Water SA 52(3) 242–253 / Jul 2026
https://doi.org/10.17159/wsa/2026.v52.i3.4220

Research paper

Modelling soil water content in different tillage systems and soil types using machine learning

W Mupangwa[1,2], L Chipindu[3,4], B Ncube[5] and TP Tauro[6,7]

[1]*International Water Management Institute, 127 Sunil Mawatha, Battaramulla, Colombo, Sri Lanka.*
[2]*Marondera University of Agricultural Sciences and Technology, Centre for Research and Postgraduate Studies, P.O. Box 35, Marondera, Zimbabwe*
[3]*International Maize and Wheat Improvement Centre, P.O. Box MP 163, Mount Pleasant, Harare, Zimbabwe*
[4]*School of Mathematical Sciences, College of Science, Engineering and Food Science, University College Cork, Ireland*
[5]*Centre for Water and Sanitation Research, Faculty of Engineering and the Built Environment, Cape Peninsula University of Technology, P.O. Box 1906, Bellville 7535, Cape Town, South Africa*
[6]*Department of Soil Science and Productivity, Marondera University of Agricultural Sciences and Technology, P.O. Box 35, Marondera, Zimbabwe*
[7]*Department of Agriculture and Environmental Sciences, College of Health, Agriculture and Natural Sciences, Africa University, P.O. Box 1320, Mutare, Zimbabwe*

Soil water availability is one of the major challenges in many rainfed crop production systems of the Global South. Soil water conservation practices are being promoted to enhance climate change adaptation for rainfed cropping systems of southern Africa. However, the cost and time required to develop and test appropriate modelling and simulation tools can be enormous. The objectives of this study were to: (i) test the performance of the decision tree, adaptive boosting (AdaBoost), support vector machine, neural network, stochastic gradient descent, k-nearest neighbours, random forest and linear regression machine learning models in predicting soil water under different tillage practices, soil types and depths, and (ii) assess the soil water classification and prediction capabilities of 8 models under different tillage practices, soil types and depths. The neural network, random forest and decision tree models had the best soil water prediction capabilities. The neural network, random forest and decision tree models were the best algorithms (RMSE = 15.801–16.369; MAE = 11.997–12.315; R^2 = 0.822–0.835) for predicting and classifying soil water from different soil types and depth intervals. The support vector machine learning model was the weakest algorithm (RMSE = 36.177; MAE = 30.84; R^2 = 0.133) for predicting and classifying soil water. All the algorithms poorly predicted and classified soil water based on tillage practices. All the models closely predicted soil water at 300 and 900 mm depths but poorly predicted soil water at 600 mm depth intervals. Based on this study, the neural network model is the best machine learning tool for predicting soil water in clay and sandy soils under semi-arid agroecological conditions.

CORRESPONDENCE
W Mupangwa

EMAIL
mupangwawt@gmail.com

DATES
Received: 16 May 2025
Accepted: 16 June 2026

KEYWORDS
algorithm
basins
clay
mulch
ripper
sand

INTRODUCTION

Smallholder agriculture heavily depends on total seasonal rainfall and its distribution in semi-arid sub-Saharan Africa. Annual rainfall in semi-arid south-western Zimbabwe is often less than 500 mm and exhibits extreme spatial and temporal variations (Sibanda et al., 2020). Rainfall occurs as short-duration and high-intensity convective storms that generate high surface runoff, resulting in limited recharging of the soil profiles on smallholder farms (Mamombe et al., 2017). Frequent intra-seasonal dry spells and drought make rainfed crop production risky on smallholder farms. Consequently, soil water availability in rainfed cropping systems is one of the major challenges in the semi-arid parts of southern Africa. Research has been undertaken to assess the impact of different practices on soil water availability to crops (Nyagumbo et al., 2015; Thierfelder et al., 2018) to find ways to minimise the impacts of frequent droughts and dry spells on smallholder farming. Key findings from these studies indicate that soil water availability can be prolonged in different soil types during the growing season by using conservation agriculture and reduced tillage practices (Thierfelder and Wall, 2009; Madembo and Nyagumbo, 2026).

The use of process-based modelling tools helps in understanding the long-term impact of different practices, including soil tillage and mulching, on soil water dynamics in cropping systems (Keating et al., 2003; Vogeler et al., 2023). Biophysical models such as the Agricultural Production Systems Simulator (APSIM) (Keating et al., 2003) have been used for simulating soil water patterns and surface runoff in agricultural systems (Ncube et al., 2009; Mupangwa and Jewitt, 2011; Vogeler et al., 2023). However, the capability of process-based models in predicting soil water balance in agricultural systems under different management practices and varying rainfall patterns gives mixed results (Chapagain et al., 2022). A study by Ncube et al. (2009) revealed that the APSIM model could closely predict soil water regimes in legume-cereal rotations on sandy soil under semi-arid conditions. Mupangwa et al. (2007; 2011) indicated that soil water prediction precision could vary with seasonal rainfall amounts, distribution, and depth in clay and sandy soils. In seasons with above-average rainfall, the APSIM model over-predicted soil water content in different tillage practices.

Biophysical models such as the Decision Support System for Agrotechnology Transfer (DSSAT) (Jones et al., 1998) do not accurately predict soil water during high-rainfall seasons (Mkuhlani, 2013). On the other hand, such models predict soil water with considerable accuracy during average rainfall seasons and with reduced accuracy in dry seasons (Mupangwa et al., 2011). The DSSAT model had considerable inaccuracies in predicting soil water at different depths, but the inaccuracy increased with an increase in soil depth (Mkuhlani, 2013).

The shortcomings of mechanistic biophysical tools such as APSIM and DSSAT can be reduced by using machine learning (ML) approaches to predict soil water under different agricultural water management practices (Padarian et al., 2020; de Oliveira et al., 2020). Machine learning tools are good for predicting the non-linear nature of soil water dynamics (Padarian et al., 2020). Furthermore, machine learning tools have fewer data requirements than the current mechanistic biophysical models (Young, 2006; King and Shellie, 2016). When there is limited data for the calibration and validation of the biophysical models (Wolfert et al., 2017), ML tools can be an alternative (Mishra et al., 2016; Crane-Droesch, 2018; Liakos et al., 2018) for learning complex relationships and trends (Gorni and Augusto, 2008). Machine learning comprises algorithms that allow software applications to learn and predict patterns and outcomes from systems of research interest (Crane-Droesch, 2018). The basic premise of ML is to build algorithms that can receive input data and use statistical analysis to predict an output. The advantages of ML tools include extracting more knowledge and identifying trends from big data sets (Wolfert et al., 2017). Additionally, ML algorithms can continuously improve as new data becomes available, increasing prediction efficiency (Balducci et al., 2018).

Different ML algorithms have been used to inform soil water management in regions outside southern Africa (Padarian et al., 2020; de Oliveira et al., 2021; Rani et al., 2022; Teshome et al., 2024). Machine learning techniques available for use in agriculture include regression, fuzzy cognitive map learning, artificial neural networks, classification and regression trees (CART), k-nearest neighbours (KNN), random forest and support vector machine (SVM) (Rosenblatt, 1958; Silverman and Jones, 1989; Dickerson and Kosko, 1993; Cortes and Vapnik, 1995; Breiman, 2001). Results from applications of ML algorithms for soil water prediction under different conditions are often mixed (Cai et al., 2019). Hou et al. (2016) applied the artificial neural network, and the tool closely predicted soil water variations with depth. Where the SVM was applied, soil water prediction over time had an 89% accuracy compared with the measured data (Gill et al., 2006). Machine learning tools such as fuzzy decision systems and neural networks have been applied in irrigation water management with satisfactory results (Giusti and Marsili-Libelli, 2015; King and Shellie, 2016). The prediction accuracy of ML algorithms is critical for their usefulness in agriculture. The different ML tools vary in accuracy levels, and selecting a tool to use is a crucial first step. The performance of a selected ML tool depends on various aspects, including the size of the data set used, experimental noise, and errors in data collection procedures, among others (Enke and Mehdiyev, 2012).

This study sought to answer the following research questions:

- Can the ML algorithms be used for soil water prediction in conventional and reduced tillage practices and clay and sandy soils at different soil depths?
- Which is the best ML algorithm that can be used for soil water prediction?
- How accurate and precise are the ML approaches in classifying and predicting soil water under different tillage practices, soil types and depths?

The objectives of the study were to: (i) evaluate the performance of the decision tree, adaptive boosting (AdaBoost), support vector machine, neural network, stochastic gradient descent, k-nearest neighbours, random forest and linear regression in predicting soil water under different tillage practices, soil types and depths, and (ii) evaluate the soil water prediction capabilities of decision tree, AdaBoost, support vector machine, neural network, stochastic gradient descent, k-nearest neighbours, random forest and linear regression in different tillage practices, soil types and depths.

MATERIALS AND METHODS

Description of experimental sites and simulated treatments

Machine learning and algorithmic predictions of soil water were based on experiments conducted at the Matopos Research Station in western Zimbabwe. Specifically, the experiments were performed at Westacre and Lucydale farms within the Matopos Research Station, where clay (28°30.92’ E, 20°23.32’ S, 1 344 m above sea level) and granitic sandy (28°24.46’ E, 20°25.64’ S, 1 378 m above sea level) soils were used, respectively. Matopos typically receives an average rainfall of 590 mm (Ncube et al., 2009), and the rainfall season is unimodal. The experiment consisted of 3 tillage methods as the main plot treatments: conventional mouldboard ploughing, animal traction ripping (ripper), and hand-dug planting basins. The sub-plot treatments were 4 mulch levels, namely, 0, 2, 4 and 10 $t \cdot ha^{-1}$. The planting basins were created with a hand hoe, spaced at 0.6 m × 0.9 m, and each basin had dimensions of 0.15 m (length) × 0.15 m (width) × 0.15 m (depth) (Twomlow et al., 2008). The rip lines were opened with a ripper tine attached to the beam of a donkey-drawn mouldboard plough, and the inter-row spacing was 0.9 m. The ripping depth achieved on both soil types ranged from 0.15–0.18 m. Conventional ploughing, to a depth of 0.15–0.20 m, was performed soon after the first effective rain (at least 20 to 30 mm) in December each year, using an animal traction mouldboard plough. Mulch cover was applied annually to the conventional, ripper, and planting basin tillage treatments at 0, 2, 4, and 10 $t \cdot ha^{-1}$. Soil water measurements using a portable capacitance probe (Decagon Devices, Inc.) were taken weekly at soil depths of 0–0.3 m, 0.3–0.6 m and 0.6–0.9 m during the growing season (November–March) each year. For more detailed information on the experiments, refer to studies by Mupangwa et al. (2007; 2012). Maize was grown at 37 000 plants per hectare, while sorghum and cowpea had 55 555 plants per hectare. Further details of the cropping systems and management practices used are described by Mupangwa et al. (2012).

Machine learning algorithms

Predicting soil water content is important in agriculture, hydrology and environmental sciences (Legates and McCabe, 1999; Sharma and Bharti, 2020). Machine learning offers powerful tools for tackling this complex task, due to its ability to handle non-linear relationships and complex data patterns. This study applied 8 ML algorithms for predicting soil water in 3 tillage systems and 2 soil types. The ML algorithms were decision tree, AdaBoost, support vector machine, neural network, stochastic gradient descent, k-nearest neighbours, random forest and linear regression (Ravi and Bhatia, 2021). The choice of models was influenced by their capabilities to predict soil properties such as water content (Teshome et al., 2024). Machine learning algorithms that balance several key features – high accuracy in predicting soil water, the ability to understand the factors influencing predictions (interpretability), and the efficiency to run and generate results quickly (computational efficiency) – were selected. Machine learning algorithms that can be trained and generate results quickly with the available data sets were prioritised.

These considerations were crucial given the characteristics of the data, including its size, complexity and the potential presence of non-linear relationships between variables. A detailed exposition of the algorithms' mathematical underpinnings can be found within the existing body of published research (Rosenblatt, 1958; Silverman and Jones, 1989; Dickerson and Kosko, 1993; Cortes and Vapnik, 1995; Breiman, 2001). The modelling input variables were tillage method, soil depth, soil type, and mulch rate, while soil water content (mm) was used as the target variable.

Model evaluation and selection

Soil water data were split into training, 20-fold cross-validation, and testing sets. The training set was used to train the algorithms, the validation set was used for fine-tuning hyper-parameters, and the testing set was used to evaluate the final model performance (Singh and Mallick, 2020). First each algorithm learnt from the training data to predict soil water based on input features such as tillage method, soil depth, soil type and the target variable (soil water). Stratified 20-fold cross-validation ensured the proportions of classes or categories were maintained in the training and testing data sets (Tian and Zhang, 2020). The 20-fold cross-validation involved splitting the data into 20 folds, training the algorithms on 19 folds, and evaluating it on the remaining fold. This process was repeated 20 times, using each fold for evaluation once. This approach provided a more robust estimate of model generalizability compared to a simple training–testing split.

Each algorithm's performance was assessed by estimating mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE), coefficient of determination (R^2), mean absolute percentage error (MAPE) and cross-validation root mean squared error (CVRMSE) to assess accuracy and generalisation ability (Sharma and Bharti, 2020). The MAE denotes the mean absolute deviation between predicted and actual values within the test dataset. Similar to RMSE, lower MAE signifies superior model accuracy. The MAPE resembles MAE but expresses errors as a percentage of actual values. The coefficient of determination gauges the fraction of variance in actual soil water content accounted for by the model's predictions, with values ranging from 0 to 1; higher values indicate stronger model performance (Tian and Zhang, 2020). The CVRMSE assesses model generalizability by dividing data into folds, training the model on one subset, and evaluating its performance on the others. This process is repeated across folds, with the average RMSE serving as the CVRMSE.

The major analyses were conducted in Orange-Miniconda version 3.3.36.2 (Fig. 1), which is a Python library. It was developed by the Bioinformatics Laboratory at the University of Ljubljana, Slovenia, in collaboration with an open-source community. Orange is a free and versatile data visualisation and analysis tool (Demsar et al., 2013). It empowers users to perform data-mining tasks through intuitive visual programming or by utilising Python scripting. The ML workflow shown in Fig. 1 involved connecting various model widgets to the data source. Additionally, test and score widgets were included to assess model performance. To gain deeper insights, widgets were placed between the data and models to explain predictions and extract feature importance. Finally, prediction and feature importance data were exported to R for visualisation.

RESULTS

Descriptive statistics

The violin with embedded box plots (Fig. 2) shows the distribution of soil water content (mm) across various factors influencing soil water. These factors include soil depth (300, 600 and 900 mm), tillage method (basins, plough, and ripper), soil type (clay and sand) and mulch rates (0, 2, 4, and 10 t·ha^{-1}) (Table 1). The soil water distribution across the three soil depths had a positive skew, meaning more data points with lower water content values than the mean occurred in the data set. The few data points occurring above the mean soil water content can reflect wetting cycles after rainfall events that recharged the soil profiles. The median water content also decreased with increasing soil depth, indicating that topsoil layers (300 mm) held more water than deeper layers (900 mm) when measurements were taken in the experimental plots.

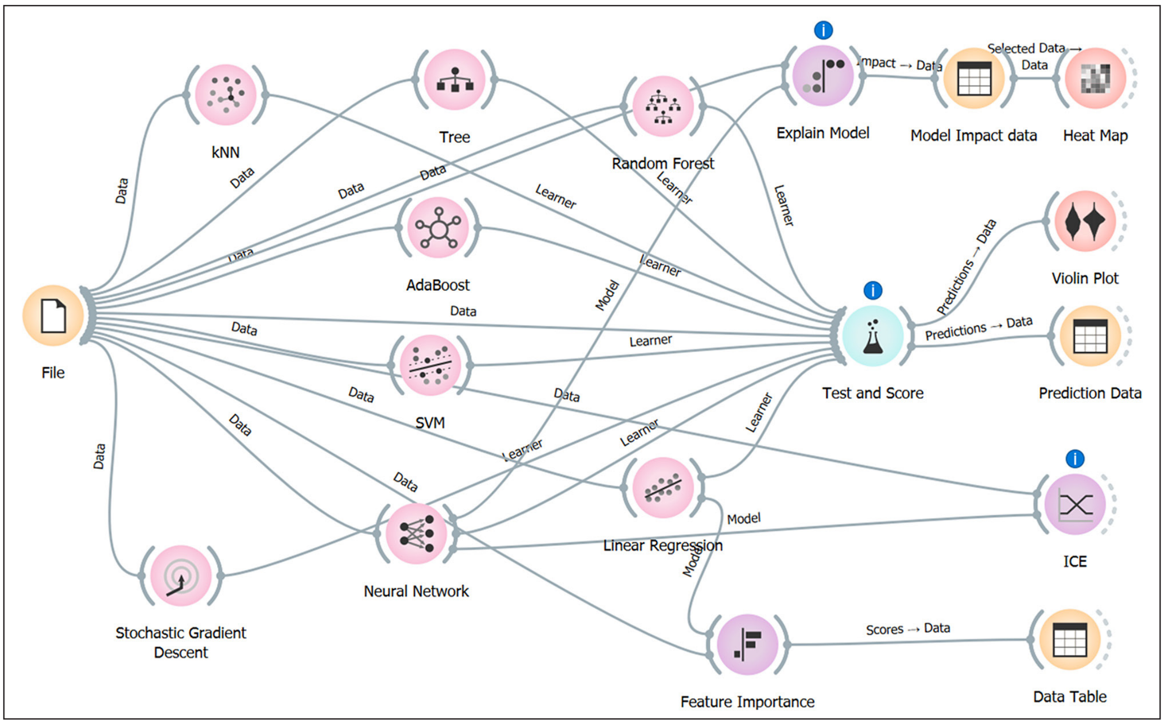

Figure 1. Model implementation architecture in Orange-Miniconda version 3.3.36.2. Notes: This workflow illustrates the complete machine-learning pipeline used for predicting soil water content across different tillage systems and soil types. The process begins with the file widget, where the dataset (soil moisture, soil type, tillage practice, and related covariates) is imported. The dataset is then simultaneously routed to multiple supervised learning algorithms.

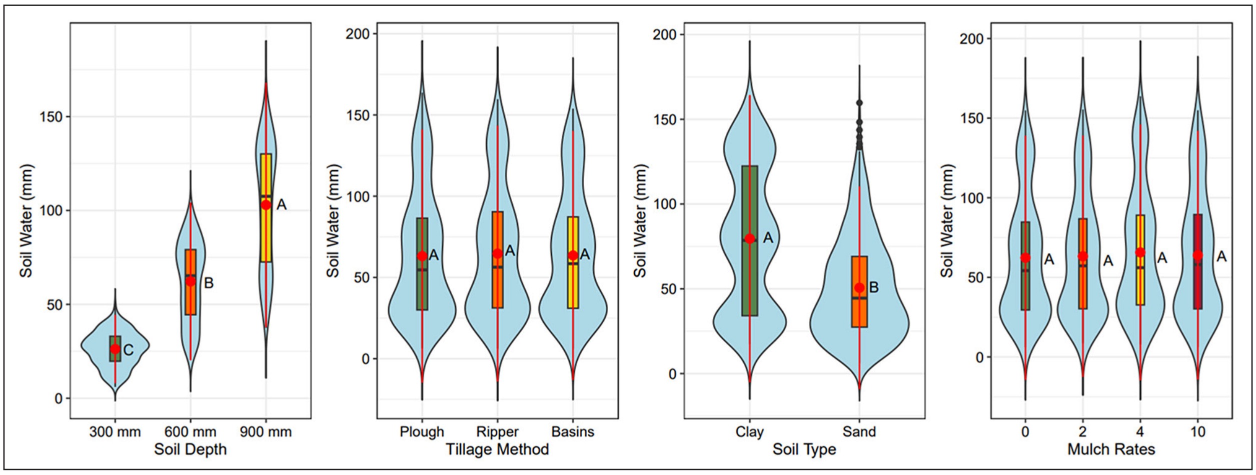

Figure 2. The distribution of soil water content (measured in millimetres) across various features influencing soil moisture (N = 1 188). Red circles inside the boxplots are the means, the horizontal line in the middle of the box represents the median, and the width of the box reflects the variability within the middle 50% of the data.

Table 1. Soil water content (mm) descriptive statistics across soil type, tillage methods, soil depth (mm) and mulch rate (t·ha^{-1})

Factor	N	Minimum	Maximum	Mean	SD	Variance
Clay	540	17.4	163.8	80	42.4	1794
Sand	648	6.1	159.7	51	29.9	895.4
Basins	396	6.1	153.9	64	38.3	1 468.4
Plough	396	6.3	163.8	63	38.9	1 519.7
Ripper	396	6.2	159.7	65	39.4	1 550.7
300	396	6.1	50.7	26	9.1	83
600	396	20.7	104.1	62	20.9	440.9
900	396	37.4	163.8	103	32.6	1 060.6
0	297	6.0	154.8	62	38.4	1 521
2	297	8.7	155.4	63	37.9	1 511
4	297	7.8	163.8	66	40.2	1 521
10	297	6.2	154.9	64	39.1	1 511

Table 2. Performance metrics of the machine-learning models for predicting soil water content, derived from the 20-fold stratified cross-validation procedure applied to the full dataset (i.e., the training-validation folds). Notes: The results presented (MSE, RMSE, MAE, MAPE, R^2, CVRMSE) reflect averaged performance across all cross-validation folds; no external hold-out test set was used for generating this table. The training and testing times shown correspond to model fitting and internal fold-level evaluation during cross-validation.

Model	*Time (s)		Model performance metrics					
	Train	Test	MSE	RMSE	MAE	MAPE	R^2	CVRMSE
Decision tree	2.744	0.003	267.164	16.345	12.315	0.273	0.823	25.629
AdaBoost	0.743	0.089	270.258	16.44	12.421	0.279	0.821	25.777
Support vector machine	0.530	0.133	1 308.8	36.177	30.840	0.869	0.133	56.726
Neural network	18.646	0.105	249.674	15.801	11.997	0.264	0.835	24.776
Stochastic gradient descent	0.292	0.106	335.448	18.315	14.711	0.321	0.778	28.718
K-nearest neighbours	0.254	0.093	302.417	17.39	12.934	0.287	0.800	27.268
Random forest	0.546	0.085	267.944	16.369	12.313	0.273	0.822	25.667
Linear regression	0.227	0.066	321.093	17.919	14.411	0.31	0.787	28.097

*Time (s): the time it took to train the model in seconds

Model performance in predicting soil water content

The models' performance in predicting soil water content is summarised in Table 2. Training time reflects how long each model took to learn from the training data. As expected, complex models like neural network and decision tree needed more training time compared to simpler models such as linear regression. Neural networks, decision trees and random forests consistently achieved the lowest test errors (RMSE, MAE) and the best fit to the soil water content data (Figs A1, A2 and A3; Appendix), as indicated by the highest R^2 value (Table 2). While other models performed adequately, these three consistently excelled and, based on these metrics, the neural network appears to be more appropriate for predicting soil water content. It achieved the lowest error rates (RMSE = 15.801, MAE = 11.997, MAPE = 0.264) and the highest R^2 value (0.835), suggesting a strong fit between the model and the data (Table 2).

Importance of factors influencing soil water content

The impact of each factor on soil water content was analysed using R^2 to understand which treatments are more influential for soil water content predictions (Fig. 3). Soil depth and soil type were the most important factors in all of the models, as they significantly decreased the models' R^2 value. The tillage method (Fig A4, Appendix) and mulch rate were less influential on soil water content than the aforementioned factors. A decision tree plot was created to gain a deeper understanding of how various factors influenced the predicted soil water content (Fig. 4). This plot revealed that the initial split considered the depth of soil water measurement. When the soil depth was 300 mm, the model analysed additional factors. At 300 mm soil depth, especially in sandy soil, soil water content was low (less than 23 mm). However, at this 300 mm soil depth, tillage method and mulch rate were more influential on soil water content in sandy soil. Individual conditional expectation (ICE) plots (Fig. 5) offer a detailed view of how predicted soil water content depends on specific factors for each data point, unlike partial dependence plots, which show one line representing the overall trend. Analysis of the soil water content data revealed the following insights: Soil water content increased with an increase in soil depth, and this is due to higher evaporation from the surface and increased uptake by plant roots concentrated in the top 300 mm soil layer.

Measured and predicted soil water based on model performance

Neural network, decision tree and random forest models exhibited a positive trend between measured and predicted soil water content. The algorithms also had similar soil water prediction power (0.91) illustrated by the Pearson correlation coefficients (R) (Fig. 6), suggesting they are all suitable for use under the soil properties and other conditions that influenced soil water content in the study environment. Notably, the three ML models achieved the highest R values (0.91), indicating a very strong ability to capture the variance in the soil water content data. This is further supported by their extremely low p-values ($p < 0.001$), signifying a statistically significant correlation between the measured and predicted values.

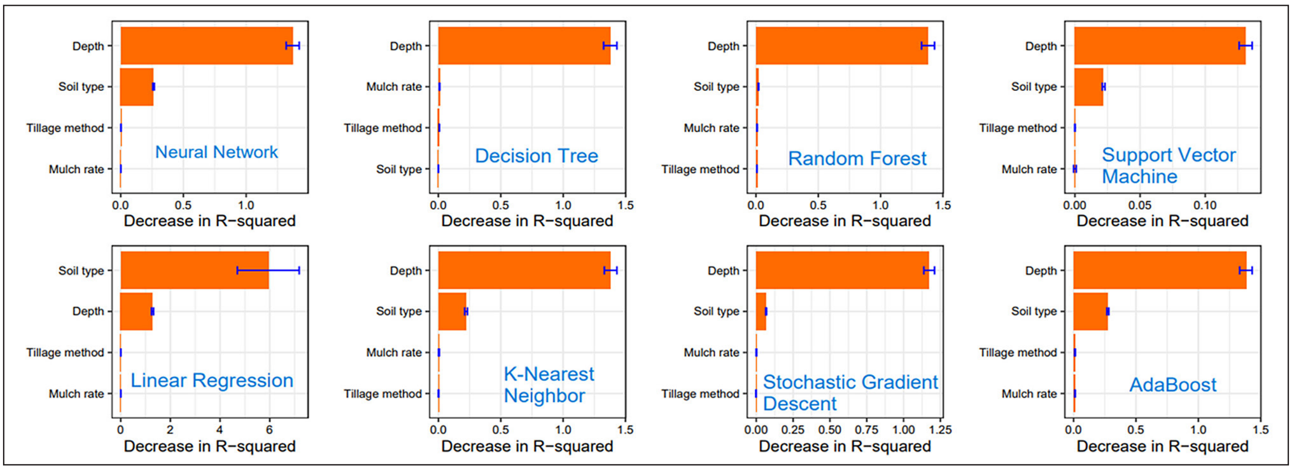

Figure 3. Feature importance evaluation based on a decrease in R^2. The error bars are indicated in blue based on the mean and standard deviation estimated per model for each performance.

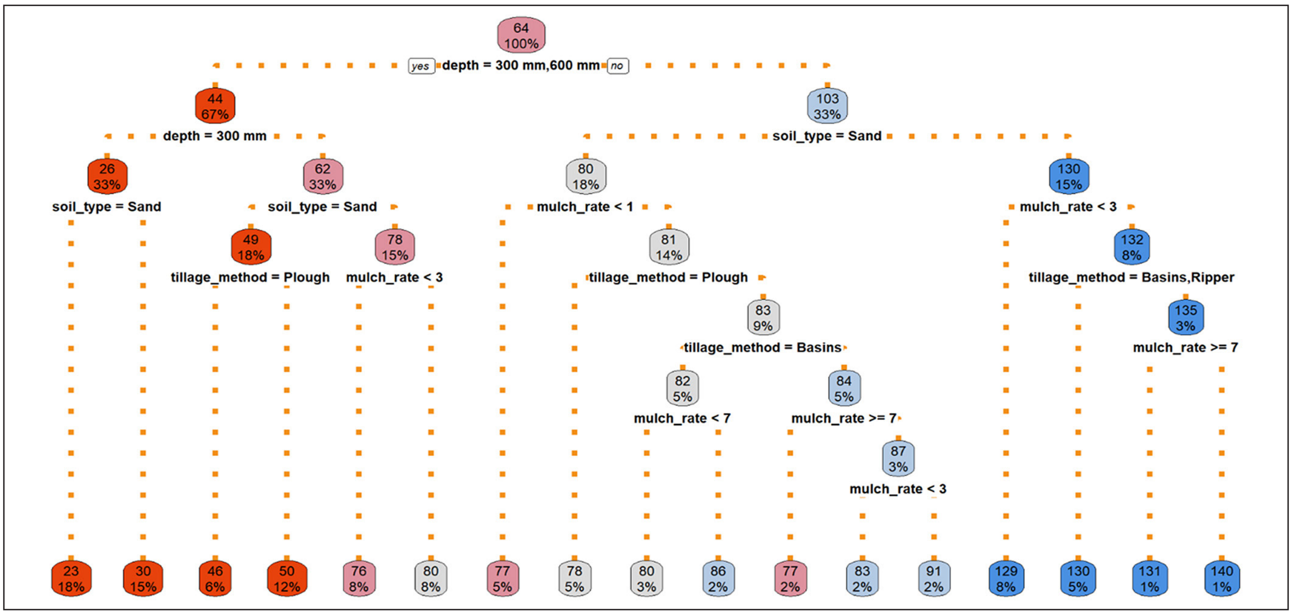

Figure 4. Decision tree for soil water prediction and feature interpretation. Note: The dependent variable is soil water content (mm). The decision tree was developed using Orange-Miniconda (version 3.3.36.2), which automatically manages internal randomization through the application's global seed settings. Although the decision tree widget does not require an explicit "random_state" parameter, reproducibility of the model was ensured by (i) setting the global random seed within Orange, and (ii) using stratified 20-fold cross-validation for consistent and repeatable model evaluation across all runs. The tree was constructed under the following parameters: minimum tree size before splitting = 20 (primary and surrogate splits based on competing variables per node); maximum improvement rate by split = 0.01 (Gini impurity); maximum tree depth = 30; data splitting method = random; maximum percentage increase for minority class size = 200.

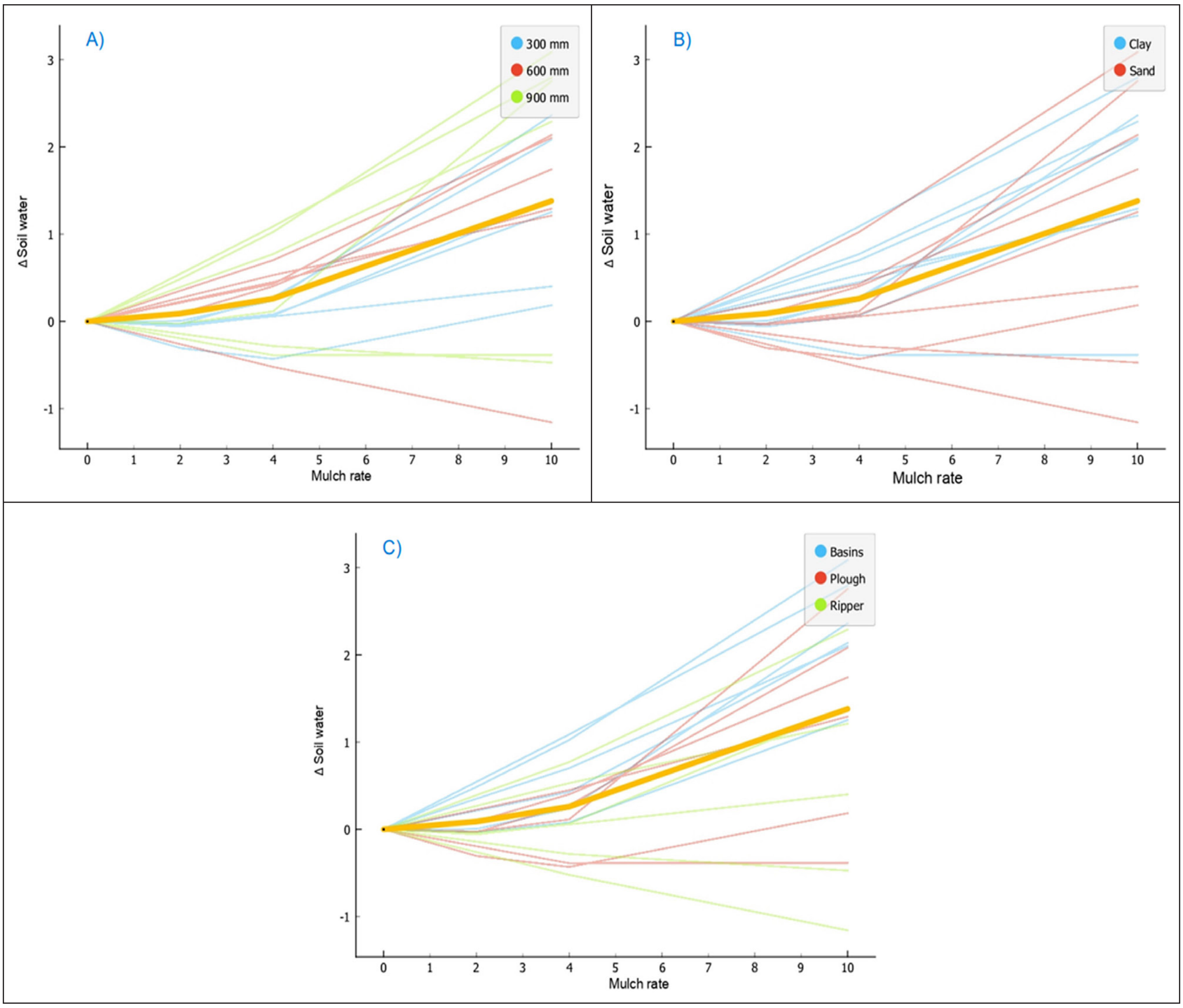

Figure 5. Individual conditional expectation (ICE) plots showing how mulch rate affects predicted changes in soil water content (Δ soil water) based on the best-performing model (neural network). Each thin coloured line represents the ICE curve for an individual observation within each category: Panel A: depth classes (300 mm, 600 mm, 900 mm), Panel B: soil types (clay, sand), Panel C: tillage methods (basins, plough, ripper). The orange line in each panel represents the averaged effect (centrally smoothed or 'mean ICE') across all individual ICE curves, providing a clearer interpretation of the overall trend. Legends correspond only to the coloured individual-observation curves; the orange line is not a separate class but the aggregate ICE trend.

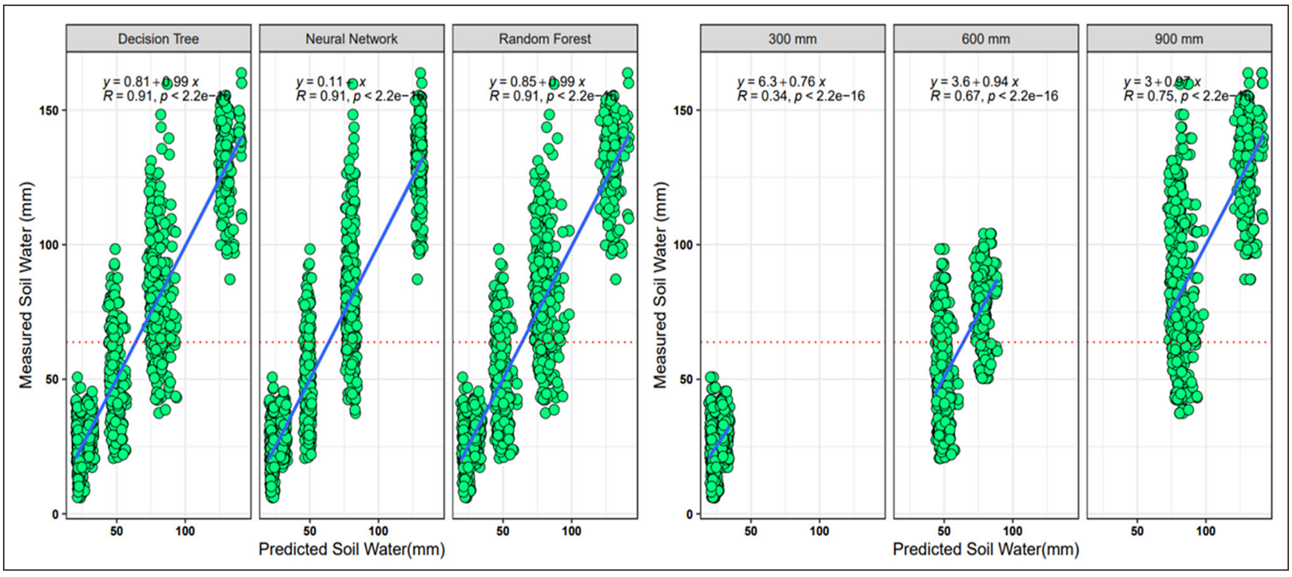

Figure 6. Relationship between measured soil water content and predicted soil water content for the three machine-learning models (decision tree, neural network, and random forest). Each subplot shows the model's scatter plot and regression line. In Orange-Miniconda, the visualisation widget automatically reports the Pearson correlation coefficient (R) and its p-value as part of its built-in plot outputs.

DISCUSSION

Soil water dynamics

More soil water in topsoil signifies recharged profiles after rainfall events (Madembo and Nyagumbo, 2025). The width of each violin plot represents the variability in soil water content at a given soil depth, with wider plots indicating greater variation. Notably, the violin plot for the 300 mm depth is wider than the one for 900 mm, signifying more variation in water content in the topsoil. More variation in topsoil could be a reflection of frequent changes in soil water content due to soil evaporation coupled with plant root extraction (Thierfelder and Wall, 2009).

There was high variability in soil water content in sandy soil, which was attributed to faster soil water loss through evaporation and deep drainage in sandy soil compared to clay soil. Soil water content was marginally higher in ripper treatment compared to the other two methods, although the overall impact of the tillage method on soil water content was minimal. Soil water content at the clay Matopos site exhibits a significantly higher and more concentrated distribution than the sandy soil at Lucydale, even displaying a trimodal structure (3 peaks). This suggests that water content at Matopos clay soil is not only higher but also exhibits distinct clusters within the overall distribution. Soils with high clay content (>30%) have high micropores and relatively low hydraulic conductivity, which help in retaining water (Al-Shayea, 2001). Soil water content was not significantly affected by different mulch rates (0, 2, 4, and 10 t·ha^{-1}), suggesting that mulch cover was not a major factor influencing soil water content compared to factors like soil depth and soil type during experimentation. However, mulch may still be important for slowing down soil evaporation from the top layer (Jat et al., 2016), which can benefit crops through the slowed drying cycle of the soil profile. In general, soil type and depth are crucial in determining the variability and overall storage of soil water in cropping systems. High variations in soil water content at the 300 mm depth can be attributed to the extraction of soil water by maize grown in the experimental plots, as well as evaporation. A study by Chikowo et al. (2004) revealed that maize root density is high for the 0–0.4 m soil depth during the critical vegetative stages of the plants.

Performance of ML algorithms

The 'best' model to use depends on the specific problem and the objective(s) of the analysis. Other factors, such as how easy it is to understand the model's predictions (interpretability), training time, and ability to handle larger datasets (scalability), also play a role in choosing the right model. The data set generated during field experiments, and used for training and validation, was adequate for the neural network, decision tree and random forest models to give good predictions of soil water under the experimental conditions of the reported study. A study by De Oliveira et al. (2020) in a forest system showed that RF is a good algorithm ($R^2 = 0.51$) for modelling spatial distribution of soil moisture. This is consistent with the findings of this study which indicated that RF can be applied to simulate ($R^2 = 0.82$) soil water patterns in agricultural system. The DT outperformed ($R^2 = 0.83$) the other algorithms tested, and this is consistent with results ($R^2 = 0.70$–0.73) from Hateffard et al. (2019) in a study conducted at landscape scale. The SVM algorithm performed poorly ($R^2 = 0.24$) in the study by De Oliveira et al. (2020), which is consistent with its simulation capability achieved ($R^2 = 0.13$) in the current study.

The conventionally ploughed treatment had lower soil water content compared to ripping and basins. The disturbed topsoil loses soil water quickly through evaporation. Traditionally, winter ploughing is often done during the April–May period in southern Africa because the dry topsoil acts as a 'mulch' for the deeper soil layers (Twomlow et al., 2006), and this helps conserve soil water during the winter and spring seasons. At the onset of the rains in October–November, the soil profile recharges faster due to the conserved residual soil water in deeper layers. When the mulch rates exceeded 3 t·ha^{-1} in basins or ripping treatments, soil water content became significantly higher than in conventional ploughing treatment. Notably, for treatments with less than 1 t·ha^{-1} mulch cover, the tillage method became a critical factor influencing soil water content. Soil water conservation in no-till and conservation agriculture-based systems has been reported in many studies from the Global North and South (Thierfelder and Wall, 2009; Li et al., 2011; Jat et al., 2013; Jat et al., 2016). Overall, the decision tree plot results align with the descriptive statistics and treatment importance analysis, highlighting the significant impact of soil depth, soil type, and their interaction, on soil water content. The plot in Fig. 4 also emphasises the importance of mulching and tillage methods in conserving soil water, particularly in treatments with lower soil water content. A clear difference exists between sand and clay soil, with the latter having higher soil water content. This is because clay particles have smaller pores which retain more water than the macropores in sandy soils. While some variations in soil water content existed between different tillage methods, the differences were not significant.

Building on this study, these ML algorithms can be coupled with biophysical models such as APSIM and DSSAT in understanding and predicting soil water dynamics and management in agricultural systems (Shahhosseini et al., 2021). The future research may focus on:

- Application of boosting techniques that can improve the performance of ML algorithms during the exploration of different issues affecting agriculture.
- Comparing the outputs of ML algorithms and biophysical models as a way of assessing the best tool to apply in understanding certain parameters such as soil water, crop yields and, pest and disease predictions in agricultural systems.
- Developing hybrid tools involving best-performing ML algorithms and biophysical models for selected parameters such as crop yields and water conservation in agricultural systems.

Measured and predicted soil water based on model performance

The correlation of predicted and measured soil water content strengthened with increase in soil depth (Fig. 6). This result confirms and is consistent with higher variability in soil water content in the top 300 mm soil layer compared to the deeper layers (Fig. 5). In essence, the results indicate that ML can be a viable tool for predicting soil water content with some degree of accuracy. However, within the scope of this experiment, the decision tree, random forest and neural network models displayed a clear advantage, in terms of both correlation strength and statistical significance, over the other models tested.

The ML algorithms used in this study were more precise in predicting soil water in both clay and sandy soils than the APSIM model in the same semi-arid agroecological conditions. A study by Ncube et al. (2007) indicated that APSIM prediction achieved an R^2 of 0.58 and 0.65 when compared to the measured soil water content at the same sandy soil research site used for our study. This level of soil water prediction accuracy by APSIM is similar to the decision tree, random forest and neural network models used in this study, making the ML models a good option for soil water studies. However, Hateffard et al. (2019) emphasised the importance of pre-testing these ML models before use because their prediction accuracy for different soil properties varies between environments; hence, universal recommendations are not appropriate. Furthermore, the selection of the ML model should be based on the soil property to be predicted and not on the prediction power of a model, because their accuracy varies with the soil property of interest (Rani et al., 2022).

CONCLUSION

The decision tree, random forest and neural network models accurately predicted soil water content under different treatments applied in this study. Therefore, decision tree, random forest and neural network models are appropriate ML tools for soil water prediction in sandy and clay soils under semi-arid agroecological conditions. All the ML models poorly predicted soil water based on tillage practices (conventional, ripper and basins) regardless of soil type. For the prediction and classification of soil water, neural network, decision tree and random forest models predicted soil water content more precisely in sandy than clay soil regardless of tillage practice and soil depth. The SVM was the weakest ML model for predicting soil water in sand and clay soils across tillage practices. The study recommends the application of neural network, decision tree and random forest models in soil water management and for informing decision making on water resource management in agricultural systems. The performance of the ML algorithms was influenced by soil type; hence there is need to conduct more studies incorporating other soil textural classes that are important for food production systems.

ACKNOWLEDGEMENTS

The authors acknowledge funding received from the WaterNet Challenge Food Program (Project Number 17) – "Integrated Water Resource Management for Improved Rural Livelihoods: Managing risk, mitigating drought and improving water productivity in the water scarce Limpopo Basin" under which the reported research was conducted, and data generated for this paper. Logistical and office space support provided by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) during the execution of the research project is highly appreciated. Field support provided by Gertrude Mpofu, Bhekimpilo Ncube, Thulani Ndlovu and Vusumuzi Ndlovu during the implementation of the research project is acknowledged.

AUTHOR CONTRIBUTIONS

Mupangwa W: conceptualisation; writing original draft. Chipindu L: data curation; formal analysis; writing methodology; validation; visualisation; writing original draft. Ncube B: editing the draft. Tauro TP: editing the draft.

FUNDING

This study was funded by the WaterNet-led Challenge Program on Water and Food, Project Number 17 (CPWF PN17) Integrated Water Resource Management for Improved Rural Livelihoods: Managing risk, mitigating drought and improving water productivity in the water scarce Limpopo Basin.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest in the research work reported in this study.

ORCIDS

W Mupangwa
https://orcid.org/0000-0001-5672-1331

L Chipindu
https://orcid.org/0000-0001-8023-1105

B Ncube
https://orcid.org/0000-0003-1936-2768

TP Tauro
https://orcid.org/0000-0002-9351-5956

REFERENCES

AL-SHAYEA NA (2001) The combined effect of clay and moisture content on the behavior of remolded unsaturated soils. *Eng. Geol.* **62** (4) 319–342. https://doi.org/10.1016/S0013-7952(01)00032-1

BALDUCCI F, IMPEDOVO D and PIRLO G (2018) Machine learning applications on agricultural datasets for smart farm enhancement. *Machines* **6**. https://doi.org/10.3390/machines6030038

BREIMAN L (2001) Random Forests. Statistics Department, University of California, Berkeley, CA 94720. 33 pp. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf (Accessed 6 March 2026).

CAI Y, ZHENG W, ZHANG X, ZHANGZHONG L and XUE X (2019) Research on soil moisture prediction model based on deep learning. *PLoS One* **14**. https://doi.org/10.1371/journal.pone.0214508

CHAPAGAIN R, REMENYI TA, HARRIS RM, MOHAMMED CL, HUTH N, WALLACH D, REZAEI EE, and OJEDA JJ (2022). Decomposing crop model uncertainty: A systematic review. *Field Crops Res.* **279** 108448. https://doi.org/10.1016/j.fcr.2022.108448

CHIKOWO R, MAPFUMO P, NYAMUGAFATA P, and GILLER KE (2004) Maize productivity and mineral N dynamics following different soil fertility management practices on a depleted sandy soil in Zimbabwe. *Agric. Ecosyst. Environ.* **102** (2) 119–131. https://doi.org/10.1016/j.agee.2003.08.009

CORTES C and VAPNIK V (1995) Support-vector networks. *Machine Learning.* **20** (3) 273–297. https://doi.org/10.1007/BF00994018

CRANE-DROESCH A (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* **13**. https://doi.org/10.1088/1748-9326/aae159

DEMSAR J, CURK T, ERJAVEC A, GORUP C, HOCEVAR T, MILUTINOVIC M, MOZINA M, POLAJNAR M, TOPLAK M, STARIC A, STAJDOHAR M, UMEK L, ZAGAR L, ZBONTAR J, ZITNIK M and ZUPAN B (2013) Orange: Data Mining Toolbox in Python. *J. Machine Learning Res.* **14** 2349–2353.

DE OLIVEIRA VA, RODRIGUES AF, MORAIS MAV, DE CASTRO NUNES SANTOS M, GUO TL and DE MELLO CR (2020) Spatiotemporal modelling of soil moisture in an Atlantic forest through machine learning algorithms. *Eur. J. Soil Sci.* **72** 1969–1987. https://doi.org/10.1111/ejss.13123

DICKERSON JA and KOSKO B (1993) Fuzzy function approximation with supervised ellipsoidal learning. World Conference on Neural Networks (WCNN'93). II, 9–17. https://sipi.usc.edu/~kosko/Virtual_Worlds_FCM.pdf (Accessed 4 December 2024).

ENKE D and MEHDIYEV N (2012) A new hybrid approach for forecasting interest rates. *Procedia Comput. Sci.* **12** 259–264. https://doi.org/10.1016/j.procs.2012.09.066

GILL MK, ASEFA T, KEMBLOWSKI MW and MCKEE M (2006) Soil moisture prediction using support vector machines. *J. Am. Water Resour. Assoc.* **42** 1033–1046. https://doi.org/10.1111/j.1752-1688.2006.tb04512.x

GIUSTI E and MARSILI-LIBELLI (2015) A Fuzzy decision support system for irrigation and water conservation in agriculture. *Environ. Model. Softw.* **63** 73–86. https://doi.org/10.1016/j.envsoft.2014.09.020

GORNI G and AUGUSTO A (2008) The application of neutral networks in the modelling of plate rolling processes. *Miner. Met. Mater. Soc.* **49** 1–4. http://www.tms.org/pubs/journals/JOM/9704/Gorni/Gorni-9704.html

HATEFFARD F, DOLATI P, HEIDARI A and ZOLFAGHARI AA (2019) Assessing the performance of decision tree and neural network models in mapping soil properties. *J. Mountain Sci.* **16** (8) 1833–1847. https://doi.org/10.1007/s11629-019-5409-8

HOU XL, FENG YH, WU GH, HE YX and CHANG DM (2016) Application research on artificial neural network dynamic prediction model of soil moisture. *Water Sav. Irrig.* **7** 70–72. https://doi.org/10.1371/journal.pone.0214508

KAUL M, HILL L and WALTHALL RC (2005) Artificial neural networks for corn and soybean yield prediction. *Agric. Syst.* **85** 1–18. https://doi.org/10.1016/j.agsy.2004.07.009

KEATING BA, CARBERRY PS, HAMMER GL, PROBERT ME, ROBERTSON MJ, HOLZWORTH D, HUTH NI, HARGREAVES JNG, MEINKE H, HOCHMAN Z, MCLEAN G, VERBURG K, SNOW V, DIMES JP, SILBURN M, WANG E, BROWN S, BRISTOW KL, ASSENG S, CHAPMAN S, MCCOWN RL, FREEBAIRN DM, and SMITH CJ (2003) An overview of APSIM, a model designed for farming systems simulation. *Eur. J. Agron.* **18** (3–4) 267–288. https://doi.org/10.1016/j.envsoft.2014.07.009

KING BA and SHELLIE KC (2016) Evaluation of neural network modeling to predict non-water-stressed leaf temperature in wine grape for calculation of crop water stress index. *Agric. Water Manage.* **167** 38–52. https://doi.org/10.1016/j.agwat.2015.12.009

JAT ML, GATHAHA MK, SAHRAWAT YS, TETAWAL JP, GUPTA R, and SINGH Y (2013) Double no-till and permanent raised beds in maize-wheat rotation of northwestern Indo-Gangetic plains of India: effects on crop yields, water productivity, profitability and soil physical properties. *Field Crops Res.* **149** 291–299. https://doi.org/10.1016/j.fcr.2013.04.024

JAT ML, DAGAR JC, SAPKOTA TB, SINGH Y, GOVAERTS B, RIDAURA SL, SAHARAWAT YS, SHARMA RK, TETARWAL JP, JAT RK, HOBBS H, and STIRLING C (2016) Climate change and agriculture: adaptation strategies and mitigation opportunities for food security in South Asia and Latin America. *Adv. Agron.* **137** 127–235. https://doi.org/10.1016/bs.agron.2015.12.005

JONES JW, TSUJI GY, HOOGENBOOM G, HUNT LA, THORNTON PK, WILKENS PW, IMAMURA DT, BOWEN WT, and SINGH U (1998) Decision support system for agrotechnology transfer: DSSAT v3. In: Tsuji GY, Hoogenboom G, Thornton PK (eds) *Understanding Options for Agricultural Production. Systems Approaches for Sustainable Agricultural Development*. 7. Springer, Dordrecht. https://doi.org/10.1007/978-94-017-3624-4_8

LEGATES DR and MCCABE GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **35** (1) 233–241. https://doi.org/10.1029/1998WR900018

LIAKOS KG, BUSATO P, MOSHOU D, PEARSON S and BOCHTIS D (2018) Machine learning in agriculture: A review. *Sensors* **18** 1–29. https://doi.org/10.3390/s18082674

LI LL, HUANG G, ZHANG R, BILL B, GUANGDI L and KWONG YC (2011) Benefits of conservation agriculture on soil and water conservation and its progress in China. *Agric. Sci. China* **10** (6) 850–859. https://doi.org/10.1016/S1671-2927(11)60071-0

MADEMBO C and NYAGUMBO I (2026) Soil fertility matters most: Rainfed maize productivity under integrated water and nutrient management on granitic sands in sub-humid Zimbabwe. *Agric. Water Manage.* **323** 110093. https://doi.org/10.1016/j.agwat.2025.110093

MAMOMBE V, KIM W and CHOI Y (2017) Rainfall variability over Zimbabwe and its relation to large-scale atmosphere-ocean processes. *Int. J. Climatol.* **37** 963–971. https://doi.org/10.1002/joc.4752

MISHRA S, MISHRA D, and SANTRA GH (2016) Applications of machine learning techniques in agricultural crop production: A review paper. *Indian J. Sci. Technol.* **9**. https://doi.org/10.17485/ijst/2016/v9i38/95032

MKUHLANI S (2013) Assessment of the effects of conservation agriculture on Maize (*Zea Mays* L.) crop yields under a changing climate. MSc. thesis, University of Zimbabwe.

MUPANGWA W, DIMES J, WALKER S and TWOMLOW S (2011) Measuring and simulating (*Zea mays* L.) yield responses to reduced tillage and mulching under semi-arid conditions. *Agric. Sci.* **2** 167–174. https://doi.org/10.4236/as.2011.23023

MUPANGWA W and JEWITT GPW (2011) Simulating the impact of no-till systems on field water fluxes and productivity under semi-arid conditions. *Phys. Chem. Earth* **36** 1004–1011. https://doi.org/10.1016/j.pce.2011.07.069

MUPANGWA W, TWOMLOW S and WALKER S (2012) Reduced tillage, mulching and rotational effects on maize (*Zea mays* L.), cowpea (*Vigna unguiculata* (Walp) L.) and sorghum (*Sorghum bicolor* L. (Moench)) yields under semi-arid conditions. *Field Crops Res.* **132** 139–148. https://doi.org/10.1016/j.fcr.2012.02.020

MUPANGWA W, TWOMLOW S, WALKER S and HOVE L (2007) Effect of minimum tillage and mulching on maize (*Zea mays* L.) yield and water content of clayey and sandy soils. *Phys. Chem. Earth* **32** 1127–1134. https://doi.org/10.1016/j.pce.2007.07.030

NCUBE B, DIMES JP, TWOMLOW SJ, MUPANGWA W and GILLER KE (2007) Raising the productivity of smallholder farms under semi-arid conditions by use of small doses of manure and nitrogen: a case of participatory research. *Nutr. Cycl. Agroecosyst.* **77** 53–67. https://doi.org/10.1007/s10705-006-9045-7

NCUBE B, DIMES JP, VAN WIJK M and TWOMLOW S (2009) Productivity and residual benefits of grain legumes to sorghum under semi-arid conditions in south-western Zimbabwe: Unravelling the effects of water and nitrogen using a simulation model. *Field Crops Res.* **110** 173–184. https://doi.org/10.1016/j.fcr.2008.08.001

NYAGUMBO I, MKUHLANI S, PISA C, KAMALONGO D, DIAS D and MEKURIA M (2015) Maize yield effects of conservation agriculture based maize–legume cropping systems in contrasting agro-ecologies of Malawi and Mozambique. *Nutr. Cycl. Agroecosyst.* **105** 275–290. https://doi.org/10.1007/s10705-015-9733-2

PADARIAN J, MINASNY B and MCBRATNEY AB (2020) Machine learning and soil sciences: A review aided by machine learning tools. *Soil* **6** 35–52. https://doi.org/10.5194/soil-6-35-2020

RANI A, KUMAR N, KUMAR J and SINHA NK (2022) Machine learning for soil moisture assessment. In: *Deep Learning for Sustainable Agriculture*. Academic Press. 143–168. https://doi.org/10.1016/B978-0-323-85214-2.00001-X

RAVI V and BHATIA R (2021) A comparative study of machine learning models for predicting river flow using meteorological variables. *Environ. Monit. Assess.* **193** (3) 1–17. URL: https://link.springer.com/journal/10661/volumes-and-issues/193-3

ROSENBLATT F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65** (6) 386–408. https://doi.org/10.1037/h0042519

SHAHHOSSEINI M, HU G, HUBER I, and ARCHONTOULIS SV (2021) Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci. Rep.* **11** 1606. https://doi.org/10.1038/s41598-020-80820-1

SHARMA S and BHARTI VK (2020) Machine learning and statistical techniques in hydrology: A review. *J. Hydrol.* **585** 124799.

SIBANDA S, GRAB SW and AHMED F (2020) Long-term rainfall characteristics in the Mzingwane catchment of south-western Zimbabwe. *Theor. Appl. Climatol.* **139** 935–948. https://doi.org/10.1007/s00704-019-03020-6

SILVERMAN BW and JONES MC (1989) E. Fix and J.L. Hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: commentary on Fix and Hodges (1951). *Int. Stat. Rev.* **57** (3) 233–38. https://doi.org/10.2307/1403796

TESHOME FT, BAYABIL HK, SCHAFFER B, AMPATZIDIS Y and HOOGENBOOM G (2024) Improving soil moisture prediction with deep learning and machine learning models. *Comput. Electron. Agric.* **226** 109414. https://doi.org/10.1016/j.compag.2024.109414

THIERFELDER C and WALL P (2009) Effects of conservation agriculture techniques on infiltration and soil water content in Zambia and Zimbabwe. *Soil Tillage Res.* **105** 217–227. https://doi.org/10.1016/j.still.2009.07.007

THIERFELDER C, BAUDRON F, SETIMELA P, NYAGUMBO I, MUPANGWA W, MHLANGA B, LEE N and GÉRARD B (2018) Complementary practices supporting conservation agriculture in southern Africa: A review. *Agron. Sustain. Dev.* **38** 1–16. https://doi.org/10.1007/s13593-018-0492-8

TIAN Y and ZHANG J (2020) Comparison of artificial intelligence models for rainfall–runoff simulation: A case study in the Loess Plateau of China. *J. Hydroinf.* **22** (2) 277–291.

TWOMLOW SJ, STEYN JT, and DU PREEZ CC (2006) Dryland farming in southern Africa. Chapter 19. In: *Dryland Agriculture* (2nd edn). Agronomy Monograph No. 23. American Society of Agronomy, Madison, Wisconsin. 769–836. https://doi.org/10.2134/agronmonogr23.2ed.c19

TWOMLOW S, UROLOV JC, JENRICH M and OLDRIEVE B (2008) Lessons from the field – Zimbabwe's Conservation Agriculture Task Force. *J. SAT* **6** 1–11. http://www.icrisat.org/journal/Volume6/aes/Twomlow.pdf (Accessed 10 December 2024).

VOGELER I, CICHOTA R, LANGER S, THOMAS S, EKANAYAKE D and WERNER A (2023) Simulating water and nitrogen runoff with APSIM. *Soil Tillage Res*. **227** 105593. https://doi.org/10.1016/j.still.2022.105593

WOLFERT S, GE L, VERDOUW C and BOGAARDT MJ (2017) Big Data in Smart Farming – A review. *Agric. Syst*. **153** 69–80. https://doi.org/10.1016/j.agsy.2017.01.023

YOUNG PC (2006) The data-based mechanistic approach to the modelling, forecasting and control of environmental systems. *Annu. Rev. Control* **30** 169–182. https://doi.org/10.1016/j.arcontrol.2006.05.002

APPENDIX

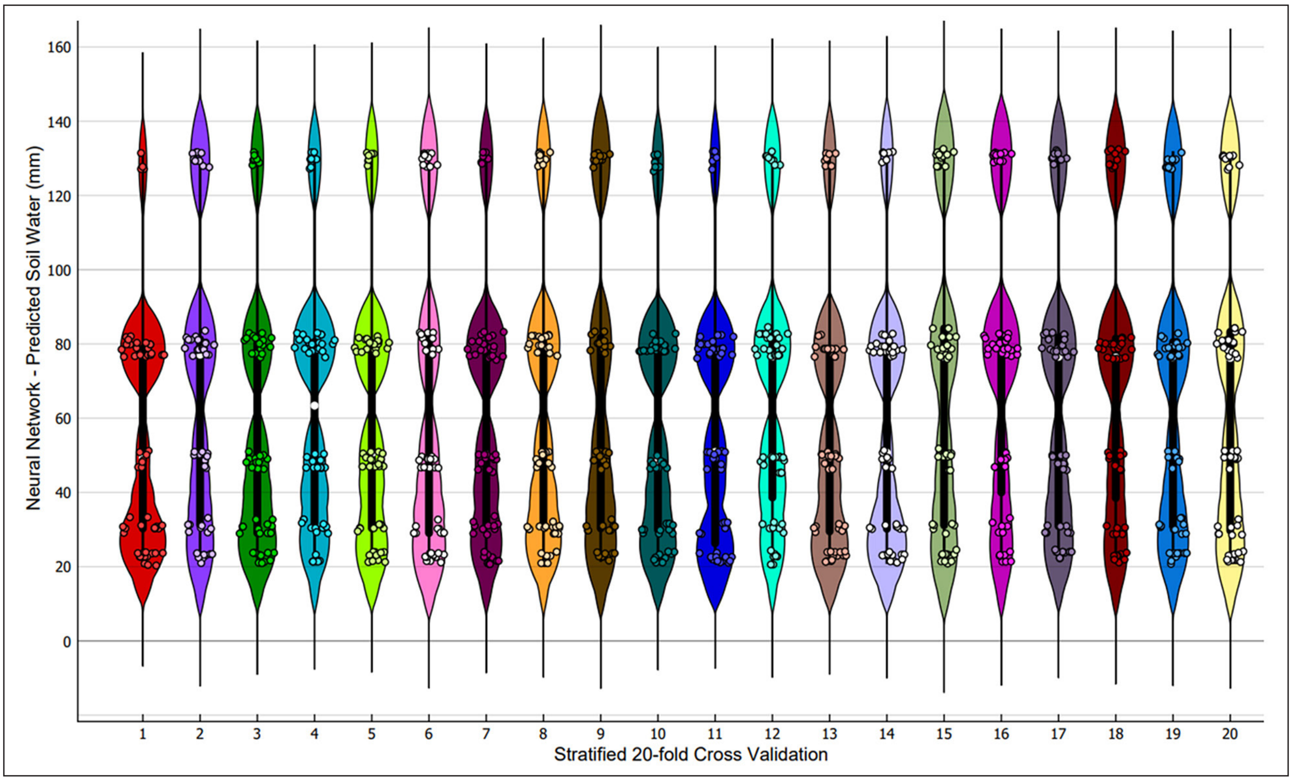

Figure A1. Distribution of neural network–predicted soil water across all stratified 20-fold cross validation

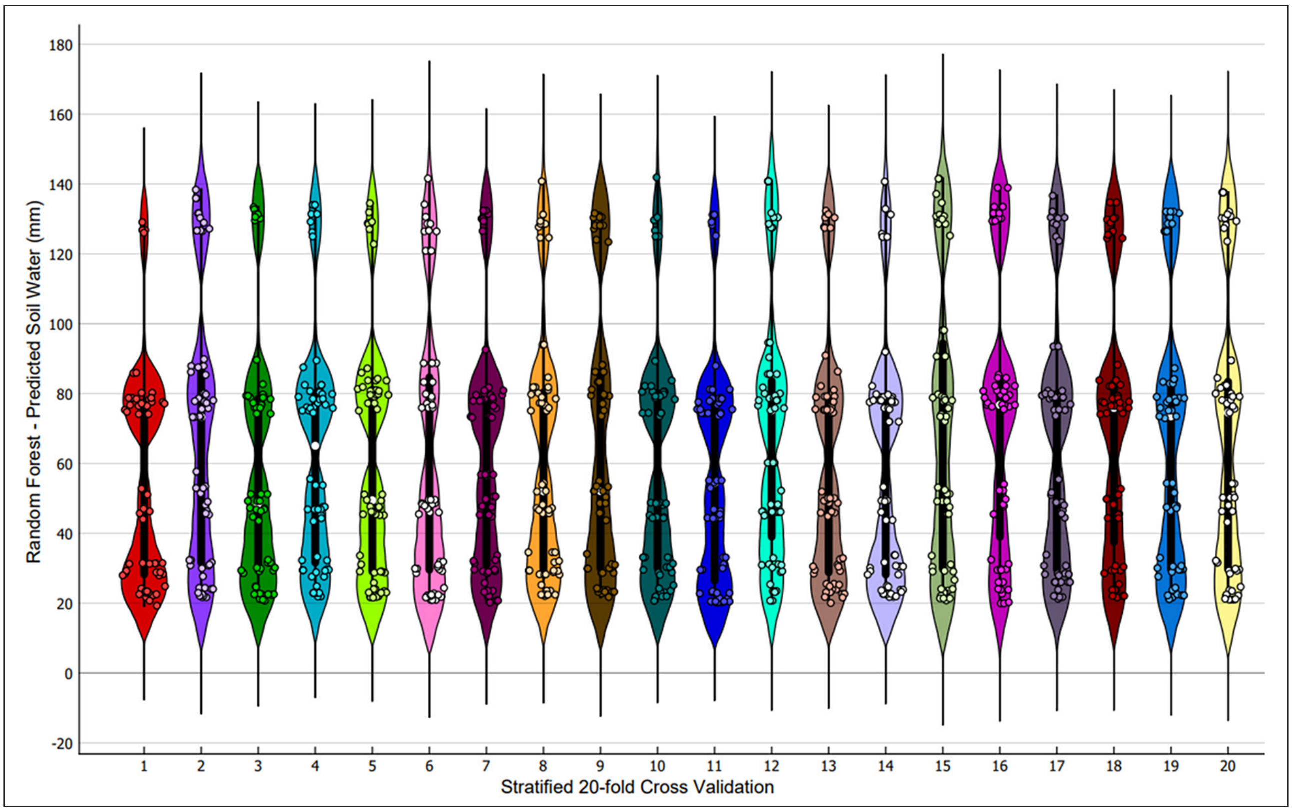

Figure A2. Distribution of random forest–predicted soil water across all stratified 20-fold cross validation

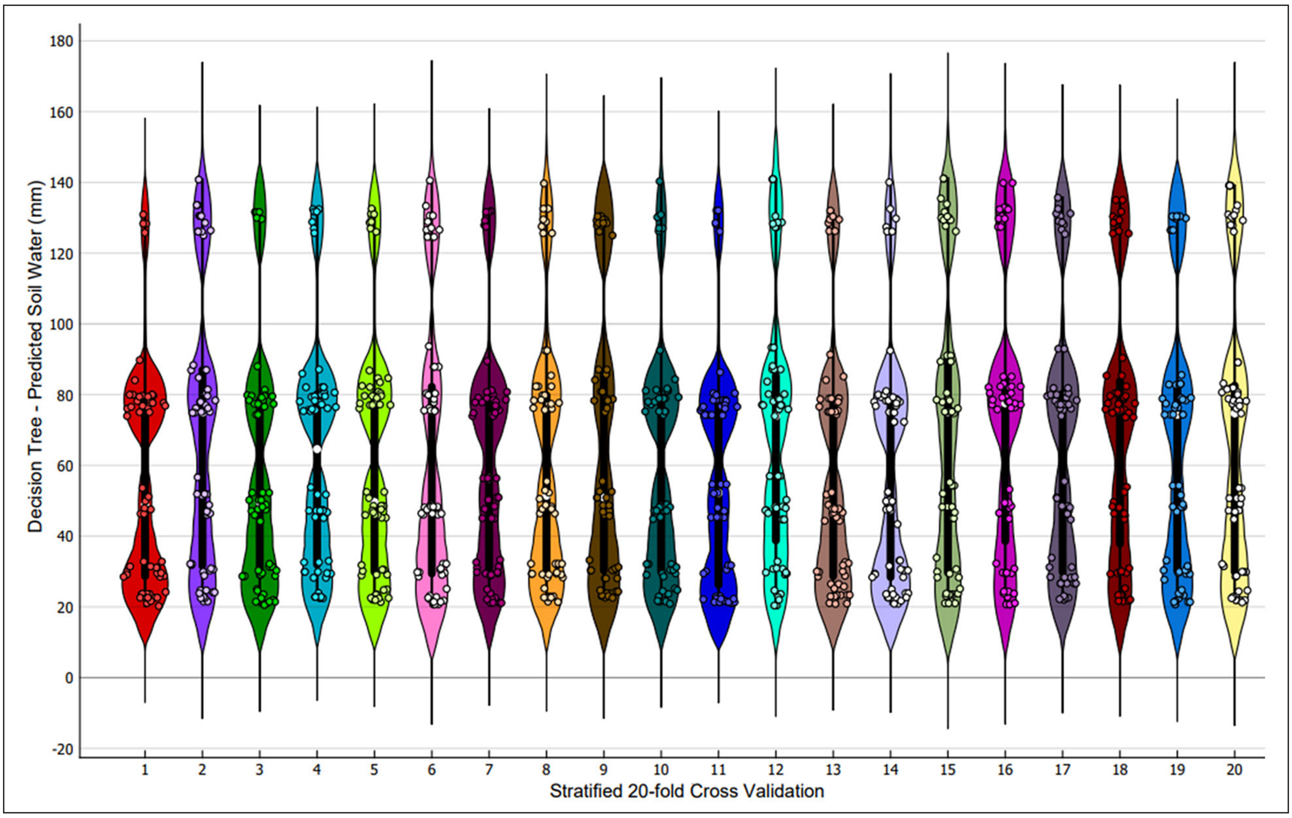

Figure A3. Distribution of decision tree–predicted soil water across all stratified 20-fold cross validation

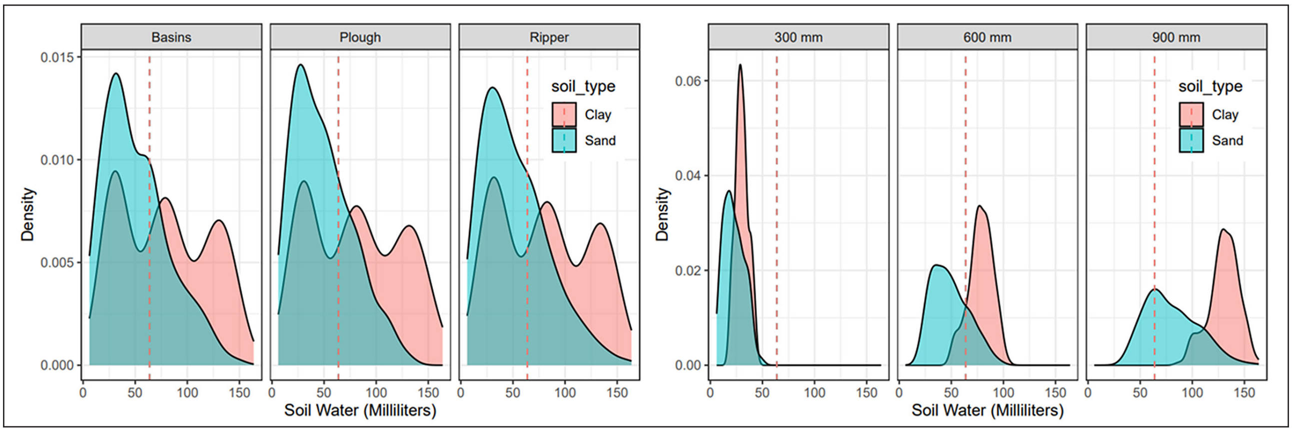

Figure A4. Soil water distribution under different tillage methods and soil depths