NAVIGATING CHALLENGES IN SPATIAL MACHINE LEARNING: VALIDATION, UNCERTAINTY, ALGORITHMS, AND REPRODUCIBILITY

Jakub Nowosad, Carmelo Bonannella, Darius Görgen, Marta Jemeljanova, Teja Kattenborn, Jan Linnenbrink, Hanna Meyer, Madlene Nussbaum, Luca Patelli, Rolf Simoes, Evelyn Uuemaa

With 3 figures and 2 tables
Received 31 October 2025· Accepted 16 May 2026

Summary: Spatial machine learning has matured considerably in recent years, providing powerful approaches for mapping and prediction. However, many challenges remain in achieving methodological robustness, transparency, and comparability. This paper examines six key methodological themes – validation, assessment, uncertainty quantification, algorithm design, software implementation, and reporting – and highlights the need for greater rigor to ensure reliable spatial predictions. We emphasize the importance of domain knowledge, appropriate sampling design, and prediction-domain adaptive cross-validation for realistic model evaluation. When training data fail to represent the prediction domain, areas beyond the model's area of applicability should be masked, and performance should be reported only within unmasked regions. Uncertainty remains a prominent challenge, requiring methodological innovation to account for spatial dependencies and improved strategies for quantifying and communicating it to diverse audiences. Furthermore, while numerous methods now incorporate spatial information into machine learning, the lack of theoretical grounding and systematic benchmarking continues to constrain comparability and progress. Finally, we highlight the critical role of open, reproducible software tools and call for standardized reporting protocols tailored to spatial prediction workflows. Addressing these methodological and practical gaps will enhance reliability, transparency, and reproducibility in spatial machine learning applications.

Keywords: Model validation, model performance metrics, uncertainty, machine learning algorithms, geospatial software, standardized protocols

1 Introduction

Spatial mapping plays a central role in geography and geosciences, where understanding the distribution of environmental variables is essential for applications ranging from climate modeling and soil mapping to biodiversity assessments. Traditionally, such maps have been generated from local field observations or measurements using geostatistical interpolation techniques that rely on spatial autocorrelation (nearby locations exhibiting similar values) and additional covariates, incorporating the associated uncertainties into the estimation process (Webster & Oliver 2007). In recent years, however, the increasing availability of high-resolution environmental data and computational power has led to a shift toward machine learning approaches. These methods are capable of modeling complex, non-linear relationships between a large number of predictor variables and the target variable (Gareth et al. 2017). In many cases, spatial autocorrelation, a cornerstone of geostatistical methods (Legendre 1993), is no longer explicitly considered, as machine learning models often capture spatial patterns implicitly through the relationships between response and predictor.

A typical geospatial machine learning workflow begins with observable reference data – such as measurements from climate stations, vegetation surveys, or soil samples – which are combined with predictor variables available for the entire area of interest. These predictors may include remote sensing products, climate data, terrain attributes, or other environmental layers. While deep learning algorithms are increasingly being applied, most current spatial mapping studies, like those discussed here, rely on traditional machine learning algorithms, which link predictors and responses at the pixel level. The chosen machine learning algorithm is often a random forest or another tree-based method, selected for its strong performance and ease of use. It is then trained to capture the relationship between the predictors and the target variable. Model performance is typically assessed using test data that have been withheld from model training or cross-validation (Piikki et al. 2021). In cross-validation, the data is divided into *n* folds, then trained on *n-1* blocks and validated on the held-out block. This is repeated until all *n* blocks have been held out once. The final trained model is then applied to produce spatial predictions across the region of interest. In

https://doi.org/10.3112/erdkunde.2026.04.01 ISSN 2702-5985 (Online)

some cases, studies also map uncertainties (Malone et al. 2016, Bastin et al. 2019) or apply methods from explainable machine learning (Molnar 2021, Ryo et al. 2021) to interpret the learned relationships and gain a better understanding of the drivers behind spatial predictions.

Despite considerable progress in recent years, many modeling strategies used in geosciences still closely resemble those from other domains, often without specific adaptations to the spatial characteristics of the data. This results in key challenges across all stages of the workflow. Challenges start with the reference data: samples are often clustered in space, collected for specific purposes, or absent over extensive regions. These patterns introduce sampling bias and spatial dependence that complicate both model development and realistic evaluation (Bruin et al. 2022). Moreover, this spatial structure can lead to issues such as difficulties in reliably measuring model performance, and limited transferability across different spatial or temporal domains. Uncertainty quantification remains underdeveloped, and standardized approaches for estimating and communicating uncertainties have yet to emerge. In addition, reproducibility is often hindered by inconsistent software implementations, the lack of structured workflows that support transparency, and insufficient documentation of the created models.

This paper outlines and critically discusses the challenges of applying machine learning for spatial mapping. We structure our discussion around six interconnected themes that span the modeling workflow: (1) validation strategies for models and spatial patterns, (2) model performance metrics, (3) uncertainty quantification and communication, (4) algorithm diversity and comparability, (5) software implementations in R and Python, and (6) standardization of model reporting (Fig. 1). These themes represent critical decision points where current practices often fall short of the methodological rigor needed for reliable spatial predictions. While reviewing recent progress, we explicitly highlight unresolved issues to encourage future research to address these pressing challenges.

2 Strategies for validation of models and spatial patterns

Unlike traditional datasets, spatial data often come with unique characteristics such as spatially uneven distribution of the sampling data (Bouasria et al. 2023), spatial autocorrelation (Legendre 1993, Koldasbayeva et al. 2024), spatial heterogeneity, i.e., variation of a phenomenon across space (Nikparvar & Thill 2021), and skewed distribution of the variables (Naveau & Allard 2005). In an ideal scenario, data for modeling would be collected through a sampling design specifically tailored to the task, with separate designs for training and validation. Under such conditions, many of the issues related to spatial dependence and heterogeneity would be minimized. However, in practice, ideal sampling designs are rarely feasible – either due to logistical constraints or because available data are aggregated from various sources, inheriting their biases and limitations. Consequently, specialized validation and assessment techniques are often necessary to evaluate that models can be reliably applied to the target spatial domain.

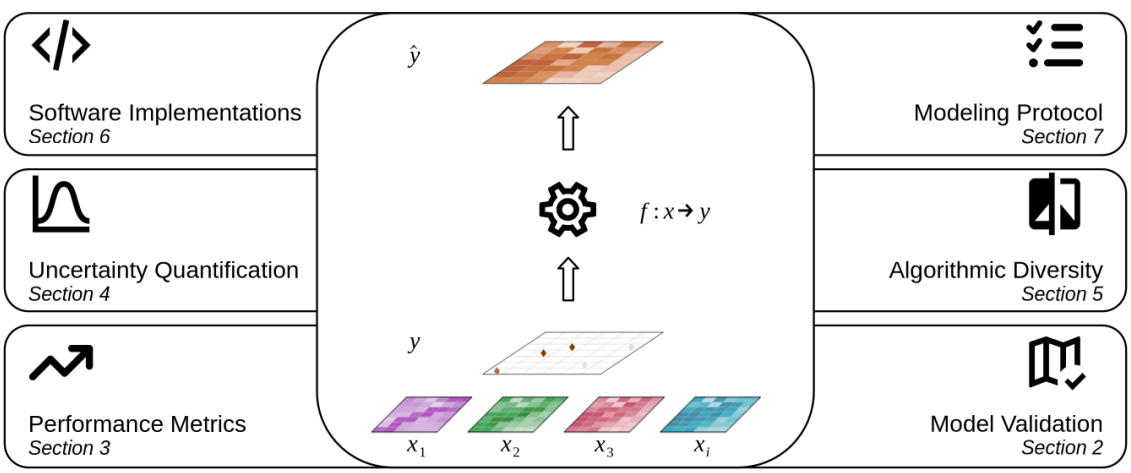

Fig. 1: A graphical overview of the paper's main themes, organized hierarchically from bottom to top to illustrate their relative order and relationships

Model evaluation in spatial machine learning should fundamentally be shaped by the representativeness and spatio-temporal structure of the training data relative to the intended prediction scenario. While prediction accuracy is generally straightforward to assess when reference data are evenly distributed and representative of the target domain (Wadoux et al. 2021, Bruin et al. 2022), many spatial modeling tasks rely on clustered or biased reference samples. In ecological and environmental modeling, spatial dependence between training and validation data is especially problematic, where reference data from multiple sources, collected using case-wise sampling designs, are combined and used. Such data – often concentrated in specific regions or environments – may not represent the spatial distances between prediction points and reference samples (Fig. 2) nor capture the full variability of the land surface or the predictor space (Tobler 1970, Meyer & Pebesma 2021, Ploton et al. 2020, Kattenborn et al. 2022). This lack of representativeness does not necessarily bias the fitted model coefficients (Pabon-Moreno et al. 2022), but due to spatial autocorrelation, test samples located near training points are often highly similar, leading to inflated performance metrics (Roberts et al. 2017, Dormann et al. 2007, Ludwig et al. 2023) if not considered in the validation strategy. Consequently, model evaluations derived from such clustered data may not generalize beyond the reference spatial domain, complicating model transferability and the interpretation of derived predictions (Yates et al. 2018, Mila et al. 2022, Sweet et al. 2023).

Ideally, we would be able to designate the locations and a timepoint of our reference samples, optimal for the given modeling purpose. Especially for mapping larger domains (national, continental, or global), however, reference data is assembled from multiple sources initially collected for other purposes exhibiting spatial clustering and sub-optimal

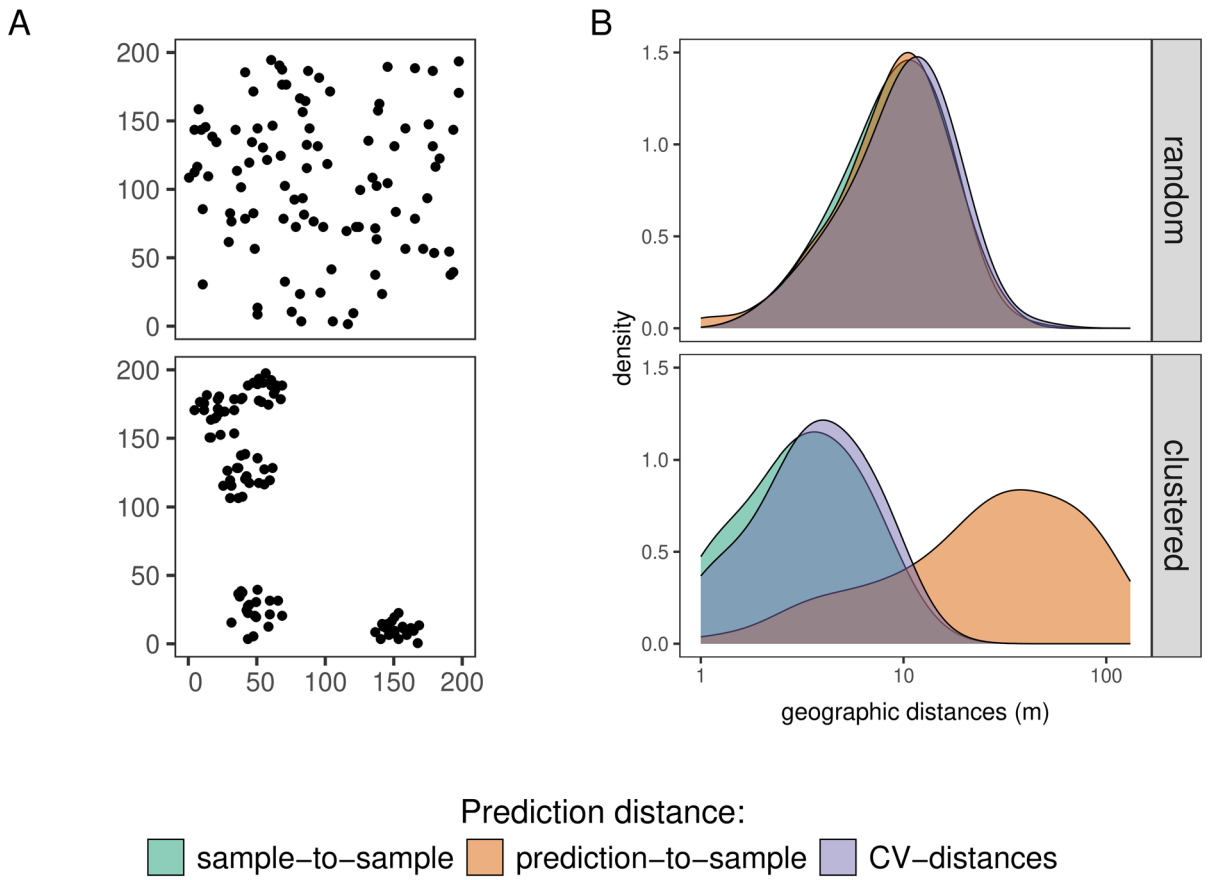

Fig. 2: (A) Sampling locations for random and clustered designs. (B) Density of nearest-neighbor distances: between samples (green), between prediction points and samples (orange), and between random cross-validation (CV) folds (blue). Prediction distances represent the typical spatial separation at which the model is evaluated. For random sampling, CV distances closely match prediction distances, providing reliable accuracy estimates. For clustered sampling, CV distances are much shorter, leading to overly optimistic and unreliable validation. Source: LOEK-RS/rs-figures, licensed under CC BY 4.0.

coverage of the characteristics of the target area. To mitigate those practical constraints, spatial cross-validation strategies can support unbiased model training and realistic model assessment for prediction scenarios of extrapolations (Roberts et al. 2017, Meyer et al. 2018, Ploton et al. 2020, Kattenborn et al. 2022). Common approaches include spatial blocking, where observations are grouped by distance-based clusters (see an example in Fig. 3), and leave-one-location-out schemes, where entire spatial units are withheld, ensuring that training and test sets are separated by a minimum autocorrelation-informed distance (Ploton et al. 2020, Kattenborn et al. 2022, Ludwig et al. 2023). Conversely, the conventional random cross-validation, where cross-validation folds are divided randomly, may be more suitable in interpolation cases when samples are randomly located in the prediction area (Wadoux et al. 2021). In practice, however, the prediction problems live in a full spectrum from random samples of the prediction area to clustered ones used for extrapolation. Thus, to tackle this issue, the recent work of Mila et al. (2022), Linnenbrink et al. (2024), and Wang et al. (2025) propose the use of prediction-domain adaptive cross-validation methods, such as dissimilarity-adaptive CV and kNN distance matching CV. In the latter technique, a set of possible CV splits from random cross-validation to spatial cross-validation is derived, and the optimal cross-validation split is then the one that best aligns with the given prediction situation. Yet, it is still unclear whether the cross-validation folds should be designed in geographical or in the feature space, and under which conditions one might be favorable over the other.

Even with careful design, validation strategies may fail to provide reliable estimates of predictive performance when the available data does not adequately represent the actual prediction scenario – for instance, when prediction locations lie far outside the spatial range of the training data. In such cases, regions not represented in the training set should be considered for masking, as the quality of the predictions are unknown in these areas due to the absence of either training and test data Validation metrics should then be reported only for the unmasked regions. Meyer & Pebesma (2021) proposed the area of applicability as a method to define the subspace of the predictor domain supported by the training data, i.e., the locations where predictions are reliable given the training data. While this approach can become computationally intensive as the study area and training dataset grow, it has recently been applied successfully in global-scale studies (Lusk et al. 2026, Ludwig et al. 2023, Maxwell et al. 2024). A less computationally demanding alternative is the thresholded anomaly score calculated using Isolation Forest (Liu et al. 2008), as demonstrated by Tian et al. (2025). Overall, explicitly recognizing and communicating the limits of model applicability is essential for ensuring that spatial predictions are trustworthy.

3 Model performance metrics

While prediction-domain adaptive cross-validation addresses how to properly structure model evaluation to minimize spatial bias, the choice of

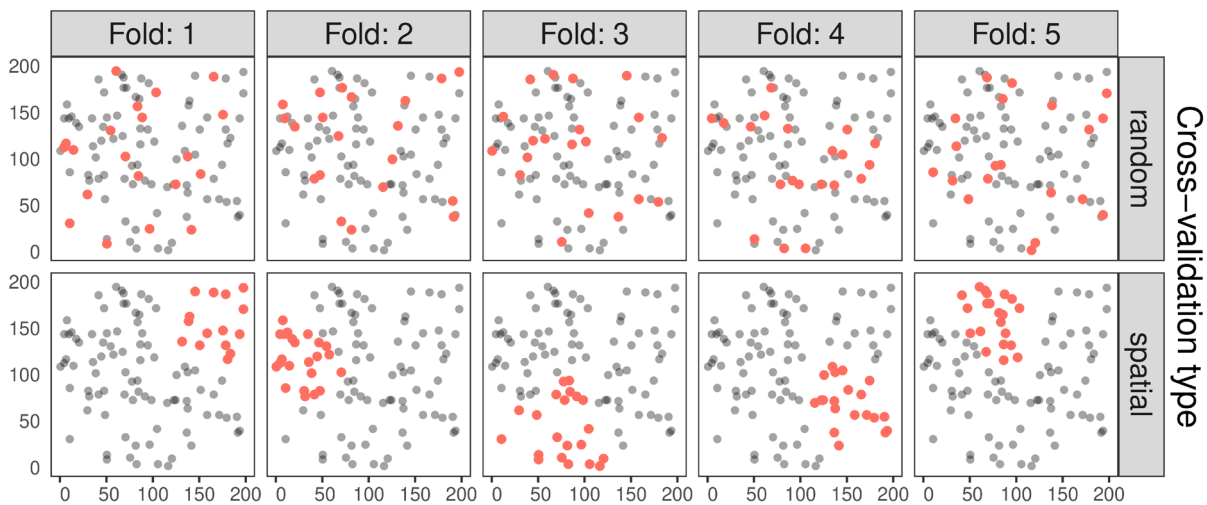

Fig. 3: Comparison between non-spatial (random) and spatial cross-validation (spatial clustering) methods. Source: LOEK-RS/rs-figures, licensed under CC BY 4.0.

assessment metrics is equally important. Even with well-designed validation schemes, relying solely on a single accuracy measure may still lead to misleading conclusions – especially when spatial dependencies and heterogeneity are ignored.

Many studies applying machine learning to spatial data report traditional accuracy metrics such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) for regression tasks and confusion matrix–based measures, Log-loss, or the Brier score for classification tasks (Zhao et al. 2024, Wei et al. 2019, Jacob & Masilamani 2025, Chrysanthopoulos & Kallioras 2025). These metrics are valuable for evaluating a model's mathematical performance, but they have notable limitations (Liemohn et al. 2021, Goodling et al. 2024). They can quantify how well a model fits the data or be used for validation by comparing the model's predictions to unseen samples. However, they are doing so without necessarily indicating whether the model prediction is meaningful or robust in real-world applications. A model might achieve a high R^2, suggesting a strong fit, but this can often be a result of overfitting, where the model captures noise rather than underlying patterns or laws of nature. Secondly, there is a need for understanding the properties and suitability of different metrics. Model evaluation metrics differ in their characteristics, which affects how comparable they are (e.g., MAE or RMSE are expressed in the units of the data, whereas R^2 is not). For classification problems, probabilistic measures (e.g., Log-loss or Brier score) should be used in cases where the model output is probabilistic, instead of, for example, confusion-matrix-based ones (e.g., True Positive Rate or F_1 score). Moreover, these metrics are the results of aggregation and therefore lack the ability to account for the spatial aspect unless they are calculated for different regions – they treat all errors the same, ignoring geographical differences. This oversight prevents them from spotting areas where errors occur consistently, which can hide local issues in model performance. For example, a model might perform well in one geographic region but struggle in other regions due to different underlying processes (Wang et al. 2024). Two models can also have similar prediction performance but yield different spatial patterns (Sterlacchini et al. 2011). Relying only on these global metrics can result in an incomplete understanding of how well the model works. Thus, the choice of metric should be informed by the study's aims (Goetz et al. 2015) and end-user requirements, and reporting several complementary metrics is recommended to fully characterize model performance.

In evaluating the performance of spatial machine learning models, it is also essential to complement traditional accuracy metrics with spatially-aware ones to ensure both a robust fit to the data and meaningful assessment across different spatial contexts. This improves the understanding of whether metrics that are near their optimal values – whether high or low – truly indicate model quality or simply overlook underlying spatial patterns by not having access to proper validation data. While some promising approaches have emerged to address spatial considerations in model evaluation, significant gaps remain in this area. Recent proposals include spatially-aware metrics such as Boundary Uncertainty Concentration in medical science, which attempts to measure segmentation uncertainty along boundaries where errors typically concentrate (Zeevi et al. 2025). Similarly, the SPACE (Spatially-Aware Calibration Error) metric represents an initial effort to quantify local alignment between predictions and actual patterns, though its ability to fully address spatial autocorrelation and regional biases in model interpretability remains to be thoroughly validated (Zeevi et al. 2025). Despite these early developments, the field still lacks comprehensive frameworks for systematically evaluating spatial performance variations, suggesting that substantial research attention is still needed to develop robust spatially-aware evaluation methodologies.

A fundamental step in diagnosing spatial performance should be the mapping of predictions and residuals (the differences between observation and model prediction) across the entire study area. This enables to visualize model generalization capabilities beyond training points, revealing spatial clusters of high residuals which indicate systematic errors in specific regions. Residual trends (e.g., directional bias) often signal omitted variables, such as uncaptured environmental drivers (Verma 2025, Gaspard et al. 2019). Spatial autocorrelation in residuals, often measured by Moran's I, may suggest missing spatially-structured predictors, though its properties and limitations in spatial machine learning are still being explored (Nowosad & Meyer 2025). Visual inspection of the map can also reveal boundary artifacts. Predictive mapping across the full spatial domain not only serves to validate the model's generalization capabilities but can also uncover practical limitations when applied at large scales, such as excessive computational demands or incomplete spatial coverage of predictor variables.

4 Quantifying and communicating uncertainty

Reliable spatiotemporal predictions hinge not only on point estimates but also on a rigorous characterization of uncertainty. In spatial machine learning, uncertainty arises from multiple sources like data errors, model structure, and inherent process variability and propagates through every stage of prediction and analysis. Below, we outline key considerations and methodological advances for quantifying, propagating, and communicating uncertainty in spatiotemporal contexts.

Uncertainty in spatiotemporal machine learning can be broadly categorized into three types based on its origin. First, data uncertainty such as measurement errors in response variables (e.g., soil samples), noise in predictors (e.g., remote-sensing products), and support mismatches (extent of sample vs. pixel size). Second, model uncertainty originates from the variability due to algorithmic choice (e.g., random forest vs. support vector machines) and model's structure (e.g., different model setups or neural network architectures). Third, process uncertainty sources from unmodeled spatial processes, temporal dynamics or boundary-condition variability (e.g., episodic land-use change, extreme weather events not captured by training data). Recognizing these sources is the first step towards rigorous uncertainty assessment (Heuvelink et al. 1989), it remains, however, often difficult to accurately quantify without vague assumptions. Carefully accounting for input data uncertainty can significantly yield relevant differences, as shown by Takoutsing et al. (2022) in the context of soil mapping. Besides assessing the average map performance (section 3), it is also important to assess whether the model's predictive uncertainty is well-calibrated. It can be done by comparing the stated prediction interval (e.g., 90%) to how often the true values fall within that interval in withheld data (prediction interval coverage probability) or by recently proposed methods that account for the former's one-sided bias (Schmidinger & Heuvelink 2023).

Quantifying uncertainties is straightforward using classical geostatistical methods, such as external drift kriging. This approach constructs the prediction surface from the spatial dependence characteristics of the observed variable – such as temperature – while incorporating an external predictor variable (the drift), such as elevation (Webster & Oliver 2007). Machine learning methods are usually not equipped with method-inherent uncertainty quantification. Their focus instead lies on improving the predictive performance without explicitly estimating parameters of the spatial autocorrelation. Thus, most machine learning methods, are inherently non-spatial and spatial autocorrelations in the residuals are, for the most part, ignored. There are exceptions to this. For example, Gaussian process models use the similarity of values at nearby locations to make predictions and provide an estimate of uncertainty (Rasmussen & Williams 2008) while quantile regression forests – an extension of random forests – return the full conditional distribution of the predictions (Meinshausen 2006), but they either run into scalability issues with high-dimensional input spaces or fail to return trustworthy prediction intervals in extrapolation situations. One possible remedy, however, which also faces scalability issues with an increasing number of samples, is to conduct computationally demanding bootstrap (Davison & Hinkley 1997), a method that estimates uncertainty by resampling the data to generate a distribution of a statistic. In the spatial context, bootstrap approaches to quantify map uncertainties further run into assumption issues (e.g., independence of reference data points, Gaussian error distribution) due to the spatial dependence of the data. To account for uncertainties in extrapolation situations, one approach is to calculate dissimilarity between the predictors observed during training and those encountered for the prediction domain (Meyer & Pebesma 2021), or to describe how many training data points similar to the prediction point have been used to train the model (Schumacher et al. 2025). As a consequence, predictions may be limited to geographic areas which are comparable to the training data and where sufficient data exists and uncertainty estimates are reliable (Meyer & Pebesma 2021).

Machine learning methods often yield univariate uncertainty estimates – such as the standard deviation from quantile regression forests for a single variable at a specific pixel. Error propagation into an aggregated end-user map (e.g. a soil quality index formed from multiple basic soil property maps) requires information on joint uncertainty distribution. The uncertainty from one base map is usually correlated to other base maps which should be accounted for to estimate the uncertainty of the final aggregated map product. Correlation across space, time and across different mapped outcomes are currently largely ignored for space-time mapping. In addition, no *off-the-shelf* solution is available, similar to block kriging (predictions over aggregated units with associated uncertainty) from the geostatistical framework (Heuvelink & Webster 2022). Methods

need to be developed that explicitly extract or model the dependencies in the data and then allow to correctly aggregate uncertainties over predefined spatial units (e.g., for a parcel or province). In this context, the initial approximation approach of Wadoux & Heuvelink (2023) represents a useful step and calls for more comprehensive investigations.

Lastly, once accurately quantified, spatial uncertainties need to be efficiently communicated to end-users (Milne et al. 2015) so that uncertain model outcomes translate into safe decision making. Long established for weather forecasts, graphical representations (e.g., dashboards) or interactive visualizations (e.g., web mapping applications) for uncertain spatial or space-time datasets are often not easily accessible to end-users. Moreover, it remains unclear how stakeholders best cognitively process reported uncertainties. This domain is currently not fully explored and clear recommendations depending on decision type (e.g., exceedance of critical threshold, local information-content-metrics to indicate further sampling) are missing. A recent exception is Courteille (2025), who investigated uncertain soil quality map representations for land use planning. By systematically identifying uncertainty sources, modeling spatio–temporal correlation, rigorously propagating errors through change-of-support, validating interval coverage (how frequently the confidence or prediction interval contains the true value), and crafting clear visualizations, spatial machine-learning practitioners can deliver predictions that are both accurate and trustworthy, crucial for applications ranging from soil mapping to climate-impact assessment.

5 Algorithm diversity and comparability

The diversity of standard machine learning algorithms provides a broad range of modeling possibilities. Furthermore, spatial machine learning has already developed numerous methods to account for spatial dependence among study sites (see Jemeljanova et al. 2024, Patelli et al. 2024, and references therein), increasing the available modeling alternatives for spatial application. The complexity of these strategies varies from techniques that are easy and straightforward, such as the inclusion of X and Y coordinates as additional predictors (so-called spatial proxies) (Brodie et al. 2020), to complex spatial adjustments that are rooted in advances in statistics and data science (e.g., Deep Kriging (Chen et al. 2022) or Generalised Least Square based Random Forest (RF-GLS) (Saha et al. 2023)). Motivations behind using these spatial adjustments include incorporating spatial autocorrelation information explicitly (e.g., by adding spatial proxies) and avoiding or mitigating bias in model estimation and evaluation arising from the presence of spatial autocorrelation (e.g., by adopting prediction domain-adaptive cross-validation). Beyond these approaches, domain-informed machine learning, where foundational domain knowledge is explicitly incorporated into models to guide and constrain them (Minasny et al. 2024), represents a promising research direction. As spatial machine learning methods become more varied and widely used, the necessity of systematic comparisons to evaluate them comprehensively becomes increasingly important.

In other fields that employ machine learning modeling (e.g., the medical field), various machine learning algorithms are typically compared based on their test predictive accuracy, with the highest-scoring algorithm chosen for the final prediction. This approach has also been adopted in the context of spatial mapping with machine learning (An et al. 2023, Kim et al. 2023, Schratz et al. 2019). However, comparative studies, such as Fernández-Delgado et al. (2014), suggest that different machine learning algorithms often result in minimal performance differences. It is also important to keep in mind that better accuracy metrics do not necessarily indicate that a model is better, as the model may be overfitted (see section 2 on the relevance of the validation strategy), or other relevant aspects (especially locally) may be hidden by the generalized metrics (Meyer & Pebesma 2022). Besides that models are sometimes compared using suboptimal assessment strategies (see also Stock et al. 2023), the validation is often relying on a single accuracy metric, which may be inappropriate in cases of the lack of independence of the training samples and non-linear relationship between response and predictors (such as R^2 (Jonard et al. 2022)). Due to these oversights, the evaluation may fail, resulting in important error structures (e.g., underrepresented areas, spatial clusters) and overly optimistic results (Kumar et al. 2025, see also section 3).

To provide a fair and informative comparison, not only should appropriate validation methods be used, but also the model-uncovered relationships should be evaluated against domain knowledge on the respective environmental processes (like in the study of Brugere et al. 2023) as it provides crucial information on models' robustness and applicability. It is also necessary to evaluate the final prediction map, as it may contain artifacts (Ahn et al.

2020, Beguin et al. 2017), not cover the actually observed value range (Baltensweiler et al. 2021) or contain incorrect patterns that may pass unobserved by standard accuracy metrics. Lastly, ease of model setup, restrictions on input data introduced by model assumptions, the time required for model training and prediction are important factors as well (Kmoch et al. 2025, Sekulić et al. 2020, Stojanova et al. 2013), as improvement over other methods may be so minimal that it does not justify the additional effort needed for method implementation. This factor is particularly relevant when researchers have limited processing power or are considering the overall simplicity and resource efficiency of their models.

The current body of literature on spatial adaptations for machine learning often lacks consistency and completeness in both methodology and reporting. For instance, a new adaptation may be introduced along with its prediction accuracy, yet without clearly demonstrating how it compares to baseline models without spatial adjustments. This omission limits our understanding of the actual contribution of the spatial adaptation method. At the same time, many studies are one-off cases that focus on specific target datasets and do not compare their adaptation methods to others. Most comparison studies typically focus on predicting one (Ploton et al. 2020) or two environmental phenomena (Bruin et al. 2022, Meyer et al. 2019), while research that compares multiple phenomena (Nussbaum et al. 2018, Patriche et al. 2023) or datasets is limited.

Furthermore, insufficient attention is given to spatial heterogeneity. Environmental phenomena exhibit varying strengths of spatial autocorrelation (e.g., temperature, precipitation) and distinct spatial distribution patterns (e.g., isotropy vs anisotropy) that influence the effectiveness of spatial modeling techniques and the performance of machine learning algorithms. Notably, comparisons that do not account for the objective of the application, such as interpolation or extrapolation, may result in misleading conclusions. For example, Bruin et al. 2022 and Mila et al. 2024 highlight how some algorithms perform well for interpolation but poorly when adopted for extrapolation.

Another challenge is the lack of well-developed theoretical foundations. Many current approaches to incorporating spatial information into machine learning are developed in an ad hoc manner, presenting new methods with limited theoretical justification, and do not yet match the formal rigor of geostatistical frameworks such as the Best Linear Unbiased Predictor (BLUP) (Cressie 1993). Comprehensive mathematical analyses and systematic simulation experiments remain relatively uncommon. Addressing these gaps could strengthen the methodological basis of spatial machine learning and support its broader and more reliable application. An example of such rigorous development can be found in Meinshausen's introduction of the quantile regression forest, which was accompanied by a formal theoretical framework (Meinshausen 2006).

The importance of evaluating different aspects of the spatial adaptation methods highlights a significant research gap: the need for a taxonomy that categorizes these methods, including factors like computational time and complexity. There is a limited number of papers that thoroughly present the available methods (Jemeļjanova et al. 2024, Patelli et al. 2024), and valuable discussions comparing these methods are often dispersed across various studies. Conducting systematic and comprehensive reviews is essential, as it helps avoid common pitfalls, particularly for newcomers in the field.

To improve the robustness and reproducibility of spatial machine learning research, the following actions are proposed. First, development of benchmark datasets with different spatial properties (summarized in Tab. 1). Such datasets should be designed to represent a wide spectrum of spatial characteristics, including a range of spatial phenomena, scale, autocorrelation patterns, and the interpolation–extrapolation context. They should also account for different data types, sampling schemes, and data quality. In addition, benchmark datasets should be openly accessible, well-documented, and accompanied by clear licensing to enable reproducible comparisons across studies. This would allow for a comprehensive evaluation and comparison of spatial machine learning methods considering the diverse characteristics of spatial phenomena. Next, scripts and data should be shared, especially when proposing new methods that need to be validated. Lastly, clear and comprehensive reporting is necessary to ensure that the proposed spatial machine learning models and methods are transparent and reproducible.

6 Spatial machine learning workflows in R and Python

R and Python offer a diverse set of frameworks for machine learning, each differing in philosophy, technical architecture, and spatial capabilities. For spatial machine learning practitioners, it is essential to understand how these frameworks handle data

Tab. 1: Aspects to consider in the construction of benchmark datasets for spatial machine learning research

Aspect	Characteristics
Phenomena	Different domains and spatial dependence characteristics.
Scope	Interpolation–Extrapolation continuum.
Scale	Local, regional, global.
Autocorrelation pattern	Strong, weak, isotropic, anisotropic, stationarity, non-stationarity.
Variable types	Continuous, categorical.
Data quality	Noise, missing values.
Sampling scheme	Uniform and clustered distribution. Low, medium, high sample density.
Accessibility	Add metadata; share through repository (e.g., data repository, software package).
Licensing	Open access, clear usage license for reproducibility.

structures, integrate spatial cross-validation, and support reproducible pipelines.

The *caret* package has long been a cornerstone of machine learning in R, offering a unified interface to numerous algorithms via its train() function (Kuhn 2008). Its simplicity and accessibility make it popular for quick prototyping without deep knowledge of underlying algorithms. It operates on native data frames that allow flexible preparation of predictors, however, spatial or temporal components must be manually converted to tabular form, as *caret* lacks explicit support for spatial objects. Packages such as *CAST* (Meyer et al. 2025) and *blockCV* (Valavi et al. 2019) add spatial and prediction domain-adaptive cross-validation and area of applicability tools. However, although it remains the most frequently applied framework in R, it is no longer being developed, which limits its potential for future studies. The *tidymodels* ecosystem takes a modular, multi-package approach that prioritizes transparency and reproducibility (Kuhn & Wickham 2020). Workflows are composed of integrated modules from separate packages for preprocessing and modeling, with explicit step declarations that enhance clarity. Like *caret*, it assumes tabular inputs and lacks direct support for spatial objects. Spatial functionality can be extended with packages such as *spatialsample*, which implements spatial resampling to address autocorrelation effects during evaluation, and *waywiser*, which provides area of applicability assessments (Mahoney et al. 2023, Mahoney 2023). While modularity fosters reproducibility, it also increases complexity and dependencies, and the absence of built-in memory management or raster partitioning limits large-scale spatial prediction, requiring external implementations of tiling, parallelization, and output reassembly. The last main framework in R, *mlr3*, implements an object-oriented design that allows flexible specification of tasks, algorithms, and resampling strategies (Lang et al. 2019). Like other frameworks, users must prepare and harmonize inputs, including spatial attributes and projections. However, unlike *caret* and *tidymodels*, it offers native spatial extensions such as *mlr3spatial* and *mlr3spatiotempcv* for spatial and spatiotemporal resampling (Becker & Schratz 2025, Schratz et al. 2024), and supports conversion of spatial objects into *mlr3*-compatible structures. Built-in distributed computing make it well-suited for intensive workflows involving hyperparameter tuning, repeated resampling, or large predictor stacks, though memory handling and tile-wise prediction for large rasters remain manual tasks. While the object-oriented design of *mlr3* introduces a steeper learning curve through concepts such as 'tasks', 'learners' and 'resamplings', it also provides enhanced flexibility and facilitates reproducible workflows.

Several specialized R packages complement general-purpose frameworks by addressing spatial and spatiotemporal prediction tasks. The *sits* package is designed for satellite time series classification, operating directly on spatiotemporal data cubes and supporting workflows from data access to large-scale prediction (Simoes et al. 2021). Cubes can also be exported to standard spatial formats for integration with other workflows, and built-in parallel processing allows scaling from local training data to regional predictions. However, its strict cube-based structure limits use with irregular data and excludes regression tasks. In contrast, several Random Forest–based packages focus on spatial regression and interpolation. *RandomForestsGLS* extends Random Forests with spatial modeling via dependency-adjusted split-

ting and residual kriging (Generalised Least Square based Random Forest, Saha et al. 2024), providing spatial estimation and prediction functions, though its syntax can be challenging for new users and requires careful data formatting. *spatialRF* (Benito 2021) generates spatial predictors from distance matrices, Moran's Eigenvector Maps, and PCA, with tools for feature (predictor) selection, evaluation, and tuning, but method choices and data preparation can be complex. *meteo* implements Random Forest Spatial Interpolation (RFSI) (Sekulić et al. 2020), integrating well with spatial R objects and supporting cross-validation and flexible prediction, but its functionality is largely limited to these core features. Each tool targets different spatial tasks, classification, regression, or interpolation, yet all handle spatial dependencies, with trade-offs in flexibility, preprocessing, and supported modeling types.

In the Python ecosystem, *scikit-learn* is one of the most widely used libraries for machine learning, particularly for classical algorithms such as classification, regression, and clustering, as well as for essential tasks including preprocessing, model training, validation, and evaluation (Pedregosa et al. 2011). It implements a broad range of algorithms, although its simple feed-forward neural network estimators are limited in flexibility and lack GPU acceleration. The package is primarily designed for tabular data, most commonly *NumPy* arrays or *Pandas* DataFrames. It does not natively support spatial geometries or raster data, requiring users to transform such datasets into compatible tabular formats. Similarly, spatial cross-validation strategies are not implemented directly. Workarounds include encoding coordinates as tabular predictors or using group-based folds created externally. Dedicated packages such as *spatialkfold* (Ghariani 2023), or spatial vector data libraries such as *geopandas* and *shapely*, can assist in these tasks. Complementary libraries such as *PySAL* provide additional functionality for spatial autocorrelation analyses (Rey & Anselin 2007).

Despite the maturity of Python's machine learning ecosystem, spatial machine learning tools remain less developed than in R. For instance, there is no widely used Python package that currently implements methods to estimate a model's area of applicability (Meyer & Pebesma 2022). As a result, task-appropriate methods are often not used because ready-to-use implementations are unavailable, and when they are applied, they must be implemented manually, reducing reproducibility. This highlights the importance of sharing processing workflows and developing dedicated Python libraries for spatial machine learning.

While R offers more mature spatial machine learning tools and Python provides a widely adopted general-purpose framework, practitioners in both languages face challenges in applying models to spatial domains. In particular, the frameworks and specialized packages in R and Python reflect a tension between general-purpose modeling tools and the specific demands of spatial prediction. Applying models trained on sparse samples to large spatial domains creates computational and deployment challenges that current software only partially supports. Most general-purpose frameworks lack built-in, memory-efficient tools for spatial prediction at scale. Users must manage tiling, parallelization, and aggregation themselves, bearing the full computational cost. Additionally, there is a lack of standardized mechanisms for describing spatial properties and harmonizing predictors, placing the burden of managing projections, metadata, and alignment squarely on the practitioner. This highlights the necessity of consistent reporting and structured workflows that allow for software implementations of common applications.

7 Standardized modeling protocols

Given the various challenges and pitfalls of spatial machine learning discussed earlier, it is essential to communicate key modeling decisions and data characteristics transparently. However, such crucial information is often lacking in spatial machine learning studies. Reporting this information using standardized model protocols might be an important step towards enhancing transparency and enabling reproducibility in spatial machine learning studies.

Standardized model protocols are available for machine learning models in general (e.g., Mitchell et al. 2019, Kapoor et al. 2024), and are widespread in scientific fields that underlie regulation, such as medicine (e.g., Collins et al. 2015, Luo et al. 2016, Mongan et al. 2020). There, they prove to be useful for reviewers of the model, end-users, and developers but also for teaching and learning purposes. For spatial machine learning modeling, such standardized model protocols are rare. To our knowledge, only one such protocol exists: the ODMAP protocol for describing species distribution models (Zurell et al. 2020) (based on the ODD protocol developed for individual-based models in ecology (Grimm et al. 2006)). A standardized model protocol suited for the general case of spatial machine learning models is lacking. As demonstrated by the success of protocols

like ODD and ODMAP, a protocol tailored towards spatial machine learning models could greatly benefit the spatio-temporal modeling community.

A protocol tailored towards spatial machine learning may enable documenting key information and encompass the challenges discussed earlier in this paper. To enable transparent comparisons of different spatial machine learning algorithms (see section 5), such a protocol should include information regarding the modeling type (classification, regression or segmentation), the aim of the model (inference or prediction), information on the response variable and its spatial structure, as well as information about the training data and study area (Tab. 2). Additionally, information about patterns and artifacts in the prediction map, an explicit quantification of the improvement over a (non-spatial) baseline model, as well as the computational time required to run the algorithm, could be included in the protocol to enable comparison of different spatial machine learning algorithms. Besides transparent comparisons of algorithms, a spatial modeling protocol can also highlight general issues, such as inappropriate cross-validation strategies (see section 2), inadequate selection of model performance metrics (see section 3), or the lack of a thorough uncertainty assessment (see section 4). Lastly, it can enable reproducibility by containing a specific implementation of software (see section 6). When completed during the model development phase, the protocol could allow these potential issues to be identified and addressed early, well before peer review.

Motivated by the lack of a protocol and its potential benefits, we are currently collaborating on developing a spatio-temporal modeling protocol. The implementation in the form of a Shiny Web App is publicly available, and the current status of the application can be monitored at https://github.com/LOEK-RS/STeMP. The protocol will include features such as warnings triggered when common pitfalls discussed in this paper are encountered. It will also support uploading model objects and spatial data to automate parts of the protocol. To enable continuous and transparent community engagement, we plan to manage the protocol via GitHub Issues and Pull Requests, where new features or criteria can be proposed and discussed by the spatial modeling community.

8 Core takeaways

Spatial machine learning has become a central tool for mapping and predicting environmental and geographic phenomena. Yet, as this perspective has shown, methodological maturity has not always kept pace with its growing application. Across validation strategies, performance assessment, uncertainty quantification, algorithm design, software implementation, and reporting practices, spatial structure is not yet fully integrated into many modeling workflows. As a result, models that appear accurate may nonetheless lack robustness, transferability, or transparency.

A recurring theme throughout this paper is that spatial machine learning cannot simply adopt conventions from non-spatial machine learning. Spatial dependence, heterogeneous sampling designs, and

Tab. 2: Aspects to report when proposing and comparing spatial machine learning methods

Aspect	Description
Spatial characteristics of the phenomenon	Data support (e.g., spatial extent, resolution), stationarity, anisotropy.
Aim of the application	Intended use: interpolation, extrapolation, temporal transfer, domain transfer.
Algorithm characteristics	Theoretical background of the model, structure, hyperparameters and tuning strategy.
Baseline/benchmark	Description of simpler and/or standardized methods to determine the improvement.
Validation strategy	Validation strategy design, adopted accuracy metrics.
Treatment of spatial dimension	How the spatial dimension was included/considered.
Resulting map/spatial pattern	Qualitative and quantitative analysis of the obtained maps. Evaluate the presence of smoothing, artifacts and uncertainty and compare with existing maps.
Computational framework	Runtime, memory, scalability, parallelization.
Reproducibility	Make the code (e.g., through packages) and data (considering data licenses) available.

the need for spatially meaningful uncertainty fundamentally alter how models should be evaluated and interpreted. Importantly, effective spatial machine learning requires more than technical proficiency. It demands contextual awareness of the spatial processes under study, careful consideration of sampling design, and explicit alignment between modeling choices and practical objectives. Domain knowledge is not optional, it is essential for interpreting predictions, identifying artifacts, and assessing plausibility.

The six interconnected themes presented here provide a conceptual framework for strengthening spatial machine learning practice. Rather than treating validation, uncertainty, algorithms, software, and reporting as isolated concerns, we argue they are better understood jointly. In addition, weaknesses in any one component can compromise the reliability of the entire workflow.

Machine learning holds considerable promise for advancing geographic and environmental research. Realizing this potential, however, requires a transition from performance-driven mapping toward spatially explicit, uncertainty-aware, and reproducible modeling. By embedding methodological rigor and transparency at every stage of the workflow, spatial machine learning can evolve beyond prediction tool into a robust scientific framework for understanding and managing complex spatial systems.

9 Future research directions

Meaningful advances in spatial machine learning demand more than algorithmic tweaks, with progress now depending on strengthening model evaluation standards, uncertainty frameworks, comparability infrastructure, software ecosystems, and reporting protocols. Future work may benefit from evaluating models with validation strategies appropriate for the intended prediction scenario, metrics that reveal local model behavior, and explicitly modeling and communicating uncertainty. These considerations underscore the need not only for methodological rigor but also for software that operationalizes these practices – scalable, reproducible workflows are essential components of both current and future frameworks. Furthermore, algorithmic innovation, presented alongside theoretical scrutiny and systematic comparisons, is essential to consolidate and expand current knowledge. Equally important is the transparent documentation of all methodological choices and the resulting model performance, which calls for the establishment of common reporting standards tailored to spatial predictive machine learning.

Building on this foundation, a first priority is the establishment of a spatially explicit evaluation framework tailored to predictive modeling with machine learning. This includes clearer guidance on when and how to apply spatial validation strategies. Beyond validation design, space-aware error metrics should be further researched and developed. Structured residual analysis and evaluation of map patterns for artifacts and domain compliance are needed to complement global performance measures. Second, uncertainty quantification requires scalable methods that move beyond pixel-level estimates toward aggregation over predefined spatial units and joint uncertainty across variables. It is also important to research how spatial uncertainty can be represented in ways that are understandable and decision-relevant for different stakeholders. Third, achieving comparability and cumulative progress depends on openly shared data and rigorously defined methods. Public benchmark datasets covering diverse spatial properties, together with a clearer taxonomy of spatial machine learning methods, would enable systematic and reproducible comparisons across algorithms and application domains. Such resources would reduce reliance on isolated case studies and support theory-informed method development. For fair and reproducible comparisons, assessments should employ guidelines that include baseline models, multiple datasets, and consistent performance metrics. Finally, software and reporting standards must evolve alongside methodological advances. The development of dedicated spatial machine learning tools is essential to reduce ad hoc implementations and enhance reproducibility. Complementing these tools, a standardized reporting protocol for spatial machine learning models would increase transparency and facilitate cross-study comparison. Such a protocol should be developed collaboratively by the community and provide clear guidance on validation strategies, treatment of uncertainty, software environments, and computational requirements.

Acknowledgement

We acknowledge that this work is partially based on discussions from the 'Advances in Spatial Machine Learning 2025' workshop and extend our sincere gratitude to all participants for their valu-

able contributions to the collaborative discussions that shaped the ideas presented in this paper. Jakub Nowosad has received the financial support of the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101147446. Carmelo Bonannella and Rolf Simoes were supported by the Open-Earth-Monitor Cyberinfrastructure project, funded by the European Union's Horizon Europe research and innovation programme (grant agreement No.101059548). Darius Görgen was supported by the TRR 391 Spatio-temporal Statistics for the Transition of Energy and Transport (520388526), and Jan Linnenbrink by the project BEyond within the priority Program 1374 "Biodiversity-Exploratories"(512284513), both funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). Teja Kattenborn was supported by the German Research Foundation (DFG) within the Emmy Noether Group PANOPS (504978936). Evelyn Uuemaa and Marta Jemeljanova were supported by the Estonian Research Agency (grant number PRG1764), Estonian Ministry of Education and Research, Centre of Excellence for Sustainable Land Use (TK232) and by the European Union (ERC, WaterSmartLand, 101125476). However, the views and opinions expressed are those of the authors only and do not necessarily reflect those of the funding agencies or affiliated institutions. Neither the funding bodies nor the supporting organizations can be held responsible.

Author contributions

JN coordinated and edited the manuscript and integrated all sections. HM drafted the introduction; TK, JL, and JN wrote the section on strategies for validation of models and spatial patterns; EU drafted the section on model performance metrics; MN and DG wrote the section on quantifying and communicating uncertainty; MJ and LP wrote the section on algorithm diversity and comparability; CB, RS, MJ and JN contributed the section on spatial machine learning workflows in R and Python; and JL prepared the section on standardized modeling protocols. MJ, EU, and JN worked on the sections about core takeaways and future research directions. All authors contributed to reviewing and revising the manuscript and approved the submitted version. JN is first author as the lead coordinator and integrator of the manuscript; the other authors are listed alphabetically.

Declaration of generative AI use

Generative AI tools assisted in the grammatical and stylistic editing of parts of this manuscript. No AI tools were employed to generate ideas, arguments, or conclusions.

References

Ahn S, Ryu D-W, Lee S (2020) A machine learning-based approach for spatial estimation using the spatial features of coordinate information. *ISPRS International Journal of Geo-information* 9: 587. https://doi.org/10.3390/ijgi9100587

An Y, Shim W, Jeong G (2023) High-resolution digital soil maps of forest soil nitrogen across South Korea using three machine learning algorithms. *Forests* 14: 1141. https://doi.org/10.3390/f14061141

Baltensweiler A, Walthert L, Hanewinkel M, Zimmermann S, Nussbaum M (2021) Machine learning based soil maps for a wide range of soil properties for the forested area of Switzerland. *Geoderma Regional* 27: e00437. https://doi.org/10.1016/j.geodrs.2021.e00437

Bastin J-F, Finegold Y, Garcia C, Mollicone D, Rezende M, Routh D, Zohner CM, Crowther TW (2019) The global tree restoration potential. *Science* 365: 76–79. https://doi.org/10.1126/science.aax0848

Becker M, Schratz P (2025) mlr3spatial: Support for spatial objects within the 'mlr3' ecosystem. https://doi.org/10.32614/CRAN.package.mlr3spatial

Beguin J, Fuglstad G-A, Mansuy N, Paré D (2017) Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. *Geoderma* 306: 195–205. https://doi.org/10.1016/j.geoderma.2017.06.016

Benito BM (2021) spatialRF: Easy spatial regression with random forest. https://doi.org/10.5281/zenodo.4745208

Bouasria A, Bouslihim Y, Gupta S, Taghizadeh-Mehrjardi R, Hengl T (2023) Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents. *Ecological Informatics* 78: 102294. https://doi.org/10.1016/j.ecoinf.2023.102294

Brodie SJ, Thorson JT, Carroll G, Hazen EL, Bograd S, Haltuch MA, Holsman KK, Kotwicki S, Samhouri JF, Willis-Norton E, Selden RL (2020) Trade-offs in covariate selection for species distribution models: A methodological comparison. *Ecography* 43: 11–24. https://doi.org/10.1111/ecog.04707

Brugere L, Kwon Y, Frazier AE, Kedron P (2023) Improved prediction of tree species richness and interpretability of environmental drivers using a machine learning approach. *Forest Ecology and Management* 539: 120972. https://doi.org/10.1016/j.foreco.2023.120972

Bruin S de, Brus DJ, Heuvelink GBM, Ebbenhorst Tengbergen T van, Wadoux AMJ-C (2022) Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecological Informatics* 69: 101665. https://doi.org/10.1016/j.ecoinf.2022.101665

Chen, W., Li, Y., Reich, B. J. a. Sun, Y. (2022) DeepKriging: Spatially dependent deep neural networks for spatial prediction. *arXiv*. https://arxiv.org/abs/2007.11972

Chrysanthopoulos E, Kallioras A (2025) Temporal and geographic extrapolation of soil moisture using machine learning algorithms. Catena 257: 109156. https://doi.org/10.1016/j.catena.2025.109156

Collins GS, Reitsma JB, Altman DG, Moons K (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Medicine* 13: 1. https://doi.org/10.1186/s12916-014-0241-z

Courteille L (2025) Impact of cartographic representation of uncertainty on decision-making : The case of soil quality in land-use planning. PhD thesis. https://doi.org/10.13140/RG.2.2.19285.95205

Cressie NAC (1993) Statistics for spatial data. New York

Davison A, Hinkley D (1997) Bootstrap methods and their application. Cambridge

Dormann C, M. McPherson J, B. Araújo M, Bivand R, Bolliger J, Carl G, G. Davies R, Hirzel A, Jetz W, Daniel Kissling W, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM, Wilson R (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* 30: 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Fernández-Delgado M, Cernadas E, Barro S, Amorim D (2014) Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15: 3133–3181. https://doi.org/10.5555/2627435.2697065

Gareth J, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning: With applications in R. Springer texts in statistics. New York; Heidelberg; Dordrecht; London

Gaspard G, Kim D, Chun Y (2019) Residual spatial autocorrelation in macroecological and biogeographical modeling: A review. *Journal of Ecology and Environment* 43: 19. https://doi.org/10.1186/s41610-019-0118-3

Ghariani, W. (2023) Spatial-kfold: A python package for spatial resampling toward more reliable cross-validation in spatial studies. https://github.com/WalidGharianiEAGLE/spatial-kfold

Goetz JN, Brenning A, Petschko H, Leopold P (2015) Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. *Computers & Geosciences* 81: 1–11. https://doi.org/10.1016/j.cageo.2015.04.007

Goodling P, Belitz K, Stackelberg P, Fleming B (2024) A spatial machine learning model developed from noisy data requires multiscale performance evaluation: Predicting depth to bedrock in the Delaware river basin, USA. *Environmental Modelling & Software* 179: 106124. https://doi.org/10.1016/j.envsoft.2024.106124

Grimm V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, Goss-Custard J, Grand T, Heinz SK, Huse G, Huth A, Jepsen JU, Jørgensen C, Mooij WM, Müller B, Pe'er G, Piou C, Railsback SF, Robbins AM, Robbins MM, Rossmanith E, Rüger N, Strand E, Souissi S, Stillman RA, Vabø R, Visser U, DeAngelis DL (2006) A standard protocol for describing individual-based and agent-based models. *Ecological Modelling* 198: 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Heuvelink GBM, Burrough PA, Stein A (1989) Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems* 3: 303–322. https://doi.org/10.1080/02693798908941518

Heuvelink G, Webster R (2022) Spatial statistics and soil mapping: A blossoming partnership under pressure. *Spatial Statistics* 50: 100639. https://doi.org/10.1016/j.spasta.2022.100639

Jacob C, Masilamani US (2025) Performance evaluation of different machine learning algorithms for prediction of nitrate in groundwater in Thiruvannamalai District. *CLEAN – Soil, Air, Water* 53: 2400060. https://doi.org/10.1002/clen.202400060

Jemeljanova M, Kmoch A, Uuemaa E (2024) Adapting machine learning for environmental spatial data - A review. *Ecological Informatics* 81: 102634. https://doi.org/10.1016/j.ecoinf.2024.102634

Jonard F, Feldman AF, Short Gianotti DJ, Entekhabi D (2022) Observed water and light limitation across global ecosystems. *Biogeosciences* 19: 5575–5590. https://doi.org/10.5194/bg-19-5575-2022

Kapoor S, Cantrell EM, Peng K, Pham TH, Bail CA, Gundersen OE, Hofman JM, Hullman J, Lones MA, Malik MM, Nanayakkara P, Poldrack RA, Raji ID, Roberts M, Salganik MJ, Serra-Garcia M, Stewart BM, Vandewiele G, Narayanan A (2024) REFORMS: Consensus-based Recommendations for Machine-learning-based Science. *Science Advances* 10: eadk3452. https://doi.org/10.1126/sciadv.adk3452

Kattenborn T, Schiefer F, Frey J, Feilhauer H, Mahecha MD, Dormann CF (2022) Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing* 5: 100018. https://doi.org/10.1016/j.ophoto.2022.100018

Kim H-J, Mawuntu KBA, Park T-W, Kim H-S, Park J-Y, Jeong Y-S (2023) Spatial autocorrelation incorporated machine learning model for geotechnical subsur-

face modeling. *Applied Sciences* 13: 4497. https://doi.org/10.3390/app13074497

Kmoch A, Harrison CT, Choi J, Uuemaa E (2025) Spatial autocorrelation in machine learning for modelling soil organic carbon. *Ecological Informatics* 86: 103057. https://doi.org/10.1016/j.ecoinf.2025.103057

Koldasbayeva D, Tregubova P, Gasanov M, Zaytsev A, Petrovskaia A, Burnaev E (2024) Challenges in data-driven geospatial modeling for environmental research and practice. *Nature Communications* 15: 10700. https://doi.org/10.1038/s41467-024-55240-8

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

Kuhn M, Wickham H (2020) Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org

Kumar C, Walton G, Santi P, Luza C (2025) Random cross-validation produces biased assessment of machine learning performance in regional landslide susceptibility prediction. *Remote Sensing* 17. https://doi.org/10.3390/rs17020213

Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B (2019) mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*. https://doi.org/10.21105/joss.01903

Legendre P (1993) Spatial autocorrelation: Trouble or new paradigm? *Ecology* 74: 1659–1673. https://doi.org/10.2307/1939924

Liemohn MW, Shane AD, Azari AR, Petersen AK, Swiger BM, Mukhopadhyay A (2021) RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics* 218: 105624. https://doi.org/10.1016/j.jastp.2021.105624

Linnenbrink J, Milà C, Ludwig M, Meyer H (2024) kNNDM CV: K-fold nearest-neighbour distance matching cross-validation for map accuracy estimation. *Geoscientific Model Development* 17: 5897–5912. https://doi.org/10.5194/gmd-17-5897-2024

Liu FT, Ting KM, Zhou Z-H (2008) Isolation Forest. *Proceedings of the 2008 eighth IEEE international conference on data mining*: 413–422. ICDM '08. USA. https://doi.org/10.1109/ICDM.2008.17

Ludwig M, Moreno-Martinez A, Hölzel N, Pebesma E, Meyer H (2023) Assessing and improving the transferability of current global spatial prediction models. *Global Ecology and Biogeography* 32: 356–368. https://doi.org/10.1111/geb.13635

Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M (2016) Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. *Journal of Medical Internet Research* 18: e323. https://doi.org/10.2196/jmir.5870

Lusk D, Wolf S, Svidzinska D, Dormann CF, Kattge J, Bruelheide H, Sabatini FM, Damasceno G, Moreno Martínez Á, Violle C, Hending D, Hähn GJA, Tabeni S, Phartyal S, Gonçalves F, Kreft H, Schmidt M, Chen H, Güler B, Dolezal J, Pielech R, Guido A, Dwyer C, Napoleone F, Willie J, Gasper AL, Macía MJ, Chytry M, Lenoir J, Thakur D, Dengler J, Świerszcz S, Altman J, Mucina L, Nerlekar AN, Kakinuma K, Rawat P, Stančić Z, Testolin R, Hatim MZ, Rodrigues F, Homeier J, Marques MCM, McCarthy JK, El-Sheikh MA, Korznikov K, Gerberding K, Kattenborn T (2026) Crowdsourced biodiversity monitoring fills gaps in global plant trait mapping. *Nature Communications* 17: 1203. https://doi.org/10.1038/s41467-026-68996-y

Mahoney, M. J. (2023) Waywiser: Ergonomic methods for assessing spatial models. *arXiv*. https://doi.org/10.48550/arXiv.2303.11312

Mahoney, M. J., Johnson, L. K., Silge, J., Frick, H., Kuhn, M. a. Beier, C. M. (2023) Assessing the performance of spatial cross-validation approaches for models of spatially structured data. *arXiv*. https://doi.org/10.48550/arXiv.2303.07334

Malone BP, Minasny B, McBratney AB (2016) Using R for digital soil mapping. Progress in soil science. Cham, Switzerland

Maxwell TL, Spalding MD, Friess DA, Murray NJ, Rogers K, Rovai AS, Smart LS, Weilguny L, Adame MF, Adams JB, Austin WEN, Copertino MS, Cott GM, Duarte De Paula Costa M, Holmquist JR, Ladd CJT, Lovelock CE, Ludwig M, Moritsch MM, Navarro A, Raw JL, Ruiz-Fernández A-C, Serrano O, Smeaton C, Van De Broek M, Windham-Myers L, Landis E, Worthington TA (2024) Soil carbon in the world's tidal marshes. *Nature Communications* 15: 10265. https://doi.org/10.1038/s41467-024-54572-9

Meinshausen N (2006) Quantile regression forests. *Journal of Machine Learning Research* 7: 983–999. http://jmlr.org/papers/v7/meinshausen06a.html

Meyer H, Milà C, Ludwig M, Linnenbrink J, Schumacher F (2025) CAST: 'Caret' applications for spatial-temporal models. https://doi.org/10.32614/CRAN.package.CAST

Meyer H, Pebesma E (2021) Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution* 12: 1620–1633. https://doi.org/10.1111/2041-210X.13650

Meyer H, Pebesma E (2022) Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications* 13: 2208. https://doi.org/10.1038/s41467-022-29838-9

Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T (2018) Improving performance of spatio-temporal ma-

chine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software* 101: 1–9. https://doi.org/10.1016/j.envsoft.2017.12.001

Meyer H, Reudenbach C, Woellauer S, Nauss T (2019) Importance of spatial predictor variable selection in machine learning applications - moving from data reproduction to spatial prediction. *Ecological Modelling* 411. https://doi.org/10.1016/j.ecolmodel.2019.108815

Mila C, Ludwig M, Pebesma E, Tonne C, Meyer H (2024) Random forests with spatial proxies for environmental modelling: Opportunities and pitfalls. *EGUsphere* 2024: 1–30. https://doi.org/10.5194/egusphere-2024-138

Mila C, Mateu J, Pebesma E, Meyer H (2022) Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. *Methods in Ecology and Evolution* 13: 1304–1316. https://doi.org/10.1111/2041-210X.13851

Milne AE, Glendining MJ, Lark RM, Perryman SAM, Gordon T, Whitmore AP (2015) Communicating the uncertainty in estimated greenhouse gas emissions from agriculture. *Journal of Environmental Management* 160: 139–153. https://doi.org/10.1016/j.jenvman.2015.05.034

Minasny B, Bandai T, Ghezzehei TA, Huang Y-C, Ma Y, McBratney AB, Ng W, Norouzi S, Padarian J, Rudiyanto, Sharififar A, Styc Q, Widyastuti M (2024) Soil science-informed machine learning. *Geoderma* 452: 117094. https://doi.org/10.1016/j.geoderma.2024.117094

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*: 220–229. https://doi.org/10.1145/3287560.3287596

Molnar C (2021) Interpretable machine learning: A guide for making black box models explainable. Victoria, Canada

Mongan J, Moy L, Kahn CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence* 2: e200029. https://doi.org/10.1148/ryai.2020200029

Naveau P, Allard D (2005) Modeling skewness in spatial data analysis without data transformation. In: *Geostatistics Banff 2004*: 929–937. https://doi.org/10.1007/978-1-4020-3610-1

Nikparvar B, Thill J-C (2021) Machine learning of spatial data. *ISPRS International Journal of Geo-Information* 10. https://doi.org/10.3390/ijgi10090600

Nowosad J, Meyer H (2025) Investigating Moran's I properties for spatial machine learning: A preliminary analysis. *AGILE: GIScience Series* 6: 1–6. https://doi.org/10.5194/agile-giss-6-40-2025

Nussbaum M, Spiess K, Baltensweiler A, Grob U, Keller A, Greiner L, Schaepman ME, Papritz A (2018) Evaluation of digital soil mapping approaches with large sets of environmental covariates. *SOIL*: 1–22. https://doi.org/10.5194/soil-4-1-2018

Pabon-Moreno DE, Migliavacca M, Reichstein M, Mahecha MD (2022) On the potential of Sentinel-2 for estimating gross primary production. *IEEE Transactions on Geoscience and Remote Sensing* 60: 1–12

Patelli L, Cameletti M, Golini N, Ignaccolo R (2024) A path in regression Random Forest looking for spatial dependence: A taxonomy and a systematic review. Knoth S, Okhrin Y, Otto P (eds) *Advanced statistical methods in process monitoring, finance, and environmental science*: 467–489. Cham. https://doi.org/10.1007/978-3-031-69111-9

Patriche CV, Roșca B, Pîrnău RG, Vasiliniuc I (2023) Spatial modelling of topsoil properties in Romania using geostatistical methods and machine learning. *Plos One* 18: e0289286. https://doi.org/10.1371/journal.pone.0289286

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay Édouard (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12: 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Piikki K, Wetterlind J, Söderström M, Stenberg B (2021) Perspectives on validation in digital soil mapping of continuous attributes - A review. *Soil Use and Management* 37: 7–21. https://doi.org/10.1111/sum.12694

Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, Lyapustin A, Gourlet-Fleury S, Pélissier R (2020) Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications* 11. https://doi.org/10.1038/s41467-020-18321-y

Rasmussen CE, Williams CKI (2008) Gaussian processes for machine learning. Adaptive computation and machine learning. Cambridge, Mass.

Rey SJ, Anselin L (2007) PySAL: A Python library of spatial analytical methods. *The Review of Regional Studies* 37: 5–27. https://doi.org/10.52324/001c.8285

Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Arroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle, BA, Hartig F, Dormann, CF (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40: 913–929. https://doi.org/10.1111/ecog.02881

Ryo M, Angelov B, Mammola S, Kass JM, Benito BM, Hartig F (2021) Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography* 44: 199–205. https://doi.org/10.1111/ecog.05360

Saha A, Basu S, Datta A (2023) Random forests for spatially dependent data. *Journal of the American Statistical Association* 118. https://doi.org/10.1080/01621459.2021.1950003

Saha A, Basu S, Datta A (2024) RandomForestsGLS: Random forests for dependent data. https://doi.org/10.32614/CRAN.package.RandomForestsGLS

Schmidinger J, Heuvelink GBM (2023) Validation of uncertainty predictions in digital soil mapping. *Geoderma* 437: 116585. https://doi.org/10.1016/j.geoderma.2023.116585

Schratz P, Becker M, Lang M, Brenning A (2024) mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in R. *Journal of Statistical Software* 111: 1–36. https://doi.org/10.18637/jss.v111.i07

Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A (2019) Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling* 406: 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Schumacher FL, Knoth C, Ludwig M, Meyer H (2025) Estimation of local training data point densities to support the assessment of spatial prediction uncertainty. *Geoscientific Model Development* 18: 10185–10202. https://doi.org/10.5194/gmd-18-10185-2025

Sekulić A, Kilibarda M, Heuvelink GBM, Nikolić M, Bajat B (2020) Random forest spatial interpolation. *Remote Sensing* 12: 1687. https://doi.org/10.3390/rs12101687

Simoes R, Camara G, Queiroz G, Souza F, Andrade P, Santos L, Carvalho A, Ferreira K (2021) Satellite image time series analysis for big earth observation data. *Remote Sensing* 13: 2428. https://doi.org/10.3390/rs13132428

Sterlacchini S, Ballabio C, Blahut J, Masetti M, Sorichetta A (2011) Spatial agreement of predicted patterns in landslide susceptibility maps. *Geomorphology* 125: 51–61. https://doi.org/10.1016/j.geomorph.2010.09.004

Stock A, Gregr EJ, Chan KMA (2023) Data leakage jeopardizes ecological applications of machine learning. *Nature Ecology & Evolution* 7: 1743–1745. https://doi.org/10.1038/s41559-023-02162-1

Stojanova D, Ceci M, Appice A, Malerba D, Džeroski S (2013) Dealing with spatial autocorrelation when learning predictive clustering trees. *Ecological Informatics* 13: 22–39. https://doi.org/10.1016/j.ecoinf.2012.10.006

Sweet L, Müller C, Anand M, Zscheischler J (2023) Cross-validation strategy impacts the performance and interpretation of machine learning models. *Artificial Intelligence for the Earth Systems* 2. https://doi.org/10.1175/aies-d-23-0026.1

Takoutsing B, Heuvelink GBM, Stoorvogel JJ, Shepherd KD, Aynekulu E (2022) Accounting for analytical and proximal soil sensing errors in digital soil mapping. *European Journal of Soil Science* 73: e13226. https://doi.org/10.1111/ejss.13226

Tian X, Bruin S de, Simoes R, Isik MS, Minarik R, Ho Y-F, Şahin M, Herold M, Consoli D, Hengl T (2025) Spatiotemporal prediction of soil organic carbon density in Europe (2000–2022) using earth observation and machine learning. *PeerJ* 13: e19605. https://doi.org/10.7717/peerj.19605

Tobler WR (1970) A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G (2019) blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution* 10: 225–232. https://doi.org/10.1111/2041-210X.13107

Verma V (2025) A comprehensive framework for residual analysis in regression and machine learning. *Journal of Information Systems Engineering and Management* 10: 34–46. https://doi.org/10.52783/jisem.v10i31s.4958

Wadoux AM-C, Heuvelink GB, De Bruin S, Brus DJ (2021) Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling* 457: 109692

Wadoux AMJ-C, Heuvelink GBM (2023) Uncertainty of spatial averages and totals of natural resource maps. *Methods in Ecology and Evolution* 14: 1320–1332. https://doi.org/10.1111/2041-210X.14106

Wang S, Gao K, Zhang L, Yu B, Easa SM (2024) Geographically weighted machine learning for modeling spatial heterogeneity in traffic crash frequency and determinants in US. *Accident Analysis & Prevention* 199: 107528. https://doi.org/10.1016/j.aap.2024.107528

Wang Y, Khodadadzadeh M, Zurita-Milla R (2025) A dissimilarity-adaptive cross-validation method for evaluating geospatial machine learning predictions with clustered samples. *Ecological Informatics* 90: 103287. https://doi.org/10.1016/j.ecoinf.2025.103287

Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Chichester, England. https://doi.org/10.1002/9780470517277

Wei J, Huang W, Li Z, Xue W, Peng Y, Sun L, Cribb M (2019) Estimating 1-km-resolution PM2.5 concentrations across China using the space-time random forest approach. *Remote Sensing of Environment* 231: 111221. https://doi.org/10.1016/j.rse.2019.111221

Yates KL, Bouchet PJ, Caley MJ, Mengersen K, Randin CF, Parnell S, Fielding AH, Bamford AJ, Ban S, Barbosa AM, others (2018) Outstanding challenges in the transferability of ecological models. *Trends in Ecology & Evolution* 33: 790–802. https://doi.org/10.1016/j.tree.2018.08.001

Zeevi T, Lieffrig EV, Staib LH, Onofrey JA (2025) Spatially-aware evaluation of segmentation uncertainty. *arXiv*. https://doi.org/10.48550/arXiv.2506.16589

Zhao W, Li Z, Li H, Li X, Yang P (2024) Soil salinity prediction in an arid area based on long time-series multispectral imaging. *Agriculture* 14. https://doi.org/10.3390/agriculture14091539

Zurell D, Franklin J, König C, Bouchet PJ, Dormann CF, Elith J, Fandos G, Feng X, Guillera-Arroita G, Guisan A, Lahoz-Monfort JJ, Leitão PJ, Park DS, Peterson AT, Rapacciuolo G, Schmatz DR, Schröder B, Serra-Diaz JM, Thuiller W, Yates KL, Zimmermann NE, Merow C (2020) A standard protocol for reporting species distribution models. *Ecography* 43: 1261–1277. https://doi.org/10.1111/ecog.04960

Authors

Dr. Jakub Nowosad
https://orcid.org/0000-0002-1057-3721
nowosad.jakub@gmail.com
Institute of Landscape Ecology
University of Münster
Heisenbergstrasse 2
D-48149 Münster
Germany
and
Institute of Geoecology and Geoinformation
Adam Mickiewicz University
ul. B. Krygowskiego 10
61-680 Poznań
Poland

Dr. Carmelo Bonannella
https://orcid.org/0000-0002-5391-8427
carmelo.bonannella@opengeohub.org
OpenGeoHub Foundation
Waldeck Pyrmontlaan 14
6865 HK Doorwerth
Netherlands

Darius Görgen
https://orcid.org/0009-0008-5503-7704
dgoergen@uni-muenster.de
Jan Linnenbrink
https://orcid.org/0000-0003-0991-8646
jan.linnenbrink@uni-muenster.de
Prof. Dr. Hanna Meyer
https://orcid.org/0000-0003-0556-0210
hanna.meyer@uni-muenster.de
Institute of Landscape Ecology
University of Münster
Heisenbergstrasse 2
D-48149 Münster
Germany

Marta Jemeljanova
https://orcid.org/0000-0002-7063-2236
marta.jemeljanova@ut.ee
Prof. Dr. Evelyn Uuemaa
https://orcid.org/0000-0002-0782-6740
evelyn.uuemaa@ut.ee
Department of Geography
University of Tartu
Vanemuise 46
Tartu 50409
Estonia

Prof. Dr. Teja Kattenborn
https://orcid.org/0000-0001-7381-3828
Teja.Kattenborn@geosense.uni-freiburg.de
Chair of Sensor-based Geoinformatics (geosense)
University of Freiburg
Tennenbacher Str. 4
79106 Freiburg
Germany

Dr. Madlene Nussbaum
https://orcid.org/0000-0002-6808-8956
m.nussbaum@uu.nl
Department of Physical Geography
Utrecht University
Princetonlaan 8a
3584 CB Utrecht
Netherlands

Dr. Luca Patelli
https://orcid.org/0000-0001-5536-1047
luca.patelli@unibg.it
Department of Economics
University of Bergamo
Via dei Caniana 2
24127 Bergamo
Italy

Dr. Rolf Simoes
https://orcid.org/0000-0003-0953-4132
rolf.simoes@fgv.br
Agribusiness Studies Center (FGV Agro)
Fundação Getulio Vargas (FGV)
Av. Paulista 542, 3º andar
01310-000 São Paulo
Brazil