Toward a Theory-Guided Framework for Spatial Data Fusion in GIS: Challenges, Methodological Insights, and an Operational Checklist

Ali Azizi[1], Parham Pahlavani[*2], and Mohammad Nakhaei[3]
[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran, azizi.ali@ut.ac.ir
[2] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran, pahlavani@ut.ac.ir
[3] Department of Applied Geology, Faculty of Earth Science, Kharazmi University, Tehran, Iran, nakhaei@khu.ac.ir

Keywords: Checklist for Spatial Data Fusion in GIS; Data Selection and Preparation; Integration Method Selection

Abstract

The synergistic integration of spatial layers (data fusion) within Geographic Information Systems (GIS) is pivotal for enhancing environmental decision-making and spatial planning. However, this process is fraught with inherent complexities, including the necessity for explicitly defined decision objectives, the judicious selection of conceptually and statistically relevant data layers, ensuring robust data quality and balance, devising optimal reclassification strategies, and rigorously validating model outputs through appropriate metrics. Addressing these multifaceted issues, this paper introduces a structured framework complemented by an operational checklist. This methodology is designed to minimize subjective biases, bolster methodological transparency, and significantly enhance the reproducibility of spatial analysis results. The proposed framework is versatile, accommodating a spectrum of data environments: from data-scarce contexts where knowledge-based or statistical approaches, such as the Analytic Hierarchy Process (AHP) or Dempster-Shafer Theory, are crucial for compensating for limited data, to data-rich settings that can leverage advanced machine learning (ML) techniques. Furthermore, the framework explicitly acknowledges inherent limitations. These include the restricted transferability of models to disparate geographical regions, the necessity for localized parameter tuning, and the fundamental understanding that computational modeling, while powerful, serves as a decision support tool rather than a replacement for indispensable field-based assessments. By delineating between addressable methodological challenges and unavoidable contextual constraints, this framework provides a practical and comprehensive guide for advanced GIS-based spatial analysis across critical domains such as natural resource management, environmental assessment, and urban planning.

1. Introduction

The integration of spatial layers—commonly referred to as data fusion—within Geographic Information Systems (GIS) is foundational for spatial analysis and environmental decision-making. By combining heterogeneous datasets from multiple sources, such as remote sensing imagery, digital elevation models (DEMs), field surveys, and thematic vector layers, spatial layer integration enables a comprehensive understanding of geospatial phenomena and supports decisions in critical domains including natural resource management, urban planning, environmental risk assessment, and sustainable development (T. Meng et al., 2020, H. Dai et al., 2025, V.P. Dheeraj et al., 2025).

Despite its potential, the spatial data fusion process in GIS remains methodologically challenging. Key issues include the need for explicitly defined decision objectives, the careful selection of conceptually and statistically relevant layers, ensuring data completeness and balance, devising robust reclassification strategies, and validating outputs using appropriate performance metrics. Additionally, integration efforts are complicated by the heterogeneity of spatial data in terms of resolution, scale, format, and accuracy, as well as the subjectivity associated with expert-based methods.

Traditional approaches such as the Weighted Linear Combination (WLC) (W. Nigussie et al., 2019), Analytic Hierarchy Process (AHP) (P. Arulbalaji et al., 2019), and fuzzy logic (S.V.R. Termeh et al., 2019) remain widely used due to their simplicity and transparency. However, their reliance on subjective judgments can introduce bias and limit generalizability. In contrast, data-driven methods—including logistic regression (H.-A. Park, 2013), Support Vector Machines (SVM), Random Forests (RF), and Deep Neural Networks (DNN)—offer the ability to model non-linear, high-dimensional relationships while reducing dependence on predefined assumptions and expert input (S.A. Naghibi et al., 2017, S.K. Roy et al., 2024). These techniques have significantly enhanced GIS-based spatial modeling, enabling robust predictions in applications such as land cover classification, hazard mapping, and hydrological modeling.

Moreover, fuzzy logic and Bayesian methods have gained traction for their capacity to manage uncertainty and reconcile conflicting or incomplete information, particularly in complex decision-making scenarios (S. Obeidavi et al., 2021). Hybrid frameworks that combine data-driven and expert-driven methods are also increasingly adopted to enhance interpretability and predictive accuracy.

Nevertheless, several critical limitations persist, such as the dependence on high-quality training data, the need for localized parameter tuning, and challenges in model transferability across different geographic regions. Furthermore, many studies lack a structured analytical framework and a comprehensive evaluation checklist that systematically address data quality, class balance, weighting methods, model validation, and spatial generalizability.

To address this gap, the present study proposes a step-by-step decision-making framework and operational checklist to guide spatial layer integration in GIS. The framework is designed to enhance methodological transparency, reduce subjective bias, and improve reproducibility across data environments—ranging from data-scarce regions reliant on expert knowledge to data-

* Corresponding author

rich contexts capable of supporting advanced machine learning workflows. In doing so, it aims to standardize spatial fusion practices and support reliable, actionable insights in geospatial decision-making.

2. Types of Data Utilised

Spatial data for integration typically comes in two primary formats:

Raster Data: Often derived from remote sensing sources or Digital Elevation Models (DEMs).

- **DEMs** :provide fundamental topographic and hydro-geomorphological parameters, including elevation, slope, aspect, curvature, Topographic Wetness Index (TWI), distance from streams, and Topographic Position Index (TPI) (W. Chen et al., 2018:, 2019:, M. Panahi et al., 2020:)
- **Remote sensing data** (from satellites like Landsat, Sentinel, MODIS or airborne sensors) offer environmental and dynamic information, yielding layers such as Vegetation Indices (e.g., NDVI, EVI), Soil Moisture Indices (e.g., NDSI), Thermal Indices (e.g., Land Surface Temperature - LST), Snow/Ice Cover Indices, and Radar Data (SAR) (W. Nigussie et al., 2019, R. Goldblatt et al., 2021).

Vector Data: Represents discrete spatial features as points, lines, and polygons, commonly derived from field surveys, base maps, or standardised databases. Examples include land use/land cover classifications, road networks, rivers and surface water bodies, sampling points (e.g., wells, meteorological stations), and human infrastructure (A. Azizi et al., 2017).

3. Methods for Spatial Layer Integration

Spatial layer integration in Geographic Information Systems (GIS) can be broadly categorized into four principal approaches: classical expert-driven frameworks, advanced data-driven or intelligent techniques (including machine learning and deep learning), Bayesian models, and hybrid strategies.

3.1 Classical Expert-Driven Approaches

Classical methods rely on well-structured and transparent frameworks that emphasize interpretability and the integration of expert knowledge. These approaches include:

Boolean Overlay: This method applies binary conditions (e.g., suitable/unsuitable) based on strict thresholds, offering high transparency. However, its simplicity limits the ability to capture nuanced or gradual spatial variations (C. Cheng, and R.G. Thompson, 2016).

Weighted Linear Combination (WLC): WLC assigns weights to spatial layers—typically derived from expert judgment—and combines them linearly. It is computationally efficient and intuitive but suffers from inherent subjectivity in weight assignment and lacks robust mechanisms for uncertainty management.

Multi-Criteria Decision Analysis (MCDA): MCDA comprises a suite of methods designed to evaluate and rank alternatives based on multiple, often conflicting, criteria, making it highly valuable in complex spatial decision-making (A.L. Achu et al., 2020). MCDA integrates weighted spatial layers, constraints, and quantitative criteria to produce final preference orders. A key challenge remains the subjectivity involved in weight determination. Prominent MCDA techniques include:

- **Analytic Hierarchy Process (AHP):** Decomposes complex decisions into hierarchical structures and employs pairwise comparisons to derive criterion weights and rank alternatives. Widely used in GIS applications such as site selection and prioritization, AHP's effectiveness depends heavily on subjective expert input (P. Arulbalaji et al., 2019).

- **Technique for Order Preference by Similarity to Ideal Solution (TOPSIS):** Ranks alternatives by their closeness to an ideal solution and distance from a negative-ideal solution. It is frequently applied alongside other optimization methods in energy, transportation, and urban planning contexts (S.K. Ray, 2025).

- **Ordered Weighted Averaging (OWA):** A flexible aggregation technique that models a spectrum of decision attitudes—from optimistic to pessimistic—by reordering criterion values before weighting. OWA is especially useful in GIS applications like landslide risk mapping and land-use suitability analysis, as it aids in managing uncertainty and conflicting information (J. Wang et al., 2024).

Fuzzy Logic: Fuzzy logic techniques model vagueness and uncertainty by assigning continuous suitability values through membership functions rather than strict binary classifications. This approach effectively manages uncertainty and conflicting evidence, particularly in heterogeneous or incomplete datasets. Defining appropriate membership functions often requires expert judgment. An important advancement is the Adaptive Neuro-Fuzzy Inference System (ANFIS), which synergizes the learning capabilities of neural networks with fuzzy logic's interpretability and uncertainty handling. ANFIS autonomously generates decision rules and uncovers complex relationships between criteria, thereby reducing reliance on pre-assigned weights. It has demonstrated strong performance in function approximation, clustering, and pattern recognition tasks (S.V.R. Termeh et al., 2019).

3.2 Advanced Data-Driven and Intelligent Approaches (Machine Learning and Deep Learning)

Advanced methods leverage computational intelligence to model complex relationships and uncertainties inherent in spatial data. These approaches are particularly effective when dealing with high-dimensional datasets, such as those derived from remote sensing or multi-source geospatial data.

Logistic Regression: This statistically grounded and interpretable method predicts binary or categorical outcomes. Despite its computational efficiency, logistic regression assumes linear relationships between predictors and outcomes and is sensitive to class imbalance, which may limit its performance in complex spatial problems (H.-A. Park, 2013).

Machine Learning (ML) Algorithms: ML techniques excel at capturing nonlinear patterns and automatically extracting relevant features from large datasets, enhancing predictive accuracy (J. Han et al., 2022). Key algorithms include:

- *Support Vector Machines (SVM):* Widely used for classification and regression tasks, SVMs are especially effective in high-dimensional feature spaces, providing robust decision boundaries (M. Bansal et al., 2022).

- *Random Forest (RF):* An ensemble learning method that constructs multiple decision trees, Random Forest improves prediction performance and reduces overfitting by aggregating diverse trees' outputs (S. Sachdeva, and B. Kumar, 2021).

- *Artificial Neural Networks (ANNs):* Inspired by biological neural architectures, ANNs can model complex, nonlinear relationships without explicit predefined equations. They have been extensively applied in remote sensing classification and other spatial modeling tasks (P.T. Nguyen et al., 2020).

Deep Learning (DL) Neural Networks: A subfield of machine learning that utilizes multi-layered architectures to learn hierarchical representations from data, enabling advanced spatial analysis:

- *Autoencoders:* Neural networks designed to learn compact and efficient encodings of input data, facilitating dimensionality reduction and anomaly detection in geospatial datasets (S. Nagar et al., 2024).

- *Multi-Layer Perceptrons (MLPs):* Basic feedforward neural networks applied in diverse areas including voice recognition and predictive analytics, adaptable for spatial data modeling (E. Dodangeh et al., 2020).

- *Convolutional Neural Networks (CNNs):* Architectures specialized for analyzing visual data, capable of detecting spatial patterns and features within imagery. Variants such as U-Net have demonstrated superior performance in precise segmentation tasks like land cover classification and change detection (W.L. Hakim et al., 2022).

- *Recurrent Neural Networks (RNNs):* Tailored for sequential data processing, RNNs maintain temporal dependencies through internal memory. Long Short-Term Memory (LSTM) networks, a widely used RNN variant, are effective in time-series forecasting, speech recognition, and spatio-temporal modeling of environmental phenomena (W.L. Hakim et al., 2022).

- *Deep Reinforcement Learning (DRL):* This approach involves an agent learning optimal actions through interactions with an environment, guided by reward maximization. DRL holds promise for dynamic spatial planning and adaptive decision-making processes in complex geospatial systems (Y. Zheng et al., 2023).

3.3 Bayesian Approaches

Bayesian methods provide a rigorous probabilistic framework for reasoning under uncertainty and incomplete information. Notable techniques include:

Dempster-Shafer Theory (DST): A mathematical framework that facilitates reasoning with uncertain and incomplete evidence by combining belief functions from multiple sources. DST allows for flexible representation of uncertainty beyond traditional probability theory (S. Obeidavi et al., 2021).

Bayesian Belief Networks (BBN): Graphical probabilistic models that encode dependencies among variables using directed acyclic graphs, enabling inference and prediction in complex systems with uncertain relationships (H. Woldesellasse, and S. Tesfamariam, 2025).

3.4 Hybrid Approaches

Hybrid approaches integrate multiple techniques to leverage their complementary strengths while mitigating the inherent limitations of individual methods. Several representative hybrid strategies include:

AHP–Fuzzy Logic Integration: This approach merges the hierarchical structuring of decision criteria provided by the Analytic Hierarchy Process (AHP) with the ability of fuzzy logic to model uncertainty and linguistic vagueness. It is particularly useful in contexts where expert knowledge is abundant but imprecise (Z. Shao et al., 2020).

AHP–Machine Learning Integration: Here, expert judgment informs initial weight assignments or feature prioritization, which are subsequently refined through machine learning algorithms. This enhances both interpretability and predictive performance while maintaining domain relevance (M. Hussain et al., 2023).

Fuzzy Logic–Machine Learning Integration: Fuzzy logic is used to preprocess spatial inputs—such as environmental or socio-economic variables—by transforming them into fuzzy membership values that represent degrees of suitability or risk. These transformed inputs are then fed into classifiers like Random Forests or Convolutional Neural Networks to detect complex, nonlinear patterns.

Ensemble Learning and Stacking: More recent advances in hybridization include stacking-based frameworks, where multiple base learners (e.g., decision trees, SVMs, gradient boosting machines) are combined through a meta-learner that synthesizes individual predictions into a unified output. This process mimics a voting mechanism and often leads to enhanced generalization performance, reduced over fitting, and improved predictive stability (S. Chen et al., 2025).

Selecting an appropriate hybrid integration strategy requires careful consideration of project goals, data characteristics, spatial resolution, and desired levels of methodological transparency and reproducibility. Table 1 summarizes the strengths and limitations of commonly used techniques in this context. A growing trend in recent GIS research is the deliberate fusion of classical knowledge-based models and modern data-driven algorithms, particularly those incorporating remote sensing data, expert knowledge, and machine learning techniques. This convergence holds promise for producing scalable, accurate, and context-sensitive decision support systems in a wide array of geospatial applications.

	Method	Description	Advantages	Limitations
3.1	**Boolean Overlay**	Applies binary conditions (e.g., suitable/unsuitable) based on strict thresholds.	Simple, transparent, intuitive.	Overly rigid; lacks nuance in spatial variation.
	Weighted Linear Combination (WLC)	Assigns expert-derived weights to layers and combines them linearly.	Easy to implement; intuitive.	Subjective weights; limited handling of uncertainty.
	Multi-Criteria Decision Analysis (MCDA)	Evaluates alternatives using multiple conflicting criteria. Includes several sub-methods:	Integrates diverse criteria; adaptable to complex decisions.	Weight assignment is subjective.
	Fuzzy Logic	Assigns continuous suitability using membership functions.	Models vagueness and uncertainty.	Defining functions requires expert input.
	* ANFIS	Combines neural networks and fuzzy inference for adaptive modelling.	Learns rules autonomously; interpretable.	Training can be complex.
3.2	**Logistic Regression**	Predicts binary outcomes with linear assumptions.	Interpretable; efficient.	Limited in non-linear relationships; sensitive to imbalance.
	Support Vector Machines (SVM)	Classifies data using optimal hyperplanes.	High performance in high-dimensional spaces.	Requires parameter tuning.
	Random Forest (RF)	Ensemble of decision trees with bootstrap aggregation.	Robust; reduces over fitting.	Less interpretable; sensitive to noise.
	Artificial Neural Networks (ANNs)	Models complex non-linear relations without predefined equations.	High adaptability.	Black-box nature; requires tuning.
	Autoencoders	Learns compressed representations of data.	Useful for anomaly detection, feature extraction.	May lose interpretability.
	MLPs	Simple feedforward neural networks.	Versatile; general-purpose.	Limited temporal/spatial specificity.
	CNNs	Specialized for spatial imagery; detects patterns and segments data.	Excellent for image classification.	Requires large datasets.
	RNNs / LSTM	Maintains memory for temporal sequences; suitable for time-series.	Effective in environmental forecasting.	Vanishing gradient issue (partially solved in LSTM).
	Deep Reinforcement Learning (DRL)	Learns optimal actions through reward-driven interactions.	Adaptive decision-making.	Complex training; data-intensive.
3.3	**Dempster-Shafer Theory (DST)**	Combines multiple belief sources under uncertainty.	Handles incomplete information.	Complex rule definition.
	Bayesian Belief Networks (BBNs)	Probabilistic graphical models encoding dependencies.	Captures causal relationships.	Needs prior distributions; parameter tuning required.
3.4	**AHP–Fuzzy Logic**	Combines structured decision-making with modelling of vagueness.	Balances structure and flexibility.	Requires expert-designed fuzzy rules.
	AHP–ML	Uses expert input to initialize ML models for better learning and interpretation.	Improved accuracy with transparency.	Complexity in integration.
	Fuzzy–ML	Transforms fuzzy data into ML-ready format for classification.	Captures uncertainty; improves learning.	Requires dual modelling.
	Stacking (Ensemble Learning)	Combines multiple base learners into a meta-model.	Reduces over fitting; improves generalization.	Needs large data; model tuning.

Table 1. Methods for Spatial Layer Integration in GIS

4. Challenges and a Proposed Checklist for Developing a Robust Scientific Framework in Spatial Layer Integration

Developing a robust scientific framework for spatial layer integration in Geographic Information Systems (GIS) requires meticulous attention to five critical components: (1) clearly defining objectives, (2) selecting conceptually relevant and high-quality data, (3) ensuring data integrity through sound sampling and preprocessing, (4) selecting appropriate integration methods, and (5) applying rigorous model validation strategies. Each component corresponds to a specific stage in the modeling workflow and addresses challenges that can

critically affect the accuracy, transparency, and generalizability of GIS-based spatial analysis.

First, a precise definition of the research objective is foundational. This should clearly articulate whether the objective is quantitative (e.g., classification, prediction) or qualitative (e.g., policy assessment, expert mapping), as this distinction directly influences the modeling paradigm. Quantitative goals typically require large, high-quality datasets and advanced data-driven methods such as Random Forest or Convolutional Neural Networks (CNN). In contrast, qualitative objectives are better addressed through expert-based approaches like the Analytic Hierarchy Process (AHP), fuzzy logic, or hybrid techniques, where interpretability and transparency are prioritized.

Second, data selection is a crucial step that directly affects model realism and reliability. Conceptually irrelevant or redundant layers can introduce noise and artificial correlations. For data-driven models, this stage involves correlation analysis, dimensionality reduction, and optimal feature selection to minimize overfitting and computational complexity. For expert-based frameworks, layer inclusion should be guided by theoretical understanding and domain knowledge. Reclassification of continuous variables must be executed with caution—poorly defined class boundaries can distort spatial distributions, introduce bias, and mislead decision-making. Post-reclassification checks and sensitivity analyses are essential to evaluate mapping balance and avoid class imbalances.

Third, ensuring data quality and designing an unbiased sampling strategy are essential for model validity. Spatial sampling must be representative and evenly distributed across the full range of environmental conditions. Recommended approaches include stratified random or spatially balanced sampling to avoid spatial autocorrelation and sampling bias. Preprocessing steps—such as normalization, outlier detection, and the temporal and spatial alignment of input layers—enhance consistency and model performance. Additionally, addressing class imbalance through resampling or weighting techniques is critical in both classification and regression tasks.
Fourth, the choice of spatial integration method should be guided by data characteristics, research goals, and the required level of interpretability. This involves evaluating trade-offs between transparency and predictive performance. In contexts where data and expert opinions may conflict, hybrid methods provide a structured means to integrate both. Furthermore, consideration of uncertainty—whether through probabilistic frameworks like Bayesian modeling or belief function theory (e.g., Dempster-Shafer)—enhances model robustness and decision confidence.

Finally, robust model validation is essential to ensure scientific credibility. For classification problems, metrics such as accuracy, F1-score, sensitivity (recall), and ROC-AUC are commonly applied, while regression tasks are evaluated using RMSE, MAE, and the coefficient of determination (R^2). In expert-based or qualitative models, practical metrics like the proportion of known high-potential zones correctly identified can be used. The validation strategy must be tailored to the spatial and temporal structure of the data—standard k-fold cross-validation is suitable for general datasets. Sensitivity analysis supports the identification of key input drivers, while uncertainty analysis ensures transparency in model limitations. Ultimately, generalizability must be confirmed through external validation on independent datasets and reporting of uncertainty bounds.
This five-stage checklist provides a comprehensive foundation for developing scientifically rigorous and operationally robust frameworks for spatial layer integration in GIS. The proposed checklist is presented in Table 2.

5. Discussion

This section critically examines the challenges and the proposed framework within the specific context of groundwater potential mapping in arid and semi-arid regions. Designing robust analytical frameworks in such environments is particularly complex due to data scarcity, environmental heterogeneity, and the diversity of decision-making objectives.
In many arid areas, spatial information is often limited to the locations of existing successful wells. This poses two major limitations for classification-based machine learning approaches. First, the small number of samples constrains the training of complex models such as Random Forest, Support Vector Machines (SVM), or Convolutional Neural Networks (CNN), increasing the risk of overfitting. Second, the lack of explicit negative samples—areas where drilling attempts failed—prevents the proper formulation of a binary classification problem, which requires both presence and absence data.
Under such conditions, expert-driven methods such as the Analytic Hierarchy Process (AHP), Multi-Criteria Decision Analysis (MCDA), and fuzzy logic offer a practical alternative. These approaches rely on expert judgment and conceptual understanding of environmental variables (e.g., slope, geology, land use) to generate relative potential maps. In the absence of formal ground-truth datasets, overall accuracy can be used to validate expert-based results, typically by calculating the proportion of existing wells that fall within high-potential zones. When expert knowledge is limited or conceptual understanding of groundwater controls is incomplete, data-compatible methods such as logistic regression or Dempster-Shafer Theory (DST) provide feasible alternatives. These techniques estimate the spatial probability of groundwater occurrence and can extract meaningful patterns even from sparse datasets. Validation metrics such as prediction accuracy, success rate analysis, or probability matching serve as useful performance indicators in such contexts.
In contrast, in data-rich environments where both successful and failed drilling data are available, advanced machine learning models become suitable. Models such as Random Forest, SVM, and CNN can effectively capture complex nonlinear relationships between environmental factors and well outcomes. Validation strategies—including confusion matrices, F1-scores, ROC-AUC, and sensitivity analysis of class areas—ensure model robustness and help mitigate class imbalance issues.
Another important consideration is the reclassification of continuous layers, such as slope or vegetation indices. Improper reclassification schemes (e.g., Natural Breaks, Equal Interval, Quantile) can distort class boundaries and influence the final output map distribution. Therefore, systematic comparison of alternative reclassification scenarios and sensitivity analysis are essential for minimizing uncertainty and ensuring meaningful classifications.
Hybrid methods that combine expert knowledge and data-driven outputs offer a promising solution. By using structured weighting strategies or rule-based fusion, these approaches can balance interpretability and predictive power while reducing bias. They also allow integration of empirical data with local

knowledge, making them particularly suitable for regions with limited ground data.

Stage	Technical Activities	Scientific Purpose / Justification
1. Define Objective	- Define clear research question and objective type (quantitative/qualitative) - Specify spatial and temporal scope - Align approach with scale and resolution requirements	Ensure appropriate model choice and consistent alignment with research goals and scale dependencies.
2. Data Selection	- Conduct exploratory and statistical analysis of variables - Remove non-informative, redundant, or low-variance features - Eliminate artificial layers and select an optimal feature subset - Choose appropriate reclassification schemes - Analyse class area distribution	Improve data relevance, reduce noise, avoid class imbalance and ensure realistic, Minimizes computational complexity, balanced final mapping outcomes.
3. Data Quality and Sampling	- Design spatially balanced, representative sampling strategy - Apply data pre-processing (normalization, outlier removal) - Check class distribution and apply resampling as needed	Minimize spatial bias, enhance data accuracy and consistency, prevent model overfitting to dominant classes.
4. Choose Integration Method	- Assess data volume, quality, and heterogeneity - Evaluate interpretability vs. predictive power needs - Address expert-data conflicts via hybrid integration - Consider uncertainty handling requirements	Ensure model selection aligns with project goals, data availability, complexity of patterns, and desired uncertainty representation.
5. Model Validation	- Select appropriate performance metrics (accuracy, RMSE, AUC, etc.) - Apply k-fold or spatial block cross-validation - Conduct sensitivity and uncertainty analysis	Rigorously evaluate model performance, robustness, and reliability under varying data conditions.

Table 2. Checklist for Spatial Modelling Workflow

Beyond groundwater-specific concerns, spatial data integration methodologies face cross-cutting challenges that broadly fall into two categories. Expert-driven approaches, while valuable for their interpretability, often suffer from subjectivity in weight assignment, vague definitions of fuzzy membership functions, and limited mechanisms for uncertainty quantification—factors that reduce reproducibility and may introduce bias. On the other hand, data-driven methods, particularly those integrating GIS with machine learning or deep learning, encounter issues related to data quality and availability, high computational costs, limited model transferability, and the "black box" nature of complex algorithms, which hampers interpretability and trust. Addressing these limitations through hybrid, transparent, and context-aware frameworks is critical for advancing robust spatial decision-making—especially in environmentally complex and data-scarce regions.

6. Conclusion

This study proposes a structured framework and operational checklist to support the integration of spatial layers in Geographic Information Systems (GIS), aiming to reduce subjective judgments and enhance methodological consistency. By clearly defining decision-making objectives, selecting conceptually and statistically relevant layers, ensuring data quality and spatial balance, managing class imbalance and reclassification schemes, and adopting appropriate validation metrics, the framework offers a practical guide for developing transparent, repeatable, and scientifically robust spatial analyses.

The discussion highlights that modeling strategies must be tailored to data availability and local conditions. In data-rich regions, advanced machine learning approaches can effectively model complex, nonlinear relationships. Conversely, data-poor environments often require expert-driven or simpler statistical methods that prioritize interpretability and conceptual understanding. Interdisciplinary collaboration among hydrogeologists, remote sensing specialists, data scientists, and decision analysts is essential to ensure that such frameworks remain scientifically rigorous while meeting real-world decision-making needs. Crucially, spatial models are not substitutes for thorough field-based assessments but serve as complementary tools to narrow search spaces, prioritize targets, and optimize exploration costs. Ensuring model transferability also requires careful parameter tuning and independent validation in new regions to avoid biases and systemic errors.

It is important to emphasize that while this checklist provides a structured foundation for GIS-based spatial integration, it is not intended as a static or final tool. Rather, it should be viewed as a living framework that must be critically reviewed, updated, and expanded over time. Future research should focus on improving this checklist by incorporating advances in data acquisition, novel integration algorithms capable of handling uncertainty and multi-scale variability, and more systematic validation strategies across diverse contexts.

References

Achu, A.L., Reghunath, G., Shivashankar, C.K., Sherif, M., Thomas, J., 2020: Multi-Criteria Decision Analysis for Delineation of Groundwater Potential Zones in a Tropical River Basin Using Remote Sensing, GIS and Analytical Hierarchy Process (AHP). Groundwater for Sustainable Development, 10, 100365.

Arulbalaji, P., Padmalal, D., Sreelash, K., 2019: GIS and AHP Techniques Based Delineation of Groundwater Potential Zones: A Case Study from Southern Western Ghats, India. Scientific Reports, 9(1), 1–17.

Azizi, A., Karimipour, F., Esmaeily, A., 2017: Time-Dependent, Activity-Based Itinerary Personal Tour Planning in Multimodal Transportation Networks. Annals of GIS, 23(1), 27–39.

Azizi, A., Pahlavani, P., Nakhaei, M., 2025: Integrating Dempster–Shafer theory and clustering algorithms for enhanced groundwater potential assessment. Stochastic Environmental Research and Risk Assessment, 39(11), 1-15.

Bansal, M., Yadav, B.K., Yadav, A.K., 2022: A Comparative Analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory Algorithms in Machine Learning. Decision Analytics Journal, 3, 100071.

Chen, S., Pan, Y., Lu, W., Zhang, Y., Liu, W., Fang, Z., 2025: Landslide Spatial Prediction Based on Cascade Forest and Stacking Ensemble Learning Algorithm. International Journal of Systems Science, 56(3), 658–670.

Chen, W., Li, X., Hou, E., Zhao, Z., Ren, X., 2018: GIS-Based Groundwater Potential Analysis Using Novel Ensemble Weights-of-Evidence with Logistic Regression and Functional Tree Models. Science of the Total Environment, 634, 853–867.

Chen, W., Tsangaratos, P., Ilia, I., Duan, Z., Chen, X., 2019: Groundwater Spring Potential Mapping Using Population-Based Evolutionary Algorithms and Data Mining Methods. Science of the Total Environment, 684, 31–49.

Cheng, C., Thompson, R.G., 2016: Application of Boolean Logic and GIS for Determining Suitable Locations for Temporary Disaster Waste Management Sites. International Journal of Disaster Risk Reduction, 20, 78–92.

Dai, H., Quddoos, A., Naz, I., Batool, A., 2025: Geospatial Decision Support System for Urban and Rural Aquifer Resilience: Integrating Remote Sensing-Based Rangeland Analysis With Groundwater Quality Assessment. Rangeland Ecology & Management, 99, 102–118.

Dheeraj, V.P., Upadhyay, R.K., Singh, A.K., Kumar, C., 2025: Hydrogeochemical Quality Investigation of Groundwater Resource Using Multivariate Statistical Methods, Water Quality Indices (WQIs), and Health Risk Assessment in Korba Coalfield Region, India. Stochastic Environmental Research and Risk Assessment, 1–22.

Dodangeh, E., Choubin, B., Eigdir, A.N., Nabipour, N., Panahi, M., Mosavi, A., 2020: Integrated Machine Learning Methods with Resampling Algorithms for Flood Susceptibility Prediction. Science of the Total Environment, 705, 135983.

Goldblatt, R., Addas, A., Stuhlmacher, M., Bright, A., de la Torre, G., Adler, B., 2021: Remotely Sensed Derived Land Surface Temperature (LST) as a Proxy for Air Temperature and Thermal Comfort at a Small Geographical Scale. Land, 10(4), 410.

Hakim, W.L., Nur, A.S., Rezaie, F., Panahi, M., Lee, C.-W., Lee, S., 2022: Convolutional Neural Network and Long Short-Term Memory Algorithms for Groundwater Potential Mapping in Anseong, South Korea. Journal of Hydrology: Regional Studies, 39, 100990.

Han, J., Pei, J., Tong, H., 2022: Data Mining: Concepts and Techniques. Morgan Kaufmann.

Hussain, M., Tayyab, M., Ullah, K., Abdullah, S.N.H.S., Pradhan, B., 2023: Development of a New Integrated Flood Resilience Model Using Machine Learning with GIS-Based Multi-Criteria Decision Analysis. Urban Climate, 50, 101589.

Meng, T., Jing, X., Yan, Z., Pedrycz, W., 2020: A Survey on Machine Learning for Data Fusion. Information Fusion, 57, 115–129.

Nagar, S., Farahbakhsh, E., Awange, J., Chandra, R., 2024: Remote Sensing Framework for Geological Mapping via Stacked Autoencoders and Clustering. Advances in Space Research, 74(10), 4502–4516.

Naghibi, S.A., Khosravi, K., Pourghasemi, H.R., 2017: Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. Water Resources Management, 31(9), 2761–2775.

Nguyen, P.T., Ha, D.H., Avand, M., Jaafari, A., Nguyen, H.D., Al-Ansari, N., Phong, T.V., Sharma, R., Kumar, P., Thanh, H.V., Leelawongtawee, J., Phong, P.V., Prakash, I., Pham, B.T., 2020: Groundwater Potential Mapping Combining Artificial Neural Network and Real AdaBoost Ensemble Technique: The DakNong Province Case-Study, Vietnam. International Journal of Environmental Research and Public Health, 17(7), 2473.

Obeidavi, S., Barzegar, R., Moghaddam, H.K., Gandomkar, M., Zeinali, E., 2021: Evaluation of Groundwater Potential Using Dempster-Shafer Model and Sensitivity Analysis of Effective Factors: A Case Study of North Khuzestan Province. Remote Sensing Applications: Society and Environment, 22, 100475.

Panahi, M., Rezaie, F., Naghibi, S.A., Khosravi, K., Lee, C.-W., Lee, S., Pradhan, B., 2020: Spatial Prediction of Groundwater Potential Mapping Based on Convolutional Neural Network (CNN) and Support Vector Regression (SVR). Journal of Hydrology, 588, 125033.

Park, H.-A., 2013: An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. Journal of Korean Academy of Nursing, 43(2), 154–164.

Ray, S.K., 2025: Unveiling Groundwater Gems: A GIS-Powered Fusion of AHP and TOPSIS for Mapping Groundwater Potential Zones. Groundwater for Sustainable Development, 29, 101431.

Roy, S.K., Hasan, M.M., Mondal, I., Alam, J., Pal, S., Pham, Q.B., Anh, D.T., Linh, N.T.T., Khedher, K.M., Elbeltagi, A., 2024: Empowered Machine Learning Algorithm to Identify Sustainable Groundwater Potential Zone Map in Jashore District, Bangladesh. Groundwater for Sustainable Development, 25, 101168.

Sachdeva, S., Kumar, B., 2021: Comparison of Gradient Boosted Decision Trees and Random Forest for Groundwater Potential Mapping in Dholpur (Rajasthan), India. Stochastic Environmental Research and Risk Assessment, 35(2), 287–306.

Shao, Z., Zhong, Q., Altan, O., Taubenböck, H., 2020: Integrated Remote Sensing and GIS Approach Using Fuzzy-AHP to Delineate and Identify Groundwater Potential Zones in Semi-Arid Shanxi Province, China. Environmental Modelling & Software, 134, 104868.

Termeh, S.V.R., Khosravi, K., Sartaj, M., Keesstra, S.D., Tsai, F.T., Dijksma, R., Pham, B.T., 2019: Optimization of an Adaptive Neuro-Fuzzy Inference System for Groundwater Potential Mapping. Hydrogeology Journal, 27(7), 2511–2534.

Wang, J., Xing, Y., Huang, C., Wang, W., Xu, L., Yan, P., Chang, M., 2024: Identification of Priority Conservation Areas for Natural Forest Protection Project in Northeastern China Based on OWA-GIS. Ecological Indicators, 160, 111718.

Woldesellasse, H., Tesfamariam, S., 2025: Risk Assessment of Gas Pipeline Using an Integrated Bayesian Belief Network and GIS: Using Bayesian Neural Networks for External Pitting Corrosion Modelling. The Canadian Journal of Chemical Engineering, 103(1), 98–109.

Zheng, Y., Lin, Y., Zhao, L., Wu, T., Li, D., Yin, Y., Wang, W., Zheng, Y., Li, Y., 2023: Spatial Planning of Urban Communities via Deep Reinforcement Learning. Nature Computational Science, 3(9), 748–762.