GIScience & Remote Sensing

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

Generalised spatial error for validation

Haiyang Liu, Yongze Song, Pengcheng Zhang, Kai Ren & Yilong Wu

To cite this article: Haiyang Liu, Yongze Song, Pengcheng Zhang, Kai Ren & Yilong Wu (2026) Generalised spatial error for validation, GIScience & Remote Sensing, 63:1, 2690327, DOI: 10.1080/15481603.2026.2690327

To link to this article: https://doi.org/10.1080/15481603.2026.2690327

Published online: 27 Jul 2026.

GISCIENCE & REMOTE SENSING
2026, VOL. 63, NO. 1, 2690327
https://doi.org/10.1080/15481603.2026.2690327

Taylor & Francis
by informa

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Generalised spatial error for validation

Haiyang Liu[a], Yongze Song[b], Pengcheng Zhang[c], Kai Ren[d,e] and Yilong Wu[b]

[a]Graduate School of Environmental Science, Hokkaido University, Sapporo, Japan; [b]School of Design and the Built Environment, Curtin University, Perth, Australia; [c]Department of Building and Real Estate, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China; [d]State Key Laboratory of Climate System Prediction and Risk Management, Nanjing Normal University, Nanjing, People's Republic of China; [e]Key Laboratory of Virtual Geographic Environment (Ministry of Education of PR China), Nanjing Normal University, Nanjing, Jiangsu, People's Republic of China

ABSTRACT

Validation of spatial prediction models is still dominated by global metrics such as root mean square error (RMSE), which summarize average model performance but can underrepresent the spatial concentration and heterogeneity of prediction errors. To complement existing residual-diagnostic approaches, we propose the Generalised Spatial Error (GSE), a validation metric that combines local RMSE with local error variance and assigns greater weight to neighborhoods with unstable performance. We evaluated GSE using nine predictive models for vascular plant species richness in the Murray–Darling Basin (MDB), Australia. Across models, GSE was consistently 20.29–24.40% higher than RMSE, reflecting the total increase from variance-weighted aggregation of local error magnitude and instability. Maps of local error variance further showed that predictive instability was concentrated mainly in the species-rich eastern part of the basin. Sensitivity analyses based on geographical area, elevation and species-richness stratification yielded consistent differences between GSE and RMSE, supporting the robustness of GSE across contrasting environmental settings. These results suggest that GSE provides a useful complementary metric for diagnosing spatial reliability beyond average prediction accuracy and for identifying locations where model predictions are less dependable for environmental assessment and decision-making.

ARTICLE HISTORY
Received 19 April 2026
Accepted 12 June 2026

KEYWORDS
Spatial validation; generalised spatial error; local error variance; spatial heterogeneity; Murray–Darling Basin

1 Introduction

Spatial prediction models (SPMs) are widely used across disciplines such as environmental science, ecology and urban analysis to support scientific inference, spatial decision-making and resource allocation (Luo et al. 2022). Their practical value, however, depends not only on predictive performance itself but also on how that performance is evaluated (Goodchild 2004). In current practice, model validation is still dominated by global, non-spatial metrics such as root mean square error (RMSE) and the coefficient of determination (R^2), which summarise model performance using a single, easily interpretable value (Chai and Draxler 2014). Although such metrics are useful for comparing average predictive accuracy, they do not explicitly account for the spatial dependence and heterogeneity that characterise most geographic systems (Anselin 1995; Comber et al. 2021; Brunsdon et al., 1996; Fotheringham et al., 2009). As a result, models with similar global accuracy may differ substantially in their spatial reliability.

This issue is particularly important because spatial prediction errors are rarely distributed randomly. Following Tobler's First Law of Geography, nearby locations often share similar environmental and geographic conditions, and model errors may therefore exhibit spatial clustering or localised instability (Tobler 1970). A model with a relatively low global RMSE may still perform poorly in specific subregions that are especially important for management, conservation or risk assessment (Dungan et al. 2002). At the same time, the spatial structure of residuals has long been recognised in geographic analysis, and a range of post hoc diagnostic approaches have been used to examine local residual patterns and spatial autocorrelation

CONTACT Yongze Song yongze.song@curtin.edu.au

(Anselin 1995; Zhang et al. 2023). Structured residuals may reflect localised nonlinearity, omitted predictors, non-stationarity, or other forms of model misspecification. However, these diagnostic approaches are typically used to identify where model failures occur, rather than to incorporate such spatial instability directly into a single validation metric for model comparison.

Here we address this need by proposing the Generalised Spatial Error (GSE), a spatially explicit validation metric that integrates both local error magnitude and local error instability. Specifically, GSE combines local RMSE with local error variance derived from spatial neighbourhoods and assigns greater influence to areas where prediction errors are more unstable (Goovaerts 1997). In this sense, GSE is not intended to replace residual diagnostics; rather, it complements them by translating spatially structured uncertainty into a summary measure that remains comparable across models. This provides a bridge between conventional global accuracy assessment and the geographically uneven error structures that often determine the practical usefulness of spatial predictions.

We apply GSE to the prediction of vascular plant species richness in Australia's Murray–Darling Basin, a region of substantial environmental heterogeneity and strong biodiversity relevance. Using this case study, we aim to: formalise the GSE framework for spatial model validation; compare GSE with conventional global metrics across multiple predictive models; and examine whether GSE more effectively reveals spatially concentrated uncertainty under different data stratification schemes. By complementing average-accuracy metrics with an explicitly spatial perspective, this study seeks to improve the evaluation of model reliability for biodiversity assessment and other spatial decision-making applications.

2 Generalised spatial error for validation (GSE)

2.1 Concept of GSE

Validation of spatial prediction models is still commonly based on global, non-spatial metrics such as RMSE, which summarise prediction error across the entire study area using a single value (Chai and Draxler 2014). Although such metrics are useful for comparing average predictive accuracy, they do not explicitly represent how model errors vary across space. In geographic applications, however, residuals often exhibit spatial heterogeneity and spatial autocorrelation, so models with similar global error levels may differ substantially in their local reliability (Cressie 2015). Moreover, the interpretation of model performance may be influenced by the spatial support and neighbourhood delineation used in analysis, a concern related to the Modifiable Areal Unit Problem (Openshaw 1984). It should also be noted that spatial residual patterns have not been ignored in previous research; local residual diagnostics and spatial autocorrelation analyses have long been used to identify clustered model errors and local model failures (Zhang et al. 2023). However, these approaches are mainly diagnostic and are not usually incorporated into a single validation metric for direct comparison across models. To address this issue, we propose the Generalised Spatial Error (GSE), which extends model validation from a purely global summary to a neighbourhood-based assessment of spatial reliability. Conceptually, GSE first partitions the study area into a series of localised neighbourhoods according to the spatial arrangement of data points (Figure 1), then quantifies two complementary dimensions of error within each neighbourhood, namely local error magnitude (local RMSE) and local error instability (local error variance), and finally aggregates these local measures through a variance-based weighting scheme that gives greater influence to areas with less stable predictive performance (Fotheringham et al. 2009). In this way, GSE complements existing residual diagnostics by translating geographically uneven error structure into a summary metric that remains comparable across models.

2.2 Calculation process of GSE

The Generalised Spatial Error (GSE) is calculated through a sequential neighbourhood-based procedure that re-aggregates local error information into a single comparable validation metric. The input is the residual vector generated by any predictive model, denoted as e_i (Equation (1)), where y_i is the observed value and $\widehat{y}_i$ is the corresponding prediction. For a given neighbourhood size K, the study area is partitioned using the k-Nearest Neighbours (k-NN) algorithm into a set of adaptive local neighbourhoods S_k. The use of k-NN ensures that each local unit contains a comparable number of observations despite spatially uneven sample density, thereby providing a stable basis for local error estimation. Within each neighbourhood S_k, two local

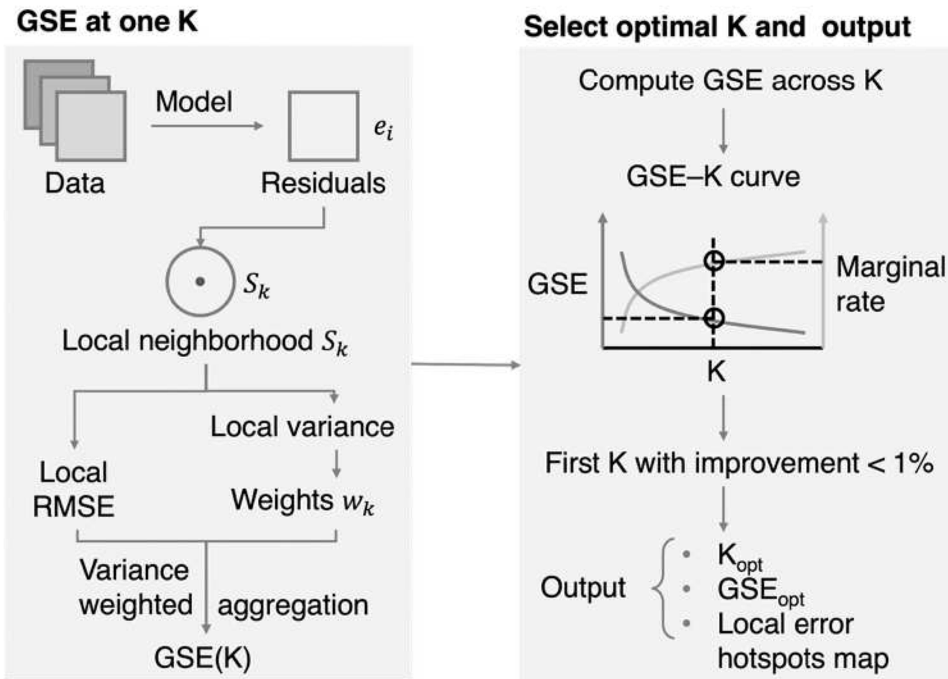

Figure 1. Computational flowchart of the Generalised Spatial Error (GSE) and the process for optimal scale determination.

statistics are computed simultaneously: the local Root Mean Square Error, $RMSE_k$ (Equation (2)), which captures the magnitude of prediction error, and the local error variance, σ_k^2 (Equation (3)), which captures the instability of model performance, where $\underline{e}_k$ is the mean residual within neighbourhood S_k. These local statistics are then re-aggregated through a variance-based weighting scheme, in which each neighbourhood receives a weight w_k proportional to its local error variance (Equation (4)), so that neighbourhoods with more unstable errors contribute more strongly to the final metric. The final GSE value is obtained as the weighted root mean square of all local $RMSE_k$ values (Equation (5)). In this way, GSE does not replace local residual diagnostics, but transforms geographically uneven error structure into a single spatially informed validation value that remains directly comparable across models.

$$e_i = y_i - \hat{y}_i \tag{1}$$

$$RMSE_k = \sqrt{\frac{1}{n_k} \sum\nolimits_{i \in S_k} e_i^2}, \quad \text{for all } i \text{ in } S_k \tag{2}$$

$$\sigma_k^2 = \frac{1}{n_k} \sum\nolimits_{i \in S_k} (e_i - \underline{e}_k)^2, \ \text{for all } i \text{ in } S_k \tag{3}$$

$$w_k = \frac{\sigma_k^2}{\sum_{j=1}^{K} \sigma_j^2}, \ \text{for all } j = 1..K \tag{4}$$

$$GSE = \sqrt{\sum\nolimits_{k=1}^{k} w_k \cdot {RMSE_k}^2}, \ \text{for all } k = 1..K \tag{5}$$

3 Case study: GSE metric for the validation of species richness prediction in the Murray-Darling Basin

The Murray–Darling Basin (MDB) is located in southeastern Australia (Figure 2a) and covers approximately 1.06×10^6 km^2, making it the country's largest river basin and a major agricultural region (Crosbie et al. 2023). The basin spans strong environmental gradients, from the temperate, high-altitude landscapes of the

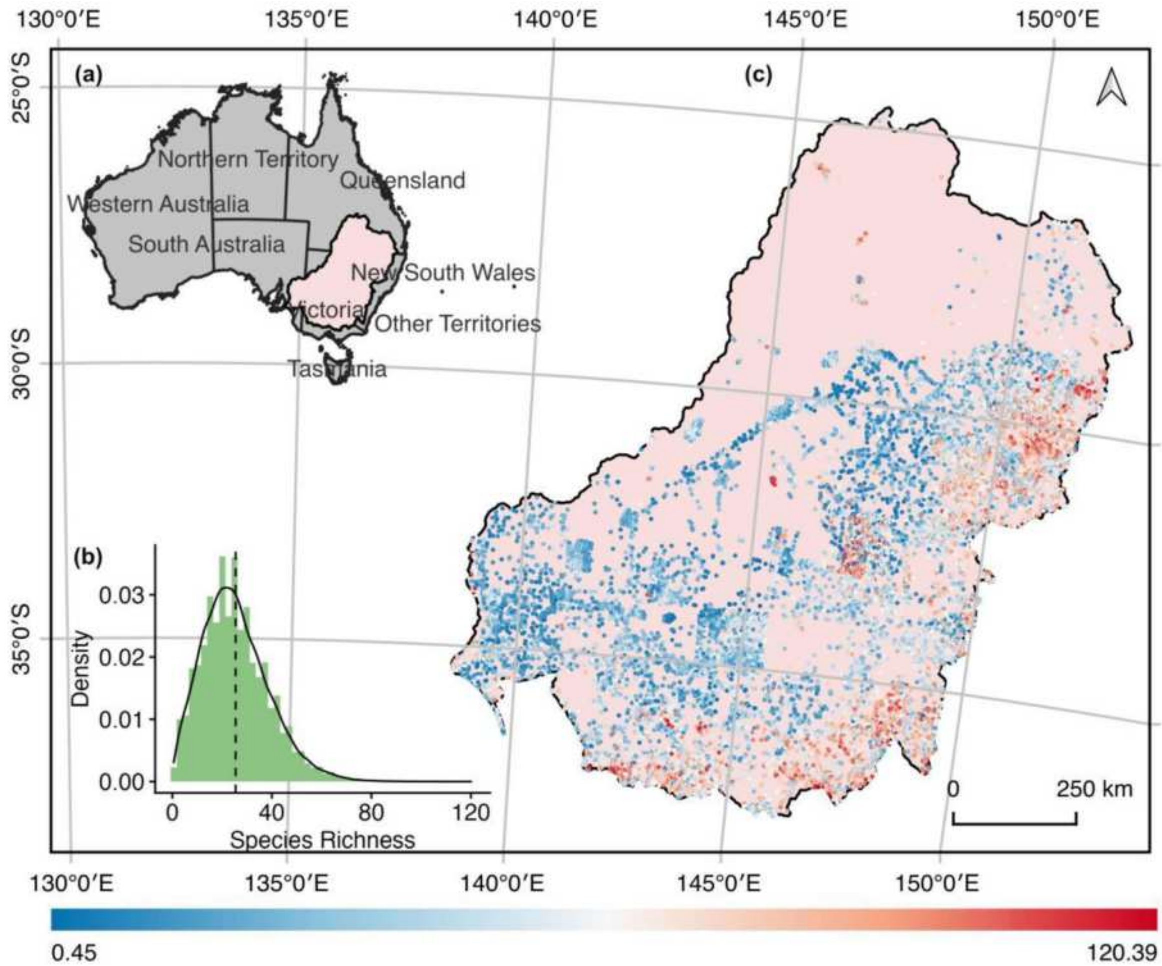

Figure 2. Study area and vascular plant species richness data used for model development and validation in the Murray–Darling Basin (MDB), Australia. (a) Geographic location of the MDB within Australia. (b) Density distribution of observed species richness across all sample plots, with the dashed vertical line indicating the mean value. (c) Spatial distribution of sample plots across the MDB, with point colour representing observed species richness (0.45–120.39 species per plot).

Great Dividing Range in the east to the semi-arid, low-lying plains in the west, resulting in substantial climatic, topographic and ecological heterogeneity (CSIRO 2012). This heterogeneity supports a wide range of ecosystems, including alpine grasslands, temperate woodlands and extensive riparian forests, and makes the MDB an important region for biodiversity conservation in Australia (Pittock and Finlayson 2011). Owing to its broad environmental gradients and spatially variable species richness patterns, the MDB provides a suitable case study for examining how model performance and prediction uncertainty vary across space.

3.1 Plant diversity data

The response variable in this study is vascular plant α-diversity, defined as the number of unique vascular plant species recorded in each sample plot. The data were obtained from the Harmonised Australian Vegetation Plot dataset (HAVPlot) developed by Mokany et al. (2022). From this database, we extracted 27,689 plot records located within the Murray–Darling Basin (MDB). Following the species–area standardisation adopted by Mokany et al. (2022), each record represents standardised species richness for a 400 m^2 sample plot (Rosenzweig 1995). Across the MDB, observed species richness has a mean of 25.44, a standard deviation of 13.40, and ranges from 0.45 to 120.39 species per plot. As shown in Figure 2b,c, the dataset exhibits both a broad richness distribution and marked spatial heterogeneity, with relatively high-richness plots concentrated mainly in the more humid eastern and southeastern parts of the basin. These data provide the observational basis for model training, validation, and subsequent GSE-based analysis.

3.2 Explanatory data

To predict vascular plant species richness, we assembled 12 explanatory variables representing climatic, topographic, edaphic, and landscape conditions in the Murray–Darling Basin (Figure 3; Table 1) (Guisan and Zimmermann 2000). Climatic variables included long-term mean short-wave radiation and wind speed derived from the FLDAS dataset at 5000 m resolution. Topographic variables included elevation, slope, sin (Aspect), and cos(Aspect), calculated from the SRTM dataset. Edaphic variables were extracted from the CSIRO TERN topsoil (0–5 cm) product at 90 m resolution and included soil depth, total nitrogen, pH, and cation exchange capacity. Landscape variables were derived from the 2015 Digital Earth Australia land-cover dataset and included distance to water bodies and distance to artificial land, representing hydrological proximity and human disturbance, respectively (Lymburner et al. 2015). Together, these variables characterise multiple environmental dimensions that are likely to influence species richness patterns across the basin. Detailed descriptions and data sources are provided in Table 1.

3.3 Experiment design

3.3.1 Data pre-processing

Prior to model construction, we conducted a Pearson correlation analysis (r) and a Variance Inflation Factor (VIF) test on all explanatory variables to ensure model stability and the interpretability of the results (Kutner 2005; Legendre and Legendre 2012). The heatmap of the Pearson correlation coefficient matrix (Figure 4) shows that all inter-variable correlation coefficients (r) are less than 0.8, and all VIF values are less than 5.

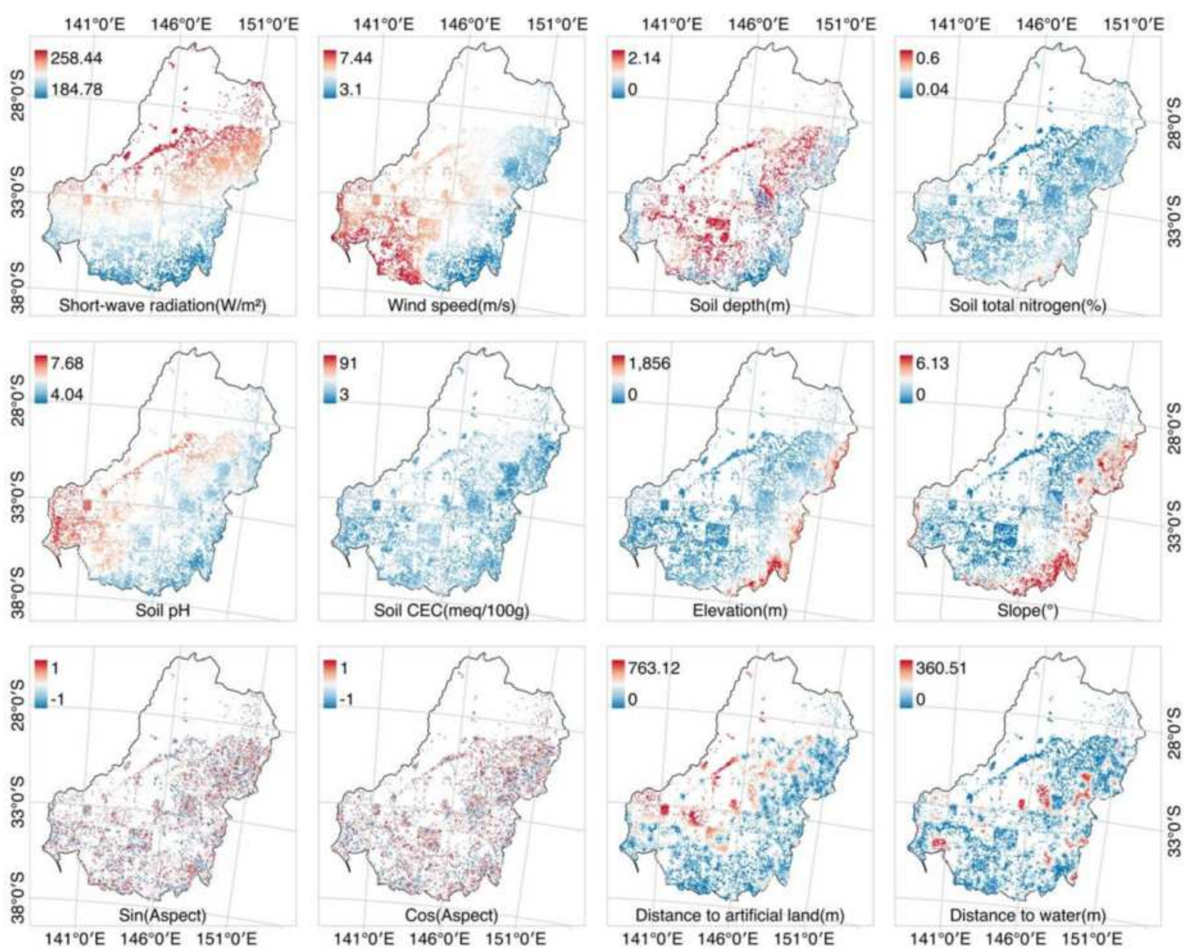

Figure 3. Spatial distribution of the explanatory variables used for constructing the vascular plant diversity prediction model in the Murray–Darling Basin (MDB) region.

Table 1. Summary of explanatory variables used for constructing the vascular plant diversity prediction model in the MDB region.

Category	Variable name	Description	Spatial resolution	Source
Climatic	Short-wave radiation	Long-term mean daily solar radiation (1982–2015), the primary energy source for the ecosystem.	5000m	FLDAS
	Wind speed	Long-term mean annual wind speed (1982–2015), influencing plant water stress and dispersal.	5000m	FLDAS
Topographic	Elevation	Altitude above sea level, serving as a proxy for broad-scale climatic conditions.	5000m	SRTM
	Slope	The gradient of the land surface, controlling surface water flow and soil erosion potential.	5000m	SRTM
	sin(Aspect)	A linear transformation of aspect quantifying the east-west exposure (eastness).	5000m	SRTM
	cos(Aspect)	A linear transformation of aspect quantifying the north-south exposure (northness), related to solar insolation.	5000m	SRTM
Edaphic	Soil depth	Depth to bedrock or impermeable layer.	90m	CSIRO
	Soil total N	Overall nitrogen concentration in the topsoil (0–5 cm).	90m	CSIRO
	Soil pH	The pH level of the topsoil (0–5 cm).	90m	CSIRO
	Soil CEC	The capacity of the topsoil (0–5 cm) to adsorb and exchange nutrient cations.	90m	CSIRO
Landscape	Distance to artificial land	Distance to the nearest artificial surface.	250m	DEA
	Distance to water	Distance to the nearest water source.	250m	DEA

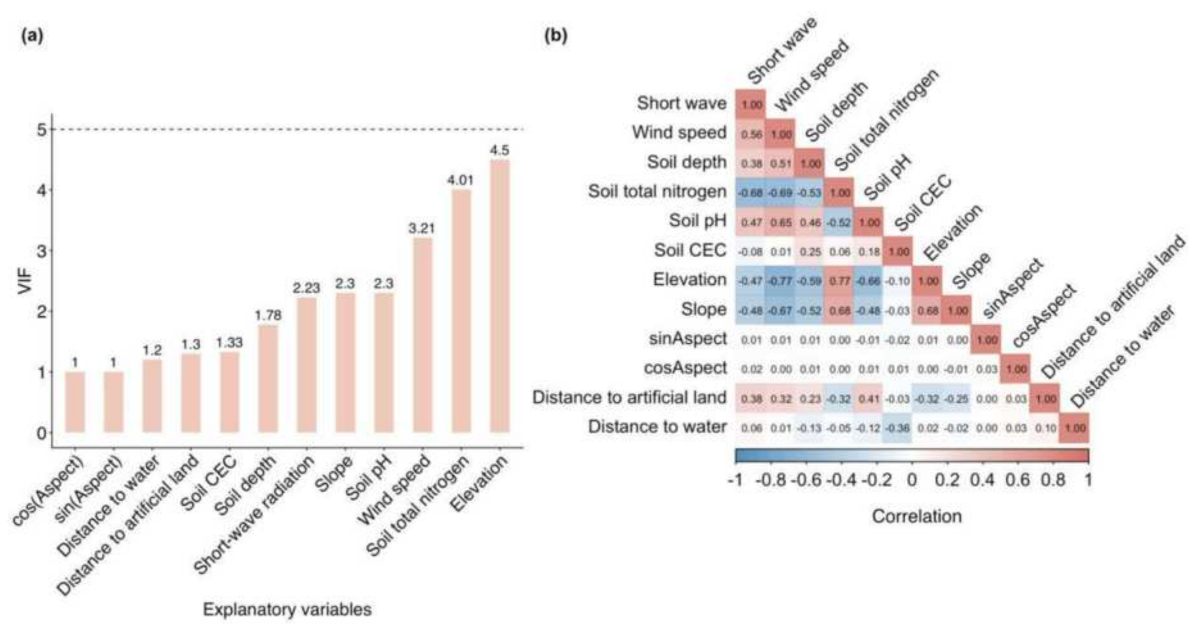

Figure 4. Multicollinearity diagnosis of the explanatory variables used for species richness prediction in the Murray–Darling Basin (MDB). (a) Variance Inflation Factor (VIF) values for all explanatory variables, with the dashed horizontal line indicating the threshold of VIF = 5. (b) Pearson correlation matrix of the explanatory variables, where colours and numeric labels indicate the strength and direction of pairwise correlations.

This indicates that no severe multicollinearity issues are present in the dataset (Daoud 2017; Dormann et al. 2013). Therefore, all 12 explanatory variables were retained for the subsequent construction of the predictive models. All raster layers were projected to a common coordinate system, and covariate values were extracted at plot locations. For variables with different native resolutions, values were used at their native support during point extraction rather than resampled to a common fine grid.

3.3.2 Model fitting

To predict vascular plant species richness, we evaluated nine statistical and machine-learning models representing different algorithmic families (Table 2). All models were implemented in R using the caret framework to ensure a consistent training and evaluation workflow across methods (Kuhn 2008). The full

Table 2. Summary of the nine predictive models used for vascular plant species richness modelling in the Murray–Darling Basin (MDB).

Type	Full Name	Acronym	R Package
Linear models	Linear Model	lm	stats
Tree-based models	Random Forest	rf	RandomForest
	Extreme Gradient Boosting	xgbTree	xgboost
	Stochastic Gradient Boosting Machine	gbm	gbm
	Cubist	cubist	Cubist
Neighbourhood-based models	k-Nearest Neighbours	knn	kknn
Non-linear and adaptive regression models	Generalised Additive Model with Splines	gam	mgcv
Support vector machine-based models	Support Vector Machine (Radial Kernel)	svmRadial	kernlab
Spatially explicit regression models	Geographically Weighted Regression	gwr	GWmodel

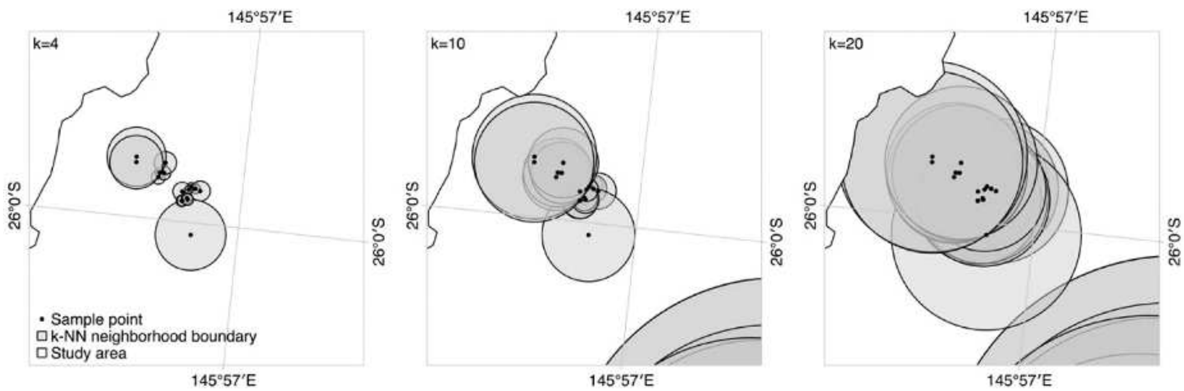

Figure 5. Schematic illustration of k-nearest neighbour (k-NN) spatial neighbourhoods under different neighbourhood sizes in the MDB. From left to right, panels show the local neighbourhood boundaries generated with K = 4, K = 10, and K = 20, respectively, overlaid on the spatial distribution of observed species richness.

dataset was randomly divided into a training set (70%) and an independent test set (30%; $n = 8{,}036$), with the test set reserved for final model assessment. Within the training set, 5-fold cross-validation was used for model fitting and hyperparameter tuning (Hastie 2009). Model performance was evaluated on the independent test set using the coefficient of determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) (Chai and Draxler 2014). This design enabled direct comparison of predictive performance across the nine models.

3.3.3 GSE-based model validation

To evaluate the spatial structure of prediction errors on the common holdout set, we determined an optimal neighbourhood size (K) for the calculation of GSE, so that local error magnitude and instability could be summarised at a comparable analysis scale (Dungan et al. 2002). For each model, GSE values were first calculated across a sequence of increasing neighbourhood sizes (Figure 5) and then smoothed using a locally weighted regression (LOESS) curve to characterise the GSE– K relationship. Based on the fitted values at the original discrete K levels, we calculated the marginal rate of change between successive neighbourhood sizes as $\left(\frac{y_{next} - y_{previous}}{y_{previous}} \times 100\right)$, where y denotes the fitted GSE value. Because GSE generally decreases with increasing K, the marginal rate is typically negative, and values closer to zero indicate diminishing change. The practical stabilisation scale was defined as the first K at which three consecutive marginal-rate values were all greater than −1%, indicating that the decline in fitted GSE had become consistently small (Song et al. 2021). This rule was used as an operational heuristic rather than a strict statistical optimum. If no such sequence was identified, K was set to the neighbourhood size corresponding to the minimum fitted GSE value. We adopted adaptive k-nearest neighbour (k-NN) neighbourhoods because they provide comparable local sample support under spatially uneven point density and yield an interpretable basis for estimating local RMSE and local error variance. Alternative neighbourhood definitions, such as distance-band neighbourhoods, Delaunay triangulation, and spatial weight matrices, are also plausible, but they may produce unequal effective local sample support or introduce additional parameter-selection issues. We therefore use adaptive k-NN here as a transparent baseline implementation of GSE that supports stable local error estimation and direct comparison with RMSE on the same holdout set. The resulting GSE_{opt} value was used for cross-model comparison, while maps of local error variance (σ_k^2) were used to identify spatial hotspots of unstable predictive performance (MacEachren and Kraak 2001). This study-wide optimal K for each model should therefore be interpreted as a practical summary scale for comparison rather than as evidence that the spatial footprint of model error is spatially uniform throughout the study area.

3.3.4 Comparing GSE and RMSE for Local Error Instability

To examine the extent to which conventional global metrics underrepresent spatially uneven prediction error, we directly compared GSE_{opt} with RMSE. RMSE summarises average error across the full study area, whereas GSE_{opt} additionally incorporates local error instability through variance-based weighting. The relative difference, $\frac{(GSE_{opt} - RMSE)}{RMSE}$, was therefore used as an operational indicator of the total increase in validation error produced by variance-weighted local aggregation. This increase was interpreted as reflecting local error magnitude and instability, rather than as a pure measure of residual spatial dependence (Krause et al. 2005). To relate this increase to an established spatial diagnostic, we also calculated Global Moran's I for the test-set residuals of each model using the same model-specific optimal kNN scale, and examined its Pearson and Spearman correlations with the relative GSE increase. This comparison quantifies how local error heterogeneity affects the effective validation error, while distinguishing GSE from residual autocorrelation diagnostics and using local error variance maps separately to identify hotspot regions of unstable predictive performance.

3.3.5 GSE validation: permutation test and sensitivity analysis based on data stratification

To examine the robustness of GSE and to separate the effect of residual spatial arrangement from residual magnitude distribution, we conducted two validation analyses based on the independent test-set residuals. First, we performed a sensitivity analysis based on data stratification (Lilburne and Tarantola, 2009). The test dataset was stratified by three criteria: (1) geographical area, with the study region divided into three spatial subregions (Figure 6a); (2) observed species richness, with samples grouped into low-, medium-, and high-richness classes based on quantiles (Figure 6b); and (3) elevation, with samples divided into low, medium, and high-elevation groups using the 33% and 66% quantiles (Figure 6c). For each subset, RMSE and GSE_{opt} were recalculated for all nine models to assess whether their differences remained consistent across contrasting spatial and environmental settings.

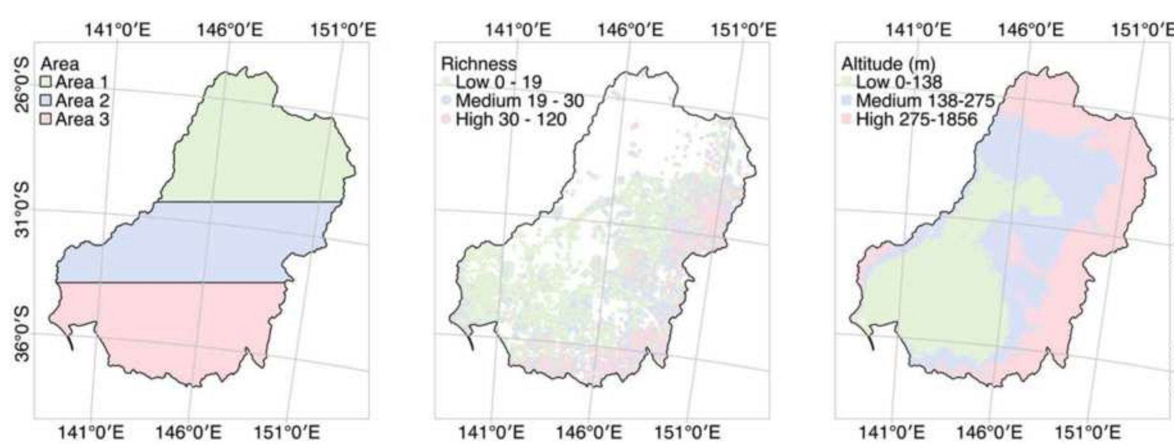

Figure 6. Data stratification schemes used for sensitivity analysis of model validation in the MDB. (a) Stratification by geographical area, where the study region was divided into three spatial subregions (Area 1–Area 3). (b) Stratification by observed species richness, where sample plots were grouped into low-, medium-, and high-richness classes. (c) Stratification by elevation, where sample plots were grouped into low-, medium-, and high-elevation classes.

Second, we conducted a permutation validation to assess whether the observed GSE exceeded the level expected from variance-weighted aggregation alone. Residual values were randomly reassigned to test locations while keeping their values unchanged, thereby preserving the residual magnitude distribution and RMSE but disrupting the spatial arrangement of errors. GSE was recalculated using the model-specific optimal kNN scale, and this procedure was repeated 999 times to generate a null distribution. The relative difference was calculated as $\frac{GSE_{obs} - GSE_{perm}}{GSE_{perm}} \times 100$, where GSE_{obs} is the observed GSE and GSE_{perm} is the mean permuted GSE.

4 Results

4.1 Model performance assessment based on global, non-spatial metrics

As a baseline evaluation, we first assessed the nine predictive models using three global, non-spatial metrics: R^2, RMSE, and MAE. The results on the independent test set are summarised in Table 3 and Figure 7. Among the nine models, rf achieved the best overall performance, with the highest R^2 (0.473) and the lowest RMSE (9.751) and MAE (7.251). Cubist and knn also performed relatively well, with R^2 values of 0.446 and 0.392, respectively. In contrast, lm showed the weakest predictive performance, with the lowest R^2 (0.229) and the highest RMSE (11.796) and MAE (9.086). The spatially explicit GWR model also performed competitively, with $R^2 = 0.450$ and RMSE = 9.981. The remaining models showed intermediate performance. Overall, these global metrics provide a baseline ranking of model accuracy on the test set, which is used in the following sections as a reference for comparison with the spatially explicit GSE-based evaluation.

4.2 Optimal scale determination and spatial error patterns based on GSE

Because GSE depends on neighbourhood size, we next examined how GSE changed with increasing K in order to determine a practical analysis scale for each model. As shown in Figure 8, GSE generally declined rapidly at small neighbourhood sizes and then decreased more gradually as K increased, indicating a diminishing change in the spatial error summary at larger neighbourhood sizes. To characterise this pattern, the GSE–K relationship was smoothed using LOESS, and marginal rates of change were calculated from fitted values at successive discrete K levels. The selected optimal K was defined as the first neighbourhood size for which three consecutive marginal-rate values were all greater than −1%, indicating that the decline in fitted GSE had become consistently small. Under this rule, the selected K values ranged from 19 to 27 across the nine models, including 19 for lm, 25 for rf, 23 for xgbTree, 23 for gbm, 25 for knn, 21 for gam, 27 for gwr, 23 for svmRadial, and 25 for cubist. Although these values differed slightly among models, they were concentrated within a relatively narrow range, supporting the use of model-specific but comparable neighbourhood scales in the subsequent GSE-based analysis. These selected scales were then used to calculate GSE_{opt}, while spatial patterns of local error variance were examined separately to identify areas of unstable predictive performance.

Table 3. Comprehensive comparison of model performance for the nine plant species richness prediction models, based on global, non-spatial metrics.

Model	R^2	RMSE	MAE
rf	0.473	9.751	7.251
cubist	0.446	10.011	7.475
knn	0.392	10.524	7.908
svmRadial	0.372	10.677	8.013
xgbTree	0.37	10.672	8.157
gbm	0.362	10.739	8.227
gam	0.324	11.041	8.44
lm	0.229	11.796	9.086
gwr	0.450	9.981	7.475

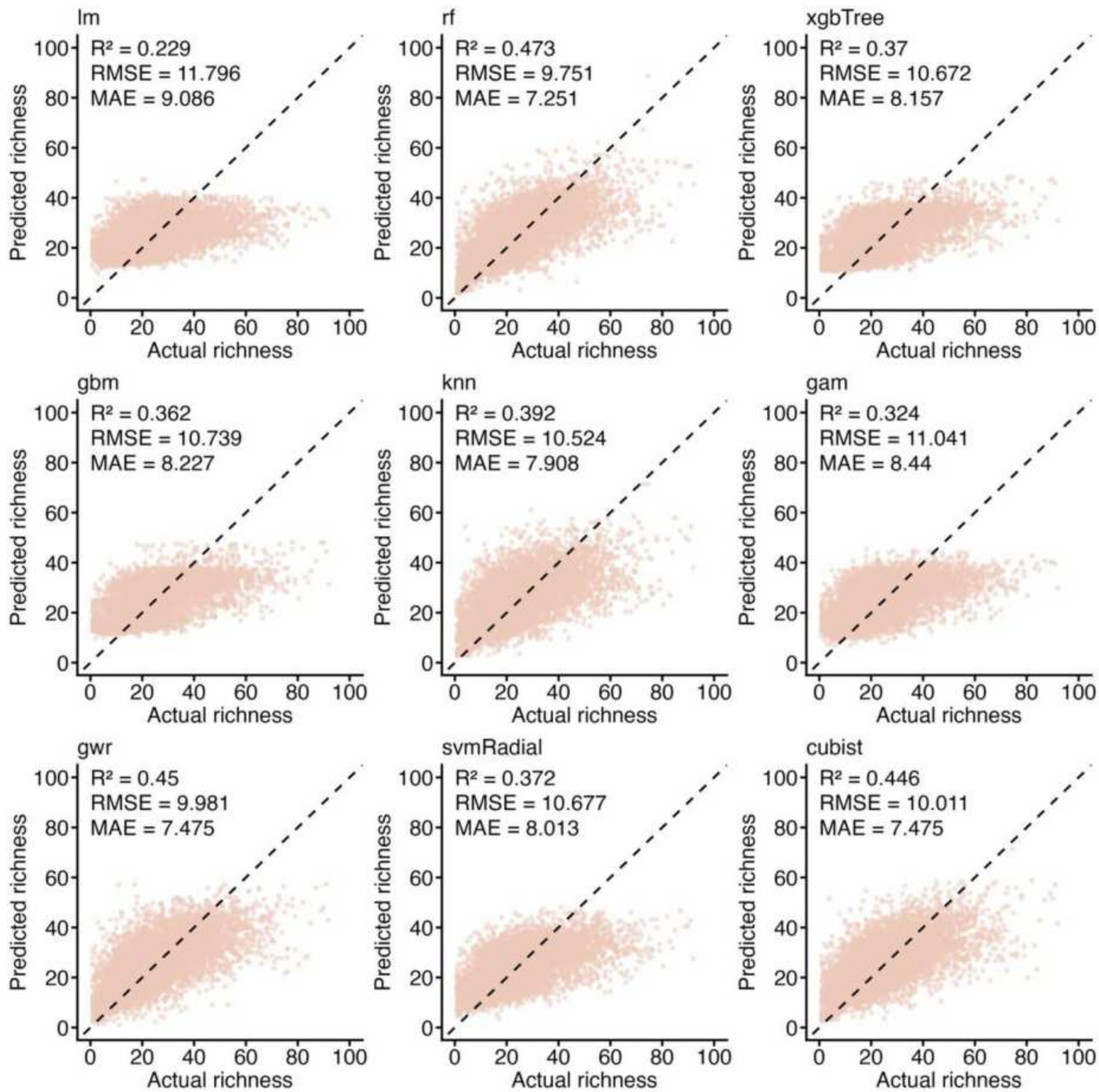

Figure 7. Scatter-plot comparison of model fit for nine vascular plant species richness prediction models in the MDB. Each panel presents actual versus predicted richness values for one model, with the dashed 1:1 line indicating perfect agreement between observations and predictions. R^2, RMSE, and MAE are provided to summarise model performance.

4.3 Spatial patterns of model error and "Hotspot" Diagnosis

Analysis of local error variance (σ_k^2) at the selected neighbourhood scales revealed a broadly consistent spatial pattern across the nine models. Given this cross-model consistency, Figure 9 presents the random forest (rf) model as a representative example, with inset maps highlighting major high-variance subregions. In all models, elevated local error variance was concentrated mainly in the eastern and southeastern parts of the Murray–Darling Basin, whereas lower values were more common in the western basin. The recurrence of these high-variance areas across models indicates that spatial instability in prediction error is not solely model-specific, but is also associated with shared characteristics of the study area. These patterns should not be interpreted as identifying the exact causes of model error. As an exploratory check, we further examined the relationship between observed species richness and local error variance using the rf model. Species richness was significantly and positively correlated with local error variance (Pearson $r = 0.407$, $p < 0.001$; Spearman $\rho = 0.430$, $p < 0.001$), suggesting that the eastern high-richness region was associated with greater local error instability. Rather, they indicate locations where prediction performance

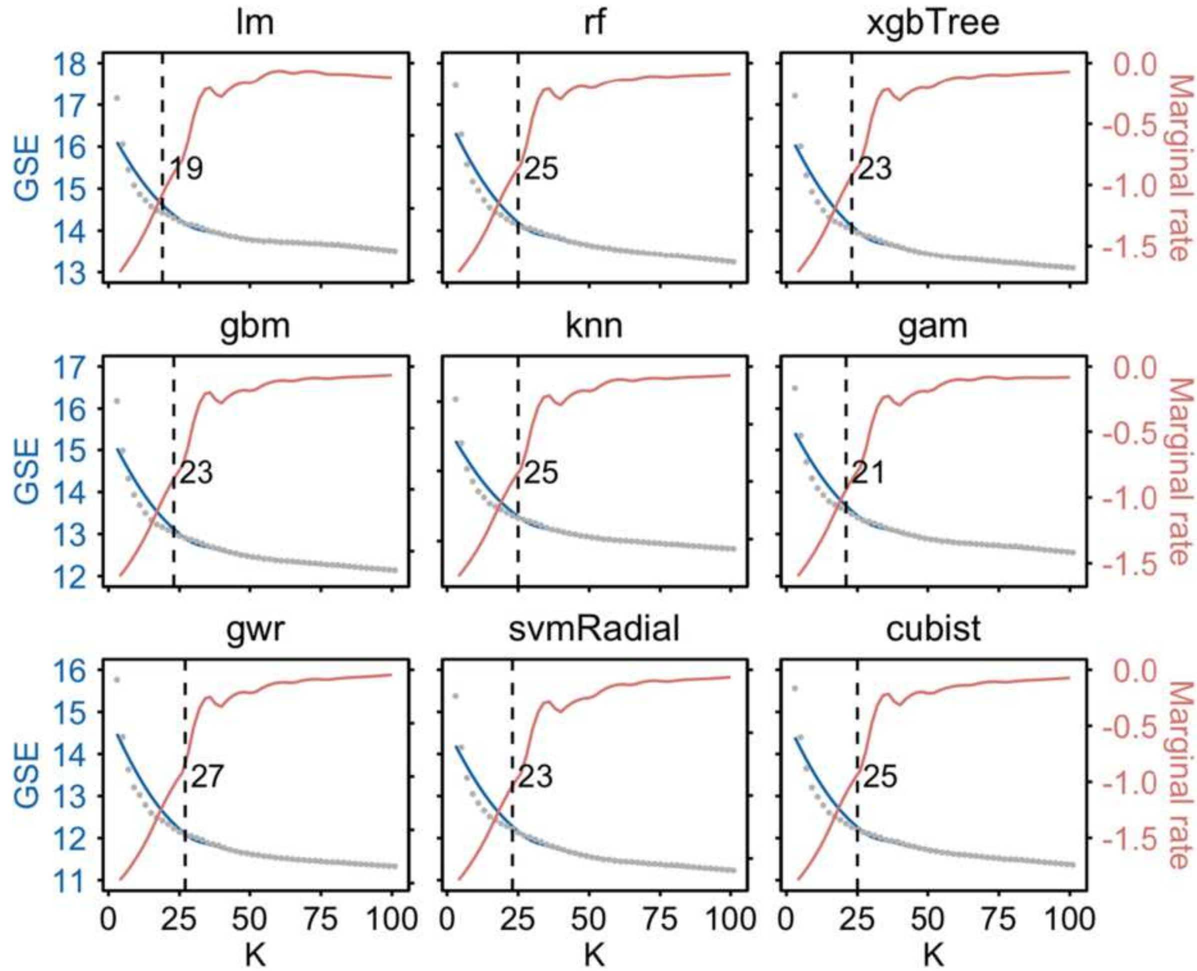

Figure 8. Model-specific determination of the optimal neighbourhood size for GSE-based validation in the Murray–Darling Basin (MDB). Each panel shows GSE values across increasing neighbourhood sizes (K; grey points), the LOESS-fitted GSE trend (blue line), and the marginal rate of change calculated from fitted values at successive discrete K levels (red line). The dashed vertical line and adjacent label indicate the selected optimal K, defined as the first neighbourhood size for which three consecutive marginal-rate values exceed −1%.

is less stable and where further investigation—such as the inclusion of additional predictors or the use of complementary residual diagnostics—may be warranted. In this sense, the GSE framework complements global accuracy metrics by linking model evaluation to the spatial distribution of predictive uncertainty.

4.4 Comparison of GSE and RMSE

To examine the extent to which global error metrics underrepresent locally uneven prediction error, we compared GSE_{opt} with RMSE across the nine predictive models. As shown in Table 4, GSE_{opt} was consistently higher than RMSE for all models, with relative differences $(GSE_{opt} - RMSE)/RMSE$ ranging from 20.29% to 24.40%. This difference reflects the total increase produced by variance-weighted aggregation of local error magnitude and instability, rather than a pure measure of residual spatial dependence. We further calculated Global Moran's I for the test-set residuals. All models showed positive and significant residual spatial autocorrelation, indicating that spatial structure remained in the prediction errors. However, the relationship between residual Moran's I and the proportional increase of GSE over RMSE was weak and not statistically significant, with a Pearson correlation of $r = 0.114$ ($p = 0.770$) and a Spearman rank correlation of $\rho = 0.200$ ($p = 0.606$). The GWR model further illustrates this distinction: despite its very low residual Moran's I value of 0.013, its GSE remained 21.12% higher than RMSE. This indicates that Moran's I and GSE capture related but non-equivalent aspects of residual structure. Moran's I measures global residual autocorrelation,

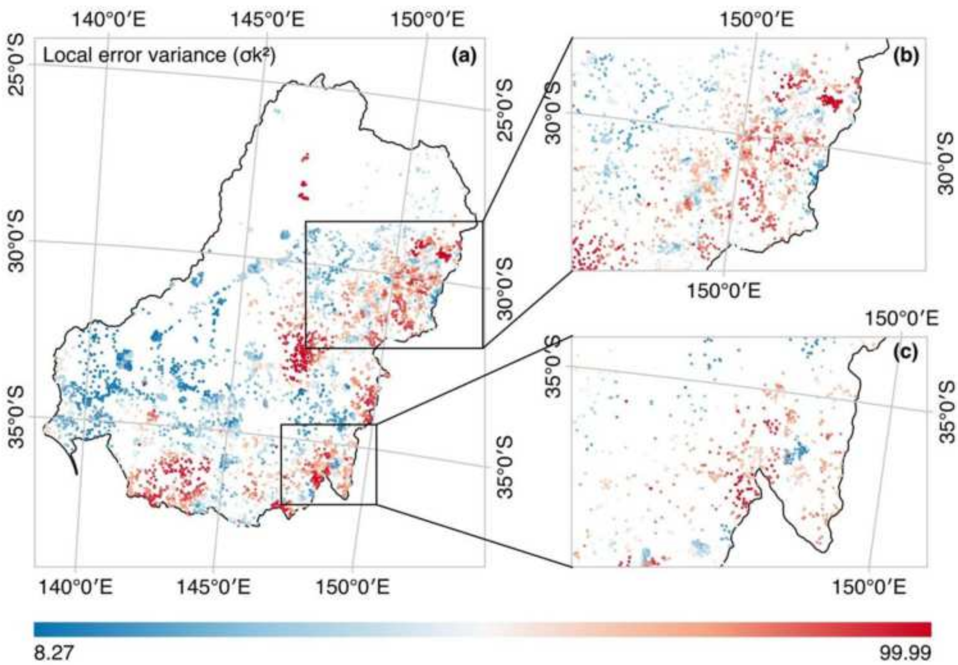

Figure 9. Spatial distribution of local error variance (σ_k^2) for the random forest model at its optimal scale, with inset maps showing major high-variance subregions.

Table 4. Comparison of GSEopt and RMSE for nine predictive models.

Model	Optimal kNN scale	GSE	Global Moran's I	RMSE	Total GSE increase over RMSE
gwr	27	12.089	0.013	9.981	21.12%
rf	25	12.130	0.027	9.751	24.40%
cubist	25	12.199	0.03	10.011	21.85%
knn	25	12.659	0.048	10.524	20.29%
xgbTree	23	13.019	0.101	10.672	22.00%
gbm	23	13.031	0.108	10.739	21.34%
svmRadial	23	13.215	0.099	10.677	23.77%
gam	21	13.552	0.142	11.041	22.74%
lm	19	14.427	0.251	11.796	22.31%

whereas GSE_{opt} integrates local error magnitude and local error instability through variance-weighted aggregation. Taken together, these results suggest that RMSE provides a useful baseline measure of average prediction accuracy, while GSE provides a complementary summary of local predictive instability.

4.5 GSE validation: sensitivity analysis based on data stratification

Across all three stratification schemes, the comparison between GSE_{opt} and RMSE showed a consistent pattern (Figure 10): all points lay above the 1:1 line, indicating that GSE_{opt} was higher than RMSE in every subset. This suggests that the additional spatially uneven error captured by GSE remained evident under area-, elevation-, and richness-based stratifications. Performance also varied among subsets. In particular, the richness-based stratification showed a clear gradient from lower errors in low-richness areas to higher errors in high-richness areas, while the elevation-based stratification showed generally higher errors in medium- and high-elevation subsets than in the low-elevation subset. Geographical stratification also revealed differences among subregions, although these were less systematic. Together, these results support the robustness of GSE as a spatially explicit complement to global accuracy metrics.

The permutation validation further separated the total GSE increase from the component associated with residual spatial arrangement (Table 5). After randomly reassigning residuals to test locations, the mean permuted GSE remained higher than RMSE, indicating that local variance-weighted aggregation still produces a larger error estimate than global averaging. More importantly, the observed GSE was

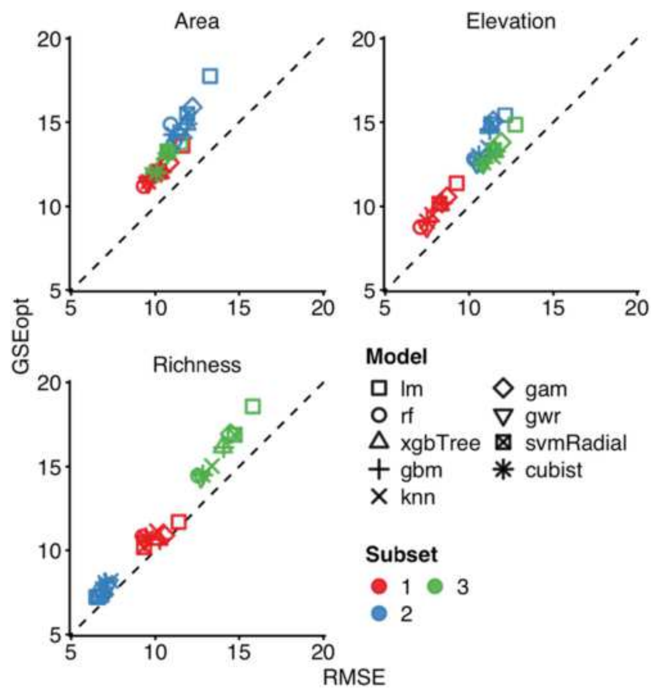

Figure 10. Sensitivity analysis of model performance based on the relationship between GSE_{opt} and RMSE under three stratification schemes in the Murray–Darling Basin (MDB). Panels show results for geographical area, elevation, and species richness stratification, respectively. Colours indicate subsets within each stratification scheme, symbols denote predictive models, and the dashed line represents the 1:1 reference line.

Table 5. Permutation validation of the incremental spatial information captured by GSE.

Model	Optimal kNN scale	Observed GSE	Mean permuted GSE	ΔGSE (%)	*p*-value
rf	25	12.130	10.530	15.194	0.001
svmRadial	23	13.215	11.546	14.455	0.001
gam	21	13.552	11.934	13.558	0.001
lm	19	14.427	12.798	12.729	0.001
xgbTree	23	13.019	11.473	13.475	0.001
cubist	25	12.199	10.798	12.975	0.001
gbm	23	13.031	11.530	13.018	0.001
knn	25	12.659	11.293	12.096	0.001
gwr	27	12.089	10.691	13.077	0.001

consistently higher than the mean permuted GSE across all nine models, with relative increases ranging from 12.096% to 15.194% and significant permutation tests in all cases ($p = 0.001$). This indicates that the observed increase in GSE is unlikely to arise from residual magnitude alone, but reflects spatially organised local error instability.

5 Discussion

5.1 Method for spatial local analysis: selection and validation of k-NN

A key methodological issue in the GSE framework is how local neighbourhoods are defined for spatial error analysis. In this study, we used adaptive k-nearest neighbour (k-NN) neighbourhoods because they provide comparable local sample support under spatially uneven point density. To assess this choice, we also compared k-NN with non-adaptive regular hexagonal and square grids (Wong 2004). In our dataset, the grid-based approaches produced less stable and less interpretable relationships between GSE and analysis scale (Figures 11 and 12), likely because fixed geometric partitions do not adjust to the observed spatial distribution of sample points (Fotheringham et al. 2009). This instability is also consistent with the Modifiable Areal Unit Problem (MAUP), in which statistical results may vary with the scale and zoning of

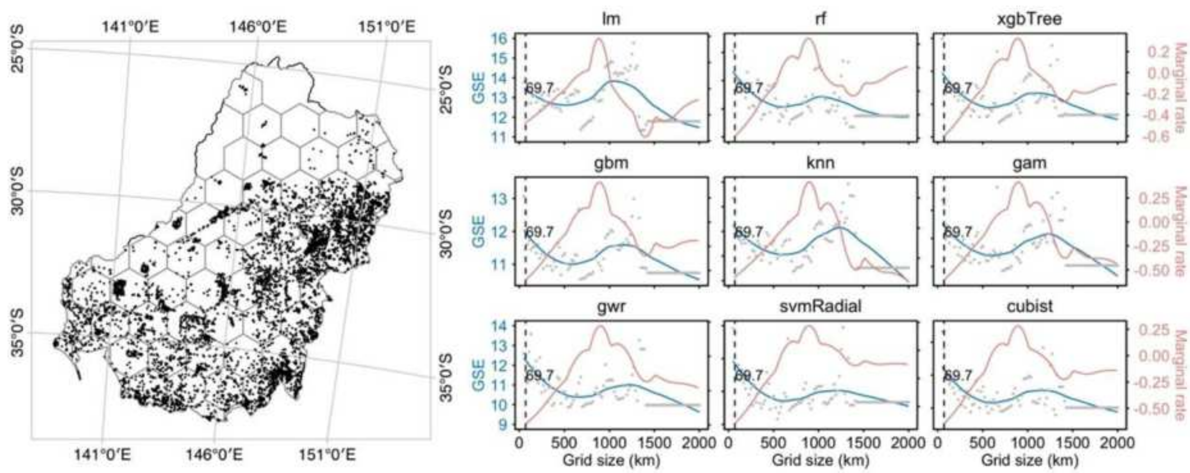

Figure 11. Relationship between GSE and spatial scale under regular hexagonal-grid local analysis in the Murray–Darling Basin (MDB). The left panel shows the hexagonal partition of the study area, and the right panels show model-specific GSE responses to increasing grid size together with the corresponding marginal-rate curves.

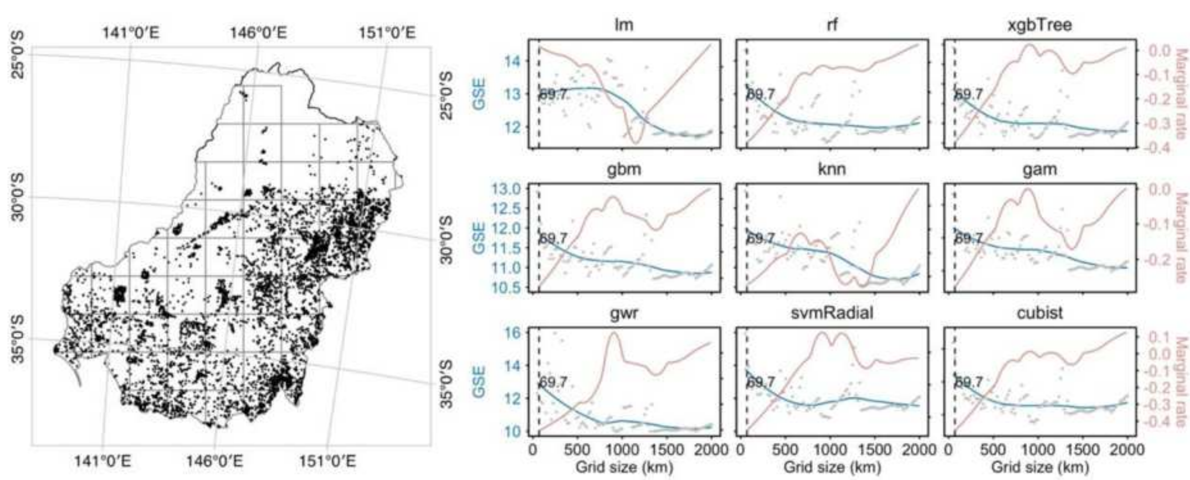

Figure 12. Relationship between GSE and spatial scale under regular square-grid local analysis in the Murray–Darling Basin (MDB). The left panel shows the square-grid partition of the study area, and the right panels show model-specific GSE responses to increasing grid size together with the corresponding marginal-rate curves.

spatial aggregation units (Openshaw 1984; Wong 2004). By contrast, k-NN neighbourhoods were better aligned with the data structure and provided a more stable basis for summarising local error patterns (Anselin 1995). The differences in selected K among models may further indicate that different modelling approaches attenuate or retain residual spatial structure at different neighbourhood scales, consistent with the scale-dependent nature of spatial analysis and marginal-change-based curve interpretation (Dungan et al. 2002; Song et al. 2021). These results support the use of adaptive k-NN as a practical method for local spatial analysis in the present study, although this does not imply that it is universally superior to all alternative neighbourhood definitions.

5.2 Advantages, limitations, and future prospects of GSE

The main contribution of GSE is that it extends model validation beyond global average accuracy to the spatial reliability of predictions. Conventional global metrics such as RMSE provide a single summary of model error, but they do not explicitly represent how that error is distributed across space and may therefore understate

Table 6. Comparison of global, local, and spatially informed summary metric paradigms.

Feature	Global metric	Local metric	Spatially informed summary metric
Concept	A single value summarising a phenomenon across the entire study area, assuming spatial stationarity.	Decomposes a global pattern into location-specific statistics, highlighting spatial heterogeneity.	Re-aggregates local statistics into a single, spatially-informed summary value, preserving heterogeneity through weighting.
Examples	RMSE, R^2, Global Moran's I	LISA, local Moran's I, local error variance	GSE
Output format	A single number	A map of statistics	A single, weighted number
Key advantages	Simplicity and direct comparability.	Reveals local "hotspots," clusters, and spatial patterns.	Combines the simplicity of a global metric with the spatial awareness of a local one.
Key limitations	Spatially blind; masks critical local variations.	Complex output; lacks a single value for direct model-to-model comparison.	The weighting scheme can add complexity; may be less intuitive than a simple average.

instability in localised regions. In this respect, GSE complements rather than replaces existing validation tools by re-aggregating local error information into a single, spatially informed metric. This idea builds on the broader shift from purely global to local spatial analysis represented by approaches such as Local Indicators of Spatial Association (LISA), which make it possible to identify localised spatial structure, but typically yield maps of local statistics rather than a single comparable value. By weighting local error magnitude with local error variance, GSE retains information on spatial heterogeneity while preserving the practical advantage of direct model-to-model comparison. Table 6 shows the characteristics of commonly used existing indicators.

Despite this advantage, GSE also has several limitations. First, it operates on model residuals but does not diagnose the underlying causes of spatially structured error, which may arise from underfitting, omitted variable bias, nonlinear predictor-response relationships, non-stationarity, or heteroskedasticity. In this sense, GSE is a validation and diagnostic-support tool for identifying spatially concentrated predictive instability, rather than evidence of intrinsic model unreliability or a substitute for model specification tests. Second, the current implementation relies on k-NN neighbourhoods and a heuristic rule for selecting the practical summary scale, both of which may be refined in future work. Third, although GSE involves additional k-NN searches and local-statistic calculations compared with RMSE, its computational cost was manageable in the present case study. Using the random forest residuals from the independent test set with 8,036 samples, GSE at a fixed K required 0.081 s, and the full GSEopt procedure required 3.25 s. For larger datasets, more efficient nearest-neighbour search strategies, such as precomputed neighbourhood structures or approximate k-NN search, could further improve scalability (Andoni and Indyk 2008). Future extensions could also expand the current framework beyond local error magnitude and variance, for example by incorporating local bias to distinguish systematic over- and under-prediction patterns (Getis and Ord 1992), or by exploring more data-driven approaches to adaptive scale selection (Loader 2006).

6 Conclusions

This study proposed the Generalised Spatial Error (GSE) as a spatially explicit validation metric for spatial prediction models. By combining local error magnitude and local error variance, GSE extends model evaluation beyond global average accuracy and captures spatially uneven predictive instability within a single comparable value. In the Murray–Darling Basin case study, GSE_{opt} was consistently higher than RMSE across all nine models, and local error variance maps further identified concentrated zones of unstable predictive performance.

These results indicate that widely used global metrics such as RMSE can underrepresent spatially structured error. GSE therefore serves as a useful complement to conventional validation metrics by incorporating spatial heterogeneity into model comparison. Although the current framework still depends on neighbourhood definition and heuristic scale selection, it provides a practical basis for more spatially explicit model validation in geospatial applications.

Author contributions

CRediT: **Haiyang Liu:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing; **Yongze Song:** Conceptualization, Methodology, Supervision, Writing – review & editing; **Pengcheng Zhang:** Data curation, Validation, Writing – review & editing; **Kai Ren:** Data curation, Supervision, Writing – review & editing; **Yilong Wu:** Data curation, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

No funding was received for conducting this study.

ORCID

Haiyang Liu 0009-0009-8701-0538
Yongze Song 0000-0003-3420-9622
Pengcheng Zhang 0009-0002-6198-7999
Kai Ren 0009-0007-1079-902X
Yilong Wu 0009-0008-0497-6904

Data availability statement

The processed data generated in this study, including vascular plant species richness, explanatory variables, model residuals, and GSE-related outputs, are publicly available on Zenodo at https://doi.org/10.5281/zenodo.20472428 (Liu et al. 2026).

References

Andoni, A., and P. Indyk. 2008. “Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions.” *Communications of the ACM* 51 (1): 117–122. https://doi.org/10.1145/1327452.1327494.

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.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. “Geographically weighted regression: a method for exploring spatial nonstationarity.” *Geographical Analysis* 28 (4): 281–298. https://doi.org/10.1111/j.1538-4632.1996.tb00936.x.

Chai, T., and R. R. Draxler. 2014. “Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature.” *Geoscientific Model Development* 7 (3): 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Comber, A. J., P. Harris, Y. Lü, L. Wu, and P. M. Atkinson. 2021. “The Forgotten Semantics of Regression Modeling in Geography.” *Geographical Analysis* 53 (1): 113–134. https://doi.org/10.1111/gean.12199.

Cressie, N. 2015. Statistics for spatial data. John Wiley & Sons.

Crosbie, R., B. Wang, S. Kim, C. Mateo, and J. Vaze. 2023. “Changes in the surface water–groundwater interactions of the Murray-Darling Basin (Australia) over the past half a century.” *Journal of Hydrology* 622: 129683. https://doi.org/10.1016/j.jhydrol.2023.129683.

CSIRO. 2012. Australian National Outlook 2012. Canberra: Commonwealth of Scientific and Industrial Research Organisation.

Daoud, J. I. 2017. “Multicollinearity and regression analysis.” *Journal of Physics: Conference Series* 949(1): 012009. IOP Publishing. https://doi.org/10.1088/1742-6596/949/1/012009.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, and T. Münkemüller. 2013. “Collinearity: a review of methods to deal with it and a simulation study evaluating their performance.” *Ecography* 36 (1): 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Dungan, J. L., J. N. Perry, M. R. Dale, P. Legendre, S. Citron-Pousty, M. J. Fortin, and M. S. Rosenberg. 2002. “A balanced view of scale in spatial statistical analysis.” *Ecography* 25 (5): 626–640. https://doi.org/10.1034/j.1600-0587.2002.250510.x.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. 2009. “Geographically weighted regression.” In Vol. 1 of The Sage handbook of spatial analysis. 243–254. https://doi.org/10.4135/9780857020130.n13.

Getis, A., and J. K. Ord. 1992. “The analysis of spatial association by use of distance statistics.” *Geographical Analysis* 24 (3): 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Goodchild, M. F. 2004. “The validity and usefulness of laws in geographic information science and geography.” *Annals of the Association of American Geographers* 94 (2): 300–303. https://doi.org/10.1111/j.1467-8306.2004.09402008.x.

Goovaerts, P. 1997. Geostatistics for natural resources evaluation. Oxford university press.

Guisan, A., and N. E. Zimmermann. 2000. “Predictive habitat distribution models in ecology.” *Ecological modelling* 135 (2-3): 147–186. https://doi.org/10.1016/S0304-3800(00)00354-9.

Hastie, T. 2009. The elements of statistical learning: data mining, inference, and prediction.

Krause, P., D. P. Boyle, and F. Bäse. 2005. “Comparison of different efficiency criteria for hydrological model assessment.” *Advances in Geosciences* 5: 89–97. https://doi.org/10.5194/adgeo-5-89-2005.

Kuhn, M. 2008. "Building predictive models in R using the caret package." *Journal of Statistical Software* 28: 1–26. https://doi.org/10.18637/jss.v028.i05.

Kutner, M. H. 2005. Applied linear statistical models.

Legendre, P., and L. Legendre. 2012. Numerical ecology. Vol 24. Elsevier.

Lilburne, L., and S. Tarantola. 2009. "Sensitivity analysis of spatial models." *International Journal of Geographical Information Science* 23 (2): 151–168. https://doi.org/10.1080/13658810802094995.

Liu, H., Song, Y., Zhang, P., Ren, K. R., and Wu, Y. 2026. Generalized spatial error for validation. https://doi.org/10.5281/zenodo.20472428

Loader, C. 2006. Local regression and likelihood. Springer Science & Business Media.

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.

Lymburner, L., P. Tan, A. McIntyre, M. Thankappan, and J. Sixsmith. 2015. Dynamic land cover dataset version 2.1. Canberra: Geoscience Australia.

MacEachren, A. M., and M. J. Kraak. 2001. "Research challenges in geovisualization." *Cartography and Geographic Information Science* 28 (1): 3–12. https://doi.org/10.1559/152304001782173970.

Mokany, K., J. K. McCarthy, D. S. Falster, R. V. Gallagher, T. D. Harwood, R. Kooyman, and M. Westoby. 2022. "Patterns and drivers of plant diversity across Australia." *Ecography* 2022 (11): e06426. https://doi.org/10.1111/ecog.06426.

Openshaw, S. 1984. The modifiable areal unit problem. Concepts and techniques in modern geography.

Pittock, J., and C. M. Finlayson. 2011. "Australia's Murray–Darling Basin: freshwater ecosystem conservation options in an era of climate change." *Marine and Freshwater Research* 62 (3): 232–243. https://doi.org/10.1071/MF09319.

Rosenzweig, M. L. 1995. Species diversity in space and time. (No Title).

Song, Y., P. Wu, K. Hampson, and C. Anumba. 2021. "Assessing block-level sustainable transport infrastructure development using a spatial trade-off relation model." *International Journal of Applied Earth Observation and Geoinformation* 105: 102585. https://doi.org/10.1016/j.jag.2021.102585.

Tobler, W. R. 1970. "A computer movie simulating urban growth in the Detroit region." *Economic Geography* 46 (sup1): 234–240. https://doi.org/10.2307/143141.

Wong, D. W. 2004. "The modifiable areal unit problem (MAUP), WorldMinds: geographical perspectives on 100 problems: commemorating the 100th anniversary of the association of American geographers 1904–2004. 571–575. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-1-4020-2352-1_93.

Zhang, Z., Y. Song, P. Luo, and P. Wu. 2023. "Geocomplexity explains spatial errors." *International Journal of Geographical Information Science* 37 (7): 1449–1469. https://doi.org/10.1080/13658816.2023.2203212.