Journal of Hydrology 666 (2026) 134797

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Impacts of validation strategy and geological complexity on machine learning-based subsurface structure modeling

Chuanjun Zhan [a,b,c], Jiu Jimmy Jiao [a,b,*], Yue Ma [c], Hao Wang [e], Mohamad Reza Soltanian [f], Kenneth C. Carroll [g], Zhenxue Dai [c,d]

[a] Department of Earth and Planetary Sciences, The University of Hong Kong, Hong Kong, China
[b] Guangdong-Hong Kong Joint Laboratory for Soil and Groundwater Pollution Control, China
[c] School of Environmental and Municipal Engineering, Qingdao University of Technology, Qingdao 266520, China
[d] College of Construction Engineering, Jilin University, Changchun 130026, China
[e] Water Resources Research Institute of Shandong Province, Jinan 250014, China
[f] Departments of Geosciences and Environmental Engineering, University of Cincinnati, Cincinnati, OH, USA
[g] Department of Plant & Environmental Science, New Mexico State University, Las Cruces, NM, USA

ARTICLE INFO

This manuscript was handled by Dan Lu, Editor-in-Chief, with the assistance of Jiangjiang Zhang, Associate Editor

ABSTRACT

Assessing model performance is critical in machine learning (ML) based subsurface structure modeling. Random cross-validation (R-CV), a commonly used evaluation method in general ML applications, typically does not consider the spatial correlation structure that exists in geological borehole data. This study systematically evaluates how this validation strategy affects performance assessment when applied to spatially correlated geological datasets. Seven ML models were tested using both a synthetic dataset and a complex case from the Pearl River Delta (PRD), comparing results from R-CV with those from spatial cross-validation (S-CV), which ensures spatial separation between training and testing data. The results show that R-CV consistently overestimates performance by measuring interpolation rather than extrapolation. Under rigorous S-CV testing, all models showed substantial accuracy decreases, providing more realistic performance benchmarks. In the complex PRD case, all models converged to a similar accuracy ceiling under S-CV. Despite comparable accuracies, their predicted geological morphologies differed significantly. This finding indicates that model selection should be based not just on accuracy, but on whether a model's inherent tendency to generate specific geological shapes aligns with prior geological knowledge. All models also had difficulty capturing complex and alternating facies distributions, which reflects the limitations of relying on sparse borehole data. Incorporating additional sources of information, such as geophysical data, is necessary to overcome these limitations and to better represent the spatial continuity of subsurface facies.

1. Introduction

Developing accurate models of subsurface geology is fundamental to a wide range of applications, including underground space planning, civil engineering design, hydrogeological assessment, mineral exploration, groundwater contamination, and geological hazard prevention (Bianchi and Zheng, 2016; Ershadnia et al., 2021; Kitanidis, 2015; Rajaram and Gelhar, 1995; Wallace et al., 2021). Traditionally, geological modeling has relied on geostatistical methods. These include two-point statistics (e.g., Kriging) and multi-point statistics (MPS) (Carle and Fogg, 1996; Remy et al., 2009; Strebelle, 2002). These methods have proven capable of statistically reproducing observed data and generating realizations that honor available measurements while providing predictions within a quantifiable range of uncertainty (Rubin, 2003; Yeh et al., 2015). However, their simplicity comes with conceptual limits. Two-point statistics reproduce correlations and global proportions but cannot encode the hierarchical organization of deposits, such as bounding surfaces or stacking patterns, that govern connectivity and flow. MPS extends this capability by incorporating training images to capture more complex spatial patterns, yet these training images are often unavailable or insufficiently representative when only sparse borehole data exist (Mariethoz and Caers, 2014). As a result, while

* Corresponding author at: Department of Earth and Planetary Sciences, The University of Hong Kong, Hong Kong, China.
E-mail address: jjiao@hku.hk (J.J. Jiao).

https://doi.org/10.1016/j.jhydrol.2025.134797
Received 12 November 2025; Received in revised form 4 December 2025; Accepted 15 December 2025
Available online 17 December 2025

geostatistical approaches remain powerful tools for data-consistent modeling, they tend to smooth the underlying architecture that controls geological realism. In this context, geological complexity—primarily sedimentary heterogeneity—plays an important role in controlling contaminant transport in porous media (Agbotui et al., 2025; Bianchi and Pedretti, 2018).

Recent advances in machine learning (ML) and deep learning (DL) have introduced powerful new approaches to subsurface structure modeling (Awawdeh et al., 2025; Galiyev et al., 2025; Hasan et al., 2025; Lin et al., 2024; Zhan et al., 2023). These data-driven methods generally fall into two main categories. The first category includes generative artificial intelligence models, such as Generative Adversarial Networks (GANs) (Cui et al., 2024; Lyu et al., 2024; Song et al., 2022), Variational Autoencoders (VAEs) (Chen et al., 2026; Kang et al., 2025), and diffusion models (Di Federico and Durlofsky, 2025; Zhan et al., 2025). These models are designed to include and work with data assimilation algorithms. They identify subsurface structures by adjusting the model's input variables based on multiple data sources (Zhang et al., 2024). While this approach has gained significant attention, the training of these core generative models also relies on a large number of reliable training images (i.e., prior geological models) (Merzoug and Pyrcz, 2025). Some studies try to generate these training samples with geostatistical methods. But this often results in the same stationarity problem of traditional methods (Zhan et al., 2022b). The second category of methods is more analogous to geostatistical methods. It uses ML or DL models to directly learn the complex, non-linear spatial distribution features of facies (i.e., bodies of sediments, see Soltanian and Ritzi, 2014) from borehole data. The models then make predictions for unsampled locations (Hang et al., 2025; Liu et al., 2025).

An emerging concept is that these two ML approaches (generative models and direct prediction) can be used in a complementary way. Geological structures predicted directly from borehole data, provided they possess both geological plausibility and rigorously-assessed reliability, can serve as training images for generative models such as GANs and VAEs. Generating high-quality training data in this manner has the potential to significantly enhance the performance of generative frameworks and associated data assimilation workflows, leading to more robust and geologically consistent predictions. As such, directly using ML or DL models to learn spatial facies patterns from borehole data is a promising direction. This approach resembles geostatistical interpolation but does not require a predefined training image. It offers a path toward greater automation and improved predictive accuracy in subsurface modeling (Guo, 2024; Hu et al., 2024; Jordão et al., 2023; Shi and Wang, 2022; Wang et al., 2025a).

As machine learning becomes more widely applied to subsurface modeling, a notable trend has emerged in how model performance is evaluated. Many studies have adopted random cross validation (RCV) as the default validation strategy. In this method, all available data points are pooled and then randomly partitioned into training and test sets. However, a critical methodological question arises from applying this standard validation strategy to geological data. By design, R-CV treats each data point as an independent sample. This assumption may not fully align with the nature of geological borehole data, which is fundamentally characterized by strong spatial autocorrelation (i.e., nearby data points are highly similar) (Juda et al., 2020; Wadoux et al., 2021). This discrepancy raises the question of what impact this methodological choice has on the interpretation of model performance. It is essential to determine whether a validation strategy that overlooks spatial correlation provides a comprehensive evaluation of a model's ability to extrapolate to new, or previously undrilled locations.

This study is motivated by a fundamental question about how model validation strategies affect the interpretation of ML performance in subsurface modeling. Rather than focusing solely on methods and algorithms, the goal of this study is to explore how validation strategy interacts with geological complexity and data sparsity to shape our understanding of model capability. To achieve this goal, we address three specific questions: (i) What is the quantitative impact of using standard R-CV compared to a validation strategy that explicitly enforces spatial separation (i.e., S-CV)? (ii) What is the realistic performance limit for true spatial extrapolation in geologically complex settings with sparse data, and under what conditions are ML-based modeling approaches most effective? (iii) What are the underlying causes of this performance limit in models that rely solely on sparse borehole observations?

2. Method

2.1. Research framework

This study employs a multi-layered, controlled experimental design to isolate and quantify the independent and combined effects of validation strategy and geological complexity (Fig. 1). The workflow has three main components: a combination of synthetic and real-world datasets, a diverse library of machine learning models, and a comparative analysis of two core validation methods.

To ensure generalizability and robustness, this study uses a dual-dataset strategy. First, a three-dimensional (3D) synthetic geological model with a completely known facies distribution was built. This model acts as an absolute "ground truth." This controlled environment allows for a precise measurement of how different experimental conditions affect the models' "true accuracy." Next, to test the findings in a real-world case, a dataset from the Pearl River Delta (PRD), China, was introduced. Its complex geology provides a challenging scenario for testing model generalization.

The study also established a comprehensive model library containing seven mainstream machine learning models. This ensures that the findings do not depend on a single specific algorithm. These models were grouped into three functional types: Neighborhood-Based Models, Global Function Fitting Models, and Ensemble Spatial Partitioning Models, covering a wide range of algorithmic approaches. This diverse selection helps determine whether the choice of algorithm is the primary factor or a secondary one in relation to data and validation methodology.

The core of the method is a comparison of two different model validation strategies (Fig. 2). The first strategy is the "point-wise split" (a standard R-CV), and the second is the "borehole-wise split" (a S-CV). This study systematically compares these two distinct strategies to quantify the impact of data partitioning choice on the final performance assessment. The detailed procedures for each strategy are described in Section 2.3.

2.2. Machine learning model

This study selected seven ML models. The selection included six simple, common ML models. It also included one complex DL model. A key goal was to compare the performance of models with different levels of complexity. In this modeling task, all models solve the same core problem. They must learn a predictive function from sparse borehole data. This function is used to classify facies at unknown locations.

The model inputs and outputs must be clearly defined. For all models, the basic query is the 3D spatial coordinate (x, y, z) of an unknown point. The final output is always the predicted facies for that specific point. However, different models use this (x, y, z) input in very different ways. Models like SVM and RF (Global Function and Ensemble models) use the (x, y, z) coordinates directly. They feed the coordinates into a pre-trained function. In contrast, Neighborhood-dependent models (e.g., KNN and AE-ResCapsNet) work differently. They use the (x, y, z) coordinates to first find nearby data points from the training set. This local neighbor information is then used to make the final prediction.

The models are grouped into three categories. This grouping is based on their core working principles and 'inductive biases'. An 'inductive

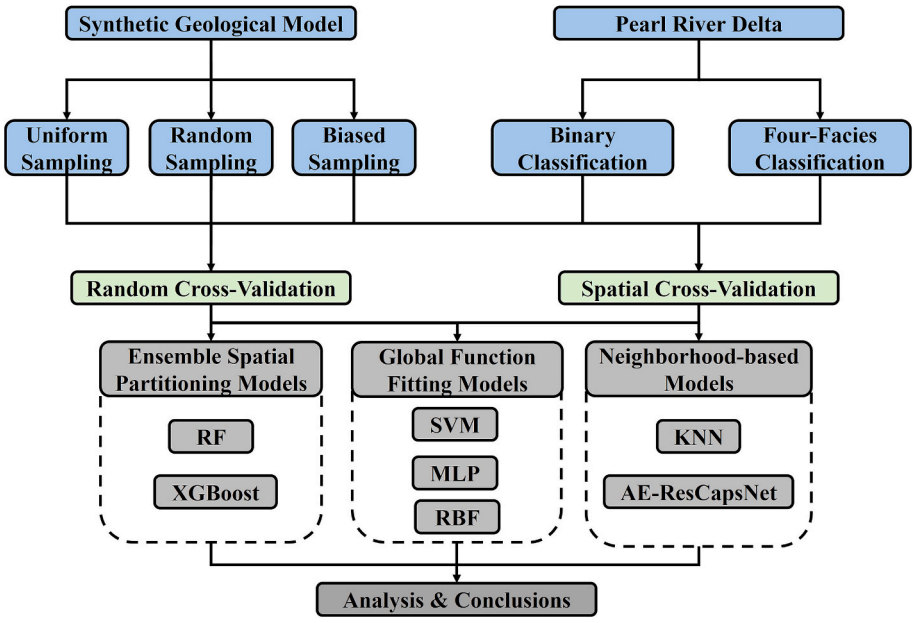

Fig. 1. Schematic of the research framework illustrating the three main components of this study: (1) the dual-dataset strategy combining a synthetic 3D geological model with known ground truth and a real-world case study from the Pearl River Delta (PRD); (2) the machine learning model library comprising seven models grouped into three functional categories (Neighborhood-Based, Global Function Fitting, and Ensemble Spatial Partitioning); and (3) the comparative validation framework contrasting point-wise split (random cross-validation, R-CV) with borehole-wise split (spatial cross-validation, S-CV) strategies. Arrows indicate the workflow from data preparation through model training and validation to performance comparison and analysis.

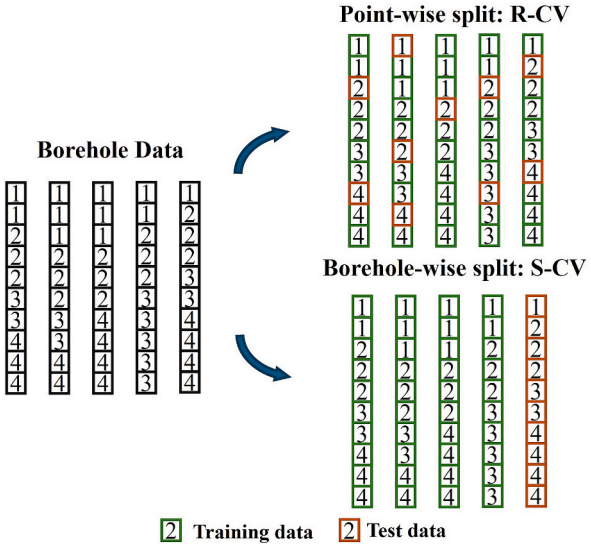

Fig. 2. The comparison of two fundamentally different model validation strategies.

bias' refers to a model's inherent tendency to learn in a specific manner. This bias is very important. It directly controls the shape of the predicted geological structures in areas with sparse data. Appendix A provides detailed descriptions of each model's mathematical and architectural components.

The categories and their predictive processes are described below.

(1) Neighborhood-based Models: This category of models relies completely on local neighborhood information. Their core assumption is simple: nearby points have similar properties (Peterson, 2009). These models do not build a single function for the whole area. Instead, their prediction process starts when a query point (x, y, z) is entered. The model first searches the training data for the closest neighbors. The final prediction is based only on these local neighbors (X. Wang et al., 2018). This category includes K-Nearest Neighbors (KNN). KNN is a simple algorithm that finds the k nearest points and takes a "vote" to decide the facies (Martín-Martín et al., 2023). It also includes the AutoEncoder with Residual Capsule Network (AE-ResCapsNet). The implementation in this study is constructed based on the architecture and parameters expressed in the work of He et al., (2025), to which the reader is referred for detailed structural diagrams and a comprehensive description. The model's process has three main stages. First, it samples a "point cloud" of many neighbors. Second, an Autoencoder processes the relative coordinates of these neighbors into a compact feature vector. Third, a Residual Capsule Network uses this feature vector to make the final classification.

(2) Global Function Fitting Models: This category of models learns a single, continuous function (x, y, z). This function covers the entire coordinate space. The inductive bias of these models favors creating smooth geological boundaries (Vaferi et al., 2011). Their prediction process has two parts. First, during training, the model fits this global function f to best separate the facies classes. Second, during prediction, a query point (x, y, z) is simply fed into this learned function. The function's output directly determines the facies class. Models in this category include the Support Vector Machine (SVM). SVM finds an optimal "hyperplane" to separate classes, using a "kernel trick" for complex data (Cervantes et al., 2020; Hearst et al., 1998; Smirnoff et al., 2008). This group also includes the Radial Basis Function (RBF) Network, a three-layer neural network (Juliani and Ellefmo, 2019; Wu et al., 2012). Finally, the Multi-Layer Perceptron (MLP) is a standard neural network that learns complex, non-linear patterns from the (x, y, z) coordinates (Marquina-Araujo et al., 2024; Pei and Zhang, 2022).

(3) Ensemble Spatial Partitioning Models: This category of models builds a final prediction by combining many simple decision trees. The prediction process for these models is based on "spatial partitioning." (Costa and Pedreira, 2023). During training, the

models repeatedly split the 3D space into smaller 'blocks' (like 'is $z < 50$ m?' or 'is $x > 100$ m?'). During prediction, a query point (x, y, z) is assigned to one of these specific blocks. The final output facies are simply the majority facies of the training samples that fell into that same block. This process creates a clear inductive bias. The model's predictions naturally form "step-like" or "blocky" boundaries. This group includes Random Forest (RF), which builds many trees on random data samples and averages their votes (Breiman, 2001; Kuhn et al., 2018; Rodriguez-Galiano et al., 2015). It also includes Extreme Gradient Boosting (XGBoost), which builds trees sequentially. Each new tree corrects the errors of the previous one (Abbas et al., 2023; Chen et al., 2015).

2.3. Experimental setup and validation strategies

The core of this study is a controlled, multi-run experimental framework. It was designed to systematically test and contrast the impact of different validation strategies on model performance assessment. This framework ensures that the conclusions are statistically robust and reduces biases from a single, random data split.

The complete experimental pipeline was independently repeated 10 times for each combination of sampling strategy and machine learning model, using different random seeds. The goal of this multi-run procedure is to get a robust estimate of each model's average performance and its stability across different data subsets. Each of the 10 independent runs followed a structured procedure. First, the dataset was split into a 90 % training set and a 10 % test set according to one of two distinct strategies. Second, a 5-fold cross-validation procedure was used on the 90 % training set for model training and hyperparameter selection (Nti et al., 2021). The training set was split into five folds, and the model was iteratively trained on four folds and validated on the remaining one. The model with the best average performance was selected. Third, the selected model was then evaluated on the independent 10 % test set. The final reported performance is the mean and standard deviation of the facies identification accuracy from these 10 independent runs.

Two distinct data splitting strategies were employed to evaluate the impact of spatial autocorrelation on model performance (Fig. 2). The first strategy, the point-wise split, treats all data points as independent units, irrespective of their borehole of origin. All data points from every borehole are first pooled together. A standard stratified random split is then performed on this pooled dataset to partition it into a 90 % training set and a 10 % test set. Because points from the same borehole can be distributed into both the training and test sets, this strategy represents a scenario where training and test points may be spatially close.

The second strategy, the borehole-wise split, is a form of spatial cross-validation. It treats each borehole as a single, indivisible unit to maintain spatial integrity (Sun et al., 2023). The procedure begins by identifying the complete list of unique boreholes. A random split is then performed at the borehole level, assigning 90 % of the boreholes to the training set and the remaining 10 % to the test set. All data points belonging to the boreholes assigned to the training set make up the final training dataset. All points from the boreholes assigned to the test set make up the final test dataset. This method ensures a complete spatial separation between the training and test data, representing a scenario focused on spatial 'extrapolation' (predicting in new areas). Before the 10-run evaluation, a one-time hyperparameter optimization was conducted for each model using 5-fold stratified cross-validation. This ensured all subsequent comparisons were based on each algorithm's optimal configuration.

3. The synthetic case study

3.1. Synthetic geological model

A 3D synthetic geological model was created to evaluate the true performance of various machine learning models. The model works as a controllable ground truth. The model covers a physical domain of 120 m $\times$ 120 m $\times$ 60 m. This domain is divided into a regular, structured grid of 60 $\times$ 60 $\times$ 60 voxels. This discretization gives a total of 216,000 voxels. Each voxel has a resolution of 2 m $\times$ 2 m $\times$ 1 m. The model comprises four facies stacked vertically in sequence, as shown in Fig. 3(a). The model includes three geological interfaces between the facies, and the interfaces were defined using sinusoidal functions:

$$z(x,y) = z_{base} + \sum_{i=1}^{N} A_i \bullet \sin\left(f_{x,i} \bullet x\right) \bullet \cos\left(f_{y,i} \bullet y\right) + \sigma \quad (1)$$

here, N represents the total number of sinusoidal terms used to generate the complexity of the geological interface (N = 4 in this study). where z_{base} is the base elevation of the interface, A_i are the amplitude coefficients, and $f_{x,i}$ and $f_{y,i}$ are the spatial frequencies in the x and y directions. This approach preserves the macroscopic layered characteristics of the model while introducing non-linear local variations, as illustrated in the internal cross-sectional view in Fig. 3(b).

To simulate different scenarios of acquiring subsurface information via drilling, three distinct borehole sampling strategies were designed and implemented on the synthetic geological model. In all strategies, boreholes are represented as vertical profiles that penetrate the entire depth of the model.

The first strategy is Uniform Sampling, where boreholes are arranged on a regular grid. This ensures systematic and unbiased coverage of the entire study area, representing an idealized survey scenario, as depicted in Fig. 3(c). The second strategy is Random Sampling, whereby borehole locations are determined by random selection from all possible positions within the study area (see Fig. 3(d)).

The third strategy is Biased Sampling. This strategy mimics realistic exploration scenarios where geological knowledge is highly asymmetric. This method divides the study area into four equal quadrants (top-left, bottom-left, top-right, bottom-right) along the horizontal mid-planes. A different number of boreholes is allocated to each quadrant. Within each quadrant, borehole locations are randomly selected from all available positions (Fig. 3(e)).

3.2. Results and analysis

3.2.1. Performance assessment under the point-wise split strategy

The point-wise split strategy was observed to systematically produce overestimated model performance. This trend is visible across all sampling strategies (Figs. 4–6). This strategy can create a potentially misleading high accuracy. However, the degree of overestimation varied. It depended on the model type and the sampling strategy.

Under uniform sampling (Fig. 4), the susceptibility of different model types became apparent. The Global Function Fitting models (SVM, RBF, MLP) showed high robustness. Their test accuracy was very close to their true accuracy. The differences ranged only from −0.9 % to +0.2 %. showed less performance overestimation. In contrast, the Ensemble Spatial Partitioning models (RF, XGBoost) showed significant overestimation. The difference between test and true accuracy reached +1.6 % for RF. It reached +3.4 % for XGBoost. The Neighborhood-based models also showed overestimation, but it was moderate. The accuracy difference for KNN was +0.8 %. For the more complex AE-ResCapsNet, the difference was only +0.4 %.

The performance overestimation became much more pronounced when the sampling strategy changed. The shift from uniform to random and biased sampling amplified the overestimation (Figs. 5 and 6). As the data distribution became less even, the overestimation increased for almost all models.

The Global Function Fitting models still showed the highest robustness among the three categories. However, their overestimation slightly increased. For example, under random sampling, the differences for SVM and MLP rose to +3.2 % and +2.5 %, respectively. The overestimation in Ensemble Spatial Partitioning models became notably

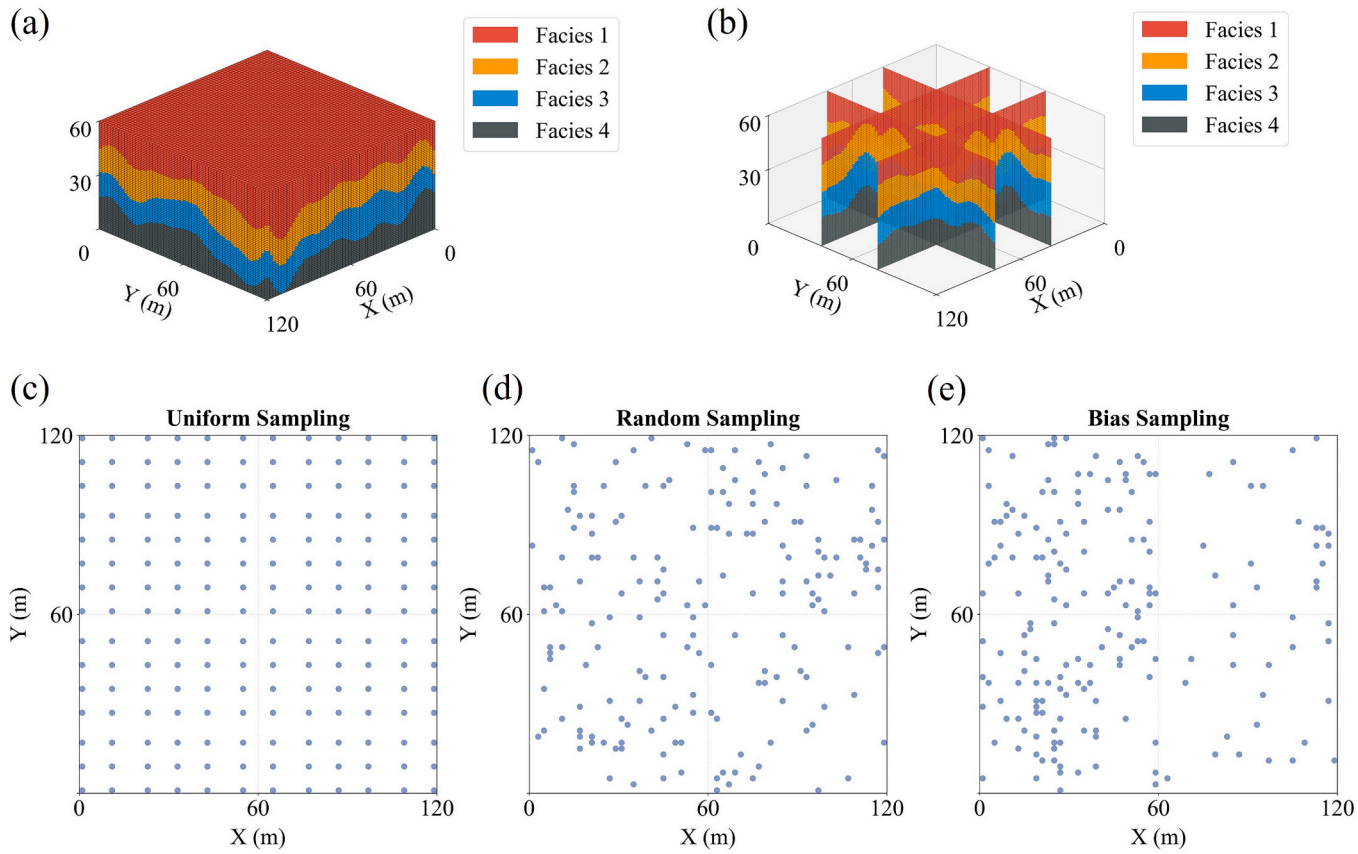

Fig. 3. Overview of the synthetic geological model and borehole sampling strategies. (a) A 3D perspective view of the synthetic ground truth model. (b) Internal cross-sections of the model. (c–e) The spatial distribution of borehole locations under the three sampling strategies: (c) uniform sampling, (d) random sampling, and (e) biased sampling.

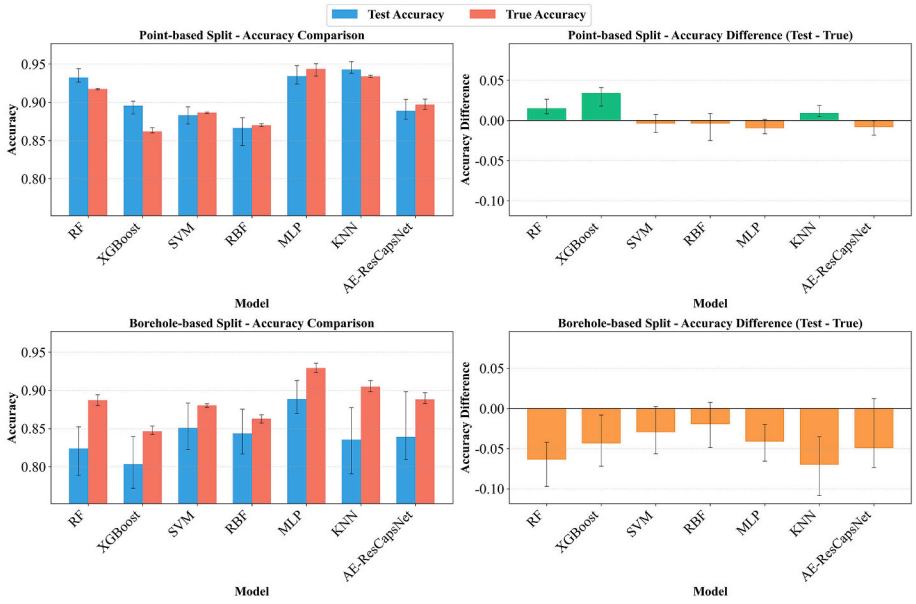

Fig. 4. Performance comparison of all models under uniform sampling.

larger. Under random sampling, the accuracy difference for RF surged to +5.7 %. For XGBoost, it surged to +6.9 %. Under biased sampling, where data are highly clustered, the difference for XGBoost reached its peak for this strategy (+7.1 %). Meanwhile, the Neighborhood-based models also showed larger differences. The accuracy difference for KNN expanded to +2.0 % (random) and +3.5 % (biased). AE-ResCapsNet again showed greater robustness compared to KNN. Its accuracy differences only increased to +1.7 % and +1.9 %.

A visual comparison of the predicted cross-sections (Fig. 7) provides structural insights. It is crucial to assess the quality of model predictions

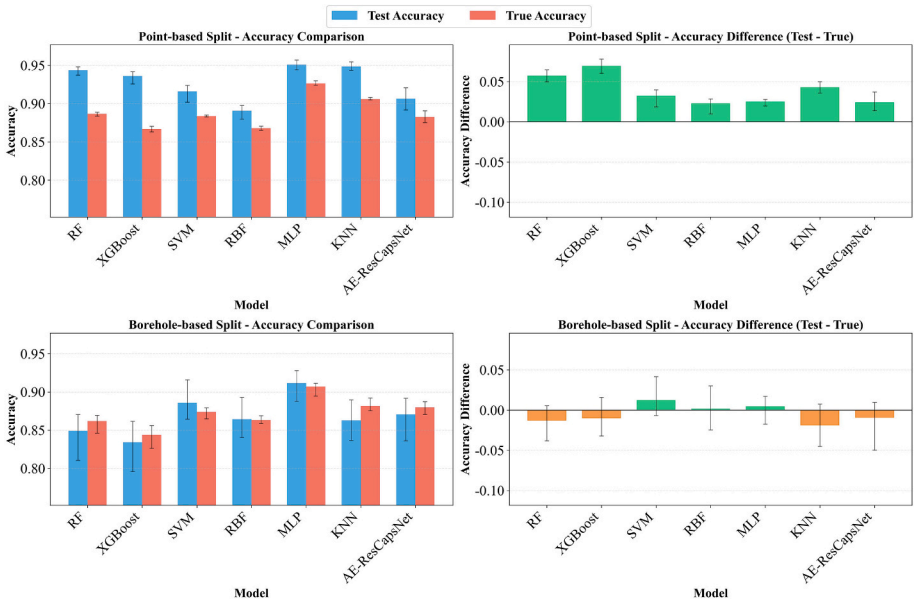

Fig. 5. Performance comparison of all models under random sampling.

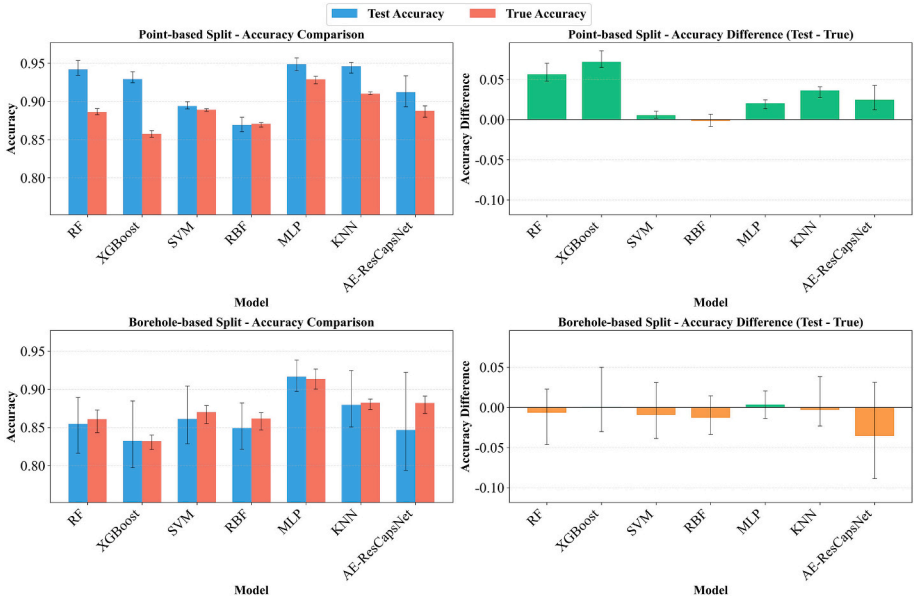

Fig. 6. Performance comparison of all models under biased sampling.

based on their structural similarity to the known ground truth. The ground truth model (Fig. 7(a)) features smooth and continuous geological interfaces. The cross-sections generated by the Global Function Fitting models (especially SVM, RBF, and MLP) closely resembled the ground truth. They produced smooth and continuous interfaces (Fig. 7(d–f)). The shapes generated by these models align well with the structural characteristics of this synthetic case. Their predictions were both numerically accurate and structurally faithful.

In fundamental contrast, the Ensemble Spatial Partitioning models (RF, XGBoost) produced distinctly "blocky" or "stair-step" structures (Fig. 7(b and c)). This morphology resulted in a significant structural mismatch with the continuously varying interfaces of the ground truth. Thus, although these models achieved high classification accuracy at discrete points, the structures they generated were morphologically inconsistent with the ground truth.

Similarly, the Neighborhood-based models produced morphologies that differed from the smooth ground truth. The KNN model's cross-sections (Fig. 7(g)) displayed "patchy" areas formed around training data points. The AE-ResCapsNet model (Fig. 7(h)), while effectively capturing the main structural trends, produced a more fragmented or "salt-and-pepper" texture at the interfaces. This visual result suggests that the inductive biases of these neighborhood-reliant models differ fundamentally from the smooth, continuous nature of this specific synthetic case.

3.2.2. *Performance assessment under the borehole-wise split strategy*

The borehole-wise split provides a more rigorous benchmark for model evaluation. This strategy changes the evaluation task from simple "interpolation" to a more challenging "extrapolation." It ensures complete spatial separation between the training and test sets. The results contrasted sharply with the point-wise split (Figs. 4–6). A central and universal observation is that the true accuracy of all models remained

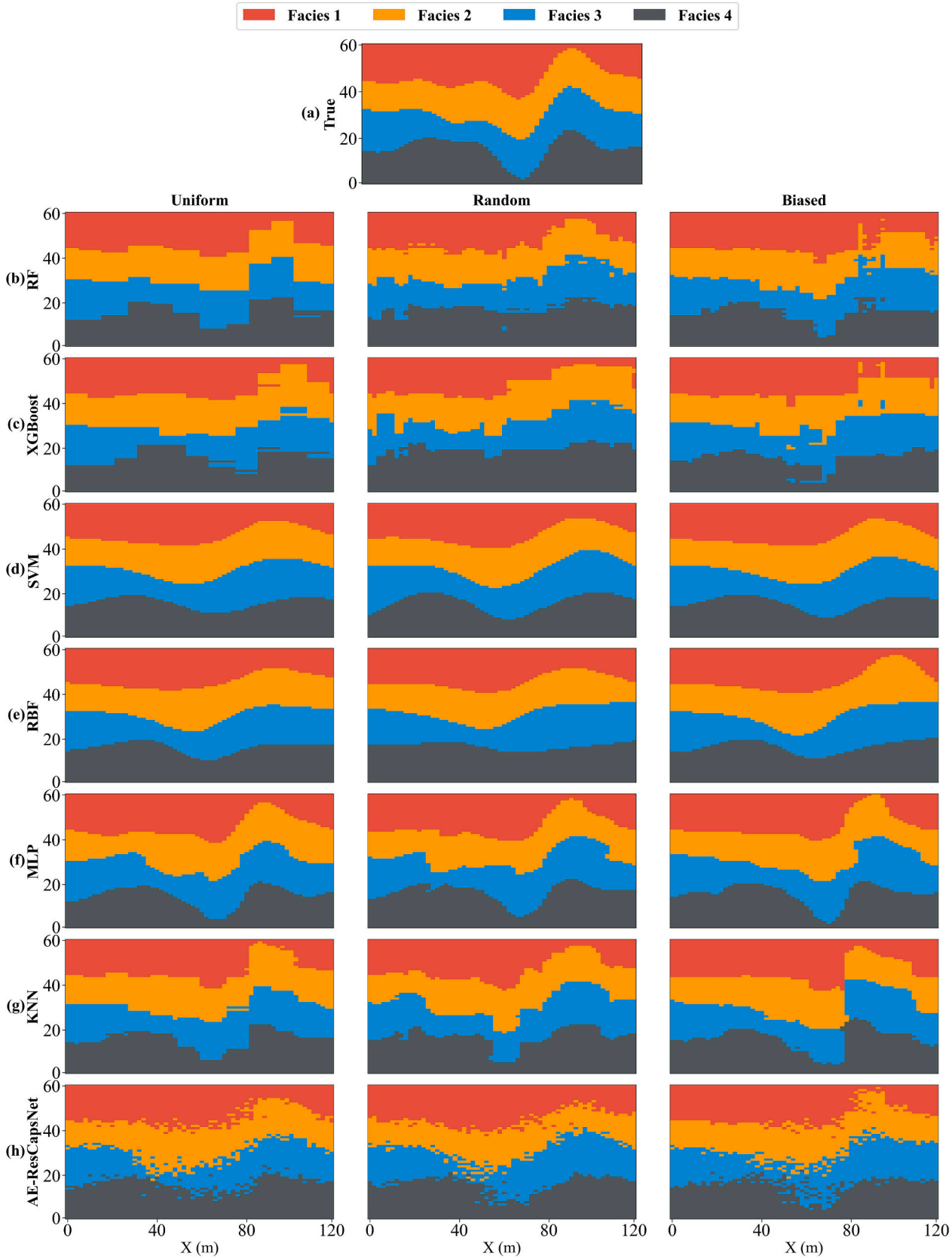

Fig. 7. Predicted geological cross-sections (Y = 39 m) from models trained using the point-wise split.

high. It was similar to the levels achieved under the point-wise split. This indicates the models' overall ability to represent the entire geological space did not change.

However, the test accuracy changed dramatically. Under the borehole-wise split, the test accuracy was universally lower than the corresponding true accuracy. This resulted in a consistent "negative difference" (Test minus True) across all models and scenarios. The magnitude of this negative difference varied with the sampling strategy. Under uniform sampling (Fig. 4), nearly all models exhibited the largest negative difference. When the sampling strategy shifted to random and biased (Figs. 5 and 6), the negative difference generally decreased for all models. Furthermore, analysis across multiple runs revealed that Ensemble Spatial Partitioning models exhibited higher variance across different data splits compared to Global Function Fitting models, indicating their greater sensitivity to data sparsity.

The predicted cross-sections showed notable structural degradation under the rigorous borehole-wise split condition (Fig. 8). A general decrease in structural quality was observed for all models compared to the point-wise split results (Fig. 7). The Global Function Fitting models (SVM, RBF, MLP) could still roughly maintain the macro-scale morphology of the geological interfaces. However, their smoothness and continuity were significantly compromised (see Figs. 8(d), (e), (f)).

The predicted cross-sections of the Ensemble Spatial Partitioning models (RF, XGBoost) and KNN deteriorated further. Their inherent "blocky" structures became more fragmented and disordered (Figs. 8(b), (c), (g)). They almost completely lost their geological continuity. The AE-ResCapsNet model showed a unique behavioral pattern. Visually, it successfully captured the overall macro-scale trends of the strata, exhibiting a behavior similar to that of the Global Function Fitting models (Fig. 8(h)). However, its complex neighborhood-based feature-learning process also produced a fine-scale fragmented texture (or "salt-and-pepper noise"). This visual result contrasts sharply with the smooth and continuous interfaces from Global Function Fitting models (like MLP or SVM). This highlights a key difference in inductive bias: the Global Function models default to smooth interpolation in data-sparse regions, whereas the AE-ResCapsNet's bias, when applied to sparse data, results in this more pixelated local variability. Post-processing techniques such as majority filtering or MRF-based smoothing can be applied to reduce these artifacts and improve structural continuity.

4. The Pearl River Delta case study

4.1. Study area, dataset, and experimental design

Section 3 used a synthetic model. This provided a controlled "laboratory" setting. It revealed basic model behaviors and evaluation biases. Now, the study moves to a real-world case. The goal is to test whether those findings hold true in complex geological settings. A second goal is to find the performance limits of the models in practice. This study introduces the Pearl River Delta (PRD) case.

The PRD region was selected because it poses a significant challenge for the models. First, the PRD is a typical alluvial plain. Its Quaternary sediments were formed by complex interactions between rivers, deltas, and the ocean. This process caused rapid changes in facies, both horizontally (side-to-side) and vertically (up-and-down). The facies relationships are complex. The area lacks the large-scale, continuous layers seen in the synthetic model. This natural geological complexity provides a highly challenging testbed. It helps examine the models' ability to predict new areas (generalization) and maintain structural accuracy.

Second, the distribution of real-world borehole data is naturally uneven and biased. This is unlike the idealized, uniform or random sampling in the synthetic case. As shown in Fig. 9, the 796 engineering geological boreholes collected for this study cover the core plains of the PRD. However, their locations are not evenly spread. Urban development and engineering activities clearly influenced the distribution. The boreholes cluster along transportation lines and in developed urban areas. This uneven and "biased" data distribution directly matches the most challenging "biased sampling" scenario from the synthetic case. This allows for further validation of the model behaviors observed in Section 3 using real-world data.

The modeling domain for this study is shown in Fig. 9. The modeling depth was set from 5 m to 115 m below the ground surface. This covers the main Quaternary deposits. For 3D modeling, the entire domain was divided into a regular grid. The cell resolution was 2 km $\times$ 2 km $\times$ 2 m. This resulted in a total of 1,985,400 grid cells. From the 796 boreholes, 220,170 cells with known facies labels were extracted. This means only 11.1 % of the entire 3D modeling space has known data. This high degree of data sparsity (lack of data) is a common problem faced in the practical application of machine learning for geological modeling.

Geological complexity strongly impacts model performance. To study this impact, two classification tasks were designed. The tasks have increasing difficulty. The first task is a binary classification. It aims to identify the spatial distribution of two major facies classes: sediments and bedrock. This large-scale task involves a relatively clear geological boundary. Structurally, it is similar to the layered synthetic model. This task is designed to act as a "bridge" from the idealized synthetic world to the real world. It tests whether the models still show the basic behaviors seen in the synthetic case. This test utilizes real-world data characteristics, including sampling bias and data sparsity.

The second task is a four-facies classification. It builds upon the first by further dividing the facies. Sediments are classified into two main types: clay and sand. Bedrock is divided into weathered bedrock and unweathered bedrock. This task is much more difficult. The spatial distribution of these four smaller-scale facies is complex. They show high heterogeneity and non-stationarity. This reflects the complex depositional and weathering processes of the PRD. This allows for precise measurement of how much the models' performance ceiling drops when facing real, complex facies relationships. Therefore, the PRD case is not merely an extension from a synthetic to a real-world application. It is an ultimate challenge. It combines geological complexity, uneven sampling, and sparse data. The results from this case will provide critical evidence. They will help achieve a deep understanding of the true potential and limitations of ML in the geosciences.

4.2. Results and analysis

4.2.1. Binary classification

The first task was a binary classification. It involved identifying two large-scale facies units: sediments and bedrock. The results clearly showed the strong impact of the validation strategy.

When using the point-wise split strategy, all models achieved very high test accuracy (Fig. 10, left panel). The Ensemble Spatial Partitioning model (XGBoost, 96.7 %) and the simple Neighborhood-based model (KNN, 96.2 %) achieved the highest scores. However, when switching to the rigorous borehole-wise split strategy, a large drop in accuracy was observed for all models (Fig. 10, right panel). A critical observation is the convergence of model performance. The differences between the models became very small. The average accuracies of all seven models fell within a narrow range of about 1 % (from 89.5 % to 90.6 %). All models reached a performance level of approximately 90 %.

The models with the highest scores under the point-wise split showed the largest drops. XGBoost showed a 7.0 % drop. KNN showed a 6.2 % drop. This trend matches the findings from the synthetic case study. The predicted cross-sections (Fig. 11) provide visual confirmation. A notable finding is that the large-scale shape (morphology) of the bedrock surface looked very similar for both validation strategies. The models generated similar macro-structures under both the point-wise and borehole-wise

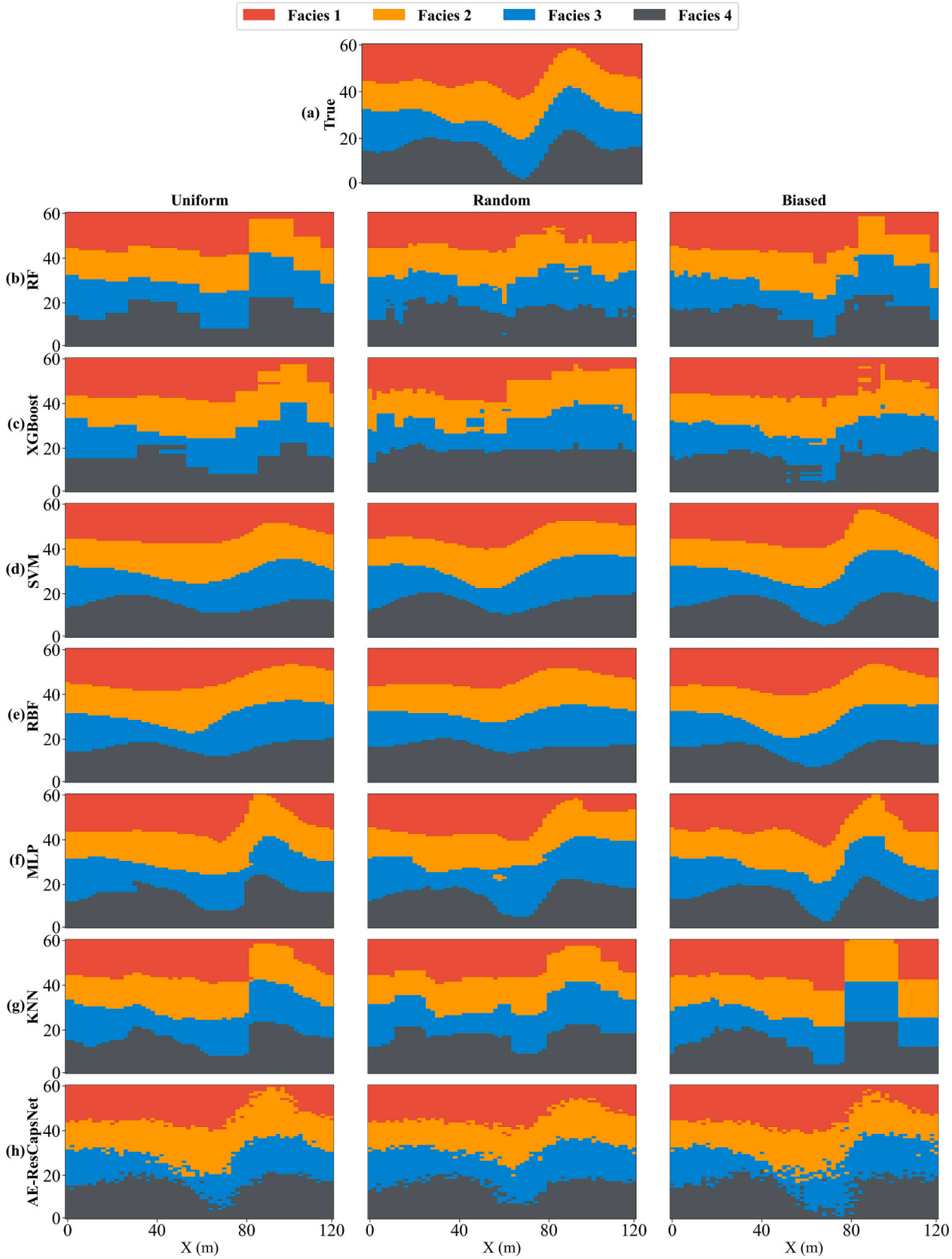

Fig. 8. Predicted geological cross-sections (Y = 39 m) from models trained using the borehole-wise split.

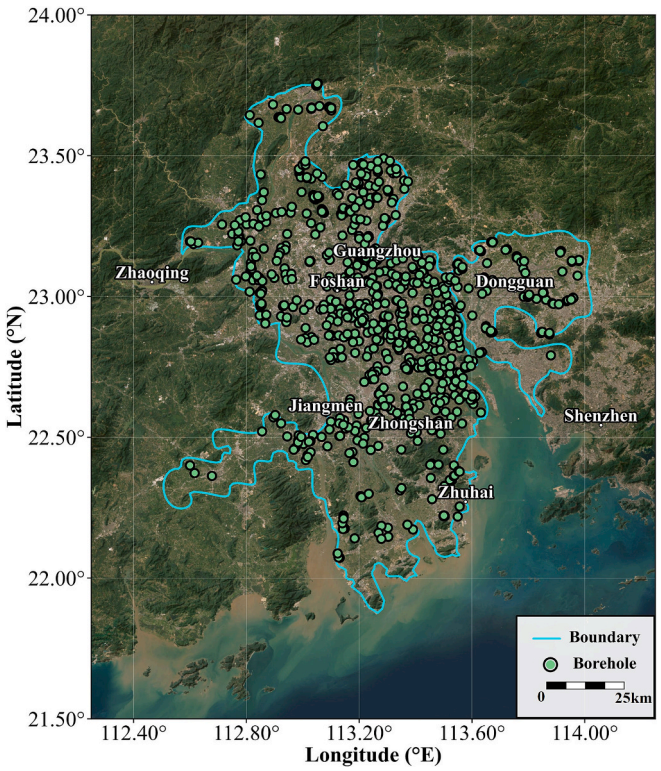

Fig. 9. The geographical location of the study area in the Pearl River Delta (PRD) and the spatial distribution of the 796 boreholes used for modeling. The red line delineates the boundary of the modeling domain. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

splits. This visual similarity contrasts sharply with the large differences in accuracy scores between the two strategies.

Upon closer inspection of the structures, the different tendencies of the models were visible. The Global Function Fitting models (SVM, RBF, MLP) produced smoother interfaces (Figs. 11(c), (d), (e)). The Ensemble Spatial Partitioning models (RF, XGBoost) and Neighborhood-based models (KNN) generated more localized and irregular shapes (Figs. 11 (a), (b), (f)). The AE-ResCapsNet model produced relatively smooth boundaries with some local variations (Fig. 11(g)).

4.2.2. *Four-facies classification*

To evaluate the impact of geological complexity, the task difficulty was increased. The study moved to a four-facies classification. This involved dividing the sediments and bedrock into four smaller-scale facies. This change introduced much more heterogeneity and complexity into the modeling problem.

The quantitative results clearly revealed that geological complexity limited model performance (Fig. 10). Under the point-wise split strategy, the accuracy decreased significantly compared to the binary task. The best-performing model in the four-facies classification (KNN, 89.8 %) was about 7 % lower than the best model in the binary classification (XGBoost, 96.7 %).

When the rigorous borehole-wise split strategy was applied, the impact of geological complexity was stronger. The average accuracy of all models dropped to about 75 %. The models again showed a high degree of convergence. The performance range was only about 2 %. The accuracy drop from the binary task to the four-facies task under the borehole-wise split was significant. For the KNN and RF models, the accuracy dropped by 15.1 % and 13.6 %, respectively. These drops were significantly larger than those observed in the binary task when

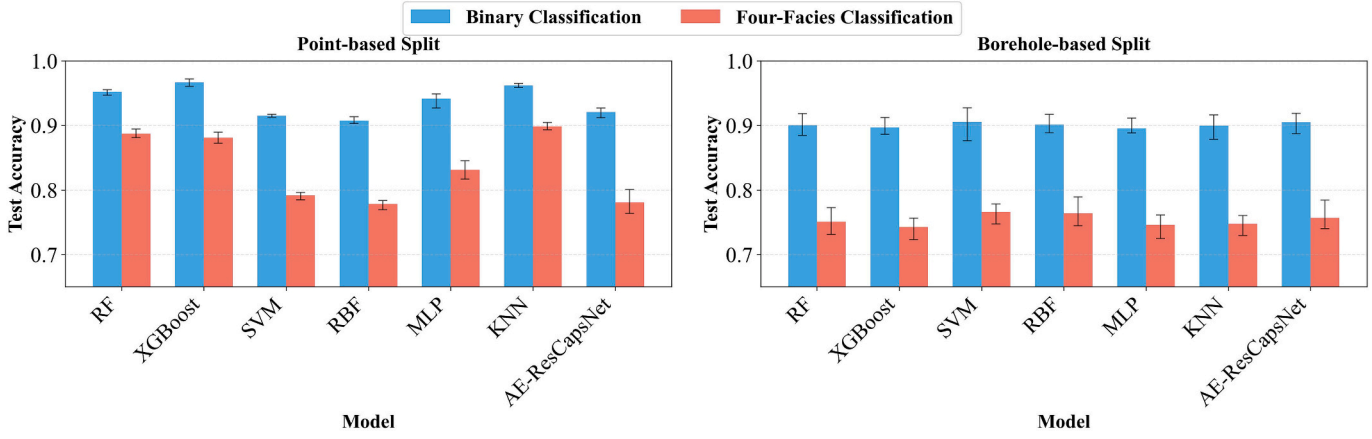

Fig. 10. Comparison of test accuracies for the seven machine learning models in the Pearl River Delta case study.

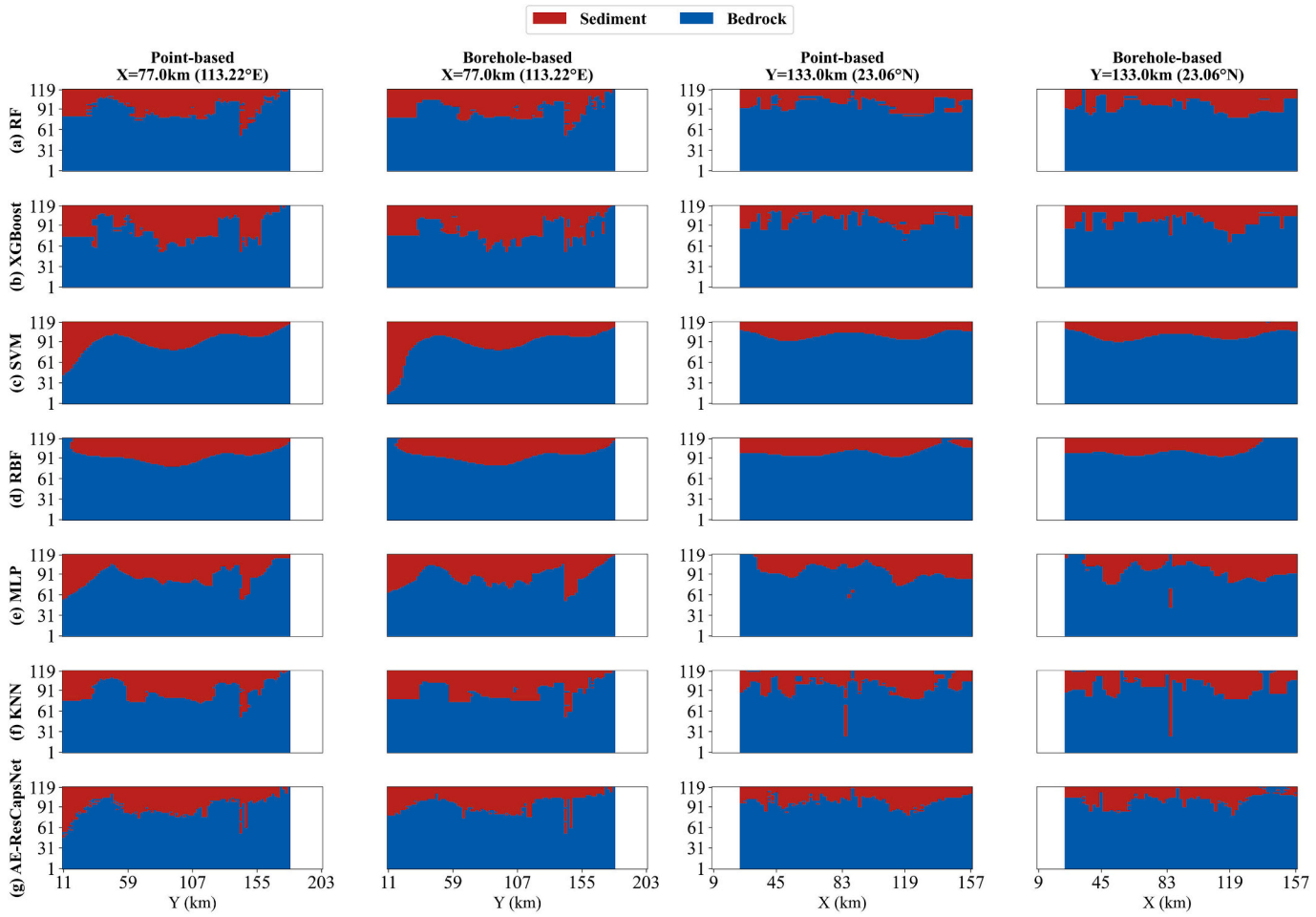

Fig. 11. Predicted geological cross-sections for the binary classification task (sediment vs. bedrock) at two representative locations (X = 77.0 km and Y = 133.0 km).

changing the validation strategy (which were 6.2 % and 5.1 %, respectively).

These large performance drops appeared in the predicted cross-sections (Fig. 12) as complex visual patterns. The behaviors of the different models diverged when facing this challenge. A key observation is that the predicted cross-sections of the Global Function Fitting models (SVM, RBF, MLP) and AE-ResCapsNet showed little structural difference between the point-wise and borehole-wise splits (Fig. 12(c)–(e) and (g)). These models tended to generate relatively large and smooth facies patches. They maintained a form of structural continuity even under the rigorous borehole-wise split.

In contrast, the Ensemble Spatial Partitioning models (RF, XGBoost) and KNN showed a marked difference (Figs. 12(a), (b), (f)). They generated somewhat structured clusters under the point-wise split. However, their predictions degraded into finer and more chaotic "pixelated" or fragmented distributions under the borehole-wise split.

In summary, the visual results showed distinct behaviors under the borehole-wise split. Some models produce smooth, continuous structures. Other models produced highly fragmented or pixelated structures. These significant visual differences occurred despite all models having similar accuracy scores (around 75 %).

5. Discussion

5.1. What does validation strategy really measure?

Results from both the synthetic and real-world cases consistently show that the choice of validation strategy is not merely a technical detail but fundamentally determines which aspects of model performance are being assessed. In this study, R-CV (implemented as a "point-wise split") consistently led to overestimated performance metrics (Kumar et al., 2025). This pattern does not mean that R-CV is intrinsically flawed, but rather reflects a conceptual mismatch: R-CV is designed to evaluate spatial interpolation and does not fully account for spatial autocorrelation (the fact that nearby data points are similar). By design, R-CV measures a model's ability to reproduce patterns between existing data points (Valavi et al., 2018). This is primarily measuring "goodness-of-fit" or "pattern reproduction." In this scenario, high accuracy (e.g., >90 %) is expected. But this high accuracy may say relatively little about the model's ability to predict in new areas.

In contrast, S-CV, implemented as a "borehole-wise split," evaluates true spatial extrapolation. By enforcing spatial separation, S-CV rigorously tests a model's ability to apply learned rules to entirely unseen locations (Sun et al., 2023; Wadoux et al., 2021).

The synthetic case study clearly demonstrated how validation interacts with data sampling. Under R-CV, the degree of performance overestimation was not constant. It was directly amplified by spatial bias in the sampling data. This inflation was most severe in the "biased" scenario (as seen in Figs. 4–6). The reason is clear. Biased sampling creates dense clusters of data in some regions, leaving others empty. R-CV randomly selects individual points. This creates a test set where most points are also from those dense clusters. These test points are very close to training points. The model is therefore mainly tested on easy predictions near known data. Its potential underperformance in the data-sparse regions may be obscured. This results in optimistic scores that may not reflect true predictive power. Our results suggest that the

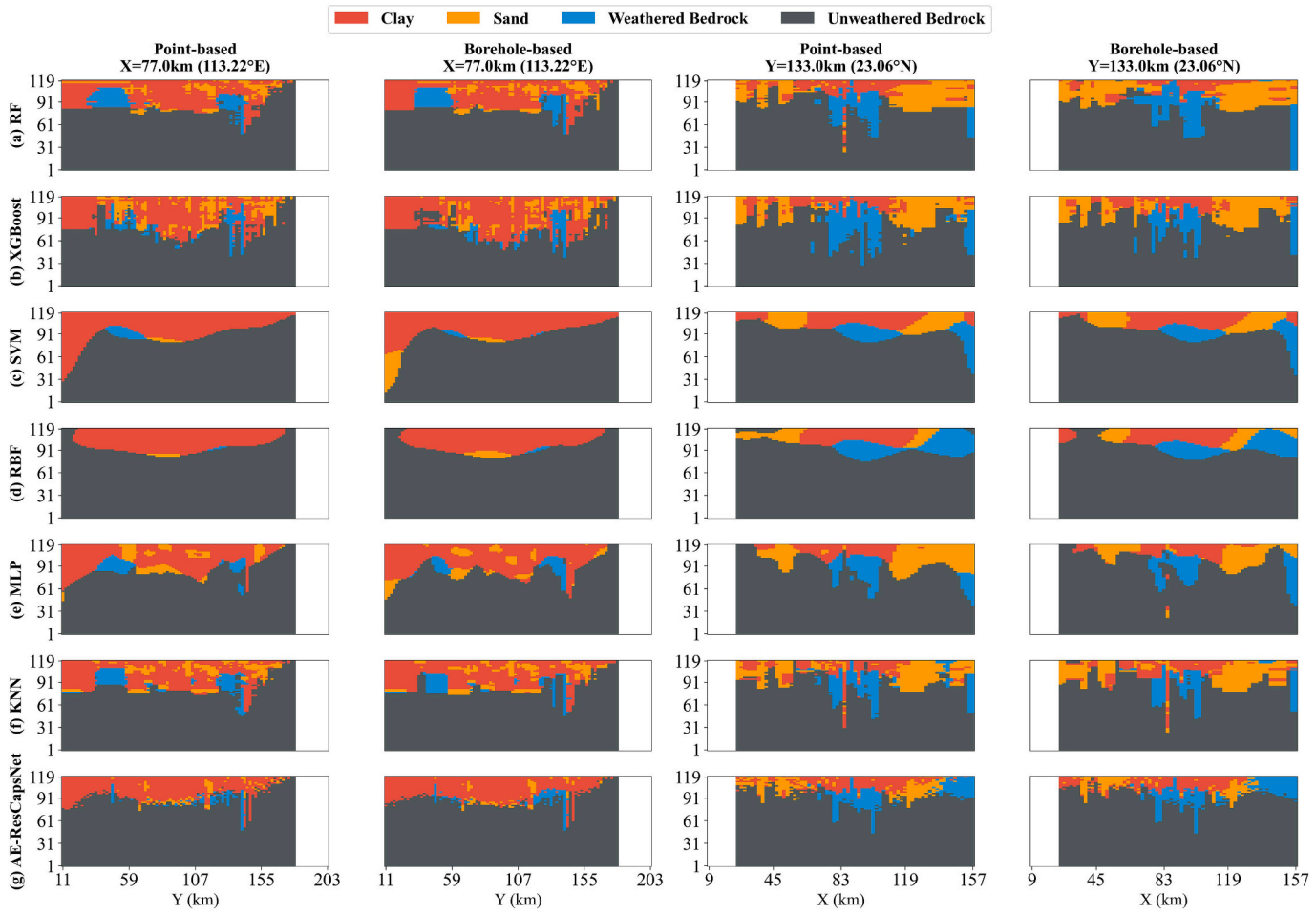

Fig. 12. Predicted geological cross-sections for the four-facies classification task (clay, sand, weathered bedrock, and unweathered bedrock) at two representative locations (X = 77.0 km and Y = 133.0 km).

magnitude of R-CV overestimation depends more on clustering geometry than on overall spatial density. Comparing the overestimation patterns across sampling strategies (Figs. 4–6), the transition from uniform to random sampling changed spatial density patterns but produced only moderate increases in overestimation. However, the transition to biased sampling introduced strong spatial clustering and resulted in the most severe overestimation. Conversely, S-CV demonstrated its robustness. Under S-CV, none of the models showed severe performance overestimation, regardless of the sampling strategy. This confirms that S-CV provides a stable, conservative benchmark (Pohjankukka et al., 2017; Wang et al., 2023).

From an engineering perspective, these differences in validation strategy are highly consequential. Real-world borehole data often display uneven spatial distributions, with sampling concentrated in specific areas such as infrastructure corridors. Under such conditions, R-CV may yield relatively high accuracy scores that are overly optimistic for regions with limited data. This optimism could influence decision-making in underground construction or hydrogeological assessments. In contrast, S-CV tends to provide a more conservative estimate, helping practitioners better evaluate reliability and manage uncertainty in data-sparse areas.

5.2. Model selection based on inductive bias

A key finding of this study is the performance convergence under rigorous S-CV. All model types tested converged to a similar "performance ceiling" in the PRD case (approximately 90 % for the binary task and approximately 75 % for the four-facies task). This convergence indicates that S-CV accuracy has limited utility as a primary criterion for model selection. If all algorithms, from ML to DL models, produce similar S-CV scores, then the "model race" to achieve marginally higher S-CV accuracy may be counterproductive.

This finding shifts the evaluation focus from "which model is most accurate?" to "which model's bias is most appropriate?". The experiments demonstrate that a model's "inductive bias" significantly influences the shape (morphology) of the predicted geological structures. Inductive bias refers to the set of built-in assumptions inherent in the algorithm.

First, the synthetic case itself favored certain models. It was generated using smooth sinusoidal functions. This inherently favored the Global Function Fitting models (e.g., SVM, RBF, MLP) (Vaferi et al., 2011). Their tendency to learn smooth boundaries aligned perfectly with this specific ground truth (Bigdeli et al., 2024). This results in structures that are both numerically accurate and morphologically faithful.

In contrast, the intrinsic mechanism of Ensemble Spatial Partitioning models (e.g., RF, XGBoost) involves repeatedly dividing the space into rectangular regions (Saljoughi and Hezarkhani, 2018). This leads to predicted results showing "blocky" or "stair-step" shapes. In the context of the smooth synthetic model, this "blocky" output presented a significant structural mismatch.

However, this highlights a deeper insight. This apparent distortion is not a "failure" of the ensemble models. Instead, it is a fundamental "mismatch" between their inductive bias and the specific nature of this test case. If the ground truth were a fault-block system or a sharp, channelized river system, the "blocky" structure of RF/XGBoost might

provide a more realistic result. In that case, the "over-smoothing" from an MLP or SVM would become the source of structural distortion.

The PRD deltaic aquifer has more complex and variable boundaries than the synthetic model. In the four-facies task, RF and XGBoost achieved slightly higher accuracies than the Global Function Fitting models, suggesting that their inductive bias better matched the variable geology of the PRD. Future synthetic benchmarks should therefore include diverse geological architectures, such as channelized or faulted geometries, to provide broader guidance for model selection. In turn, Global Function Fitting models may be more suitable for smaller-scale stratigraphic structures with regular sedimentary patterns, where interfaces between lithologies are distinct and transitions are relatively smooth.

Furthermore, ML methods' strength primarily lies in delineating geological bodies with relatively clear spatial boundaries, as demonstrated in the synthetic case and the PRD binary task. However, they may struggle to effectively identify facies that change frequently in the vertical direction, a common feature in complex sedimentary sequences (Bayer et al., 2015; Dai et al., 2019). In such cases, traditional stochastic methods, which are explicitly designed to reproduce statistical properties of complex spatial arrangements, may offer complementary strengths (Carle and Fogg, 1997; Zhan et al., 2022a). This suggests that a hybrid approach or using "direct prediction" models to generate high-quality Training Images (TIs) for generative models (e.g., GANs) represents a promising workflow for addressing data shortages and capturing complex heterogeneity (Singh et al., 2022; Sun, 2018; Zhan et al., 2022b; Zheng et al., 2023).

Therefore, in practical applications S-CV accuracy should be treated as a threshold metric, used to verify that models have reached the performance ceiling imposed by data availability and geological complexity. Once this threshold has been reached by multiple algorithms, model selection should be guided primarily by how well their inductive biases align with existing geological understanding of the target subsurface structure.

The convergence of model performance in the PRD case does not imply that data volume dominates model selection. In fact, the PRD case has more data (approximately 10 % of total grid cells) than the synthetic case (approximately 5 %). The lower accuracy in the PRD case is due to stronger spatial heterogeneity of facies, indicating that geological complexity, rather than data volume, is the primary factor affecting prediction accuracy under S-CV.

5.3. Limitations and future directions

Several limitations of this study should be acknowledged. First, while the seven ML models tested represent major types of algorithms, they do not make a complete list. Recent years have seen the development of numerous advanced models. These include improvements on these basic algorithms and novel architectures for geological modeling based on Graph Neural Networks (GNNs) or Transformers (Hillier et al., 2021; Wang et al., 2025b). However, many of these advanced methods share underlying principles with the models evaluated in this study. Therefore, it is likely that the core conclusions regarding the importance of validation strategy and data limitations can be generalized.

Second, and more critically, the root causes of the "performance ceiling" must be analyzed. The conclusion is that it stems from two intertwined limitations: "data sparsity" and "feature poverty." "Feature poverty" refers to the models' reliance on only sparse borehole data as input. This limitation effectively constrains all the tested models to function primarily as spatial interpolators. They learn the spatial proximity of data (e.g., "points near A are also A"). Meanwhile, "data sparsity" (i.e., the limited number and uneven distribution of boreholes) means the models lack sufficient data support when predicting in large unknown areas.

Therefore, the performance ceiling identified in the complex PRD case (approximately 75 %) is best interpreted as a practical limit imposed jointly by sparse and unevenly distributed borehole data, geological complexity, and the interpolation nature of the task, rather than as a failure of the ML algorithms themselves.

Several methodological considerations may help address the limitations discussed above. For model evaluation, this study relied on visual comparison to assess structural differences among models. While this approach effectively demonstrates inductive bias, quantitative morphology metrics such as boundary roughness indices and connectivity functions could provide more objective comparisons in future studies. For input data, sample reweighting strategies may help reduce the influence of clustered boreholes without discarding valuable observations.

In practical geological modeling, integrating multi-source information (such as geophysical data and hydrogeological observations) is not just an optional improvement (Guo et al., 2025). It is the main and necessary way to fundamentally overcome the "feature poverty" limitation. Geophysical data can provide continuous spatial characteristics of subsurface structures that sparse boreholes cannot capture. By integrating such multi-source data, models can learn the physical correlations between different data types and facies distributions. In our previous work on a synthetic binary aquifer structure, stochastic simulation using only borehole data achieved approximately 80 % accuracy, while ERT-based inversion achieved approximately 95 % accuracy (Xia et al., 2026). This enables models to capture more continuous and complex spatial patterns of geological structures, rather than merely relying on the proximity of sparse point measurements. This approach represents the primary pathway to break through the current performance ceiling and achieve more accurate and reliable subsurface structure modeling.

6. Conclusions

Through controlled experiments in both synthetic and real-world geological settings, this study systematically examined the impact of different evaluation strategies on machine learning-based facies modeling. A core finding of this study is that applying standard R-CV may lead to an overestimation of prediction accuracy. This overestimation is often associated with the use of R-CV. This strategy mixes spatially close data points into both training and test sets. As a result, it may not rigorously test a model's ability to predict at new, unseen locations. In contrast, S-CV enforces spatial separation, resulting in a lower accuracy but providing a more realistic measure of true predictive performance. Therefore, S-CV is demonstrated to be a more rigorous and realistic validation method for subsurface structure modeling.

In the real-world PRD case, this study found that all tested models converged to a similar 'performance ceiling'. This finding, resulting from the high geological complexity and sparse data, suggests that small differences in S-CV accuracy are not a reliable basis for model selection. This shifts the evaluation focus: instead of asking "which model is most accurate?", the question becomes which model's predicted geological structures best match prior geological knowledge? Based solely on sparse borehole data, these ML models are more suitable for delineating geological bodies with distinct boundaries. They are less effective for complex zones where different facies are frequently layered together.

This performance ceiling is not a failure of the ML algorithms. Instead, it is the practical limit of the task itself, set by data sparsity, geological complexity, and "feature poverty". The primary path to overcome this limit is to integrate multi-source data, such as geophysical data. Sparse borehole data alone may lack continuous spatial characteristics of subsurface structures. Geophysical data can provide these continuous features, enabling models to learn more continuous and complex spatial distributions of facies. This feature-enriched approach, guided by the S-CV and 'inductive bias' workflow, may also help generate better TIs. This could help address the critical data shortage for generative models, such as GANs and VAEs.

CRediT authorship contribution statement

Chuanjun Zhan: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Funding acquisition, Conceptualization. **Jiu Jimmy Jiao:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Yue Ma:** Writing – review & editing, Validation. **Hao Wang:** Writing – review & editing, Validation. **Mohamad Reza Soltanian:** Writing – review & editing, Methodology, Conceptualization. **Kenneth C. Carroll:** Writing – review & editing, Visualization, Conceptualization. **Zhenxue Dai:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work is funded by the National Key R&D Program of China (2024YFC3713800), Guangdong-Hong Kong Joint Laboratory for Soil and Groundwater Pollution Control (No. 2023B1212120001), the National Natural Science Foundation of China (NSFC: 42502245), the Research Grants Council of the Hong Kong SAR Government (AoE/E-603/18), Shandong Key Water Conservancy Science and Technology Project (2024370203001957), and Shandong Provincial Natural Science Foundation (ZR2024QD095). The U.S. co-authors (M.R. Soltanian and K.C. Carroll) received no financial support from these projects.

Appendix A. Detailed machine learning model descriptions

This appendix provides detailed descriptions of the seven machine learning models used in this study. The models are grouped into three categories based on their 'inductive bias', which is the model's inherent tendency to learn patterns in a specific way.

A.1. Neighborhood-based models

Neighborhood-Based Models learn by storing training samples. They make predictions based on the idea that spatially close points have similar properties. Their complexity grows with the amount of training data. Their predictions depend entirely on the local distribution and labels of training samples near a query point.

KNN is a non-parametric, instance-based algorithm. For a given query point x_q, the algorithm first identifies the k_{nn} nearest neighboring points from the training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ to form a neighborhood set $N_{k_{nn}}(x_q)$. This is based on a predefined distance metric, where k_{nn} represents the number of neighbors to use. The distance is typically the Minkowski distance, where the exponent p determines the distance metric:

$$d(x_i, x_q) = \left(\sum_{j=1}^{3} |x_{i,j} - x_{q,j}|^p \right)^{1/p} \tag{A.1}$$

here, $p = 1$ corresponds to the Manhattan distance, and $p = 2$ corresponds to the Euclidean distance. The predicted class $\widehat{y_q}$ is determined by a vote among the samples in the neighborhood set $N_{k_{nn}}(\boldsymbol{x_q})$, with the voting method controlled by the weights scheme w_i:

$$\widehat{y_q} = argmax \sum_{x_i \in N_k(x_q)} w_i \cdot I(y_i = c) \tag{A.2}$$

where $I(\cdot)$ is the indicator function, and c represents all possible facies classes. If weights = 'uniform', all neighbors have equal weight ($w_i = 1$). If weights = 'distance', the weight is inversely proportional to the distance ($w_i = 1/d(x_i, x_q)$), giving closer neighbors greater influence.

AE-ResCapsNet's predictive process involves three main stages: data self-organization, feature extraction, and facies prediction.

First, a data self-organization method is employed. For a target unit with unknown facies, its functional characteristics are internalized by coupling it with $N_{sample} = 2048$ units randomly selected from the known borehole sampling data.

Second, this process generates a feature matrix $F_q \in R^{K \times 4}$ for the query point $\boldsymbol{x_q}$. A key implementation detail is the use of relative coordinates to enhance the model's focus on spatial relationships. Each row i of the matrix is defined as the vector $f_{q,i} = (\Delta x_i, \Delta y_i, \Delta z_i, l_i')$, $where(\Delta x_k, \Delta y_k, \Delta z_k)$ are the relative coordinates between the i-th sample point x_i and the query point x_q, and l_k' is the normalized facies label of the sample. This representation explicitly encodes the displacement vectors to the surrounding samples, providing a more direct input for learning spatial patterns.

Third, a convolutional Autoencoder performs unsupervised feature extraction on this matrix. The encoder, $g_\phi(\cdot)$, maps the input matrix to a compressed latent representation $z_q = g_\phi(F_q)$. The autoencoder is trained by minimizing the mean squared reconstruction error between the original input and the output of the decoder $f_\theta(\cdot)$:

$$L_{AE}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} \left(F_q^{(i)} - f_\theta\left(g_\phi\left(F_q^{(i)}\right)\right) \right)^2 \tag{A.3}$$

Finally, a Residual Capsule Network (ResCapsNet) performs supervised classification using the extracted deep feature map z_q as input. The network is trained by minimizing the Margin Loss, which is standard for capsule networks and designed to penalize incorrect classifications based on the length of the output capsule vectors. The loss for each class capsule j is defined as:

$$L_j = T_j max(0, m^+ - ||v_j||)^2 + \lambda_{cap}(1 - T_j) max(0, ||v_j|| - m^-)^2 \tag{A.4}$$

here, T_j is an indicator function (1 if the true class is j, 0 otherwise). $||v_j||$ is the length of the output vector for capsule j, representing the probability of class j. The hyperparameters m^+ and m^- are the upper and lower margins (typically 0.9 and 0.1, respectively), and λ_{cap} is a down-weighting factor (e.

g., 0.5) for the loss associated with absent classes. This loss function encourages the vector length for the correct class to be above m^+ while pushing the lengths for incorrect classes below m^-. The training process is governed by key hyperparameters: the learning rate η_{lr}, batch size B_s, and dropout probability p_{drop}.

A.2. Global function fitting models

This category of models, along with Ensemble Spatial Partitioning Models, falls under the umbrella of Model-Based Learning. This approach induces a general model or function from training data that captures underlying patterns. Once trained, the model makes predictions without referencing the original data. Global Function Fitting Models specifically learn a single, continuous decision function $f(x)$ that covers the entire coordinate space. The inductive bias of these models favors the generation of smooth decision surfaces.

SVM: The core objective of SVM is to find one or more optimal hyperplanes in a high-dimensional feature space that maximize the margin between different facies classes. For non-linearly separable data, the "kernel trick" is used. The dual optimization problem for a soft-margin SVM is formulated as:

$$\max_{\alpha}\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j)\right) \tag{A.5}$$

subject to the constraints $\sum_{i=1}^{N}\alpha_i y_i=0$ and $0\le\alpha_i\le C$. Here, $y_i\in\{-1,1\}$ are the class labels and α_i are the Lagrange multipliers. The regularization parameter C balances the trade-off between maximizing the margin and minimizing training error. A smaller C results in a wider margin (higher regularization), while a larger C aims to classify all training samples correctly (lower regularization). The type of kernel is specified by the K_{type} (kernel) parameter. This study employs the Radial Basis Function (RBF) kernel, defined as:

$$K(x_i,x_j)=\exp\left(-\gamma_{sum}|x_i-x_j|^2\right) \tag{A.6}$$

The kernel coefficient γ_{sum} defines the influence of a single training example. A smaller γ implies a larger influence radius and a smoother decision boundary, whereas a larger γ_{sum} results in a smaller influence radius and a more complex boundary. For a new query point x_q, the predicted class is determined by the decision function:

$$\widehat{y_q}=sign\left(\sum_{i=1}^{N}\alpha_i y_i K(x_i,x_q)+b\right) \tag{A.7}$$

where the summation is performed only over the support vectors (i.e., training samples with $\alpha_i>0$).

RBF: An RBF network is a three-layer feedforward neural network whose hidden layer consists of RBF neurons, each forming a localized response region around a "center". The network's output is a weighted sum of the activations of these hidden neurons. For a query point x_q, the predicted value is given by:

$$f(x_q)=\sum_{j=1}^{N_c}w_j\phi_j(x_q)+b \tag{A.8}$$

here, N_c is the number of neurons in the hidden layer, w_j is the weight from the j-th hidden neuron to the output layer, and b is the bias term. The activation function of the j-th hidden neuron, $\phi_j(x_q)$, is typically a Gaussian function:

$$\phi_j(x_q)=\exp\left(-\gamma|x_q-c_j|^2\right) \tag{A.9}$$

where c_j is the center vector of the j-th neuron. The kernel width parameter γ controls the width of the Gaussian function; a smaller γ corresponds to a wider basis function and a smoother overall function approximation. The training process typically involves first determining the centers c_j via unsupervised K-means clustering, then solving for the weights w_k through supervised learning.

MLP: The MLP is a feedforward artificial neural network composed of multiple fully connected layers of neurons. For an input vector x_q, the output vector $h^{(l)}$ of the l-th hidden layer is calculated as:

$$h^{(l)}=g\left(W^{(l)}h^{(l-1)}+b^{(l)}\right) \tag{A.10}$$

where $h^{(0)}=x_q$, $W^{(l)}$ and $b^{(l)}$ are the weight matrix and bias vector of the l-th layer, and $g(\cdot)$ is a non-linear activation function (e.g., ReLU) specified by the A_{func} parameter. The network's architecture is defined by the structure of its hidden layers H_{struct}. For a multi-class task with N_{class} classes, the output layer uses a Softmax function to convert the output values into a probability distribution:

$$P(y=c|x_q)=\text{softmax}(z)_c=\frac{\exp(z_c)}{\sum_{j=1}^{C}\exp(z_j)} \tag{A.11}$$

where z is the input vector to the output layer. The model is trained using backpropagation to minimize the Cross-Entropy Loss, with an added L2 regularization term:

$$L(W,b) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{N_{class}} y_{ic}\log(p_{ic}) + \frac{\alpha}{2}\sum_{l}|W^{(l)}|_F^2 \tag{A.12}$$

here, y_{ic} is the true label and p_{ic} is the predicted probability. The L_2 regularization strength α controls the penalty on large weights. The learning rate η_{lr} is a hyperparameter of the optimizer (e.g., Adam).

A.3. Ensemble spatial partitioning models

This category of models also falls under Model-Based Learning, but it constructs a predictive model by aggregating numerous "weak learners" (typically decision trees). Each tree learns local rules by recursively partitioning the feature space with axis-parallel splits. The final model's inductive bias produces "step-like" or "blocky" decision boundaries composed of hyper-rectangles (cuboids in 3D space).

RF: RF is an ensemble algorithm that constructs a multitude of decision trees and aggregates their votes to make a prediction. The model consists of B decision trees $\{T_b(x)\}_{b=1}^{B}$, where B represents the number of trees in the forest. Each tree is trained on a bootstrap sample of the original dataset, and at each node, the best split is chosen from a random subset of features. The growth of each tree is controlled by hyperparameters such as the maximum depth D_{max}, the minimum number of samples for a split S_{min_split}, and the minimum number of samples per leaf L_{min_leaf}, which act as pre-pruning mechanisms to prevent overfitting. For a query point x_q, the final prediction is the mode of the predictions from all trees, determined by majority voting:

$$\widehat{y_q} = mode\{C_1(x_q), C_2(x_q), \cdots, C_B(x_q)\} \tag{A.13}$$

XGBoost: XGBoost is a gradient boosting framework that builds decision trees sequentially, with each new tree aiming to correct the residual errors of the preceding ensemble. The prediction model is an additive one, composed of the sum of predictions from M trees (where M is the number of estimators). The prediction for sample x_i after the t-th iteration is:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta \cdot f_t(x_i) \tag{A.14}$$

where $f_t(x_i)$ is the prediction of the new tree and η_{lr} is the learning rate, which scales the contribution of each tree. At each iteration t, the algorithm minimizes a regularized objective function:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \widehat{y_i^{(t-1)}} + f_t(x_i)\right) + \Omega(f_t) \tag{A.15}$$

here, $l(\cdot,\cdot)$ is the loss function, and $\Omega(f_t)$ is a regularization term that penalizes tree complexity. The regularization term is defined as:

$$\Omega(f_t) = \tau N_{leaf} + \frac{1}{2}\lambda_{L2}\sum_{j=1}^{N_{leaf}} w_j^2 \tag{A.16}$$

In this term, N_{leaf} is the number of leaf nodes, and w_j is the weight of the j-th leaf. The minimum loss reduction required to make a split is denoted by τ, acting as a pruning parameter. The L_2 regularization coefficient on the leaf weights is λ_{L2}, which helps prevent overfitting. Other key hyperparameters include the maximum tree depth D_{max}, the subsample ratio ρ_{sample} and the column subsample ratio $\rho_{feature}$, which introduce randomness to prevent overfitting. In this implementation, τ and λ_{L2} use their default values ($\tau = 0$, $\lambda_{L2} = 1$), while the hyperparameter optimization focuses on M, η, D_{max}, ρ_{sample} , and $\rho_{feature}$.

This focused tuning strategy is a pragmatic approach rooted in balancing computational cost with performance impact. The parameters M, η, and D_{max} are widely considered to have the most substantial influence on the model's performance, as they directly control the fundamental aspects of the boosting process and the structural complexity of the individual trees. The ρ_{sample} and $\rho_{feature}$ parameters are also critical for managing overfitting by introducing stochasticity. In contrast, τ and λ_{L2} serve as finer-grained regularization controls. While important, their default values often provide a reasonable baseline of regularization. By fixing these two parameters, the computationally expensive hyperparameter search can concentrate its resources on the more impactful parameters, allowing for a more thorough exploration of their optimal combination. This hierarchical approach to tuning—first optimizing for core structure and boosting behavior, then potentially fine-tuning regularization—is an efficient and effective method for achieving a high-performing model without incurring the prohibitive cost of an exhaustive search across the entire hyperparameter space.

Data availability

Data will be made available on request.

References

Abbas, M.A., Al-Mudhafar, W.J., Wood, D.A., 2023. Improving permeability prediction in carbonate reservoirs through gradient boosting hyperparameter tuning. Earth Sci. Inf. 16 (4), 3417–3432. https://doi.org/10.1007/s12145-023-01099-0.

Agbotui, P.Y., Firouzbehi, F., Medici, G., 2025. Review of effective porosity in sandstone aquifers: insights for representation of contaminant transport. Sustainability 17 (14), 6469. https://doi.org/10.3390/su17146469.

Awawdeh, A.R.M., Yasarer, H., Ghaffari, Z., Yarbrough, L.D., 2025. Downscaling GRACE data for improved groundwater forecasting using artificial neural networks. Civ. Eng. J. 11 (2), 406–419. https://doi.org/10.28991/CEJ-2025-011-02-01.

Bayer, P., Comunian, A., Höyng, D., Mariethoz, G., 2015. High resolution multi-facies realizations of sedimentary reservoir and aquifer analogs. Sci. Data 2 (1), 150033. https://doi.org/10.1038/sdata.2015.33.

Bianchi, M., Pedretti, D., 2018. An entrogram-based approach to describe spatial heterogeneity with applications to solute transport in porous media. Water Resour. Res. 54 (7), 4432–4448. https://doi.org/10.1029/2018WR022827.

Bianchi, M., Zheng, C., 2016. A lithofacies approach for modeling non-Fickian solute transport in a heterogeneous alluvial aquifer. Water Resour. Res. 52 (1), 552–565. https://doi.org/10.1002/2015WR018186.

Bigdeli, A., Maghsoudi, A., Ghezelbash, R., 2024. A comparative study of the XGBoost ensemble learning and multilayer perceptron in mineral prospectivity modeling: a case study of the Torud-Chahshirin belt, NE Iran. Earth Sci. Inf. 17 (1), 483–499. https://doi.org/10.1007/s12145-023-01184-4.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Carle, S.F., Fogg, G.E., 1996. Transition probability-based indicator geostatistics. Math. Geol. 28 (4), 453–476. https://doi.org/10.1007/BF02083656.

Carle, S.F., Fogg, G.E., 1997. Modeling spatial variability with one and multidimensional continuous-lag Markov chains. Math. Geol. 29 (7), 891–918. https://doi.org/10.1023/A:1022303706942.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., Lopez, A., 2020. A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118.

Chen, Q., Zhou, R., Chen, D., Cui, Z., Ma, X., Liu, G., 2026. A conditional masked autoencoder network based on efficient multiple-head self-attention for characterizing heterogeneous reservoirs. Expert Syst. Appl. 296, 128973. https://doi.org/10.1016/j.eswa.2025.128973.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., 2015. Xgboost: extreme gradient boosting. R Package Version 0.4-2 1 (4), 1–4.

Costa, V.G., Pedreira, C.E., 2023. Recent advances in decision trees: an updated survey. Artif. Intell. Rev. 56 (5), 4765–4800. https://doi.org/10.1007/s10462-022-10275-5.

Cui, Z., Chen, Q., Luo, J., Ma, X., Liu, G., 2024. Characterizing subsurface structures from hard and soft data with multiple-condition fusion neural network. Water Resour. Res. 60 (11), e2024WR038170. https://doi.org/10.1029/2024WR038170.

Dai, Z., Zhan, C., Soltanian, M.R., Ritzi, R.W., Zhang, X., 2019. Identifying spatial correlation structure of multimodal permeability in hierarchical media with Markov chain approach. J. Hydrol. 568, 703–715. https://doi.org/10.1016/j.jhydrol.2018.11.032.

Di Federico, G., Durlofsky, L.J. (2025). 3D latent diffusion models for parameterizing and history matching multiscenario facies systems. *arXiv Preprint arXiv:2508.16621*.

Ershadnia, R., Wallace, C.D., Hosseini, S.A., Dai, Z., Soltanian, M.R., 2021. Capillary heterogeneity linked to methane lateral migration in shallow unconfined aquifers. Geophys. Res. Lett. 48 (23), e2021GL095685. https://doi.org/10.1029/2021GL095685.

Galiyev, S., Galiyev, D., Uteshov, Y., Shabelbnikov, Y., Makhonin, V., Maldynova, A., Tekenova, A., Axanaliyev, N., 2025. Enhancing energy and operational efficiency of geotechnological complexes using geoinformation technologies. Emerg. Sci. J. 9 (3), 1372–1387. https://doi.org/10.28991/ESJ-2025-09-03-013.

Guo, B., 2024. Ensemble learning approach for accurate virtual borehole prediction in 3D geological modeling. Int. J. Digital Earth 17 (1), 1–27. https://doi.org/10.1080/17538947.2024.2409964.

Guo, L., Hermans, T., Benoit, N., Dudal, D., Van De Vijver, E., Madsen, R., Nørgaard, J., Deleersnyder, W., 2025. Enhanced Markov-type categorical prediction with geophysical soft constraints for hydrostratigraphic modeling. Egusphere 2025, 1–41. https://doi.org/10.5194/egusphere-2025-3160.

Hang, Z., Xue, T., Chen, J., Shi, Y., Yin, Z., Cui, Z., Zhou, G., 2025. A 3D geological modeling method using the transformer model: a solution for sparse borehole data. Minerals 15 (3), 301. https://doi.org/10.3390/min15030301.

Hasan, M.F.R., Susilo, A., Sutan Haji, A.T., Suryo, E.A., Agung, P.A.M., Idmi, M.H., Musta, B., 2025. Subsurface mapping and geotechnical design for landslide mitigation. Civ. Eng. J. 11 (9), 3811–3825. https://doi.org/10.28991/CEJ-2025-011-09-015.

He, Z., Xu, X., Peng, P., Wang, L., Tian, S., 2025. A deep learning-driven three-dimensional geological modeling method using sparse borehole sampling data. Measurement 256, 118461. https://doi.org/10.1016/j.measurement.2025.118461.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. Appl. 13 (4), 18–28. https://doi.org/10.1109/5254.708428.

Hillier, M., Wellmann, F., Brodaric, B., de Kemp, E., Schetselaar, E., 2021. Three-dimensional structural geological modeling using graph neural networks. Math. Geosci. 53 (8), 1725–1749. https://doi.org/10.1007/s11004-021-09945-x.

Hu, Y., Wang, Z.Z., Guo, X., Kek, H.Y., Ku, T., Goh, S.H., Leung, C.F., Tan, E., Zhang, Y., 2024. Three-dimensional reconstruction of subsurface stratigraphy using machine learning with neighborhood aggregation. Eng. Geol. 337, 107588. https://doi.org/10.1016/j.enggeo.2024.107588.

Jordão, H., Sousa, A.J., Soares, A., 2023. Using Bayesian neural networks for uncertainty assessment of ore type boundaries in complex geological models. Nat. Resour. Res. 32 (6), 2495–2514. https://doi.org/10.1007/s11053-023-10265-6.

Juda, P., Renard, P., Straubhaar, J., 2020. A framework for the cross-validation of categorical geostatistical simulations. Earth Space Sci. 7 (8), e2020EA001152. https://doi.org/10.1029/2020EA001152.

Juliani, C., Ellefmo, S.L., 2019. Prospectivity Mapping of mineral deposits in northern Norway using radial basis function neural networks. Minerals 9 (2), 131. https://doi.org/10.3390/min9020131.

Kang, X., Kokkinaki, A., Grana, D., Zong, C., Shi, X., Deng, Y., Li, W., Wu, J., 2025. Effects of prior reconstruction resolution and data density on DNAPL imaging and mass-discharge prediction using a variational autoencoder. Math. Geosci. https://doi.org/10.1007/s11004-025-10233-1.

Kitanidis, P.K., 2015. Persistent questions of heterogeneity, uncertainty, and scale in subsurface flow and transport. Water Resour. Res. 51 (8), 5888–5904. https://doi.org/10.1002/2015WR017639.

Kuhn, S., Cracknell, M.J., Reading, A.M., 2018. Lithologic mapping using random forests applied to geophysical and remote-sensing data: a demonstration study from the eastern goldfields of Australia. Geophysics 83 (4), B183–B193. https://doi.org/10.1190/geo2017-0590.1.

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 Sens. (Basel) 17 (2), 213. https://doi.org/10.3390/rs17020213.

Lin, L., Zhong, Z., Li, C., Gorman, A., Wei, H., Kuang, Y., Wen, S., Cai, Z., Hao, F., 2024. Machine learning for subsurface geological feature identification from seismic data: methods, datasets, challenges, and opportunities. Earth Sci. Rev. 257, 104887. https://doi.org/10.1016/j.earscirev.2024.104887.

Liu, Q., Liu, R., He, T., Zhang, H., 2025. Prediction of lithofacies and three-dimensional geological modeling based on CNN-LSTM spatiotemporal networks. Math. Geosci. 57 (7), 1357–1377. https://doi.org/10.1007/s11004-025-10198-1.

Lyu, B., Wang, Y., Shi, C., 2024. Multi-scale generative adversarial networks (GAN) for generation of three-dimensional subsurface geological models from limited boreholes and prior geological knowledge. Comput. Geotech. 170, 106336. https://doi.org/10.1016/j.compgeo.2024.106336.

Mariethoz, G., Caers, J., 2014. Multiple-Point Geostatistics: Stochastic Modeling with Training Images. John Wiley & Sons.

Marquina-Araujo, J.J., Cotrina-Teatino, M.A., Mamani-Quispe, J.N., Soto-Juscamayta, L. M., Ccatamayo-Barrios, J.H., Ortiz-Quintanilla, S.M., Cruz-Galvez, J.A., 2024. Application of multilayer perceptron neural network in geological modeling of categorical variables: a case study in Peru. Math. Modell. Eng. Probl. 11 (6), 1463–1472. https://doi.org/10.18280/mmep.110607.

Martín-Martín, M., Bullejos, M., Cabezas, D., Alcalá, F.J., 2023. Using python libraries and k-Nearest neighbors algorithms to delineate syn-sedimentary faults in sedimentary porous media. Mar. Pet. Geol. 153, 106283. https://doi.org/10.1016/j.marpetgeo.2023.106283.

Merzoug, A., Pyrcz, M., 2025. Conditional generative adversarial networks for multivariate Gaussian subsurface modeling: how good are they? Math. Geosci. 57 (4), 733–757. https://doi.org/10.1007/s11004-025-10176-7.

Nti, I.K., Nyarko-Boateng, O., Aning, J., 2021. Performance of machine learning algorithms with different K values in K-fold cross validation. Int. J. Inf. Technol. Comput. Sci. 13 (6), 61–71. https://doi.org/10.5815/ijitcs.2021.06.05.

Pei, J., Zhang, Y., 2022. Prediction of reservoir fracture parameters based on the multi-layer perceptron machine-learning method: a case study of Ordovician and Cambrian carbonate rocks in Nanpu Sag, Bohai Bay Basin, China. Processes 10 (11), 2445. https://doi.org/10.3390/pr10112445.

Peterson, L., 2009. K-nearest neighbor. Scholarpedia 4 (2), 1883. https://doi.org/10.4249/scholarpedia.1883.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., Heikkonen, J., 2017. Estimating the prediction performance of spatial models via spatial k-fold cross validation. Int. J. Geogr. Inf. Sci. 31 (10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.

Rajaram, H., Gelhar, L.W., 1995. Plume-scale dependent dispersion in aquifers with a wide range of scales of heterogeneity. Water Resour. Res. 31 (10), 2469–2482. https://doi.org/10.1029/95WR01723.

Remy, N., Boucher, A., Wu, J., 2009. Applied Geostatistics with SGeMS: A User's Guide. Cambridge University Press.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001.

Rubin, Y., 2003. Applied Stochastic Hydrogeology. Oxford University Press.

Saljoughi, B.S., Hezarkhani, A., 2018. A comparative analysis of artificial neural network (ANN), wavelet neural network (WNN), and support vector machine (SVM) data-driven models to mineral potential mapping for copper mineralizations in the Shahr-e-Babak region, Kerman, Iran. Appl. Geomat. 10 (3), 229–256. https://doi.org/10.1007/s12518-018-0229-z.

Shi, C., Wang, Y., 2022. Machine learning of three-dimensional subsurface geological model for a reclamation site in Hong Kong. Bull. Eng. Geol. Environ. 81 (12), 504. https://doi.org/10.1007/s10064-022-03009-y.

Singh, R.K., Ray, D., Sarkar, B.C., 2022. Mineral deposit grade assessment using a hybrid model of kriging and generalized regression neural network. Neural Comput. Appl. 34 (13), 10611–10627. https://doi.org/10.1007/s00521-022-06951-w.

Smirnoff, A., Boisvert, E., Paradis, S.J., 2008. Support vector machine for 3D modelling from sparse geological information of various origins. Comput. Geosci. 34 (2), 127–143. https://doi.org/10.1016/j.cageo.2006.12.008.

Soltanian, M.R., Ritzi, R.W., 2014. A new method for analysis of variance of the hydraulic and reactive attributes of aquifers as linked to hierarchical and multiscaled sedimentary architecture. Water Resour. Res. 50 (12), 9766–9776. https://doi.org/10.1002/2014WR015468.

Song, S., Mukerji, T., Hou, J., Zhang, D., Lyu, X., 2022. GANSim-3D for conditional geomodeling: theory and field application. Water Resour. Res. 58 (7), e2021WR031865. https://doi.org/10.1029/2021WR031865.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21. https://doi.org/10.1023/A:1014009426274.

Sun, A.Y., 2018. Discovering state-parameter mappings in subsurface models using generative adversarial networks. Geophys. Res. Lett. 45 (20), 11–137. https://doi.org/10.1029/2018GL080404.

Sun, K., Hu, Y., Lakhanpal, G., & Zhou, R.Z. (2023). Spatial cross-validation for GeoAI. In: *Handbook of geospatial artificial intelligence* (pp. 201–214). CRC Press.

Vaferi, B., Eslamloueyan, R., Ayatollahi, S., 2011. Automatic recognition of oil reservoir models from well testing data by using multi-layer perceptron networks. J. Petrol. Sci. Eng. 77 (3–4), 254–262. https://doi.org/10.1016/j.petrol.2011.03.002.

Valavi, R., Elith, J., Lahoz-Monfort, J.J., Guillera-Arroita, G., 2018. blockCV: an R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. bioRxiv, 357798. https://doi.org/10.1101/357798.

Wadoux, A.-M.-J.-C., Heuvelink, G.B.M., de Bruin, S., Brus, D.J., 2021. Spatial cross-validation is not the right way to evaluate map accuracy. Ecol. Model. 457, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692.

Wallace, C.D., Tonina, D., McGarr, J.T., de Barros, F.P.J., Soltanian, M.R., 2021. Spatiotemporal dynamics of nitrous oxide emission hotspots in heterogeneous riparian sediments. Water Resour. Res. 57 (12), e2021WR030496. https://doi.org/10.1029/2021WR030496.

Wang, J., Hopkins, L., Hallman, T., Robinson, W.D., Hutchinson, R., 2023. Cross-validation for geospatial data: estimating generalization performance in geostatistical problems. Trans. Mach. Learn. Res.

Wang, L., Gao, Y., Pan, Q., Wang, S., Phoon, K.-K., 2025a. Coupled geological modeling using multi-source data: a K-dimensional tree-graph convolutional neural process approach. Comput. Geotech. 187, 107509. https://doi.org/10.1016/j.compgeo.2025.107509.

Wang, L., Pan, Q., Su, D., Huang, S., 2025b. Three-dimensional voxel geological modelling for subsurface stratigraphy: a graph convolutional network approach. Can. Geotech. J. 62, 1–15. https://doi.org/10.1139/cgj-2024-0191.

Wang, X., Yang, S., Zhao, Y., Wang, Y., 2018. Lithology identification using an optimized KNN clustering method based on entropy-weighed cosine distance in Mesozoic strata of Gaoqing field, Jiyang depression. J. Petrol. Sci. Eng. 166, 157–174. https://doi.org/10.1016/j.petrol.2018.03.034.

Wu, Y., Wang, H., Zhang, B., Du, K.-L., 2012. Using radial basis function networks for function approximation and classification. Int. Scholar. Res. Notices 2012 (1), 324194.

Xia, Y., Zhan, C., Dai, Z., Wu, J., Zhang, X., Yin, H., Zhu, L., Yan, J., Wang, Z., Soltanian, M.R., Carroll, K.C., 2026. Combining hydrologic, chemical, and geophysical deep learning-based inversion for heterogeneous aquifer structure identification. J. Hydrol. 665, 134701. https://doi.org/10.1016/j.jhydrol.2025.134701.

Yeh, T.-C., Khaleel, R., Carroll, K.C., 2015. Flow Through Heterogeneous Geologic Media. Cambridge University Press.

Zhan, C., Dai, Z., Jiao, J.J., Soltanian, M.R., Yin, H., Carroll, K.C., 2025. Toward artificial general intelligence in hydrogeological modeling with an integrated latent diffusion framework. Geophys. Res. Lett. 52 (3), e2024GL114298. https://doi.org/10.1029/2024GL114298.

Zhan, C., Dai, Z., Soltanian, M.R., De Barros, F.P.J., 2022a. Data-worth analysis for heterogeneous subsurface structure identification with a stochastic deep learning framework. Water Resour. Res. 58 (11), e2022WR033241. https://doi.org/10.1029/2022WR033241.

Zhan, C., Dai, Z., Soltanian, M.R., Zhang, X., 2022b. Stage-wise stochastic deep learning inversion framework for subsurface sedimentary structure identification. Geophys. Res. Lett. 49 (1), e2021GL095823. https://doi.org/10.1029/2021GL095823.

Zhan, C., Dai, Z., Yang, Z., Zhang, X., Ma, Z., Thanh, H.V., Soltanian, M.R., 2023. Subsurface sedimentary structure identification using deep learning: a review. Earth Sci. Rev. 239, 104370. https://doi.org/10.1016/j.earscirev.2023.104370.

Zhang, J., Cao, C., Nan, T., Ju, L., Zhou, H., Zeng, L., 2024. A novel deep learning approach for data assimilation of complex hydrological systems. Water Resour. Res. 60 (2), e2023WR035389. https://doi.org/10.1029/2023WR035389.

Zheng, N., Jiang, S., Xia, X., Kong, W., Li, Z., Gu, S., Wu, Z., 2023. Efficient estimation of groundwater contaminant source and hydraulic conductivity by an ILUES framework combining GAN and CNN. J. Hydrol. 621, 129677. https://doi.org/10.1016/j.jhydrol.2023.129677.