GIScience & Remote Sensing

ISSN: 1548-1603 (Print) 1943-7226 (Online) Journal homepage: www.tandfonline.com/journals/tgrs20

Integrating local spatial outliers into context-aware spatial prediction models

Jinyu Meng , Ziqin Zheng & Yongze Song

To cite this article: Jinyu Meng , Ziqin Zheng & Yongze Song (2026) Integrating local spatial outliers into context-aware spatial prediction models, GIScience & Remote Sensing, 63:1, 2700126, DOI: 10.1080/15481603.2026.2700126

To link to this article: https://doi.org/10.1080/15481603.2026.2700126

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 24 Jul 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tgrs20

GISCIENCE & REMOTE SENSING
2026, VOL. 63, NO. 1, 2700126
https://doi.org/10.1080/15481603.2026.2700126

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Integrating local spatial outliers into context-aware spatial prediction models

Jinyu Meng[a], Ziqin Zheng[b] and Yongze Song[c]

[a]School of Atmospheric Sciences, Nanjing University of Information Science and Technology, Nanjing, People's Republic of China; [b]College of Hydrology and Water Resources, Hohai University, Nanjing, People's Republic of China; [c]School of Design and the Built Environment, Curtin University, Perth, Australia

ABSTRACT

Improving spatial prediction remains a key challenge in geographical and environmental modeling, particularly in regions with strong spatial heterogeneity. This study proposes a local-outlier-based context-aware model, named LOCA, for precipitation spatial prediction. The novelty of LOCA lies not in spatial stratification or local modeling alone, but in formalizing local outliers of explanatory variables as continuous spatial contextual predictors and integrating them with spatial stratification and stratum-specific machine-learning models. Unlike conventional approaches that mainly treat outliers as errors or noise, LOCA uses positive and negative departures from the surrounding spatial background to represent local anomaly information. The model was evaluated through a case study of precipitation prediction in Qinghai Province, China, using CRU TS precipitation data, ERA5 meteorological variables, and topographic data for 2022. Because the target precipitation product was derived from CRU TS at 0.5° resolution, this case study should be interpreted as a statistical downscaling and fine-resolution spatial prediction experiment rather than direct validation against independent 1 km precipitation observations. K-means, Gaussian mixture model, and spectral clustering were used for spatial stratification, and multiple machine-learning models were compared within different strata. Model performance was assessed using spatial block cross-validation and compared with global machine-learning models, stratified context-aware models without local outlier features, and additional spatial reference models. The best-performing configuration, GMM-based LOCA, achieved an R^2 of 0.858 ± 0.125, RMSE of 77.67 ± 15.43, and MAE of 57.12 ± 14.44. The improvement was associated with two complementary mechanisms: spatial stratification enhanced regional adaptability, while local outlier-derived features provided supplementary information on local spatial contrasts. These findings suggest that explanatory-variable local outliers can provide useful contextual information for statistical downscaling and spatial prediction in heterogeneous environments.

ARTICLE HISTORY
Received 30 March 2026
Accepted 3 July 2026

KEYWORDS
Spatial local outliers; spatial context-aware models; spatial stratification; precipitation

1 Introduction

The spatial context-aware model adapts prediction strategies to different geographical contexts by considering spatial variations in environmental factors, data distributions, and model parameters (Asghari and Nasseri 2015; Meyer et al. 2019; Shekhar et al. 2002). Such models are important for geographical and environmental prediction because spatial data often exhibit strong heterogeneity, non-stationarity, and regional dependence. Improving context-aware spatial modelling therefore requires not only diagnosing spatial structure, but also transforming spatial information into effective model inputs and adaptive modelling strategies.

Spatial differentiation provides an important basis for context-aware spatial modelling. Existing methods for describing spatial differentiation can be broadly classified into spatial partitioning analysis, which divides the study area into internally similar sub-regions (Lubbe et al. 2020; Subbu and Ray 2008); spatial stratified heterogeneity analysis, which evaluates between-region differences across spatial strata or scales

CONTACT Yongze Song yongze.song@curtin.edu.au
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15481603.2026.2700126.
© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

(Duan et al. 2024; Wang and Wan 2022); and GIS-based spatial analysis, which supports the visualisation, processing, and statistical diagnosis of spatial patterns (Supriadi and Oswari 2020). Among these methods, spatial stratified heterogeneity modelling is particularly useful for constructing context-aware models because it allows prediction relationships to vary among different spatial regimes.

Geographic detection models are widely used to quantify spatial stratified heterogeneity by comparing within-stratum and between-stratum variance (Liu et al. 2021; Su et al. 2020). Representative extensions include the optimal parameter-based geographic detector (OPGD) (Cen, Zhou, and Qiu 2024; Song et al. 2020), interactive detector for spatial association (IDSA) (Zhang et al. 2023), spatial association detector (SPADE) (Cang and Luo 2018), geographic optimal zoning-based heterogeneity (GOZH) (Luo et al. 2022), robust geographic detector (RGD) (Zhang, Song, and Wu 2022), robust interaction detector (RID) (Zhang et al. 2024), and heterogeneous spatial autocorrelation (Zhang, Li, and Song 2024). These methods improve the diagnosis and interpretation of spatial heterogeneity, but their primary purpose is not spatial prediction itself.

It is therefore necessary to distinguish heterogeneity-oriented detectors from association-oriented methods. Heterogeneity-oriented detectors mainly evaluate whether spatial stratification explains the variance of a target variable, whereas association-oriented statistics, such as Moran's I, local indicators of spatial association (LISA), and spatial association detectors, focus more on spatial dependence, local clustering, co-location, or interactions among geographical variables or neighbouring observations (Anselin 1995; Cang and Luo 2018; Moran 1950). Recent studies have further combined LISA, geographically weighted models, and machine-learning methods to analyse spatially heterogeneous processes, such as forest loss, urban green-space accessibility, and environmental remediation (Adorno, Pereira, and Amaral 2025; O'Leary et al. 2023; Park, Lee, and Kim 2021). However, for spatial prediction, diagnosing heterogeneity or association is only one step; the key challenge is how to convert such spatial structural information into predictive features or adaptive modelling strategies.

Many spatial methods explicitly consider spatial characteristics such as autocorrelation, heterogeneity, anisotropy, spatial regimes, and local non-stationarity. For example, geographically weighted regression models spatially varying relationships through local regression coefficients (Fotheringham, Brunsdon, and Charlton 2002), whereas robust kriging and related geostatistical methods aim to reduce the adverse influence of abnormal observations in spatial prediction (Sun et al. 2019). Nevertheless, in many machine-learning-based spatial prediction workflows, spatial characteristics are still more often used for exploratory analysis, model diagnosis, or post hoc interpretation than as explicit predictive features. Spatial characteristics, including autocorrelation (Getis 2009), heterogeneity (Ben-Moshe and Itzkovitz 2019), complexity (Carpenter and Brock 2004), similarity (Li and Fonseca 2006), and spatial outliers (Singh and Lalitha 2018), may also contain useful contextual information for improving prediction.

Spatial outliers are an important but underused type of spatial contextual information. Traditionally, outliers are often regarded as errors, noise, or observations that may reduce model stability. From a geographical perspective, however, not all outliers are meaningless. In environmental systems, local anomalies may reflect meaningful spatial processes, such as orographic precipitation enhancement, land-water contrasts, localised runoff or evaporation conditions, and terrain- or circulation-related effects. It is also necessary to distinguish global outliers from local outliers. Global outliers deviate from the overall distribution of a variable across the entire study area, whereas local outliers are unusually high or low relative to neighbouring locations or the local spatial background. For spatial prediction, local outliers may be more informative because they represent regional departures from surrounding environmental conditions rather than only extreme values in the global distribution.

Existing studies have considered outliers in spatial modelling, such as robust kriging and outlier-aware spatial prediction (Sun et al. 2019). However, most studies focus on detecting anomalous observations or reducing their adverse influence. In contrast, the potential role of local outlier-derived information as explicit explanatory features within a context-aware spatial prediction framework remains relatively limited. This gap motivates the present study.

This study develops a local-outlier-based context-aware model, named LOCA, for spatial prediction. The novelty of LOCA does not lie in replacing existing spatial stratification, spatial association, or local modelling methods. Instead, it lies in integrating three components into a unified predictive workflow: first, constructing local outlier-derived features from explanatory variables to represent local spatial anomaly information; second, combining these features with spatial stratification to characterise heterogeneous spatial regimes; and third, selecting stratum-specific machine-learning models to improve adaptive prediction under different spatial contexts. In this sense, LOCA bridges spatial structure diagnosis and spatial machine-learning prediction by transforming local anomaly information into model inputs rather than only treating outliers as observations to be removed or down-weighted.

Specifically, this study aims to answer three research questions: (1) Can local outlier-derived features provide useful spatial contextual information for environmental spatial prediction? (2) Can the integration of local outlier-derived features, spatial stratification, and stratum-specific model selection improve prediction accuracy compared with global models and stratified models without local outlier features? (3) What is the overall contribution of local outlier-derived features to model prediction? By answering these questions, this study seeks to clarify the role of local spatial outliers in context-aware spatial prediction and to provide a methodological framework for incorporating local anomaly information into spatial machine-learning models.

2 The local-outlier-based context-aware (LOCA) model

2.1 Concepts

Geospatial context-aware models demonstrate the growing need for different spatial algorithms and methods across various regions. The regionally varied methods are not just geospatial techniques that can be applied in different areas with adjusted or modified parameters (Rahang and Patowari 2016), such as various spatial scales for the geographically weighted regression, but also highlight the necessity for distinct algorithms across space. For instance, in the spatial prediction of precipitation, support vector machine may be the optimal algorithm in one region, while random forest may be the best-performed approach in another area (Qi et al. 2021). Thus, variability and heterogeneity are common within spatial data, meaning that regionally adaptive and specialised models (i.e. spatial context-aware models) are widely required to improve their cross-regional performance in spatial prediction.

Local outliers, a type of spatially local characteristic of geographical variables, contain spatially stratified characteristics of data for more accurate and effective spatial prediction, in addition to their spatially local nature. Local outliers represent extremely high or low values within a small or local region compared with nearby observations. This definition emphasises the relative nature of local outliers: a value may not be extreme in the global distribution but may still be anomalous within its immediate spatial context. Therefore, local outliers should not be understood only as abnormal observations in a statistical sense, but also as localised departures from the surrounding geographical background. Most statistical and spatial models contain mathematical assumptions that outliers should be avoided due to their impact on model fit. Such treatment is reasonable when outliers originate from measurement errors, data processing problems, or random noise. However, in geographical and environmental systems, some local outliers may be generated by real spatial processes and may therefore contain useful contextual information. From a spatial perspective, local outliers may contain essential spatial information for explaining regional geographical and environmental phenomena, such as high trace element concentrations (Song 2023), high disease incidence (Ge et al. 2017), or heat waves (Gaitán 2016).

2.2 The proposed LOCA model

This study develops a local-outlier-based context-aware model, named LOCA, for spatial prediction by integrating local outlier-derived features, spatial stratification, and stratum-specific machine-learning models. Figure 1 illustrates the basic architecture of the LOCA model, which consists of three key steps. First, local outlier-derived features are constructed from explanatory variables to represent local spatial anomaly information. Second, spatial stratification is performed to characterise heterogeneous spatial regimes. Finally, stratum-specific models are selected and integrated to generate spatial predictions. Each step of the LOCA model is described in the following sections.

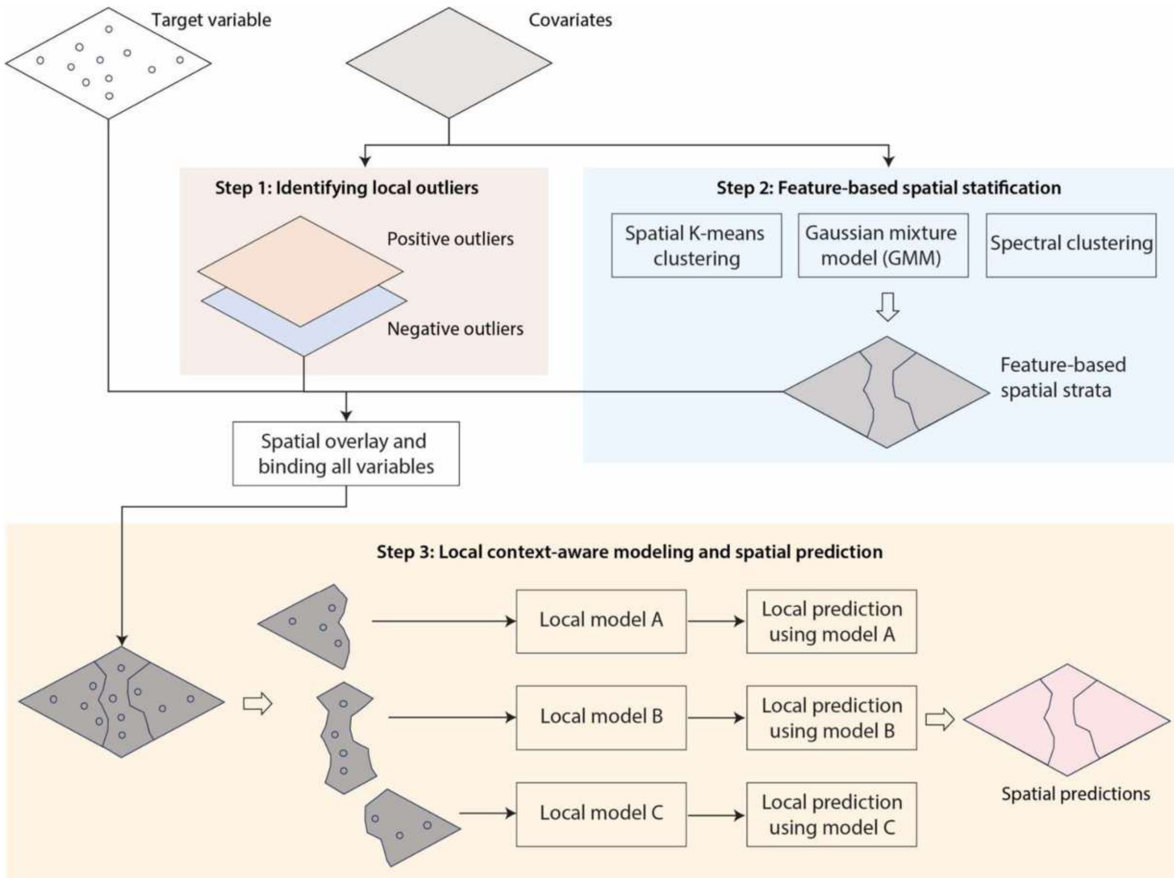

Figure 1. Schematic overview of the local outliers for context-aware (LOCA) model for spatial prediction.

2.2.1 Identify local outliers

The first step of LOCA is to construct local outlier-derived features from the explanatory variables. These features were calculated only from the explanatory variables, rather than from the target precipitation variable, because the purpose is to provide additional spatial contextual information for prediction and to avoid using target-variable information in feature construction.

For a given explanatory variable m, let x_{im} denote its value at spatial location s_i, where s_i represents the coordinate of location i. The local neighbourhood of location i is defined according to spatial distance. For each location, the local mean and local standard deviation are calculated using the valid neighbouring values of the same explanatory variable. Based on the local mean and standard deviation, the positive and negative local outlier-derived features are defined as normalised magnitude-weighted exceedance statistics.

$$N_i(r) = \{j \mid 0 < d(s_i, s_j) \leq r\} \tag{1}$$

$$\mu_{im} = \frac{1}{n_i} \sum_{j \in N_i(r)} x_{jm} \tag{2}$$

$$\sigma_{im} = \sqrt{\frac{1}{n_i - 1} \sum_{j \in N_i(r)} (x_{jm} - \mu_{im})^2} \tag{3}$$

$$O_{im}^{+}(r, k) = \frac{1}{n_i} \sum_{j \in N_i(r)} I(x_{jm} > \mu_{im} + k\sigma_{im}) \cdot (x_{jm} - \mu_{im} - k\sigma_{im}) \tag{4}$$

$$O_{im}^{-}(r,k) = \frac{1}{n_i} \sum_{j \in N_i(r)} I(x_{jm} < \mu_{im} - k\sigma_{im}) \cdot (\mu_{im} - k\sigma_{im} - x_{jm}) \tag{5}$$

where $d(s_i, s_j)$ is the spatial distance between locations i and j, r is the neighbourhood radius, n_i is the number of valid neighbouring locations, *I(.)* is an indicator function, and k is the threshold coefficient. If the number of valid neighbours within $N_i(r)$ was smaller than $n_{\min}$, the neighbourhood was expanded by including the nearest valid locations until the minimum requirement was satisfied. In this study, $k = 2.0$ was used as the baseline threshold following the local z-score logic. If no neighbouring values satisfy the positive or negative outlier condition, the corresponding feature is set to zero. Features with zero variance were removed before variable selection and model training. Sensitivity analyses were further conducted using alternative r and k values, and the results are reported in Table S3. The detailed workflow for constructing local outlier-derived features is provided in Algorithm S1 in the Supplementary Materials.

2.2.2 Feature-based spatial stratification

To construct a spatial context-aware model, spatial stratification is a crucial step because it allows the prediction relationships to vary across different environmental contexts. Spatial stratification can be classified into univariate and multivariate stratification according to the number of variables used, or into hard and soft spatial adjacency stratification according to the spatial continuity requirement of strata (Guo et al. 2022). Hard adjacency stratification requires locations within the same stratum to be spatially contiguous, whereas soft adjacency stratification allows locations with similar environmental characteristics to be grouped together even if they are not directly adjacent. In this study, a soft adjacency strategy was adopted because the main purpose was to identify regions with similar precipitation-forming environments rather than to enforce strictly continuous geographical boundaries. For example, high-elevation areas located in different parts of Qinghai Province may share similar terrain, wind speed, evaporation, and runoff conditions; therefore, they can be assigned to the same stratum even when they are geographically separated. This strategy helps the model capture similar environmental controls on precipitation across non-contiguous regions.

Common multivariate stratification methods include spatial K-means clustering (Brus, De Gruijter, and Van Groenigen 2006), Gaussian mixture model (GMM) (Zhao et al. 2016), and spectral clustering. These methods were used in this study to construct alternative spatial stratification schemes based on the selected explanatory variables. Because K-means and related clustering methods are sensitive to feature scaling, all variables used for stratification were standardised before clustering. The final selected covariate set contained only four original explanatory variables before local outlier expansion; therefore, high dimensionality was not a major concern in the present application. Instead, the main methodological concerns were feature scaling, cluster shape, initialisation sensitivity, and the selection of an appropriate number of strata. To address these issues, different stratification methods were compared, and the number of strata was evaluated using objective clustering diagnostics, including the silhouette score, Davies-Bouldin index, and Calinski-Harabasz index.

2.2.3 Spatial context-aware modelling

In complex geographical spaces, significant differences often exist in the characteristics and data distribution across different regions. To effectively capture this spatial heterogeneity, spatial stratification techniques are employed to divide the study area into multiple sub-regions, with adaptive models constructed for each sub-region. The variations among different sub-regions in aspects such as climate and topography indicate that a single model may struggle to accurately describe their complex dynamic patterns and spatial structures.

For the characteristics of each sub-region, the choice of models can be diverse. Commonly used models include machine learning models, such as random forest (RF) and support vector machines (SVM), which are suitable for handling complex nonlinear data; neural network models, such as convolutional neural networks (CNN) and long short-term memory networks (LSTM), which are effective for addressing complex dependencies in spatial and temporal data; linear regression models, which are appropriate for regions with simple relationships among variables; and spatial interpolation models, such as kriging, which are used in areas with significant spatial autocorrelation. In addition, Bayesian models and decision tree models can also be employed to model sub-regions with different characteristics.

The selection of models should be flexibly adjusted based on the characteristics of the sub-regions, data properties, and research objectives. By constructing adaptive models for each sub-region individually, the overall accuracy and predictive capability of spatial data analysis can be enhanced. Ultimately, by integrating the model results from each sub-region, high-precision predictions for the entire study area can be achieved.

3 Case study: LOCA for the spatial prediction of precipitation

3.1 Study area and data

In this study, precipitation data were obtained from the CRU TS dataset, which provides gridded climate variables at a spatial resolution of 0.5° and has been widely used in regional climate studies (Harris et al. 2020). The CRU TS precipitation dataset has also been validated in mainland China by previous studies (Shi, Li, and Wei 2017). Meteorological variables, including evaporation, temperature, radiation, wind speed, and surface pressure, were obtained from the ERA5 dataset (Hersbach et al. 2020), which has shown good applicability in northwestern China (Fang et al. 2023). Geographic data, including the Digital Elevation Model (DEM) and terrain slope, were obtained from the China 1 km resolution digital elevation map provided by the Spatial-Temporal Triple Environment Big Data Platform (Tang 2019). All meteorological variables were calculated as annual averages for 2022.

Because the original CRU TS precipitation data are coarser than the 1 km topographic predictors, the task in this study should be interpreted as a fine-resolution statistical downscaling and spatial prediction experiment rather than direct observation-based mapping at 1 km resolution. To construct a consistent modelling dataset, all variables were spatially matched to the 1 km prediction grid. The topographic variables were used as fine-resolution predictors, while the coarser CRU TS precipitation and ERA5 meteorological variables were resampled or spatially aligned to the same grid using a consistent interpolation/resampling procedure. Therefore, the 1 km output represents a model-based fine-resolution estimate constrained by the coarse-resolution precipitation product and should not be interpreted as independent 1 km precipitation observations.

The average annual temperature in Qinghai Province ranges from −5.1 °C to 9.0 °C, while annual precipitation varies from 15 mm to 750 mm, with most areas receiving less than 400 mm of precipitation. Figure 2(b) illustrates the spatial distribution of precipitation, showing a decreasing trend from south to north. Figure 2(c) presents the statistical characteristics of precipitation, with an average of 330.34 mm, ranging from 9.7 to 826.4 mm, and 50% of the precipitation distributed between 135.72 mm and 515.18 mm.

The use of a single-year average provides a consistent cross-sectional dataset for evaluating the spatial behaviour of the LOCA framework under the climatic conditions of 2022. However, precipitation–environment relationships may vary among years due to interannual climate variability, such as changes in atmospheric circulation, moisture transport, and local land–atmosphere interactions. Therefore, the present case study mainly evaluates the spatial applicability of LOCA for the selected year rather than claiming full temporal generalisability across different climatic years.

3.2 Experiment design

Figure 3 shows the workflow of the LOCA model for predicting the spatial distribution of precipitation in Qinghai Province at a 1 km grid. Because the target precipitation field is derived from a coarser gridded product, the workflow should be interpreted as fine-resolution statistical downscaling and spatial prediction. The process includes four main steps: selecting and preprocessing meteorological and geographic explanatory variables, constructing local outlier-derived features and spatial strata, performing spatial prediction using global and context-aware models, and validating the models using spatial block cross-validation.

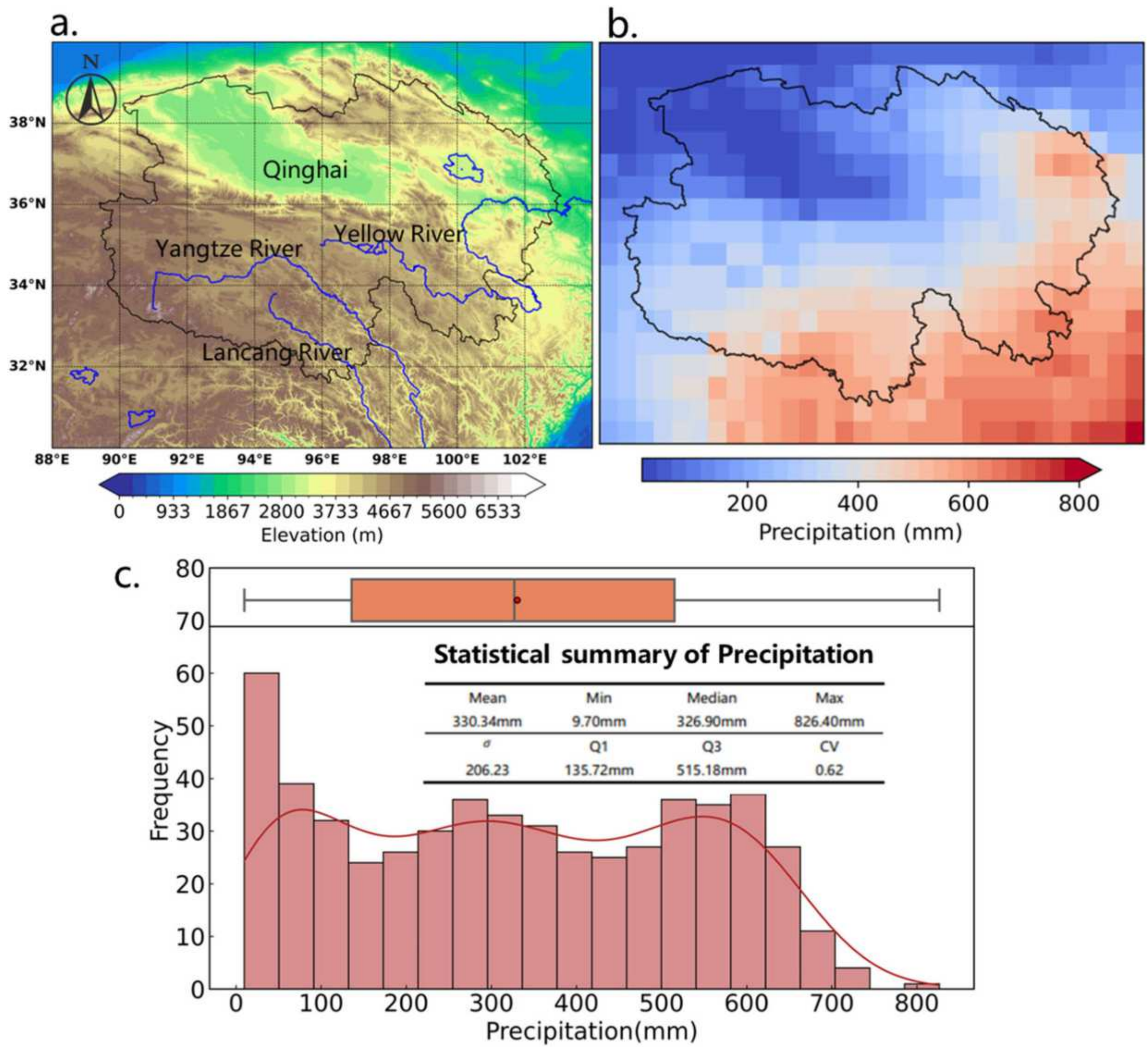

Figure 2. Geographical characteristics of the study area. (a) Location of the study area in China, (b) Spatial distribution of precipitation in the study area, (c) Statistics on precipitation in the study area.

The first step involves the selection of explanatory variables. Preliminary predictors included evaporation, temperature, radiation, wind speed, atmospheric pressure, elevation, terrain slope, and surface runoff because these factors influence moisture supply, atmospheric transport, uplift, cooling, and precipitation formation. Pearson correlation analysis was used as a preliminary screening step, with $|r| >= 0.3$ used as a practical criterion to identify variables with at least moderate association with precipitation. This threshold was combined with physical interpretability and multicollinearity diagnosis, so the final variable set was not determined solely by correlation magnitude. To avoid information leakage, variable selection and multicollinearity screening were conducted using the training data only within each validation procedure, while the testing data were reserved only for final model evaluation.

The second step is to utilise the LOCA model for precipitation modelling. First, the positive and negative local outlier-derived features of each explanatory variable are calculated using Equations (1)–(5). Next, the explanatory variables are subjected to spatial stratification using K-means, Gaussian mixture model (GMM), and spectral clustering. Finally, random forest (RF), linear regression (LR), support vector regression (SVR), XGBoost, and K-nearest neighbours (KNN) models are applied to each clustered sub-region and to the overall region. All candidate models were implemented using a consistent modelling workflow. Continuous variables were standardised for scale-sensitive models, including LR, SVR, and KNN, whereas tree-based models were fitted without additional scaling. A fixed random seed was used for data splitting and model

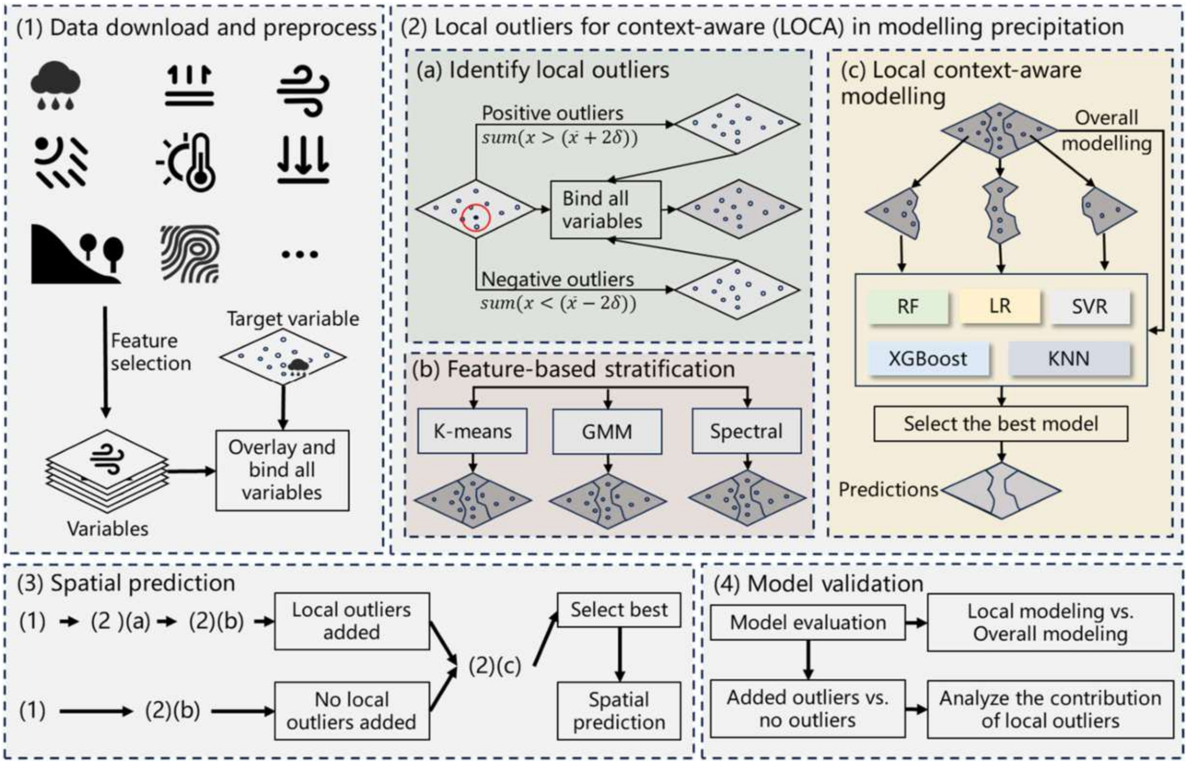

Figure 3. The process of LOCA spatial prediction of precipitation.

training where applicable. Model selection was conducted within the training data of each validation procedure, and the testing data were used only for final evaluation. The same preprocessing and model-selection strategy were applied to all strata to ensure fair comparison among algorithms.

The third step is to perform spatial prediction. Based on the modelling results from the second step, the initially selected explanatory variables can be used directly for spatial prediction, or the local outlier-derived features can be added as supplementary predictors. This study compares single machine-learning models, machine-learning models with local outlier-derived features, spatial context-aware models without local outlier-derived features, and LOCA models that combine spatial stratification and local outlier-derived features.

The fourth step is model validation. To reduce the risk of overoptimistic accuracy estimates caused by spatial autocorrelation, the original random training-testing split was replaced by spatial block cross-validation. The study area was divided into spatial blocks, and samples from the same block were assigned to the same validation fold. In each fold, one group of spatial blocks was used as the validation set, while the remaining blocks were used for model training. The same spatial block folds were used for all model comparisons. Model performance was reported as mean ± SD across spatial folds using the coefficient of determination (R^2, Equation (6)), root mean square error (RMSE, Equation (7)), and mean absolute error (MAE, Equation (8)) (Meng et al. 2024).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(o_i - y_i)^2}{\sum_{i=1}^{n}(o_i - \underset{\cdot}{o})^2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\, y_i - o_i \,| \tag{8}$$

Where y is the predicted data and o is the observed data.

Feature contribution was interpreted using SHAP values calculated on held-out validation samples under spatial block cross-validation. To provide a unified interpretation, SHAP analysis was conducted using a global random forest model with both original explanatory variables and local outlier-derived features. The predictors were grouped into original variables, positive local outlier-derived features, and negative local outlier-derived features.

4 Results

4.1 Factors of precipitation

Setting the correlation coefficient threshold at 0.3, seven factors were identified as pre-selected explanatory variables through variable selection: elevation, evaporation, potential evaporation, surface pressure, surface runoff, temperature, and wind speed. A multicollinearity test was conducted on these seven variables, and the results showed that the variance inflation factors (VIF) for potential evaporation, surface pressure, and temperature were 5.54, 18.15, and 15.55, respectively, all of which exceeded the threshold of 2.5. Consequently, these three explanatory variables were removed, leaving four explanatory variables. The statistical characteristics of these remaining variables are shown in Table 1. Among them, r is the correlation coefficient between the explanatory variables and precipitation. Elevation and surface runoff were positively correlated with precipitation, while evaporation and wind speed were negatively correlated with precipitation. In addition, the VIF values for the remaining variables were below 2, indicating no severe multicollinearity.

4.2 Identify outliers

In Figure 4, the calculated negative local outlier-derived feature of surface runoff is zero at all locations, resulting in undefined correlation coefficients with other variables. This indicates that no neighbouring values of surface runoff fell below the local mean minus the selected threshold in the study area. In other words, compared with their local spatial background, surface runoff values rarely showed distinct negative departures. This result is related to the distributional characteristics of surface runoff in Qinghai Province, where many grid cells have low or near-zero runoff values. Under such conditions, the lower tail of the local distribution is strongly constrained by the zero boundary, making it difficult to identify meaningful negative local anomalies. Therefore, the negative local outlier-derived feature of surface runoff was excluded from subsequent modelling. This exclusion does not imply that surface runoff itself is unimportant for precipitation prediction; rather, it indicates that its negative local outlier-derived feature is degenerate and contains no spatial variation for model learning. By contrast, the original surface runoff variable and its positive local outlier-derived feature were retained when they satisfied the variable-selection and multicollinearity criteria. After removing this degenerate feature, the VIF values of the remaining variables were below 2 and did not exceed the threshold of 2.5, indicating that the local outlier-derived features could be used for subsequent modelling.

4.3 Spatial stratification

The spatial stratification of explanatory variables was performed using K-means, GMM, and spectral clustering, and the stratification results are shown in Figure 5. To avoid determining the number of strata only according to the minimum sample-size rule, additional clustering diagnostics were calculated for

Table 1. Statistical summary of explanatory variables for the prediction of precipitation.

Variable	Mean	Min	Median	Max	σ	ρ	VIF
Elevation	3872	796	4156	5548	1021	0.36***	1.55
Evaporation	339.74	14.6	387.54	618.41	156.92	−0.79***	1.75
Surface runoff	38.08	0	16.44	436.80	53.54	0.67***	1.24
Wind speed	0.93	0.02	0.81	3.41	0.58	−0.34***	1.25

***indicates significance level indicated by a p-value less than 0.001.

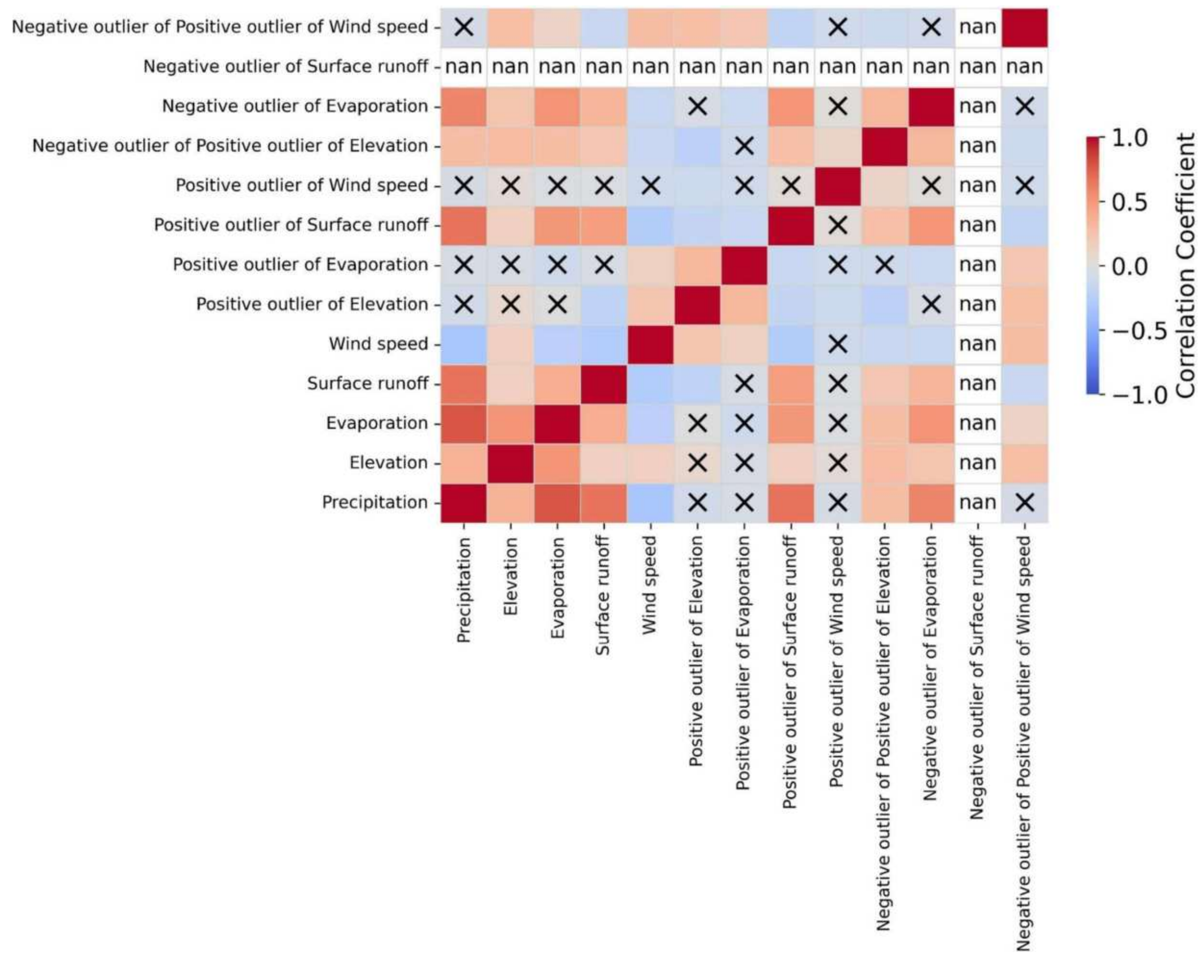

Figure 4. Correlation matrix between precipitation and selected explanatory variables and outliers of explanatory variables. Note: "x" indicates that the correlation is not significant.

different numbers of strata from K = 3 to K = 6. The diagnostic metrics included the silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. For GMM, AIC and BIC were additionally calculated to evaluate the trade-off between model fit and complexity. The results are reported in Table S1.

The diagnostic results supported the use of K = 5 as a balanced stratification setting. For K-means, K = 5 achieved the highest silhouette score, the lowest Davies-Bouldin index, and the highest Calinski-Harabasz index among the tested K values, indicating better overall clustering quality. For GMM, K = 5 achieved the highest silhouette score and Calinski-Harabasz index, although AIC and BIC continued to decrease at K = 6. This suggests that K = 6 may improve likelihood-based fit but also increases model complexity and may reduce the stability of local model training. For spectral clustering, K = 5 achieved the highest Calinski-Harabasz index and a silhouette score close to the maximum value at K = 4, while K = 6 showed the lowest Davies-Bouldin index. Therefore, K = 5 was selected as the baseline stratification number by jointly considering clustering quality, sample-size adequacy, spatial interpretability, model complexity, and subsequent spatial block cross-validation performance.

As shown in Figure 5(d), when K = 5, the number of samples in each stratum was greater than the minimum requirement of 40, which ensured that each stratum had sufficient samples for local model training and model selection. Among the three stratification methods, the spatial distributions of the K-means and spectral clustering results were more similar, and the sample sizes within their strata were relatively balanced. This similarity was partly reflected in the subsequent model performance, as both K-means- and spectral-based context-aware models outperformed their corresponding single machine-learning models under spatial block cross-validation. However, similar visual stratification patterns did not lead to identical prediction accuracy. As shown in Table 2, the GMM-based LOCA achieved the best

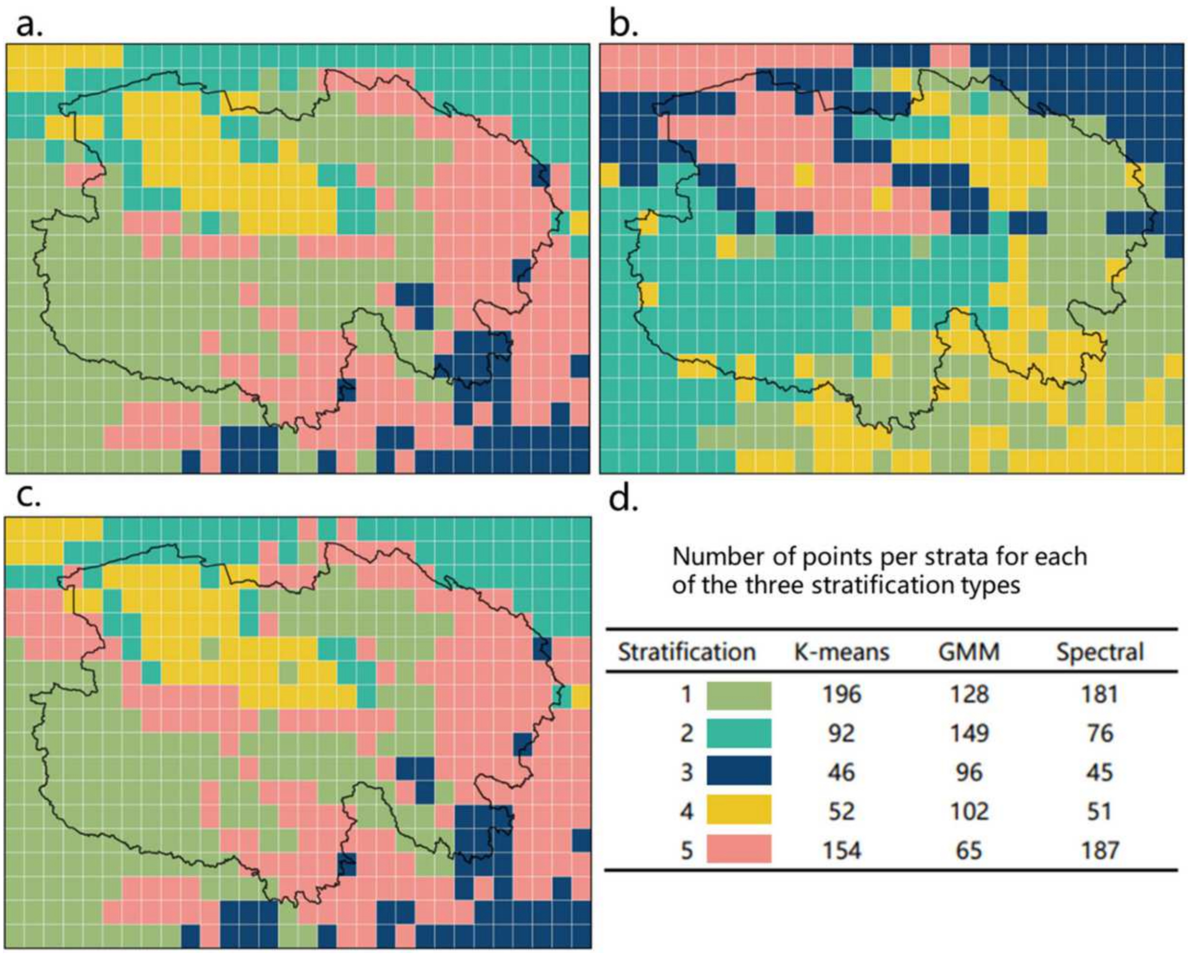

Figure 5. Spatial distribution of the three types of stratification and number of points per strata. (a) K-means, (b) GMM, (c) Spectral, (d) Number of points per strata for each of the three stratification types.

Table 2. Model validation results under spatial block cross-validation. Metrics are reported as mean ± SD across spatial folds.

Spatial stratification method	Spatial prediction model	R^2	RMSE	MAE
K-means	Single machine learning (ML)	0.761 ± 0.140	100.90 ± 15.26	75.20 ± 12.93
	Spatial context-aware ML	0.796 ± 0.154	93.24 ± 17.21	69.66 ± 15.93
	LOCA	**0.850** ± 0.100	**79.84** ± 9.41	**62.34** ± 10.29
Gaussian mixture model (GMM)	Single machine learning (ML)	0.760 ± 0.143	101.11 ± 15.70	75.01 ± 13.06
	Spatial context-aware ML	0.791 ± 0.133	94.20 ± 14.39	70.84 ± 13.02
	LOCA	**0.858** ± 0.125	**77.67** ± 15.43	**57.12** ± 14.44
Spectral clustering	Single machine learning (ML)	0.764 ± 0.126	100.24 ± 12.06	75.80 ± 10.31
	Spatial context-aware ML	0.779 ± 0.143	96.95 ± 13.79	72.07 ± 11.41
	LOCA	**0.845** ± 0.126	**81.14** ± 15.21	**64.69** ± 14.43

overall performance, with the highest R^2 and the lowest RMSE and MAE. This indicates that model performance was determined not only by visual similarity of strata or sample-size balance, but also by the internal feature homogeneity of each stratum, the suitability of stratum-specific model selection, and the contribution of local outlier-derived features.

4.4 Spatial prediction

Figure 6 shows the precipitation spatial prediction results based on Machine Learning, Machine Learning with Local Outliers, Spatial Context-Aware Machine Learning, and LOCA. The results indicate that there are clear differences in the predicted spatial patterns among the four modelling strategies. The predictions

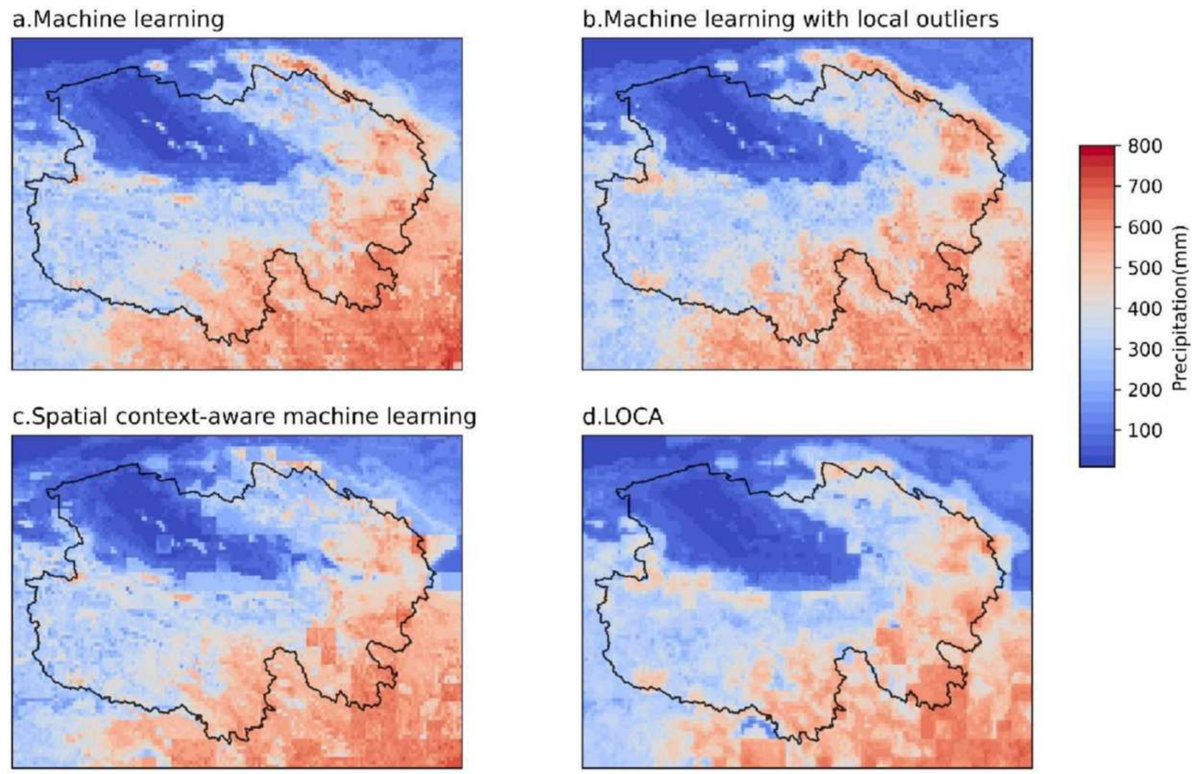

Figure 6. Spatial predictions of precipitation in the study area.

from Machine Learning and Machine Learning with Local Outliers show relatively smooth spatial variations, whereas Spatial Context-Aware Machine Learning and LOCA exhibit more local changes.

This difference is mainly related to the adopted modelling structure. In the global Machine Learning and Machine Learning with Local Outliers models, a single prediction function is applied to the entire study area, which tends to produce smoother spatial outputs. In contrast, Spatial Context-Aware Machine Learning and LOCA construct separate prediction models for different strata. Therefore, when neighbouring locations belong to different strata, predicted values may change more rapidly because of the transition between stratum-specific models and their fitted relationships.

The inclusion of local outlier-derived features further affects the spatial pattern of LOCA. These features describe positive or negative departures from the surrounding spatial background and may enhance local spatial contrasts in the prediction results. Therefore, the spatial fluctuations in LOCA may arise from two sources: the transition of models across stratum boundaries and the additional local anomaly information introduced by outlier-derived features. However, such fluctuations should not be automatically interpreted as improved spatial detail, because part of them may also reflect artificial discontinuities caused by stratified modelling. Therefore, the spatial prediction maps should be interpreted together with validation metrics and residual patterns.

To further understand the spatial prediction differences among the models, a comparison of predicted values along a profile line from the northwest to the southeast of the study area was conducted (Figure 7). The profile comparison shows that the global models produce smoother changes along the transect, whereas the spatial context-aware models and LOCA show more pronounced local variations. In particular, LOCA presents a wider prediction range along the section from 89.0°E to 97.9°E, suggesting that the inclusion of local outlier-derived features increases the model's sensitivity to local spatial variation. Nevertheless, a wider prediction range is not necessarily superior by itself. Its reliability should be evaluated based on whether the predicted variations are consistent with observed precipitation, validation performance, and residual distributions. Therefore, the profile analysis is used here as a complementary spatial interpretation rather than as independent evidence of model superiority.

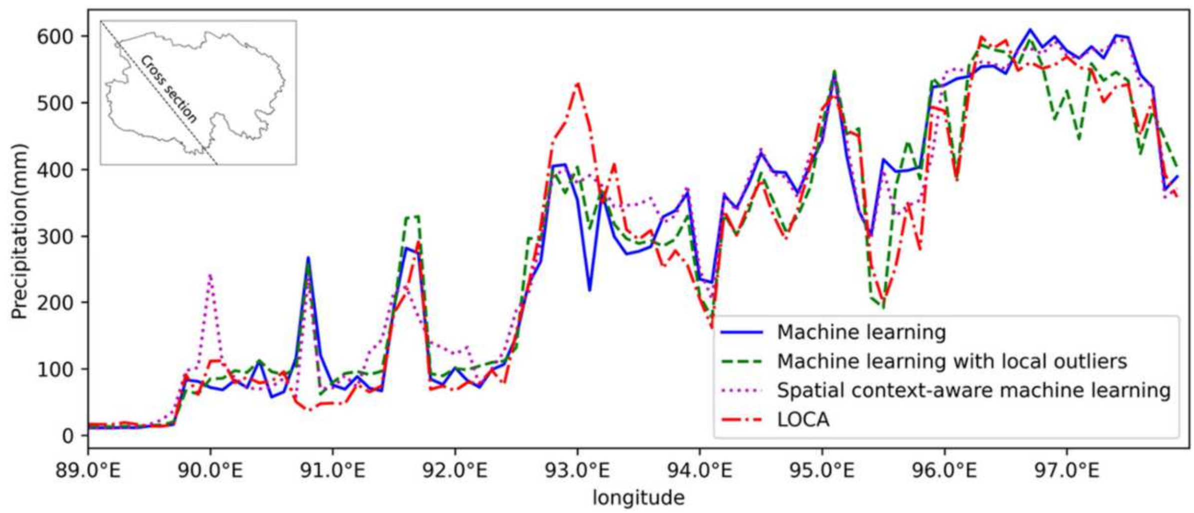

Figure 7. Comparisons of the four models' predictions along the cross-section from the northwest to the southeast of the study area.

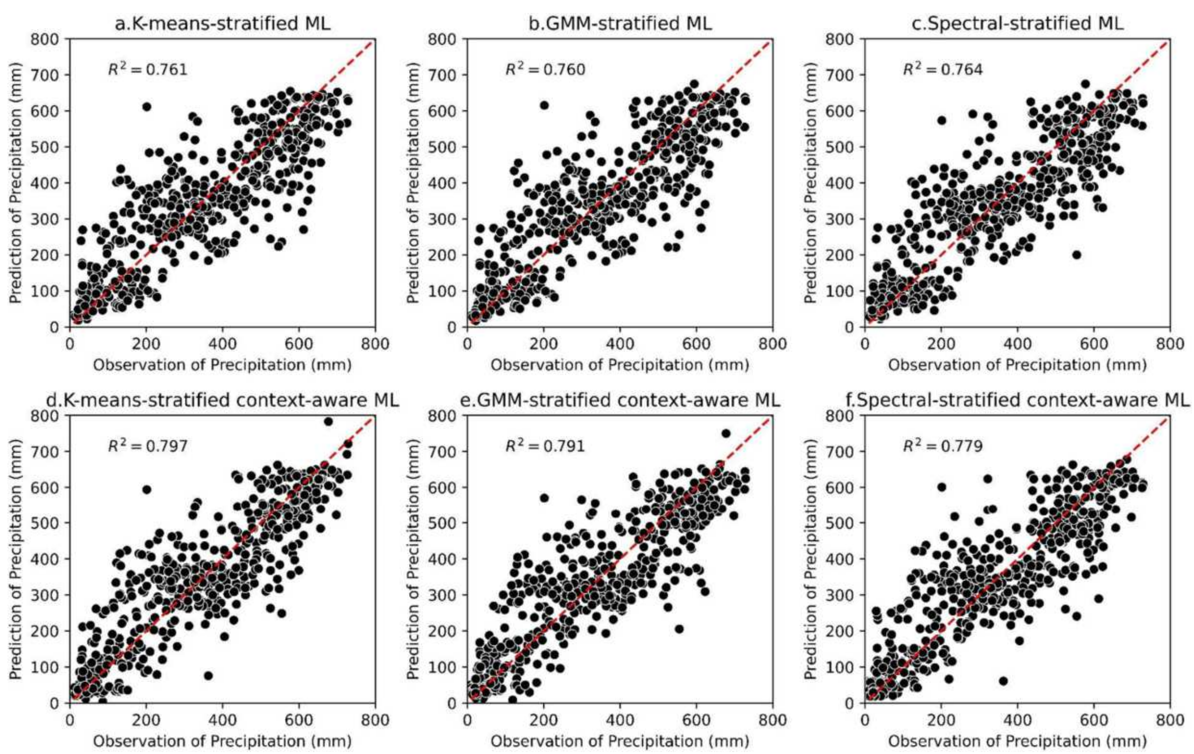

Figure 8. Comparison of predictions and observations in the validation set of precipitation data for context-aware and using the single machine learning model.

4.5 Model validation

4.5.1 Model improvement by spatial context-aware models

Model validation was conducted using spatial block cross-validation, which separates training and validation samples by spatial blocks and provides a stricter assessment of spatial generalisation than random splitting. All validation metrics are reported as mean ± SD across spatial folds.

Figure 8 compares observed and predicted precipitation values for the single machine-learning models and the spatial context-aware models. The spatial context-aware models generally showed closer agreement with the observations than their corresponding single machine-learning models across the three stratification methods. As summarised in Table 2, compared with the single machine-learning models, the spatial context-aware models increased R^2 by 4.6%, 4.1%, and 2.0% for K-means, GMM, and spectral clustering, respectively. Meanwhile, RMSE decreased by 7.6%, 6.8%, and 3.3%, and MAE decreased by 7.4%, 5.6%, and 4.9%, respectively.

Among the three spatial context-aware models, the K-means-based context-aware model achieved the best overall performance, with an R^2 of 0.796 ± 0.154, RMSE of 93.24 ± 17.21, and MAE of 69.66 ± 15.93. These results indicate that spatial stratification and stratum-specific model selection can improve precipitation prediction by allowing different regions to use locally suitable predictive relationships. The relatively large SD values also suggest that prediction difficulty varied among spatial folds, reflecting the strong spatial heterogeneity of precipitation and explanatory variables in Qinghai Province.

4.5.2 Model improvement by local outliers

Figure 9 compares observed and predicted precipitation values for LOCA and the corresponding spatial context-aware models without local outlier-derived features. LOCA generally produced more accurate predictions across the three stratification methods, indicating that local outlier-derived features provided additional spatial contextual information beyond stratification alone.

As shown in Table 2, LOCA further improved model performance compared with the spatial context-aware models. For K-means, GMM, and spectral clustering, R^2 increased by 6.8%, 8.5%, and 8.5%, respectively. RMSE decreased by 14.4%, 17.5%, and 16.3%, and MAE decreased by 10.5%, 19.4%, and 10.2%, respectively. Compared with the single machine-learning models, LOCA increased R^2 by 10.6%–12.9%, reduced RMSE by 19.1%–23.2%, and reduced MAE by 14.7%–23.9%. These improvements suggest that LOCA benefits from two complementary mechanisms: spatial stratification improves regional adaptability, while local outlier-derived features enhance the representation of local spatial contrasts.

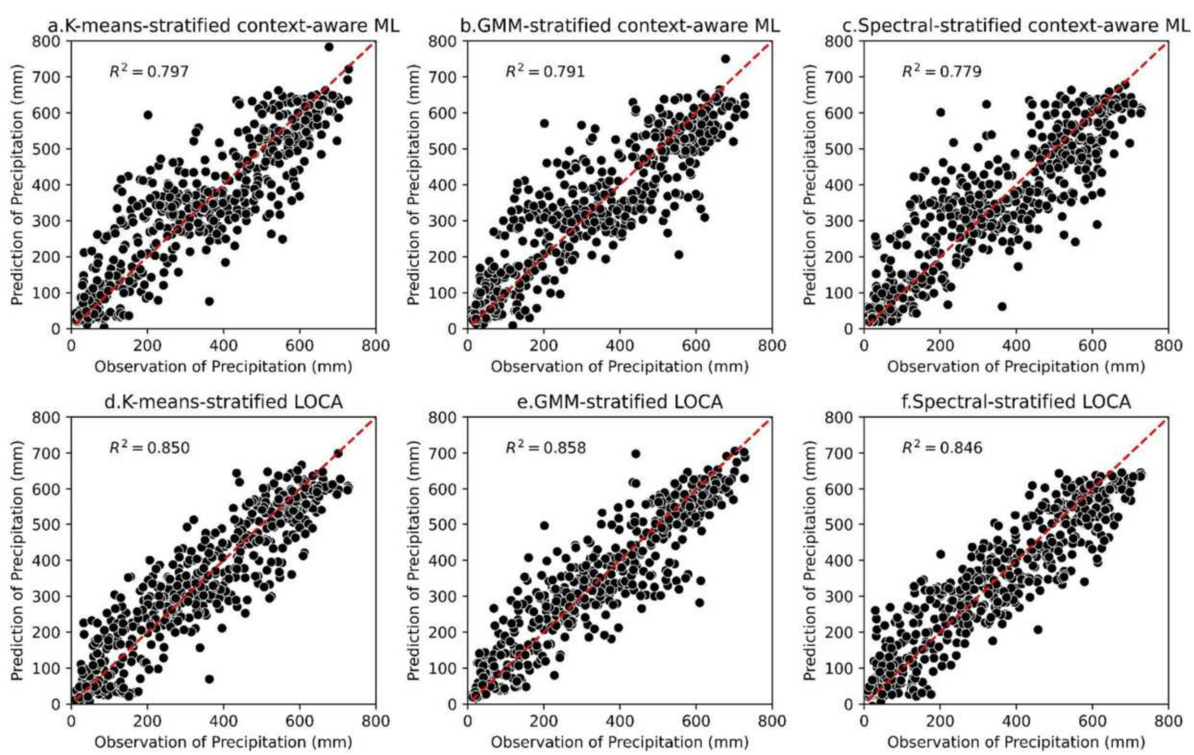

Figure 9. Comparison of predictions and observations in the validation set of precipitation data for LOCA and spatial context-aware models.

Among the three LOCA models, the GMM-based LOCA achieved the best overall performance, with an R^2 of 0.858 ± 0.125, RMSE of 77.67 ± 15.43, and MAE of 57.12 ± 14.44. All three LOCA configurations outperformed their corresponding single machine-learning and spatial context-aware models, indicating that the benefit of local outlier-derived features was not limited to a specific stratification method.

Additional reference models were further evaluated to examine whether local outlier-derived features provide predictive value independent of spatial stratification. As shown in Table S2, adding local outlier-derived features improved OLS regression, random forest, and GWR. For example, the R^2 of random forest increased from 0.504 ± 0.430 to 0.638 ± 0.250, and the R^2 of GWR increased from 0.736 ± 0.303 to 0.796 ± 0.218. In contrast, regression kriging showed little additional improvement after adding local outlier-derived features, with R^2 changing from 0.681 ± 0.265 to 0.667 ± 0.285. This may be because part of the residual spatial structure had already been represented by the local outlier-derived predictors, reducing the additional benefit of residual kriging. Overall, these additional comparisons confirm that local outlier-derived features generally provide useful supplementary spatial information, while their effectiveness depends on the model structure.

4.5.3 SHAP-based contribution of local outlier-derived features

To provide a unified interpretation of feature contribution, SHAP analysis was conducted using a global random forest model with both original explanatory variables and local outlier-derived features. This analysis was used to provide an overall assessment of feature contribution rather than a stratum-specific interpretation of each LOCA model. SHAP values were calculated on the held-out validation samples under spatial block cross-validation.

As shown in Figure 10a, the original explanatory variables remained the dominant predictors. Evaporation and surface runoff showed the largest SHAP effects, indicating that background hydroclimatic conditions were the primary controls on precipitation prediction. Wind speed also contributed to the model output, although its SHAP values were more concentrated around zero. Among the local outlier-derived features, the positive outlier of surface runoff showed a clear contribution and ranked higher than several original and outlier-derived variables, suggesting that local anomaly information can provide additional predictive value beyond the original variables.

The grouped SHAP contribution in Figure 10b further confirms this result. Original explanatory variables accounted for 69.1% of the total SHAP contribution, while positive and negative local outlier-derived features accounted for 25.6% and 5.4%, respectively. Therefore, local outlier-derived features contributed approximately 31.0% of the total SHAP importance. This indicates that the original variables remained the main source of predictive information, but local outlier-derived features provided a non-negligible supplementary contribution. The stronger contribution of positive outlier features suggests that local high-value anomalies were more informative than local low-value anomalies in the present precipitation prediction case.

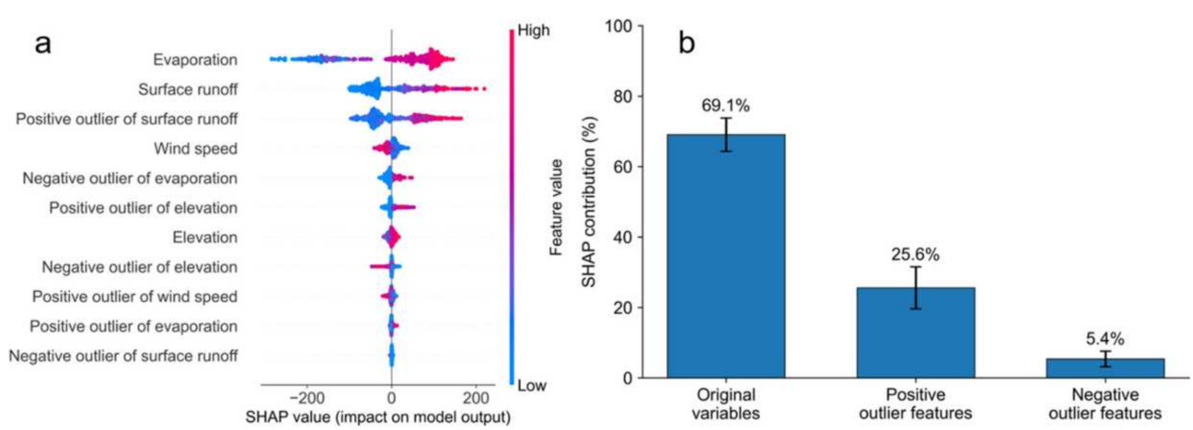

Figure 10. Overall SHAP-based contribution of original and local outlier-derived features.

These results support the interpretation that local outlier-derived features enhance model performance by representing local spatial contrasts. However, their contribution should not be interpreted as a fixed proportion across all models or strata. Accordingly, the revised manuscript no longer claims that local outliers contribute up to 50%; instead, it shows that local outlier-derived features provide useful but model-dependent supplementary information for precipitation spatial prediction.

5 Discussion

This study proposed a local-outlier-based context-aware model (LOCA) and applied it to precipitation spatial prediction in Qinghai Province. The revised validation results based on spatial block cross-validation showed that LOCA generally improved prediction performance compared with single machine-learning models and spatial context-aware models without local outlier-derived features. This improvement can be attributed to two complementary mechanisms. First, spatial stratification improves regional adaptability by allowing different strata to use locally suitable predictive relationships. Second, local outlier-derived features provide additional information on local spatial contrasts, which helps the model represent anomaly-related environmental conditions that may not be fully captured by original explanatory variables alone.

The additional reference models and SHAP analysis further support this interpretation. Models with local outlier-derived features generally outperformed their counterparts without these features, and the SHAP-based contribution analysis showed that original variables remained the dominant predictors, while local outlier-derived features provided non-negligible supplementary information. Therefore, local outliers should not be interpreted as universally dominant predictors, but as additional spatial contextual features whose usefulness depends on the target variable, environmental setting, and model structure.

Several limitations should be acknowledged. First, the current case study used annual precipitation data for Qinghai Province in 2022. Although Qinghai Province has strong topographic variation and clear precipitation gradients, the results should be interpreted as evidence from a single case rather than proof of universal superiority. Multi-year experiments and applications in other regions are needed to further evaluate the temporal and spatial transferability of LOCA.

Second, although spatial block cross-validation provides a stricter validation strategy than random splitting, the validation is still based on gridded precipitation data rather than independent high-resolution station observations. Because the target precipitation product has a coarser resolution than some explanatory variables, the present task should be interpreted as a statistical downscaling and spatial prediction experiment. Future studies should further evaluate LOCA using independent station observations or other high-quality precipitation datasets.

Third, local outlier-derived features may represent meaningful spatial signals, but they may also introduce noise if local anomalies are caused by measurement errors, resampling artifacts, or unstable neighbourhood statistics. The exclusion of the degenerate negative surface runoff outlier feature in this study also shows that local outlier construction should be combined with zero-variance checks, multicollinearity screening, and sensitivity analysis. Therefore, local outliers should be regarded as potentially useful contextual features rather than automatically beneficial predictors.

Fourth, LOCA may introduce artificial discontinuities near stratum boundaries. Because different strata may use different prediction models, neighbouring locations assigned to different strata can have abrupt prediction differences. Such discontinuities may reflect real changes in environmental controls, but they may also result from model transitions across strata. Therefore, spatial prediction maps should be interpreted together with spatial block validation metrics, residual patterns, and feature contribution analysis, rather than being used alone as evidence of improved local detail.

The applicability of LOCA may depend on data characteristics. The framework is likely to be more useful for variables with strong spatial heterogeneity, nonlinear relationships, and meaningful local anomaly structures, such as precipitation, runoff, soil properties, vegetation conditions, and environmental pollution. In contrast, LOCA may provide limited benefits for variables with weak spatial heterogeneity, very smooth gradients, sparse observations, or dominant temporal nonstationarity.

Future studies can further refine both the stratification and prediction components of LOCA. Other clustering methods may be explored to identify more flexible spatial structures, but their parameter sensitivity, spatial continuity, and potential boundary effects should be carefully evaluated. In addition, complex geospatial models such as Kriging (Jeong, Murayama, and Yamamoto 2005), regression kriging, geographically weighted regression (GWR) (Fotheringham, Brunsdon, and Charlton 2009), and geographically weighted machine-learning models may be incorporated into future context-aware modelling frameworks. Such extensions would help further examine how local anomaly information, spatial dependence, and spatially varying relationships can be jointly represented in spatial prediction.

6 Conclusions

This study proposed a local-outlier-based context-aware modelling framework and evaluated its applicability to precipitation spatial prediction in Qinghai Province in 2022. Under spatial block cross-validation, LOCA improved prediction performance compared with single machine-learning models and spatial context-aware models without local outlier-derived features. The results suggest that spatial stratification improves regional adaptability, while local outlier-derived features provide supplementary information for representing local spatial contrasts. However, the current study is limited to one study area, one target variable, and one year. Therefore, the results should be interpreted as evidence from a case study rather than as proof of universal superiority. Future studies should test LOCA using multi-year datasets, different environmental variables, independent observations, and additional spatial validation strategies.

Author contributions

CRediT: **Jinyu Meng:** Methodology, Software, Writing – original draft; **Ziqin Zheng:** Data curation, Formal analysis, Visualization; **Yongze Song:** Conceptualization, Resources, Writing – review & editing.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID

Yongze Song 0000-0003-3420-9622

Data availability statement

The data that support the findings of this study are openly available in zenodo at https://doi.org/10.5281/zenodo.20726612 and are cited in the reference list as Meng et al. (2026).

References

Adorno, G., R. H. Pereira, and S. Amaral. 2025. "Combining Spatial Clustering and Spatial Regression Models to Understand Distributional Inequities in Access to Urban Green Spaces." *Landscape and Urban Planning* 256: 105297. https://doi.org/10.1016/j.landurbplan.2025.105297.

Anselin, L. 1995. "Local Indicators of Spatial Association—Lisa." *Geographical Analysis* 27 (2): 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Asghari, K., and M. Nasseri. 2015. "Spatial Rainfall Prediction Using Optimal Features Selection Approaches." *Hydrology Research* 46 (3): 343–355. https://doi.org/10.2166/nh.2014.178.

Ben-Moshe, S., and S. Itzkovitz. 2019. "Spatial Heterogeneity in the Mammalian Liver." *Nature Reviews Gastroenterology & Hepatology* 16 (7): 395–410. https://doi.org/10.1038/s41575-019-0134-x.

Brus, D. J., J. J. De Gruijter, and J. W. Van Groenigen. 2006. "Designing Spatial Coverage Samples Using the k-means Clustering Algorithm." *Developments in Soil Science* 31: 183–192. https://doi.org/10.1016/S0166-2481(06)31014-8.

Cang, X., and W. Luo. 2018. "Spatial Association Detector (Spade)." *International Journal of Geographical Information Science* 32 (10): 2055–2075. https://doi.org/10.1080/13658816.2018.1476693.

Carpenter, S. R., and W. A. Brock. 2004. "Spatial Complexity, Resilience, and Policy Diversity: Fishing on Lake-Rich Landscapes." *Ecology and Society* 9(1): 8. https://doi.org/10.5751/ES-00622-090108.

Cen, Q., X. Zhou, and H. Qiu. 2024. "Exploration of Urban Neighborhood Blue-Green Space Quality Patterns and Influencing Factors in Waterfront Cities Based on MGWR and OPGD Models." *Urban Climate* 55: 101942. https://doi.org/10.1016/j.uclim.2024.101942.

Duan, R., L. Zhong, G. Huang, Y. Li, and F. Wang. 2024. "Quantitative Analysis of the Sensitivity and Spatial Stratified Heterogeneity of Extreme Precipitation Across River Basins." *Journal of Hydrology* 643: 131939. https://doi.org/10.1016/j.jhydrol.2024.131939.

Fang, W., C. Yang, D. Liu, Q. Huang, B. Ming, L. Cheng, L. Wang, G. Feng, and J. Shang. 2023. "Assessment of Wind and Solar Power Potential and Their Temporal Complementarity in China's Northwestern Provinces: Insights from ERA5 Reanalysis." *Energies* 16 (20): 7109. https://doi.org/10.3390/en16207109.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester: Wiley.

Fotheringham, A. S., C. Brunsdon, and M. E. Charlton. 2009. "Geographically Weighted Regression." *The Sage Handbook of Spatial Analysis* 1: 243–254. https://doi.org/10.4135/9780857020130.n13.

Gaitán, C. F. 2016. "Effects of Variance Adjustment Techniques and Time-Invariant Transfer Functions on Heat Wave Duration Indices and Other Metrics Derived from Downscaled Time-Series." *Study Case: Montreal, Canada Natural Hazards* 83 (3): 1661–1681.

Ge, Y., Y. Song, J. Wang, W. Liu, Z. Ren, J. Peng, and B. Lu. 2017. "Geographically Weighted Regression-Based Determinants of Malaria Incidences in Northern China." *Transactions in GIS* 21 (5): 934–953. https://doi.org/10.1111/tgis.12259.

Getis, A. 2009. Spatial Autocorrelation. In *Handbook of applied spatial analysis: Software tools, methods and applications*255–278, Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03647-7_14.

Guo, J., J. Wang, C. Xu, and Y. Song. 2022. "Modeling of Spatial Stratified Heterogeneity." *GIScience & Remote Sensing* 59 (1): 1660–1677. https://doi.org/10.1080/15481603.2022.2126375.

Harris, I., T. J. Osborn, P. Jones, and D. Lister. 2020. "Version 4 of the CRU TS Monthly High-Resolution Gridded Multivariate Climate Dataset." *Scientific Data* 7 (1): 109. https://doi.org/10.1038/s41597-020-0453-3.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, et al. 2020. "The ERA5 Global Reanalysis." *Quarterly Journal of the Royal Meteorological Society* 146 (730): 1999–2049. https://doi.org/10.1002/qj.3803.

Jeong, S., M. Murayama, and K. Yamamoto. 2005. "Efficient Optimization Design Method Using Kriging Model." *Journal of Aircraft* 42 (2): 413–420. https://doi.org/10.2514/1.6386.

Li, B., and F. Fonseca. 2006. "TDD: A Comprehensive Model for Qualitative Spatial Similarity Assessment." *Spatial Cognition and Computation* 6 (1): 31–62. https://doi.org/10.1207/s15427633scc0601_2.

Liu, Y., Y. Chen, Z. Wu, and B. Wang. 2021. "Geographical Detector-Based Stratified Regression Kriging Strategy for Mapping Soil Organic Carbon with High Spatial Heterogeneity." *Catena* 196: 104953. https://doi.org/10.1016/j.catena.2020.104953.

Lubbe, R., W. Xu, D. N. Wilke, P. Pizette, and N. Govender. 2020. "Analysis of Parallel Spatial Partitioning Algorithms for GPU Based Dem." *Computers and Geotechnics* 125: 103708. https://doi.org/10.1016/j.compgeo.2020.103708.

Luo, P., Y. Song, X. Huang, H. Ma, J. Liu, Y. Yao, and L. Meng. 2022. "Identifying Determinants of Spatio-Temporal Disparities in Soil Moisture of the Northern Hemisphere Using a Geographically Optimal Zones-Based Heterogeneity Model." *ISPRS Journal of Photogrammetry and Remote Sensing* 185: 111–128. https://doi.org/10.1016/j.isprsjprs.2022.01.009.

Meng, J., Zheng, Z., and Song, Y. 2026. Precipitation Prediction Dataset of Qinghai Province [Data set]. Zenodo https://doi.org/10.5281/zenodo.20726612.

Meng, J., Z. Dong, G. Fu, S. Zhu, Y. Shao, S. Wu, and Z. Li. 2024. "Spatial and Temporal Evolution of Precipitation in the Bahr El Ghazal River Basin, Africa." *Remote Sensing* 16 (9): 1638. https://doi.org/10.3390/rs16091638.

Meyer, H., C. Reudenbach, S. Wöllauer, and T. Nauss. 2019. "Importance of Spatial Predictor Variable Selection in Machine Learning applications–Moving from Data Reproduction to Spatial Prediction." *Ecological Modelling* 411: 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815.

Moran, P. A. P. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37 (1/2): 17–23. https://doi.org/10.1093/biomet/37.1-2.17.

O'Leary, B. F., A. B. Hill, C. Linn, M. Lu, C. J. Miller, A. Newman, F. G. Sperone, and Q. Zhang. 2023. "Exploring the Association of Brownfield Remediation Status with Socioeconomic Conditions in Wayne County, Mi." *Environmental Science and Pollution Research* 30: 60768–60776. https://doi.org/10.1007/s11356-023-26666-2.

Park, J., S. Lee, and K. Kim. 2021. "Analysis of Factors Influencing Forest Loss in South Korea: Statistical Models and Machine-Learning Model." *Forests* 12(12): 1636. https://doi.org/10.3390/f12121636.

Qi, W. Y., J. Chen, L. Li, C. Xu, Y. Xiang, S. Zhang, and H. Wang. 2021. "Impact of the Number of Donor Catchments and the Efficiency Threshold on Regionalization Performance of Hydrological Models." *Journal of Hydrology* 601: 126680. https://doi.org/10.1016/j.jhydrol.2021.126680.

Rahang, M., and P. K. Patowari. 2016. "Parametric Optimization for Selective Surface Modification in EDM Using Taguchi Analysis." *Materials and Manufacturing Processes* 31 (4): 422–431. https://doi.org/10.1080/10426914.2015.1037921.

Shekhar, S., P.R. Schrater, R.R. Vatsavai, W.l. Wu, and S. Chawla. 2002. "Spatial Contextual Classification and Prediction Models for Mining Geospatial Data." *IEEE Transactions on Multimedia* 4(2): 174–188. https://doi.org/10.1109/TMM.2002.1017732.

Shi, H., T. Li, and J. Wei. 2017. "Evaluation of the Gridded CRU TS Precipitation Dataset with the Point Raingauge Records over the Three-River Headwaters Region." *Journal of Hydrology* 548: 322–332. https://doi.org/10.1016/j.jhydrol.2017.03.017.

Singh, A. K., and S. Lalitha. 2018. "A Novel Spatial Outlier Detection Technique." *Communications in Statistics-Theory and Methods* 47 (1): 247–257. https://doi.org/10.1080/03610926.2017.1301477.

Song, Y. 2023. "Geographically Optimal Similarity." *Mathematical Geosciences* 55 (3): 295–320. https://doi.org/10.1007/s11004-022-10036-8.

Song, Y., J. Wang, Y. Ge, and C. Xu. 2020. "An Optimal Parameters-Based Geographical Detector Model Enhances Geographic Characteristics of Explanatory Variables for Spatial Heterogeneity Analysis: Cases with Different Types of Spatial Data." *GIScience & Remote Sensing* 57 (5): 593–610. https://doi.org/10.1080/15481603.2020.1760434.

Su, Y., T. Li, S. Cheng, and X. Wang. 2020. "Spatial Distribution Exploration and Driving Factor Identification for Soil Salinisation Based on Geodetector Models in Coastal Area." *Ecological Engineering* 156: 105961. https://doi.org/10.1016/j.ecoleng.2020.105961.

Subbu, A., and A. Ray. 2008. "Space Partitioning Via Hilbert Transform for Symbolic Time Series Analysis." *Applied Physics Letters* 92(8): 084107. https://doi.org/10.1063/1.2883958.

Sun, X. L., Y. Wu, C. Zhang, and H. Wang. 2019. "Performance of Median Kriging with Robust Estimators of the Variogram in Outlier Identification and Spatial Prediction for Soil Pollution at a Field Scale." *Science of the Total Environment* 666: 902–914. https://doi.org/10.1016/j.scitotenv.2019.02.231.

Supriadi, A., and T. Oswari. 2020. "Analysis of Geographical Information System (GIS) Design Aplication in the Fire Department of Depok City." *Technium Social Sciences Journal* 8: 1.

Tang, G. 2019. Digital elevation model of China (1KM). A Big Earth Data Platform for Three Poles.

Wang, Y., and Z. Wan. 2022. "Spatial Autocorrelation and Stratified Heterogeneity in the Evaluation of Breast Cancer Risk Inequity and Socioeconomic Factors Analysis in China: Evidence from Nanchang, Jiangxi Province." *Geospatial Health* 17(1): 1078. https://doi.org/10.4081/gh.2022.1078.

Zhang, Z., Y. Song, and P. Wu. 2022. "Robust Geographical Detector." *International Journal of Applied Earth Observation and Geoinformation* 109: 102782. https://doi.org/10.1016/j.jag.2022.102782.

Zhang, Z., Z. Li, and Y. Song. 2024. "On Ignoring the Heterogeneity in Spatial Autocorrelation: Consequences and Solutions." *International Journal of Geographical Information Science* 38(12): 2545–2571.

Zhang, Y., T. He, S. Liang, and Z. Zhao. 2023. "A Framework for Estimating Actual Evapotranspiration through Spatial Heterogeneity-Based Machine Learning Approaches." *Agricultural Water Management* 289: 108499. https://doi.org/10.1016/j.agwat.2023.108499.

Zhang, Z., Y. Song, L. Karunaratne, and P. Wu. 2024. "Robust Interaction Detector: Case of Road Life Expectancy Analysis." *Spatial Statistics* 59: 100814. https://doi.org/10.1016/j.spasta.2024.100814.

Zhao, B., Y. Zhong, A. Ma, and L. Zhang. 2016. "A Spatial Gaussian Mixture Model for Optical Remote Sensing Image Clustering." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9 (12): 5748–5759. https://doi.org/10.1109/JSTARS.2016.2546918.