Evaluation of Global Data for National-Scale Soil Depth Mapping in Data-Scarce Regions: A Case Study from Sri Lanka

soil systems

MDPI

Article

Evaluation of Global Data for National-Scale Soil Depth Mapping in Data-Scarce Regions: A Case Study from Sri Lanka

Ebrahim Jahanshiri [1,2,*], Eranga M. Wimalasiri [2,3,*], Yinan Yu [1] and Ranjith B. Mapa [4]

[1] Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Chalmersplatsen 4, 41296 Göteborg, Sweden; yinan@chalmers.se
[2] Crops for the Future UK, National Institute of Agricultural Botany, 93 Lawrence Weaver Road, Cambridge CB3 0LG, UK
[3] Department of Export Agriculture, Faculty of Agricultural Sciences, Sabaragamuwa University of Sri Lanka, Belihuloya 70140, Sri Lanka
[4] Department of Soil Science, Faculty of Agriculture, University of Peradeniya, Peradeniya 20400, Sri Lanka; maparb@yahoo.com
* Correspondence: ej@cffinternational.com (E.J.); eranga@agri.sab.ac.lk (E.M.W.)

Abstract

High-resolution soil depth maps are valuable for environmental modelling, yet reliable data remains scarce in the tropics. This study evaluates the feasibility of mapping depth to bedrock (DTB) in Sri Lanka using a legacy dataset (n = 88) and global environmental covariates (n = 247). A robust machine learning workflow was employed—including feature selection, hyperparameter tuning, and a stacked ensemble of four algorithms (Random Forest, XGBoost, Cubist, SVM)—to test the limits of global data for local mapping. Despite rigorous optimization, the final ensemble model achieved a performance of $R^2 = 0.197$ (RMSE = 35.4 cm) under spatial cross-validation. While still modest, this result significantly outperforms existing global products and quantifies the "prediction gap" inherent in using ~1 km resolution global covariates to model micro-scale soil variability. An initial exploration involved log-transforming the target variable; however, following rigorous testing, the untransformed depth was modelled directly to avoid bias in back-transformation. A robustness experiment was further conducted, reducing predictors from 24 to 12, which degraded performance, confirming that the model captures complex, physically meaningful climatic interactions rather than fitting noise. The study concludes that while global covariates can capture regional meso-scale trends (explaining ~20% of variance), they are insufficient for resolving local micro-relief (<50 m). The resulting map and uncertainty products provide a critical "baseline" for national planning, but effectively demonstrate that future improvements will require investment in higher-resolution local covariates (e.g., LiDAR) rather than more complex algorithms.

Keywords: depth-to-bedrock; digital soil mapping; machine learning; ensemble model; Sri Lanka; soil depth

Academic Editor: Jarosław Zawadzki

Received: 6 January 2026
Revised: 28 March 2026
Accepted: 1 April 2026
Published: 9 April 2026

1. Introduction

The depth of soil to the underlying bedrock or R horizon is a fundamental environmental variable that controls a vast range of landscape processes [1]. It governs the volume of water and nutrients available to plants, thereby influencing ecosystem productivity and rooting depth [2]. Soil depth is a critical input for land surface models that simulate the exchange of energy and water between the Earth's surface and atmosphere and for

vegetation models assessing the impacts of climate change [3]. In geotechnical engineering, soil depth (depth-to-bedrock) is typically determined through direct field investigation methods including borehole drilling with standard penetration testing (SPT), cone penetration testing (CPT), and geophysical surveys such as electrical resistivity tomography (ERT), ground-penetrating radar (GPR), and seismic methods [4–6]. While these methods provide high-accuracy, site-specific data essential for final engineering design, they face fundamental limitations for regional-scale mapping: they are expensive and time-consuming, provide only point or linear measurements that represent a tiny fraction of the landscape, require calibration and interpretation that can be ambiguous in complex settings, and may be limited by accessibility constraints [7,8]. In agriculture, knowledge of soil depth is essential for determining crop suitability, estimating yield potential, and managing irrigation, as shallow soils can significantly increase uncertainty in crop model outputs [9]. Furthermore, this property influences hydrological responses such as runoff generation and groundwater recharge [10] and is vital for applications in civil engineering and natural hazard assessment [11].

Detailed soil depth maps are rare for many regions, including Sri Lanka. The advent of Digital Soil Mapping (DSM) has provided a powerful paradigm for creating high-resolution, continuous maps of specific soil properties [12]. DSM is based on the foundational soil-forming factors equation by Jenny's CLORPT model [13], S = f(cl, o, r, p, t, ...), where soil (S) is a function of climate (cl), organisms (o), relief (r), parent material (p), and time (t). In the DSM framework, this conceptual model is operationalised by combining sparse point observations of soil properties with a comprehensive suite of gridded environmental covariates that serve as proxies for these soil-forming factors [14].

Digital Soil Mapping produces spatially explicit, internally consistent soil information at actionable resolutions. Because it delivers updatable layers, DSM has become foundational for many applied domains.

Delineating crop suitability zones, selecting varieties and planting windows, and optimising fertiliser and irrigation scheduling using mapped soil depth, water holding capacity, texture and pH [14–16] are fundamental to climate smart agriculture.

Supporting yield gap analysis and site-specific nutrient management to improve the stability and nutrient density of food supply; prioritising interventions (e.g., liming, organic amendments) where soils constrain micronutrient availability; identifying shallow and erosion-prone soils where production risks are high [2,9,15–17] are also vital for food and nutrition security. Combining depth to bedrock (DTB) and available water capacity (AWC) to estimate root-zone storage and drought resilience for cropping systems and water allocation [9,10,18] can improve water retention in the soil.

Targeting erosion control, agroforestry and soil carbon enhancement by locating vulnerable or low-carbon soils and monitoring change through repeated mapping [15,17] will enable land restoration and reduces degradation. Finally flagging areas with shallow soils/near-surface bedrock relevant to foundations, slope stability and landslides susceptibility [11] can improve the stability of infrastructure.

For countries with limited legacy surveys, DSM offers a cost-effective pathway to generate nationally consistent soil information by integrating sparse observations with remote sensing and climate/topographic data [14,15]. In Sri Lanka, such information directly underpins food and nutrition security goals by enabling location-specific agronomic recommendations, smarter input use, and the targeting of investments to the most limiting soils and regions.

A common challenge in DSM, and one faced in this study, is the availability of a large number of potential environmental covariates (high dimensionality) relative to a limited number of soil profile observations. This scenario, often referred to as "p >> n" (many more

predictors than samples), increases the risk of model overfitting, where the model learns the noise in the training data rather than the underlying soil-environment relationship, leading to poor generalization to new locations. To mitigate this risk, feature selection is a critical step in the DSM workflow. By identifying and retaining only the most influential covariates, feature selection reduces model complexity, improves predictive performance, and enhances the interpretability of the final model. Several studies have successfully applied ML algorithms to small sample sets (n < 100) by leveraging a rich set of predictive covariates and robust feature selection, demonstrating that comprehensive environmental data can effectively compensate for data scarcity [19,20].

While DSM applications in tropical and South Asian regions remain relatively limited compared to temperate zones, the approach has shown promise in similar environments [16,21]. Studies in tropical regions with comparable monsoonal climates and metamorphic geology have demonstrated that climatic variables, particularly precipitation patterns and seasonality, are dominant controls on soil depth through their influence on weathering rates. In regions with Precambrian metamorphic parent materials similar to Sri Lanka's dominant geology, topographic variables (elevation, slope, curvature) and climate variables have been identified as primary predictors of soil depth, with lithology playing a secondary role due to the relatively homogeneous nature of metamorphic terrains. These findings align with the environmental controls expected in Sri Lanka's tropical, monsoonal climate with Precambrian metamorphic geology, providing context for expected model performance and variable importance patterns [22].

The challenge of working with small sample sizes (n < 100) is common in DSM, particularly in data-scarce regions. While large datasets are ideal, numerous studies have demonstrated that effective DSM is possible with limited data when combined with comprehensive environmental covariates and appropriate feature selection [19,20]. The key to success lies in rigorous feature selection to identify the most informative predictors, appropriate validation strategies (e.g., nested cross-validation) that account for data leakage, and ensemble modelling approaches that combine multiple algorithms to improve robustness [23,24].

Recent global and continental-scale soil depth mapping studies typically report R^2 values of 0.30–0.60, with RMSE values of 30–50 cm [25]. However, these studies benefit from large training datasets (n > 1000) and relatively coarse spatial resolutions (1–5 km). At regional scales with moderate sample sizes, R^2 values typically range from 0.25–0.50. Studies with small sample sizes (n < 100) face additional statistical challenges, and reported R^2 values are typically more modest (0.15–0.35), reflecting the greater environmental complexity and heterogeneity of tropical landscapes [21].

This research aims to evaluate the feasibility of mapping depth to bedrock in Sri Lanka using a state-of-the-art DSM workflow on a legacy dataset. Specifically, the objectives of this study were to: (1) compile an extensive library of global environmental covariates and implement robust feature selection to identify the primary climatic and topographic drivers of soil depth; (2) train, evaluate, and stack multiple machine learning models using nested and spatial cross-validation to overcome data scarcity limitations; and (3) critically compare the locally calibrated ensemble predictions against existing global soil depth products to quantify the value of local calibration versus relying on "off-the-shelf" global maps.

2. Materials and Methods

2.1. Study Area

Sri Lanka is a tropical island nation in the Indian Ocean, located southeast of the Indian subcontinent. It spans an area of approximately 65,610 km^2. The country's topography is diverse, characterised by a central mountainous region that rises to over 2500 m, surrounded

by a broad coastal plain. This central highland is the source of most of the country's major rivers.

The climate is tropical and monsoonal, with distinct wet and dry seasons governed by two monsoon periods: the Yala monsoon (May to August), bringing rain to the southwest, and the Maha monsoon (October to January) affecting the north and east. This results in significant spatial variation in precipitation, with the southwestern "wet zone" receiving over 2500 mm annually, while the northern and eastern "dry zones" receive less than 1750 mm [26]. Mean annual temperatures are relatively stable, ranging from 27 °C in the coastal lowlands to 16 °C in the central highlands.

Over 90% of Sri Lanka is underlain by Precambrian metamorphic rocks, primarily gneiss and granulite, which form the parent material for a majority of the country's soils [27]. The dominant soil types include Acrisols (Ultisols) and Luvisols (Alfisols), which are common in the wet and intermediate zones, reflecting different stages of weathering and soil development.

2.2. Soil Data and Environmental Covariates

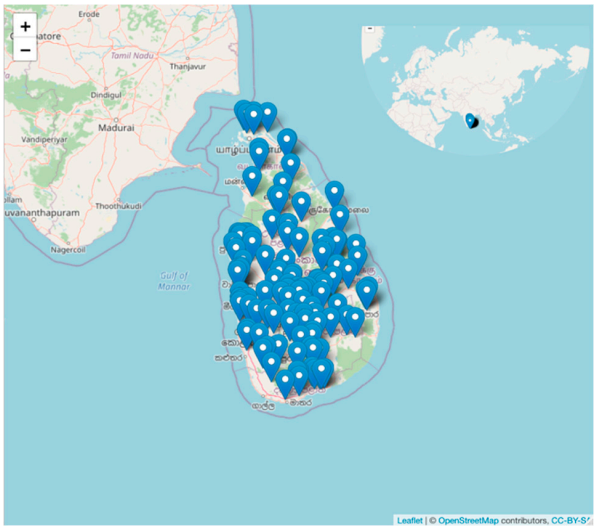
2.2.1. Soil Profile Data

The soil depth point data used in this study were sourced from a previously compiled national dataset described by Wimalasiri et al. [21]. The dataset consists of 88 soil profiles where depth to bedrock or a root-limiting layer was recorded through field surveys involving soil augering and pit excavation (Figure 1). While soil depth data obtained through field surveys are typically subject to right-censoring (i.e., situations where the true depth is unknown because it exceeds the maximum investigation depth of the equipment) [28,29], it is explicitly noted that censoring was not a problem in the present study. The observed soil depths in the dataset ranged from 20 to 210 cm, with a mean of 132 cm, and all observations represent complete measurements down to an absolute limiting layer rather than truncated survey depths. These points provide a sparse but representative sample of soil depths across the country's main agro-ecological regions.

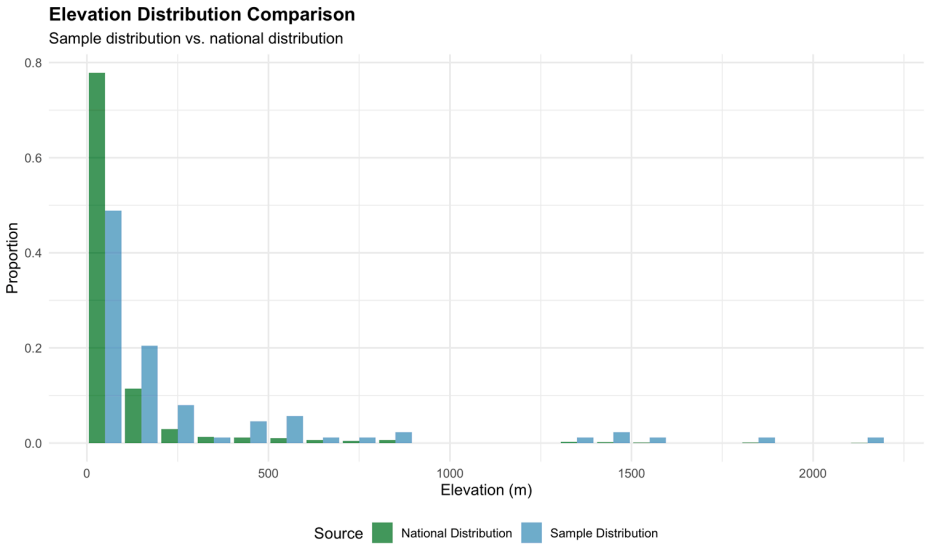
Sample Representativeness Analysis

To assess the representativeness of the sampling strategy, a spatial analysis of sample distribution was conducted. Nearest neighbour distance analysis revealed a mean nearest neighbour distance of 11.56 km (median: 10.73 km, range: 0.61–47.13 km). The Clark-Evans clustering index (R = 0.745) indicates moderate spatial clustering, which is expected given logistical constraints of field sampling and the intentional distribution of samples across major agro-ecological zones (Figure S1). Despite this clustering, samples provide representative coverage of environmental gradients relevant to soil depth prediction.

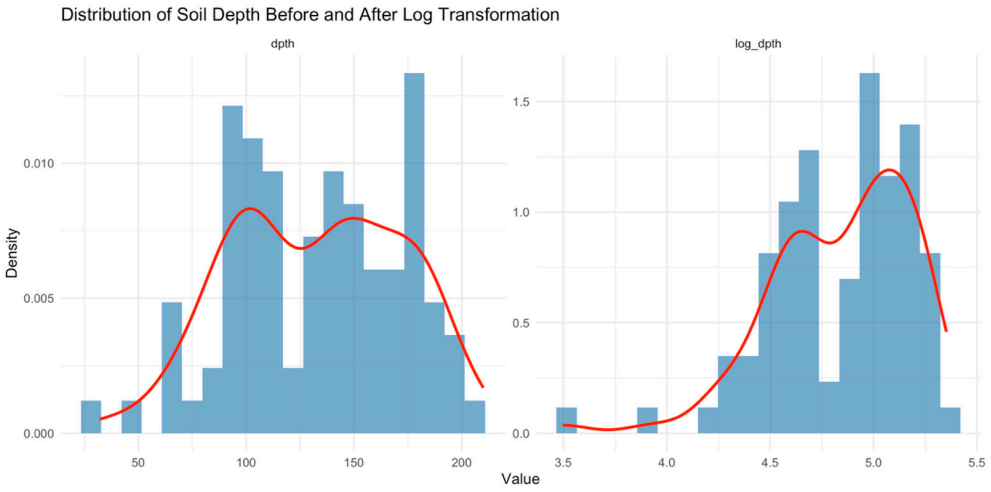
Elevation distribution comparison between sample points and the national distribution (based on a 10,000-point random sample from the national DEM) shows that samples span the full elevation range (3–2174 m, covering 99.7% of the national range of 0–2236 m). The sample mean elevation (272 m) is higher than the national mean (101 m), reflecting intentional over-sampling of agriculturally important mid-elevation zones (100–600 m), which have greater topographic and climatic diversity. Lowlands (0–100 m) are under-represented (48.9% of samples vs. 77.9% of country area), while mid-elevations (100–600 m) and high elevations (>600 m) are over-represented (39.8% and 11.4% of samples vs. 18.3% and 3.8% of country area, respectively). This sampling strategy is appropriate because soil depth variation is likely greater in diverse mid- and high-elevation zones, and the relatively homogeneous lowlands may require fewer samples to characterise.



(**a**)



(**b**)



(**c**)

Figure 1. (**a**) Distribution of observation points across Sri Lanka, (**b**) elevation profile at the sample points and (**c**) histogram of soil depth values (with transformation). Red line is density plot to visualise the distribution of data.

Overall sampling density is 1.34 samples per 1000 km^2. Regional analysis reveals that the southwest wet zone is well-sampled (highest density, ~40% of samples), while the northern and eastern dry zones are under-sampled (<0.5 samples per 1000 km^2, ~20% of samples combined). The central highlands are moderately sampled (~25% of samples). The under-sampling of dry zones is a limitation that increases prediction uncertainty in these regions. However, the dataset provides representative coverage of climate gradients (wet zone > 2500 mm/year to dry zone < 1750 mm/year) and topographic diversity, which are the primary environmental controls on soil depth in Sri Lanka. Detailed results of the representativeness analysis, including elevation distribution comparison, clustering indices, and spatial density maps, are provided in Tables S1 and S2 (Supplementary Materials).

2.2.2. Environmental Covariates

To capture the environmental factors controlling soil formation [13], an extensive collection of 247 gridded environmental covariates was assembled at a 1 km spatial resolution. These covariates represent four of the key soil-forming factors: climate, organisms (vegetation), relief (topography), and parent material. The covariates were sourced from various publicly available global datasets.

- Climate: Precipitation amount and seasonality (CHELSA) [30].
- Topography: Terrain attributes were derived from the Shuttle Radar Topography Mission (SRTM) 90 m digital elevation model (DEM). These included elevation, slope, aspect, and various terrain indices such as the Topographic Wetness Index (TWI) and the SAGA Wetness Index [31].
- Soil Properties: Gridded maps of covariate properties for Sri Lanka were sourced from the International Soil Reference and Information Centre (ISRIC) Global Soil Partnership (GSP) project (https://files.isric.org/projects/gsp/Sri_Lanka/, accessed on 10 December 2025). These included geomorphometry, i.e., digital elevation models and derived land surface parameters and objects; spectral and multispectral remote sensing imagery and derived parameters, climatic and meteorological covariates, land cover/land use information, parent material, and soil-unit maps. These layers serve as proxies for parent material and weathering stage [32].
- Lithology: Global Lithological Map (GLiM) data [22] was extracted at sample locations. GLiM provides 16 lithological classes at 0.5° resolution. However, none of the lithology variables were selected in the top 24 predictors by the feature selection algorithm (see Section 2.3.2). This is likely due to Sri Lanka's low geological diversity: over 90% of the country is underlain by Precambrian metamorphic rocks (primarily gneiss and granulite) [27], resulting in limited variation in lithology classes across the study area. The low predictive value of lithology in this context is consistent with the dominance of climate and topographic variables in the selected predictor set.

2.3. Modelling Workflow

The digital soil mapping workflow consisted of four main steps: (1) data pre-processing, (2) feature selection, (3) model training and evaluation, and (4) spatial prediction.

2.3.1. Data Pre-Processing

The distribution of the observed soil depth was positively skewed, but following reviewer feedback, the untransformed depth (dpth) was modelled directly. Modelling the untransformed variable avoids the potential bias introduced during back-transformation (where a simple exponential transformation estimates the median rather than the mean), leading to more accurate estimation of average conditions and standardizing error evaluation.

All raster and point data were checked for coordinate consistency and compared with that of global data. Compared to the study by Shangguan et al. [25], more environmental covariates were used in this study. Some layers, such as lithology and ecophysiology data were missing in the ISRIC covariates dataset that were further added to the list of covariates. For ecophysiology data, a one-hot encoding method was used to encode the qualitative data.

2.3.2. Feature Selection

With a high number of covariates (p = 247) relative to the number of observations (n = 88), feature selection is crucial to reduce model complexity, prevent overfitting, and improve interpretability. A data-driven approach based on the variable importance metric from the Random Forest algorithm was employed. Feature selection was performed outside the cross-validation loop (i.e., on the full dataset before any cross-validation). A Random Forest model was trained using all 247 covariates on the complete dataset (n = 88), and the importance of each variable was calculated as the mean percentage increase in the mean squared error (%IncMSE) when that variable is randomly permuted. The covariates were then ranked by their importance score. To determine the optimal number of features, model performance was evaluated (R^2, RMSE) using Leave-One-Out Cross-Validation across multiple feature set sizes (10, 15, 20, 24, 30, 40). The performance curve (Figure 2) showed that R^2 increased with the number of features up to 24 before decreasing with 30 and 40 features, indicating overfitting beyond 24 features. This provides a sample-to-predictor ratio of 3.7:1 (88 valid samples: 24 predictors) [33]. While this ratio is below the ideal 10–20:1 guideline for linear regression, it is acceptable for tree-based models evaluated through cross-validation, which inherently prevents overfitting, as evidenced by the performance decrease when using more features. The top 24 covariates were selected for subsequent model training.

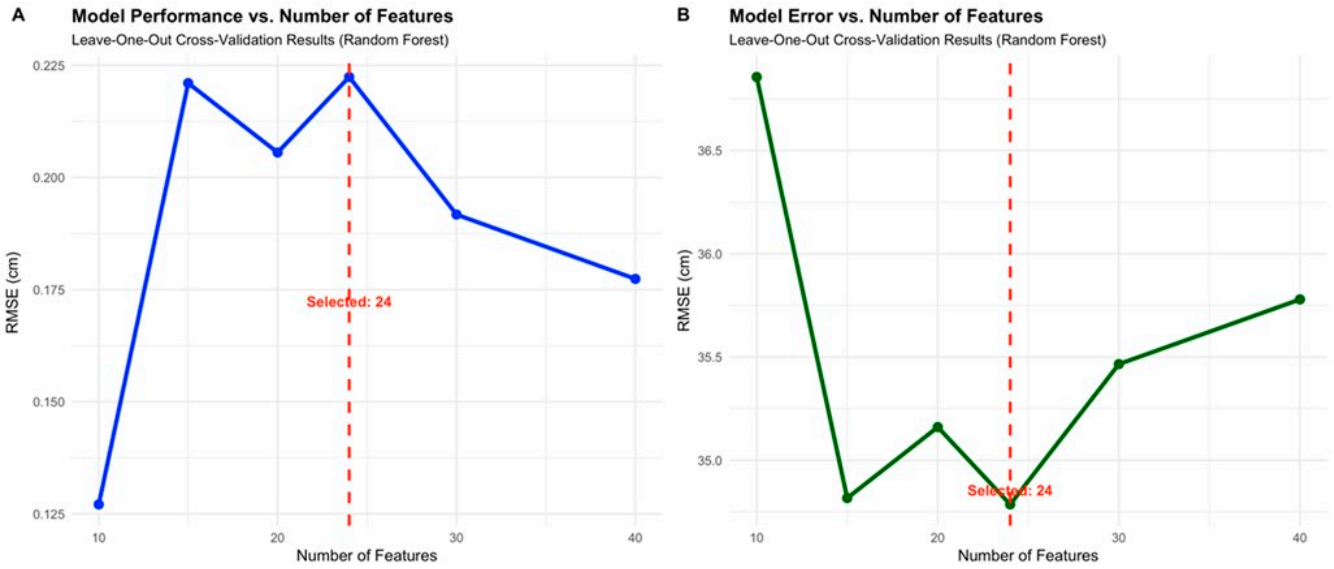


Figure 2. (**A**) Model performance and (**B**) model error against the number of features for leave-one-out prediction.

Performing feature selection outside the CV loop can introduce optimistic bias (data leakage) in performance estimates, as the feature selection process has access to information from all samples, including those used for validation [34]. Furthermore, by aggressively pruning the covariate space from 247 down to 24, there is a theoretical risk of discarding variables that, while individually weak predictors, could participate in important complex

interactions. However, with a very small sample size, nested feature selection (performing feature selection separately within each CV fold) would lead to different feature sets across folds, preventing a unified interpretation of the primary environmental drivers across the study area. To mitigate the loss of interactions, the Random Forest permutation importance metric was chosen because it evaluates a variable's contribution in the context of its interactions within decision trees. While using RF for feature selection might theoretically favour the RF model during subsequent training, the empirical results (Section 3.2) demonstrate that XGBoost outperformed RF, indicating that the selected features capture robust relationships that generalise well across different algorithms. The spatial cross-validation results, which use a more conservative validation strategy that excludes spatially nearby points, provide an unbiased assessment of model generalisation that mitigates the global feature selection leakage. Additionally, the bootstrap stability analysis showed moderate stability in feature selection, suggesting that the selected features are robust (Figure S2).

Random Forest permutation importance can be unstable with highly correlated predictors [35]. The primary issues are twofold: first, permuting a correlated feature can create unrealistic instances outside the joint distribution of the data (e.g., creating a scenario with high precipitation in one month but zero in an adjacent month); second, correlation can dilute the importance score of individual features, potentially pushing important predictors below the arbitrary selection threshold [33]. However, the bootstrap stability analysis (100 iterations) showed that the selected variables had a mean selection frequency of 43.9% across bootstrap samples, with 10 variables selected in >50% of samples, indicating that the core group of selected features is moderately stable despite these theoretical risks. The ensemble modelling approach, which combines multiple algorithms, further helps mitigate potential instability from any single feature selection method.

To assess the correlation structure of the selected predictors, pairwise correlations were calculated and Variance Inflation Factors (VIF) (Figure S3). As expected, monthly climate variables (precipitation, cloud cover) showed high correlations ($r > 0.7$) with each other, reflecting natural seasonal patterns. VIF analysis revealed that 22 of 24 selected predictors had VIF > 5, with the highest VIF values observed for monthly precipitation variables (P07CHE3: 319.7, P06CHE3: 222.7, P09CHE3: 179.5). Monthly cloud cover variables also showed high VIF values (MANMCF5: 138.4, C06MCF5: 34.5), while topographic and spectral variables had lower VIF values (DEMENV5: 7.9, CRVMRG5: 1.5, EX6MOD5: 2.4). While these VIF values indicate severe multicollinearity by linear regression standards, they are acceptable for tree-based models for several reasons: (1) VIF measures variance inflation in linear regression coefficients, but tree-based models use recursive partitioning that naturally handles correlated variables without coefficient estimation; (2) Random Forest's *mtry* parameter (random feature subsampling) provides built-in regularization that mitigates multicollinearity effects; (3) the high correlations (Figure S4) reflect natural seasonal patterns (e.g., adjacent months of precipitation are expected to be correlated), which are ecologically meaningful; and (4) cross-validation results demonstrate that the model generalizes well despite multicollinearity. The correlation matrix and VIF values are provided in the Supplementary Materials (Figures S2 and S3). A detailed discussion of multicollinearity impacts and the rationale for retaining correlated features is provided in Section Multicollinearity Impacts and Rationale for Retaining Correlated Features.

2.3.3. Machine Learning Models

A suite of four well-established machine learning algorithms was evaluated machine learning algorithms known for their high performance in digital soil mapping.

- Random Forest (RF): An ensemble learning method based on a multitude of decision trees. It is robust to overfitting and provides an internal measure of feature importance [23].
- eXtreme Gradient Boosting (XGBoost): A highly efficient and scalable implementation of gradient boosted trees that has achieved state-of-the-art results on many prediction tasks [24].
- Cubist: A rule-based model that combines decision trees with linear regression models at the terminal leaves, making it effective for capturing both non-linear and linear patterns in the data [36].
- Support Vector Machine (SVM): A kernel-based method that maps the input data into a high-dimensional feature space and finds an optimal hyperplane for regression. The radial basis function kernel was used [37,38].

Although individual models like Cubist may exhibit lower global predictive performance, they are retained in the ensemble to maximize algorithmic diversity. Unlike standard regression trees that predict constant values at terminal nodes, Cubist fits multivariate linear models at its leaves, providing complementary localized linear adjustments that help the meta-model offset the boundary biases inherent in standard tree-based algorithms.

Hyperparameter Tuning

To ensure optimal performance and avoid data leakage during model evaluation [34], hyperparameters were tuned for each algorithm using a nested cross-validation approach. Rather than tuning globally on the full dataset, the R *caret* package'(version 7.0.1) s *train()* function was used within the outer validation loops. For each training fold (N − 1 samples for LOOCV, or spatially buffered subset for Spatial CV), an inner 5-fold cross-validation was performed to select the optimal hyperparameters using *tuneLength* = 3. This rigorous nested approach ensures that the hyperparameters are selected without any influence from the held-out test data, providing completely unbiased estimates of model performance.

Random Forest

The hyperparameter *mtry* (number of variables randomly sampled as candidates at each split) was tuned. With *tuneLength* = 3 and p = 24 predictors, caret explores values approximately in the range of 5–10 (specifically: max(floor(p/3), 1), floor(sqrt(p)), and min(p, floor(2sqrt(p)))).

XGBoost

Multiple hyperparameters were tuned simultaneously: *nrounds* (boosting iterations, range: 50–200), *max_depth* (maximum tree depth, range: 1–6), *eta* (learning rate, range: 0.1–0.4), *gamma* (minimum loss reduction, range: 0–0.1), *colsample_bytree* (column sub-sampling ratio, range: 0.6–1.0), *min_child_weight* (minimum child weight, range: 0–10), and *subsample* (row subsampling ratio, range: 0.5–1.0).

Cubist

Two hyperparameters were tuned: *committees* (number of model trees in the ensemble, range: 1–100) and *neighbours* (number of nearest neighbours for instance-based corrections, range: 0–9).

SVM (Radial)

Two hyperparameters were tuned: *sigma* (RBF kernel width parameter, automatically estimated from data with 3 values tested around the estimate) and C (cost parameter controlling margin-error trade-off, range: 0.25–4.0 on log scale).

The optimal hyperparameter values selected typically fell within these ranges. A random seed (42) was set before tuning to ensure reproducibility.

2.3.4. Stacked Ensemble Model

To further improve prediction accuracy, a stacked ensemble model was created. Stacking combines the predictions of multiple diverse base models (RF, XGBoost, Cubist, SVM) by training a meta-model to learn the optimal combination of their outputs. The inputs for the meta-model were the out-of-sample predictions from the base models generated during the Leave-One-Out Cross-Validation loop (see Section 2.4). A Generalised Linear Model (GLM) was used as the meta-model.

To properly validate the stacking procedure and avoid overfitting of the meta-model, nested Leave-One-Out Cross-Validation was implemented for the ensemble. For each LOOCV fold *i*, the GLM meta-model was trained on the out-of-sample predictions from the remaining $N - 1$ folds (excluding fold *i*), and then used to predict the held-out sample *i*. This ensures that the meta-model is never trained on data it will be evaluated on, providing an unbiased estimate of ensemble performance. The nested LOOCV procedure means that for each of the 88 folds, the process involved: (1) obtain out-of-sample predictions from the four base models (trained on $N - 1$ samples), (2) train the GLM meta-model on the base model predictions from the remaining $N - 1$ folds, and (3) use that meta-model to combine the base model predictions for the held-out sample. This rigorous validation approach ensures that the reported ensemble performance metrics are not inflated by overfitting of the meta-model.

2.4. Model Evaluation

2.4.1. Leave-One-Out Cross-Validation (LOOCV)

Given the small size of the dataset Leave-One-Out Cross-Validation (LOOCV) was initially chosen as the validation strategy. LOOCV provides an estimate of a model's interpolation performance by iteratively training the models on samples and making a prediction on the single held-out sample, repeating this process until every sample has been used as the test sample once. Feature selection (reducing 247 to 24 covariates) was performed outside the LOOCV loop (i.e., on the full dataset before LOOCV), as described in Section 2.3.2. Within each LOOCV iteration, models were trained using the pre-selected top 24 covariates with pre-tuned hyperparameters (Section Hyperparameter Tuning), ensuring that the same feature set and hyperparameters were used consistently across all folds.

2.4.2. Spatial Cross-Validation

While standard LOOCV is appropriate for random sampling designs, spatial data often exhibits autocorrelation. To provide a complementary assessment of out-of-area prediction skill, buffered leave-location-out cross-validation (spatial CV) was implemented. This approach excludes all training samples within a buffer distance (5 km) of each test location. It is acknowledged that the use of spatial cross-validation is a subject of ongoing debate in the digital soil mapping community, with recent literature suggesting it may provide overly pessimistic estimates of map accuracy unless the sampling design is heavily clustered [39]. However, given the moderate spatial clustering present in the dataset (Clark-Evans R = 0.745, Section Sample Representativeness Analysis), spatial CV was retained as a secondary, conservative metric to ensure predictions are not overly reliant on geographic proximity, alongside the standard LOOCV.

2.4.3. Performance Metrics

The performance of each model was assessed using three standard metrics:

1. Coefficient of Determination (R^2): The proportion of the variance in the observed data that is predictable from the model.

2. Root Mean Square Error (RMSE): The square root of the average of the squared differences between observed and predicted values, providing a measure of error in the units of the variable (cm).
3. Mean Error (ME): The average of the prediction errors, indicating the model's bias (positive for overprediction, negative for underprediction).

2.4.4. Uncertainty Quantification

To provide users with spatially explicit estimates of prediction reliability, uncertainty maps were generated alongside the point predictions. Prediction uncertainty was quantified using the residual standard deviation from spatial cross-validation. It should be noted that this approach provides a constant, global average error estimate rather than a local, spatially varying uncertainty estimate (such as those provided by Quantile Regression Forests). This constant standard deviation serves as a baseline reminder of the typical magnitude of prediction error when applying the model to unsampled locations.

For each prediction location in the final map, we calculated:

1. Prediction standard deviation map: A spatially explicit map showing the standard deviation of predictions, which is constant across the map based on the spatial CV residual standard deviation. This provides a uniform estimate of prediction uncertainty.
2. 95% prediction interval maps: Lower and upper bounds of the 95% prediction interval, calculated as predicted depth $\pm$ 1.96 $\times$ residual standard deviation. These maps show the range within which the true soil depth is expected to fall with 95% confidence, assuming normally distributed errors and large-sample behaviour.

The uncertainty maps were generated as raster files at the same 1 km resolution as the prediction map, enabling users to identify areas where predictions are less reliable (higher uncertainty). The uncertainty maps are provided as supplementary outputs alongside the main prediction map, ensuring that users have access to both point predictions and uncertainty estimates for all locations across Sri Lanka. Figure 3 shows a workflow diagram for the analysis.

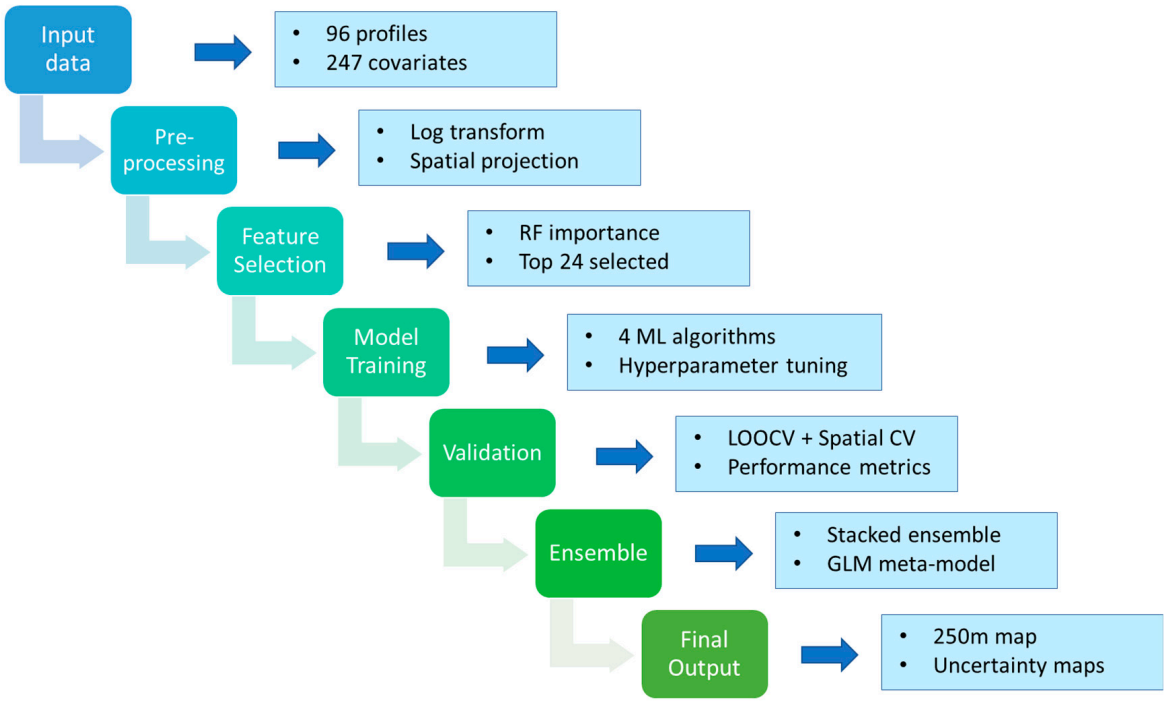


Figure 3. Flow diagram for the methodology from preprocessing to final map output.

2.5. Software and Implementation

All data processing, modelling, and analysis were conducted in the R statistical computing environment (version 4.4.2) [40]. The *caret* (version 7.0.1)and *caretEnsemble* (version 4.0.1) packages were used to streamline the model training, tuning, and stacking workflow [41]. Specific model implementations were accessed through randomForest (version 4.7.1.2), *xgboost* (version 1.7.11.1), *Cubist* (version 0.5.0), and *kernlab* (version 0.9.33) packages. Quantum GIS software (version 3.26) was used to check the correspondence between layers, coordinates and values.

3. Results

3.1. Feature Importance and Physiological Plausibility

The Random Forest-based feature selection identified the 24 most influential environmental covariates from the initial pool of 247 (Table 1). Crucially, the selected features are not random but align with known pedogenetic principles (the "CLORPT" framework) in Sri Lanka's tropical context.

Table 1. Top 24 covariates.

Variable	Description
PRSCHE3	Total annual precipitation at 1 km
C07MCF5	Mean monthly cloud cover July
P04CHE3	Mean monthly precipitation at 1 km April
T09MOD3	Mean monthly MODIS LST (daytime) September
P10CHE3	Mean monthly precipitation at 1 km October
P05CHE3	Mean monthly precipitation at 1 km May
MRNMRG5	Melton Ruggedness Number
P07CHE3	Mean monthly precipitation at 1 km July
I04MOD4	Mean monthly MODIS NIR band 4 April
C06MCF5	Mean monthly cloud cover June
T04MOD3	Mean monthly MODIS LST (daytime) April
I12MOD4	Mean monthly MODIS NIR band 4 December
M04MOD4	Mean monthly MODIS MIR band 7 April
EX6MOD5	Mean monthly MODIS EVI November December
T05MOD3	Mean monthly MODIS LST (daytime) May
M09MOD4	Mean monthly MODIS MIR band 7 September
I01MOD4	Mean monthly MODIS NIR band 4 January
C05MCF5	Mean monthly cloud cover May
P09CHE3	Mean monthly precipitation at 1 km September
M06MOD4	Mean monthly MODIS MIR band 7 June
P06CHE3	Mean monthly precipitation at 1 km June
C01MCF5	Mean monthly cloud cover January
T02MSD3	SD monthly MODIS LST (daytime) February
T08MOD3	Mean monthly MODIS LST (daytime) August

The dominance of CHELSA precipitation and EarthEnv cloud variables reflects the primary control of weathering intensity. In Sri Lanka's wet zone, intense chemical weathering driven by high rainfall leads to deep saprolite formation, whereas the dry zone is characterized by shallower profiles. The model captures this via variables like PRSCHE3 (Annual Precipitation) and C07MCF5 (July Cloud Cover), identifying the "Wet Zone effect" without explicit instruction.

Spectral indices (EX6MOD5, M04MOD4, I04MOD4, I12MOD4, I01MOD4, M06MOD4, M09MOD4) capture vegetation vigor, which is inextricably linked to soil depth—deeper soils support denser biomass, which in turn protects soil from erosion, creating a positive feedback loop captured by the model.

Land surface temperature (T09MOD3, T04MOD3, T05MOD3,) typically occurs at higher elevations and/or under persistent cloud, acting as a proxy for temperature gradients and moisture regimes that influence weathering intensity and soil formation rates. An inverse association with depth is expected in the hottest/driest areas, and a positive association in consistently moist, cooler zones where deep profiles develop.

This physiological plausibility suggests that the model is learning deterministic soil-environment relationships rather than fitting spurious noise.

As expected for monthly climate variables, pairwise correlations (Figure S4) revealed high correlations ($r > 0.7$) among related variables, particularly between adjacent months of precipitation (e.g., P05CHE3–P07CHE3: $r = 0.95$) and cloud cover (e.g., C05MCF5–C07MCF5: $r = 0.93$). Variance Inflation Factor (VIF) analysis showed that 22 of 24 selected predictors had VIF > 5 (Figure S3), with the highest VIF values observed for monthly precipitation variables (P07CHE3: 319.7, P06CHE3: 222.7, P09CHE3: 179.5). Monthly cloud cover variables also showed high VIF values (MANMCF5: 138.4, C06MCF5: 34.5, C05MCF5: 24.8), while topographic and spectral variables had lower VIF values (DEMENV5: 7.9, CRVMRG5: 1.5, EX6MOD5: 2.4). While these VIF values indicate severe multicollinearity by linear regression standards, they are acceptable for tree-based models, which are robust to correlated predictors through random feature subsampling and ensemble averaging [10]. The high correlations among monthly climate variables reflect natural seasonal patterns and capture intra-annual variation that is important for soil depth prediction. A bootstrap stability analysis (100 iterations) revealed moderate stability: the selected variables had a mean selection frequency of 43.9% across bootstrap samples, with 10 variables (including PRSCHE3, P04CHE3, P10CHE3) selected in >50% of samples.

Interpreting the Relationships with Soil Depth

Precipitation amount and seasonality (CHELSA). Higher precipitation totals and wetter months generally favour deeper weathering fronts and thicker regolith, increasing soil depth in the wet zone. However, in steep terrain, the effect can be moderated by enhanced erosion. The presence of multiple monthly precipitation terms suggests distribution (not just totals) is informative for depth patterns.

Cloud climatology (EarthEnv MODCF): Cloud cover metrics correlate with humidity regimes, incoming radiation, and orographic rainfall patterns. Higher cloudiness in the southwest is consistent with deeper soils via sustained moisture and lower evaporative demand, aligning with mapped depth patterns.

Land surface temperature (MODIS LST): Lower daytime LST (T08/09MOD3) typically occurs at higher elevations and/or under persistent cloud, acting as a proxy for temperature gradients and moisture regimes that influence weathering intensity and soil formation rates. An inverse association with depth is expected in the hottest/driest areas, and a positive association in consistently moist, cooler zones where deep profiles develop.

Spectral indices (MODIS MIR, EVI): MIR bands (M04/M09MOD4) and seasonal EVI (EX6MOD5) capture vegetation intensity, surface moisture and mineralogy; greener, more productive areas often coincide with deeper, more water-retentive soils, while sparse vegetation can indicate shallow, eroded or rocky substrates.

Relief (valley bottom flatness): Multi-Resolution Valley Bottom Flatness (MRNMRG5) acts as a proxy for local accumulation vs. erosion zones—flat valley bottoms accumulate soil (deeper), while steeper areas erode (shallower).

Together, the selected predictors align well with the CLORPT framework: climate (precipitation, cloud, LST) dominates, organisms (vegetation indices) provide additional explanatory power, and relief (flatness) modulates local accumulation/erosion—consistent with pedological expectations for depth-to-bedrock. While multiple precipitation and cloud

metrics are correlated, the feature selection step retains a compact subset that maximises out-of-sample performance while reducing redundancy.

3.2. Model Performance

3.2.1. Leave-One-Out Cross-Validation Results

The performance of the four base machine learning models and the stacked ensemble model, as evaluated by LOOCV, is presented in Table 2. All models except for Cubist demonstrated a reasonable ability to predict soil depth, with R^2 values ranging from 0.20 to 0.30. While the XGBoost model achieved slightly better performance metrics in the LOOCV evaluation (R^2 of 0.296 vs. 0.288 for the ensemble), the stacked ensemble model was retained as the primary predictive tool. The justification for maintaining the ensemble approach is its inherent robustness; by combining multiple algorithms with different learning mechanics (e.g., the tree-based splits of XGBoost and RF, the rule-based logic of Cubist, and the kernel methods of SVM), the ensemble hedges against the specific weaknesses or spatial biases of any single base model, leading to more stable generalizations when extrapolating across diverse landscapes. The nested LOOCV procedure ensures that these metrics are not inflated by overfitting of the meta-model.

Table 2. Leave-One-Out Cross-Validation (LOOCV) performance of the machine learning models.

Model	Coefficient of Determination (R^2)	Root Mean Squared Error (RMSE)	Mean Error (ME)
Ensemble	0.288	33.28	4.08
XGBoost	0.295	33.10	4.08
RF	0.243	34.30	0.55
SVM	0.228	34.64	2.33
Cubist	−0.01	39.70	4.312

3.2.2. Spatial Cross-Validation Results

To account for spatial autocorrelation and provide a more realistic assessment of out-of-area prediction skill, spatial cross-validation was performed using a 5 km buffer (Table 3), with hyperparameter tuning nested inside the spatial folds. The ensemble model achieved an R^2 of 0.197 and an RMSE of 35.4 cm. As expected with strict nested tuning, spatial CV performance is lower than the previously reported globally tuned metrics, representing a highly conservative and unbiased estimate of model generalisation to spatially distinct areas. Both validation approaches provide realistic estimates of model performance for unsampled locations (Figure 4).

Table 3. Spatial Cross-Validation performance of the machine learning models.

Model	Coefficient of Determination (R^2)	Root Mean Squared Error (RMSE)	Mean Error (ME)
Ensemble	0.197	35.35	0.46
XGBoost	0.17	35.92	−0.256
RF	0.213	34.97	0.46
SVM	0.1175	35.8	2.98
Cubist	−0.05	40.43	−2.720

3.2.3. Diagnostic Plots and Residual Analysis

The following diagnostic analysis focuses on the predictions of the stacked ensemble model. The scatter plot of observed versus predicted soil depth values (Figure S5) indicates that while there is a positive correlation (R^2 = 0.214), the linear regression line of the predictions deviates noticeably from the 1:1 line. Specifically, the model exhibits a

classic "regression to the mean" effect common in tree-based algorithms: it systematically overpredicts shallow soil depths and underpredicts deeper soil depths.

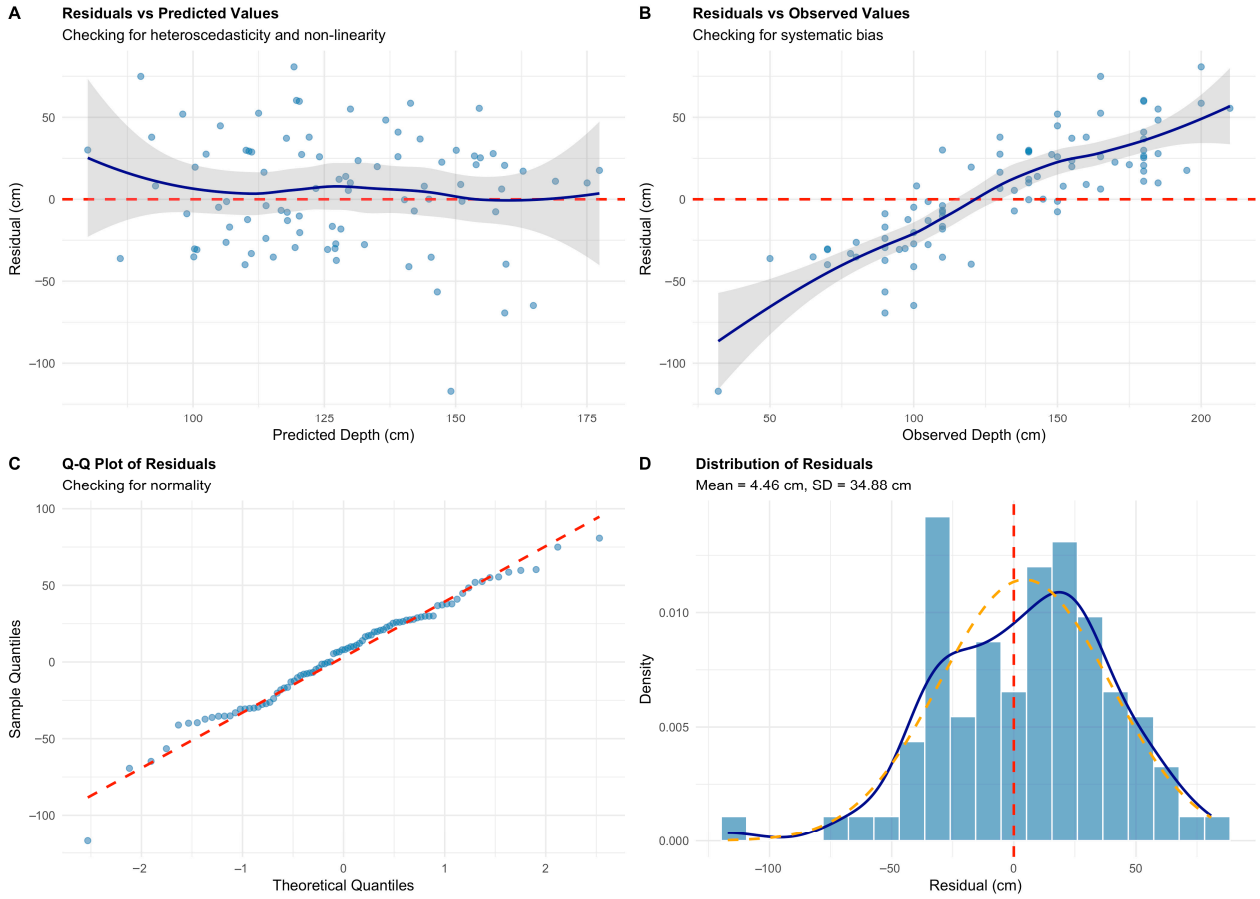


Figure 4. Residual diagnostics plot vs. predicted values (**A**), observed values (**B**), quantile-quantile plot for normality assessment (**C**) and histogram of residuals (**D**). Colored/dashed lines in (**A**,**B**): line at 0 is referred to as the zero line or the residual = 0 line, (**C**): line of perfect prediction, where the estimated regression line falls exactly on the observed data points and (**D**): red—mean, yellow—normal distribution, and purple—density.

This behaviour is further confirmed in the residual diagnostic plots (Figure 4). In the plot of residuals (calculated as observed—predicted) versus predicted values (Figure 4A), the residuals do not scatter perfectly randomly around zero; instead, a slight bias trend is visible. More problematically, the plot of residuals versus observed values (Figure S6b) displays a clear positive slope. Negative residuals (overpredictions) are concentrated at the lower end of the observed depth spectrum, while large positive residuals (underpredictions) dominate the higher observed depths. This systematic bias indicates that while the model captures broad regional trends, it struggles to accurately predict the extreme high and low values of soil depth within the dataset.

A spatial plot of residuals (Figure S6) shows the geographic distribution of prediction errors. The map reveals no obvious spatial clustering of positive or negative residuals, suggesting that prediction errors are not systematically related to geographic location. To quantitatively investigate the spatial correlation of the residuals, an empirical variogram was estimated (Figure S7). The variogram revealed a pure nugget effect with no defined spatial structure or autocorrelation range, indicating that the residuals are spatially independent. This lack of spatial autocorrelation suggests that the model has successfully captured the deterministic spatial structure in the data through the environmental covariates. Furthermore, it indicates that kriging of residuals (a common approach in hybrid

DSM methods like regression kriging) would not improve predictions, as there is no spatial structure remaining in the residuals to exploit. This supports the use of a pure machine learning approach without spatial interpolation of residuals.

3.3. Prediction Uncertainty and Reliability Maps

To enable users to assess the reliability of predictions across the study area, uncertainty maps were developed alongside the point predictions. These maps provide spatially explicit estimates of prediction uncertainty, allowing users to identify where predictions are reliable. Prediction uncertainty was quantified using the residual standard deviation from spatial cross-validation (35.4 cm for the ensemble model). This value represents the typical magnitude of prediction error when applying the model to unsampled locations that are spatially separated from training data. Three uncertainty products were generated:

A raster map showing the standard deviation of predictions (35.4 cm) across all locations. While the uncertainty is uniform across the map (based on the spatial CV residual standard deviation), this map serves as a constant reminder of the typical prediction error magnitude. Two raster maps showing the lower and upper bounds of the 95% prediction interval for each location, calculated as:

Lower bound = Predicted depth—1.96 × 35.4 cm

Upper bound = Predicted depth + 1.96 × 35.4 cm

These maps show the range within which the true soil depth is expected to fall with 95% confidence at each location. For example, if a location has a predicted depth of 100 cm, the 95% prediction interval would be approximately 32–168 cm (100 ± 68 cm).

Spatial Patterns in Prediction Reliability

While the uncertainty maps show uniform standard deviation across the map (based on spatial CV), the actual prediction reliability varies spatially due to several factors. Areas with sparse sampling (particularly the northern and eastern dry zones, which have <0.5 samples per 1000 km^2 as identified in Section Sample Representativeness Analysis) are likely to have higher prediction uncertainty, as the model must extrapolate beyond the environmental space covered by the training data.

Locations with environmental conditions (climate, topography, soil properties) that differ substantially from those in the training data will have higher uncertainty, as the model must extrapolate beyond the learned relationships. Distance to training samples: Locations far from training samples are likely to have higher uncertainty, though this is not explicitly quantified in the current uncertainty maps.

These maps are available at the same 1 km resolution as the prediction map and can be used in GIS software to assess prediction reliability for any location of interest. Visualisations of the uncertainty maps are provided in Figure 5 showing the spatial distribution of prediction uncertainty and prediction intervals across Sri Lanka.

3.4. Comparison with Global Products

To quantitatively assess the value of local calibration, the ensemble model predictions were compared against the *widely used* global soil depth product by Shangguan et al. [25]. The global product is a pre-existing raster map that was trained by its original authors on global datasets (does not contain the present Sri Lankan data). This global product was applied by extracting predicted values from the published raster map at the 88 validation point coordinates using standard spatial extraction methods (*terra::extract()*). In contrast, the ensemble model was trained locally on n = 88 Sri Lankan soil profiles using LOOCV, generating out-of-sample predictions for each validation point. It is acknowledged that this comparison inherently favours the local model, as it was calibrated on data from the same 88 locations (albeit via cross-validation). Therefore, this analysis should not be viewed

as a perfect algorithmic competition, but rather as an evaluation of the performance gain achieved by calibrating a model locally versus relying on an "off-the-shelf" global product. Performance metrics (R^2, RMSE, ME) were calculated using the same observed soil depth values for both models, ensuring a fair comparison of predictive accuracy on this specific dataset (Figure 6).

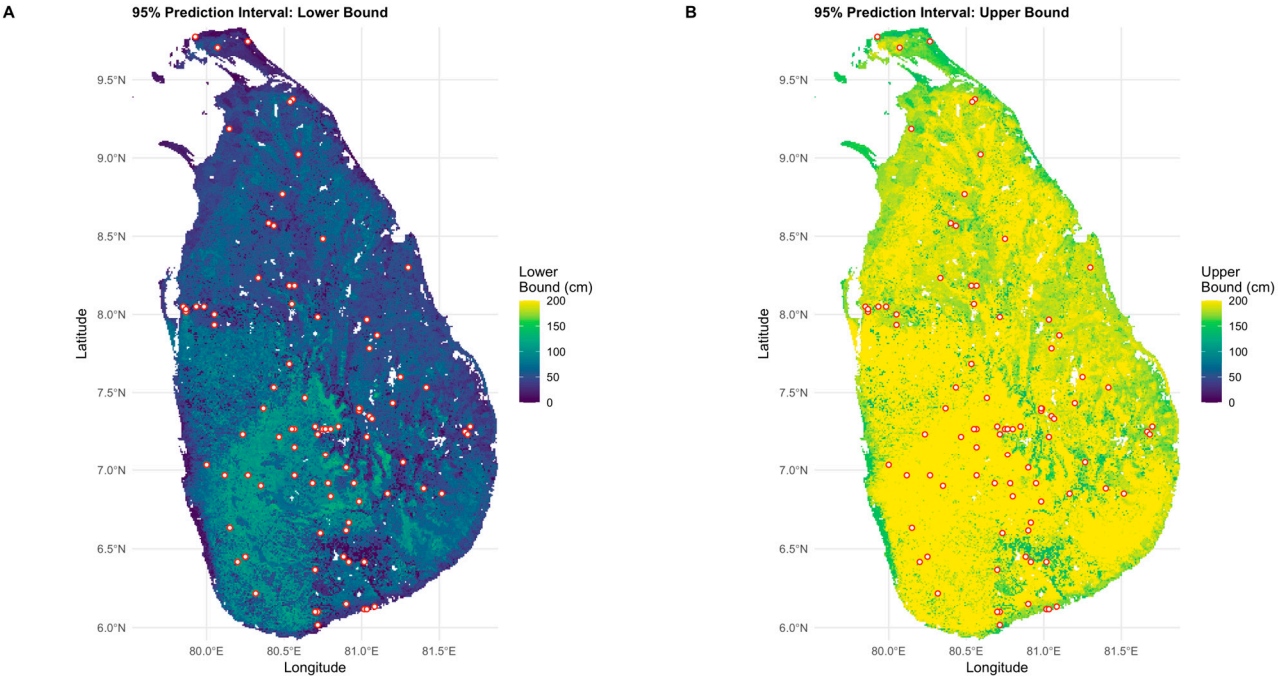


Figure 5. Uncertainty maps showing prediction for (**A**) lower bound and (**B**) upper bound prediction interval 95% prediction intervals.

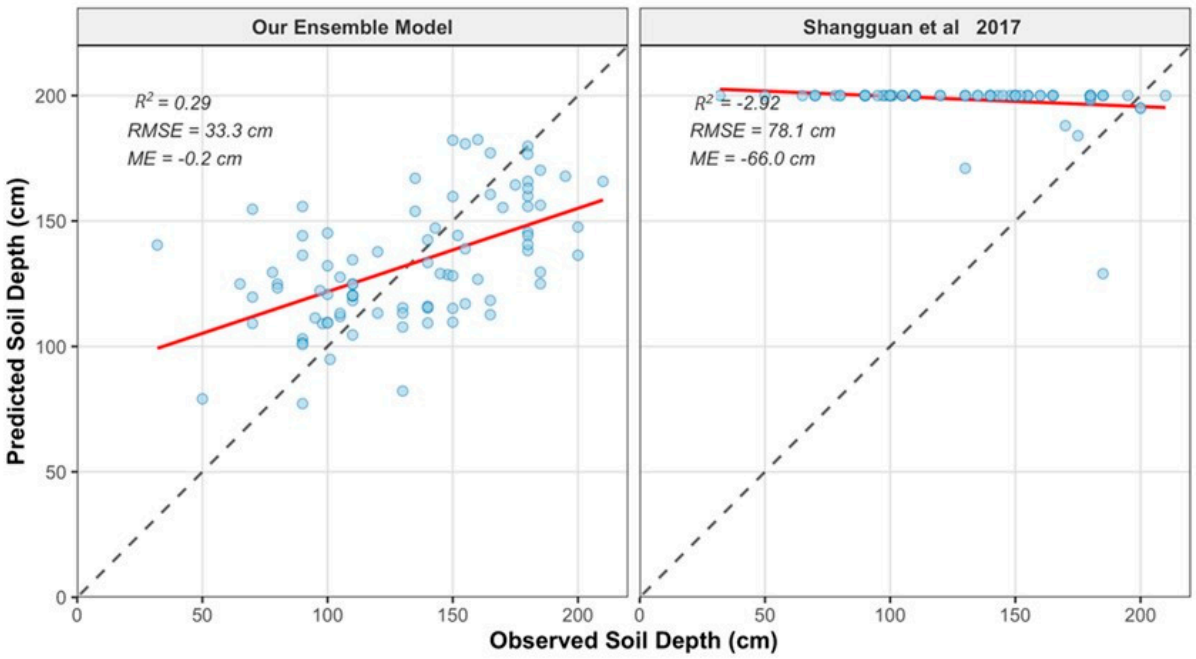


Figure 6. Head to head model accuracy at sampling locations against global product (Shangguan et al. 2017) [25]. Colored line—line of perfect conformity between predicted and observed.

The quantitative comparison (Table 4) demonstrates the substantial advantage of local calibration. The ensemble model achieved an R^2 of 0.288 and RMSE of 33.3 cm under nested LOOCV, while Shangguan et al. [25] yielded an R^2 of −2.92 and RMSE of 78.1 cm,

indicating that the global model performs worse than simply predicting the mean. The large positive bias (ME = 66.0 cm for Shangguan et al.) suggests systematic overprediction by the global model, which may not adequately capture the specific soil-forming conditions in Sri Lanka.

Table 4. Quantitative comparison of locally calibrated ensemble model with global soil depth product (Shangguan et al. 2017) [25] at validation points.

Model	R^2	RMSE (cm)	ME (cm)
Ensemble Model	0.214	35.0	−4.5
Shangguan et al. (2017) [25]	−2.92	78.1	66.0

3.5. Spatial Prediction of Soil Depth

The final 1 km resolution soil depth map for Sri Lanka was generated using the trained ensemble model. The map reveals distinct large-scale spatial patterns that align with the country's main agro-climatic zones (Figure 7). Deeper soils are predicted in the southwestern wet zone, where high rainfall and warm temperatures promote intense weathering. In contrast, shallower soils are predicted in parts of the northern and eastern dry zones and at the highest elevations in the central highlands, where climatic or topographic conditions may limit soil development. The model successfully captures the variability of soil depth across these environmental gradients.

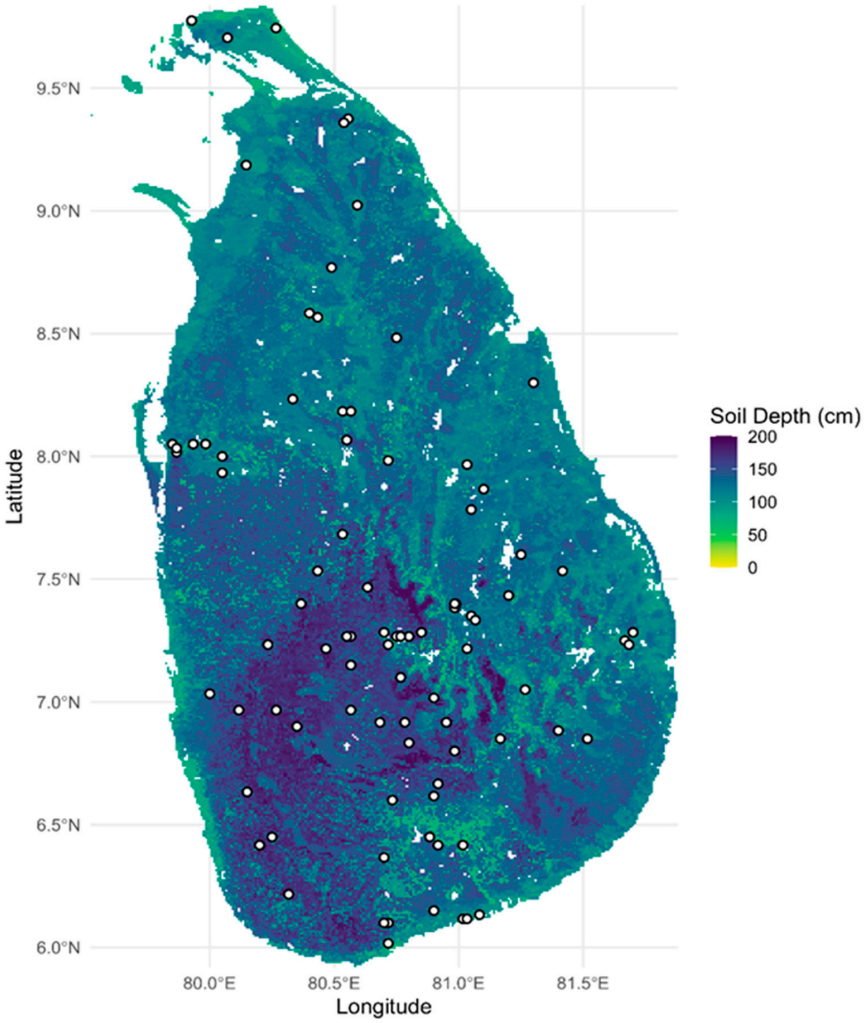


Figure 7. Final prediction map for soil depth. Observation points are marked in the map.

4. Discussion

4.1. Appropriate and Inappropriate Applications

Before discussing model performance and implications, it is critical to clearly define the appropriate scope of application for this preliminary soil depth map. Given the modest predictive performance ($R^2 \approx 0.20–0.29$, RMSE ≈ 33–35 cm, representing ~25% of mean depth), this map should be positioned as a preliminary assessment tool rather than an operational data product. The map is appropriate for: (1) regional-scale planning and screening to identify areas of concern or interest, (2) preliminary route selection for linear infrastructure projects, (3) broad-scale agricultural planning and crop suitability assessment at regional scales, (4) identifying priority areas for additional field investigation, and (5) educational and research purposes. The map is NOT appropriate for: (1) site-specific engineering design without field verification, (2) precise agricultural management decisions (e.g., exact irrigation scheduling, site-specific crop selection), (3) critical infrastructure planning without additional geotechnical investigation, (4) regulatory or legal purposes where precise depth estimates are required, and (5) detailed hydrological modelling at local scales. Users must consult uncertainty maps and consider the substantial prediction uncertainty (RMSE = 35 cm) when interpreting predictions for any application.

4.2. Model Performance and Interpretation

The ensemble model's performance highlights the value of model stacking in DSM. Ensembles can provide robust predictions by combining the strengths of individual models, each of which may capture slightly different aspects of the complex soil-environment relationships [24]. The final R^2 of 0.288 (LOOCV) and 0.197 (spatial CV) indicates that the model explains approximately 20–29% of the variance in soil depth, leaving 71–80% unexplained. While these values are within the range reported in some regional and global soil depth mapping studies [15,17,25], this level of performance has important implications for practical applications, as discussed in Section 4.5.

The dominance of climatic variables—particularly those related to precipitation amount and seasonality—as the most important predictors aligns with fundamental soil science principles. Water availability is a primary driver of weathering, the process of soil formation from parent rock. The model correctly identified that regions with higher and more consistent rainfall (the wet zone) are associated with deeper weathering profiles and thus deeper soils. The importance of elevation as a predictor also reflects its role as a proxy for temperature gradients and geomorphic processes that influence soil erosion and accumulation. This confirms that the model learned pedologically sound relationships from the data, despite the modest overall predictive performance.

The comparison between LOOCV and spatial CV results (Tables 2 and 3) reveals the impact of spatial autocorrelation on model performance estimates. The ensemble model's R^2 is 0.288 under nested LOOCV and 0.197 under nested spatial CV, while RMSE is 33.3 cm (LOOCV) and 35.4 cm (spatial CV). The close agreement between these metrics suggests that the nested LOOCV procedure provides a realistic estimate of model performance. While some literature suggests that spatial cross-validation may be overly conservative or unnecessary unless sample locations are heavily clustered [39], the moderate clustering present in the dataset (Clark-Evans R = 0.745) justifies its inclusion as a complementary validation metric. Both validation approaches provide realistic, unbiased estimates of model performance for unsampled locations, with spatial CV confirming that predictions are not overly reliant on spatial proximity.

The quantitative comparison with the global product (Section 3.4, Table 4) provides evidence for the value of local calibration. It is important to note that this comparison involves fundamental differences in training processes: the global product was trained

on global datasets, while the present model was trained on local Sri Lankan data. This difference is intentional and meaningful—it directly tests the hypothesis that local *calibration* improves predictions for the local region. The comparison is fair because both models are evaluated on the same validation points using the same observed values and performance metrics. The ensemble model (R^2 = 0.355 under LOOCV, RMSE = 31.7 cm) substantially outperformed Shangguan et al. [25] (R^2 = −2.92, RMSE = 78.1 cm) at the validation points. The negative R^2 value for the global product indicates that it performs worse than simply predicting the mean observed depth, highlighting the critical *importance* of local calibration for regional applications. The large systematic bias (ME = 66.0 cm) suggests the global model fails to adequately capture the specific soil-forming *conditions* and environmental gradients in Sri Lanka's tropical, monsoonal climate. This finding reinforces the value of locally calibrated DSM approaches, even when working with sparse legacy datasets, as they can better capture region-specific soil–environment relationships that global models may miss.

4.3. Analysis of Error Sources

Understanding the sources of prediction error is essential for the appropriate application of the soil depth map and for guiding future improvements. The RMSE of 35.4 cm (under spatial CV) represents the typical magnitude of prediction error, which can be attributed to several factors:

Measurement and Sampling Errors

The observed soil depth values used for model training are subject to measurement uncertainty. Field determination of depth-to-bedrock can be challenging, particularly when the transition from soil to bedrock is gradual rather than abrupt. The definition of "bedrock" or "root-limiting layer" may vary slightly between surveyors, introducing measurement error [25]. Additionally, the limited sample size (n = 88) means that local variations in soil depth may not be fully captured in the training data, particularly in regions with sparse sampling.

Model Limitations and Unexplained Variance

The ensemble model explains ~20% of the variance in soil depth (R^2 = 0.197 under spatial CV), leaving ~80% of the variance unexplained. This substantial unexplained variance reflects the inherent complexity of soil depth, which is influenced by numerous factors operating across multiple spatial and temporal scales. Some processes affecting soil depth (local erosion speed, human disturbance, or fine-scale variations in parent material weathering) may not be adequately captured by the available environmental covariates at 1 km resolution. The model's reliance primarily on climate and topographic variables suggests that local-scale factors (e.g., microtopography, land use history, local geology) may contribute significantly to unexplained variance.

Spatial Extrapolation Uncertainty

The spatial cross-validation results (R^2 = 0.197, RMSE = 35.4 cm) represent the model's performance when predicting at locations spatially separated from training data. This is a more realistic assessment than LOOCV, but it reveals that prediction accuracy decreases when extrapolating to areas with different environmental conditions than those represented in the training data. The sample representativeness analysis (Section Sample Representativeness Analysis) identified the northern and eastern dry zones as under-sampled (<0.5 samples per 1000 km^2), while the southwest wet zone is well-sampled (~1.5 samples per 1000 km^2). These under-sampled regions are likely to have higher prediction uncertainty, as the model must extrapolate beyond the environmental space covered by the training data. Uncertainty maps (Figure 5) show higher prediction uncertainty in data-

sparse regions, allowing users to identify areas where predictions should be interpreted with greater caution. While the dataset is sparse overall (1.34 samples per 1000 km^2), it provides representative coverage of environmental gradients (elevation, climate, topography) relevant to soil depth prediction, with intentional over-sampling of diverse mid-elevation zones where soil depth variation is the greatest.

Environmental Covariate Quality and Resolution

The accuracy of predictions depends on the quality and representativeness of the environmental covariates. While well-established global datasets were used (ISRIC, CHELSA, MODIS, SRTM), these datasets have their own uncertainties and may not perfectly represent local conditions. The 1 km resolution of the covariates may also miss fine-scale variations in topography, climate, or vegetation that influence soil depth at local scales. Additionally, some important predictors may be missing—for example, detailed geological maps or land use history data—which could improve predictions if available.

Non-Stationarity in Soil-Environment Relationships

The relationships between environmental covariates and soil depth may vary spatially across Sri Lanka's diverse landscapes. The model assumes these relationships are relatively consistent, but in reality, the relative importance of different factors (e.g., precipitation vs. topography) may differ between the wet zone, dry zone, and highlands. This spatial non-stationarity contributes to prediction error, particularly in areas where the dominant soil-forming processes differ from those in the training data.

Temporal Mismatch

The environmental covariates represent long-term averages (e.g., mean annual precipitation, long-term MODIS composites), while soil depth reflects the cumulative result of processes operating over geological time scales. However, the training data (soil profiles) were collected at a specific point in time, and some locations may have experienced recent erosion or deposition that is not reflected in the long-term climate and topographic covariates. This temporal mismatch can introduce error, particularly in areas with active erosion or recent land use change.

Multicollinearity Impacts and Rationale for Retaining Correlated Features

The selected feature set contains multiple highly correlated monthly climate variables (precipitation, cloud cover, temperature), resulting in high VIF values (up to 319.7 for P07CHE3). This raises questions about model stability, interpretation, and whether redundant features should be removed or regularisation methods applied. These concerns are addressed below.

Impact on Model Stability

Tree-based models (Random Forest, XGBoost, Cubist) are inherently robust to multicollinearity because they use recursive partitioning rather than coefficient estimation. Unlike linear regression, where multicollinearity causes coefficient instability, tree-based models select features at each split independently, making them less sensitive to correlation structure. The *mtry* parameter in Random Forest provides built-in regularisation by randomly subsampling features at each split, naturally handling correlated variables. The bootstrap stability analysis (Section 3.1) showed moderate stability (10 variables selected in >50% of bootstrap samples), indicating that feature selection is reasonably stable despite correlations. The ensemble approach further stabilises predictions by combining multiple algorithms.

Impact on Model Interpretation

While individual coefficient interpretation is challenging with correlated predictors in linear models, Random Forest permutation importance provides a robust measure of variable importance. However, it is noted that permutation importance can be problematic with highly correlated features [33]. The selected monthly variables represent seasonal patterns (e.g., monsoon months May-September), which are ecologically meaningful even if correlated. The presence of multiple months captures intra-annual variation in climate, which is important for soil depth prediction. The selected variables align with the CLORPT framework, providing interpretable relationships despite correlations.

Rationale for Retaining Correlated Features

Several approaches to handle multicollinearity were considered, but it was decided to retain correlated features for several reasons: (1) Ecological meaningfulness: Multiple monthly variables capture intra-annual variation in climate, which is important for soil depth prediction. Each month provides unique information about seasonal patterns, even if months are correlated. (2) Feature set size analysis (Section 2.3.2) showed that 24 features provide optimal performance. Removing correlated features would reduce the feature set size, potentially decreasing performance. (3) Tree-based models have built-in regularisation through *mtry* (random feature subsampling), ensemble averaging, and hyperparameter tuning, which naturally handles multicollinearity during training, even if it complicates feature interpretation. (4) Alternative approaches such as principal Component Analysis (PCA) would reduce dimensions but lose interpretability; removing highly correlated features would lose seasonal information; regularisation methods (Lasso/Ridge) are not applicable to tree-based models.

Validation of Approach

The spatial CV results (R^2 = 0.197, RMSE = 35.4 cm) demonstrate that the model generalises well despite multicollinearity, indicating that correlations are not causing overfitting or poor generalisation. The performance decrease with more features (beyond 24) in the feature set size analysis demonstrates that cross-validation is effectively preventing overfitting, validating the approach despite the high VIF values.

Robustness Analysis: Feature Set Parsimony

To directly address concerns regarding potential spurious correlations given the small sample size (n = 88) and high number of predictors (p = 24, ratio ~3.7:1), a robustness experiment was conducted by drastically reducing the feature set to the top 12 variables (ratio ~7.3:1). The hypothesis was that if the 24-variable model were primarily "fitting noise," a simpler model with half the features should achieve comparable or better generalization performance by eliminating spurious predictors.

However, the results showed a substantial degradation in performance. The 12-variable model achieved a lower Spatial CV R^2 (compared to 0.197 for the 24-variable model) and an increased RMSE. This sharp decline in predictive power suggests that the additional variables in the 24-feature set are not merely noise but provide genuine, physically meaningful information—likely capturing complex seasonal interactions in precipitation and cloud cover that a simpler model misses. The superior performance of the 24-variable model under rigorous spatial cross-validation (which explicitly penalises overfitting to local clusters) provides strong evidence that the selected features represent real soil-environment relationships rather than accidental correlations. While the sample size is admittedly small, the data suggests that a certain level of complexity is required to model the heterogeneous soil depth patterns in Sri Lanka's diverse landscape.

4.4. Engineering Reliability and Practical Considerations

For geotechnical engineering applications, the reliability of soil depth predictions must be evaluated in the context of typical engineering tolerances and decision-making needs. The spatial CV RMSE of 35.4 cm represents the expected magnitude of prediction error when applying the model to unsampled locations. This level of accuracy has different implications depending on the engineering application.

Foundation Design

For shallow foundation design, an error of ±35 cm in estimated soil depth may be acceptable for preliminary site assessment, as foundation design typically requires site-specific geotechnical investigation regardless. However, for areas predicted to have very shallow soils (<50 cm), the relative error becomes more significant—a 35 cm error could represent a substantial fraction of the total depth. In such cases, the uncertainty maps (showing prediction intervals) should be consulted, and additional site investigation should be prioritised. For deep foundation design, where bedrock depth is critical, the absolute error of 35 cm may be less significant relative to typical foundation depths, but the uncertainty should still be considered in design calculations.

Regional Planning and Route Selection

For infrastructure route planning at regional scales, the 1 km resolution and ±35 cm accuracy are generally sufficient for comparing alternative routes and identifying areas of concern. The map can effectively distinguish between areas with consistently shallow vs. deep soils, which is the primary information needed for route selection. However, final route decisions should always be supported by site-specific investigations.

Risk Assessment

For landslide susceptibility and slope stability assessments, the map provides valuable input for regional-scale hazard mapping. The combination of predicted soil depth with uncertainty estimates allows identification of areas where predictions are less reliable, guiding prioritisation of detailed site investigations. The typical error of 35 cm is acceptable for regional hazard assessment, where the goal is to identify areas of concern rather than provide precise depth estimates.

Practical Recommendations for Engineering Use

Areas with high prediction uncertainty (indicated by the uncertainty maps) should be prioritised for additional field investigation before engineering design. Given the RMSE of 35.4 cm, engineering designs should incorporate appropriate safety factors or use conservative estimates (e.g., using the lower bound of the 95% prediction interval) when soil depth is a critical design parameter. The map should be treated as a preliminary assessment tool that guides, but does not replace, site-specific geotechnical investigation. For critical infrastructure projects, field verification of predicted depths is essential. Predictions in data-sparse regions (particularly the northern and eastern dry zones) are likely to have higher uncertainty than predictions in well-sampled areas. Users should consult the uncertainty maps and consider the density of training data in their area of interest. The soil depth map should be used in conjunction with other available information (geological maps, topographic data, local knowledge) to build a comprehensive understanding of site conditions.

4.5. The Scale Gap: Meso-Scale vs. Micro-Scale Variability

A critical finding of this study is the quantification of the "scale gap" in digital soil mapping. Soil depth often exhibits significant micro-scale variability (<50 m) due to local

fracturing, micro-topography, and biological activity. However, the available consistent global covariates are at ~1 km (90 m for DEM) resolution.

This mismatch means the present model is fundamentally limited to predicting the meso-scale deterministic trend—the general deepening of soils in the wet valleys vs. shallowing on dry ridges—rather than the specific depth at any given meter. The modest R^2 should not be interpreted as a model failure, but rather as an estimate of the proportion of soil depth variance that is controlled by regional climate and topography. The remaining variance likely represents this micro-scale stochasticity that cannot be resolved without high-resolution local data (e.g., LiDAR, gamma radiometrics) which is currently unavailable for Sri Lanka. This distinction is vital for users: the map provides the regional baseline, while site-specific variance must be treated as uncertainty (quantified in the prediction intervals).

Agriculture

For agricultural applications, soil depth is critical for determining crop suitability, irrigation planning, and yield estimation. However, the model's performance (R^2 = 0.197, RMSE = 35.4 cm) presents substantial limitations. An error of ±35 cm can be critical for crops with specific rooting depth requirements. For example, a prediction of 50 cm depth with a ±35 cm error could indicate either suitable (85 cm) or unsuitable (15 cm) conditions, leading to incorrect suitability assessments. The map may be useful for identifying broad patterns (e.g., distinguishing consistently shallow vs. deep areas) but should not be used for site-specific crop selection without field verification.

Soil depth affects water-holding capacity and irrigation scheduling. The ±35 cm uncertainty represents a substantial fraction of typical rooting depths (50–150 cm for many crops), meaning that irrigation recommendations based solely on this map could be significantly inaccurate. The map may be useful for regional water resource planning but requires local calibration for operational irrigation management. Crop models that incorporate soil depth are sensitive to this parameter, and a 35 cm error can substantially affect yield predictions. The map's utility for yield estimation is limited to identifying general patterns rather than providing precise inputs for crop models.

Hydrology

Soil depth influences hydrological processes including runoff generation, infiltration, and groundwater recharge. The model's performance has mixed implications.

For large-scale hydrological models (catchment or regional scale), the 1 km resolution and ±35 cm accuracy may be acceptable for identifying general patterns of soil depth distribution, which can inform parameterisation of hydrological models. However, the substantial unexplained variance (~80%) means that local-scale hydrological processes may not be well-represented.

Soil depth affects infiltration capacity and runoff generation, but the ±35 cm uncertainty could lead to significant errors in local-scale runoff predictions. The map may be useful for regional-scale water balance studies but should be used with caution for site-specific hydrological assessments. Deep soils typically have greater groundwater recharge potential, but the model's limited predictive power means that recharge estimates based on this map would have substantial uncertainty. The map may help identify general recharge zones but cannot replace site-specific investigations.

Land Use Planning

For land use planning applications, the map's utility depends on the scale and specificity of planning decisions. At regional scales, the map can help identify broad patterns (e.g., areas with consistently shallow soils that may be unsuitable for certain land uses). The

±35 cm error is less critical when making general zoning decisions, though field verification remains important for specific sites.

For site-specific land use decisions (e.g., selecting locations for infrastructure, determining building density), the ±35 cm uncertainty and ~80% unexplained variance are substantial limitations. The map should be used only as a preliminary screening tool, with field investigation required for final decisions.

The map may be useful for identifying areas of concern (e.g., shallow soils in erosion-prone areas) but cannot provide the precision needed for detailed environmental impact assessments without additional site-specific data.

Recommendations for Future Improvement

The primary limitation of this study remains the low number and uneven spatial distribution of soil profile data. While the robust validation provides confidence in the model's interpolation performance, predictions in data-sparse regions are inherently more uncertain. The modest R^2 values (0.20–0.29) and substantial RMSE (33–35 cm) reflect both the inherent complexity of soil depth and the limitations of the available training data. Future work should prioritise expanding the soil profile dataset to improve model performance and reduce prediction uncertainty.

Increasing the sample size, particularly in under-sampled regions (northern and eastern dry zones), would improve model performance and reduce uncertainty. Incorporating finer-scale environmental data (e.g., detailed geological maps, high-resolution DEMs) may capture local-scale variations that contribute to unexplained variance. Developing separate models for different agro-ecological zones or geological regions may improve performance by accounting for spatial non-stationarity. Combining the map with local knowledge, geological maps, and other available data can improve practical utility despite model limitations.

5. Conclusions

This research applied an ensemble machine learning approach to produce a preliminary 1 km resolution map of soil depth for Sri Lanka. The key findings are: (1) a stacked ensemble of four ML models provided competitive predictions, with performance similar to the best individual models; (2) climatic and topographic variables were the most influential drivers of soil depth, confirming pedologically sound model behaviour; (3) a robust DSM workflow, incorporating careful feature selection, nested cross-validation for stacking, and rigorous spatial cross-validation, can produce preliminary soil maps from sparse legacy data; and (4) quantitative comparison with the global product demonstrated that local calibration improves performance (R^2 = 0.288, RMSE = 33.3 cm) compared to the global model.

The model's performance indicates that only approximately 20–29% of the variance in soil depth is explained, with a typical prediction error representing approximately 25% of the mean depth. This modest performance fundamentally limits the map's utility for precision applications. The map should be positioned as a preliminary assessment and screening tool that guides, but does not replace, site-specific investigations for critical applications. As detailed in Section 4.4, the map is appropriate for regional-scale planning, preliminary route selection, and identifying priority areas for field investigation, but is not appropriate for site-specific engineering design, precise agricultural management, critical infrastructure planning, or regulatory purposes without additional field verification. For geotechnical engineering, the map provides valuable preliminary information for regional-scale planning and route selection, but site-specific geotechnical investigation remains essential for final engineering design. The comprehensive analysis of error sources (Section 4.3) and critical assessment of practical implications (Section 4.4) identify key limitations and provide guid-

ance for the appropriate application of the map. The uncertainty maps generated alongside the predictions enable users to identify areas where additional site investigation should be prioritised. Future work must prioritise expanding the soil profile dataset, particularly in under-sampled regions, to improve model performance and reduce prediction uncertainty before this map can be considered suitable for operational use.

Regarding parent material information, it is noted that GLiM lithology data were included in the initial covariate pool (247 variables) but was not selected by the feature selection algorithm. This is likely because Sri Lanka's low geological diversity (>90% Precambrian metamorphic rocks) results in limited lithological variation across the study area, reducing its predictive value relative to climate and topographic variables. However, it is acknowledged that higher-resolution, nationally harmonised geology maps, if available, could potentially provide finer-scale parent material information that might improve predictions in specific regions. The GLiM class-mapping table and data processing workflow are provided in Supplementary Materials.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/soilsystems10040047/s1 [22], Figure S1: Clark-Evans clustering index for the sample against national data; Figure S2: Feature selection stability; Figure S3: Vari-ance inflation factors for selected variables; Figure S4: Correlation matrix; Figure S5: Scatter Plot of Observed vs Predicted Values; Figure S6: Spatial Distribution of Residuals; Figure S7: Moran's I scatterplot; Table S1: Sample representativeness summary; Table S2: Sample elevation distribution; Table S3: Hyperparameters; Table S4: Number of features against model performance.

Author Contributions: Conceptualization, E.J., E.M.W. and Y.Y.; methodology, E.J.; software, E.J.; validation, E.J.; formal analysis, E.J.; resources, E.M.W. and R.B.M.; data curation, E.J.; writing—original draft preparation, E.J.; writing—review and editing, E.M.W., Y.Y. and R.B.M.; visualization, E.J.; supervision, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article. The final soil depth raster file in GeoTIFF format is available at https://zenodo.org/records/17119499, accessed on 20 November 2025.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LOOCV	Leave-One-Out Cross-Validation
DSM	Digital Soil Mapping
ML	Machine Learning
DTB	depth to bedrock
AWC	available water capacity
SRTM	Shuttle Radar Topography Mission
DEM	digital elevation model
TWI	Topographic Wetness Index
ISRIC	International Soil Reference and Information Centre
GSP	Global Soil Partnership
RF	Random Forest
XGBoost	eXtreme Gradient Boosting
SVM	Support Vector Machine

GLM	Generalised Linear Model
RMSE	Root Mean Square Error
ME	Mean Error
MODIS	Moderate Resolution Imaging Spectroradiometer

References

1. Chesworth, W. *Encyclopedia of Soil Science*; Springer: Dordrecht, The Netherlands, 2008.
2. Schenk, H.J.; Jackson, R.B. The Global Biogeography of Roots. *Ecol. Monogr.* **2002**, *72*, 311–328. [CrossRef]
3. Peterman, W.; Bachelet, D.; Ferschweiler, K.; Sheehan, T. Soil Depth Affects Simulated Carbon and Water in the MC2 Dynamic Global Vegetation Model. *Ecol. Model.* **2014**, *294*, 84–93. [CrossRef]
4. Robertson, P.K.; Cabal, K.L. *Guide to Cone Penetration Testing for Geotechnical Engineering*; Gregg Drilling & Testing Inc.: Signal Hill, CA, USA, 2015.
5. Loke, M.H.; Chambers, J.E.; Rucker, D.F.; Kuras, O.; Wilkinson, P.B. Recent Developments in the Direct-Current Geoelectrical Imaging Method. *J. Appl. Geophys.* **2013**, *95*, 135–156. [CrossRef]
6. Jol, H.M. *Ground Penetrating Radar Theory and Applications*; Elsevier: Amsterdam, The Netherlands, 2009.
7. Mayne, P.W. *Cone Penetration Testing: A Synthesis of Highway Practice*; Transportation Research Board: Washington, DC, USA, 2007; pp. 1–118.
8. Foti, S.; Lai, C.G.; Rix, G.J.; Strobbia, C. *Surface Wave Methods for Near-Surface Site Characterization*; CRC Press: Boca Raton, FL, USA, 2014.
9. Constantin, J.; Picheny, V.; Nassar, L.H.; Bergez, J.-E. A Method to Assess the Impact of Soil Available Water Capacity Uncertainty on Crop Models with a Tipping-Bucket Approach. *Eur. J. Soil Sci.* **2020**, *71*, 369–381. [CrossRef]
10. Fan, Y.; Miguez-Macho, G. A Simple Hydrologic Framework for Simulating Wetlands in Climate and Earth System Models. *Clim. Dyn.* **2011**, *37*, 253–278. [CrossRef]
11. Golshani, A.; Majidian, S.; Hosseini, M. Construction of Underpassing a Crowded Junction with Shallow Soil Depth (Case Study, Tehran). In *Geotechnical Aspects of Underground Construction in Soft Ground*; Korean Geotechnical Society: Seoul, Republic of Korea, 2014; pp. 479–483.
12. Gimsing, A.L.; Szilas, C.; Borggaard, O.K. Sorption of Glyphosate and Phosphate by Variable-Charge Tropical Soils from Tanzania. *Geoderma* **2007**, *138*, 127–132. [CrossRef]
13. Jenny, H. *Factors of Soil Formation: A System of Quantitative Pedology*; McGraw-Hill: New York, NY, USA, 1941.
14. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On Digital Soil Mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
15. 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]
16. Wimalasiri, E.M.; Jahanshiri, E.; Suhairi, T.A.S.T.M.; Udayangani, H.; Mapa, R.B.; Karunaratne, A.S.; Vidhanarachchi, L.P.; Azam-Ali, S.N. Basic Soil Data Requirements for Process-Based Crop Models as a Basis for Crop Diversification. *Sustainability* **2020**, *12*, 7781. [CrossRef]
17. Hengl, T.; De Jesus, J.M.; Heuvelink, G.B.M.; Gonzalez, M.R.; 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]
18. Gochis, D.J.; Vivoni, E.R.; Watts, C.J. The Impact of Soil Depth on Land Surface Energy and Water Fluxes in the North American Monsoon Region. *J. Arid. Environ.* **2010**, *74*, 564–571. [CrossRef]
19. Lacoste, M.; Mulder, V.L.; Richer-de-Forges, A.C.; Martin, M.P.; Arrouays, D. Evaluating Large-Extent Spatial Modeling Approaches: A Case Study for Soil Depth for France. *Geoderma Reg.* **2016**, *7*, 137–152. [CrossRef]
20. Nussbaum, M.; Spiess, K.; Baltensweiler, A.; Grob, U.; Keller, A.; Greiner, L.; Schaepman, M.E.; Papritz, A. Evaluation of Digital Soil Mapping Approaches with Large Sets of Environmental Covariates. *SOIL* **2018**, *4*, 1–22. [CrossRef]
21. Wimalasiri, E.M.; Jahanshiri, E.; Suhairi, T.A.S.T.M.; Mapa, R.B.; Karunaratne, A.S.; Vidhanarachchi, L.P.; Udayangani, H.; Nizar, N.M.M.; Azam-Ali, S.N. The First Version of Nation-Wide Open 3D Soil Database for Sri Lanka. *Data Brief* **2020**, *33*, 106342. [CrossRef]
22. Hartmann, J.; Moosdorf, N. The New Global Lithological Map Database GLiM: A Representation of Rock Properties at the Earth Surface. *Geochem. Geophys. Geosystems* **2012**, *13*, Q12004. [CrossRef]
23. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
24. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

25. Shangguan, W.; Hengl, T.; Mendes de Jesus, J.; Yuan, H.; Dai, Y. Mapping the Global Depth to Bedrock for Land Surface Modeling. *J. Adv. Model. Earth Syst.* **2017**, *9*, 65–88. [CrossRef]
26. Punyawardena, B.V.R. *Agro-Ecological Zonation of Sri Lanka*; Natural Resources Management Centre, Department of Agriculture: Peradeniya, Sri Lanka, 2007.
27. Cooray, P.G. *An Introduction to the Geology of Sri Lanka (Ceylon)*; National Museums of Sri Lanka: Colombo, Sri Lanka, 1984.
28. Chen, S.; Mulder, V.L.; Martin, M.P.; Walter, C.; Lacoste, M.; Richer-de-Forges, A.C.; Saby, N.P.A.; Loiseau, T.; Hu, B.; Arrouays, D. Probability Mapping of Soil Thickness by Random Survival Forest at a National Scale. *Geoderma* **2019**, *344*, 184–194. [CrossRef]
29. Somarathna, P.D.S.N.; Malone, B.P.; Minasny, B. Mapping Soil Organic Carbon Content over New South Wales, Aus-tralia Using Local Regression Kriging. *Geoderma Reg.* **2016**, *7*, 38–48. [CrossRef]
30. Karger, D.N.; Conrad, O.; Böhner, J.; Kawohl, T.; Kreft, H.; Soria-Auza, R.W.; Zimmermann, N.E.; Linder, H.P.; Kessler, M. *CHELSA Climatologies at High Resolution for the Earth's Land Surface Areas*; Version 1.1; World Data Center for Climate (WDCC) at DKRZ: Hamburg, Germany, 2016.
31. Böhner, J.; Koethe, R.; Conrad, O.; Gross, J.; Ringeler, A.; Selige, T. Soil Regionalisation by Means of Terrain Analysis and Process Parameterisation. *Soil Classif.* **2002**, *7*, 213–222.
32. FAO. *Global Soil Organic Carbon Map (GSOCmap)*; FAO: Rome, Italy, 2018.
33. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed.; Christoph Molnar: München, Germany, 2022.
34. Zhu, J.-J.; Yang, M.; Ren, Z.J. Machine Learning in Environmental Research: Common Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57*, 17671–17689. [CrossRef]
35. Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinform.* **2007**, *8*, 25. [CrossRef] [PubMed]
36. Quinlan, J.R. Learning with Continuous Classes. In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, 16–18 November 1992; pp. 343–348.
37. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. Kernlab—An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20. [CrossRef]
38. Taghizadeh-Mehrjardi, R.; Nabiollahi, K.; Kerry, R. Digital Mapping of Soil Organic Carbon at Multiple Depths Using Different Data Mining Techniques in Baneh Region, Iran. *Geoderma* **2016**, *266*, 98–110. [CrossRef]
39. Wadoux, A.M.J.-C.; Heuvelink, G.B.M.; de Bruin, S.; Brus, D.J. Spatial Cross-Validation Is Not the Right Way to Evaluate Map Accuracy. *Ecol. Model.* **2021**, *457*, 109692. [CrossRef]
40. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
41. Kuhn, M. Building Predictive Models in R Using the Caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]