Contents lists available at ScienceDirect

Geoderma Regional

journal homepage: www.elsevier.com/locate/geodrs

Unravelling spatial drivers of topsoil total carbon variability in tropical paddy soils of Sri Lanka

T.M. Paranavithana [a], S.B. Karunaratne [a,b,**], N. Wimalathunge [c], B.P. Malone [b], B. Macdonald [b], T.F.A. Bishop [c], R.R. Ratnayake [a,*]

[a] National Institute of Fundamental Studies, Hantana Road, Kandy 20000, Sri Lanka
[b] CSIRO Agriculture and Food, Butler Laboratory, Black Mountain, Clunies Ross Street, Acton, ACT 2601, Australia
[c] Sydney Institute of Agriculture, School of Life & Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

ARTICLE INFO

Keywords:
Covariate selection
Digital soil mapping
Spatial modelling
Random forest
Area of applicability
Machine learning
Histosols
Ultisols
Luvisols
Fluvisols
Gleysols
Cambisols

ABSTRACT

This study aimed to map and identify the spatial drivers of total carbon (TC) concentration in topsoil (0–15 cm) across paddy-growing regions in tropical climates using Sri Lanka as a case study. For model calibration, a total of 888 sampling locations were sampled using the conditioned Latin Hypercube sampling approach with a sample density of one sample per 11 km^2. Additionally, 99 sampling sites were selected using a design-based (probabilistic) stratified random strategy for independent evaluation of the developed models. Two distinct spatial random forest (RF) models were developed using a variety of environmental covariates: Model 1: using all environmental covariates without variable selection; Model 2: only incorporated covariates selected based on the forward selection process. Evaluation of model quality using fully independent validation sites revealed that both Model 1 and Model 2 performed similarly. Based on the spatial estimates of Model 1 across the paddy-growing regions of Sri Lanka, the predicted TC concentration varied from 0.89% to 13.15%. The highest predicted TC concentration range was in the Wet zone (2.06% to 13.15%), followed by the Intermediate zone (1.18% to 7.23%), and the lowest was reported in the Dry zone (0.86% to 4.30%). In the spatial estimates of Model 2, the predicted values varied between 0.86% and 13.29% and were similar to Model 1. The highest predicted TC concentration range was in the Wet zone (2.09% to 13.29%), followed by the Intermediate zone (1.08% to 6.99%), and the lowest was reported in the Dry zone (0.86% to 4.30%) following the similar pattern to Model 1. In fact, this clearly showed the importance of mean annual rainfall on the dynamics of TC in tropical rice production systems. Furthermore, the variable importance plot of the RF models revealed that out of all considered environmental covariates, the mean annual rainfall was identified as being the most important variable in the developed spatial prediction function. Moreover, we deployed an area of applicability (AOA) calculation to quantify and identify regions where prediction is less reliable and quantified the prediction uncertainty using a bootstrapping approach. Additionally, we assessed the influence of increasing the number of calibration sites on model prediction quality and reliability using user defined sequence of calibration sites. Independent evaluations of each model indicated that model performance quality indices were improved up to $n = 400$ and thereafter stagnated. For AOA results, an improvement in model reliability is observed for Wet and Intermediate zones when models are developed using 400 calibration sites. Derived estimates of TC can be used for regional-scale planning to enhance the soil carbon and provide a baseline for designing a future land-based carbon accounting system for Sri Lanka.

1. Introduction

The Paris Climate Agreement was produced at the 21st Conference of the United Nations Framework Convention on Climate Change (UNFCCC) as an attempt to avert the impacts of climate change. It is anticipated that soil carbon will play a vital role in keeping rise in global temperature to below 2 °C (preferably to 1.5 °C) (Minasny et al., 2017). Soil contains the largest terrestrial carbon pool (Scharlemann et al.,

* Corresponding author at: National Institute of Fundamental Studies, Hantana Road, Kandy 2000, Sri Lanka
** Corresponding author at: CSIRO Agriculture and Food, Butler Laboratory, Black Mountain, Clunies Ross Street, Acton, ACT 2601, Australia
E-mail addresses: Senani.Karunaratne@csiro.au (S.B. Karunaratne), renuka.ra@nifs.ac.lk (R.R. Ratnayake).

https://doi.org/10.1016/j.geodrs.2023.e00745
Received 17 July 2022; Received in revised form 14 December 2023; Accepted 14 December 2023
Available online 19 December 2023

2014), and there are two forms of soil carbon that are prevalent: soil organic carbon (SOC) and soil inorganic carbon (SIC) (Sreenivas et al., 2016). Whether organic or inorganic, the global soil carbon pool is crucial in maintaining soil ecosystem function and productivity (Raza et al., 2020; Qadir et al., 2006).

Submerged paddy fields are recognised as an important agro-ecosystem for global carbon cycling (Meetei et al., 2020). Rice is the primary food source for more than half of the global population (Rajkishore et al., 2015). With the rising demand for rice globally (Haque et al., 2020), attention should be paid to increasing productivity upon the limited land resources where it is grown. As pressure on the limited cultivable lands increases, maintaining and improving soil quality is vital to sustaining agricultural productivity and environmental quality in those areas. As a primary natural resource in paddy-growing ecosystems, the soil should have sufficient physical, chemical, and biological qualities to increase rice production, along with other management practices (Komatsuzaki and Ohta, 2007). Rahman and Parkinson (2007) reported that a combination of bio-physical-chemical factors are important in increasing soil fertility, that would lead to an increase in rice production. Soil organic carbon, which relates to soil physical, chemical, and biological fertility, and available soil N, P, and K, all of which limit rice yields, were included in their analysis. Furthermore, Girsang et al. (2019) demonstrated that the soil bulk density, saturated hydraulic conductivity, soil water-filled space and N mineralisation significantly affect the grain yield of rice. Soil management also determines the productivity of the land, with common practices including conservation tillage (Ghimire et al., 2017; Wissing et al., 2013), manure application, retaining crop residue (Gattinger et al., 2012; Zhang et al., 2022) and crop rotation (Paranavithana et al., 2020; Ratnayake et al., 2017) all improving soil carbon status by improving carbon inflows and reduction of losses. Ratnayake et al. (2014) in the Northern region of Sri Lanka showed that organic fertilisation that was maintained for 10 years and minimum tillage practices significantly increased SOC and carbon stocks in different annual cropping systems.

In submerged paddy-growing soil systems, SOC accumulation rates are significantly high owing to some inherent mechanisms such as subjecting soils to periodic anaerobic conditions (Xu et al., 2020), production of microbial activity inhibitors, incomplete decomposition and decreased humification of the organic matter (Ratnayake et al., 2017; Sahrawat, 2004). Higher silt and clay concentrations in lowland paddy soils also stabilise SOC because those particles act as chemical (Yan et al., 2013) and physical (Huang et al., 2010) protectors against carbon mineralisation. For example, Song et al. (2020), in their study in Jiangxi Province, subtropical China, found that soil organic matter concentration in paddy soils was higher than the amount recorded in upland and forest soils in the same region.

The current study focuses on the quantification of the spatial drivers and landscape-scale modelling of total carbon (TC) for tropical rice production systems. In general, the spatial variability of TC has often been reported as quite high across landscapes due to a combination of edaphic environmental and climatic factors together with land management practices (De Blecourt et al., 2017). Usually, under natural environmental conditions, soil characteristics are strongly influenced by the inter-relationships between soil parent material, climatic conditions and landform characteristics and features (Liu and Liu, 2014). Along with these, other environmental features, such as vegetation and related indices such as type, density, diversity, and patterning (both spatially and temporally), have been adopted to develop soil carbon spatial models at different scales (Shi-Hang et al., 2011).

Machine Learning (ML) techniques have been used in digital soil mapping (DSM) by enabling the inference of relationships between soil properties and environmental covariates (Khaledian and Miller, 2020; Wadoux et al., 2020). Several ML techniques have emerged that could potentially facilitate greater predictive power despite the complexity of the variation in soil carbon. The ML approaches utilised in soil carbon modelling encompass a diverse array of techniques and applications. These include the use of support vector machines (Song et al., 2022; Peng et al., 2014), artificial neural networks (Tiwari et al., 2015), regression trees (Rentschler et al., 2019), random forests (RF) (Wang et al., 2023; Zhang et al., 2017; Hengl et al., 2015), extreme gradient boosting (Taghizadeh-Mehrjardi et al., 2020; Forkuor et al., 2017), and neural networks (Aitkenhead and Coull, 2016) for advancing prediction models of soil carbon. Out of all those different algorithms RF algorithm is the most widely used ML algorithm for soil carbon modelling work (Lamichhane et al., 2019). As a result, linear regression models can easily be replaced with ML algorithms to account for more complex soil-environment relationships (Hengl et al., 2015).

In the development of a spatial prediction function for soil carbon, key elements in model development include not only a set of environmental covariates that are used as model drivers but also the distribution of model calibration sites across the landscape and the number of sites required to develop an optimum model with higher model quality. Due to the associated cost of field data collection and the need to capture the inherent variation of environmental covariates through sampling sites, algorithms such as conditioned Latin hypercube sampling (cLHS) (Minasny and McBratney, 2006) are commonly deployed. Somarathna et al. (2017) stated that the uncertainty of model predictions decreases with increasing calibration sample size. Furthermore, prediction uncertainty in soil carbon modelling can be significantly influenced by various factors, including the spatial heterogeneity of soil carbon, the choice of modelling algorithm (Somarathna et al., 2017), and environmental and landscape characteristics (Sun et al., 2022). Additionally, Mishra et al. (2022) and Saurette et al. (2022) emphasized the importance of the selection and inclusion of environmental covariates, which also could control the uncertainty of soil carbon prediction. Therefore, it is imperative for soil carbon modelling studies to carefully consider and address these factors to improve the robustness and reliability of predictions.

The current study aims to understand the drivers of total carbon concentration in tropical paddy-growing soils. Annually Sri Lanka cultivates approximately 708,000 ha of paddy soils across the country (two primary seasons), accounting for 34% of the country's total agricultural land extent. Currently, there is no consistent baseline dataset on TC concentrations across the major paddy cultivation regions in Sri Lanka. One exception is the regional-scale study conducted by Ratnayake et al. (2016), one of the first spatially explicit studies carried out to estimate SOC concentration in the Northern paddy-growing region of Sri Lanka. Furthermore, a national-scale study conducted by Vitharana et al. (2019) focused on the spatial distribution of SOC stocks throughout the country with a limited number of ground truth data locations scattered across a large land extent ($n = 122$, area $= 64,610$ km^2). Therefore, the current study aims to:

1. Undertake a detailed field sampling campaign to collate ground truth datasets covering paddy-growing soils in Sri Lanka
2. Develop spatially explicit machine learning model/s to identify drivers of TC in tropical rice production systems and assess the quality of the model using a fully independent dataset
3. Evaluate the reliability of the generated models across the landscape using Area of Applicability (AOA), as outlined by Meyer and Pebesma (2021). The AOA provides guidance on the applicability of model extension across entire mapping extent.
4. Assess the relative impact of sample site number on model evaluation perfromance.

2. Materials and methods

2.1. Description of the study area

Sri Lanka is located between 5.9° and 9.87° North and 79.65° and 81.88° East. There are three major climatic zones in the country, which are essentially defined on the basis of annual rainfall; Dry zone (<1750 mm); Intermediate zone (1750–2500 mm); and Wet zone (>2500 mm) (Mapa, 2020). The Wet zone experiences relatively high mean annual rainfall without any pronounced dry periods, whereas the Dry zone experiences relatively lower mean annual rainfall, with a distinct dry season from May to September. Compared to the Dry zone, there is a short and less prominent dry season in the Intermediate zone (Punyawardena, 2020). A large proportion of paddy-growing lands are located in the country's Dry zone, which contains two-thirds of the country's entire paddy-growing area. Compared to other paddy-growing countries, Sri Lanka cultivates paddy under various hydrogeological regimes, climatic conditions, terrain conditions (e.g., under significant variations in altitude/elevation and slope) and soil types that differ throughout the country. The maximum annual rainfall in the Wet zone of the country has been recorded as 6000 mm, while values as low as 600 mm have been reported for the dry and arid regions. The altitude of the country ranges from mean sea level (MSL) to 2575 m above MSL, and the average temperature values vary within a range of 15-30 °C across the elevation gradient. Paddy is cultivated across all agro-ecological regions except the high massif areas above 1200 m (Dhanapala, 2007). Two different cultivation seasons prevail within the country, depending on the monsoon's rainfall patterns. The two main seasons are known as *'Maha kannaya'* (falling during the second inter-monsoon and northeast monsoon season from September to February) and *'Yala kannaya'* (falling during the first inter-monsoon and southwest monsoon season between March and August (Sathischandra et al., 2014). This study covers current paddy-growing areas in all 25 administrative districts of Sri Lanka.

2.2. Designing soil sampling schemes for model calibration and validation

Two distinct sampling strategies were used to collate soil samples for model calibration and validation. The cLHS strategy was used to determine the model calibration sites. The cLHS algorithm selects sample sites from a Latin hypercube in the feature space (Minasny and McBratney, 2006). For example, for k continuous variables, each X component is divided into n (sample sites) equally probable strata based on their distributions, and x is a sub-sample of X. The cLHS algorithm is based on heuristic rules with an annealing schedule (Minasny and McBratney, 2006). The cLHS sampling design is an effective sampling technique for identifying sampling locations that represent the variation of different environmental covariates. In this study, a variety of environmental covariates were used to capture the inherent variability of the landscape that affects the carbon inflows and out-flows (Table 1). Hence, the considered environmental covariates directly or indirectly affect the TC concentrations in the study region. Additionally, a fully independent validation dataset using design-based sampling principles was collated to assess the model prediction quality (Brus et al., 2011). As a design-based sampling scheme, the stratified random sampling (SRS) approach was adopted. For stratification, the same environmental covariates listed in Table 1 were clustered (using the k-means clustering algorithm). In each stratum, simple random sampling was performed, with each stratum being considered as a sampling zone.

In total, 1000 sampling sites were selected as model calibration and validation sites. Among them, 800 sampling locations were allocated across the landscape using the cLHS algorithm. In addition to those 800 calibration sampling sites, another 100 soil samples were taken at an approximate distance of 100–150 m away from the main sampling sites, similar to the approach described by Karunaratne et al. (2014). The additional calibration sampling sites were used to capture the inherent short-range soil variability. For the independent validation of the model, 100 sampling sites were randomly assigned across strata generated using the SRS strategy. In the SRS approach, a set of environmental covariates (Table 1) is stratified into 25 strata, and four samples are allocated for each stratum. Despite the sampling locations being predetermined, reaching the exact sampling location was quite challenging during the sampling stage due to practical issues such as site accessibility. Therefore, 888 calibration samples out of 900 sites and 99 validation samples out of 100 sites were sampled. Fig. 1 depicts the spatial distribution of sampling locations and paddy-growing areas within the country across the major climatic zones, namely Wet, Intermediate and Dry zones. The soil samples were collected at a soil depth of 0–15 cm soil depth level using a soil augur with a diameter of 5 cm. At each sampling site, soil samples were collected from three points in a triangular path with a distance of approximately 10 m between sampling points and composited to form a representative sample. The GPS locations of all the sampling sites were recorded using a Garmin eTrex 30 handheld GPS receiver.

Table 1
Summary of the environmental covariates used for the study.

Scorpan factor	Environmental variable	Units	Reference/ data source
Climate (*c*)	Mean annual rainfall	mm	Wordclim http://www.worldclim.org/
	Temperature (annual average mean, annual average maximum, annual average minimum)	°C	Wordclim http://www.worldclim.org/
	Vapour Pressure Deficient (VPD)	K pa	Wordclim http://www.worldclim.org/
Relief (*r*)	Elevation	m	NASA SRTM data http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database v4-1
	SAGA Wetness Index (WI)	Unit less	Derived from NASA STRM (secondary terrain attribute)
	Slope	Degrees	Derived from NASA STRM (primary terrain attribute)
Organism (*o*)	MODIS Enhanced Vegetation Index (EVI)	Unit less	NASA https://modis.gsfc.nasa.gov/data/dataprod/mod13.php. Derived from taking mean annual EVI data from 2005 to 2014.

2.3. Soil sample analysis

All visible organic debris, plant roots, and stones were removed by handpicking prior to the analysis of the composited soil samples. The moist soil samples from the field were analysed for soil pH (1:2.5 soil: water suspension) (Anderson and Ingram, 1993). The remaining soil samples were air-dried and sieved using a sieve with a 2 mm mesh. Soil samples were then ground to size of <0.15 mm to obtain a uniform particle size. Before determination of the soil carbon concentration, another portion of powder with a size <0.15 mm was again ground and sieved through a 42-μm mesh sieve. Then, soil carbon concentration (%) was analysed using an automated dry combustion method via a 2400 Series II CHN Elemental Analyser (Fadeeva et al., 2008; Skeen, 1994). The measured TC concentrations are reported as oven-dry equivalent (ODE) using the following equations (Eqs. (1) and (2)).

ODE correction factor (θ_m) is given by:

$$\theta_m = \frac{\mathrm{Mw}}{\mathrm{Ms}} \tag{1}$$

where Mw = mass of water in the air-dried sample and Ms. = the total mass of the oven-dried soil.

$$\mathrm{TC}_{\mathrm{OD}} = \mathrm{TC}_{\mathrm{AD}} \times (1+\theta_m)\left[\frac{g\ OC}{kg\ ODsoil}\right] \tag{2}$$

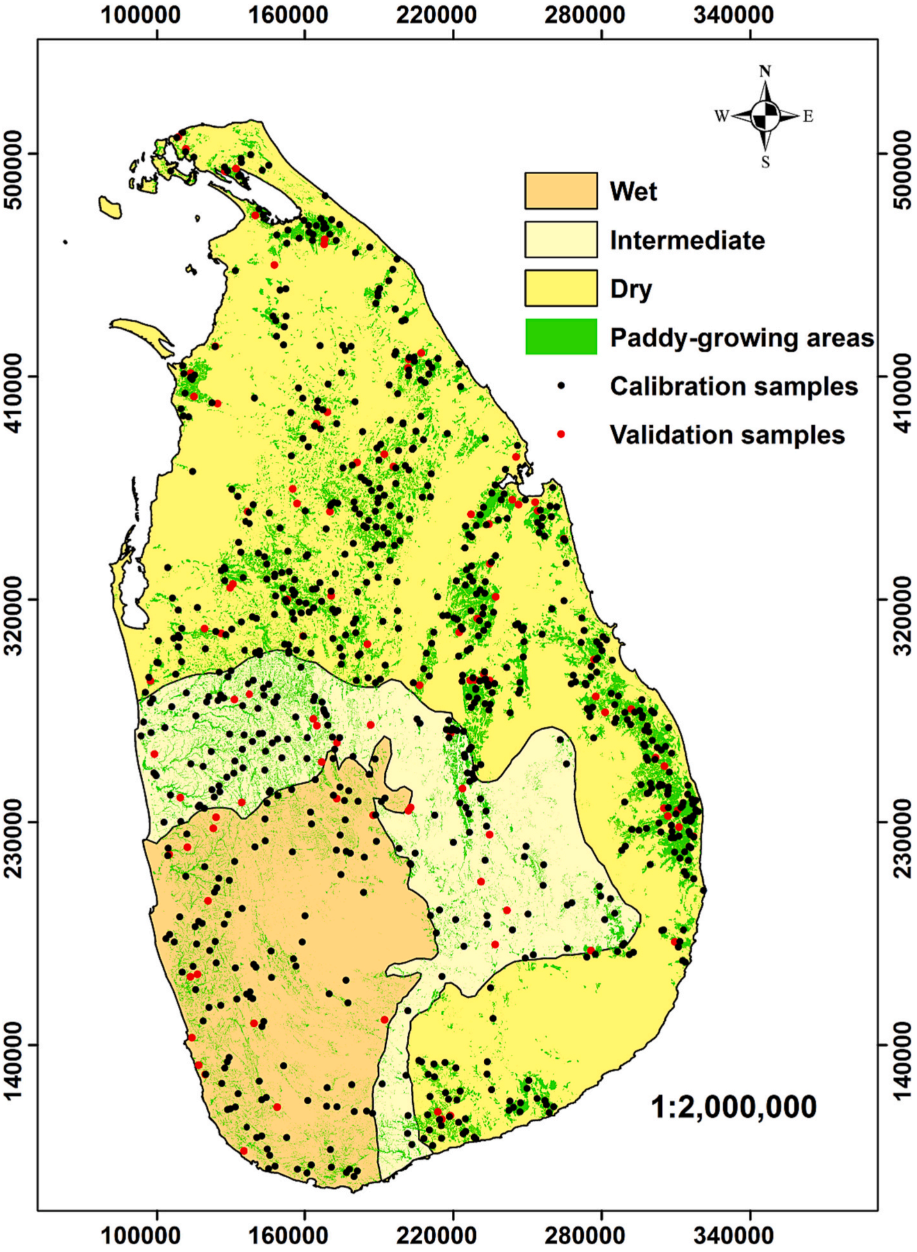

Fig. 1. The paddy-growing areas of Sri Lanka (green shaded areas) and sampling sites with overlapping major climatic zones: calibration sample sites are indicated in black colour, while validation sample sites are indicated in red colour (coordinate system: *Kandawala* Sri Lanka Grid).

where TC_{OD} total carbon concentration in g C/kg oven-dried (OD) soil; and TC_{AD} total carbon concentration in g C/kg air-dried (AD) soil.

2.4. *Preparation of environmental covariates for spatial modelling of total carbon*

The development of the spatial model was performed based on the *scopan* digital soil mapping framework (Eq. (3)), as outlined by McBratney et al. (2003). The *scopan* model describes the quantitative relationships prevailing among TC and environmental covariates by developing a spatial soil prediction function. A variety of environmental covariates are considered in the current study including slope, the SAGA wetness index (WI)) (Bohner and Selige, 2006), and other terrain attributes such as hydrologically corrected elevation data derived from NASA Shuttle Radar topographic mission, MODIS Enhanced Vegetation Index (EVI), Vapour Pressure Deficit (VPD), annual average mean temperature, annual average maximum temperature, annual average minimum temperature and mean annual rainfall. All environmental covariates were standardised (resampled) to a spatial resolution of 100 m prior to spatial analysis. A summary of the environmental covariates used in this study is presented in Table 1, the details of which were presented in Rajapaksha et al. (2020).

$$S f = (s, c, o, r, p, a, n) + e \tag{3}$$

where, *S* represents TC concentration, soil (*s*), climate(*c*), organisms (*o*), relief (*r*), parent materials (*p*), age (*a*), and spatial position (*n*); and where *e* is the error. A random forest modelling framework represents the *f* in the current study.

2.5. *Geospatial modelling*

The RF model can be used either as a classifier or for regression. For the current modelling work, a regression RF model was adapted in which the importance of each predictor variable was determined by a regression loss function on the basis of mean square error (MSE) (Dewi and Chen, 2019). The RF algorithm is capable of handling both linear and nonlinear complex relationships and multicollinearity among the considered parameters (Karunaratne et al., 2020). At each binary split, a random subset of covariates is selected to provide the best split. The number of variables available for splitting at each tree node is referred to as the m_{try} parameter. Heung et al. (2014) reported that the m_{try} parameter in the RF model is the main tuning parameter that requiring optimisation. In the current study, the RF model's m_{try} parameter was optimised using repeated 10-fold cross-validation. The best model parameter for the m_{try} was determined using the return value with the lowest RMSE value obtained via 10-fold cross-validation. The cross-validation was based on the calibration dataset.

Two different forms of RF model were tested in the current study. The only difference between Model 1 and Model 2 is that the latter used forward selection of the variables, as described by Meyer et al. (2019). In summary, Model 2 is trained with all possible pairs of predictor variables and keeps the best pair as the initial model. Then, each of the remaining predictor variables is iteratively added and tested for improvement with the best model. The process stops if none of the remaining variables increases the model performance when added to the current best model (Meyer et al., 2018). The purpose of utilising forward selection for Model 2 is to overcome model overfitting issues by removing highly correlated variables (Meyer et al., 2019). A summary of the two RF models tested in the current study is provided in Table 2.

2.5.1. Evaluation of model quality

The model performances were evaluated using the Nash-Sutcliffe model efficiency coefficient (NSE) (Eq. (4)), Root- Mean Square Error (RMSE) (Eq. (5)) and Lin's Concordance Correlation Coefficient (LCCC) (Eq. (6)). The NSE measures the improvement made by the model based on the magnitude of the residual variance compared to the measured data variance. The RMSE provides an indicator for the accuracy of the model, while the LCCC indicates how well the measured and predicted values deviate from a 1:1 line (i.e., a 45-degree line). The best models are those with the lowest RMSE values and comparatively higher LCCC and NSE values. Models with LCCC and NSE values close to 1 are considered to be those with the best performance. The models were validated using independent datasets collated using a design-based sampling strategy, as explained in Section 2.2. A schematic diagram of the TC measurement and data modelling pipeline is presented in Fig. 2.

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(p-o)^2}{\sum_{i=1}^{n}(o-\bar{o})^2} \tag{4}$$

where p is the difference between predicted(p) and observed(o) values, o-$\bar{o}$ is the difference between the observed (o) value and the mean of the observed ($\bar{o}$) values and n refers to the number of observations.

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}(p-o)^2} \tag{5}$$

where p, o refer to predicted and observed values, and n refers to the number of observations.

$$\text{LCCC }(\rho_C) = \frac{2\rho\sigma_X\sigma_Y}{\sigma^2{}_X + \sigma^2{}_Y(\mu_X-\mu_Y)^2} \tag{6}$$

Table 2
Summary of the specific techniques employed in model development.

Model name	RF model optimisation	Cross-validation	Variable selection
Model 1	m_{try} parameter	10-fold CV	NA
Model 2	m_{try} parameter	10-fold CV	Forward selection

where ρ_C is the estimated LCCC, μ_X and μ_y are the means for the measured and predicted parameters, and $\sigma^2{}_X$ and $\sigma^2{}_y$ are the corresponding variances of the measured and predicted parameters. ρ is the Pearson's correlation coefficient between the measured and predicted values.

2.5.2. Spatial estimates of total carbon

The area of applicability (AOA) of the two models was calculated, as explained by Meyer and Pebesma (2021). The AOA provides a quantitative assessment of the reliability of current prediction quality using the existing measurement datasets, and the function can be found in the CAST R package (Meyer et al., 2020). The AOA approach identifies the areas in which the model is likely to be problematic as a result of the dataset used in the modelling not capturing the environmental and spatial features of the area in which the model is being applied. The AOA is determined on the basis of the dissimilarity index (DI), which is a unitless measurement for detecting the deviation of new data cases (a prediction location) from the training data. The DI is calculated by considering the cross-validation folds and using a threshold, which is by default is the 95% quantile of the DI of all training data, and then returns the AOA statistics. The patterns in the DI are in general agreement with the true prediction error, i.e., very high DI values indicate areas that are not covered by the training data. For prediction areas in which the DI values are over the threshold, the predictions are assumed to be unreliable. They should be excluded from further analysis, as the values of the predictors at the locations of the training data do not represent the values of the predictors where the prediction is being made (Meyer and Pebesma, 2021).

Furthermore, if distances were calculated based on the standardised covariates, all variables would be treated as being equally important. However, distances are not equally relevant within the predictor space; some variables are more important than others (as indicated by the variable importance in machine learning algorithms). Therefore, scaled variables are multiplied by the weighting estimate derived from the variable importance of the RF model for each variable before distance calculation. The training data set is created for our 888 sampling locations on the basis of the environmental covariate dataset (Appendix, Fig. A1). In addition to AOA analysis as a measure of the reliability of current prediction quality, 100 bootstrapped models, resulting from 100 possible mapped outputs, were used to generate the lower (5%) and upper (95%) predictions (Gray et al., 2019). The thus-obtained prediction intervals were used to calculate the 90% prediction interval (Appendix, Fig. A2).

2.6. Impact of number of calibration sites used for model training: model performance and reliability

Considering the advantage of having a large number of calibration sites ($n = 888$), we evaluated the impact of model prediction quality and reliability with increased calibration sites in a sequence. To assess the effect of the number of calibration sites, sequentially increasing numbers of data cases were used for model calibration, starting at $n = 200$ and then increasing by 100 each time up to 800. Samples for each configuration were selected using the cLHS algorithm and selecting from the 888 sampling locations available for model development. For each of these chosen sequences, Section 2.5 was repeated, and each of these models was independently evaluated. The individual model quality was performed using the validation sites ($n = 99$), as noted above, enabling an unbiased comparison of the model prediction quality. Finally, the percentage of AOA was calculated for each model to identify the model reliability.

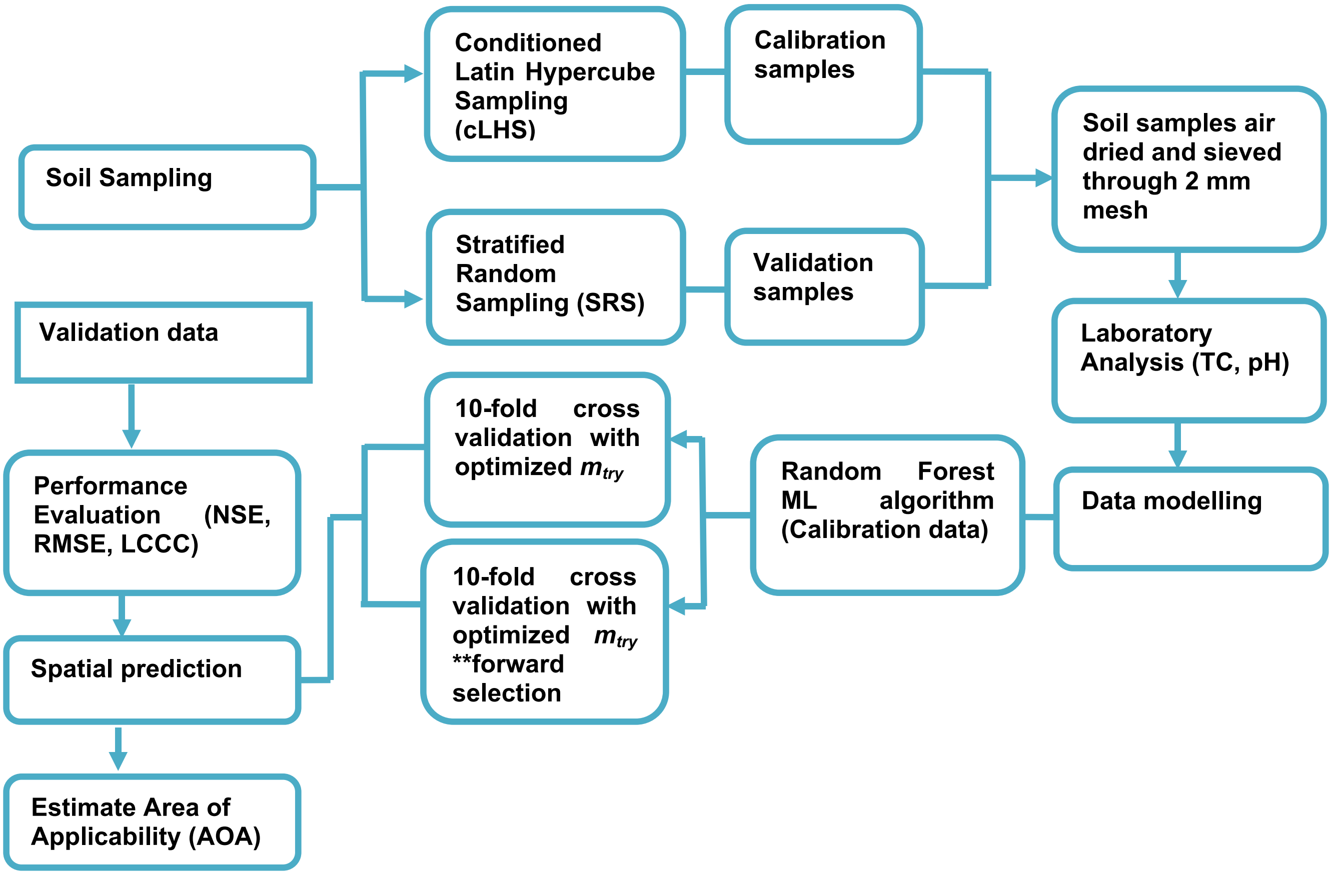
Soil Sampling
Conditioned Latin Hypercube Sampling (cLHS)
Stratified Random Sampling (SRS)
Calibration samples
Validation samples
Soil samples air dried and sieved through 2 mm mesh
Laboratory Analysis (TC, pH)
Data modelling
Random Forest ML algorithm (Calibration data)
10-fold cross validation with optimized m_{try}
10-fold cross validation with optimized m_{try} **forward selection
Validation data
Performance Evaluation (NSE, RMSE, LCCC)
Spatial prediction
Estimate Area of Applicability (AOA)

Fig. 2. A schematic diagram for the measurement and modelling of the soil total carbon adopted in the current study.

3. Results and discussion

3.1. Descriptive analysis of the total carbon concentrations in paddy-growing soils

The descriptive statistics for TC% in paddy-growing soils of Sri Lanka are presented in Table 3. The summary of the statistics reveals that the mean TC% of the paddy soil was 2.44 ± 1.73%. The reported skewness value was 3.53, which is indicative of the positively skewed, unimodal distribution of the measured TC concentrations. This implies that high concentrations of TC are stored in a few locations, whereas only a relatively small amount of carbon is stored in most of the other locations on the landscape (Delgado-Baquerizo et al., 2018). Further analysis considering the different climatic zones reported mean TC% values of 5.21 ± 2.78, 2.24 ± 0.75, and 1.89 ± 0.79 for the Wet, Intermediate, and Dry zones, respectively. In contrast to the other two zones in the country, TC% values for the Wet zone are significantly higher, and the associated soil pH values are significantly lower, as shown in the box plots (Fig. 3). At higher soil pH, the bonds between organic constituents and clay particles in the soil can be easily broken (Neina, 2019), leading to an increase in soil carbon mineralisation, whereas the decarboxylation of organic acid anions during the organic matter decomposition could lead to an increase in soil pH, as explained by Ding et al. (2019).

3.2. Relationships between total soil carbon concentration and environmental covariates

The analysis of the Pearson's correlation coefficient is summarised in Fig. 4. The Pearson's correlation coefficient was calculated in agreement with the linear relationships between TC concentrations and

Table 3
Descriptive statistics for total carbon in paddy soils across the country and in major climatic zones of Sri Lanka.

Variable	n	Mean	SD	Median	Min	Max	Skewness	SE	Q1	Q3
Whole country TC%	987	2.44	1.73	2.04	0.30	17.85	3.53	0.06	1.47	2.69
Wet Zone TC%	145	5.21	2.78	4.63	1.36	17.85	1.93	0.23	3.19	6.45
Intermediate Zone TC%	176	2.24	0.75	2.21	0.78	5.04	0.81	0.06	1.76	2.65
Dry Zone TC%	666	1.89	0.79	1.77	0.30	5.33	1.15	0.03	1.34	2.31

Note: n: number of samples; SD: Standard Deviation; Min: minimum; Max: maximum; SE: Standard Error; Q1: first quartile; Q3: third quartile.

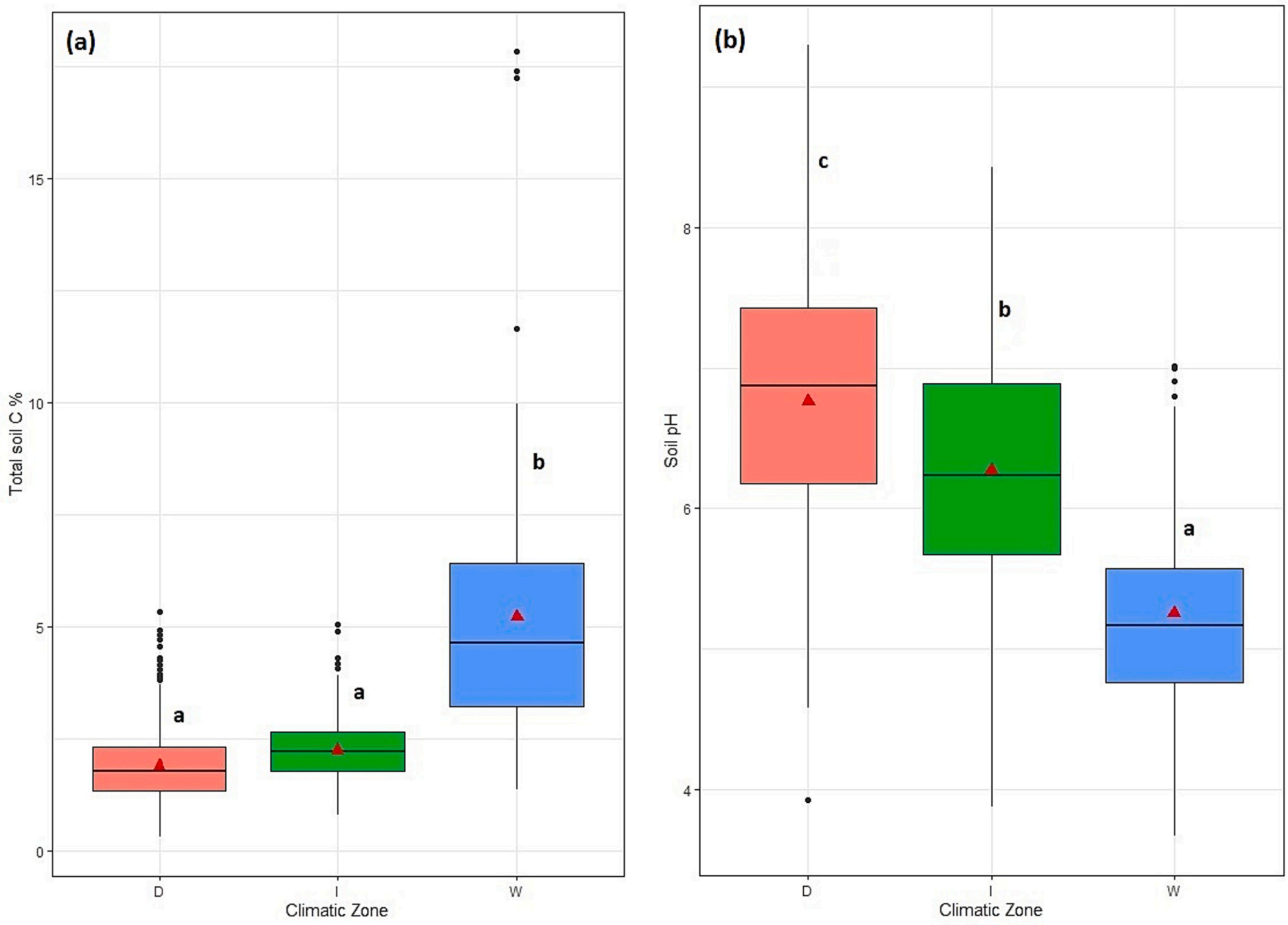

Fig. 3. Distribution of the measured total carbon concentration (Fig. 3a) and pH (Fig. 3b) values with respect to major climatic zones, abbreviated as D: Dry zone; I: Intermediate zone; and W: Wet zone.
Note: The solid horizontal line in the boxplots indicates estimated median TC concentrations. The ends of the boxes indicate the inter-quartile range, while the whiskers represent the maximum and minimum values, excluding any outliers, and outliers are depicted as 'dots', the red-coloured triangles indicate the category means.

environmental covariates. A strong positive correlation was observed between TC concentration and mean annual rainfall (r = 0.64). Furthermore, positive correlations were observed between TC concentration and MODIS EVI (r = 0.34) and TC concentration and the slope of the landscape position (r = 0.09). Negative correlations were observed between TC concentration and annual average maximum temperature (r = −0.36), annual average mean temperature (r = −0.22) and annual average minimum temperature (r = −0.07).

Climate variables are among the key drivers of TC concentration in paddy-growing soils. In general, higher rainfall and lower temperature provide the conditions necessary for increasing soil carbon levels (Fantappie et al., 2011). However, water availability during the rainy season affects both carbon accumulation through primary production and carbon loss through decomposition, which ultimately balances the long-term storage of soil carbon, which also depends on the rate of carbon inflows into the system. Furthermore, MODIS EVI data showed a positive correlation with the soil TC concentrations. In fact, MODIS EVI acts as a proxy for land productivity (biomass production), and the quantitative connection between the amounts of carbon added to soils.

3.3. *Identification of the drivers of soil total carbon concentration across paddy-growing soils*

The environmental covariates that explain the TC concentration across the landscape are divided into three main categories: climatic (rainfall, temperature, VDP), relief (elevation, WI, slope degree), and organism (MODIS EVI). The variable importance plot (VIP) obtained using the RF model was used to identify the key model drivers (Fig. 5). A summary of the fitted RF model with all covariates (Model 1) is presented in Fig. 5a, while a summary of the forward-selected variables employed in Model 2 is presented in Fig. 5b. Rainfall was the key environmental driver affecting the spatial distribution of TC concentrations, as observed by the VIP plots for both models (Fig. 5). The slope angle was the least important variable in both Model 1 and Model 2 when predicting TC concentrations. This may be due to the lower prevalence of land variability in the flat terrain areas used in rice production. Emphasising the smaller degree of variability in the relatively flat landscapes of paddy-growing paddocks, >90% of the paddy-growing areas were scattered within the narrow range of 0^0 to 2.5^0. However, in Model 2, which used forward selection of the variables, as explained by Meyer et al. (2019), MODIS EVI, elevation and WI were not selected in the final model.

Among the most important model drivers used for spatial prediction, rainfall, temperature, and evapotranspiration are considered as the primary climatic covariates involved in SOC storage fluctuations (Delgado-Baquerizo et al., 2018). Both the rainfall and temperature undeniably regulate the soil TC dynamics of ecosystems. The VPD is closely related to the evapotranspiration rate in the area of interest (Zheng

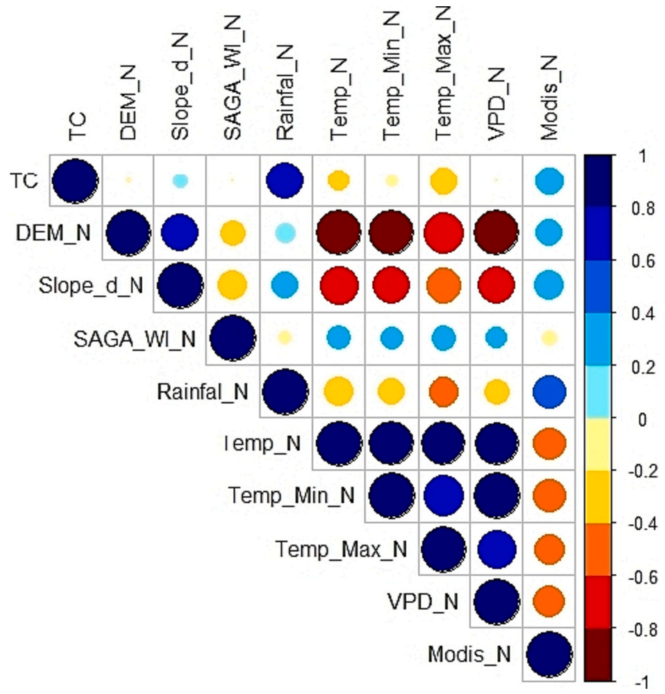

Fig. 4. Pearson's correlation coefficient matrix of soil total carbon concentration and climatic, terrain, and imaging attributes of paddy-soils in Sri Lanka. *Abbreviations*: TC: total carbon concentration; Rainfal_N: mean annual rainfall; Temp_Min_N: annual average minimum temperature; DEM_N: elevation; VDP_N: vapour pressure deficient; Temp_N: annual average mean temperature; Temp_Max_N: annual average maximum temperature; Modis_N: MODIS EVI; Slope_d_N: slope; SAGA_WI_N: SAGA wetness index.

et al., 2014). Accordingly, the increased temperature levels lead to enhanced evapotranspiration rates resulting in a nonlinear rise in VPD. Furthermore, the higher VPD increases soil evapotranspiration, affecting plant growth and soil productivity (Breshears et al., 2013). Elevation plays a crucial role among topographic variables in determining soil carbon distribution by altering the micro- and macro-environmental conditions (Martin et al., 2014; Tsui et al., 2013). The MODIS EVI (MODIS–Terra sensor) is an important time series vegetation index capable of monitoring substantial changes in the ecosystem, providing new insight into the mechanisms of the carbon cycle (inflows of carbon into the soil system). The MODIS EVI data are directly related to plant productivity and act as a proxy for the carbon inflows into the soils.

Several previous studies conducted in tropical climatic regions have reported similar results corroborating with the current research. For example, Hinge et al. (2018) predicted SOC stocks using an RF model with climatic and remotely sensed datasets in India covering different land use types, including croplands and forest areas. They found that, although the topographic parameters, slope, and multi-resolution index of valley bottom flatness were relevant to surface SOC distribution, the most important factors were elevation and land use. Furthermore, Hinge et al. (2018) reported that the decrease in temperature with rising elevation, as well as changes in rainfall distribution, might affect the decomposition rate of soil organic matter. Therefore, the combined contribution of elevation, rainfall and temperature towards the regulation of plant productivity and organic matter decomposition is emphasized. Dharumarajan et al. (2017), in their study, performed spatial prediction of SOC in the semi-arid tropics of India, incorporating five major land use types (single crop, double-crop, fallow land, scrub and forest), and showed that EVI and normalised different vegetation index (NDVI) were the most critical determiners of SOC distribution. In addition to productivity, the contribution of vegetation in controlling high-temperature levels through evapotranspiration may be the underlying reason for the preservation of high levels of carbon in soil.

3.4. Independent model evaluation

A summary of the fully independent model validation is presented in Table 4. The NSE values of models 1 and 2 were 0.29 and 0.27, respectively. The RMSE (%) and LCCC values of the two distinct predictive models reported identical values of 1.35 (%) (Table 4). The high LCCC value of 0.75 for the two fitted models indicated considerable agreement between measured and predicted TC concentrations. In summary, it can be concluded that the NSE, RMSE, and LCCC values related to the performance of the two models were more-or-less similar.

Table 4
Performance of predicted soil carbon models according to fully independent validation.

Model	NSE	RMSE (%)	LCCC
Model 1	0.29	1.35	0.75
Model 2	0.27	1.35	0.75

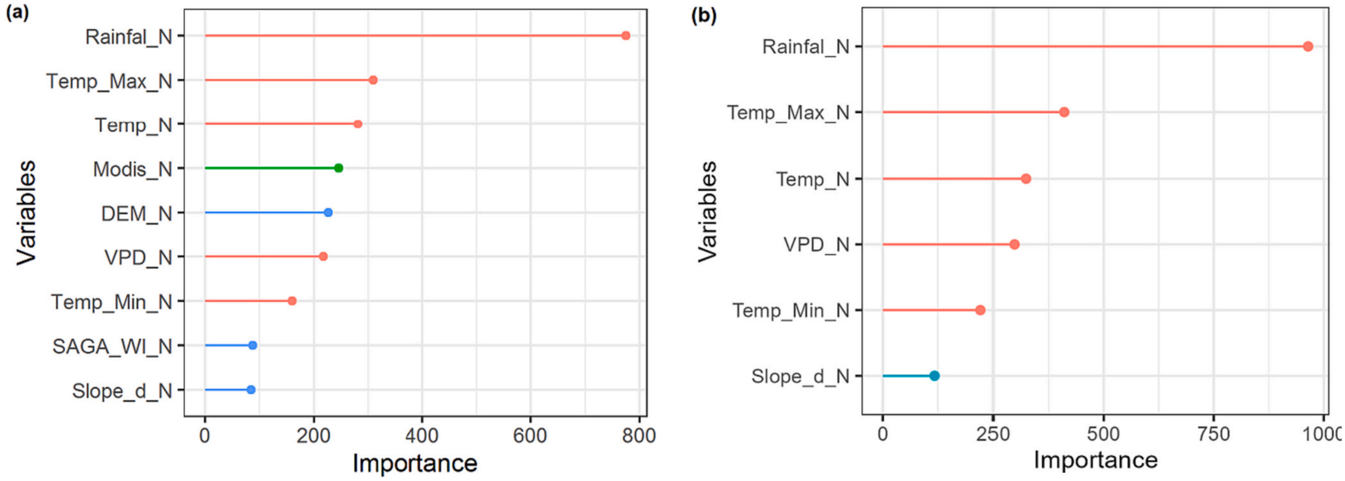

Fig. 5. Relative importance of variables of each soil carbon model on the basis of the random forest algorithm: (a) Variable Importance Plot of Model 1, (b) Variable Importance Plot of Model 2 *Abbreviations*: Rainfall_N: mean annual rainfall; Temp_Min_N: annual average minimum temperature; DEM_N: elevation; VDP_N: Vapour pressure Deficient; Temp_N: annual average mean temperature; Temp_Max_N: annual average maximum temperature; Modis_N: MODIS EVI; slope_d_N: slope; SAGA_WI_N: SAGA Wetness Index. *Variable groups*: red colour: climatic; green colour: organism; blue colour: relief.

Scatter plots for the observed TC vs. predicted TC concentrations are predicted in Fig. 6. On the basis of results of the independent validation, Model 1 and Model 2 can be concluded to exhibit the same model quality. However, Model 2 incorporates a smaller number of environment covariates relative to Model 1 (Fig. 5), thus having lower computational requirements when performing predictions across the landscape.

Hengl et al. (2015) used the RF modelling approach to model and map a variety of soil properties, including soil carbon across the African continent, at a spatial resolution of 250 m. Random Forest was proved to be a more accurate prediction approach than comparatively simpler multiple linear regression models, with an average improvement of mapping accuracy of 20% when performing predictions across a range of climatic conditions from tropical wet climates to hyper-arid climates (Hengl et al., 2015). Similarly, Taghizadeh-Mehrjardi et al. (2016) identified the efficacy of the RF Model for the prediction of the SOC topsoil (0–15 cm) in semi-arid regions in Iran with an LCCC value of 0.66. Dharumarajan et al. (2017), in the semi-arid tropics of India, reported an LCCC value for SOC prediction of 0.38, for a model developed for use within a depth range of 0–30 cm. In comparison, in the current study, both models tested for TC reported much higher LCCC values than those in the studies carried out by Taghizadeh-Mehrjardi et al. (2016) and Dharumarajan et al. (2017).

The sampling density used in the current study for calibration was one site per 11 km^2 ($n = 888$), and the sampling density for validation was one site per 96 km^2 ($n = 99$), where the paddy extent was around 9516 km^2. The sample densities calculated for calibration and validation in different climatic zones in Sri Lanka are depicted in Table 6. Keskin et al. (2019) reported a calibration sampling density of one site per 211 km^2 ($n = 710$), and a validation sampling density of one site per 493 km^2 ($n = 304$) in a study performed across an area of 150,000 km^2 in Florida, United States. Martin et al. (2011) reported a sampling density of one site per 247 km^2 ($n = 2200$) in a study performed across an area of 543,965 km^2 in France. Moreover, Bui et al. (2009), in their study across Australian agricultural zones (2,765,000 km^2), reported a sampling density of one site per 250 km^2. Therefore, the sample density employed in the current study is considerably better than those used in previous studies.

3.5. *Mapping the total soil carbon concentrations across the paddy-growing regions in Sri Lanka*

The distribution of the predicted TC concentrations is depicted in Fig. 7, overlaid across the major climate zones (Wet, Intermediate, and Dry), as derived from both Model 1 and Model 2. A high TC concentration was recorded for paddy fields of the southwestern part that belong to the Wet zone (Table 5). Theoretically, the equilibrium between carbon inputs and decomposition basically governs the sequestration or degradation of organic substances in the soil systems. The high organic-matter soils or Histosols found in general across the paddy-growing soils in the Wet zone form as permanently waterlogged soils. Sahrawat (2004) reported that the loss rate of organic matter in Histosols is slower than its accumulation. Ultisols are the dominant soil type in the Wet zone of Sri Lanka, both in the lowlands and in the central highlands. Ultisols are also found in the Intermediate zone of the country (Moorman and Panabokke, 1961). Despite this, the depressions common to this soil group have been naturally displaced by hydromorphic soil types or Histosols, which create a more suitable environment for paddy cultivation. Furthermore, the Wet zone placed on the windward side of the country receives a high amount of rainfall during the southeast monsoon. Relatively low temperatures prevail throughout the year, and a long period of anoxic conditions resulting in low pH values and associated low decomposition rates may also contribute to the accumulation of high TC concentration in this region, as previously reported by Delgado-Baquerizo et al. (2018). High plant productivity and litter decomposition rates are also seen in areas with high mean annual rainfall, eventually contributing to high atmospheric carbon-fixation rates and SOC accumulation (García-Palacios et al., 2013).

In the Dry zone, high maximum temperatures would contribute towards the storage of less TC compared to in the Wet and Intermediate zones of the country (Fig. 7, Table 5). In addition, in the Dry zone experiencing high evapotranspiration or high VPD often results in a decrease in plant productivity, thereby restricting carbon inflows into the soil system, resulting in low soil carbon storage (Delgado-Baquerizo et al., 2013). Reddish-brown earths are the dominant great soil group in the Dry zone climatic region (USDA Taxonomy: Alfisols; WRB legend: Luvisols). These soils experiences free drainage, and in the geographical depressions or valley areas, this great soil group is substituted by hydromorphic soils such as alluvial soils (USDA Taxonomy: Entisols; WRB legend Fluvisols) and Low-Humic Gley soils (USDA Taxonomy: Alfisols; WRB legend: Gleysols) (Moorman and Panabokke, 1961). To facilitate rice production, low-lying paddy-growing areas in the Dry zone mostly consist of Alfisols and hydromorphic associations, inheriting poorly drained soil characteristics. Eastern Sri Lanka (i.e., the Ampara and Batticaloa administrative units) exhibits lower TC distributions than other paddy-growing areas. According to Moorman and Panabokke

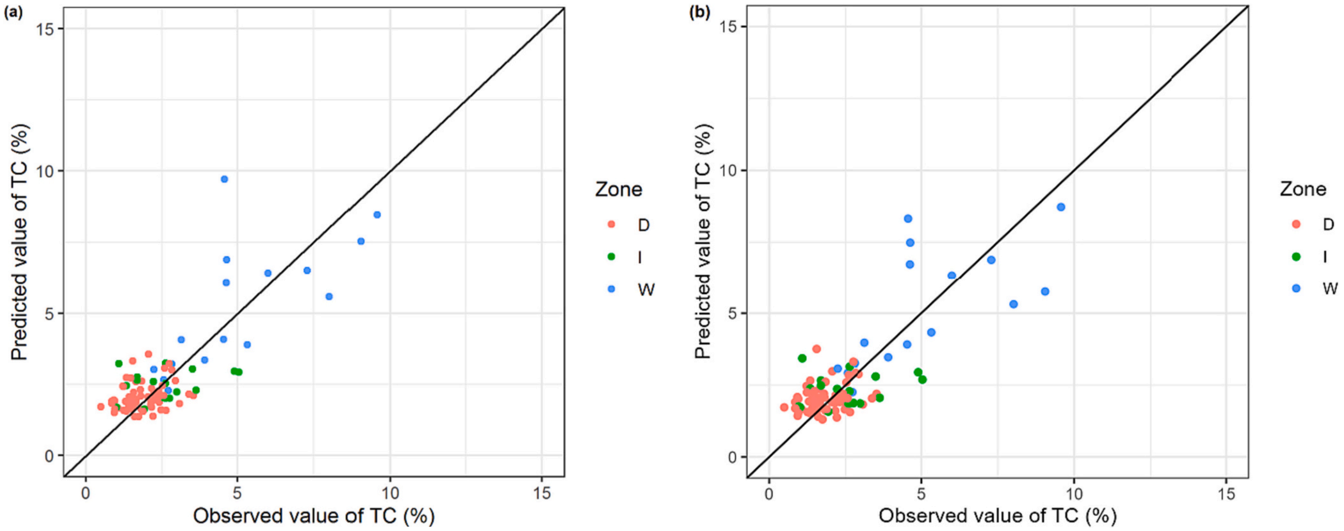

Fig. 6. Scatter plots of observed TC values vs. predicted TC values: Model 1 (a); and Model 2 (b). Observed TC values are related to the independent validation dataset.

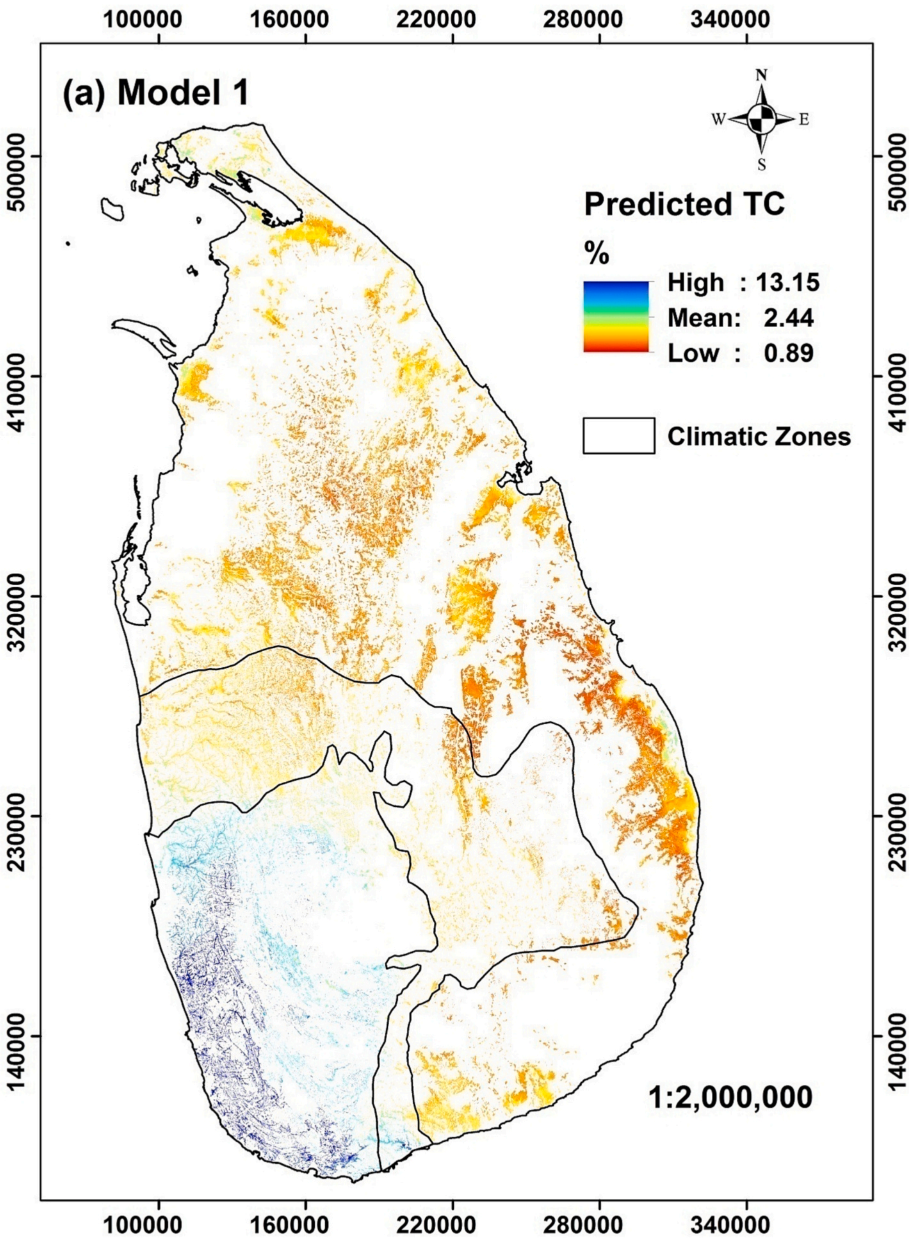

Fig. 7. Spatial distribution of predicted TC concentrations (%) in paddy-growing soils across Sri Lanka with the major climatic zone boundaries. The areas remaining as white colour patches are non-paddy areas, (a) spatial distribution of TC according to Model 1, and (b) spatial distribution of TC according to Model 2.

(1961), the Low-Humic Gley soils associated with the Non-Calcic Brown soils (USDA Taxonomy: Alfisols; WRB legend: Cambisols and Gleysols) in the eastern province have a coarser texture, which can commonly be recognised as being a sandy loam to loamy sand texture and exhibits moderately well-drained characteristics. Those soil characteristics reduce carbon retention ability (mineral-associated carbon) compared to the other soil types in the Dry zone. The predicted values were higher in the North, Northeast, East, and Northwest coastal regions compared to the other areas in the Dry zone. The soils of these areas are formed from recent and older marine sand, lagoons, and shallow seabed deposits. Furthermore, these coastal areas, which are vulnerable to high tide submergence from previous events, are rich in marine clay with the previous decomposition materials of calcareous contents (Dassanayake et al., 2020).

3.6. How reliable are the spatial estimates of the soil total carbon concentrations?

The AOA analysis aided in determining the reliability of current TC predictions (Fig. 8). The AOA function demarcates and shows us the area/land extent to which the predicted model can successfully be applied (Meyer and Pebesma, 2021). The percentages of AOA in each primary climatic zone of Sri Lanka, considering both models, are depicted in Table 6. It can be observed that the Model 1 predictions were reliable for 89.56% of paddy-growing areas across Sri Lanka and unreliable for only 10.44%. Furthermore, the Model 2 predictions were reliable for 89.62% of the area and unreliable for 10.38%. Similar reliability was reported across all paddy-growing regions for both tested models. The spatial predictions of soil TC concentration in the Dry zone can be considered more reliable, followed by the Intermediate and Wet zones, respectively (Table 6). However, Model 1 achieved slightly

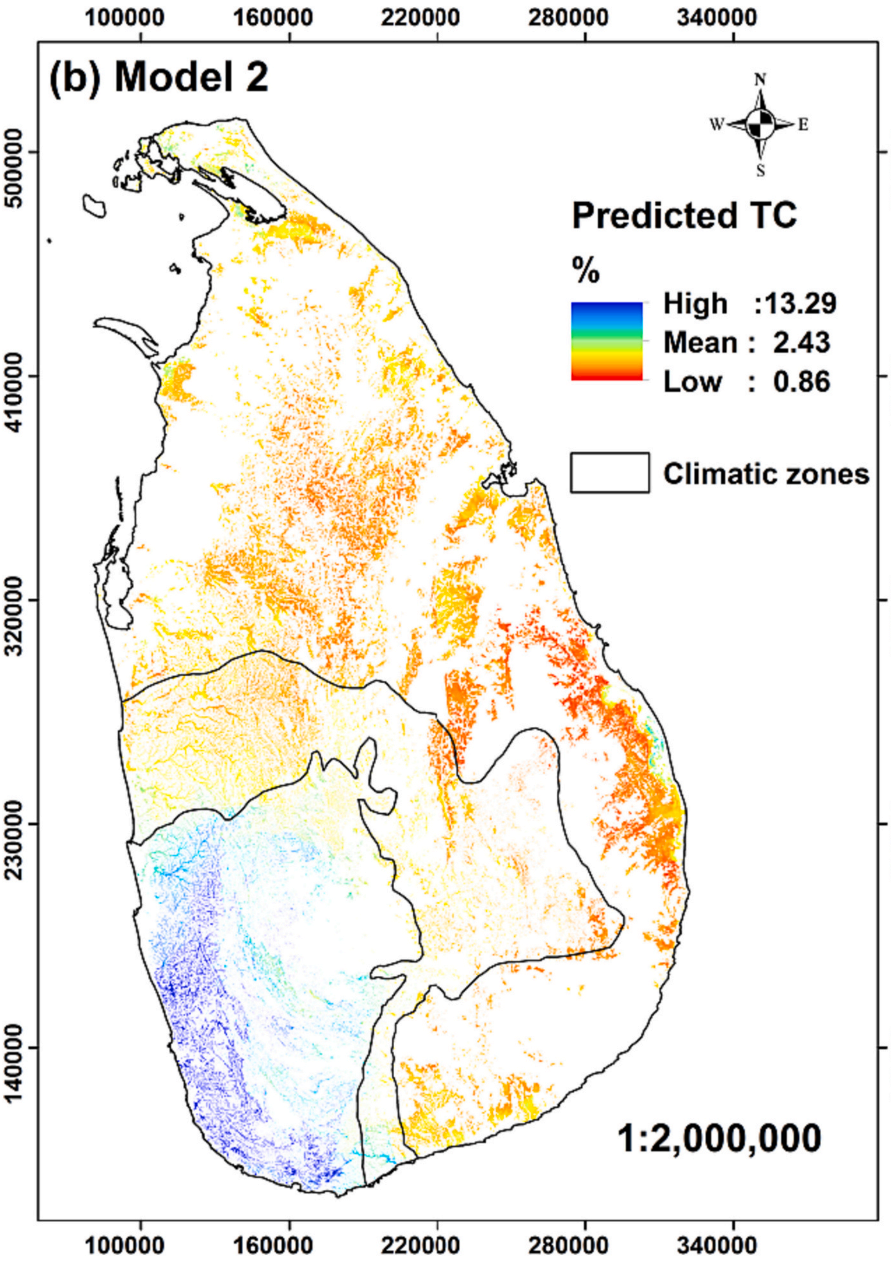

Fig. 7. (*continued*).

Table 5
Summary of predicted total carbon % values in paddy soils across the country and in major climatic zones of Sri Lanka.

	Model 1						Model 2					
Variable	Mean	Min	Max	SD	Q1	Q3	Mean	Min	Max	SD	Q1	Q3
Whole country	2.44	0.89	13.15	1.35	1.74	2.46	2.43	0.86	13.29	1.38	1.71	2.44
Wet	5.27	2.06	13.15	1.84	3.97	6.55	5.32	2.09	13.29	1.91	3.95	6.48
Intermediate	2.43	1.18	7.23	0.71	2.01	2.65	2.37	1.08	6.99	0.71	1.96	2.55
Dry	1.91	0.86	4.30	0.40	1.63	2.14	1.91	0.86	4.30	0.40	1.63	2.14

Note: Min: minimum; Max: maximum; SD: standard deviation; Q1: first quartile; Q3: third quartile.

higher reliability for the Wet zone than Model 2. The Wet zone of Sri Lanka possesses a unique topography and higher temperature variation due to the elevation gradient and annual cumulative rainfall. The report of a less reliable area is most likely due to the current sampling scheme not being able to capture this inherent variability of the environmental covariates that govern the variation of the soil TC concentration in the Wet zone. The AOA results show higher unreliable TC estimates across the Wet zone, which is further supported by the higher model uncertainty values for the same region as depicted by the calculated 90% prediction interval (Appendix, Fig. A2, Table A1). The highest uncertainty of the prediction was recorded in the Wet zone of the country while the lowest was recorded in the Dry zone.

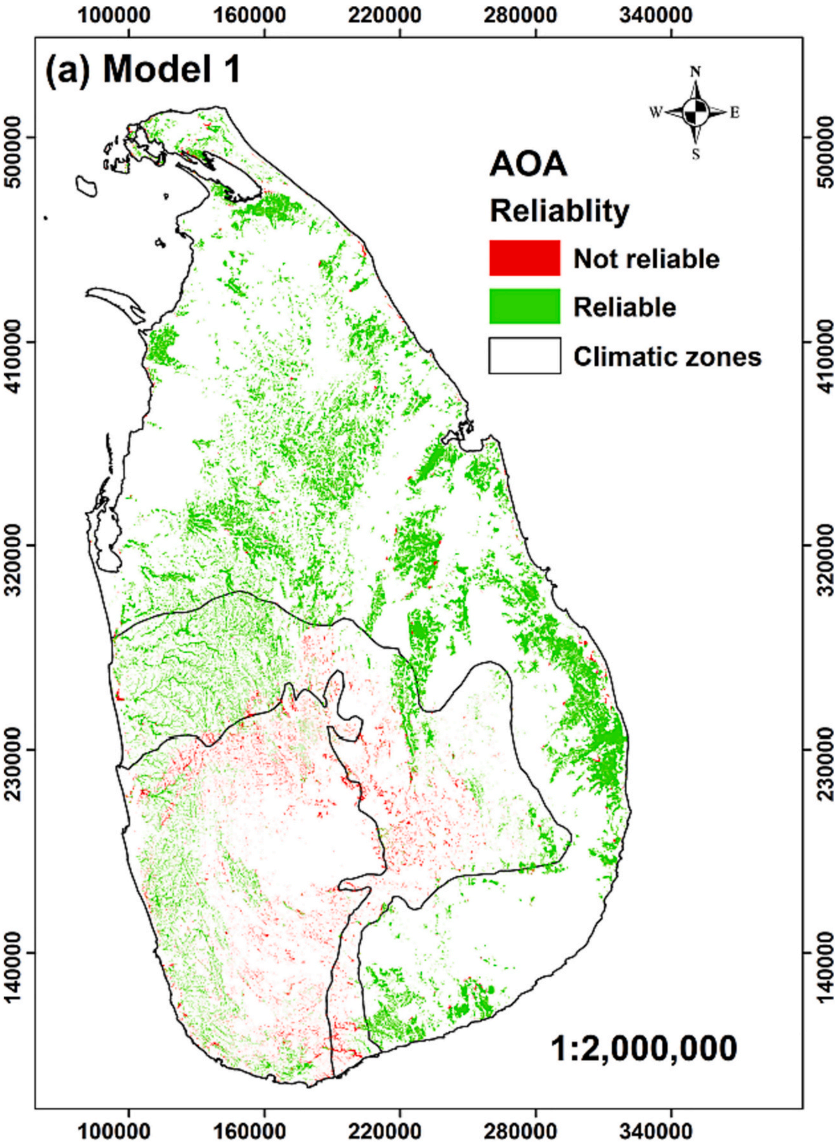

Fig. 8. Area of applicability (AOA) of the soil carbon prediction for the paddy-growing areas across Sri Lanka estimated using Model 1 and Model 2.

In previous studies on paddy soils in the tropical and subtropical regions of the world, Xu et al. (2020) used different multivariate techniques to compare their ability to estimate SOC across soil profiles. As per the superior model prediction, shale contained the highest SOC concentration ranging from 29.42 to 1.73 g kg^{-1} from top to bottom, and quaternary red clay exhibited the lowest, from 22.45 to 0.27 g kg^{-1}. Song et al., 2020 stated that SOC stock at soil depths of 0–20 cm was 27.6 g kg^{-1} in Jiangxi Province, China. Furthermore, several other studies in subtropical and tropical climatic regions studying different soil carbon pools of paddy-growing soils are summarised below, and their results compared with those of current study (Table 7). Relatively lower values of TC concentration were reported in paddy soils in the southeastern part of China recorded relatively lower values of 0.5–1.5% for the depths of 0–15 cm. The mean TC concentration in the Wet zone of Sri Lanka was relatively higher than the soil carbon values reported in other countries, except for Selangor Malaysia, at the same depth level (Aishah et al., 2010). The range of SOC values reported for Lombok Island, Indonesia was quite similar to the values reported for both the Dry and Intermediate zones of Sri Lanka. The TC pool of paddy soil in Madagascar was recorded to be 2.18 ± 1.16% (Kawamura et al., 2017), and this value is greater than the Dry zone mean value in Sri Lanka and less than the mean soil TC concentration in the Wet zone. Furthermore, the recorded SOC% in the Bara district, Nepal, is indicated to be 2.13% (±1.5) (Panday et al., 2018); this value is in accordance with the paddy soils in the Intermediate zone of Sri Lanka.

3.7. Model performances due to number of data cases used for model calibration

The summary of the model performance quality using NSE, RMSE and LCCC for the sequence of calibration models tested with varying calibration sites and the full calibration sites is demonstrated in Fig. 9. As denoted by the model performance quality (Fig. 9), there is an increase in the NSE and LCCC values, and a reduction of RMSE value can be observed for $n = 400$. Beyond this point, the model performance quality indices are stagnated. The NSE values for Model 1 exhibited a slight improvement from 0.30 to 0.40, and for Model 2, the respective values increased from 0.35 to 0.42 as the calibration sample size increased from 400 to 800. Simultaneously, the LCCC values for Model 1 showed enhancement from 0.75 to 0.79, whereas for Model 2, the value remained unchanged at 0.76. The RMSE values ranged from 1.36 to 1.31 in Model 1 and from 1.36 to 1.39 in Model 2 when the calibration sample

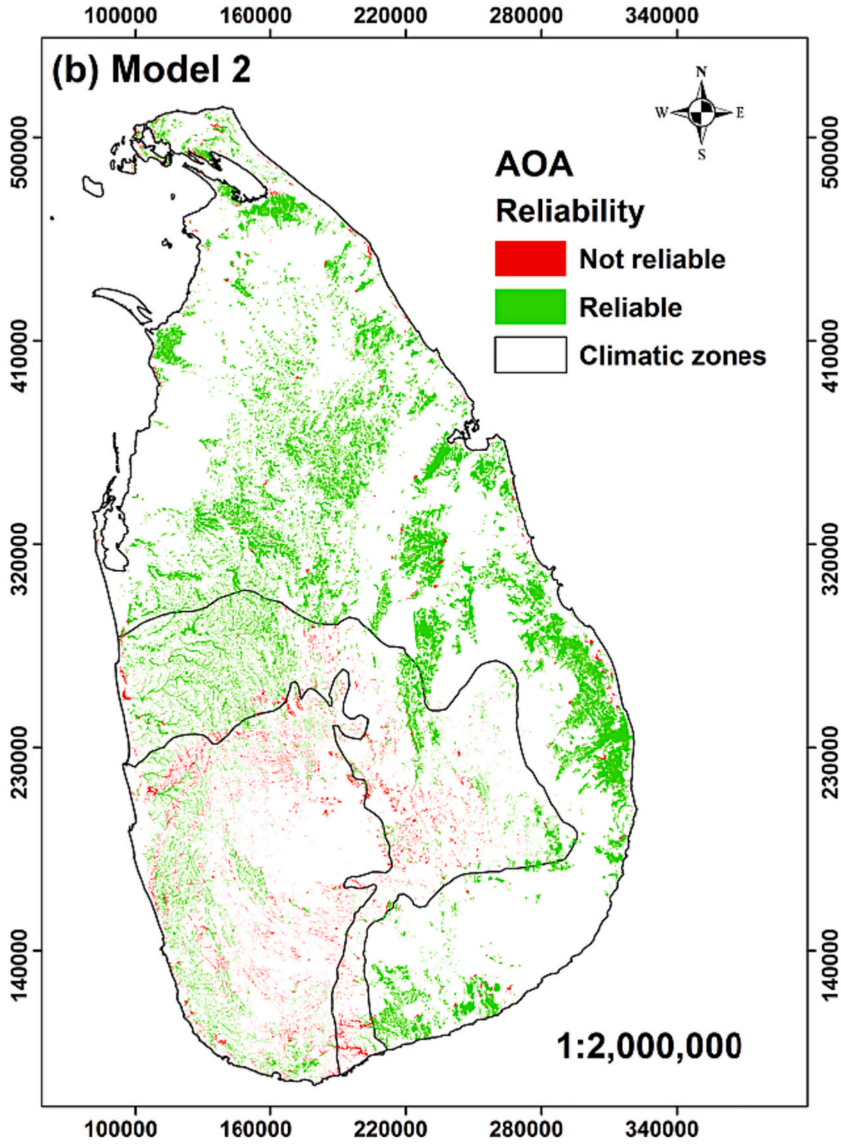

Fig. 8. (*continued*).

Table 6
Summary of sampling densities and percentage area of applicability.

Climatic zone	Area of paddy/km^2	No. of calibration sampling locations	No. of validation sampling locations	Calibration density of sampling/one site per *x*km^2	Validation density of sampling/ one site per *x* km^2	AOA% (Model 1)		AOA% (Model 2)	
						Reliable	Unreliable	Reliable	Unreliable
Wet	1201.94	128	17	9	71	60.74	39.26	53.75	46.25
Intermediate	1760.77	158	18	11	98	76.47	23.53	76.74	23.26
Dry	6562.25	602	64	11	102	98.65	1.35	99.60	0.40

size increased from 400 to 800, as presented in Fig. 9.

As demonstrated by Lagacherie et al. (2020) and Somarathna et al. (2017), the increasing sample size leads to a rise in prediction accuracy at a decreasing rate, irrespective of the specific model employed for the analyses. Morgan et al. (2003) utilised a decision tree-based data mining tool to investigate the impact of sample size on modelling accuracy, revealing that the rate of improvement in model accuracy reaches a plateau after a certain point. Moreover, Saurette et al. (2022) compared Cubist and RF models to ordinary Kriging, and their findings indicated that all three models showed a similar pattern of improvement with increasing sample size aligning with the findings of Morgan et al. (2003). Therefore, the results of the current study are consistent with those of previous studies. However, as Sun et al. (2022) and Long et al. (2020) suggested, the improvement of model performances with increasing sampling sites could also be specific to the landform characteristics of the region.

When developed models are applied across the landscapes, the reliability of models also varies with the number of sites chosen to develop calibration models using cLHS strategy. Technically, selected sites for each calibration model using the cLHS strategy should represent the marginal distribution of global calibration datasets ($n = 888$). Nevertheless, the increasing number of sites within calibration sites has influenced capturing the complex variation across the landscape features. At the country scale comparison, whether the variable selection is

Table 7
Summary of relevant studies on the modelling and mapping of soil carbon.

Region	Land use	Soil depth(cm)	No: of samples	Model	Soil carbon predictor	Soil carbon pool	Mean soil carbon stocks/ concentration in paddy soil	References
Yujiang County, Jiangxi Province, China	Paddy	100 ± 5 (vertical distribution from top to bottom)	306 (calibration = 214, validation = 92)	PLSR, ANN, Cubist, GPR, and SVMR with the CARS		SOC	In different parent materials: • Red sandstone: 24.76–0.51 g kg^{-1} • Shale: 29.42–1.73 g kg^{-1} • River alluvium: 26.61–0.65 g kg^{-1} • Quaternary red clay: 22.45–0.27 g kg^{-1}	(Xu et al., 2020)
Jiangxi Province, China	Paddy, Upland soil, Forest	0–20 20–40	256	MLR, RK	Land use, Elevation, Parent material	SOC	0–20: 27.6 g kg^{-1} 20–40: 12.11 g kg^{-1}	(Song et al., 2020)
Jinjing catchment, China	woodlands, paddy fields and tea fields	0–20	1033	GWR, OK, IDW, LMR, LMM,	Elevation, Slope, TWI, Land use	SOC	3.50 kg^{-2}	(Liu et al., 2017)
South eastern part of China	Paddy	0–15	212	MLR, OK, SK, RK	NDVI, Elevation, Elevation above nearest drainage path, TWI	TC	0.5–1.5%	(Sumfleth and Duttmann, 2008)
Selangor, Malaysia	Paddy	0–20	138, 30 extra points for validation	K	–	SOC	3–5%	(Aishah et al., 2010)
Lombok Island, Indonesia	Paddy	0–10	150	PLSR	–	SOC	0.90–2.98%	(Kusumo et al., 2018)
Central highland of Madagascar, Sothern Africa	Paddy	0–10 (mainly)	59	Vis-NIR diffuse reflectance spectroscopy, PLS	–	TC	2.18% (±1.16)	(Kawamura et al., 2017)
Bara district, Nepal	Paddy	0–15	109	OK	–	SOC	2.13% (±1.5)	(Panday et al., 2018)
Sri Lanka, Northern Province	Paddy	0–15 15–30	83	LMM	DEM, WI, ARF, MT, NDVI	SOC	0–15: 1.78% (±0.78) 15–30:1.03% (±0.47)	(Ratnayake et al., 2016)
Current study (whole Sri Lanka)	Paddy	0–15	987	RF	Rainfall, temperature, VPD. MODIS EVI, slope, TWI	TC	Wet zone: 5.36% (±2.07) Intermediate zone: 2.40% (±0.0.74) Dry zone: 1.89% (±0.41)	Current study

Note: *Soil carbon pool abbreviations*: Soil organic carbon (SOC), total carbon (TC); *Model name abbreviations*: Partial Least Square Regression (PLSR), Artificial Neural Network (ANN), Gaussian process regression (GPR), Support Vector Machine Regression (SVMR), Competitive Adaptive Reweighted Sampling (CARS), Multiple Linear Regression (MLR), Regression Kriging (RK), kriging (K), Ordinary Kriging (OK), Simple Kriging (SK), Geographically Weighted Regression (GWR), Inverse Distance Weighted (IDW), visible and near-infrared (Vis-NIR), Linear Mixed-effects Model (LMM), Random Forest (RF); *SOC predictor abbreviations*: laboratory-based hyper-spectral imaging (HSI), topographic wetness index (TWI), normalised difference vegetation Index (NDVI), digital elevation model (DEM), wetness index (WI), annual rainfall (ARF), mean temperature (MT), vapour pressure deficient (VPD), enhanced vegetation index (EVI).

performed (Model 2) or not (Model 1) resulted in similar reliability but model 2 was more stable. Notably, the Dry zone of Sri Lanka, which has undulating terrain with less variability with landscape and climatic variation, resulted in almost no change in the estimates' reliability with an increase of the calibration sites for the two tested models (Fig. 10). The model reliability decreases for the Intermediate and Wet zones with the increasing number of model calibration sites after $n > 400$ sites. This coincides with the calibration model quality evaluated using the fully independent dataset also revealed the model prediction quality stagnated after $n > 400$ calibration sites (Fig. 9).

As the study area expands to a larger spatial scale, the soil carbon distribution may become more heterogeneous, making capturing all variations challenging. The relationship between sample size and modelling performance can be influenced by various factors, such as the spatial scale of the study and the heterogeneity of the soil carbon distribution (Stevens et al., 2013). Therefore, the relationship between sample size and modelling performance may not always be linear. In some cases, the improvement in modelling performance reaches a point beyond which further increases in sample size do not significantly enhance the model performances (Wartini et al., 2020). This saturation may occur when the existing sample size adequately captures the dominant predictors used for the soil carbon modelling and additional samples do not provide substantial new information. Therefore, these results indicate that the locations selected in $n = 400$ adequately represent the study area's soil-environment relationships under the current modelling framework. The variable importance plot of Model 2 for $n = 400$ showed that almost all the variables used in Model 1 were utilised in developing Model 2, except for the slope angle (Fig. S2). It was also revealed that despite changing calibration sites, top three key variables that were identified includes mean annual rainfall, annual average maximum temperature and annual average mean temperature. Furthermore, relatively lesser significance of the slope angle and the

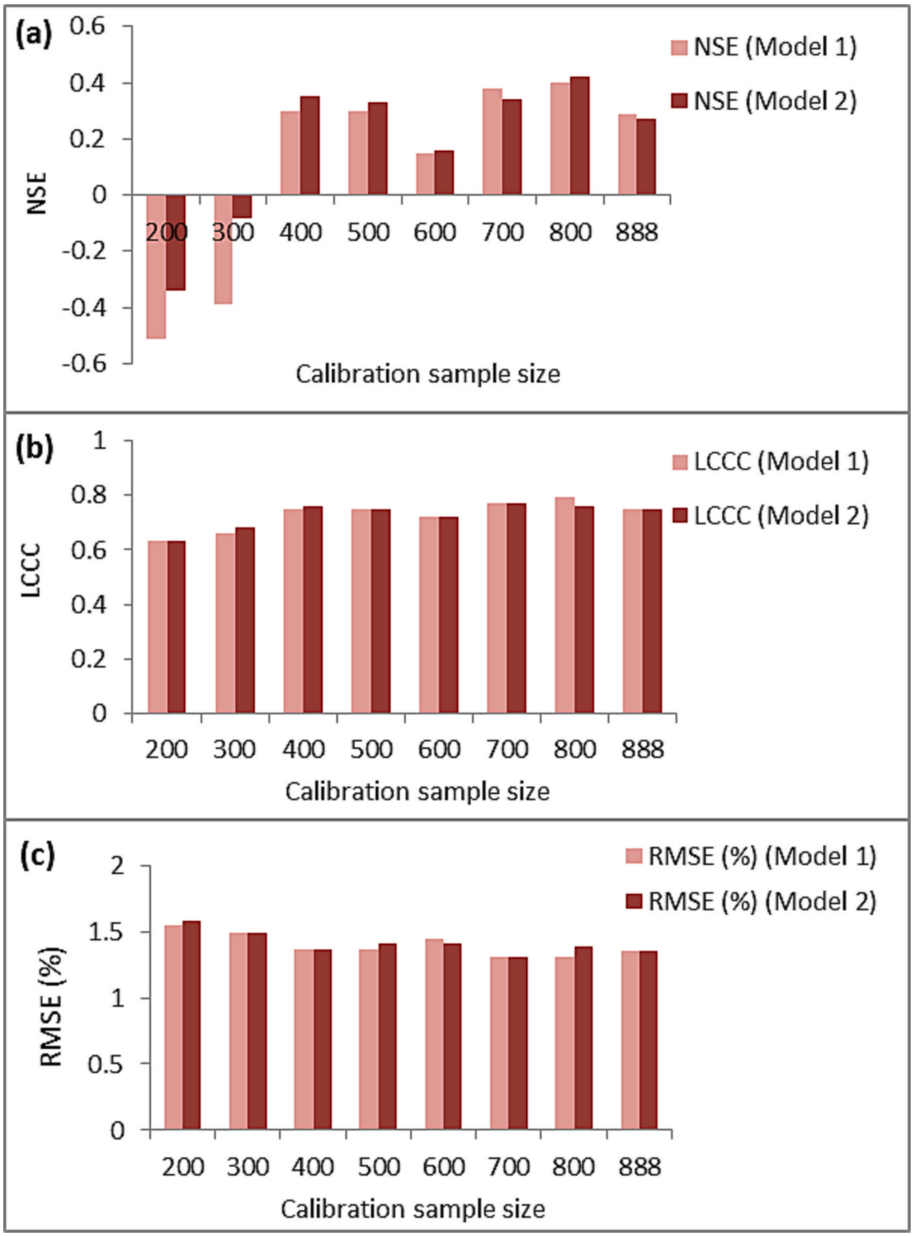

Fig. 9. Comparison of the model quality with the full calibration dataset with sequence of increase of the calibration dataset for Model 1 and Model 2, (a) Nash-Sutcliffe model efficiency coefficient (NSE), (b) Lin's Concordance Correlation Coefficient (LCCC), (c) Root-Mean Square Error (RMSE%).

SAGA wetness index is evident.

The supplementary material includes the spatial distribution of the calibration sites sequence used for further comparison with the full calibration sites (Fig. S1), the site distribution with major climatic zones (Table S1), key drivers identified using variable importance plot for Model 1 and Model 2 (Fig. S2), and fully independent validation plots (Fig. S3).

3.8. *Model prediction uncertainty*

The prediction uncertainty associated with spatial output layers generated using ML algorithms can be assessed through variety of approaches. Commonly used methods for analysing prediction uncertainty in ML algorithms include: (a) model-embedded quantile regression, such as the quantile random forest estimator (Wadoux et al., 2023), enabling the quantification of prediction interval coverage (Wadoux, 2019); (b) ensemble model development using bootstrapping (Rossel et al., 2014); and more recently (c) utilising the AOA concept (Meyer and Pebesma, 2021). In the current study, two different approaches were employed to quantify prediction uncertainties, namely bootstrapping and the AOA concept. Both tested methods resulted in similar overall patterns of uncertainty estimation across the landscape (Figs. 8 and A2). The AOA concept, as presented in Meyer and Pebesma (2021), provides a systematic approach to assess the applicability and uncertainty of spatial prediction models. The AOA function delineates regions where spatial prediction models offer reliable and accurate estimations, enhancing our understanding of the model's prediction limitations and associated uncertainties.

3.9. *The practical implication of the derived outputs*

The derived spatial estimates across all rice production regions in Sri Lanka provide the first-ever detailed country-scale assessment of TC concentration. These spatial estimates can be used to formulate an

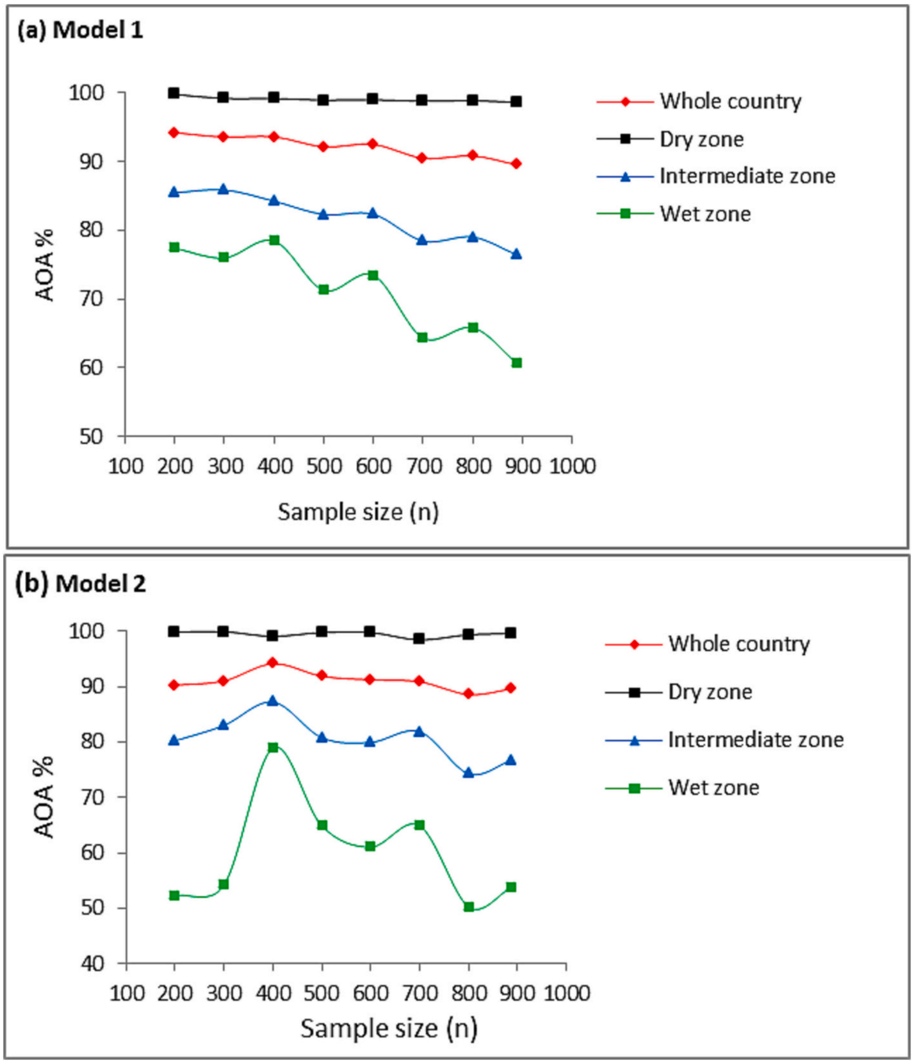

Fig. 10. Comparison of the model percentage area of reliability using area of applicability (AOA%) concept considering broad climate zones and whole country using two different model fitting processes.

integrated management strategy to enhance rice productivity by increasing the TC concentration, mainly through SOC. In designing such strategies, the best-performing lands within a given region (based on the drivers of the TC concentrations) should be considered as determining the practically attainable TC concentrations. Hence, management strategies should be developed using the thus-defined attainable TC concentrations by considering both strategic tillage and improved stubble management. This strategy will lead to the development of other land parcels with low levels of TC compared to existing land parcels in the same geographic region in order to move closer towards the attainable TC limit. This approach is important, as the values of TC concentration are clearly distinct across the major climatic zones, being predominantly affected by annual rainfall.

Furthermore, spatial TC concentration estimates provide a carbon baseline for future carbon trading, which relies on greenhouse gas emissions and subsequent storage capacity. Hence, the project can help identify and prioritise potential locations for soil-based carbon sequestration projects. In fact, the generated baseline datasets will be immensely useful for Sri Lanka to develop IPCC Tier 3 carbon accounting model for the land sector in the near future. The estimated spatial soil TC concentration values within the areas that were identified as being less reliable through AOA should be interpreted with caution, due to the associated uncertainty of the estimates. Furthermore, those regions can be used to define future targeted field sampling campaigns to improve the reliability of the spatial estimates of the soil TC concentrations.

3.10. *Limitations of the current study*

Even though the best possible environmental covariates at 100 m resolution were used for the current predictions, there are practical limitations to how much improvement can be achieved through their combination. Integrating target variable data and environmental covariates is crucial for accurately understanding and predicting the spatial distribution of soil carbon. However, the combination of these data reaches a saturation point in terms of model improvement. Beyond this point, introducing more data can even have a counterproductive effect, potentially leading to a decrease in model reliability. When additional data points are introduced, they can create complex or diverse relationships that the existing model structure might not be capable of capturing. The intricacies of these new relationships may be difficult to align with the patterns identified by the initial covariate dataset, which can lead to an unexpected outcome in the overall performance of the model. Therefore, some limitations associated with selected environmental covariates and their combined effect could reduce the model's performance and reliability. In such cases, introducing alternative or new sets of covariates could be a potential strategy to address this limitation. However, the success of this approach is cannot be guaranteed.

The efficacy of these alternative covariates depends on whether the new relationships they introduce align with the underlying dynamics of the soil carbon in the region.

4. Conclusions

Both RF models fitted in the current study exhibited similar model quality, while Model 2 incorporates a smaller number of environmental covariates compared to Model 1. A series of calibration models with varying sample sizes were evaluated, and their performances improved until $n = 400$, after which they stagnated. In the country-scale comparison, both Model 1 and Model 2 yielded similar reliability, while Model 2 was more stable. In the Dry zone, both models exhibited comparable reliability with an increasing number of calibration sites. However, in the Wet and Intermediate zones, model reliability decreased after, $n > 400$. The results suggested that the locations selected in $n = 400$ adequately reflect the study area's soil-environment relationships under the current modelling framework. The derived AOA maps be used to target additional samples first to improve model quality, followed by spatial estimates. In this case, recognizing the potential challenges and intricacies of incorporating additional samples or environmental predictors is pivotal in maintaining realistic outcomes regarding model performance. In this study, the first-ever detailed baseline spatial estimates of TC concentration across the paddy-growing regions in Sri Lanka were derived. The derived maps will be pivotal for allocating resources to enhance the TC, mainly through SOC management with the aim of enhancing soil health and rice productivity. The regional differences in soil carbon distribution could be helpful in Sri Lanka for planning site-specific fertilizer recommendations for rice cultivation. Well-demarcated AOA maps are vital to avoiding possible deception when utilising such predictive maps to assist decision making. Furthermore, enhancing the TC concentration will act as an offset strategy for the mitigation of climate change.

CRediT authorship contribution statement

T.M. Paranavithana: Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **S.B. Karunaratne:** Conceptualization, Data curation, Formal analysis, Funding acquisition, Supervision, Validation, Visualization, Writing – review & editing, Methodology. **N. Wimalathunge:** Formal analysis, Validation, Visualization, Writing – review & editing. **B.P. Malone:** Investigation, Writing – review & editing. **B. Macdonald:** Writing – review & editing. **T.F.A. Bishop:** Conceptualization, Writing – review & editing. **R.R. Ratnayake:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing, Investigation.

Declaration of Competing Interest

None.

Data availability

The data that support the findings of this study are openly available in Mendeley data repository at Paranavithana, Thilakshi; Karunaratne, Senani; Wimalathunge, Niranjan; Malone, Brendan; Macdonald, Ben; Bishop, Thomas; Ratnayake, Renuka (2023), "Raw dataset and R scripts for: Unravelling spatial drivers of topsoil total carbon variability in tropical paddy soils of Sri Lanka", Mendeley Data, V3, doi: 10.17632/dnw2v82r8y.3 NB:The raw data set, R scripts, and raster files of environmental covariates used for the analyses will be publicly available within two working days.

Acknowledgments

We wish to acknowledge the financial assistance from the National Research Council (NRC), Sri Lanka (Grant no: 17-011). Further, we wish to thank Ms. K. Karunaratne, Mr. A. Pathirana for technical assistance, and Mr. L. Mahadeva, National Institute of Fundamental Studies (NIFS), Kandy, for supporting soil sampling.

Appendix A

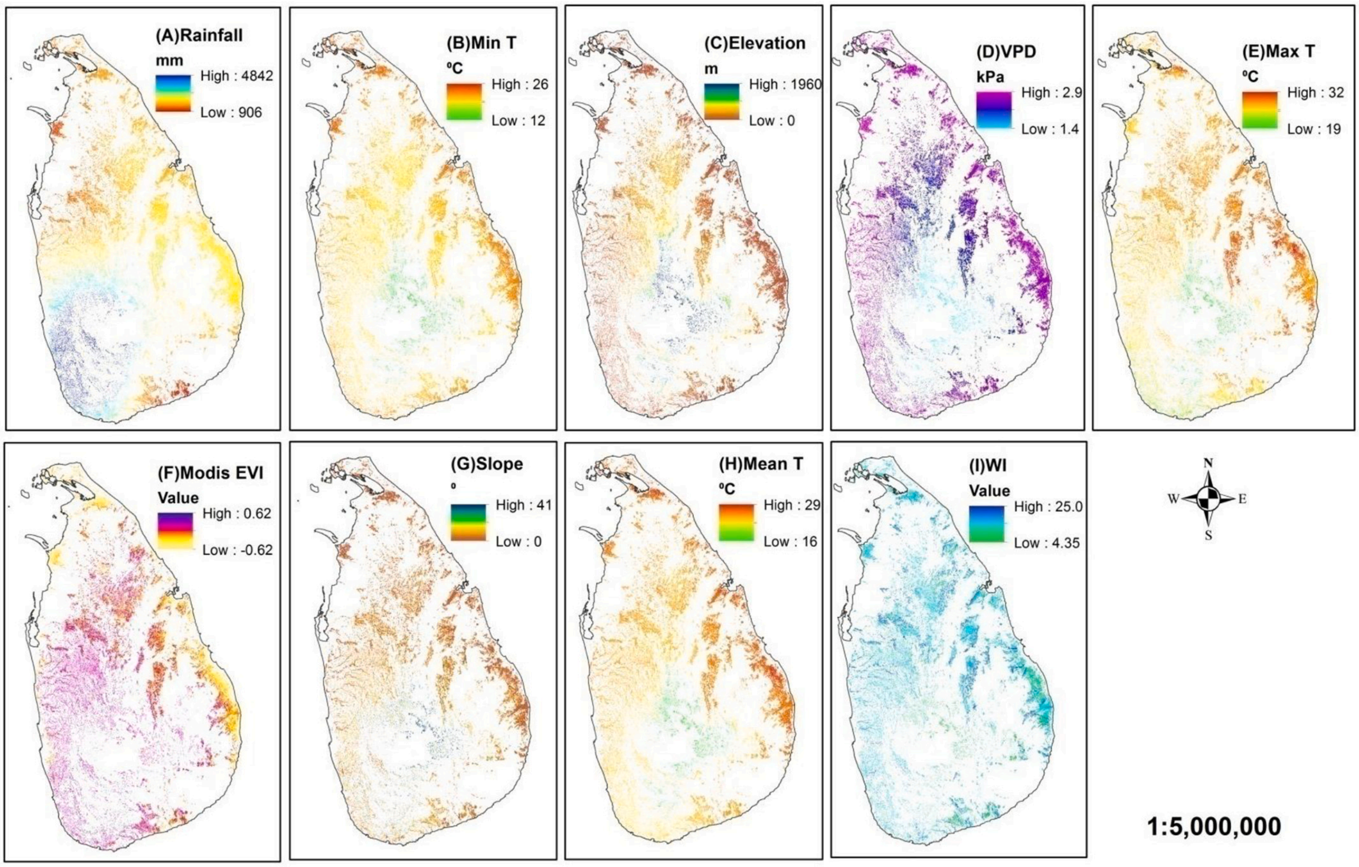

Fig. A1. Spatial distribution of environmental covariates (A) Mean annual Rainfall (Rainfall) (B) Annual average minimum Temperature (Min T) (C) Elevation (D) Vapour Pressure Deficient (VPD) (E) Annual average maximum Temperature (Max T) (F) MODIS Enhanced Vegetation Index (Modis EVI) (G) Slope angle (H) Annual average mean Temperature (MIT) (I) SAGA Wetness Index (WI).

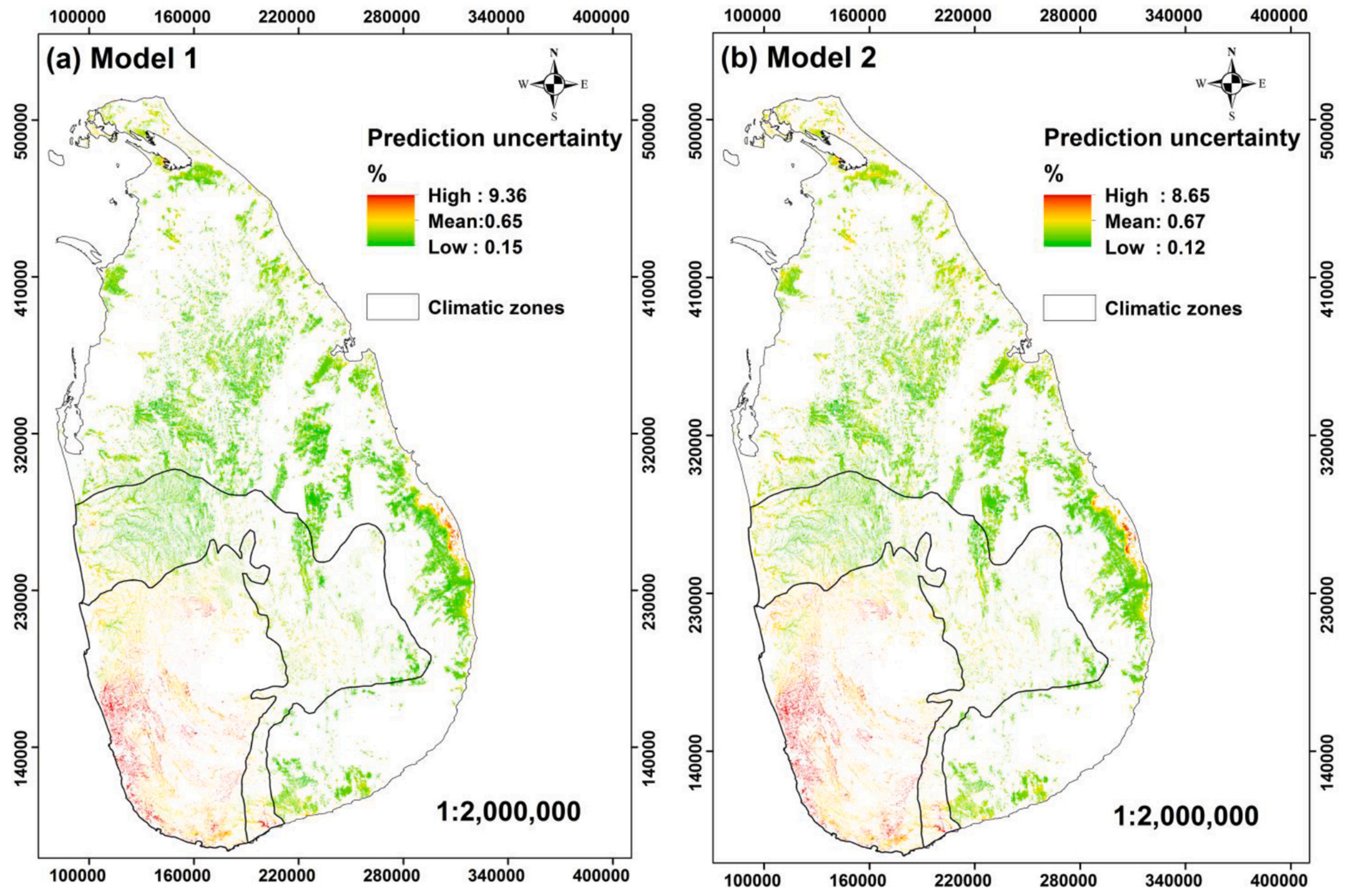

Fig. A2. Uncertainty of TC prediction (%) in paddy-growing soils across Sri Lanka with the major climatic zone boundaries derived through a calculated 90% prediction interval. The areas remaining as white colour patches are non-paddy areas, (a) Uncertainty of prediction according to Model 1, and (b) Uncertainty of prediction according to Model 2.

Table A1
Descriptive statistics for the uncertainty of total carbon prediction (%) in paddy soils across the country and in major climatic zones of Sri Lanka.

	Model 1						Model 2					
Variable	Mean	Min	Max	SD	Q1	Q3	Mean	Min	Max	SD	Q1	Q3
Whole country	0.65	0.15	9.36	0.58	0.37	0.67	0.67	0.12	8.65	0.52	0.40	0.72
Wet	1.62	0.26	9.36	1.06	0.99	1.84	1.56	0.30	8.65	0.91	1.02	1.81
Intermediate	0.57	0.16	3.78	0.33	0.36	0.66	0.60	0.13	3.79	0.32	0.39	0.70
Dry	0.49	0.15	3.10	0.20	0.36	0.55	0.52	0.12	3.17	0.21	0.39	0.61

Note: Min: Minimum, Max: Maximum, SD: Standard Deviation, Q1: first quartile, Q3: third quartile.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geodrs.2023.e00745.

References

Aishah, A., Zauyah, S., Anuar, A., Fauziah, C., 2010. Spatial variability of selected chemical characteristics of paddy soils in Sawah Sempadan, Selangor, Malaysia. Malays. J. Soil Sci. 14, 27–39.

Aitkenhead, M.J., Coull, M.C., 2016. Mapping soil carbon stocks across Scotland using a neural network model. Geoderma 262, 187–198.

Anderson, J., Ingram, J., 1993. Tropical Soil Biological and Fertility: A Handbook of Methods, 2 ed. CAB International, Wallingford, p. 221.

Bohner, J., Selige, T., 2006. Spatial Prediction of Soil Attributes Using Terrain Analysis and Climate Regionalisation. SAGA-Analyses and modelling applications, Goltze.

Breshears, D.D., Adams, H.D., Eamus, D., McDowell, N., Law, D.J., Will, R.E., Williams, A.P., Zou, C.B., 2013. The critical amplifying role of increasing atmospheric moisture demand on tree mortality and associated regional die-off. Front. Plant Sci. 4, 266.

Brus, D., Kempen, B., Heuvelink, G., 2011. Sampling for validation of digital soil maps. Eur. J. Soil Sci. 62 (3), 394–407.

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian soil resource information system database to inform soil carbon mapping in Australia. Glob. Biogeochem. Cycles 23 (4).

Dassanayake, A., De Silva, G., Mapa, R.B., 2020. Major soils of the dry zone and their classification. In: The Soils of Sri Lanka. Springer, pp. 49–67.

De Blecourt, M., Corre, M.D., Paudel, E., Harrison, R.D., Brumme, R., Veldkamp, E., 2017. Spatial variability in soil organic carbon in a tropical montane landscape: associations between soil organic carbon and land use, soil properties, vegetation, and topography vary across plot to landscape scales. Soil 3 (3), 123–137.

Delgado-Baquerizo, M., Maestre, F.T., Gallardo, A., Bowker, M.A., Wallenstein, M.D., Quero, J.L., Ochoa, V., Gozalo, B., García-Gómez, M., Soliveres, S., 2013. Decoupling of soil nutrient cycles as a function of aridity in global drylands. Nature 502 (7473), 672–676.

Delgado-Baquerizo, M., Karunaratne, S.B., Trivedi, P., Singh, B.K., 2018. Climate, geography, and soil abiotic properties as modulators of soil carbon storage. In: Soil Carbon Storage. Elsevier, pp. 137–165.

Dewi, C., Chen, R.C., 2019. Random forest and support vector machine on features selection for regression analysis. Int. J. Innov. Comput. Inf. Control 15, 2027–2037.

Dhanapala, M.P., 2007. Bridging the Rice Yield Gap in Sri Lanka. https://coin.fao.org/coin-static/cms/media/9/13171760277090/2000_16_high.pdf#page=141/ (accessed 26 February 2021).

Dharumarajan, S., Hegde, R., Singh, S., 2017. Spatial prediction of major soil properties using random forest techniques-a case study in semi-arid tropics of South India. Geoderma Reg. 10, 154–162.

Ding, C., Du, S., Ma, Y., Li, X., Zhang, T., Wang, X., 2019. Changes in the pH of paddy soils after flooding and drainage: modeling and validation. Geoderma 337, 511–513.

Fadeeva, V., Tikhova, V., Nikulicheva, O., 2008. Elemental analysis of organic compounds with the use of automated CHNS analysers. J. Anal. Chem. 63 (11), 1094–1106.

Fantappie, M., L'Abate, G., Costantini, E., 2011. The influence of climate change on the soil organic carbon content in Italy from 1961 to 2008. Geomorphology 135 (3–4), 343–352.

Forkuor, G., Hounkpatin, O.K., Welp, G., Thiel, M., 2017. High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple linear regression models. PLoS One 12 (1), e0170478.

García-Palacios, P., Maestre, F.T., Kattge, J., Wall, D.H., 2013. Climate and litter quality differently modulate the effects of soil fauna on litter decomposition across biomes. Ecol. lett. 16 (8), 1045–1053.

Gattinger, A., Muller, A., Haeni, M., Skinner, C., Fliessbach, A., Buchmann, N., Mäder, P., Stolze, M., Smith, P., Scialabba, N.E.-H., 2012. Enhanced top soil carbon stocks under organic farming. Proc. Natl. Acad. Sci. 109 (44), 18226–18231.

Ghimire, R., Lamichhane, S., Acharya, B.S., Bista, P., Sainju, U.M., 2017. Tillage, crop residue, and nutrient management effects on soil organic carbon in rice-based cropping systems: a review. J. Integr. Agric. 16 (1), 1–15.

Girsang, S.S., Quilty, J.R., Correa Jr., T.Q., Sanchez, P.B., Buresh, R.J., 2019. Rice yield and relationships to soil properties for production using overhead sprinkler irrigation without soil submergence. Geoderma 352, 277–288.

Gray, J., Karunaratne, S., Bishop, T., Wilson, B., Veeragathipillai, M., 2019. Driving factors of soil organic carbon fractions over New South Wales, Australia. Geoderma 353, 213–226.

Haque, M.M., Biswas, J., Maniruzaman, M., Akhter, S., Kabir, M., 2020. Carbon sequestration in paddy soil as influenced by organic and inorganic amendments. Carbon Manag. 11 (3), 231–239.

Hengl, T., Heuvelink, G.B., Kempen, B., Leenaars, J.G., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. PLoS One 10 (6), e0125814.

Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. Geoderma 214, 141–154.

Hinge, G., Surampalli, R.Y., Goyal, M.K., 2018. Prediction of soil organic carbon stock using digital mapping approach in humid India. Environ. Earth Sci. 77 (5), 1–10.

Huang, S., Rui, W., Peng, X., Huang, Q., Zhang, W., 2010. Organic carbon fractions affected by long-term fertilization in a subtropical paddy soil. Nutr. Cycl. Agroecosyst. 86 (1), 153–160.

Karunaratne, S., Bishop, T., Odeh, I., Baldock, J., Marchant, B., 2014. Estimating change in soil organic carbon using legacy data as the baseline: issues, approaches and lessons to learn. Soil Res. 52 (4), 349–365.

Karunaratne, S., Thomson, A., Morse-McNabb, E., Wijesingha, J., Stayches, D., Copland, A., Jacobs, J., 2020. The fusion of spectral and structural datasets derived from an airborne multispectral sensor for estimation of pasture dry matter yield at paddock scale with time. Remote Sens. 12 (12), 2017.

Kawamura, K., Tsujimoto, Y., Rabenarivo, M., Asai, H., Andriamananjara, A., Rakotoson, T., 2017. Vis-NIR spectroscopy and PLS regression with waveband selection for estimating the total C and N of paddy soils in Madagascar. Remote Sens. 9 (10), 1081.

Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. Geoderma 339, 40–58.

Khaledian, Y., Miller, B.A., 2020. Selecting appropriate machine learning methods for digital soil mapping. Appl. Math. Model. 81, 401–418.

Komatsuzaki, M., Ohta, H., 2007. Soil management practices for sustainable agro-ecosystems. Sustain. Sci. 2 (1), 103–120.

Kusumo, B., Sukartono, S., Bustan, B., 2018. Rapid measurement of soil carbon in rice paddy field of Lombok Island Indonesia using near infrared technology. In: IOP Conference Series: Materials Science and Engineering, p. 012014.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Nkuba-Kasanda, L., 2020. Analysing the impact of soil spatial sampling on the performances of digital soil mapping models and their evaluation: a numerical experiment on quantile random forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. Geoderma 375, 114503.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review. Geoderma 352, 395–413.

Liu, Z., Liu, Q., 2014. Magnetic properties of two soil profiles from Yan'an, Shaanxi Province and their implications for paleorainfall reconstruction. Sci. China Earth Sci. 57 (4), 719–728.

Liu, H., Zhou, J., Feng, Q., Li, Y., Li, Y., Wu, J., 2017. Effects of land use and topography on spatial variety of soil organic carbon density in a hilly, subtropical catchment of China. Soil Res. 55 (2), 134–144.

Long, J., Smith, K., Chen, W., 2020. Landform characteristics and their influence on model performances. Geogr. Res. 18 (2), 134–149.

Mapa, R.B., 2020. Soil research and soil mapping history. In: The Soils of Sri Lanka. Springer, pp. 1–11.

Martin, M., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8 (5), 1053–1065.

Martin, M., Orton, T., Lacarce, E., Meersmans, J., Saby, N., Paroissien, J., Jolivet, C., Boulonne, L., Arrouays, D., 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. Geoderma 223, 97–107.

McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.

Meetei, T.T., Kundu, M.C., Devi, Y.B., 2020. Long-term effect of rice-based cropping systems on pools of soil organic carbon in farmer's field in hilly agro-ecosystem of Manipur, India. Environ. Monit. Assess. 192 (4), 1–17.

Meyer, H., Pebesma, E., 2021. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods Ecol. Evol. 12 (9), 1620–1633.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., Nauss, T., 2018. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ. Model Softw. 101, 1–9.

Meyer, H., Reudenbach, C., Wöllauer, S., Nauss, T., 2019. Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. Ecol. Model. 411, 108815.

Meyer, H., Reudenbach, C., Ludwig, M., Nauss, T., Meyer, M.H., 2020. Package 'CAST', 13. CARN, Wien. https://mirror.lyrahosting.com/CARN/web/packages/CAST/CAST.pdf/ (accessed 04 November 2021).

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32 (9), 1378–1388.

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86.

Mishra, U., Yeo, K., Adhikari, K., Riley, W.J., Hoffman, F.M., Hudson, C., Gautam, S., 2022. Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning. Soil Sci. Soc. Am. J. 86 (6), 1611–1624.

Moorman, F., Panabokke, C., 1961. Soils of Ceylon. Trop. Agric. 117 (I), 22–23.

Morgan, J., Daugherty, R., Hilchie, A., Carey, B., 2003. Sample size and modeling accuracy of decision tree based data mining tools. J. Manag. Inf. Decis. Sci. 6 (2), 77–91.

Neina, D., 2019. The role of soil pH in plant nutrition and soil remediation. Appl. Environ. Soil Sci. 2019.

Panday, D., Maharjan, B., Chalise, D., Shrestha, R.K., Twanabasu, B., 2018. Digital soil mapping in the Bara district of Nepal using kriging tool in ArcGIS. PLoS One 13 (10), e0206350.

Paranavithana, T.M., Marasinghe, S., Perera, G.A.D., Ratnayake, R.R., 2020. Effects of crop rotation on enhanced occurrence of arbuscular mycorrhizal fungi and soil carbon stocks of lowland paddy fields in seasonaly dry tropics. Paddy Water Environ. 1–10.

Peng, X., Shi, T., Song, A., Chen, Y., Gao, W., 2014. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. Remote Sens. 6 (4), 2699–2717.

Punyawardena, B.V.R., 2020. Climate. In: The Soils of Sri Lanka. Springer, pp. 13–22.

Qadir, M., Noble, A., Schubert, S., Thomas, R.J., Arslan, A., 2006. Sodicity-induced land degradation and its sustainable management: problems and prospects. Land Degrad. Dev. 17 (6), 661–676.

Rahman, S., Parkinson, R.J., 2007. Productivity and soil fertility relationships in rice production systems, Bangladesh. Agric. Syst. 92 (1-3), 318–333.

Rajapaksha, R., Karunaratne, S., Biswas, A., Paul, K., Madawala, H., Gunathilake, S., Ratnayake, R., 2020. Identifying the spatial drivers and scale-specific variations of soil organic carbon in tropical ecosystems: a case study from Knuckles forest reserve in Sri Lanka. For. Ecol. Manag. 474, 118285.

Rajkishore, S., Natarajan, S., Manikandan, A., Vignesh, N., Balusamy, A., 2015. Carbon Sequestration in Rice Soils–A Review. https://www.researchgate.net/publication/281935281_carbon_sequestration_in_rice_soils_a_review (accessed 20 April 2020).

Ratnayake, R., Kugendren, T., Gnanavelrajah, N., 2014. Changes in soil carbon stocks under different agricultural management practices in North Sri Lanka. J. Natl. Sci. Found. Sri Lanka 42 (1).

Ratnayake, R., Karunaratne, S., Lessels, J., Yogenthiran, N., Rajapaksha, R., Gnanavelrajah, N., 2016. Digital soil mapping of organic carbon concentration in paddy growing soils of northern Sri Lanka. Geoderma Reg. 7 (2), 167–176.

Ratnayake, R., Perera, B., Rajapaksha, R., Ekanayake, E., Kumara, R., Gunaratne, H., 2017. Soil carbon sequestration and nutrient status of tropical rice based cropping systems: rice-rice, rice-soya, rice-onion and rice-tobacco in Sri Lanka. Catena 150, 17–23.

Raza, S., Miao, N., Wang, P., Ju, X., Chen, Z., Zhou, J., Kuzyakov, Y., 2020. Dramatic loss of inorganic carbon by nitrogen-induced soil acidification in Chinese croplands. Glob. Chang. Biol. 26 (6), 3738–3751.

Rentschler, T., Gries, P., Behrens, T., Bruelheide, H., Kuhn, P., Seitz, S., Shi, X., Trogisch, S., Scholten, T., Schmidt, K., 2019. Comparison of catchment scale 3D and 2.5 D modelling of soil organic carbon stocks in Jiangxi Province, PR China. PLoS One 14 (8), e0220881.

Rossel, V.R.A., Webster, R., Bui, E.N., Baldock, J.A., 2014. Baseline map of organic carbon in Australian soil to support national carbon accounting and monitoring under climate change. Glob. Chang. Biol. 20 (9), 2953–2970.

Sahrawat, K.L., 2004. Organic matter accumulation in submerged soils. Adv. Agron. 81, 170–203.

Sathischandra, H., Marambe, B., Punyawardena, R., 2014. Seasonal changes in temperature and rainfall and its relationship with the incidence of weeds and insect pests in rice (*Oryza sativa* L) cultivation in Sri Lanka. Clim. Change Environ. Sustain. 2 (2), 105–115.

Saurette, D.D., Berg, A.A., Laamrani, A., Heck, R.J., Gillespie, A.W., Voroney, P., Biswas, A., 2022. Effects of sample size and covariate resolution on field-scale predictive digital mapping of soil carbon. Geoderma 425, 116054.

Scharlemann, J.P., Tanner, E.V., Hiederer, R., Kapos, V., 2014. Global soil carbon: understanding and managing the largest terrestrial carbon pool. Carbon Manag. 5 (1), 81–91.

Shi-Hang, W., Xue-Zheng, S., Yong-Cun, Z., Weindorf, D., Dong-Sheng, Y., Sheng-Xiang, X., Man-Zhi, T., Wei-Xia, S., 2011. Regional simulation of soil organic carbon dynamics for dry farmland in East China by coupling a 1: 500 000 soil database with the century model. Pedosphere 21 (3), 277–287.

Skeen, C.J., 1994. Carbon, hydrogen, and nitrogen by a CHN elemental analyser. In: Analytical Methods Manual for the Mineral Resource Surveys Program US Geological Survey, p. 186.

Somarathna, P.D.S.N., Minasny, B., Malone, B.P., 2017. More data or a better model? Figuring out what matters most for the spatial prediction of soil carbon. Soil Sci. Soc. Am. J. 81 (6), 1413–1426.

Song, F.F., Xu, M.G., Duan, Y.H., Cai, Z.J., Wen, S.L., Chen, X.N., Shi, W.Q., Colinet, G., 2020. Spatial variability of soil properties in red soil and its implications for site-specific fertiliser management. J. Integr. Agric. 19 (9), 2313–2325.

Song, J., Gao, J., Zhang, Y., Li, F., Man, W., Liu, M., Wang, J., Li, M., Zheng, H., Yang, X., Li, C., 2022. Estimation of soil organic carbon content in coastal wetlands with measured VIS-NIR spectroscopy using optimized support vector machines and random forests. Remote Sens. 14 (17), 4372.

Sreenivas, K., Dadhwal, V., Kumar, S., Harsha, G.S., Mitran, T., Sujatha, G., Suresh, G.J. R., Fyzee, M., Ravisankar, T., 2016. Digital mapping of soil organic and inorganic carbon status in India. Geoderma 269, 160–173.

Stevens, A., Nocita, M., Toth, G., Montanarella, L., Van Wesemael, B., 2013. Prediction of soil organic carbon at the European scale by visible and near infrared reflectance spectroscopy. PLoS One 8 (6), e66409.

Sumfleth, K., Duttmann, R., 2008. Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. Ecol. Indic. 8 (5), 485–501.

Sun, Y., Zhang, L., Wang, H., 2022. The impact of sampling sites on model performances. J. Environ. Sci. 45 (3), 211–225.

Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. Geoderma 266, 98–110.

Taghizadeh-Mehrjardi, R., Schmidt, K., Amirian-Chakan, A., Rentschler, T., Zeraatpisheh, M., Sarmadian, F., Valavi, R., Davatgar, N., Behrens, T., Scholten, T., 2020. Improving the spatial prediction of soil organic carbon content in two contrasting climatic regions by stacking machine learning models and rescanning covariate space. Remote Sens. 12 (7), 1095.

Tiwari, S.K., Saha, S.K., Kumar, S., 2015. Prediction modeling and mapping of soil carbon content using artificial neural network, hyperspectral satellite data and field spectroscopy. Adv. Remote Sens. 4 (01), 63.

Tsui, C.C., Tsai, C.C., Chen, Z.S., 2013. Soil organic carbon stocks in relation to elevation gradients in volcanic ash soils of Taiwan. Geoderma 209, 119–127.

Vitharana, U.W.A., Mishra, U., Mapa, R.B., 2019. National soil organic carbon estimates can improve global estimates. Geoderma 337, 55–64.

Wadoux, A.M.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70.

Wadoux, A.M.C., Minasny, B., McBratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359.

Wadoux, A.M.C., Roman Dobarco, M., Malone, B., Minasny, B., McBratney, A.B., Searle, R., 2023. Baseline high-resolution maps of organic carbon content in Australian soils. Sci. Data 10 (1).

Wang, L., Wang, X., Kooch, Y., Song, K., Zheng, S., Wu, D., 2023. Remote estimation of soil organic carbon under different land use types in agroecosystems of eastern China. Catena 231, 107369.

Wartini, N.G., Minasny, B., Mendes, W.D.S., Demattê, J.A.M., 2020. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. Soil 6, 565–578.

Wissing, L., Kolbl, A., Hausler, W., Schad, P., Cao, Z.H., Kogel-Knabner, I., 2013. Management-induced organic carbon accumulation in paddy soils: the role of organo-mineral associations. Soil Tillage Res. 126, 60–71.

Xu, S., Wang, M., Shi, X., 2020. Hyperspectral imaging for high-resolution mapping of soil carbon fractions in intact paddy soil profiles with multivariate techniques and variable selection. Geoderma 370, 114358.

Yan, X., Zhou, H., Zhu, Q., Wang, X., Zhang, Y., Yu, X., Peng, X., 2013. Carbon sequestration efficiency in paddy soil and upland soil under long-term fertilisation in southern China. Soil Tillage Res. 130, 42–51.

Zhang, H., Wu, P., Yin, A., Yang, X., Zhang, M., Gao, C., 2017. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: a comparison of multiple linear regressions and the random forest model. Sci. Total Environ. 592, 704–713.

Zhang, Z.H., Jun, N., Liang, H., Wei, C.L., Yun, W., Liao, Y.L., Lu, Y.H., Zhou, G.P., Gao, S.J., Cao, W.D., 2022. The effects of co-utilizing green manure and rice straw on soil aggregates and soil carbon stability in a paddy soil in South China. J. Integr. Agric. 22 (5), 1529–1545.

Zheng, H., Wang, Q., Zhu, X., Li, Y., Yu, G., 2014. Hysteresis responses of evapotranspiration to meteorological factors at a diel timescale: patterns and causes. PLoS One 9 (6), e98857.