Received: 16 May 2023 | Revised: 11 August 2023 | Accepted: 28 August 2023

DOI: 10.1111/sum.12966

Soil Use and Management | WILEY

RESEARCH PAPER

Effects of spatial data resolution on the modelling and mapping of soil organic carbon content in hill country grassland landscapes

Duy X. Tran[1] | Estelle Dominati[1] | John Lowry[2] | Alec Mackay[1] | Ronaldo Vibart[1] | Diane Pearson[3] | Brian Devantier[1] | Emma Noakes[1]

[1]AgResearch Grasslands, Palmerston North, New Zealand

[2]School of People, Environment and Planning, Massey University, Palmerston North, New Zealand

[3]School of Agriculture and Environment, Massey University, Palmerston North, New Zealand

Correspondence
Duy X. Tran, AgResearch Grasslands, Palmerston North, New Zealand.
Email: duy.tran@agresearch.co.nz

Funding information
AgResearch Strategic Science Investment Fund "NZ Bioeconomy in the Digital Age" (NZBIDA)

Abstract

Limited use has been made of spatially explicit modelling of soil organic carbon (SOC) in highly complex farmed landscapes to advance current mapping efforts. This study aimed to address this gap in knowledge by evaluating the spatial prediction of SOC content in the 0–75 mm soil depth in hill country landscapes in New Zealand (NZ) using point-based training data, along with topographic covariates and Sentinel 2 spectral band ratios using an automated set of machine learning (AutoML) tools in ArcGIS. Subsequently, it also focused on quantifying the effects of spatial data resolution (i.e., 1, 8, 15, and 25 m) in terms of predicted map accuracy. Farmlets with contrasting phosphorus fertilizer and sheep grazing histories located at the Ballantrae Hill Country Research Station, NZ were selected to conduct the research. Six candidate algorithms incorporated in the AutoML tools (i.e., XGBoost, LightGBM, linear regression, decision trees, extra trees, random forest) and ensemble model were utilized to model the spatial pattern of SOC content. The results show that the ensemble model that combine predictions of various algorithms applied for 1 m data resolution enables the highest performance and accuracy (i.e., $R^2 = .76$, RMSE = 0.66%). Among the predictive variables used in the model, slope, wetness, and topographic position indices were found to be the most important topographical features that explain SOC patterns in the study area. Inclusion of spectral indices derived from remote sensing, including surface soil moisture and clay minerals ratio, made further improvement to the SOC content prediction. The study reveals that a decrease in the resolution of the geospatial data does not substantively affect the mean SOC content estimation of a farm-scale modelling. However, using coarser resolution data reduces the ability of the model to predict changes in the spatial pattern of SOC content across a hill country grassland landscape.

KEYWORDS
automated machine learning, grassland, hill country, soil organic carbon, spatial data resolution

1 | INTRODUCTION

Soil organic carbon (SOC) is an important component of soil health playing a key role in supporting soil structure through plant growth (Bronick & Lal, 2005; De Graaff et al., 2006; Herrick & Wander, 2018; Lal, 2004; Van Cleve & Powers, 1995). Improvements in soil health achieved by increasing SOC content, when depleted, can increase agricultural productivity, contribute to carbon storage, and enhance the resilience of agricultural systems to extreme weather events (Hontoria et al., 2016; Jayaraman et al., 2021; Manns et al., 2016; Pramanick et al., 2021; Stevenson et al., 2016). The storage of carbon in the soil is also an important part of the global carbon cycle and therefore a key component of any global climate change mitigation strategy (Khan et al., 2021; Ryals et al., 2015). Increases in SOC stocks offer an option for offsetting agricultural greenhouse gas (GHG) emissions (Huang et al., 2011; Paustian et al., 2016; Smith, 2016). Studying the spatial and temporal patterns of SOC content and stocks is essential to provide information for managing soil health and mitigating the effects of climate change.

Modelling and mapping SOC content has been relatively well studied at regional and national scales (Aitkenhead & Coull, 2016; Kaczynski et al., 2017; Minasny et al., 2013; Whitehead et al., 2021), whereas limited use has been made of spatial modelling at farm level to advance SOC mapping options. The ability to quantify the spatially explicit pattern of SOC in a complex farmed landscape provides a picture of SOC content across space, and this information can be utilized to prioritize target areas for soil carbon management towards maintaining and enhancing soil health. Some studies have applied straightforward interpolation techniques or down-scaling approaches to produce farm level SOC map information (Dewage et al., 2020; Malone et al., 2017; Van Huynh et al., 2022). However, these methods are not very effective in describing the pattern of SOC in areas with heterogenous landscapes.

Soil organic carbon can vary significantly across space depending on physical settings that include a wider variety of topographical features and land use and management practices (Doetterl et al., 2015; Mackay et al., 2018; Román-Sánchez et al., 2018; Senthilkumar et al., 2009). Among various types of land use land cover (LULC) classes and ecosystems, grasslands in New Zealand (NZ) store a significant amount of the national sequestered organic carbon stocks (Stockmann et al., 2013). A large proportion of grassland in NZ is in hill country grasslands, which is defined as "land with slopes above 15° and located below an altitude of 1000 m above sea level" (Cameron, 2016). This landscape is highly heterogenous, with a diversity of soils, slopes, aspects, and altitude can be found over short distances (Cumberland, 1941). As such, the spatial pattern of SOC content and stock patterns in this landscape can be highly variable (Hedley et al., 2015).

At the farm level, SOC distribution may differ significantly over short distances, even within a land unit or a single paddock (Franzluebbers et al., 1999; Qiu et al., 2011). The traditional method of reporting on SOC content involves in-situ point measurement of selected locations in the landscapes. This approach provides the most accurate farm-scale SOC information, however, can be both time-consuming and expensive (Acharya et al., 2022; Whitehead et al., 2012). This limits the opportunity to obtain enough samples to be able to describe changes in SOC patterns in complex grassland landscapes.

The development of geospatial technologies, including remote sensing and advanced spatial statistics and machine learning and/or deep learning, enable point source SOC information to be used to add a spatial dimension (Angelopoulou et al., 2019; Emadi et al., 2020; Minasny et al., 2013; Paul et al., 2020). This has the potential to produce a more accurate picture of the SOC pattern across a landscape than using just traditional regression and geostatistical methods. Using remotely sensed data in combination with in-situ point source data to estimate SOC over a large area creates the opportunity to add value to the investment in the in-situ point source approach, by providing timely and low-cost data with continuous surface coverage and repeated capturing over time (Xu et al., 2017). Machine learning and deep learning algorithms utilizing advanced statistical and data-mining methods (Keskin et al., 2019; Shen et al., 2022; Wang, Guan, et al., 2022; Zhou et al., 2020) add another capability to identify patterns and relationships which have the potential to add further improvement to SOC models (Odebiri et al., 2021).

Despite these successes and contributions, gaps in the literature remains. It is recognized that SOC mapping studies carried from global to catchment scales (Chartin et al., 2017; Gray et al., 2022; Heil et al., 2022; Holmquist et al., 2018; Sanderman et al., 2018) still offer limited

information on SOC patterns and dynamics suitable for forming land use planning and management practices at the farm level. This can be explained in part by the fact that the spatial resolution used at national and regional scales does not capture the micro topographical and biophysical pattern in highly complex landscapes, that can vary widely in landforms, slope, soils and vegetative cover types (Burkitt & Bretherton, 2022; Manderson & Palmer, 2006). Lemercier et al. (2022) found that the ability to predict soil properties, including SOC content, well increased from global to national to regional digital soil mapping products, but none of the products tested were able to predict satisfactorily at the most local scale (1:25,000). At the farm scale, understanding the spatial resolution required that best utilizes in situ point source data and maximizes the value advanced spatial statistics and machine learning offers in developing spatial mapping products, would be invaluable. This is because the quality of information obtained from a spatial model has a direct impact on the development and efficiency of on-farm options to offset agricultural GHG emissions (e.g., Reisinger et al., 2017).

The aims of this study were to (1) evaluate the potential value of geospatial technologies, including remote sensing and advanced spatial statistics and machine learning and/or deep learning, combined with point source SOC data to map SOC content in the 0–75 mm soil depth and (2) quantify the effect the resolution geospatial data layers has on the spatial accuracy and pattern of the SOC content predictions in a grassland hill country landscape. We used farmlets with contrasting phosphorus fertilizer and associated sheep management histories in landscapes with a complex mix of soils, slopes, and aspects, located on the Ballantrae Hill Country Research Station (Mackay et al., 2021) as a case study.

2 | MATERIALS AND METHODS

2.1 | Study area

The study was conducted at AgResearch's Ballantrae Hill Country Research Station, located in Southern Hawke's Bay, New Zealand (Figure 1). The study site covers an area of 32.8 ha of pasture, with altitude ranging between 196 and 359 m above sea level (Figure 1b). The study area is typical of hill and steep land landscapes with a prevalence of slope classes greater than 25° (58.5%) and between 15 and 25° (21.8%) (Figure 2a). It forms part of a long-term (1975–2023) phosphorus (P) fertilizer and sheep grazing experiment, adding the influence of widely different P fertilizer and sheep grazing practices to the management of these landscapes (Mackay et al., 2021). The landscapes are highly heterogenous with different soils, landforms, and topographical features (Mackay et al., 2021). Soil groups covered in the study area vary from Andic Dystrochrept (Ngamoka silt loam soils, i.e., Ng, NgR), Typic Dystrochrept (Ngamoka hill and steepland soils, i.e., NgHA, NgHB, NgSR, NgSRA; Whetukura hill soils, i.e., WtHR, WtHB; and Mangamahu steepland soil, i.e., MmS), Andic Haplumbrept (Makara steepland, i.e., MkS), through to an Aquic Hapludand (Raumati silt loam soil, i.e., Rt) (Figure 2k). The four farmlets in the study area have very different P fertilizer and sheep management histories dating back to 1975. One farmlet has a history of receiving 125 kg single superphosphate $ha^{-1}year^{-1}$ (LF), one has a history of receiving 375 kg single superphosphate $ha^{-1}year^{-1}$ (HF), and two farmlets both with a history of no P fertilizer input (NF) since 1980 have been combine as part of this study (Figure 2l). The annual nominal mean stocking rates are 6.0, 10.6, and 16.1 SU ha^{-1} for the NF, LF, and HF farmlets, respectively (Hoogendoorn et al., 2016).

2.2 | Data used

2.2.1 | Soil carbon data

In this study, we used SOC data collected from the field survey previously reported by Mackay et al. (2021). The locations of the sites for assessing SOC content in the topsoil (0–75 mm) (Figure 1a) were based on three criteria: slope (low, medium, and high), aspect (Northeast, Northwest, and Southwest), and previous P fertilizer and sheep grazing histories (NF, LF, and HF). Soil carbon data used in this study were collected in September 2021. Soil samples to a depth of 0–75 mm were taken from all 72 sites. For a detailed description of soil sampling process, see Mackay et al. (2021).

2.2.2 | Soil carbon predictors

This study utilized remotely sensed, land and environmental data variables for mapping the SOC content (see Table 1). The DEM data with original spatial resolutions of 1, 8, 15, and 25 m were used to derive a wide range of topographical variables. These are slope, aspect, Topographical Wetness Index (TWI), Terrain Ruggedness Index (TRI), and Topographic Position Index (TPI) at each resolution (Figures 2a–d). For detailed information on the definition and method for calculating these topographical indices, see Chowdhury (2023). ArcGIS topography/surface analysis tools were employed to

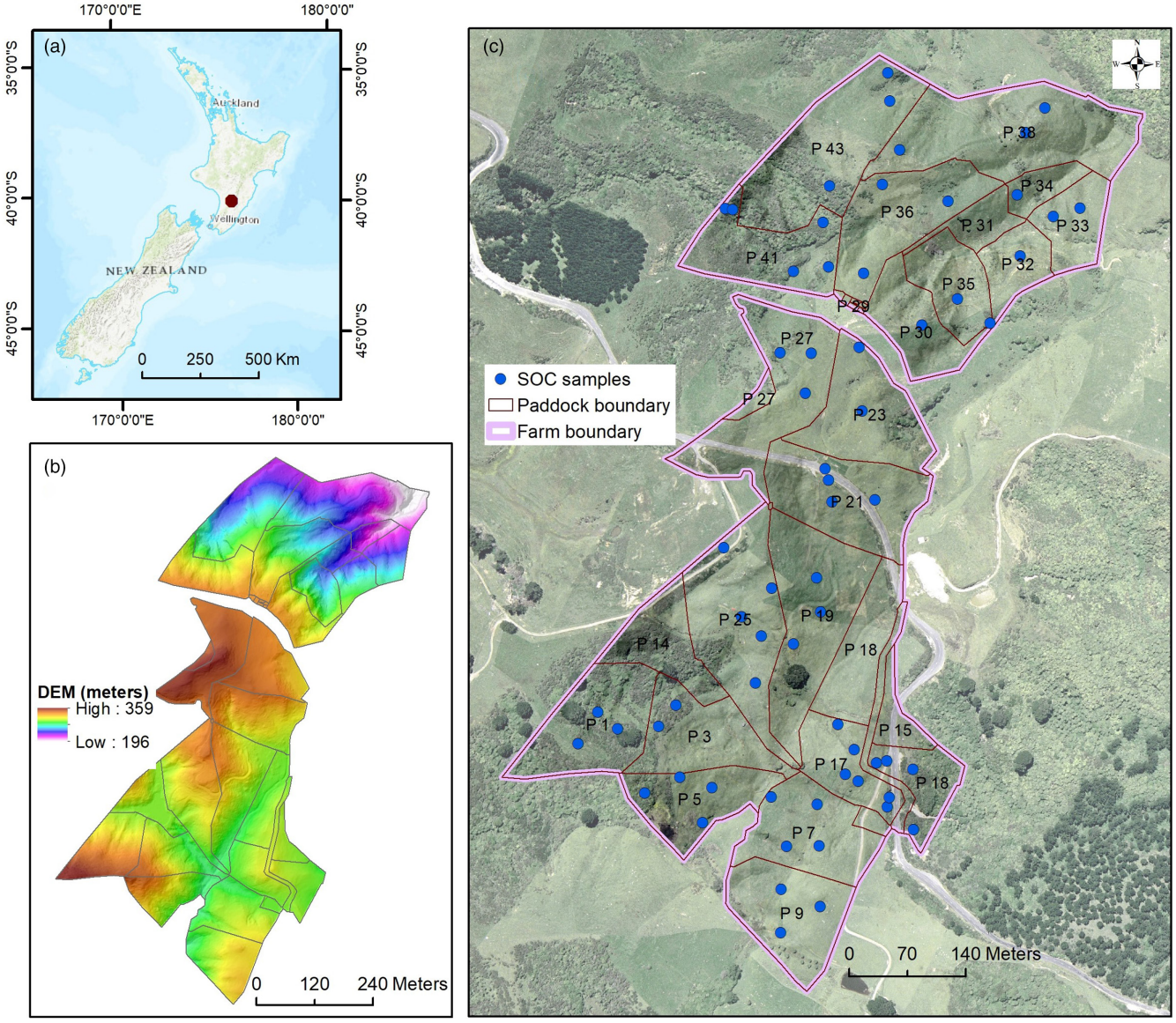


FIGURE 1 (a) Location of the study site in New Zealand, (b) digital elevation model (DEM), and (c) paddock numbers and soil sampling locations.

calculate topographic indices from DEM data. The landforms were classified following the approach proposed by Schmidt and Hewitt (2004). Types of landforms are hilltops, ridges, hill slope, valley, and flat areas. Hill slope includes concave surfaces (hollow/drainage) and convex surfaces (spur).

Cloud-free surface reflectance Sentinel 2 data covered the study site and were acquired through the climate engine web portal (https://app.climateengine.com/). The mean values of images acquired in September 2021 (i.e., the month SOC content was sampled) for near infrared, red, short-wave infrared, and blue bands were downloaded and used for calculating surface biophysical indices. These indices are Normalized Difference Vegetation Index (NDVI), Normalized Difference Moisture Index (NDMI), Bare soil index (BSI), and Clay Minerals Ratio (CMR) (Figures 2f–i). For the classification of the pattern of pasture clusters in the study area (Figure 2j), monthly time-series NDVI images from 2016 to 2021 and time-series clustering tool in ArcGIS Pro (version

FIGURE 2 Spatial pattern of SOC content predictors in the study area: (a) slope, (b) aspect, (c) topographical wetness index (TWI), (d) terrain ruggedness index (TRI), (e) topographic position index (TPI), (f) normalized difference vegetation index (NDVI), (g) normalized difference moisture index (NDMI), (h) bare soil index (BSI), (i) clay minerals ratio (CMR), (j) pasture clusters, (k) soil group, and (l) management practices.

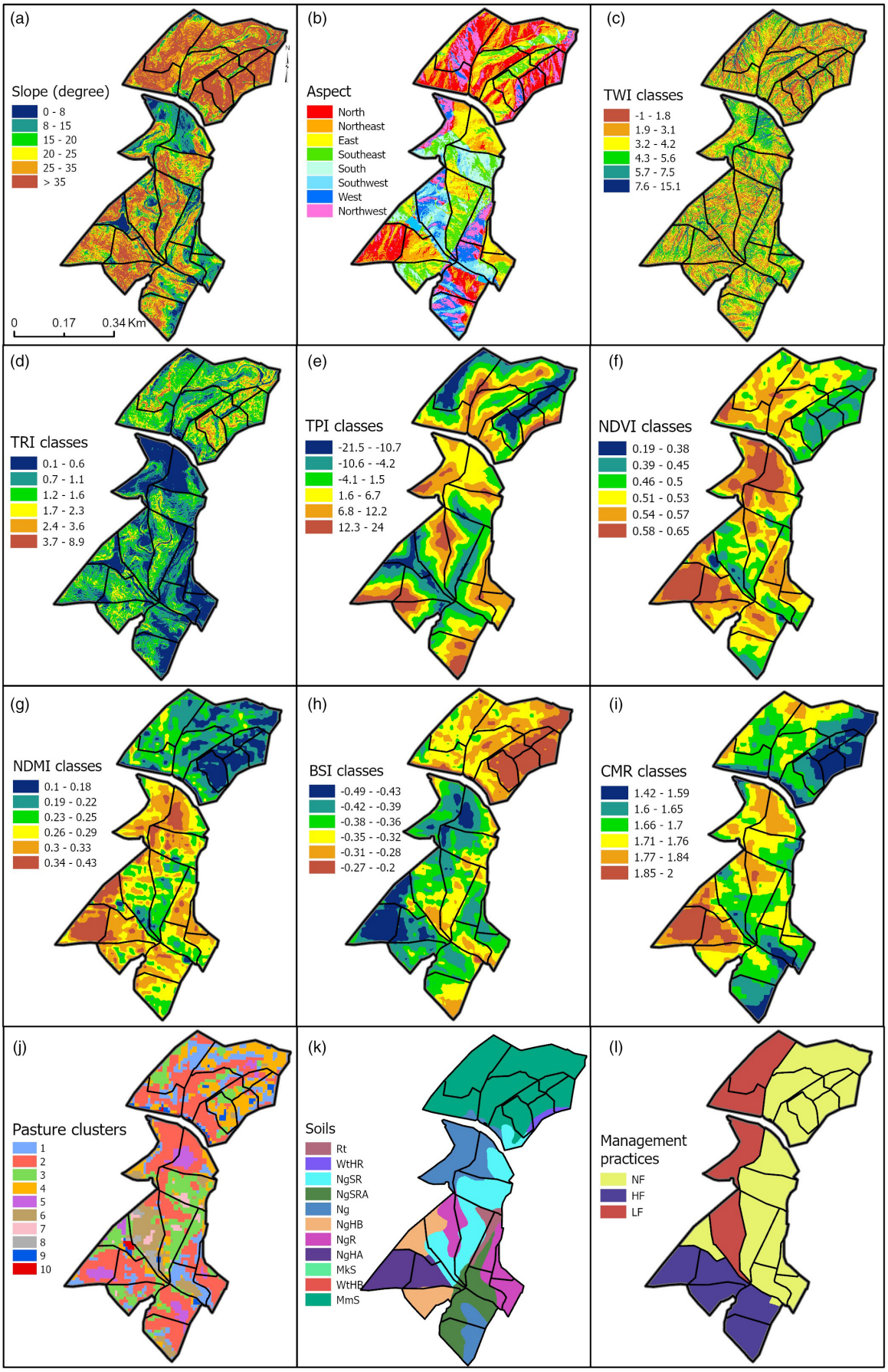


TABLE 1 Prediction variables used for landscape modelling of soil organic carbon content.

Data	Description	Source/reference
Topographical data		
DEM	Digital elevation model	1, 8, 15, and 25 m DEM were acquired from Ballantrae, LINZ, University of Otago, NZLRI, respectively
Slope	Calculated from DEM	
Aspect	Calculated from DEM	
Topographical wetness index	Calculated from DEM	
Terrain ruggedness index	Calculated from DEM	
Topographic position index	Calculated from DEM	
Sentinel 2 data		
Normalized difference vegetation index	$\frac{NIR-R}{NIR+R}$	Rouse et al. (1974)
Normalized difference moisture index	$\frac{NIR-SWIR1}{NIR+SWIR1}$	Klemas and Smart (1983)
Bare soil index	$\frac{(SWIR1+R)-(NIR+B)}{(SWIR1+R)+(SWIR1+B)} \times 100 + 100$	Rikimaru et al. (2002)
Clay minerals ratio	$\frac{SWIR1}{SWIR2}$	Carranza and Hale (2002)
Pasture clusters	Time-series NDVI data	
Soils	Soil groups	Ballantrae datasets
Land management practices	Super phosphate fertilizer inputs and sheep grazing histories (including HF, LF, and NF farmlets)	Ballantrae datasets

Abbreviations: HF, high P fertilizer input; LF, low P fertilizer input; LINZ, Land Information New Zealand; NF, no P fertilizer input; NIR, R, SWIR, B refer to the near infrared, red, short-wave infrared, and blue bands of Sentinel 2; NZLRI, New Zealand Land Resource Inventory.

3.0) were used. This tool uses clustering algorithms (e.g., k-medoids) to classify time-series NDVI data to different clusters based on their similarity (Barazzetti et al., 2022). Because the spatial resolution of Sentinel 2 data is 10 m, these images were resampled, using Empirical Bayes Kriging interpolation, to be consistent with the DEM resolutions. Soil data and phosphorus management practice histories were obtained from the Ballantrae database (Mackay et al., 2021). These data layers originally stored in shapefile (vector) format were converted to raster datasets using the data conversion tool in ArcGIS Pro (Figure 2k,l). Four levels of spatial resolution for output raster data were assigned (1, 8, 15, and 25 m) and the extent of these data were adjusted to match the cell alignment of the topographical and spectral data.

Figure 3 and Table 2 illustrate changes in spatial patterns and proportion of slope and aspect classes resulting from different resolutions of DEM data. A change in data resolution results in a large difference in the area and distribution of the lowest and highest slope classes. For instance, areas with slope values <15° increase 2–3 times, from 6.5 ha (19.8%) to 16.8 ha (51.2%), 15.5 ha (47.4%), and 12.3 ha (37.3%) for 1, 8, 15, and 25 m, respectively. Steep slope areas that have slope values >35° decrease substantially concurrent with a change from fine to coarse resolutions. For example, the 1 m DEM data has 10.6 ha (32.4% of the total area) of slopes >35°, which is much higher than that of the 8, 15, and 25 m data, which are 0.1 ha (0.4%), 0.2 ha (0.7%), and 2 ha (6.1%), respectively.

2.3 | Automated machine learning

We used automated machine learning (AutoML) available in ArcGIS Pro version 2.9 to predict SOC content. This software provides a user-friendly and effective tool for spatially explicit machine learning, making it ideal for spatial modelling of SOC content. AutoML incorporates major steps of the machine learning process into an integrated and iterated framework (Zöller & Huber, 2021). These steps include data pre-processing and engineering, model training, hyperparameter tuning, and model evaluation (Figure 4). This is an iterative process, and the optimal model is often only reached after multiple iterations and experiments. In practice, identifying the model that best fits the data takes time, effort, and expertise in the machine learning process. The AutoML tools automate this workflow and identify the best algorithm with the best set of hyperparameters that fit the data. The tool supports an automated machine learning application through two steps: (1) Train using AutoML and (2) Predict using AutoML. The machine learning algorithms that are used in this tool include extreme gradient boosting (XGBoost) algorithm (Chen & Guestrin, 2016), light gradient-boosting

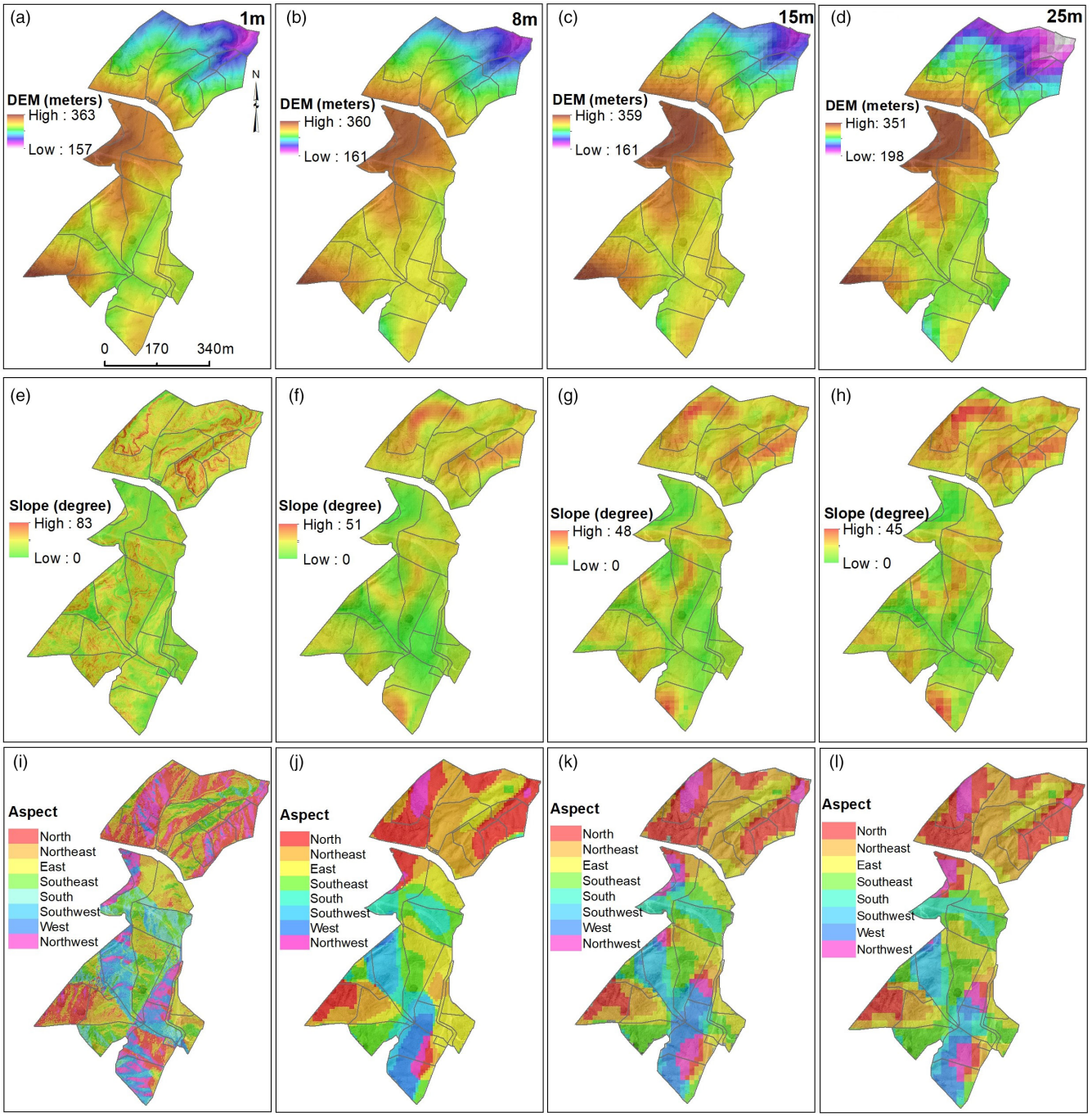


FIGURE 3 The digital elevation model (DEM) at a resolution of (a) 1 m, (b) 8 m, (c) 15 m, and (d) 25 m and the corresponding slope (e–h), and aspect (i–l), respectively, derived at each of the four spatial resolutions.

machine (LightGBM) algorithm (Ke et al., 2017), decision trees (Quinlan, 2014), linear regression, extra trees (Geurts et al., 2006), random forest (Liaw & Wiener, 2002), and ensemble model (Yang, 2016). Ensemble approaches combine the results of various models by applying different weights to the outputs. Bagging and boosting are common strategies used for creating an ensemble which combine the outputs from the models belonging to the same algorithms. Techniques, for example, stacking, use the outputs from different algorithms (Caruana et al., 2004). The Auto-ML tools in ArcGIS utilize both types of ensembles. This is an automated process in which the tools combine outputs from models using the same algorithm or diverse models to achieve the best results. See ESRI (2022) for detailed description of these machine learning algorithms and the ensemble process.

For the selection of soil samples to train the machine learning models, we cleaned and de-noized the data which may significantly affect the model's performance (Soh & Singh, 2020). This included removing points that

TABLE 2 Influence of the changes in the resolution of DEM on the area (ha) and percentage (%) of each slope and aspect class in the study area.

	1 m		8 m		15 m		25 m	
Resolution	**Area**	**%**	**Area**	**%**	**Area**	**%**	**Area**	**%**
Slope								
0–8°	2.0	6.1	8.4	25.7	7.4	22.6	5.9	17.8
8–15°	4.5	13.7	8.4	25.5	8.1	24.8	6.4	19.5
15–20°	3.4	10.5	9.1	27.7	8.8	27.0	4.4	13.4
20–25°	3.7	11.3	4.1	12.5	5.3	16.2	6.5	19.7
25–35°	8.6	26.1	2.7	8.3	2.9	8.7	7.7	23.5
>35°	10.6	32.4	0.1	0.4	0.2	0.7	2.0	6.1
Aspect								
North	6.3	19.1	6.8	20.5	6.4	19.5	6.8	20.7
Northeast	5.5	16.8	7.3	22.3	7.2	22.1	7.0	21.3
East	5.6	16.9	6.7	20.3	5.5	16.9	5.5	16.8
Southeast	4.1	12.6	4.0	12.1	4.4	13.3	5.0	15.2
South	2.8	8.6	2.8	8.5	2.3	7.1	2.8	8.6
Southwest	2.0	6.2	2.4	7.3	2.2	6.6	2.1	6.4
West	2.4	7.2	1.8	5.4	2.3	7.1	1.5	4.6
Northwest	4.2	12.8	1.1	3.5	2.4	7.3	2.1	6.3

Source: Calculated from DEM data.

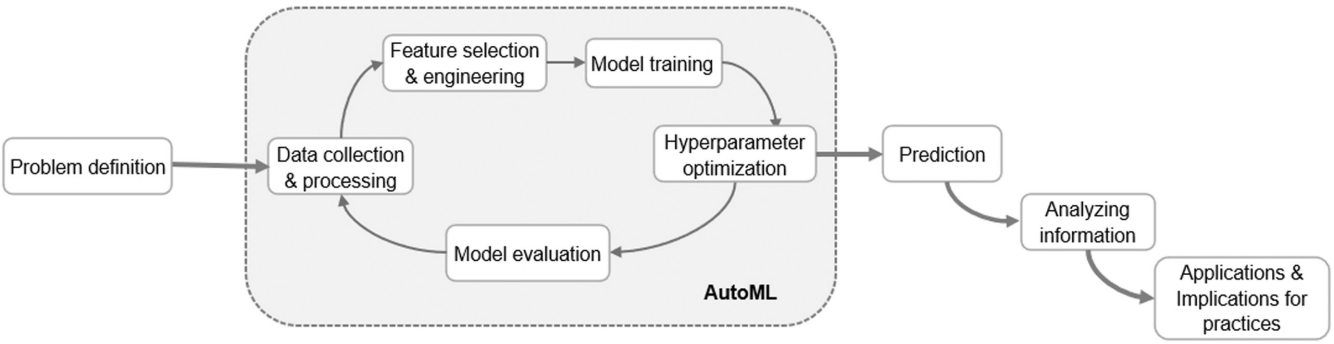


FIGURE 4 Typical steps of an application using machine learning (adapted from ESRI, 2022). The steps in the grey box can be automated by AutoML, data collection, and processing step is done by utilizing GIS and remote sensing, and the analysis of model outcomes is carried out by GIS tools and methods.

are spatial outliers and inconsistent with the description of sampling sites and data derived from a LiDAR DEM. For the spatial outlier's detection, a local outlier is identified as a soil sample point that is further away from its neighbours than would be expected by the density of points in the study area. In terms of the inconsistency between the field and GIS data, if a sample site is labelled as low slope but the slope value derived from the DEM falls within a high slope, the sample from this site was removed. In total, eight samples or sample sites were removed. This ensures that the prediction algorithms used in the model can be effectively trained and learned from the training sample sites. After this stage, 47 sample sites were selected to train the SOC content prediction models in the study area. A set of 17 samples that were not involved in the training process (~25% of total samples) were then used as a validation set to assess the prediction performance independent of the training data. The purpose of the independent sample was to reveal how effective the AutoML model for predicting SOC in nearby landscapes.

AutoML tool provides several indicators for evaluating prediction performance. These are the coefficient of determination, i.e., root square (R^2), root mean squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), and mean absolute percentage

error (MAPE) values. Of those indicators, AutoML used RMSE to optimize model output (i.e., prediction performance). This is a common indicator employed in model assessment studies, to evaluate the SOC content prediction performance among machine learning algorithms (Christie & Neill, 2021). Root mean squared error is a measure of the differences between values predicted by a model, or an estimator and the values observed (Willmott, 1981). It indicates how far predictions fall from measured true values using Euclidean distance. In addition to RMSE, we used R^2 to compare the prediction performance of different models. The best model for SOC content prediction is one with the lowest RMSE and highest R^2 values. The output from the best models was used for further analysis of the spatial variation of SOC predictions and the effects of the spatial data resolution of input data for spatially predicting SOC content. In addition to the model performance information, we used permutation feature importance measure to understand the importance of each explanatory variable used in SOC content prediction model. This approach measures feature importance by observing changes in model performance (i.e., a decrease in prediction performance) when each predictor variable is randomly shuffled (Gómez-Ramírez et al., 2020). AutoML tools produce a table containing permutation feature importance for the best SOC content prediction model. The table shows the variable importance scores, in which higher values represent the more important SOC content predictors.

2.4 | Spatial analysis of SOC content

Spatial clusters of high, low, and non-clustered SOC content were identified using hot spot analysis by Getis-Ord Gi* method (Getis & Ord, 1992). The predicted SOC content data obtained from the best prediction model (i.e., the most accurate SOC content prediction) were used as the input for hot spot analysis. Getis-Ord Gi* is a method to measure spatial clustering of values, in this case SOC content. Hot spots are areas that have high SOC content and are surrounded by high SOC content values. Cold spots are low SOC content surrounded by low SOC content values. Non-clustered areas may have high or low values but are not surrounded by like values. When a cluster of SOC content values is larger or smaller (hot or low) than the global mean, and the difference is greater than would be expected by random chance, the cluster is identified as statistically significant within intervals of 90%, 95%, and 99% confidence. This provides a better understanding of the spatial distribution of SOC content in the study area.

Because higher spatial resolution data would be expected to better capture patterns of topography (Maleki et al., 2020; Vaze et al., 2010), we used 1 m spatial resolution data as a benchmark from which to evaluate the effect of different resolutions (i.e., 8, 15, and 25 m) on predicted SOC content. The deviations were computed using the following equation:

$$\mathrm{SOC}_{\mathrm{difference}} = \frac{\mathrm{SOC}_{\mathrm{coarse}} - \mathrm{SOC}_{1\ \mathrm{m}}}{\mathrm{SOC}_{1\ \mathrm{m}}} \times 100 \qquad (1)$$

where $\mathrm{SOC}_{\mathrm{difference}}$ is the percentage difference (i.e., percentage increase or decrease) in SOC content value between two maps at different spatial resolutions and $\mathrm{SOC}_{\mathrm{coarse}}$ is the SOC content value associated with the comparator resolution (i.e., 8, 15, or 25 m). The equation used therefore provides a quantitative measure of how resulting SOC maps differ from the benchmark 1 m resolution model. The change in SOC prediction is represented as a percentage. Negative values (i.e., a percentage decrease) mean that predicted SOC content using coarser data is lower than ones using the 1 m data (i.e., an underestimation), whereas positive values (i.e., a percentage increase) indicate that predicted SOC content using coarse data is higher than that obtained using the 1 m data (i.e., an over-estimation).

3 | RESULTS

3.1 | Model performance for SOC content prediction

Information on the model's ability to predict SOC content with 1, 8, 15, and 25 m data resolutions and candidate machine learning algorithms (Table 3) demonstrates that the ensemble model combining predictions from several algorithms outperformed individual models (i.e., the ensemble model had the lowest RMSE and highest R^2 values). In terms of resolution, models using finer resolution data to predict SOC content achieved better performance. Among individual machine learning algorithms, Xgboost and random forest often had superior efficiency than the others in SOC content prediction.

The 1 m data resolution had the closest fit with the field data, shown by a mean R^2 of .76 and mean RMSE of 0.66 (Figure 5). Soil organic carbon content prediction using coarser resolution datasets (i.e., 8–25 m) had significantly lower R^2 and higher RMSE values, ranging from .34 to .51 and from 0.94 to 1.09, respectively. Results also demonstrate that the SOC content prediction tends to be reduced with a decrease in data resolution

TABLE 3 Model's performance in SOC content prediction using different data resolutions.

ML algorithms	1 m		8 m		15 m		25 m	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
XgBoost	0.68	.74	0.98	.46	1.10	.32	1.15	.26
Linear	0.98	.48	1.10	.31	1.28	.09	1.18	.23
LightGBM	0.78	.67	1.04	.39	1.14	.28	1.24	.14
Decision tree	0.77	.67	0.99	.45	1.19	.21	1.24	.14
Extra tree	0.87	.58	1.06	.37	1.17	.24	1.16	.25
Random forest	0.73	.71	1.03	.40	1.14	.27	1.10	.32
Ensemble[a]	0.66	.76	0.94	.51	1.08	.35	1.09	.34

Abbreviations: R^2, the coefficient of determination, i.e., root square; RMSE, root mean square error.

[a]Ensemble model structure: 1 m (XgBoost and Random Forest), 8 m (Decision Tree, Random Forest, and Extra Tree), 15 m (XgBoost, Random Forest, and LightGBM), 25 m (XgBoost and Random Forest).

(i.e., the R^2 decreased and the RMSE increased when the data resolution decreased). Results from the validation data set also demonstrated similar trends, in which the R^2 values were .62, .44, .41, .20, and the RMSE were 0.75, 0.91, 0.93, 1.09 for the 1, 8, 15, and 25 m data resolutions, respectively (Figure 5). Given that it had the best prediction performance, the ensemble model was used to predict and map the spatial SOC content in the study area and results obtained from using this model were used for further analysis.

Slope is the topographical variable contributed most to the SOC content prediction, indicating by the highest variable importance value (Table 4). Topographical wetness and position indices were also important variables that contributed substantially to the prediction of SOC content. Spectral indices including soil moisture (NDMI) and clay minerals ratio (CMR) were the top surface biophysical variables that substantially impacted the SOC prediction. In the study area, soils, pasture clusters, aspect, and P fertilizer and associated sheep management history were found to be less important than the topographies and surface biophysical variables in predicting the SOC content.

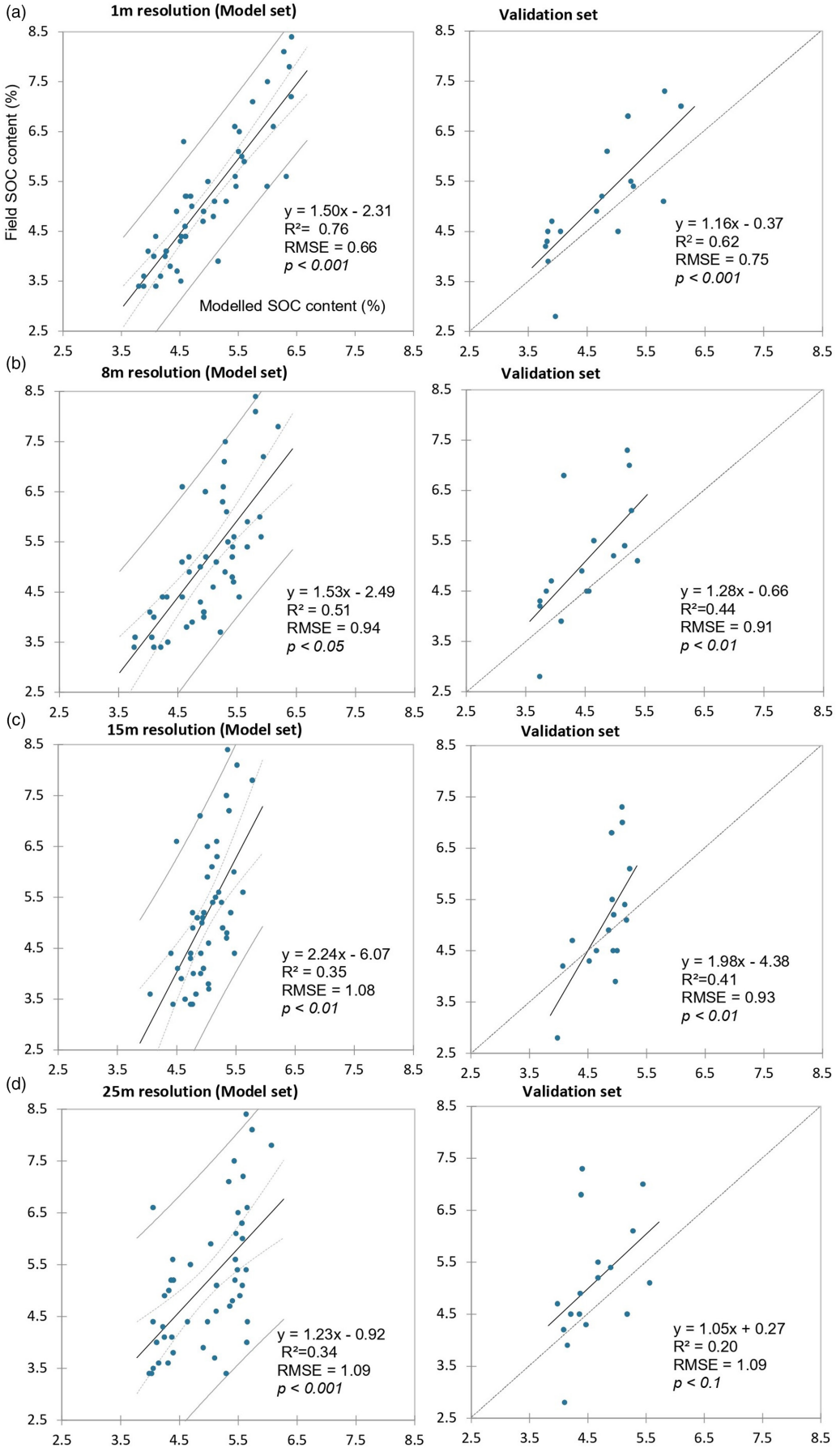
3.2 | Change in the pattern of predicted SOC content using different spatial resolutions

The mean SOC content for the 0–75 mm soil depth modelled using the 1, 8, 15, and 25 m data were 5.2%, 5.1%, 4.9%, and 4.8%, respectively. Using the spatial data at 1 m resolution (Figure 7a) resulted in the most accurate prediction, showing that the mean SOC content calculated from the model aligned with the mean SOC value of 5.2% from the field measurements. Not only was the mean value close to the field value, the ensemble model with the 1 m resolution was able to better quantify the spatial pattern of SOC distribution, which is highly heterogeneous across the study area. For example, the lowest SOC content level (SOC content <4%) that appeared in the predominantly high slope areas in the northern part of the study site at the 1 m resolution (Figure 6a), that aligns very closely with the field observations, no longer exists at the 25 m resolution (Figure 6d). Also, the high SOC content zone (SOC content >6%) seen in several areas in the low and medium slope areas in the southern region of the study area (Figure 6a) are not observed in the 15 and 25 m resolution data (Figures 6b–d).

Table 5 provides additional information about the impact of the resolution of the spatial data on SOC content calculated by the model. A substantial change is shown in the area and proportion of SOC content resulting from the change in data resolutions. For instance, at 1 m resolution, the SOC content of 4.0%–4.5% occupied the largest area (8.7 ha, ca. 26.5%). However, at 8, 15 and 25 m resolutions, the SOC contents that occupy the largest area are higher ranged from 5.0% to 5.5%, 4.5% to 5.0% and 5.0% to 5.5%, respectively.

The mean SOC content values calculated using the ensemble model with the data of varying resolutions at the paddock level showed that SOC varies between paddocks and that variation is also affected by the spatial data resolution. For example, paddocks with the highest mean SOC

FIGURE 5 The SOC content obtained from the ensemble models using (a) 1 m, (b) 8 m, (c) 15 m, and (d) 25 m resolutions data against SOC content in the 0–75 mm soil depth from the 47 field sampling sites (model set), representing the linear relationship between field and predicted SOC (black lines), bounded by 95% confidence (dashed lines), and 95% prediction intervals (grey lines). The remaining 17 field sampling sites were used as the validation set and assessed against a 1:1 line (dash lines) in predicted versus measured graph (solid lines).



content (e.g., P15, P23, and P27) were about 1.5 times higher than those with the lowest value (e.g., P33, P34, and P35). In addition, the mean SOC content value for any given paddock was up to 9% lower (e.g., P3) or 11% higher (e.g., P23) when the data resolution decreased from 1 to 25 m.

The results obtained from ensemble model were utilized to evaluate the effects of changing data resolution on mapping SOC content. The spatial pattern and distribution of the percentage change in SOC content resulting from changes in data resolution, calculated using Equation 1, is presented in Figure 7. Compared with the SOC content modelled at a 1 m resolution, the use of 8, 15, and 25 m resolution data resulted in significantly lower SOC content levels in the flat areas and valley floors, while higher levels were observed in steeper areas (e.g., hill slopes) located in the north and southwest of the farm. In other words, the prediction values were substantially underestimated in flat areas (i.e., low slope) and overestimated in steep slopes.

Figure 8 presents the mean percent change in SOC content using 8, 15, and 25 m resolution data to 1 m modelled data for slope, landform, and aspect. By slope class, the mean value of modelled SOC content using coarser resolution data decreased substantially (i.e., underestimation) in areas with slope <15°. The percentage change in SOC prediction in this case was about 9.5%–11.1% lower than the result from 1 m data. In contrast, using coarser resolution data on the prediction of SOC content in slopes >25° (i.e., steep-lands) revealed an

TABLE 4 Top five variables contributing to the final SOC content prediction model (i.e., ensemble model).

SOC content predictors	Variable importance
Slope	0.678
Normalized difference moisture index	0.243
Topographical wetness index	0.174
Clay minerals ratio	0.069
Topographic position index	0.010

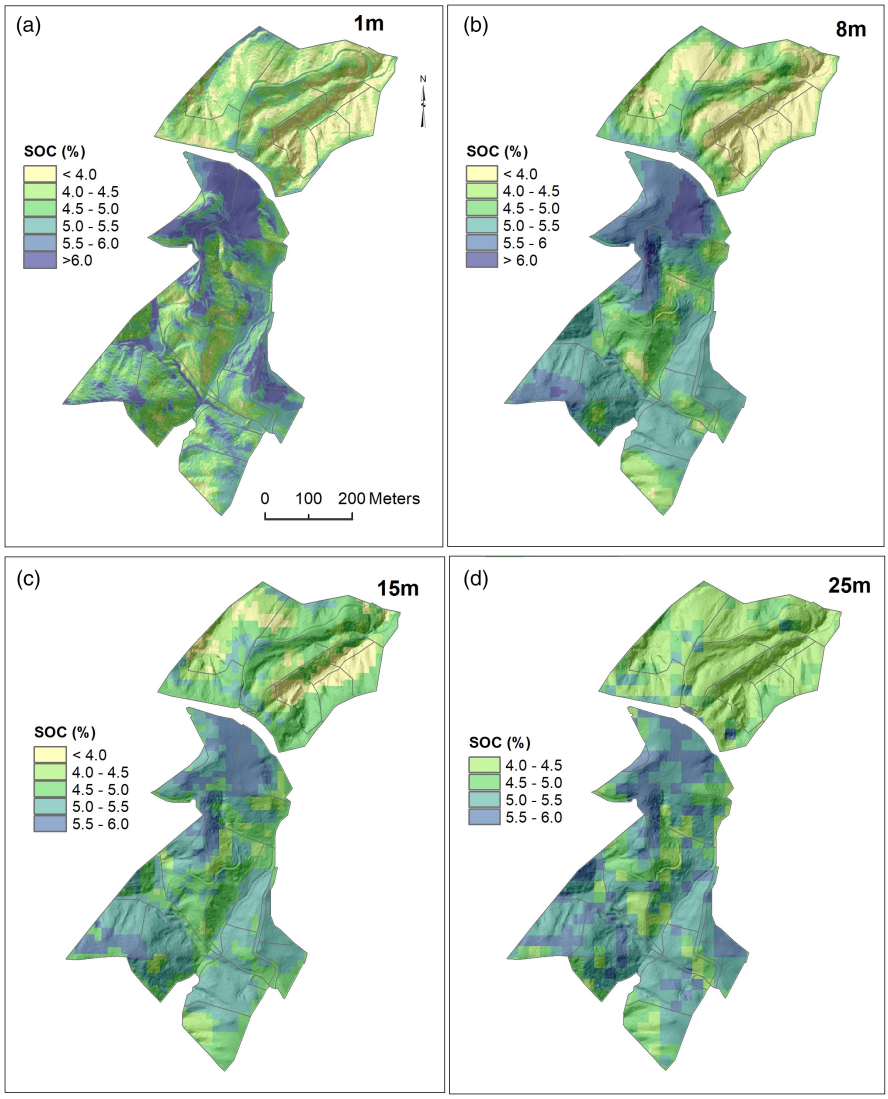


FIGURE 6 Spatial pattern of SOC content generated by the ensemble model using (a) 1 m, (b) 8 m, (c) 15 m, and (d) 25 m resolution data.

TABLE 5 The effects of changing from 1, 8, 15, and 25 m data on the area (ha) and percentage (%) of the soil organic carbon (SOC) content calculated using the ensemble models in the study area.

	1 m		8 m		15 m		25 m	
SOC content	**Area**	**%**	**Area**	**%**	**Area**	**%**	**Area**	**%**
<4	5.8	17.8	6.9	21.0	2.3	6.9	n/a	n/a
4.0–4.5	8.7	26.5	4.5	13.6	4.2	12.7	11.0	33.5
4.5–5.0	6.8	20.8	5.9	17.9	11.9	36.1	6.0	18.1
5.0–5.5	4.1	12.6	10.2	30.9	11.3	34.3	10.7	32.7
5.5–6.0	3.5	10.7	4.1	12.5	3.3	10.0	5.1	15.7
>6.0	3.8	11.7	1.3	4.0	n/a	n/a	n/a	n/a

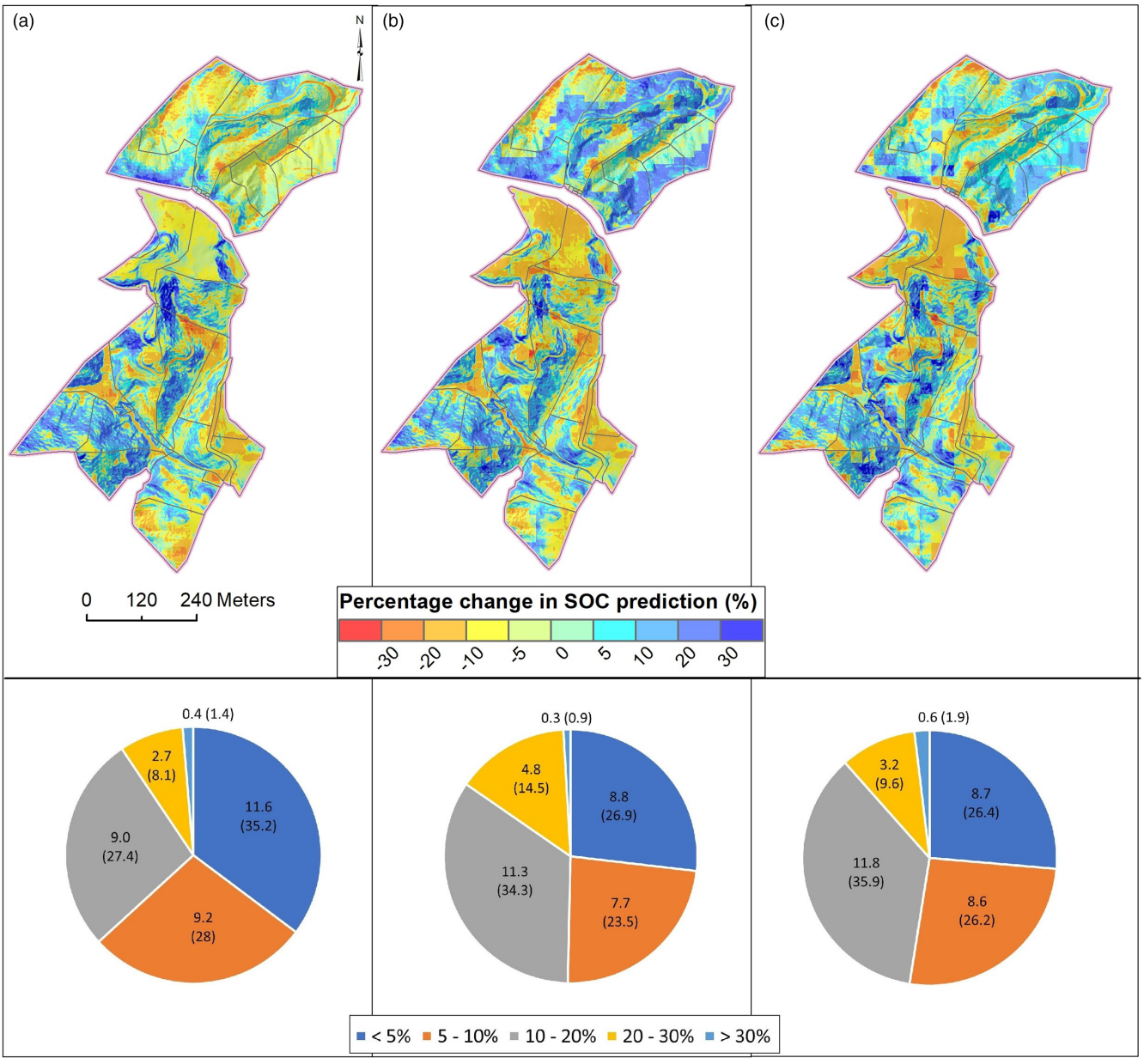


FIGURE 7 Percentage changes in the spatial pattern of the SOC content derived from (a) 8 m, (b) 15 m, and (c) 25 m when compared with the 1 m data resolutions. The pie charts present area and percentage (values in bracket) of SOC difference.

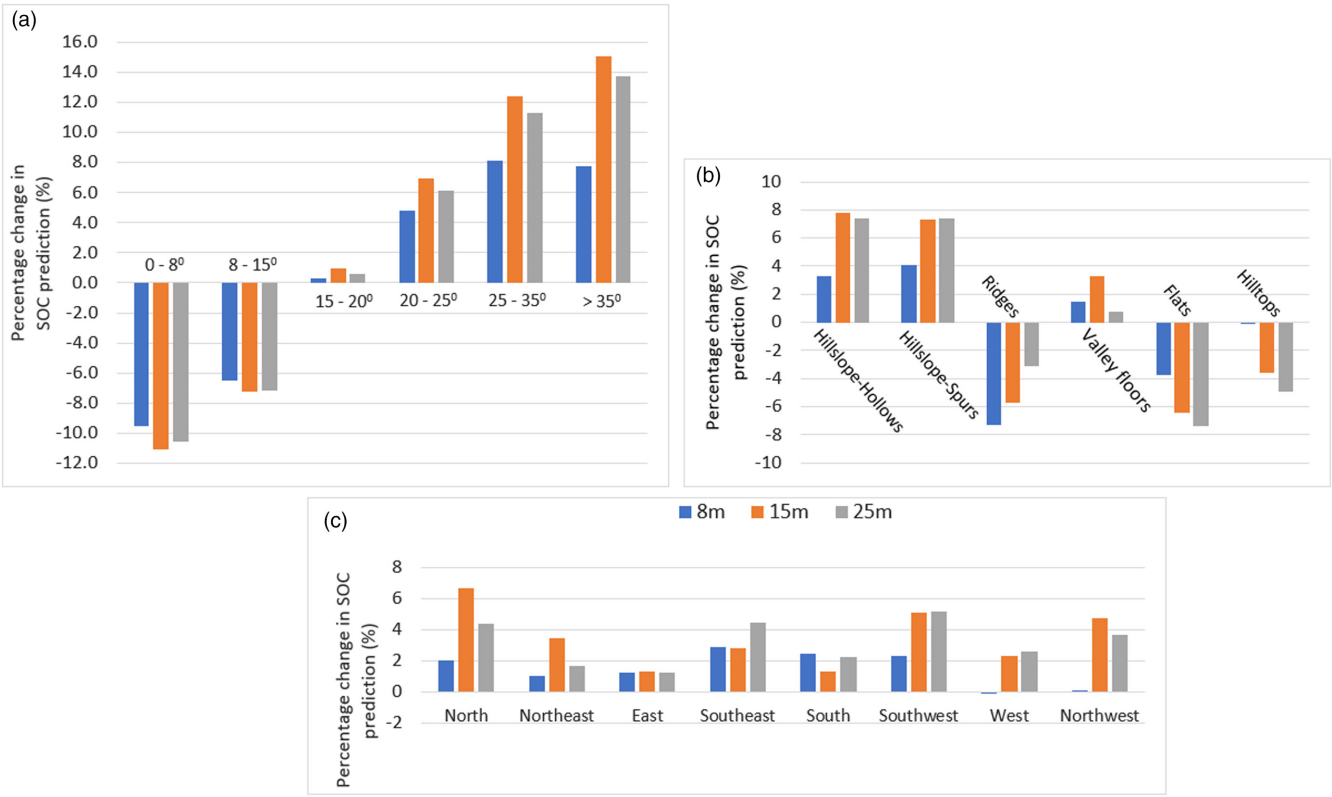


FIGURE 8 Mean percentage change in SOC content derived at 8, 15, 25 m when compared with the 1 m data, based on (a) slope class, (b) landforms, and (c) aspect.

overestimation effect (e.g., up to 15.1% of SOC difference compared to 1 m-based information). It is interesting that the effect of using lower resolution DEM on SOC content prediction had no effect in areas with slopes of 15–20°, showing a very low value in the rate of prediction difference (e.g., 0.3%–1%). In terms of aspect, the changes from fine to coarser resolution often resulted in an increase in SOC content, however, the magnitude of this change was relatively small, with a percentage change in SOC level of less than 5%. Moreover, the effect was found to be larger in the north and southwest aspect, whereas there was no difference among the other aspects. With landforms, the effects of coarser data resolution on predicting SOC content were an overestimation in hillslope and an underestimation in valley floors and flat areas. The effects were relatively low in other landform types. Among the different levels of coarse resolution, SOC content obtained from 8 m data presents the least difference compared to that of 1 m data.

3.3 | Insights into the SOC content pattern across the study area

Results described above demonstrate the highly variable nature of modelled SOC content across space and substantially different amounts of SOC level as a result of modelling with different spatial data resolutions. Given that 1 m data provides the most accurate SOC pattern with ensemble model, this model is used to provide further insight into spatial variation of SOC content in the study area. Figure 9 illustrates modelled SOC pattern for different types of topographical and biophysical features. Figure 9a shows that SOC content decreases with increasing slope. SOC content is highest in areas with slopes less than 15°, and decreases substantially on the steeper lands, and reached the lowest value in the slope greater than 35°.

Patterns of SOC content by different levels of NDMI, TWI, and CMR are similar to those of slope classes but in an inverse trend. In general, SOC content increased remarkably followed by the rising in surface soil moisture (NDMI), topographic wetness (TWI), and clay mineral ratio (CMR). Among other factors, the SOC content varies considerably with landforms (Figure 9d) and soil groups (Figure 9g). For example, flat areas, valley floors, and hilltops are landforms with high SOC content in which the mean SOC is about 1.2–1.4 times higher than other landforms such as hillslopes and ridges. Similarly, SOC content in Ng soil is much more than that of MmS soil (i.e., 5.7% and 4.2%, respectively). The variation in SOC content across aspects is negligible, showing that the mean SOC values for all aspects were mainly between 4.8% and

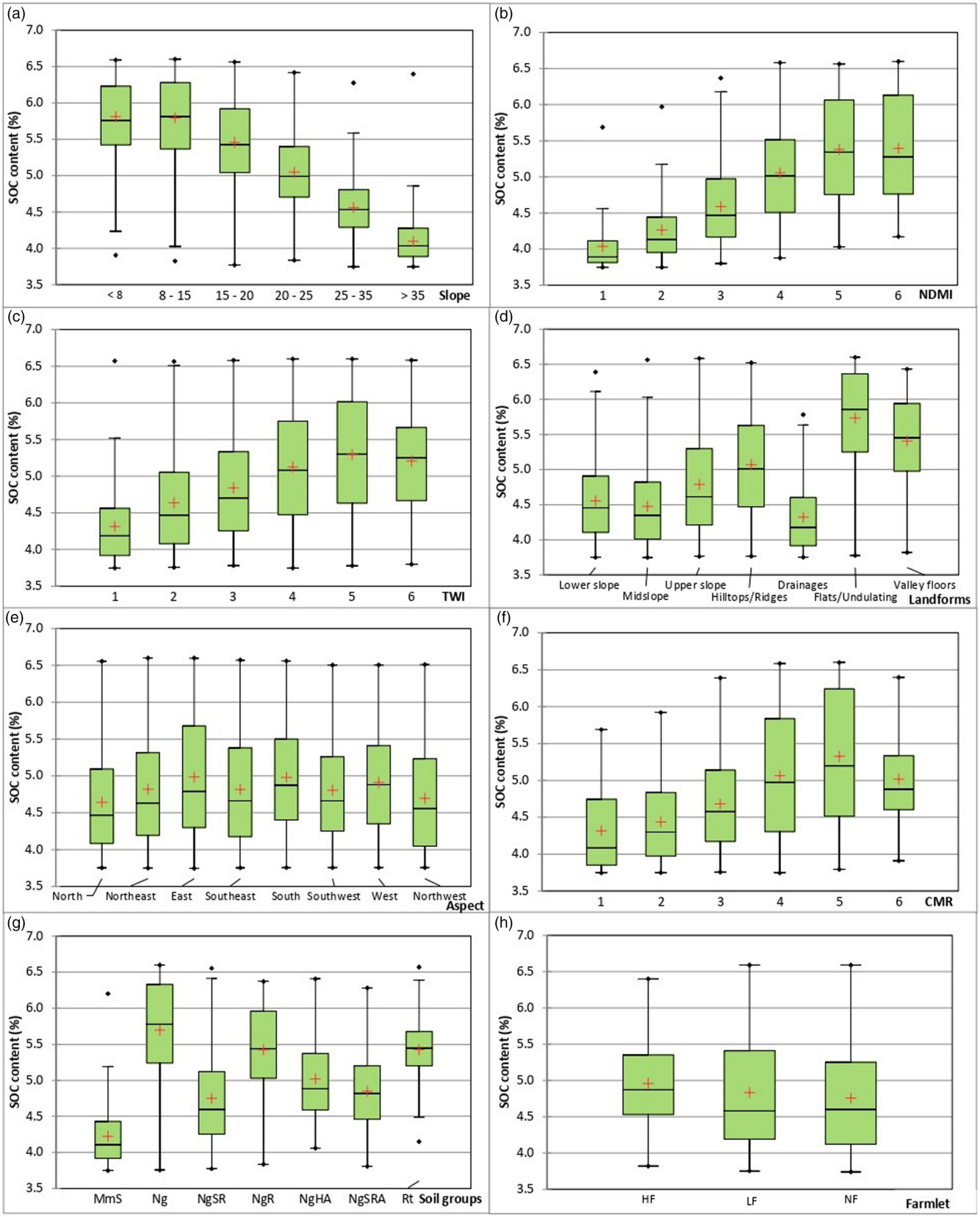


FIGURE 9 The SOC content calculated using the ensemble model using different clusters of topographical and biophysical factors: (a) slope classes, (b) normalized difference soil moisture index (NDMI), (c) topographical wetness index (TWI), (d) landforms, (e) aspect, (f) clay minerals ratio (CMR), (g) soil groups, and (h) management practice histories (HF, high phosphorus; LF, low phosphorus; NF, no phosphorus). The box plots present maximum (upper fence), median (horizontal line inside the box), mean (the plus sign), minimum (lower fence), and the middle 50% of SOC content values (the rectangle box).

5.0%. Similarly, the difference because of P fertilizer and sheep grazing histories are not as important as topographical features and other factors, with the mean SOC content values of 5.0%, 4.8%, and 4.7% for the HF, LF, and NF farmlets, respectively. Among three farmlets, the variation in SOC content in HF farmlet is lower than that of LF or NF farmlets. Notably, it is found that variation in SOC content within a particular category of topography, environmental factor, or management history could be significant and was higher than that of variation between categories/groups. For example, the difference between the aspects with lowest and highest mean SOC content (i.e., Northwest and East, respectively) was only 0.3%. However, the variation within the East aspect was much greater (i.e., 1.4%). This is further illustrated in Figure 10, showing that the SOC content not only differs between landforms but also varies within a single landform. The difference between low and high SOC content within a hilltop, valley floor, or hillslope area is notably higher (e.g., 1.5–2 times).

Hot spot and cold spot patterns of SOC content in the study area are presented in Figure 11. This map reveals clustered areas in the landscape where SOC content is statistically higher, lower, or not significantly different to mean SOC for the study area. Overall, cold spots (clusters of SOC lower than the mean) of SOC content occupy 46.2% total land area, a higher proportion than hot spots of SOC content (clusters of SOC higher than the mean) which occupy 33.9%. It is shown that high SOC content cluster areas (in red) occur mainly in the low and medium slope areas in the west, southeast, and southwest of the study site. The low SOC content cluster areas (in blue) are found in the predominantly high slope areas located in the northern part, central south, and the south of the study area. The area with SOC content that is not significantly clustered (in grey) accounts for 19.9% and is distributed widely across the study site.

4 | DISCUSSION

4.1 | Impacts of spatial data resolution on prediction of SOC content

The study revealed that the spatial resolution of explanatory variables significantly impacts the accuracy with which is possible to model and predict SOC content and patterns in a hill country grassland landscape. Of the four levels of data resolution examined in the modelling of the spatial pattern of SOC content in the study area, 1 m data provided the closest fit to the field data based on a linear regression output and RMSE (Figure 5). Using model output from the 1 m data as a benchmark, coarser spatial resolution DEM derived data and remotely sensed data limited the capability of the model to describe and predict variation in SOC content across the landscape. Similar effects were reported by Lemercier et al. (2022) who found that the ability to predict SOC significantly increased as the scale of the digital soil mapping products progressed from global to national and regional data; however, it was still poor at predicting at the local level.

Results from this study showed that modelling SOC content with coarser spatial resolution progressively reduces the mean value of the modelled SOC content from 5.2% to 5.1%, 4.9%, and 4.8% for the 1, 8, 15, and 25 m datasets, respectively. While these differences are small, the study shows that the spatial resolution of the explanatory variables impacts significantly on the ability of the model to predict the spatial distribution and variation of the SOC content in the study area (Figure 6). The finer resolution data in capturing microtopographic pattern and variation in surface biophysical factors (e.g., changes in slope, surface roughness, and aspect) was better able to capture the influence of the most important variables, in this case slope, on SOC content. With coarser spatial resolution data, the topographical patterns are generalized, and landscape complexity is diminished (Figure 3). The net effect is the numerous small areas of low slopes with high SOC content level (e.g., slopes 0–8°) slowly merged into higher slope classes containing lower level of SOC content (Table 2), resulting in an underestimation of SOC content, as resolution is coarser.

4.2 | Spatial variations in SOC content across the study area

In general, the SOC content quantified in this study (Figure 6) was strongly aligned with the patterns of topographical and surface biophysical features (Figures 1–3, 6 and 10). Of those features, SOC content across slope classes, level of soil moisture (NDMI), topographic wetness (TWI), clay mineral ratio (CMR), and landform is seen more variable than others. This strongly aligns with the information obtained from the model output which reveals the top variables contributing to SOC content prediction (Table 4). The SOC content level was high in low and medium slope areas of the southern paddocks of the study area that were also had high soil moisture levels. Most of the low SOC content areas were in the predominantly hill and steep-land areas with low soil moisture content and high bulk density located in the north, central south, and the south of the study site. These conditions also explain the variations in SOC content, as also reported in several other studies (Kerr & Ochsner, 2020; Patton, Lohse, Seyfried, Godsey, et al., 2019; Patton, Lohse, Seyfried, Will, et al., 2019). In addition, steep-land

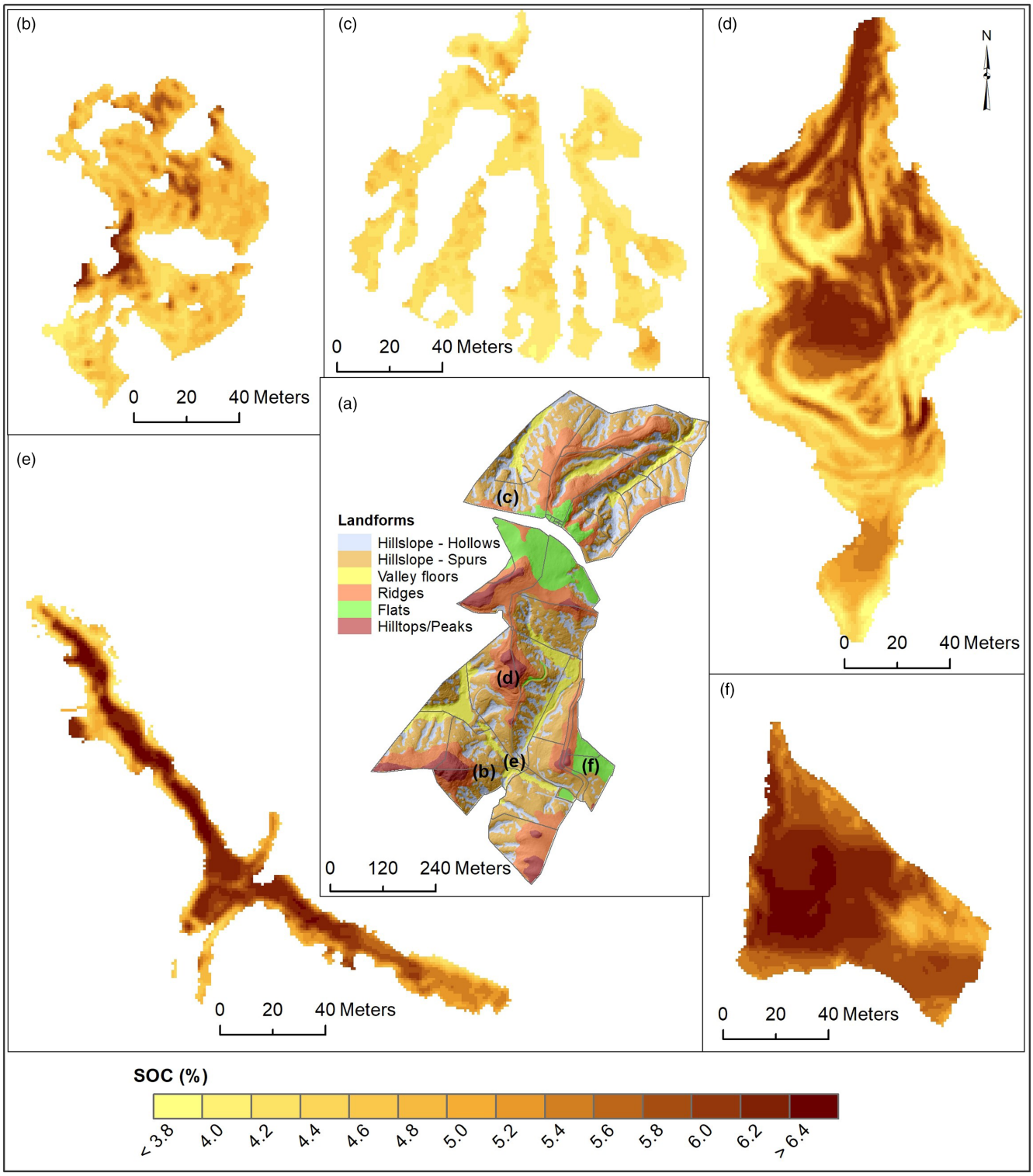


FIGURE 10 Study area showing the variation in soil organic carbon content by (a) landforms classification and soil carbon content by (b) convex hillslope, (c) concave hillslope, (d) hilltops, (e) valley floor, and (f) flat area.

areas have been much more affected by soil erosion and this is a major contribution to the reduction of SOC stock in hill country (Lambert & Roberts, 1976; Schipper et al., 2011) since the high soil C content in topsoil take a long time to rebuild in the exposed subsoil after the topsoil is lost (Basher et al., 2018). The SOC content also varied among landforms, indicating that a high SOC level was mainly in flat landscapes including hilltops and valley floors, whereas a low level of SOC was observed in steeper landforms such as hillslope areas. Our finding

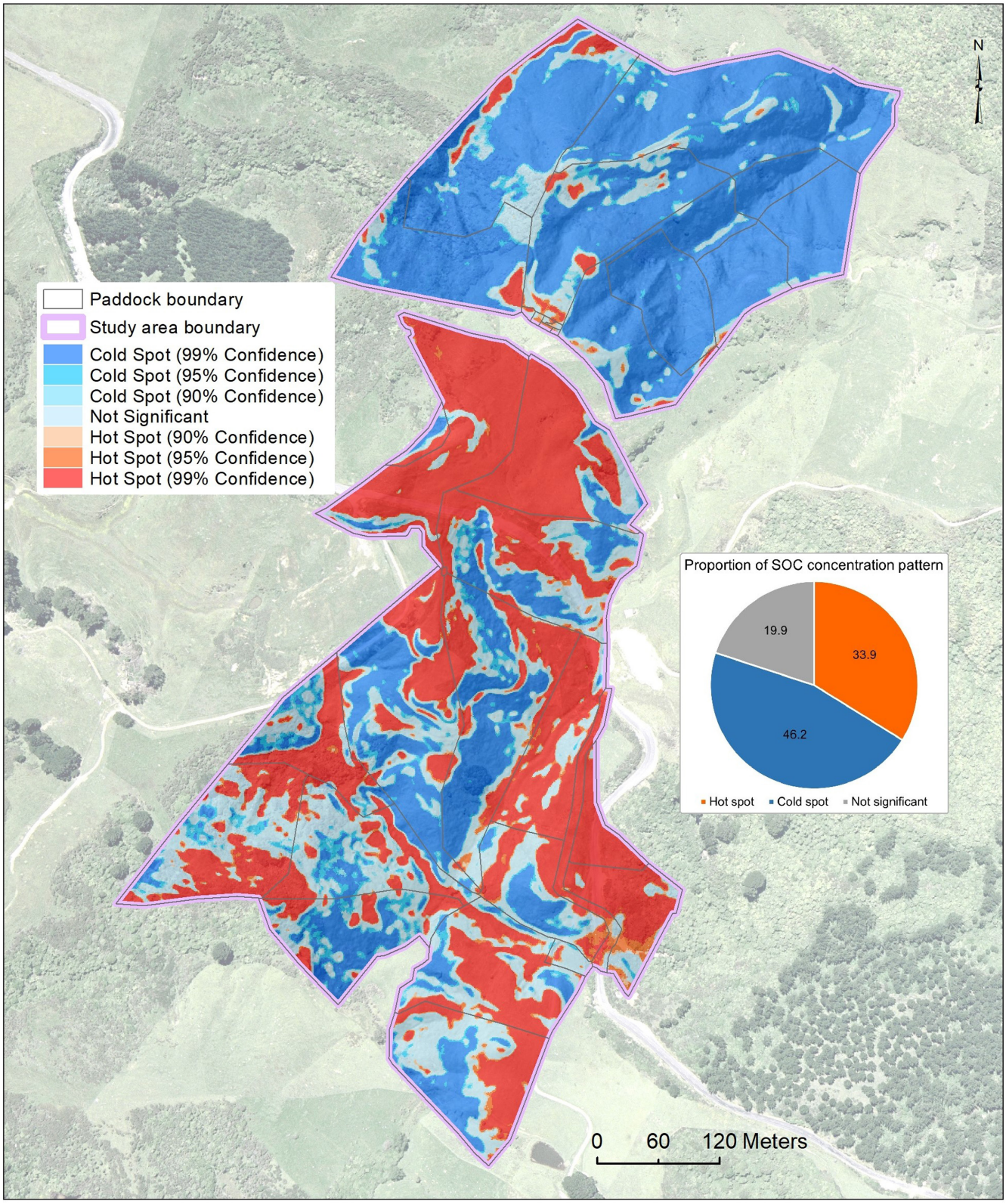


FIGURE 11 The statistically high (hot spot) or low (cold spot) soil organic carbon (SOC) content clustering pattern of the study area, based on the Getis & Ord Gi* index.

is in line with the conclusion from research published by others (Patton, Lohse, Seyfried, Godsey, et al., 2019; Wang, Li, et al., 2022) that highlighted the topographic drivers of SOC content and stocks.

Our study demonstrated that SOC content can be highly variable within a short distance and within a specific landform, slope class, soil, aspect, or a land management. The reason for the difference in SOC content across

space can be attributed to the micro-variations in topography and biophysical conditions of the land surface. In this highly heterogenous landscape SOC content is therefore not simply impacted by a single variable but a product of a combination of complex interactions between the multiple underlying factors and ecological process (Basher et al., 2018; Mackay et al., 2018; Román-Sánchez et al., 2018; Vos et al., 2019). For instance, SOC content differed significantly in areas even with similar land management practice history (e.g., high P fertilizer) because of variation in slopes. In this case, the impact of slope on SOC level is more significant than that of P fertilizer and sheep management history, so this topographical factor controls the SOC variation. The complexity of this pattern and variation may be difficult to observe when simply using the mean value of SOC content. Mapping spatially explicit SOC information in which the SOC content is visualized in a continuous surface is therefore advantageous compared with using information obtained from limited number of point source samples. The successful application of remotely sensed data and topography in predicting agricultural SOC has been demonstrated in research at the regional, national, and global scale (Dvornikov et al., 2021; Meng et al., 2021; Paul et al., 2020; Zhou et al., 2020). Our study is one of the first attempts to demonstrate that an integrated approach combining remote sensing, GIS, and advanced machine learning, coupled with point sourced data provides comprehensive spatial information of on-farm SOC content in hill country grassland landscapes.

4.3 | Implications for soil C modelling and management practices

Results from our study demonstrate that integrating prediction capabilities from several machine learning algorithms and using a wide range of explanatory factors, in combination with point sourced data, is an effective method to improve SOC modelling in complex pastoral farmed landscapes, in line with several recent studies (Mishra et al., 2020; Nguyen et al., 2022; Tajik et al., 2020). Our study provides evidence supporting the use of advanced machine learning to extend the learning from the current in situ single sampling site method. Automated machine learning integrated into a GIS enables users, especially those without technical expertise, to reap the advantage of advanced geospatial technologies (a combination of GIS, remote sensing, and advanced machine learning) to spatially quantify and visualize SOC content at multiple scales (i.e., from a per-pixel level to farm scale). People who are not experts in machine learning can focus on applying their domain expertise to a specific business problem or domain application, rather than on the machine learning workflow itself (Hutter et al., 2019). Spatial modelling and assessment of on-farm SOC provides a comprehensive understanding of the SOC pattern across space and reveals insights in the relationship between SOC pattern and underlying topographical and environmental factors that may not be achievable via a more traditional approach (e.g., point sample-based analysis).

Of the machine learning algorithms evaluated within the study areas, the ensemble method and its ability to integrate multiple machine learning algorithms accounted for more of the variation in SOC content and was mostly closely correlated with the SOC content in the 0–75 mm soil depth measured in the 47 field sampling sites. Given that the prediction performance often varies between machine learning algorithms and may differ between study areas (Keskin et al., 2019; Nguyen et al., 2022; Pham et al., 2021; Shirazi et al., 2023), we recommend the use of ensemble model that integrates different machine learning algorithms to obtain the better SOC prediction results. Doing so allows the most accurate prediction to be acquired and results from this step can be effectively used for further analysis.

Our study revealed that moving from finer to coarser data has a significant effect on mapping the spatial pattern of SOC content across the farmed landscapes. Given that a decrease in data resolution reduces the ability of the model to capture the changes in spatial pattern of SOC content in complex landscape, we suggest that SOC modelling and mapping practice needs to consider this effect so that information obtained from the model can be interpreted appropriately and with a certain level of confidence. Specifically, we recommend the use of high spatial resolution data to obtain the most detailed and accurate SOC pattern, and this can therefore be generalized to different levels of management (e.g., 1-ha, sub paddock, paddock). Whist high-resolution DEM such as 1 m LiDAR data is not widely available, current studies and mapping practices often use DEM data with the spatial resolution ≥30 m for SOC modelling (Ahmed et al., 2022; de Castro Padilha et al., 2020; Dharumarajan et al., 2021; He et al., 2021). As such, it is important that uncertainty in SOC information achieved from these data needs to be quantified. In other words, sensitivity analysis should be carried out to define the level of confidence in SOC prediction so that the results will be interpreted and applied appropriately.

Considering that SOC pattern varies across space and strongly relates to the pattern of topography, land surface features, soils, and land management, it is important that these factors are considered in identifying the sites for sampling SOC, especially in highly complex topography landscapes where variation in SOC content is significant. Given that SOC sampling can be costly, especially in the complex landscape often found in hill

country, an insight into the influence of the number and location of the in situ point source data had in the development of the algorithms would be invaluable. For instance, the effects of topography and surface biophysical factors on mapping SOC content where more significant than soil groups and management practice history a greater number of samples would be allocated to these locations. Therefore, we suggested the design of a stratified and probability-based soil carbon sampling regime by statistically and spatially interrogating multiple data layers to better capture SOC pattern in NZ hill country landscapes. This will provide better in situ data for spatial model.

Spatially modelling SOC content provides the scientific basis to achieve one of the key objectives of soil carbon study, which is to identify potential areas within a farm where soil C sequestration is most likely to be achievable or where soil C loss risk is greatest. Also, quantified spatial SOC information (i.e., mapped SOC content) provides a valuable tool for developing farm management strategies that can help maintain and improve SOC content and therefore enhancing soil health. A SOC content map with multiple sources of information, such as topography, land use and land cover (LULC), and pastoral production, can help develop targeted management strategies that are tailored to specific areas within a farm. For instance, in steep-land areas with low level of SOC content and low pasture productivity, tree planting rather than soil management practices (e.g., biochar and tillage practices) should be applied to improve future carbon stock. In addition, spatial information on the variability of SOC content is also useful for developing variable-rate fertilizer applications to target areas with low SOC content. Having spatially explicit SOC information enables the application of multiple levels of management, including location-based and paddock-based management, to ensure that management practices are targeted to the specific needs of each area. This is an advantage, given that most of the current management practices at the farm scale are often developed based on average information for the paddock or land management unit.

4.4 | Limitations of the study

Our study focuses on mapping SOC patterns under a long-term permanent pasture. We did not consider the influence of perturbations, such as cultivation, an erosion event, or an extreme livestock treading event, that could have impacted on SOC content across the landscape over time. The ability to quantify temporal, in addition to spatial changes in SOC content is important as it allows the development of more effective SOC management strategies. For instance, knowing the areas where SOC content is declining, or increased overtime would be valuable information to identify targeted management areas and suitable management practices to optimize carbon sequestration over time.

In this study, we utilized an existing field SOC information from soil samples collected from varying slopes, aspects, and areas with different management histories. It is important to recognize that these samples may not capture the full range of soils and landscape conditions or adequately represent the spatial variability within the study site. It is important that future studies assess the effects of sampling methods on the performance on spatial modelling of SOC. Furthermore, integration with a process-based model that accounts for nutrient transfer by grazing animals would be useful to capture the full complexity of SOC dynamics. In our study, different spatial resolutions of remotely sensed data were obtained by resampling from Sentinel 2 data with a native resolution of 10 m. It should be acknowledged that resampling the imagery data to match the DEM data resolutions is not the same as using imagery with resolutions in their original or native spatial resolution. As such, it would be helpful to use remote sensing data acquired from different sensors so resulting from data resampling is minimized.

5 | CONCLUSIONS

We assessed the effect that geospatial data resolution has on the prediction performance of a range of models utilizing machine learning, to describe and predict the pattern of SOC content in topographically complex grassland landscape. Results from this study demonstrate the effectiveness of the applied machine learning approach, showing that (1) ensemble-based learning that combines prediction results from several algorithms outperformed an individual model in the estimates of pastoral SOC content; (2) the use of multi-sourced data that include a wide range of predictors covering topographical features, surface biophysical pattern, and soils and land management practices, enables better prediction outcomes; and (3) an increase in spatial data resolution results in significant improvements in the prediction and measured spatial pattern on the change in spatial pattern (i.e., distribution and variation) of SOC content in complex hill country landscapes.

Our results indicate that by applying advanced digital and machine learning technologies, the spatial pattern of agricultural SOC can be quantified from existing in situ samples collected from a small number of sites. This provides important information to better understand SOC pattern at the required scale and can be used to undertake

spatial monitoring and management of soil carbon. The ability to quantify the spatially explicit pattern and evaluate the effects data quality used has on estimates of SOC in a complex farmed landscape is invaluable. Utilizing spatial-based information leads to the development of more effective and sustainable carbon management and adaptation strategies at a farm scale that can contribute to the improvement of soil health and abatement of agricultural GHG emissions.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers who provided valuable comments and suggestions to improve this manuscript. Thank you to Dr. Mike Dodd and Dr. Jiafa Luo for an internal review of the manuscript. This study was supported by the AgResearch Strategic Science Investment Fund "NZ Bioeconomy in the Digital Age" (NZBIDA). Open access publishing facilitated by AgResearch Ltd, as part of the Wiley - AgResearch Ltd agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

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

ORCID

Duy X. Tran https://orcid.org/0000-0003-3807-9925

REFERENCES

Acharya, U., Lal, R., & Chandra, R. (2022). Data driven approach on in-situ soil carbon measurement. *Carbon Management*, *13*(1), 401–419.

Ahmed, I. S., Hassan, F. A., Sulieman, M. M., Keshavarzi, A., Elmobarak, A. A., Yousif, K. M., & Brevik, E. C. (2022). Using environmental covariates to predict soil organic carbon stocks in Vertisols of Sudan. *Geoderma Regional*, *31*, e00578.

Aitkenhead, M., & Coull, M. (2016). Mapping soil carbon stocks across Scotland using a neural network model. *Geoderma*, *262*, 187–198.

Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G., & Bochtis, D. (2019). Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sensing*, *11*(6), 676.

Barazzetti, L., Previtali, M., & Roncoroni, F. (2022). Visualization and processing of structural monitoring data using space-time cubes. In O. Gervasi, B. Murgante, E. M. T. Hendrix, D. Taniar, & B. O. Apduhan (Eds.), *International conference on computational science and its applications* (pp. 19–31). Springer International Publishing.

Basher, L., Betts, H., Lynn, I., Marden, M., McNeill, S., Page, M., & Rosser, B. (2018). A preliminary assessment of the impact of landslide, earthflow, and gully erosion on soil carbon stocks in New Zealand. *Geomorphology*, *307*, 93–106.

Bronick, C. J., & Lal, R. (2005). Soil structure and management: A review. *Geoderma*, *124*(1–2), 3–22.

Burkitt, L., & Bretherton, M. (2022). The importance of incorporating geology, soil, and landscape knowledge in freshwater farm planning in Aotearoa New Zealand. *Frontiers in Soil Science*, *2*, 956692.

Cameron, D. (2016). Sustaining the productivity of New Zealand's hill country-A land manager's view. *NZGA: Research and Practice Series*, *16*, 151–155.

Carranza, E., & Hale, M. (2002). Mineral imaging with Landsat Thematic Mapper data for hydrothermal alteration mapping in heavily vegetated terrane. *International Journal of Remote Sensing*, *23*(22), 4827–4852.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the 21st International Conference on Machine Learning* (p. 18).

Chartin, C., Stevens, A., Goidts, E., Krüger, I., Carnol, M., & van Wesemael, B. (2017). Mapping Soil Organic Carbon stocks and estimating uncertainties at the regional scale following a legacy sampling strategy (Southern Belgium, Wallonia). *Geoderma Regional*, *9*, 73–86.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chowdhury, M. S. (2023). Modelling hydrological factors from DEM using GIS. *MethodsX*, *10*, 102062.

Christie, D., & Neill, S. (2021). Measuring and observing the ocean renewable energy resource. In T. M. Letcher (Ed.), *Reference module in earth systems and environmental sciences* (pp. 149–175). Elsevier.

Cumberland, K. B. (1941). A century's change: Natural to cultural vegetation in New Zealand. *Geographical Review*, *31*(4), 529–554.

de Castro Padilha, M. C., Vicente, L. E., Demattê, J. A., Loebmann, D. G. D. S. W., Vicente, A. K., Salazar, D. F., & Guimarães, C. C. B. (2020). Using Landsat and soil clay content to map soil organic carbon of oxisols and Ultisols near São Paulo, Brazil. *Geoderma Regional*, *21*, e00253.

De Graaff, M. A., Van Groenigen, K. J., Six, J., Hungate, B., & van Kessel, C. (2006). Interactions between plant growth and soil nutrient cycling under elevated CO2: A meta-analysis. *Global Change Biology*, *12*(11), 2077–2091.

Dewage, S. N. S. P., Minasny, B., & Malone, B. (2020). Disaggregating a regional-extent digital soil map using Bayesian area-to-point regression kriging for farm-scale soil carbon assessment. *The Soil*, *6*(2), 359–369.

Dharumarajan, S., Kalaiselvi, B., Suputhra, A., Lalitha, M., Vasundhara, R., Kumar, K. A., Nair, K. M., Hegde, R., Singh, S. K., & Lagacherie, P. (2021). Digital soil mapping of soil organic carbon stocks in Western Ghats, South India. *Geoderma Regional*, *25*, e00387.

Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Casanova Pinto, M., Casanova-Katny, A., Muñoz, C., Boudin, M., & Zagal Venegas, E. (2015). Soil carbon storage controlled by interactions between geochemistry and climate. *Nature Geoscience*, *8*(10), 780–783.

Dvornikov, Y. A., Vasenev, V. I., Romzaykina, O. N., Grigorieva, V. E., Litvinov, Y. A., Gorbov, S. N., Dolgikh, A. V., Korneykova, M. V., & Gosse, D. D. (2021). Projecting the urbanization effect on soil

organic carbon stocks in polar and steppe areas of European Russia by remote sensing. *Geoderma*, *399*, 115039.

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., & Scholten, T. (2020). Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sensing*, *12*(14), 2234.

ESRI. (2022). *How AutoML works?* Retrieved March 31, 2023, from https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-automl-works.htm

Franzluebbers, A., Nazih, N., Stuedemann, J., Fuhrmann, J., Schomberg, H., & Hartel, P. (1999). Soil carbon and nitrogen pools under low-and high-endophyte-infected tall fescue. *Soil Science Society of America Journal*, *63*(6), 1687–1694.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, *24*(3), 189–206.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*, 3–42.

Gómez-Ramírez, J., Ávila-Villanueva, M., & Fernández-Blázquez, M. Á. (2020). Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. *Scientific Reports*, *10*(1), 20630.

Gray, J. M., Wang, B., Waters, C. M., Orgill, S. E., Cowie, A. L., & Ng, E. L. (2022). Digital mapping of soil carbon sequestration potential with enhanced vegetation cover over New South Wales, Australia. *Soil Use and Management*, *38*(1), 229–247.

He, X., Yang, L., Li, A., Zhang, L., Shen, F., Cai, Y., & Zhou, C. (2021). Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. *Catena*, *205*, 105442.

Hedley, C., Manderson, A., Mudge, P., Roudier, P., Fraser, S., Parfitt, R., Smaill, S., Schipper, L., & Kelliher, F. (2015). *Improved measurements of hill country soil carbon—To assist carbon change studies* (p. 2). Ministry for Primary Industries.

Heil, J., Jörges, C., & Stumpe, B. (2022). Evaluation of using digital photography as a cost-effective tool for the rapid assessment of soil organic carbon at a regional scale. *Soil Security*, *6*, 100023.

Herrick, J. E., & Wander, M. M. (2018). Relationships between soil organic carbon and soil quality in cropped and rangeland soils: The importance of distribution, composition, and soil biological activity. In R. Lal, J. Kimble, R. Follet, & B. A. Stewart (Eds.), *Soil processes and the carbon cycle* (pp. 405–425). CRC Press.

Holmquist, J. R., Windham-Myers, L., Bliss, N., Crooks, S., Morris, J. T., Megonigal, J. P., Troxler, T., Weller, D., Callaway, J., & Drexler, J. (2018). Accuracy and precision of tidal wetland soil carbon mapping in the conterminous United States. *Scientific Reports*, *8*(1), 9478.

Hontoria, C., Gómez-Paccard, C., Mariscal-Sancho, I., Benito, M., Pérez, J., & Espejo, R. (2016). Aggregate size distribution and associated organic C and N under different tillage systems and Ca-amendment in a degraded Ultisol. *Soil and Tillage Research*, *160*, 42–52.

Hoogendoorn, C. J., Newton, P. C. D., Devantier, B. P., Rolle, B. A., Theobald, P. W., & Lloyd-West, C. M. (2016). Grazing intensity and micro-topographical effects on some nitrogen and carbon pools and fluxes in sheep-grazed hill country in New Zealand. *Agriculture, Ecosystems & Environment*, *217*, 22–32.

Huang, P. M., Li, Y., & Sumner, M. E. (2011). *Handbook of soil sciences: Properties and processes*. CRC Press.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges* (p. 219). Springer Nature.

Jayaraman, S., & Dalal, R. C. (2021). Conservation agriculture: Carbon turnover and carbon sequestration for enhancing soil sustainability and mitigation of climate change. *Conservation Agriculture: A Sustainable Approach for Soil Health and Food Security: Conservation Agriculture for Sustainable Agriculture*, 289–298.

Kaczynski, R., Siebielec, G., Hanegraaf, M. C., & Korevaar, H. (2017). Modelling soil carbon trends for agriculture development scenarios at regional level. *Geoderma*, *286*, 104–115.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, *30*, 3149–3157.

Kerr, D. D., & Ochsner, T. E. (2020). Soil organic carbon more strongly related to soil moisture than soil temperature in temperate grasslands. *Soil Science Society of America Journal*, *84*(2), 587–596.

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

Khan, N., Jhariya, M. K., Raj, A., Banerjee, A., & Meena, R. S. (2021). Soil carbon stock and sequestration: Implications for climate change adaptation and mitigation. In M. K. Jhariya, R. S. Meena, & A. Banerjee (Eds.), *Ecological intensification of natural resources for sustainable agriculture* (pp. 461–489). Springer.

Klemas, V., & Smart, R. (1983). The influence of soil salinity, growth form, and leaf moisture on-the spectral radiance of. *Photogrammetric Engineering and Remote Sensing*, *49*, 77–83.

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. *Science*, *304*(5677), 1623–1627.

Lambert, M. G., & Roberts, E. (1976). Aspect differences in an unimproved hill country pasture: I. Climatic differences. *New Zealand Journal of Agricultural Research*, *19*(4), 459–467.

Lemercier, B., Lagacherie, P., Amelin, J., Sauter, J., Pichelin, P., Richer-De-Forges, A. C., & Arrouays, D. (2022). Multiscale evaluations of global, national and regional digital soil mapping products in France. *Geoderma*, *425*, 116052.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Mackay, A. D., Vibart, R., & McKenzie, C. (2018). Changes in soil carbon in hill-country under contrasting phosphorus fertiliser and sheep stocking rates. *Journal of New Zealand Grasslands*, *80*, 263–268.

Mackay, A. D., Vibart, R., McKenzie, C., Costall, D., Bilotto, F., & Kelliher, F. M. (2021). Soil organic carbon stocks in hill country pastures under contrasting phosphorus fertiliser and sheep stocking regimes, and topographical features. *Agricultural Systems*, *186*, 102980.

Maleki, S., Khormali, F., Mohammadi, J., Bogaert, P., & Bodaghabadi, M. B. (2020). Effect of the accuracy of topographic data on improving digital soil mapping predictions with limited soil data: An application to the Iranian loess plateau. *Catena*, *195*, 104810.

Malone, B. P., Styc, Q., Minasny, B., & McBratney, A. B. (2017). Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. *Geoderma*, *290*, 91–99.

Manderson, A., & Palmer, A. (2006). Soil information for agricultural decision making: A New Zealand perspective. *Soil Use and Management*, *22*(4), 393–400.

Manns, H. R., Parkin, G. W., & Martin, R. C. (2016). Evidence of a union between organic carbon and water content in soil. *Canadian Journal of Soil Science*, *96*(3), 305–316.

Meng, X., Bao, Y., Liu, J., Liu, H., Zhang, X., Zhang, Y., Wang, P., Tang, H., & Kong, F. (2020). Regional soil organic carbon prediction model based on a discrete wavelet analysis of hyperspectral satellite data. *International Journal of Applied Earth Observation and Geoinformation*, *89*, 102111.

Minasny, B., McBratney, A. B., Malone, B. P., & Wheeler, I. (2013). Digital mapping of soil carbon. *Advances in Agronomy*, *118*, 1–47.

Mishra, U., Gautam, S., Riley, W. J., & Hoffman, F. M. (2020). Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. *Frontiers in Big Data*, *3*, 528441.

Nguyen, T. T., Pham, T. D., Nguyen, C. T., Delfos, J., Archibald, R., Dang, K. B., Hoang, N. B., Guo, W., & Ngo, H. H. (2022). A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of the Total Environment*, *804*, 150187.

Odebiri, O., Odindi, J., & Mutanga, O. (2021). Basic and deep learning models in remote sensing of soil organic carbon estimation: A brief review. *International Journal of Applied Earth Observation and Geoinformation*, *102*, 102389.

Patton, N. R., Lohse, K. A., Seyfried, M., Will, R., & Benner, S. G. (2019). Lithology and coarse fraction adjusted bulk density estimates for determining total organic carbon stocks in dryland soils. *Geoderma*, *337*, 844–852.

Patton, N. R., Lohse, K. A., Seyfried, M. S., Godsey, S. E., & Parsons, S. B. (2019). Topographic controls of soil organic carbon on soil-mantled landscapes. *Scientific Reports*, *9*(1), 1–15.

Paul, S., Coops, N., Johnson, M., Krzic, M., Chandna, A., & Smukler, S. (2020). Mapping soil organic carbon and clay using remote sensing to predict soil workability for enhanced climate change adaptation. *Geoderma*, *363*, 114177.

Paustian, K., Lehmann, J., Ogle, S., Reay, D., Robertson, G. P., & Smith, P. (2016). Climate-smart soils. *Nature*, *532*(7597), 49–57.

Pham, T. D., Yokoya, N., Nguyen, T. T. T., Le, N. N., Ha, N. T., Xia, J., Takeuchi, W., & Pham, T. D. (2021). Improvement of mangrove soil carbon stocks estimation in North Vietnam using Sentinel-2 data and machine learning approach. *GIScience & Remote Sensing*, *58*(1), 68–87.

Pramanick, B., Kumar, M., Singh, S. K., Sapna, K., & Maitra, S. (2021). Soil-centric approaches towards climate-resilient agriculture. In A. Rakshit, S. K. Singh, P. C. Abhilash, & A. Biswas (Eds.), *Soil science: Fundamentals to recent advances* (pp. 333–359). Springer.

Qiu, W., Curtin, D., & Beare, M. (2011). *Spatial variability of available nutrients and soil carbon under arable cropping in Canterbury* (pp. 1–7). The New Zealand Institute for Plant and Food Research Limited.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

Reisinger, A., Clark, H., Journeaux, P., Clark, D., & Lambert, G. (2017). *On-farm options to reduce agricultural GHG emissions in New Zealand* (p. 68). Report to the Biological Emissions Reference Group. Retrieved from https://www.mpi.govt.nz/dmsdocument/32158-berg-current-mitigaiton-potential-final

Rikimaru, A., Roy, P., & Miyatake, S. (2002). Tropical forest cover density mapping. *Tropical Ecology*, *43*(1), 39–47.

Román-Sánchez, A., Vanwalleghem, T., Peña, A., Laguna, A., & Giráldez, J. (2018). Controls on soil carbon storage from topography and vegetation in a rocky, semi-arid landscapes. *Geoderma*, *311*, 159–166.

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS. *NASA Special Publications*, *351*(1), 309.

Ryals, R., Hartman, M. D., Parton, W. J., DeLonge, M. S., & Silver, W. L. (2015). Long-term climate change mitigation potential with organic matter management on grasslands. *Ecological Applications*, *25*(2), 531–545.

Sanderman, J., Hengl, T., Fiske, G., Solvik, K., Adame, M. F., Benson, L., Bukoski, J. J., Carnell, P., Cifuentes-Jara, M., & Donato, D. (2018). A global map of mangrove forest soil carbon at 30 m spatial resolution. *Environmental Research Letters*, *13*(5), 055002.

Schipper, L. A., Dodd, M., Fisk, L., Power, I., Parenzee, J., & Arnold, G. (2011). Trends in soil carbon and nutrients of hill-country pastures receiving different phosphorus fertilizer loadings for 20 years. *Biogeochemistry*, *104*, 35–48.

Schmidt, J., & Hewitt, A. (2004). Fuzzy land element classification from DTMs based on geometry and terrain position. *Geoderma*, *121*(3–4), 243–256.

Senthilkumar, S., Kravchenko, A., & Robertson, G. (2009). Topography influences management system effects on total soil carbon and nitrogen. *Soil Science Society of America Journal*, *73*(6), 2059–2067.

Shen, Z., Ramirez-Lopez, L., Behrens, T., Cui, L., Zhang, M., Walden, L., Wetterlind, J., Shi, Z., Sudduth, K. A., Baumann, P., Song, Y., Catambay, K., & Rossel, R. A. V. (2022). Deep transfer learning of global spectra for local soil carbon monitoring. *ISPRS Journal of Photogrammetry and Remote Sensing*, *188*, 190–200.

Shirazi, F. R. A., Shahbazi, F., Rezaei, H., & Biswas, A. (2023). Digital assessments of soil organic carbon storage using digital maps provided by static and dynamic environmental covariates. *Soil Use and Management*, *39*, 948–974.

Smith, P. (2016). Soil carbon sequestration and biochar as negative emission technologies. *Global Change Biology*, *22*(3), 1315–1324.

Soh, J., & Singh, P. (2020). Machine learning operations. In *Data science solutions on Azure: Tools and techniques using Databricks and MLOps* (pp. 259–279). Apress. https://link.springer.com/book/10.1007/978-1-4842-6405-8#about-this-book

Stevenson, B. A., Sarmah, A. K., Smernik, R., Hunter, D. W., & Fraser, S. (2016). Soil carbon characterization and nutrient ratios across land uses on two contrasting soils: Their relationships to microbial biomass and function. *Soil Biology and Biochemistry*, *97*, 50–62.

Stockmann, U., Adams, M. A., Crawford, J. W., Field, D. J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A. B., De Courcelles, V. R., & Singh, K. (2013). The knowns, known unknowns and unknowns of sequestration of soil organic carbon. *Agriculture, Ecosystems & Environment*, *164*, 80–99.

Tajik, S., Ayoubi, S., & Zeraatpisheh, M. (2020). Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. *Geoderma Regional*, *20*, e00256.

Van Cleve, K., & Powers, R. F. (1995). Soil carbon, soil formation, and ecosystem development. In W. W. McFee & J. M. Kelly (Eds.), *Carbon forms and functions in forest soils* (pp. 155–200). Soil Science Society of America Inc.

Van Huynh, C., Pham, T. G., Nguyen, L. H. K., Nguyen, H. T., Nguyen, P. T., Le, Q. N. P., Tran, P. T., Nguyen, M. T. H., & Tran, T. T. A. (2022). Application GIS and remote sensing for soil organic carbon mapping in a farm-scale in the hilly area of central Vietnam. *Air, Soil and Water Research*, *15*, 11786221221114777.

Vaze, J., Teng, J., & Spencer, G. (2010). Impact of DEM accuracy and resolution on topographic indices. *Environmental Modelling & Software*, *25*(10), 1086–1098.

Vos, C., Don, A., Hobley, E. U., Prietz, R., Heidkamp, A., & Freibauer, A. (2019). Factors controlling the variation in organic carbon stocks in agricultural soils of Germany. *European Journal of Soil Science*, *70*(3), 550–564.

Wang, S., Guan, K., Zhang, C., Lee, D., Margenot, A. J., Ge, Y., Peng, J., Zhou, W., Zhou, Q., & Huang, Y. (2022). Using soil library hyperspectral reflectance and machine learning to predict soil organic carbon: Assessing potential of airborne and spaceborne optical soil sensing. *Remote Sensing of Environment*, *271*, 112914.

Wang, Z. P., Li, X. P., Pelletier, R., Chang, S. X., & Bork, E. W. (2022). Grassland soil organic carbon and the effects of irrigated cropping in Alberta, Canada. *Soil Use and Management*, *38*(2), 1189–1202.

Whitehead, D., Baisden, T., Campbell, D., Curtin, D., Davis, M. R., Hedley, C. B., Beare, D. M., Jones, H., Kelliher, F. M., & Saggar, S. (2012). *Review of soil carbon measurement methodologies and technologies, including nature and intensity of sampling, their uncertainties and costs*. Ministry for Primary Industries.

Whitehead, D., McNeill, S. J., & Mudge, P. L. (2021). Regional and national changes in soil carbon stocks with land-use change from 1990 to 2016 for New Zealand. *Regional Environmental Change*, *21*, 1–13.

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194.

Xu, Y., Smith, S. E., Grunwald, S., Abd-Elrahman, A., & Wani, S. P. (2017). Incorporation of satellite remote sensing pan-sharpened imagery into digital soil prediction and mapping models to characterize soil property variability in small agricultural fields. *ISPRS Journal of Photogrammetry and Remote Sensing*, *123*, 1–19.

Yang, Y. (2016). *Temporal data mining via unsupervised ensemble learning*. Elsevier.

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., & Lausch, A. (2020). High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. *Science of the Total Environment*, *729*, 138244.

Zöller, M. A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research*, *70*, 409–472.

How to cite this article: Tran, D. X., Dominati, E., Lowry, J., Mackay, A., Vibart, R., Pearson, D., Devantier, B., & Noakes, E. (2024). Effects of spatial data resolution on the modelling and mapping of soil organic carbon content in hill country grassland landscapes. *Soil Use and Management*, *40*, e12966. https://doi.org/10.1111/sum.12966