Journal of Geovisualization and Spatial Analysis (2026) 10:37
https://doi.org/10.1007/s41651-026-00279-y

RESEARCH

Pattern Similarity Learning for Spatially Adaptive Soil Organic Carbon Mapping

Haiyang Liu[1] · Yongze Song[2] · Pengcheng Zhang[3] · Ning Yang[4,5]

Received: 5 April 2026 / Accepted: 12 July 2026

Abstract

Predicting continuous environmental variables from sparse spatial observations remains challenging because global machine learning models can capture nonlinear relationships but often have limited ability to adapt to local spatial heterogeneity, whereas existing similarity-based methods still rely mainly on weighted-average prediction. This study proposes Pattern Similarity Learning (PSL), which uses pattern similarity as a spatially adaptive indicator to guide local nonlinear prediction. PSL first augments the original environmental covariates with three pattern-derived indicators: geocomplexity, positive outlier strength, and negative outlier strength. These indicators describe local heterogeneity and local anomalousness, allowing each covariate to carry both its original value and its local spatial context. For each target location, PSL then measures covariate-space similarity to identify relevant training samples and fits a similarity-weighted local Random Forest using the pattern-enhanced features. In a soil organic carbon prediction case across the conterminous United States (CONUS), five-fold spatial cross-validation shows that PSL achieved the best overall pooled performance among the six models. Compared with Random Forest, PSL increased R^2 by 0.017 (7.11%), reduced RMSE by 0.017 (1.14%), and reduced MAE by 0.012 (1.69%). These results suggest that using pattern similarity to link local spatial context, sample relevance, and nonlinear learning can improve the spatial adaptability of continuous environmental prediction.

Keywords Spatial prediction · Pattern similarity · Random forest · Geocomplexity · Soil · Organic carbon · Local modelling

✉ Yongze Song
Yongze.song@curtin.edu.au

Haiyang Liu
haiyang.liu.q6@elms.hokudai.ac.jp

Pengcheng Zhang
pengcheng.zhang@connect.polyu.hk

Ning Yang
yangn.20s@igsnrr.ac.cn

1 Graduate School of Environmental Science, Hokkaido University, Sapporo, Japan

2 School of Design and the Built Environment, Curtin University, Perth, Australia

3 Department of Building and Real Estate, Hong Kong Polytechnic University, Hong Kong, China

4 University of Chinese Academy of Sciences, Beijing, China

5 National Key Laboratory of Geographic Information Science and Technology, Institute of Geographic Sciences and Natural Resources Research, Beijing, China

Introduction

Existing methods for spatial prediction can be broadly grouped into three related streams: classical geostatistical methods, global machine learning methods, and similarity-based local spatial models. Classical spatial prediction methods, such as ordinary kriging and regression kriging, make use of spatial autocorrelation by describing how correlations between sample points change with distance (Matheron 1963; Odeh et al. 1995; Heuvelink and Webster 2001). Regression kriging usually first uses environmental covariates to model the overall trend of the target variable, and then applies kriging to the remaining spatial residuals. As a result, it often performs well when spatial autocorrelation is strong (Hengl et al. 2007; Keskin and Grunwald 2018). However, these methods usually assume that spatial dependence is broadly stable across the study area and require its form to be specified in advance (Atkinson and Lloyd 2007).

Published online: 18 July 2026

A second stream is global machine learning, which emphasizes nonlinear covariate–response relationships but usually has limited spatial adaptivity. Compared with classical spatial statistical methods, machine learning methods, especially Random Forest (RF), have shown strong competitiveness in spatial prediction because they can model nonlinear relationships and covariate interactions by aggregating multiple decision trees built from bootstrap samples (Breiman 2001; Misiuk and Brown 2023; Amin et al. 2024). However, in spatial applications, RF generally treats training observations as independent and does not directly incorporate local associations among neighboring samples. It is also difficult for RF to adapt its learning process to the local environmental conditions around different prediction locations (Jemeļjanova et al. 2024; Xie et al. 2025; Fouedjio and Arya 2024). This becomes limiting when the target variable is shaped by neighborhood structure, local heterogeneity, or surrounding sample configuration (Sinha et al. 2019; Luo et al. 2023).

A third stream uses similarity, proximity, or local weighting to make spatial predictions more locally adaptive. Beyond weighted averaging, similarity and proximity have also been incorporated directly into model structures. For example, kernel-based learning represents nonlinear relationships through kernel functions and sample similarities (Smola and Schölkopf 2004), while locally weighted learning uses distance- or similarity-based weights to fit query-specific local models (Atkeson et al. 1997). In spatial analysis, geographically weighted regression addresses spatial nonstationarity by calibrating local regression models with spatially varying weights (Brunsdon et al. 1996). More recently, similarity and geographically weighted regression (SGWR) combined attribute similarity with geographical proximity in local weight construction (Lessani and Li 2024). Similar ideas have been extended to machine learning through geographical and geo-aware random forests, which use local or spatially partitioned RF models to address spatial heterogeneity (Georganos et al. 2021; Xie et al. 2025). However, these studies do not explicitly integrate spatial pattern features, covariate-space similarity search, and local nonlinear ensemble learning.

Pattern-based similarity methods further extend this local perspective by describing not only whether two locations have similar covariate values, but also whether their surrounding spatial contexts are similar. Pattern-based similarity methods provide another way to use spatial information. These methods quantify similarity among locations using environmental covariates and local spatial context (Zhu et al. 2018; Song 2023). For example, geographically optimal similarity (GOS) uses a product-Gaussian kernel to measure covariate similarity and predicts values through similarity-weighted averaging (Song 2023). Later studies introduced local structural attributes, such as geographic complexity and outlier strength, to better describe spatial context (He et al. 2024; Ren et al. 2025). However, these spatial features are usually used only for similarity calculation, and prediction still mainly relies on weighted averaging. This limits their ability to represent complex nonlinear relationships.

To address this gap, we propose Pattern Similarity Learning (PSL), which links pattern-based spatial features with similarity-weighted local Random Forest prediction. PSL augments each original covariate with geocomplexity and positive and negative outlier-strength features, expanding the original (n) covariates, where (n) denotes the number of explanatory variables, into a (4n)-dimensional pattern-enhanced feature space. For each prediction location, PSL first computes similarity in the original covariate space and selects the most similar training samples as a candidate pool. It then fits a local Random Forest using the pattern-enhanced features, with similarity scores used as sample weights. Therefore, the contribution of PSL is not the general use of similarity in model fitting, but the integration of covariate-space similarity search, pattern-derived spatial feature representation, and similarity-weighted local Random Forest learning in a single spatial prediction model.

Pattern Similarity Learning (PSL) Model

Concept of Pattern Similarity Learning (PSL)

Pattern similarity learning (PSL) is a local learning model for predicting continuous spatial variables (Fig. 1). In PSL, the local spatial pattern is described by three additional features for each covariate: geocomplexity, positive outlier strength, and negative outlier strength. These features indicate whether the covariate is locally stable, spatially heterogeneous, or unusually high or low compared with nearby locations. For each target location, PSL first uses the original covariates to find the most similar training samples. It then uses these selected samples, together with the expanded spatial pattern features, to train a local Random Forest. Instead of simply averaging the values of similar samples, PSL uses them to build a local nonlinear model for each prediction location.

Calculation Process of PSL

Construction of Spatial Pattern Features

The first step of PSL is to expand each original explanatory variable into a set of spatial pattern features so that both the variable value itself and its local spatial context can be represented. For the $k-th$ explanatory variable $\mathrm{x_k}$ at

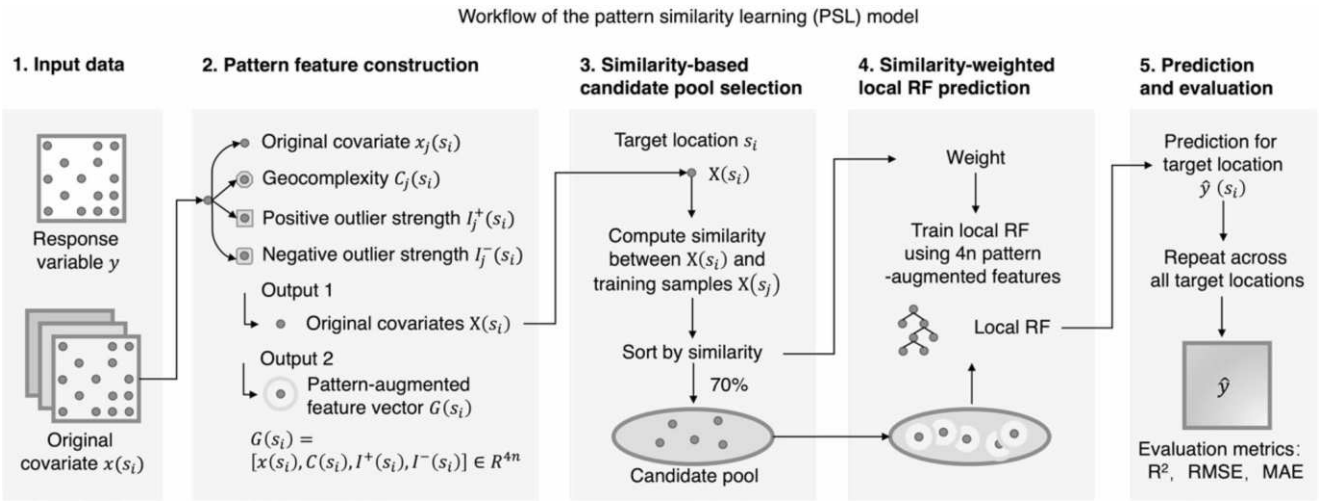

Fig. 1 Overview of the pattern similarity learning (PSL) model, from spatial pattern feature construction through candidate pool selection to similarity-weighted local random forest prediction

location s_i, a common m-nearest-neighbour set $\mathrm{N}(S_i)$ is first defined, and the same neighborhood window is used to derive three additional features: geocomplexity, positive outlier strength, and negative outlier strength. In this study, the number of neighbors is set to $\mathrm{m} = 12$, which is sufficient to capture fine-scale local spatial heterogeneity while also providing a relatively stable local reference for estimating outlier strength.

For each explanatory variable, three derived spatial features are calculated at every location. First, geocomplexity is used to quantify the degree of local spatial heterogeneity of the variable within the neighbourhood (Zhang et al. 2023). The geocomplexity index is defined as

$$P_k(s_i) = -\frac{1}{m}\sum_{j=1}^{m}\left[z_k(s_i)\, z_k(s_j) + \frac{1}{v_j}\sum_{l=1}^{v_j} z_k(s_j)\, z_k(s_l)\right] \quad (1)$$

where $z_k(s_i)$, $z_k(s_j)$, and $z_k(s_l)$ are the standardized values of the $k-th$ explanatory variable at the focal location, its neighbour, and the shared neighbour, respectively. Here, m denotes the fixed number of nearest neighbours used to define the neighbourhood of the focal location s_i. In contrast, v_j denotes the number of shared neighbours between the focal location s_i and one of its neighbours s_j. Thus, m controls the overall neighbourhood size, whereas v_j is a pair-specific quantity used in the shared-neighbour term of the geocomplexity calculation. The index is then rescaled using min–max normalization, yielding

$$C_k(s_i) = \frac{P_k(s_i) - \min_i P_k(s_i)}{\max_i P_k(s_i) - \min_i P_k(s_i)},\ C_k(s_i) \in [0,1] \quad (2)$$

Higher values of $C_k(s_i)$ indicate that the location lies in a transitional zone with stronger local spatial heterogeneity, where the variable changes rapidly over short distances; lower values of $C_k(s_i)$ indicate that the surrounding neighbourhood is relatively homogeneous (Zhang et al. 2023).

Next, outlier strength is used to characterize whether a location shows an unusually high or low value relative to its local neighbourhood (Ren et al. 2025). Let $\mu_k(s_i)$ and $\sigma_k(s_i)$ denote the mean and standard deviation of variable x_k within the neighbourhood $N(s_i)$, respectively. The positive outlier strength and negative outlier strength are then defined as follows:

$$I_k^+(s_i) = \max\left(0, x_k(s_i) - \mu_k(s_i) - 2\sigma_k(s_i)\right) \quad (3)$$

$$I_k^-(s_i) = \max\left(0, \mu_k(s_i) - 2\sigma_k(s_i) - x_k(s_i)\right) \quad (4)$$

When the value of a variable at a given location falls within the normal range of variation in its neighbourhood, both $I_k^+(s_i)$ and $I_k^-(s_i)$ are equal to 0. $I_k^+(s_i)$ takes a positive value only when the variable exceeds the neighbourhood mean by more than two standard deviations, whereas $I_k^-(s_i)$ takes a positive value only when the variable falls below the neighbourhood mean by more than two standard deviations. Therefore, $I_k^+(s_i)$ and $I_k^-(s_i)$ jointly encode both the direction and magnitude of local anomalies.

Finally, the original n explanatory variables are combined with the corresponding geocomplexity, positive outlier strength, and negative outlier strength for each variable to form the expanded feature vector at location s_i.

$$G(s_i) = \left(x_1, \ldots, x_n, C_1, \ldots, C_n, I_1^+, \ldots, I_n^+, I_1^-, \ldots, I_n^-\right)^T \in R^{4n} \quad (5)$$

Thus, the original n-dimensional explanatory variable space is expanded into a $4n$-dimensional pattern feature space, adding information on local spatial heterogeneity and anomalousness. For new prediction locations, neighbourhoods are defined from the available reference covariate locations, geocomplexity is normalized using reference-data parameters, and positive/negative outlier strength is calculated from local covariate means and standard deviations. These enhanced features are then used in the local Random Forest, allowing PSL to learn from both covariate values and their spatial context.

Similarity-Based Candidate Pool Construction

The second step of PSL is to identify, for each prediction location, a set of the most relevant candidate samples from the training set so as to localize the subsequent modelling process. In the updated PSL model, this step is performed in the original covariate space rather than in the pattern-enhanced feature space constructed in Sect. 2.2.1. This design separates similarity search from local learning by using the original covariates to identify the most similar training samples and reserving the pattern-enhanced features for the subsequent local random forest modelling.

For any location s, let its original covariate vector be denoted by $x(s) \in R^n$. Before similarity calculation, each covariate is standardized using the mean and standard deviation of the training data, so as to remove the influence of differences in scale and value range among variables. For a prediction location s_0 and a training location s_i, their similarity in the original covariate space is defined using a product-Gaussian kernel (Song 2023):

$$S(s_i, s_0) = \prod_{k=1}^{n} \exp\left(-\frac{(x_k(s_i) - x_k(s_0))^2}{2h^2}\right) \tag{6}$$

where $x_k(s_i)$ and $x_k(s_0)$ denote the standardized values of the $k-th$ original covariate for the training location and the prediction location, respectively, and h is the common kernel bandwidth. In this study, all original covariates are standardized and a common bandwidth of $h = 1.0$ is used. Under this setting, similarity decreases smoothly as differences in original covariate values increase.

After computing the similarity between s_0 and all training samples, the training samples are ranked in descending order of $S(s_i, s_0)$, and the top proportion κ of the most similar samples is retained to form the candidate pool $L(s_0)$ for prediction location s_0. This is defined as follows:

$$L(s_0) = \{s_i \in D_{train} : S(s_i, s_0) \geq Q_{1-\kappa}\left(\{S(s_j, s_0)\}_{j=1}^{N_{train}}\right)\} \tag{7}$$

Here, D_{train} denotes the training sample set, N_{train} is the total number of training samples, and $Q_{1-\kappa}$ is the $1-\kappa$ quantile of the similarity distribution. In this study, $\kappa = 0.7$, meaning that, for each prediction location, the top 70% of training samples with the highest similarity are retained as the candidate pool. The similarity scores of these retained candidate samples are then carried forward to the next step as observation-level weights in the local Random Forest modelling process.

Similarity-Weighted Local Random Forest Prediction

The third step of PSL is to fit a local Random Forest on the selected candidate pool for each prediction location. In this step, PSL follows the local learning stream of the model: the candidate pool and similarity scores are inherited from the original-covariate similarity search in Sect. 2.2.2, whereas model fitting is conducted in the pattern-enhanced feature space constructed in Sect. 2.2.1. Thus, similarity-based sample selection and local non-linear learning are separated but connected within the same prediction model.

For each prediction location s_0, PSL constructs a separate local model whose training samples are drawn from the candidate pool $L(s_0)$ (Atkeson et al. 1997). The input variables are the $4n$-dimensional pattern feature vectors $z(s_i)$, and the response variable is the observed value $y(s_i)$ at training location s_i. The similarity scores $S(s_i, s_0)$ computed in Sect. 2.2.2 are directly used as case weights in the local Random Forest, so that samples more similar to the target location exert greater influence on local model fitting.

Let the local Random Forest contain B regression trees. For prediction location s_0, the model is trained only on the candidate pool $L(s_0)$, using the corresponding pattern-enhanced feature vectors and similarity-derived weights. The final prediction is then obtained by averaging the predictions from all trees:

$$\widehat{y}_L(s_0) = \frac{1}{B}\sum_{b=1}^{B} T_b(z(s_0)) \tag{8}$$

where $T_b(\cdot)$ denotes the b-th regression tree, and $z(s_0)$ is the $4n$-dimensional pattern feature vector of the prediction location. In this study, the local Random Forest is trained with $B = 100$ trees. Through this design, PSL uses similarity in the original covariate space to localize the training samples, and exploits the expanded pattern feature space to learn complex non-linear relationships for each target location.

Cross-Validation Evaluation

To evaluate predictive performance, we used 5-fold spatial cross-validation (Roberts et al. 2017). Spatial folds were generated by hexagonal blocking (50 km cells, EPSG:5070), with hexagons assigned to folds to balance sample counts (Valavi et al. 2018). For each fold, SOC observations from the validation fold were excluded from model training, similarity-based candidate selection, and local Random Forest fitting. Spatial pattern features were derived only from explanatory covariates and spatial locations, which are available across the prediction domain in practical mapping, and thus represent covariate-derived rather than response-derived spatial contextual information. Standardization parameters for similarity calculation and model fitting were estimated from the training data and applied to the validation data. Five baseline models were included, forming a six-model comparison with PSL. After aggregating predictions from all test samples, performance was evaluated using R^2, RMSE, and MAE.

In addition to prediction accuracy, residual spatial autocorrelation was evaluated for each model using global Moran's I to assess whether the model better accounted for spatial structure (Liu et al. 2026; Zhang et al. 2024). The test was based on the held-out predictions from the 5-fold spatial cross-validation. A row-standardized m-nearest-neighbour spatial weight matrix was constructed from the sampling coordinates, with $m = 12$ to match the neighbourhood size used for spatial pattern feature construction. Statistical significance was assessed using 999 random permutations. Lower Moran's I values indicate less remaining spatial structure in model errors.

Case Study: Spatial Prediction of Soil Organic Carbon Across the Conterminous United States Using PSL

The conterminous United States (CONUS), consisting of the 48 contiguous states, was used as the case study to evaluate the proposed PSL model. CONUS spans a broad geographic extent and exhibits strong environmental gradients, covering diverse climate–topography–vegetation combinations, from humid to arid regions, from plains to mountains, and from cropland to forest and grassland. Across this domain, soil-forming environments, vegetation productivity, and surface processes vary markedly among regions, and provide an appropriate setting for testing the applicability of spatial prediction methods under large-scale and geographically complex conditions (Cao et al. 2019).

Response Variable: Soil Organic Carbon (SOC)

In this study, soil organic carbon (SOC) content was used as the response variable. SOC is a key indicator of soil quality, nutrient retention, and terrestrial carbon sequestration, and is therefore important for agricultural management, ecological restoration, and global change research (Minasny et al. 2017; Amelung et al. 2020; Adhikari and Hartemink 2016). The SOC point data were obtained from the NCSS Soil Characterization Database (Lab Data Mart) of the U.S. Department of Agriculture (USDA) National Cooperative Soil Survey. We extracted SOC observations from the 0–30 cm layer within the conterminous United States (CONUS), yielding 7,940 valid sampling locations. SOC was expressed as percentage content (%), with a mean of 1.58% and a range of 0.02%–47%. Its distribution is shown in Fig. 2. The large numerical dispersion and spatial complexity of SOC across CONUS provide a suitable empirical basis for evaluating PSL under large-scale and geographically complex conditions.

Explanatory Variables and Data Preprocessing

To represent major environmental controls on soil organic carbon (SOC), ten explanatory variables were selected to cover topography, climate, and vegetation productivity. Specifically, Slope, Sin(Aspect), Cos(Aspect), and TPI describe terrain conditions; Temperature, Precipitation, Solar radiation, and Wind speed represent regional moisture–energy conditions; and Biomass and NPP indicate vegetation carbon storage and potential organic matter input. These variables are widely recognized as important controls on SOC variation across landscapes (Lamichhane et al. 2019; Ding et al. 2025). Their spatial distributions and basic properties are presented in Fig. 3; Table 1. Biomass is also used in Fig. 4 to illustrate how derived spatial pattern features differ from original variable values. TRI, NDVI, EVI, and Elevation were initially considered but removed after pairwise correlation analysis and VIF screening to reduce redundancy and multicollinearity. Although Wind speed showed weak pairwise correlation with SOC, its low VIF justified retention because weak marginal correlation does not necessarily exclude nonlinear effects or interactions (James et al. 2013).

All explanatory variables were acquired through Google Earth Engine and resampled to a common spatial resolution of 1000 m. Temperature and Precipitation were derived from the PRISM Norm91m dataset, using the multi-year mean of the monthly tmean band and the annual accumulation of the monthly ppt band, respectively. Solar radiation was derived from the multi-year mean of the srad band in DAYMET V4, and Wind speed was derived from

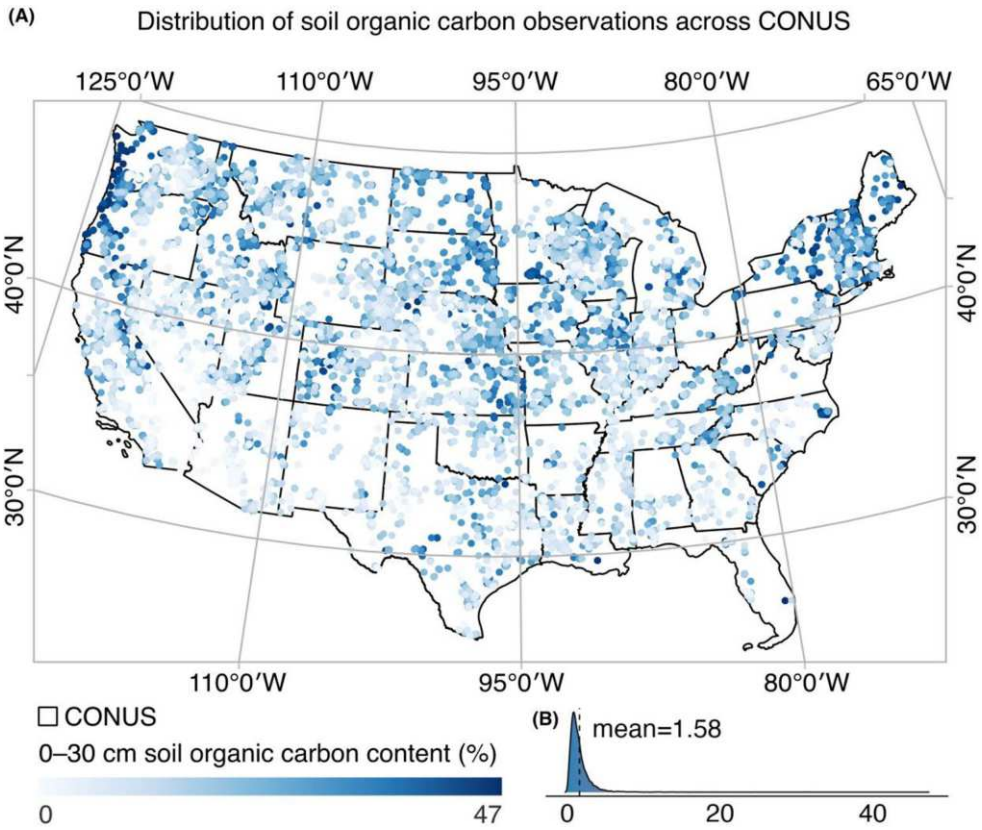

Fig. 2 Panel A shows the spatial distribution of the 7,940 soil organic carbon (SOC) observations used in this study across the conterminous United States (CONUS). Point colour represents SOC content (%) in the 0–30 cm layer using a light-to-dark sequential scale. Panel B shows the frequency distribution of SOC values, with the dashed line indicating the mean value (1.58%)

the multi-year mean of the 10 m wind speed band vs. in GRIDMET. Biomass was obtained from the agb band of the NASA/ORNL biomass carbon density v1 dataset, and NPP was derived from the multi-year mean of the Npp band in MODIS MOD17A3HGF V6.1.

The elevation data were obtained from the USGS 3DEP 10 m DEM and exported in EPSG:5070 at a 1000 m resolution to match the modelling resolution. Topographic variables, including Slope, Topographic Position Index (TPI), Sin(Aspect), and Cos(Aspect), were then derived from this 1000 m elevation layer in QGIS. Aspect was decomposed into Sin(Aspect) and Cos(Aspect) to avoid the numerical discontinuity of the original angular measure at the 0°/360° boundary, thereby improving the stability of model fitting (Amatulli et al. 2018).

Before modelling, all explanatory variables were standardized using z-scores to eliminate the effects of differences in scale and value range on similarity measurement and model training (James et al. 2013). To reduce the potential influence of multicollinearity, a correlation matrix was first computed for all explanatory variables, and variance inflation factors (VIFs) were then used to further assess variable redundancy (Dormann et al. 2013). All pairwise correlation coefficients were below 0.8 and all VIF values were below 5 (Fig. 5), indicating that no strong collinearity was present among the variables.

Experiment Design

To evaluate the proposed PSL model, six models were compared under the same 5-fold spatial cross-validation scheme: LM, RF, GOS, SGWR, PS, and PSL. LM and RF were used as conventional global baselines, representing global linear and nonlinear modelling with the original 10 environmental covariates. To ensure a fair comparison, the RF baseline was tuned under the same five-fold spatial cross-validation by searching mtry = 2, 3, 4, 5, 6, 8, 10; ntree = 500, 1000, 1500; and min.node.size = 1, 3, 5, 10. The best RF configuration was ntree = 1500, mtry = 2, and min.node.size = 1.

GOS was included as a traditional similarity-based model that predicts through similarity-weighted averaging in the original covariate space (Song 2023). SGWR was added as a similarity-based local regression baseline because it combines attribute similarity and geographical proximity in local weight construction. To isolate the role of spatial pattern features, PS used the expanded 40-dimensional pattern feature space but retained the similarity-weighted averaging mechanism. In contrast, PSL first computed similarity in the original 10-dimensional covariate space to select local candidate samples, and then fitted a similarity-weighted local Random Forest using the 40-dimensional pattern-enhanced features. The detailed model configurations are summarized in Table 2.

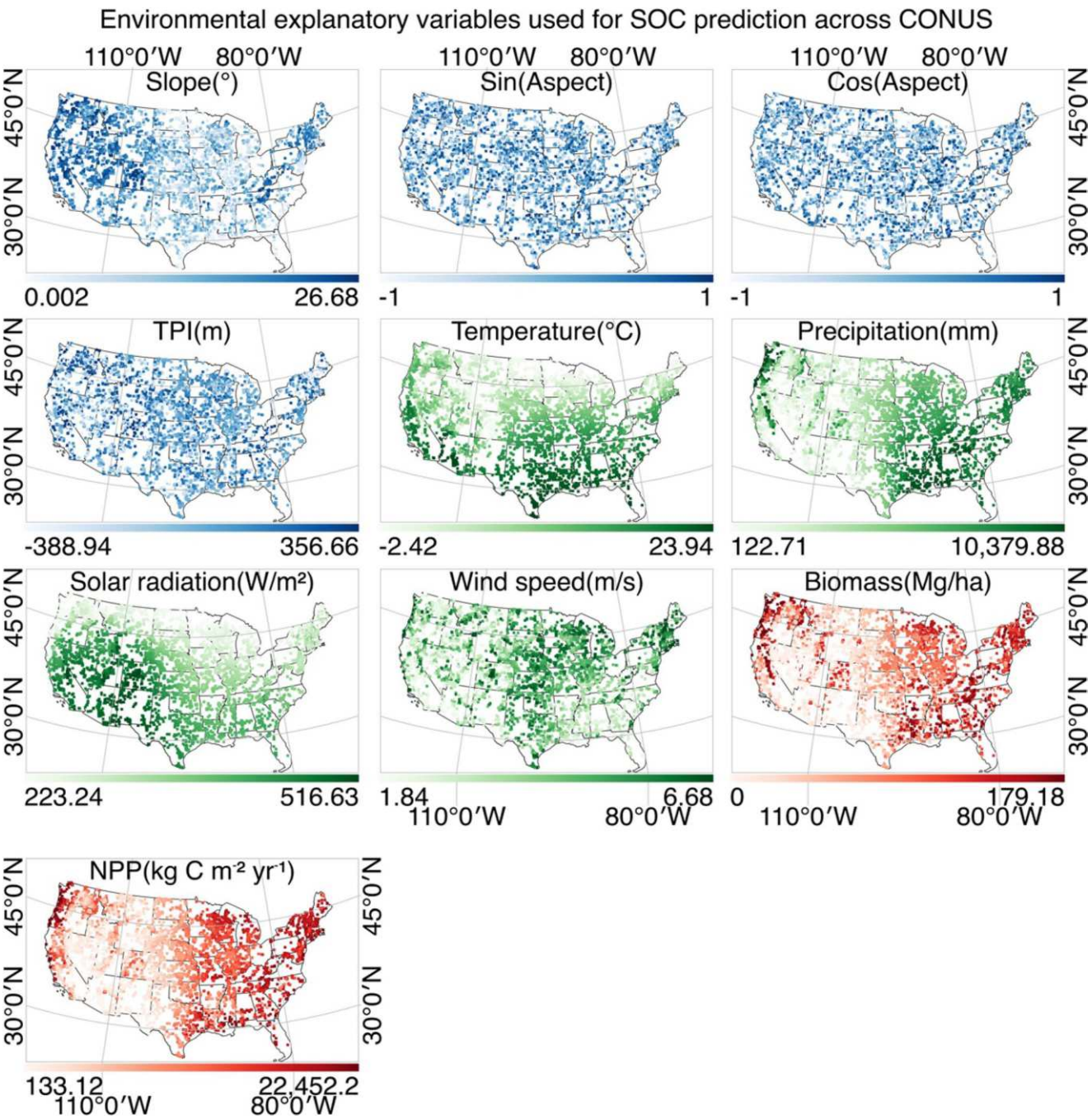

Fig. 3 Spatial patterns of the ten environmental explanatory variables used for SOC prediction across the conterminous United States (CONUS). The variables include four topographic predictors (Slope, Sin(Aspect), Cos(Aspect), and TPI), four climatic predictors (Temperature, Precipitation, Solar radiation, and Wind speed), and two productivity-related predictors (Biomass and NPP). All layers were resampled to a common spatial resolution of 1000 m

Results

Cross-Validation Prediction Accuracy

PSL achieved the highest overall predictive accuracy among the six models under 5-fold spatial cross-validation, with performance gains shown relative to all comparison models (Figs. 6 and 7). Overall, PSL achieved the highest R^2 (0.256) and the lowest RMSE (1.468) and MAE (0.699), making it the best-performing model among the six. RF was the baseline model closest to PSL in performance, but its R^2 was still 0.017 lower, with slightly higher RMSE and MAE. The fold-level paired differences shown in Fig. 7 further indicate that PSL achieved modest improvements over RF, with

Table 1 Environmental explanatory variables used for SOC prediction

Category	Covariate	Unit	Modelling resolution (m)	Description	Data source
Topography	Slope	°	1000	Represents terrain steepness.	USGS 3DEP DEM
	Sin(Aspect)	/	1000	Sine transformation of aspect, used to represent the east–west component of slope orientation and to avoid the discontinuity of angular variables.	USGS 3DEP DEM
	Cos(Aspect)	/	1000	Cosine transformation of aspect, used to represent the north–south component of slope orientation and to avoid the discontinuity of angular variables.	USGS 3DEP DEM
	TPI	m	1000	Represents the relative topographic position of a location with respect to its surrounding neighbourhood.	USGS 3DEP DEM
Climate	Temperature	°C	1000	Multi-year mean annual temperature, representing regional thermal conditions.	PRISM(1991–2020)
	Precipitation	Mm	1000	Multi-year mean annual precipitation, representing regional water input conditions.	PRISM(1991–2020)
	Solar radiation	W/m²	1000	Multi-year mean shortwave radiation, representing surface energy input.	DAYMET V4(1980–2024)
	Wind speed	m/s	1000	Multi-year mean wind speed, representing near-surface atmospheric dynamic conditions.	GRIDMET(1979–2026)
Vegetation	Biomass	Mg/ha	1000	Aboveground biomass carbon density, representing vegetation carbon storage and potential organic matter input.	NASA ORNL Biomass Carbon Density(2010)
	NPP	kg C m^{-2} yr^{-1}	1000	Multi-year mean net primary productivity, representing vegetation carbon fixation capacity.	MODIS MOD17A3HGF(2001–2025)

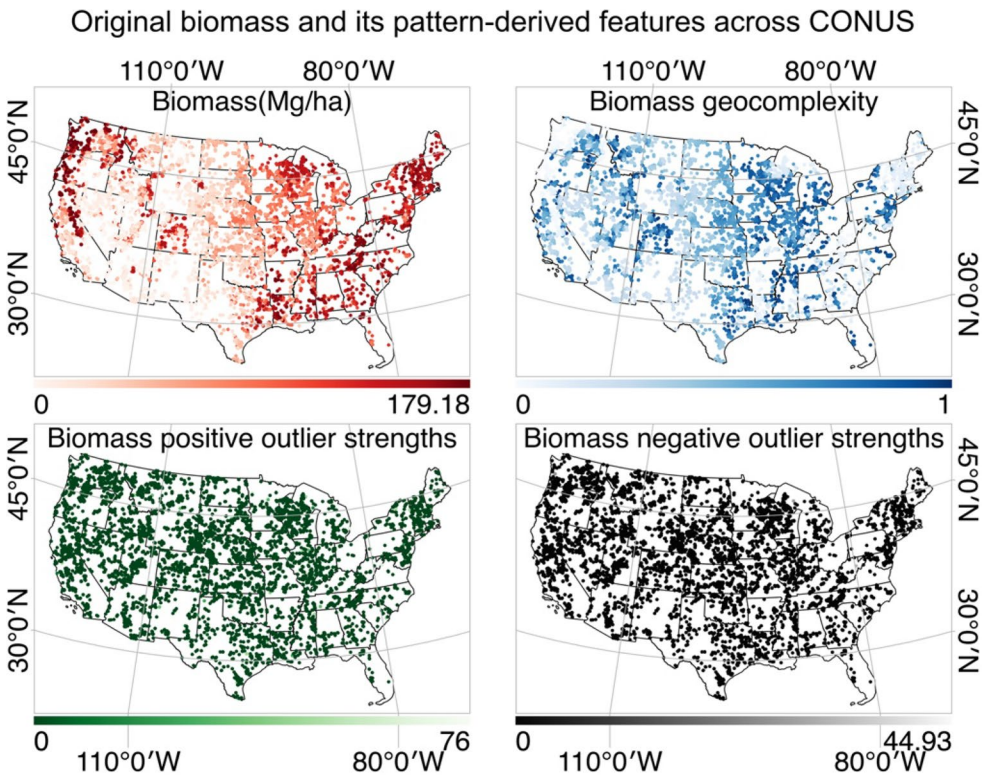

Fig. 4 Biomass is used here as an example to illustrate the spatial patterns of an original explanatory variable and its three pattern-derived features across the conterminous United States (CONUS). The derived features include geocomplexity, positive outlier strength, and negative outlier strength, which together represent local spatial heterogeneity and local anomalousness associated with the variable

mean gains of 0.017 in R^2, 0.017 in RMSE reduction, and 0.012 in MAE reduction across the five spatial folds. These intervals include zero, indicating that the fold-level differences were not statistically significant at the 95% confidence level, although PSL showed positive mean improvements and outperformed RF in four of five folds for each metric.

LM performed worse than RF, whereas SGWR, GOS, and PS showed substantially lower overall accuracy. In

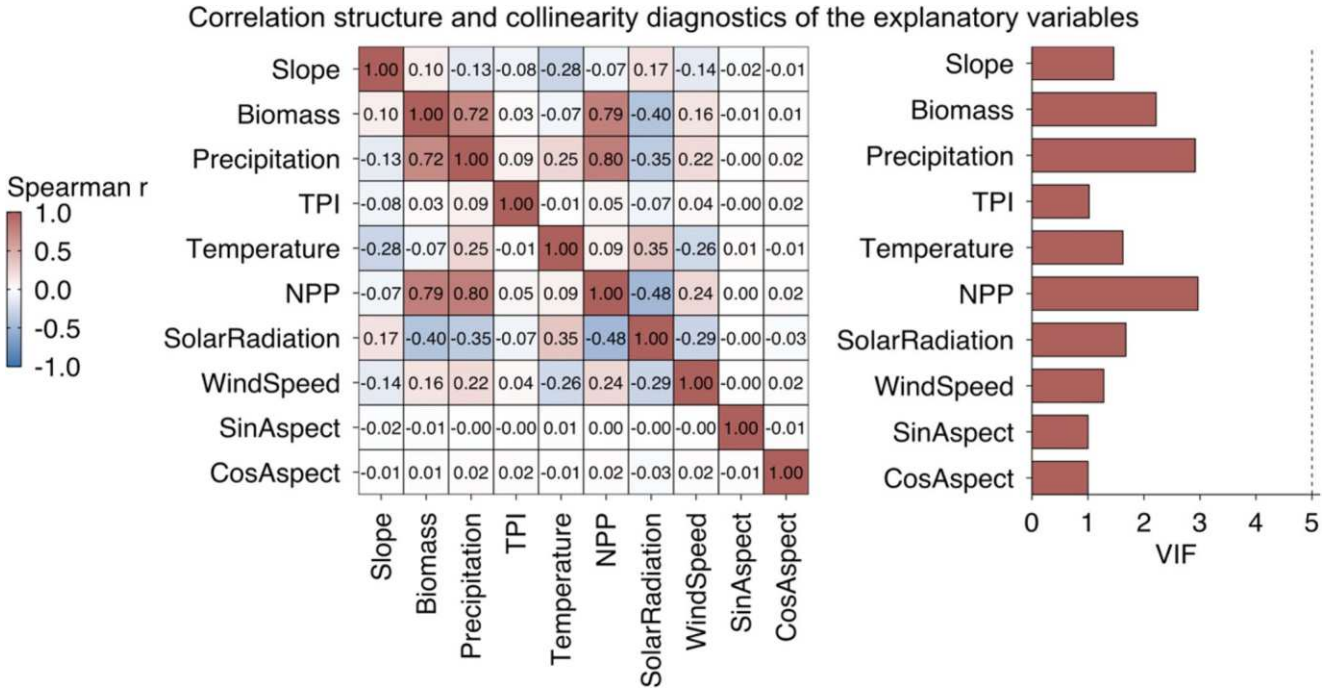

Fig. 5 Correlation structure and collinearity diagnostics for the ten explanatory variables used for SOC prediction. Panel A shows the pairwise correlation matrix, and Panel B shows the variance inflation factor (VIF) values for each variable. All pairwise correlation coefficients are below 0.8 and all VIF values are below 5, indicating the absence of strong multicollinearity

particular, the R^2 values of SGWR and GOS were relatively low, indicating that simple similarity-weighted averaging is insufficient to effectively capture SOC variation. PSL predictions were generally closer to the 1:1 reference line within the main concentration range of SOC values, although high SOC values remained underestimated (Fig. 6).

Spatial Prediction and Residual Patterns

The six models differed in their spatial residual patterns, absolute residual magnitudes, and predicted SOC patterns at the validation locations (Figs. 8, 9 and 10). Overall, residuals from all models exhibited a certain degree of spatial heterogeneity, but their spatial organization differed substantially. The residuals of SGWR, GOS, and LM formed more obvious spatial clusters in multiple regions (Fig. 8). PS improved upon the former two models to some extent, but still retained many localized patches of error. In contrast, the residuals of RF and PSL were more spatially dispersed, with PSL showing a relatively more even spatial pattern, suggesting that it more effectively absorbed prediction bias associated with local environmental variation.

Error magnitude comparisons further show the advantage of PSL (Fig. 9). PSL had the lowest median absolute residual (Med = 0.428) and the smallest interquartile range (IQR = 0.612), followed by RF (Med = 0.449; IQR = 0.636). PS, SGWR, and GOS showed larger median errors or wider spreads, while LM had the largest median residual (Med = 0.560; IQR = 0.704). Overall, PSL produced the most compact absolute residual distribution among the six models.

The spatial distributions of observed and cross-validated predicted SOC values further illustrate the prediction patterns of RF and PSL at the validation locations (Fig. 10). Both RF and PSL reproduced the broad spatial distribution of SOC across CONUS. Compared with RF, PSL showed similar large-scale spatial patterns but produced local differences in predicted SOC at many validation locations. The PSL − RF difference map indicates where PSL generated higher or lower predictions than RF, and is interpreted as a comparison of prediction patterns rather than as independent evidence of broad spatial improvement.

In addition to the visual comparison, global Moran's I was used to quantify residual spatial autocorrelation (Table 3). All models showed significant positive spatial autocorrelation in their residuals, indicating that spatially structured errors remained after prediction. SGWR had the lowest Moran's I value (0.0734), followed by PSL (0.0837) and RF (0.1000). This suggests that PSL reduced residual spatial autocorrelation more effectively than RF, GOS, LM, and PS, although SGWR showed the weakest residual spatial clustering among the six models.

Table 2 Summary of the six models used in the comparative experiment

Model abbreviation	Full Name	Input feature space	Prediction mechanism	Core parameter settings
LM	Linear Model	Original 10-dimensional covariates	Global linear fitting	Default OLS fitting
RF	Random Forest	Original 10-dimensional covariates	Global non-linear ensemble learning	ntree = 1500; mtry = 2; min. node.size = 1
SGWR	Similarity and Geographically Weighted Regression	Original 10-dimensional covariates	Similarity- and geographically weighted local linear regression	geo_k = 300; h_attr = 1.0; ridge = 1e-6; min_n = 30
GOS	Geographically Optimal Similarity	Original 10-dimensional covariates	Local similarity-weighted averaging	kappa = 0.10
PS	Pattern Similarity	Spatial pattern features (40 dimensions)	Local similarity-weighted averaging	top_k = 12; h_pattern = 3
PSL	Pattern Similarity Learning	spatial pattern features (40 dimensions) for local learning	Similarity-weighted local Random Forest	h_x = 1.0; candidate pool = 70%; num.trees = 100; mtry = 6; min. node.size = 5

Sensitivity Analysis of κ and m Parameters

PSL showed generally low sensitivity to both the candidate-pool proportion parameter κ and the neighbourhood size m across the tested ranges (Fig. 11). Overall, model performance varied only slightly as the two parameters changed. Across κ = 0.05–1.00 and m = 4–52, R^2 varied mainly between 0.224 and 0.256, while RMSE and MAE remained within narrow ranges of approximately 1.470–1.500 and 0.692–0.702, respectively. These results indicate that PSL is generally stable with respect to both candidate-pool size and neighbourhood scale, and does not rely on a narrowly tuned parameter setting to achieve good predictive performance.

In terms of performance trends, R^2 increased initially and then fluctuated slightly as κ increased, with relatively high values around $\kappa \approx 0.50$–0.80. RMSE remained low across most κ values after $\kappa \approx 0.25$, whereas MAE was relatively stable over the middle range of κ. The sensitivity to m was also limited, although very small or very large neighbourhoods produced slightly less favourable results in some metrics. Considering all three metrics together, κ = 0.70 and m = 12 were retained as the final settings, providing a good balance between predictive accuracy, locality, and stability. Very small candidate pools may provide too few local samples for effective nonlinear learning, whereas overly broad pools may weaken locality.

Discussion

Among all the comparative models, PSL achieved the best overall predictive accuracy, and its improvement over RF and PS demonstrated two advantages. First, PSL improved R^2 over PS by 0.190, indicating that replacing similarity-weighted averaging with similarity-weighted local Random Forest learning substantially improves predictive performance. Although PS can identify the training samples most relevant to the target location through similarity ranking, its prediction is still essentially a weighted linear combination of observed response values, making it difficult to capture the non-linear relationships and covariate interactions that govern SOC variation (Breiman 2001; Cutler et al. 2007). In contrast, PSL retains the spatial selectivity of PS in sample selection, and uses a local random forest to adaptively model non-linear relationships at each prediction location, thereby achieving higher accuracy.

The improvement of PSL over global RF mainly arises from richer spatial pattern information and a more localized training strategy. First, pattern features such as geo-complexity and positive and negative outlier strength encode local spatial context beyond the original covariates (He et al. 2024; Ren et al. 2025). Different covariates may show different degrees of local heterogeneity within the same neighbourhood, and these differences are retained as complementary pattern information rather than treated as contradictions. This context is particularly useful in spatial transition zones and complex areas, where it helps the model adjust local splitting rules more effectively. Second, PSL does not fit a single model using all training samples. Instead, it restricts training to the candidate pool most similar to the target location, reducing sample heterogeneity and improving local fitting. Thus, PSL's advantage comes from both feature expansion and similarity-guided local learning, producing consistent improvements across validation folds and performance metrics.

The spatial results suggest that PSL partially improved the representation of local SOC variation and showed some local differences in predicted SOC patterns compared with RF. PSL residuals were also more spatially dispersed, suggesting that the method partially reduced local modelling bias under spatial heterogeneity. However, both RF and PSL still showed range compression, with overestimation of low

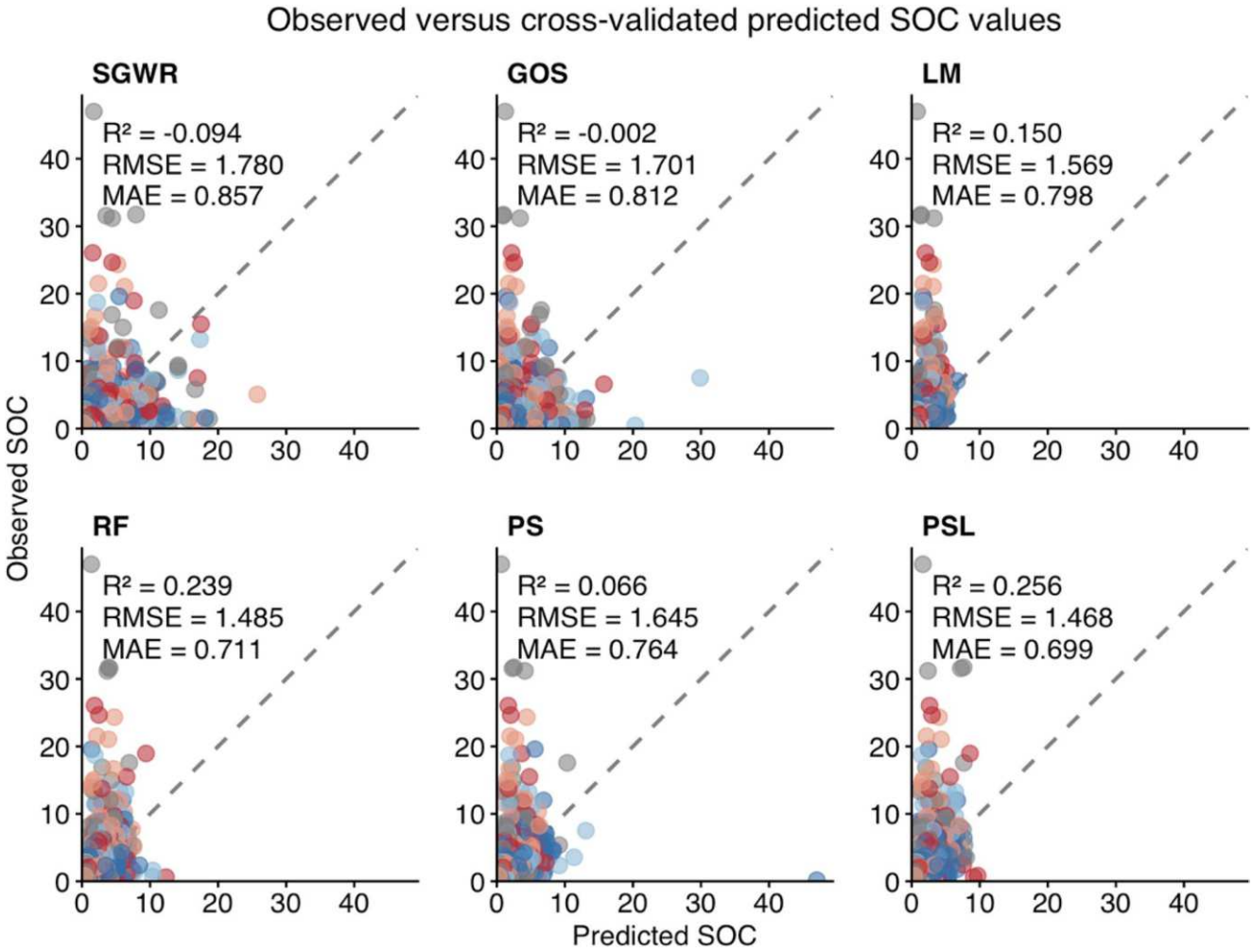

Fig. 6 Observed versus cross-validated predicted SOC values for the six models. Each panel compares observed and predicted SOC values, with the 1:1 line indicating perfect agreement. The figure summarizes prediction patterns and overall cross-validation performance across models

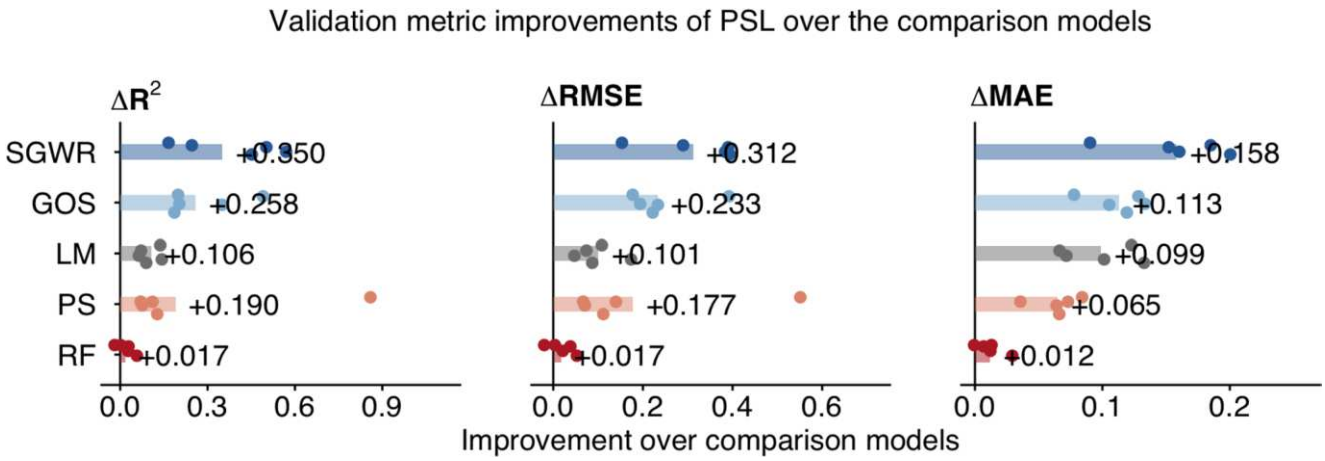

Fig. 7 Improvements of PSL over the comparison models in three validation metrics: ΔR^2, ΔRMSE, and ΔMAE. Each row corresponds to one comparison model. The horizontal bars represent the overall improvement calculated from the aggregated cross-validated predictions, whereas the dots represent the improvement obtained in each individual fold. Positive values indicate better performance of PSL than the corresponding comparison model, that is, higher R^2 and lower RMSE or MAE

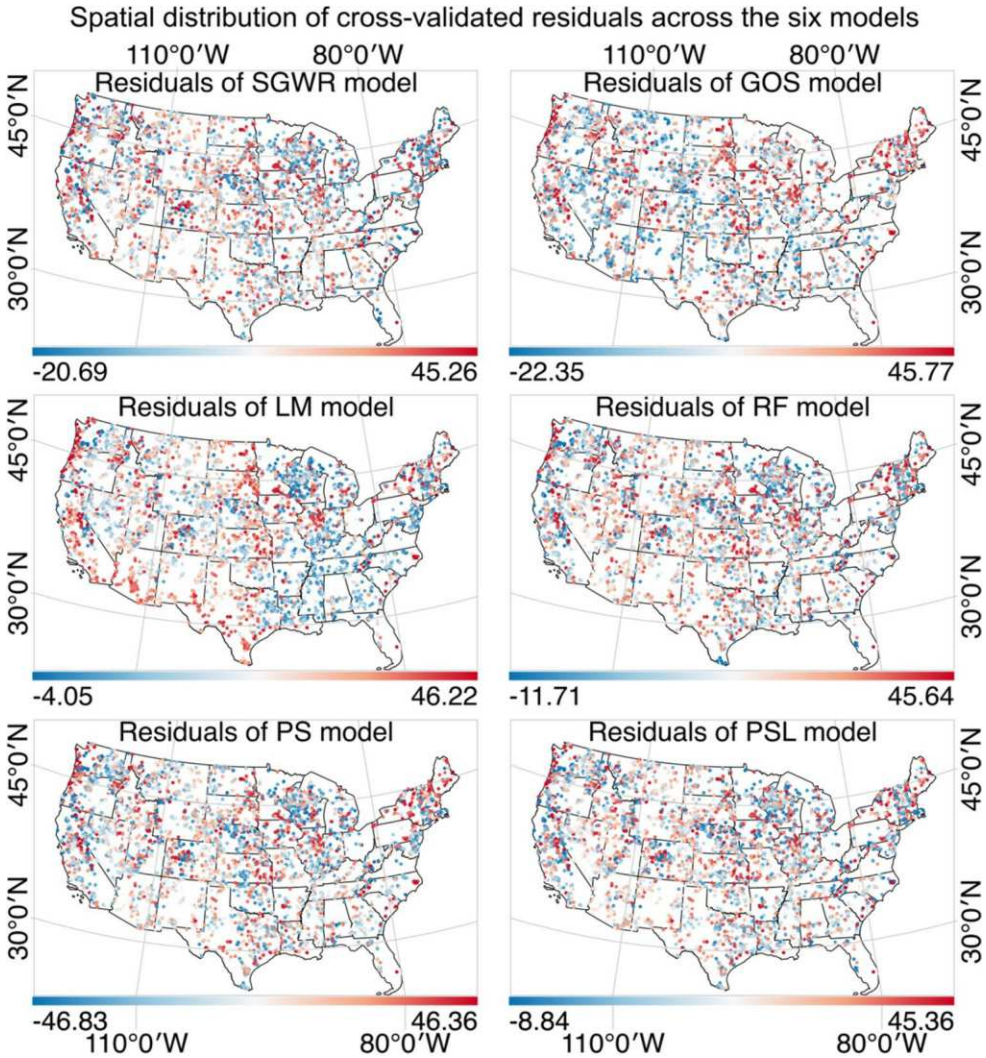

Fig. 8 Spatial distribution of cross-validated residuals for the six models across the conterminous United States (CONUS). Each panel shows the residual map for one model using the same spatial sample locations. The figure is used to compare the spatial organisation, clustering patterns, and local heterogeneity of model residuals among the six models

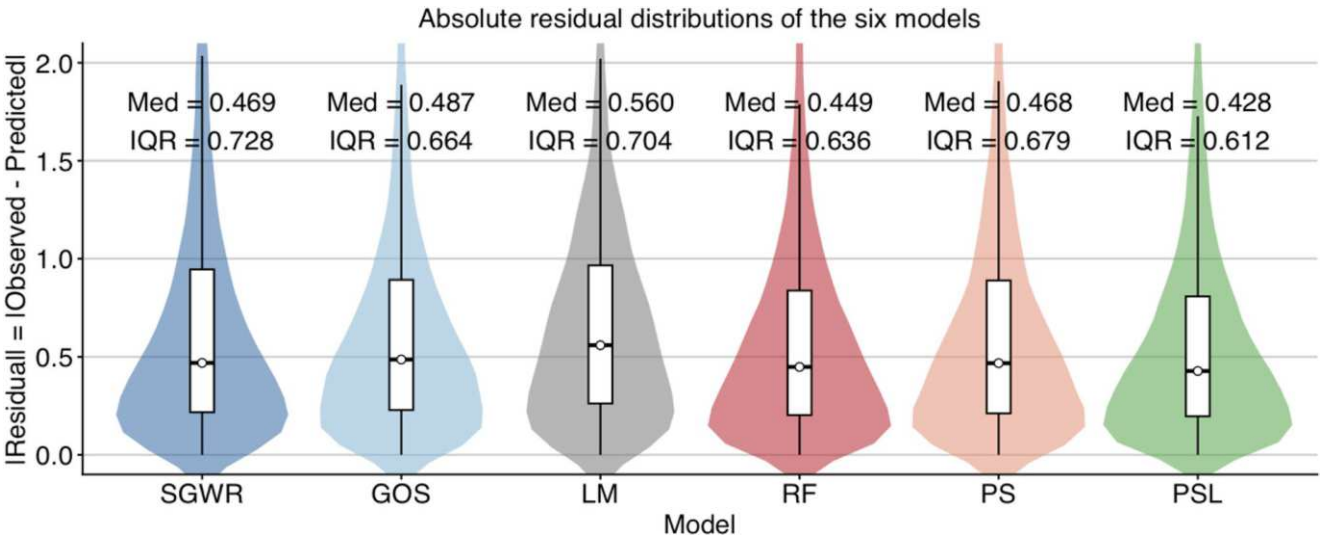

Fig. 9 Distribution of absolute residuals for the six models. Each violin plot summarizes the distribution of $|Observed - Predicted|$ values, with the embedded boxplot indicating the median and interquartile range (IQR). The figure is used to compare the central tendency, spread, and overall error concentration of the six models

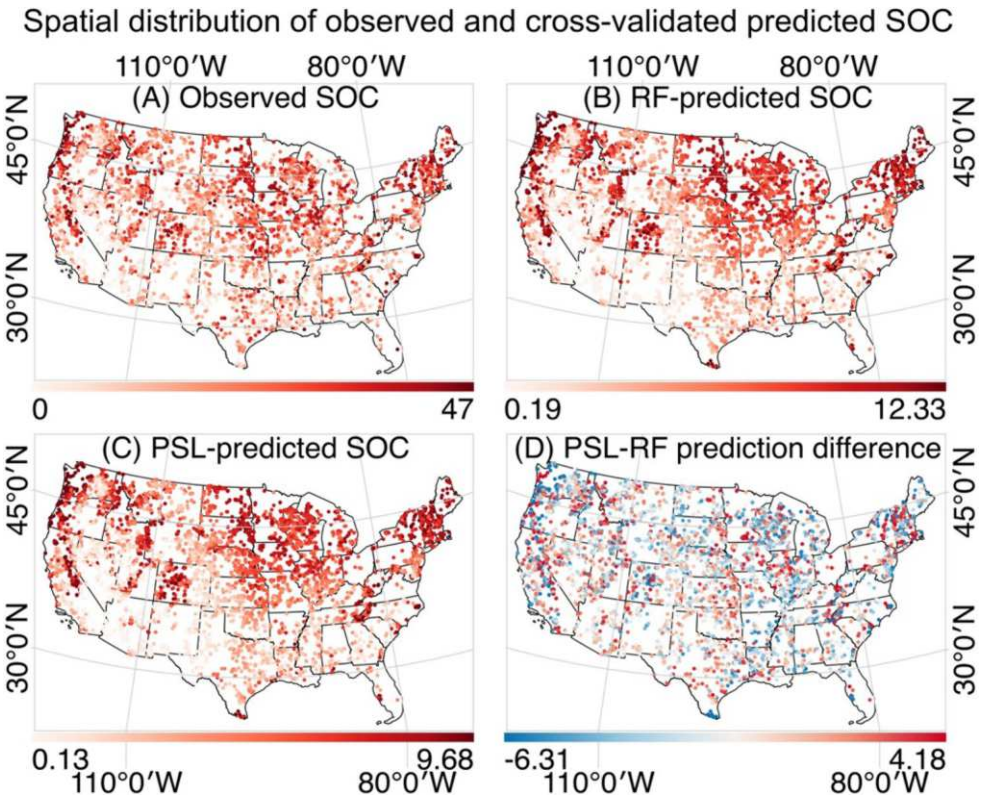

Fig. 10 Spatial distribution of observed and cross-validated predicted SOC across CONUS. Panel A shows observed SOC, Panels B and C show RF- and PSL-predicted SOC, respectively, and Panel D shows the PSL − RF prediction difference. The figure compares prediction patterns rather than independently validating spatial improvement

Table 3 Global Moran's I test for cross-validated residuals of the six models

Model	Moran's I	*p*-value
SGWR	0.0734	0.001
GOS	0.1180	0.001
LM	0.1890	0.001
RF	0.1000	0.001
PS	0.1180	0.001
PSL	0.0837	0.001

SOC values and underestimation of high SOC values, which is consistent with the known bias of Random Forest regression toward averaged predictions (Zhang and Lu 2012). Therefore, the advantage of PSL should be interpreted as achieving the strongest overall predictive accuracy while reducing residual spatial structure relative to RF and most other baselines, rather than being the best model for every individual diagnostic. The modest absolute R^2 also indicates that SOC variation at the CONUS scale is strongly constrained by missing covariates, measurement uncertainty, spatial scale mismatch between point SOC observations and 1000 m predictors, and the inherent complexity of SOC processes.

Thus, the practical value of PSL lies in its relative improvement over RF and its spatially adaptive modelling strategy, rather than in achieving a high absolute R^2. The main limitation of PSL is its computational cost, which mainly comes from fitting a separate local Random Forest

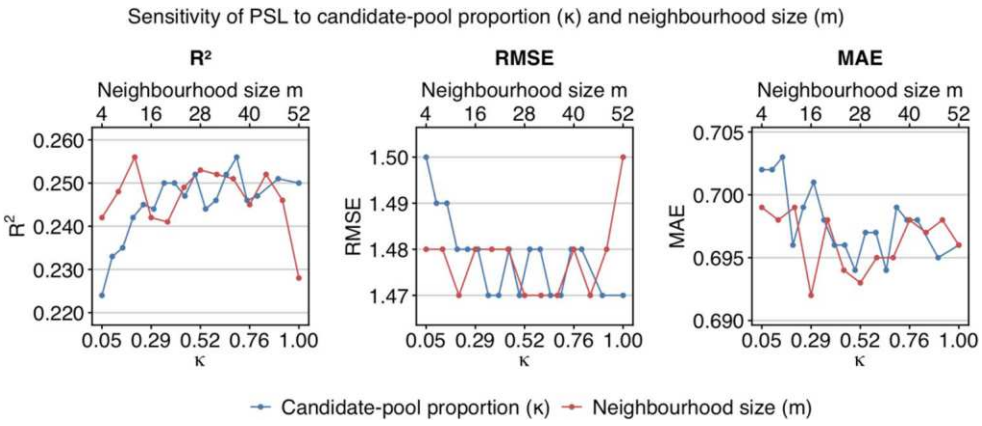

Fig. 11 Sensitivity of PSL to the candidate-pool proportion parameter (κ) and neighbourhood size (m). The three panels show changes in cross-validated R^2, RMSE, and MAE across the tested ranges of κ and m. Blue lines indicate the sensitivity to κ, and red lines indicate the sensitivity to m. The figure is used to assess whether PSL performance depends strongly on candidate-pool size or neighbourhood scale

for each prediction location. If N is the number of training samples, M is the number of prediction locations, n is the number of covariates, κN is the candidate-pool size, and B is the number of trees, the main cost approximately scales with $M\{O(Nn)+O(B\kappa Nlog(\kappa N))\}$. However, because the local models are independent across prediction locations, this process can be parallelized. For larger-scale mapping applications, efficiency could be improved through parallelization or by reducing the candidate-pool size in stable local environments. Future work could also incorporate richer soil, land-use, management, and process-related covariates, or apply spatial residual correction, such as ordinary kriging on PSL residuals, to further improve predictive performance (Matheron 1963; Hengl et al. 2004).

Conclusions

This study proposed Pattern Similarity Learning (PSL) for predicting continuous spatial variables. PSL expands the original explanatory variables with geocomplexity and positive and negative outlier strength, and combines similarity-based sample selection with local Random Forest learning. Using 5-fold spatial cross-validation with 7,940 SOC observations, PSL achieved the best pooled performance among the compared models. However, its improvement over the tuned RF baseline was modest and not statistically significant at the fold level. PSL also reduced residual spatial autocorrelation relative to RF and several other baselines.

Overall, PSL provides a potentially useful spatially adaptive approach for continuous environmental prediction, although its computational cost remains relatively high. Future work should focus on improving efficiency and incorporating residual correction or related extensions.

Author Contributions LHY was responsible for the methodology development, data processing, paper writing, and revision of this paper; SYZ was responsible for the design and methodology development of this paper; ZPC and YN were responsible for reviewing this paper and some data processing.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Data Availability The code and data for this study are archived on https://doi.org/10.6084/m9.figshare.33006431.

Declarations

Competing interests The authors declare no competing interests.

References

Adhikari K, Hartemink AE (2016) Linking soils to ecosystem services. A global review. Geoderma 262:101–111

Amatulli G, Domisch S, Tuanmu MN, Parmentier B, Ranipeta A, Malczyk J, Jetz W (2018) A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci data 5(1):180040

Amelung W, Bossio D, de Vries W, Kögel-Knabner I, Lehmann J, Amundson R, Bol R, Collins C, Lal R, Leifeld JJNC, Minasny B (2020) Towards a global-scale soil climate mitigation strategy. Nat Commun 11(1):5427

Amin G, Imtiaz I, Haroon E, Saqib NU, Shahzad MI, Nazeer M (2024) Assessment of machine learning algorithms for land cover classification in a complex mountainous landscape. J Geovisualization Spat Anal 8(2):34

Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artif Intell Rev 11(1):11–73

Atkinson PM, Lloyd CD (2007) Non-stationary variogram models for geostatistical sampling optimisation: An empirical investigation using elevation data. Comput Geosci 33(10):1285–1300

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical Anal 28(4):281–298

Cao B, Domke GM, Russell MB, Walters BF (2019) Spatial modeling of litter and soil carbon stocks on forest land in the conterminous United States. Sci Total Environ 654:94–106

Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88(11):2783–2792

Ding Z, Liu K, Grunwald S, Smith P, Ciais P, Wang B, Wadoux AMC, Ferreira C, Karunaratne S, Shurpali N, Yin X (2025) Advancing Soil Organic Carbon Prediction: A Comprehensive Review of Technologies, AI, Process-Based and Hybrid Modelling Approaches. Adv Sci 12(31):e04152

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1):27–46

Fouedjio F, Arya E (2024) Locally varying geostatistical machine learning for spatial prediction. Artif Intell Geosci 5:100081

Georganos S, Grippa T, Niang Gadiaga A, Linard C, Lennert M, Vanhuysse S, Mboga N, Wolff E, Kalogirou S (2021) Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. Geocarto Int 36(2):121–136

He W, Li L, Gao X (2024) Geocomplexity statistical indicator to enhance multiclass semantic segmentation of remotely sensed data with less sampling bias. Remote Sens 16(11):1987

Hengl T, Heuvelink GB, Stein A (2004) A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120(1–2):75–93

Hengl T, Heuvelink GB, Rossiter DG (2007) About regression-kriging: From equations to case studies. Comput Geosci 33(10):1301–1315

Heuvelink GBM, Webster R (2001) Modelling soil variation: past, present, and future. Geoderma 100(3–4):269–301

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning: with applications in R, vol 103. springer, New York

Jemeljanova M, Kmoch A, Uuemaa E (2024) Adapting machine learning for environmental spatial data-A review. Ecol Inf 81:102634

Keskin H, Grunwald S (2018) Regression kriging as a workhorse in the digital soil mapper's toolbox. Geoderma 326:22–41

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

Lessani MN, Li Z (2024) SGWR: similarity and geographically weighted regression. Int J Geogr Inf Sci 38(7):1232–1255

Liu H, Song Y, Yi W (2026) Degree of spatial interpretability. Int J Geogr Inf Sci 1–21

Luo P, Song Y, Zhu D, Cheng J, Meng L (2023) A generalized heterogeneity model for spatial interpolation. Int J Geogr Inf Sci 37(3):634–659

Matheron G (1963) Principles of geostatistics. Econ Geol 58:1246–1266

Minasny B, Malone BP, McBratney AB, Angers DA, Arrouays D, Chambers A, Chaplot V, Chen ZS, Cheng K, Das BS, Field DJ (2017) Soil carbon 4 per mille. Geoderma 292:59–86

Misiuk B, Brown CJ (2023) Improved environmental mapping and validation using bagging models with spatially clustered data. Ecol Inf 77:102181

Odeh IO, McBratney AB, Chittleborough DJ (1995) Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. Geoderma 67(3–4):215–226

Ren K, Song Y, Yu Q (2025) Second-dimension outliers for spatial prediction. Int J Geogr Inf Sci, 1–28

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929

Sinha P, Gaughan AE, Stevens FR, Nieves JJ, Sorichetta A, Tatem AJ (2019) Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling. Comp Environ Urban Syst 75:132–145

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222

Song Y (2023) Geographically optimal similarity. Math Geosci 55(3):295–320

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2018) blockCV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Biorxiv 357798

Xie Y, Nhu AN, Song XP, Jia X, Skakun S, Li H, Wang Z (2025) Accounting for spatial variability with geo-aware random forest: A case study for US major crop mapping. Remote Sens Environ 319:114585

Zhang G, Lu Y (2012) Bias-corrected random forests in regression. J Applied Statistics 39(1):151–160

Zhang Z, Song Y, Zhang P (2023) Geocomplexity explains spatial errors. Int J Geogr Inf Sci 37(8):1684–1712

Zhang Z, Li Z, Song Y (2024) On ignoring the heterogeneity in spatial autocorrelation: consequences and solutions. Int J Geogr Inf Sci 38(12):2545–2571

Zhu AX, Lu G, Liu J, Qin CZ, Zhou C (2018) Spatial prediction based on Third Law of Geography. Ann GIS 24(4):225–240