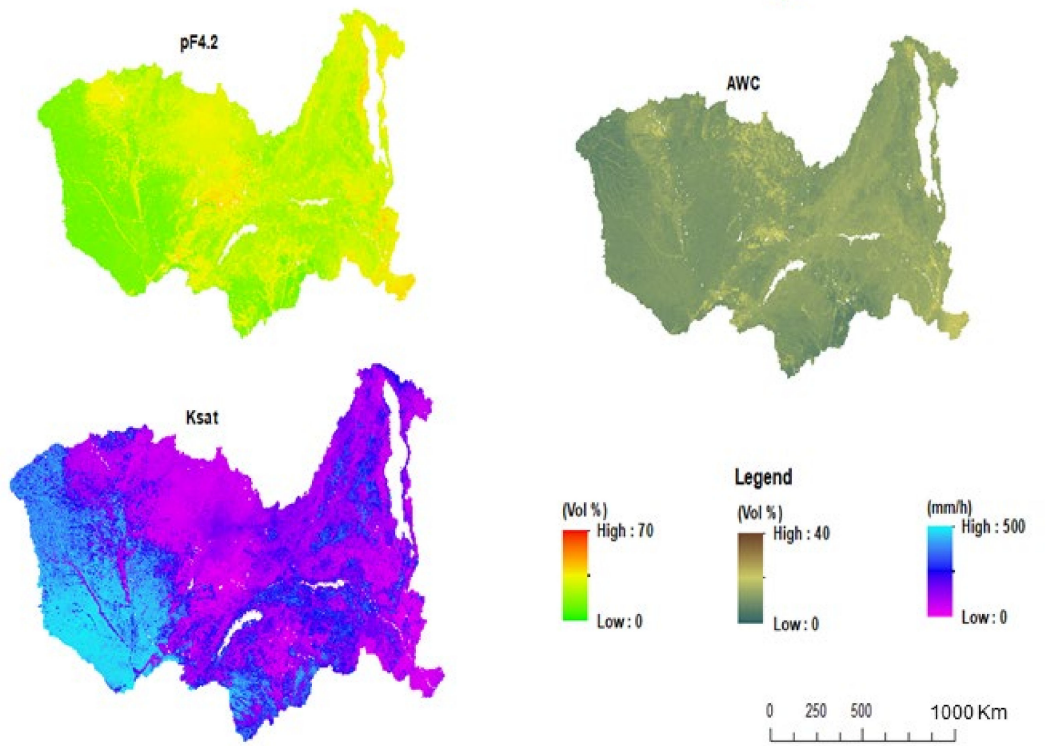

Figure 9. Soil hydraulic properties at 30 cm depth for the Zambezi River Basin. The data has a spatial resolution of 90 m, and the maps were generated using a random forest regression model based on legacy soil data, complemented with own field data, and environmental covariates reflecting soil classes, climate, landcover, relief, and lithology.

4. Discussion

According to [12], kriging of model residuals improves overall DSM performance. This was also confirmed in this study, where we observed that kriging of the deterministic model residuals considerably improves the performance of the DSM. We observed that the R^2 increased on average by more than 40%. Other studies dealing with soil hydraulic properties, such as SoilGrids [29], also found the RF model (overall average R^2 = 0.61) to perform better than other machine learning techniques for estimating water content at pF0.0 and the AWC. Ref. [33] also found that RF as a deterministic model complemented with residual kriging performed better for estimating water content at pF2.0 and pF4.2, with a RMSE ranging on average between 2 to 7 (Vol %), which was also the range found in this study. Still, in our study, the addition of the kriged model residuals to the deterministic model estimations led to a higher increase of the R^2, as compared to previous studies.

Although the relative performances were rather erratic in the validation dataset (dataset #2 and dataset #3) of Figures 6–8, the RF model was still better than the other models. This rather erratic performance on the validation dataset could be related to the relatively small size (n = 119, dataset #2) of the validation set coming from a relatively smaller sub-basin (2426 km^2), which was less heterogeneous than the whole Zambezi River Basin (1.6 million km^2), represented by a sample size (n = 1481, dataset #1) used for the training and test datasets. Furthermore, dataset #1 was already a smoothened dataset of point location estimates generated through artificial neural network-based pedotransfer functions (ANN-PTFs) [11], while datasets #2 and #3 encompass measured soil hydraulic property values obtained from Upper Mulungushi subbasin and the African Soil Profile database, whereby the sampling and measurement techniques were comparable but not completely identical between the datasets. Although we did not perform an error propagation analysis and uncertainty quantification in this study, we realize that there are error propagations and uncertainties starting from our three datasets all the way through to our models and modelling techniques.

Sand, silt, clay content, and soil organic carbon content as soil environmental covariates, climate variables such as the average annual rainfall, the average annual reference evapotranspiration (ETo), and the mean temperature of the coldest quarter, as well as topographic elevation, are all strong predictors of water content at pF0.0, pF2.0, pF4.2, the AWC, and Ksat. Moreover, the sand, silt, clay content, and soil organic carbon content layers are obtained from the SoilGrids databases, hence are themselves the result of machine learning based DSM. Inevitably, they contribute to error propagations to our estimated soil hydraulic properties.

The water retention or water content at pF0.0, pF2.0, pF4.2, and the AWC have predominately lower values and higher Ksat values in the southwestern part of the basin, where sandy soils are dominant, where lower rainfall and slightly higher evapotranspiration and temperature also prevail. In contrast, higher values for the water content at pF0.0, pF2.0, pF4.2, and the AWC, as well as lower Ksat values, are obtained in the north central and southeastern part of the basin, characterized by slightly higher rainfall, potential evapotranspiration, and temperature. Furthermore, soils with high clay, silt, and soil organic carbon content are dominant in valley bottoms, wetland areas, and the delta region, which are soils with high values of the water retention. The highest Ksat values are mostly found in the Arenosols and Podzols soil groups in the southwestern region of the basin, which also have lower clay, silt, and soil organic carbon concentrations, as well as very high sand content. In the northcentral and southeastern parts of the basin, dominated by the soils of Alisols, Andosols, Fluvisols, Histosols, Umbrisols, Gleysols, and Vertisols which are predominately found in the valley bottoms, wetland or dambo areas, as well as in the delta region, high water retention and lower Ksat values are mostly associated with soils with higher clay, silt, and soil organic carbon concentrations, as well as much lower sand content.

5. Conclusions

In this paper, we evaluated the performance of 10 approaches for the digital mapping of 5 soil hydraulic properties throughout the Zambezi River Basin. The ten approaches consisted of multiple linear regression, artificial neural network, gradient boosted regression trees, random forest, and support vector machine as the deterministic component of DSM, combined or not combined with the kriging of the model residuals. A total of 67 freely available soil environmental covariates were considered, from which 20 potential covariates were selected to train, test, and validate each of the 10 approaches. The sand, silt, clay, and soil organic carbon content of the topsoil and average annual rainfall, average annual reference evapotranspiration, and mean temperature of the coldest quarter, as well as the elevation, were all pertinent predictors for each of the hydraulic properties.

Spatial autocorrelation of the model residuals was only observed in the random forest model for the water content at pF0.0, pF2.0, pF4.2, and the available water capacity. There was no spatial autocorrelation of the model residuals for any of the five models for the saturated hydraulic conductivity. Once spatial autocorrelation in the model residuals was confirmed for the random forest model, the kriged model residuals were added to the model estimations, and this addition resulted into a better performance of the deterministic models, whereby the R^2 was observed to improve substantially.

The overall best DSM approach was found to encompass random forest as a deterministic model, complemented with residual kriging. With this approach, we developed maps of soil hydraulic properties at a depth of 30 cm with a spatial resolution of 90 m for the whole Zambezi River Basin, with R^2 ranging from 0.40 to 0.80 in the training and test data sets before adding model residuals, and from 0.80 to 0.95 after adding model residuals. The random forest model was found to be the best performing model after comparison, for the depth of 30 cm, with multiple linear regression and three other ML techniques for water content at pF0.0, pF2.0, pF4.2, the available water capacity, and saturated hydraulic conductivity. The resulting maps can be used by hydrologists and agronomists, as well as other researchers and extension workers conducting various land performance or environmental impact studies in the Zambezi River Basin.

Author Contributions: M.K.: Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing—original draft; writing—review and editing. E.N.: Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing—review and editing. I.N.: Conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; writing—review and editing. S.D.: Conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; writing—review and editing. J.V.O.: Conceptualization; funding acquisition; methodology; project administration; resources; supervision; validation; writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by an IRO-PhD-scholarship (contract no Contract 000000091676) pro-vided by the University of Leuven. Part of the field work was funded by the Decision Analytic Framework 'DAFNE' EU H2020-project (grant no. 690268).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: This research was financed by an IRO-PhD-scholarship provided by the University of Leuven in Belgium (KU Leuven), and part of the fieldwork was funded by the 'DAFNE' EU H2020-project (grant no. 690268). A special word of thanks to Ine Rosier from the Division of Forest, Nature, and Landscape (KU Leuven) for her assistance in processing the data.

Conflicts of Interest: We can confirm that there is no conflict of interest for this manuscript.

References

1. FAO Challenges and Policies for the World Agricultural and Food Economy in the 2050 Perspective. In *Looking Ahead in World Food and Agriculture: Perspectives to 2050*; Piero, C., Sarris, A., Eds.; Office of Knowledge Exchange, Research and Extension, FAO: Rome, Italy, 2011; pp. 233–275.
2. Montzka, C.; Herbst, M.; Weihermüller, L.; Verhoef, A.; Vereecken, H. A Global Data Set of Soil Hydraulic Properties and Sub-Grid Variability of Soil Water Retention and Hydraulic Conductivity Curves. *Earth Syst. Sci. Data* **2017**, *9*, 529–543. [CrossRef]
3. Cai, X.; Altchenko, Y.; Chavula, G. The Zambezi River Basin: Water and Sustainable Development, Earthscan/IWMI Series on Major River Basins of the World. In *Availability and Use of Water Resources*; Lautze, J., Phiri, Z., Smakhtin, V., Saruchera, D., Eds.; Routledge: London, UK; Taylor & Francis Group: New York, NY, USA, 2017; pp. 7–28.
4. World Bank. *The Zambezi River Basin: A Multi-Sector Investment Opportunity Volume 2 Basin Development Scenarios*; The World Bank: Washington, DC, USA, 2010; pp. 1–106.
5. Beilfuss, R. *A Risky Climate for Southern African Hydro: Assessing Hydrological Risks and Consequences for Zambezi River Basin Dams*; International Rivers: Berkeley, CA, USA, 2012; pp. 1–40.
6. Sinclair, S.; Kleinschroth, F.; Koroleva, K.; Miranda, D.; Micotti, M.; Battista, G.; Hillen, R.; Giuliani, M.; Calamita, E.; Burlando, P. *A Decision-Analytic Framework to Explore the Water-Energy-Food NExus in Complex and Transboundary Water Resources Systems of Fast Growing Developing Countries. Integrated Model of the Wef Nexus*; EU H2020 Project Grant No. 690268; Politecnico di Milano: Milano, Italy, 2019; pp. 1–45.
7. Ciarapica, L.; Todini, E. TOPKAPI: A Model for the Representation of the Rainfall-Runoff Process at Different Scales. *Hydrol. Processes* **2002**, *16*, 207–229. [CrossRef]
8. Peng, D.; Zhijia, L.; Zhiyu, L. Numerical Algorithm of Distributed TOPKAPI Model and Its Application. *Water Sci. Eng.* **2008**, *1*, 14–21. [CrossRef]
9. Todini, E. The ARNO Rainfall-Runoff Model. *J. Hydrol.* **1996**, *175*, 339–382. [CrossRef]
10. Steduto, P.; Hsiao, T.C.; Raes, D.; Fereres, E. Aquacrop—The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles. *Agron. J.* **2009**, *101*, 426–437. [CrossRef]
11. Kalumba, M.; Bamps, B.; Nyambe, I.; Dondeyne, S.; Van Orshoven, J. Development and Functional Evaluation of Pedotransfer Functions for Soil Hydraulic Properties for the Zambezi River Basin. *Eur. J. Soil Sci.* **2020**, *72*, 1559–1574. [CrossRef]
12. McBratney, A.B.; Mendonça-Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
13. Maza León, A.P. *Comparing Digital Soil Mapping Techniques to Predict Soil Organic Carbon Content and Stock in the Lower Shire Valley of Malawi (Master Dissertation)*; University of Leuven: Leuven, Belgium, 2020.
14. Jenny, H. *Factors of Soil Formation. A System of Quantitative Pedology*; Dover Publications, Inc.: New York, NY, USA, 1941; ISBN 0486681289.
15. Nussbaum, M.; Spiess, K.; Baltensweiler, A.; Grob, U.; Keller, A.; Greiner, L. Evaluation of Digital Soil Mapping Approaches with Large Sets of Environmental Covariates. *Soil* **2018**, *4*, 1–22. [CrossRef]
16. Cressie, N.A. *Statistics for Spatial Data*; John Wiley & Sons, Inc.: New York, NY, USA, 1991.

17. Isaaks, E.H.; Srivastava, R. *An Introduction to Applied Geostatistics*; Oxford University Press: Oxford, UK, 1989.
18. Sindayihebura, A.; Ottoy, S.; Dondeyne, S.; Van Meirvenne, M.; Orshoven, J. Van Comparing Digital Soil Mapping Techniques for Organic Carbon and Clay Content: Case Study in Burundi's Central Plateaus. *CATENA* **2017**, *156*, 161–175. [CrossRef]
19. Matheron, G. Principals of Geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
20. Hengl, T.; Heuvelink, G.B.M.; Stein, A. A Generic Framework for Spatial Prediction of Soil Variables Based on Regression-Kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
21. Goovaerts, P. *Geostatistics for Natural Resources Evaluation*; Oxford University Press: Oxford, UK, 1997.
22. Aitkenhead, M.J.; Coull, M.C. Mapping Soil Carbon Stocks across Scotland Using a Neural Network Model. *Geoderma* **2016**, *262*, 187–198. [CrossRef]
23. Khaledian, Y.; Miller, B.A. Selecting Appropriate Machine Learning Methods for Digital Soil Mapping. *Appl. Math. Model.* **2020**, *81*, 401–418. [CrossRef]
24. Minasny, B.; McBratney, A.B. Digital Soil Mapping: A Brief History and Some Lessons. *Geoderma* **2016**, *264*, 301–311. [CrossRef]
25. Wadoux, A.M.J.C.; Minasny, B.; McBratney, A.B. Machine Learning for Digital Soil Mapping: Applications, Challenges and Suggested Solutions. *Earth-Sci. Rev.* **2020**, *210*, 1–17. [CrossRef]
26. Derin, Y.; Yilmaz, K.K. Evaluation of Multiple Satellite-Based Precipitation Products over Complex Topography. *J. Hydrometeorol.* **2014**, *15*, 1498–1516. [CrossRef]
27. Mei, Y.; Nikolopoulos, E.I.; Anagnostou, E.N.; Borga, M. Evaluating Satellite Precipitation Error Propagation in Runoff Simulations of Mountainous Basins. *J. Hydrometeorol.* **2016**, *17*, 1407–1423. [CrossRef]
28. Khan, R.S.; Bhuiyan, M.A.E. Artificial Intelligence-Based Techniques for Rainfall Estimation Integrating Multisource Precipitation Datasets. *Atmosphere* **2021**, *12*, 1239. [CrossRef]
29. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.M.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]
30. Arrouays, D.; Lagacherie, P.; Hartemink, A.E. Digital Soil Mapping across the Globe. *Geoderma Reg.* **2017**, *9*, 1–4. [CrossRef]
31. Minasny, B.; McBratney, A.B.; Mckenzie, N.J.; Grundy, M.J. Predicting Soil Properties Using Pedotransfer Functions and Environmental Correlation. In *Guidelines for Surveying Soil and land Resources*; McKenzie, N.J., Grundy, M.J., Webster, R., Ringrose-Voase, A.J., Eds.; CSIRO Publishing: Collingwood, Australia, 2008; pp. 349–367.
32. Malone, B.P.; McBratney, A.B.; Minasny, B.; Laslett, G.M. Mapping Continuous Depth Functions of Soil Carbon Storage and Available Water Capacity. *Geoderma* **2009**, *154*, 138–152. [CrossRef]
33. Szabó, B.; Szatmári, G.; Takács, K.; Laborczi, A.; Makó, A.; Rajkai, K. Mapping Soil Hydraulic Properties Using Random-Forest-Based Pedotransfer Functions and Geostatistics. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 2615–2635. [CrossRef]
34. World Bank. *The Zambezi River Basin: A Multi-Sector Investment Opportunities Analysis. Volume 3, State of the Basin*; The World Bank: Washington, DC, USA, 2010; pp. 1–106.
35. Nugent, C. The Zambezi River: Tectonism, Climatic Change and Drainage Evolution. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **1990**, *78*, 55–69. [CrossRef]
36. Key, R.M.; Cotterill, F.P.D.; Moore, A.E. The Zambezi River: An Archive of Tectonic Events Linked to the Amalgamation and Disruption of Gondwana and Subsequent Evolution of The African Plate. *S. Afr. J. Geol.* **2015**, *118*, 425–438. [CrossRef]
37. Nyambe, I.A.; Dixon, O. Sedimentology of the Madumabisa Mudstone Formation (Late Permian), Lower Karoo Group, Mid-Zambezi Valley Basin, Southern Zambia. *J. Afr. Earth Sci.* **2000**, *30*, 535–553. [CrossRef]
38. Leenaers, H. *Estimating the Impact of Land Use Change on Soil Erosion Hazard in the Zambezi River Basin*; 90-024; IIASA: Laxenburg, Austria, 1990.
39. Batjes, N.H.; Ribeiro, E.; Oostrum, A. Van Standardised Soil Profile Data to Support Global Mapping and Modelling (WoSIS Snapshot 2019). *Earth Syst. Sci. Data* **2019**, 1–46. [CrossRef]
40. Leenaars, J.G.B.; van Oostrum, A.J.; Gonzalez, M.R. *Africa Soil Profiles Database, Version 1.2. A Compilation of Georeferenced and Standardised Legacy Soil Profile Data for Sub-Saharan Africa (with Dataset). ISRIC Report 2014/01. Africa Soil Information Service (AfSIS) Project and ISRIC—World Soil Inform*; ISRIC—World Soil Information: Wageningen, The Netherlands, 2014; pp. 1–166.
41. Samuel-Rosa, A.; Heuvelink, G.B.M.; Vasques, G.M.; Anjos, L.H.C. Do More Detailed Environmental Covariates Deliver More Accurate Soil Maps? *Geoderma* **2015**, *243–244*, 214–227. [CrossRef]
42. Baboo, D.S.S.; Devi, M.R. An Analysis of Different Resampling Methods in Coimbatore, District. *Glob. J. Comput. Sci. Technol.* **2010**, *10*, 61–66.
43. Dewitte, O.; Jones, A.; Spaargaren, O.; Breuning-Madsen, H.; Brossard, M.; Dampha, A.; Deckers, J.; Gallali, T.; Hallett, S.; Jones, R.; et al. Harmonisation of the Soil Map of Africa at the Continental Scale. *Geoderma* **2013**, *211–212*, 138–153. [CrossRef]
44. Dijkshoorn, J.A. *SOTER Database for Southern Africa (SOTERAF)*; ISRIC—World Soil Information: Wageningen, The Netherlands, 2003.
45. Dijkshoorn, J.A.; Huting, J.; Kempen, B. *Soil and Terrain Database of the Republic of Malawi*; Report 2016/01; ISRIC—World Soil Information: Wageningen, The Netherlands, 2016.
46. IUSS Working Group WRB. *World Reference Base for Soil Resources 2006, First Update 2007*; FAO: Rome, Italy, 2007.
47. Boettinger, J.L.; Ramsey, R.D.; Bodily, J.M.; Cole, N.J.; Kienast-Brown, S.; Nield, S.J.; Saunders, A.M.; Stum, A.K. Landsat Spectral Data for Digital Soil Mapping. In *Digital Soil Mapping with Limited Data*; Hartemink, A.E., McBratney, A., Mendonça-Santos, M.D.L., Eds.; Springer: New York, NY, USA, 2008; pp. 193–202.

48. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
49. Savtchenko, A.; Ouzounov, D.; Ahmad, S.; Acker, J.; Leptoukh, G.; Koziana, J.; Nickless, D. Terra and Aqua MODIS Products Available from NASA GES DAAC. *Adv. Sp. Res.* **2004**, *34*, 710–714. [CrossRef]
50. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **2013**, *342*, 850–853. [CrossRef]
51. Conrad, O.; Bechtel, B.; Bock, M.; Dietrich, H.; Fischer, E.; Gerlitz, L.; Wehberg, J.; Wichmann, V.; Böhner, J. System for Automated Geoscientific Analyses (SAGA) v. 2.1.4. *Geosci. Model Dev.* **2015**, *8*, 2271–2312. [CrossRef]
52. Planchon, O.; Darboux, F. A Fast, Simple and Versatile Algorithm to Fill the Depressions of Digital Elevation Models. *CATENA* **2002**, *46*, 159–176. [CrossRef]
53. Poggio, L.; De Sousa, L.M.; Batjes, N.H.; Heuvelink, G.B.M.; Kempen, B.; Ribeiro, E.; Rossiter, D. SoilGrids 2.0: Producing Soil Information for the Globe with Quantified Spatial Uncertainty. *Soil* **2021**, *7*, 217–240. [CrossRef]
54. Zomer, R.J.; Trabucco, A.; Bossio, D.A.; Verchot, L.V. Climate Change Mitigation: A Spatial Analysis of Global Land Suitability for Clean Development Mechanism Afforestation and Reforestation. *Agric. Ecosyst. Environ.* **2008**, *126*, 67–80. [CrossRef]
55. CCI Land Cover (LC) Team. *CCI Land Cover—S2 Prototype Land Cover 20 m Map of Africa*. 2016. Available online: https://2016africalandcover20m.esrin.esa.int/ (accessed on 6 March 2022).
56. Didan, K. *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V 006*; 2015. Available online: https://lpdaac.usgs.gov/products/mod13q1v006/ (accessed on 6 March 2022).
57. USGS NASA Shuttle Radar Topography Mission (SRTM) Global 1 Arc Second Dataset (SRTMGL1), Digital Elevation. Available online: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1 (accessed on 6 March 2022).
58. Earth Resources Observation and Science (EROS) Center. *Shuttle Radar Topography Mis Sion (SRTM) Void Filled*; 2017. Available online: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-void (accessed on 6 March 2022).
59. U.S. Geological Survey, T.N.C. *Africa Surficial Lithology*; 2009. Available online: http://geoportal.rcmrd.org/layers/servir%3Aafrica_surface_lethology (accessed on 6 March 2022).
60. R Core Team. *A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
61. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; ISBN 978-1-4614-6848-6.
62. Pebesma, E.J. Multivariable Geostatistics in S: The Gstat Package. *Comput. Geosci.* **2004**, *30*, 683–691. [CrossRef]
63. Chen, L.; Ren, C.; Li, L.; Wang, Y.; Zhang, B.; Wang, Z.; Li, L. A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid Approaches for Mapping Topsoil Organic Carbon Content. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 174. [CrossRef]
64. Sun, W.; Minasny, B.; McBratney, A. Analysis and Prediction of Soil Properties Using Local Regression-Kriging. *Geoderma* **2012**, *171–172*, 16–23. [CrossRef]
65. Yao, X.; Sun, F.; Wang, S.; Liu, M.; Fu, B.; Lu, Y. Comparison of Four Spatial Interpolation Methods for Estimating Soil Moisture in a Complex Terrain Catchment. *PLoS ONE* **2013**, *8*, e54660. [CrossRef]