An Integrated GIS and Explainable Machine Learning Framework for Groundwater Contamination Risk Assessment and Drinking Water Security

Surabhi Shivendra, Pusplata Puja,
Raunak, Ajay Prabhakar, Prakash Kumar Sinha

Abstract: *Geogenic contamination of groundwater by arsenic and fluoride threatens drinking-water security across the Ganga basin, and Bihar is among the most affected Indian states. GIS overlay indices rely on expert weights, while machine learning (ML) models are accurate but rarely interpretable enough for public-health decisions. This paper presents an integrated framework that couples GIS-based predictor engineering, spatially cross-validated ML classification, SHapley Additive exPlanations (SHAP) and a drinking-water security index (DWSI) that converts contamination probabilities into population-weighted intervention priorities. Because explanations must be shown to recover true processes before they guide policy, the framework was evaluated on a controlled synthetic Ganga-plain benchmark (120 km × 80 km, 1,800 wells) with known arsenic and fluoride generating rules. Under spatial-block cross-validation, logistic regression (AUC 0.929), random forest (0.925) and XGBoost (0.917) performed similarly and all exceeded SVM (0.878); fluoride was predicted with an AUC of 0.870. SHAP recovered the imposed controls, namely clay-rich Holocene sediments, proximity to the river and the 20–60 m shallow-well risk zone, and it exposed a spurious rainfall signal arising from spatial correlation. The DWSI placed 53.7% of the population in high or very high priority classes on 44.7% of the area, and these classes accounted for 90.8%*

Surabhi Shivendra, Department of Civil Engineering, Saharsa College of Engineering, Saharsa, Bihar, India- 852201
Pusplata Puja, Department of Civil Engineering, Saharsa College of Engineering, Saharsa, Bihar, India- 852201
Raunak, Department of Civil Engineering, Saharsa College of Engineering, Saharsa, Bihar, India- 852201
Ajay Prabhakar, Department of Civil Engineering, Government Engineering College, Buxar, Bihar, India- 802103
Prakash Kumar Sinha, Department of Civil Engineering, Saharsa College of Engineering, Saharsa, Bihar, India- 852201
*Corresponding author Email: surabhishivendra9431@gmail.com

ISSN: 0262-8104 (print) 1756-3488 (online)
Waterlines Vol. 43 No. 2

of expected exposure. The framework is transparent and transferable to Bihar's monitoring data for well-testing and safe-source planning.

Keywords: *Groundwater contamination; arsenic; fluoride; GIS; explainable machine learning; SHAP; spatial cross-validation; drinking water security; Bihar.*

Introduction

Groundwater supplies most rural drinking water in the Indo-Gangetic plain, but its quality is limited by geogenic contaminants. Arsenic is mobilised by reductive dissolution of iron oxyhydroxides in young, organic-rich alluvial sediments [1], and an estimated hundreds of millions of people worldwide are exposed to concentrations above the World Health Organization guideline of 10 µg/L [2], [3]. Fluoride above 1.5 mg/L is common where groundwater interacts with crystalline and older sedimentary aquifers [4], [5].

Bihar illustrates both problems. Arsenic contamination in the middle Ganga plain was documented more than two decades ago [6]. The state's Public Health Engineering Department (PHED) lists districts affected by arsenic along the Ganga, by fluoride in the south and by iron in the Kosi basin [7]. In a 2024 report, the state government recorded contaminated drinking-water sources in 4,709 arsenic-affected, 3,789 fluoride-affected and 21,709 iron-affected rural wards [8]. Elevated blood-arsenic levels have also been linked with cancer incidence in Gangetic districts [9]. Because testing every hand pump is not feasible, spatial prediction is needed to target testing, safe-source development and piped supply where they matter most.

This paper develops an integrated framework that joins GIS predictor engineering, spatially validated machine learning, SHAP explanations and a population-weighted drinking-water security index (DWSI). Its contributions are: (i) an end-to-end, reproducible workflow from GIS layers to intervention priorities; (ii) a controlled synthetic benchmark with known generating rules, which tests whether explanations recover true processes rather than merely producing accurate predictions; and (iii) a decision layer that translates probabilities into drinking-water exposure. The overall workflow is shown in Figure 1.

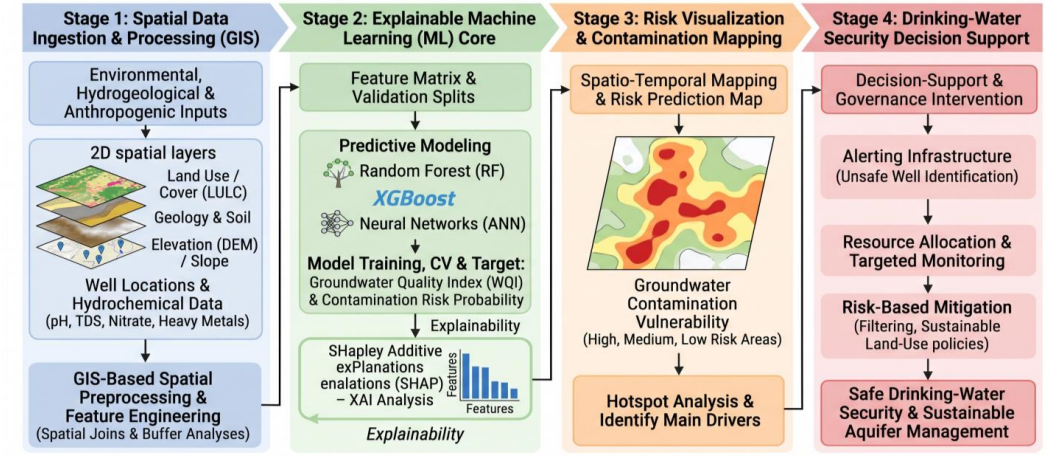


Figure 1. Integrated GIS and explainable machine learning framework for groundwater contamination risk and drinking-water security.

Literature Review

2.1 GIS Overlay and Index Methods

Overlay-index methods such as DRASTIC and GOD weight and sum expert-rated layers, often using the analytic hierarchy process [10]. They are transparent and need few data, but their weights are subjective and they describe intrinsic vulnerability rather than the observed occurrence of specific contaminants [10].

2.2 Machine Learning for Contaminant Prediction

Data-driven models trained on measured well chemistry and spatially continuous predictors have transformed contaminant mapping. Random-forest models have produced predictions for the whole of India [11] and the globe [2]. Comparable models exist for fluoride [4], [5] and nitrate [12]. Random forests [13], gradient boosting [14] and support vector machines [15] are the most common learners. However, random cross-validation of autocorrelated wells inflates accuracy [16], [17], and impurity-based importance is biased for correlated predictors [18].

2.3 Explainable AI for Water-Quality Decisions

High-stakes public-health decisions require models whose reasoning can be inspected [19]. SHAP attributes each prediction to its inputs using Shapley values from cooperative game theory [20], and TreeSHAP computes them exactly for tree ensembles [21]. Applications to groundwater contamination remain limited, and

explanations are seldom checked against known truth. The present framework addresses this gap.

Study Context and Data

3.1 Hydrogeological Setting

The framework targets the alluvial plains of Bihar. Holocene floodplain sediments flank the Ganga and its northern tributaries, older Pleistocene alluvium occupies the interfluves, and the Precambrian hard-rock margin bounds the south. Arsenic is concentrated in shallow Holocene aquifers along the Ganga [6], [9], fluoride in the southern districts and iron in the Kosi basin [7].

3.2 Predictor Layers and Synthetic Benchmark

Table 1 lists the eight GIS predictors used by the framework, their real-data sources for Bihar and the role each plays in the synthetic benchmark. The benchmark reproduces the Bihar setting on a 120 km × 80 km grid at 0.5 km resolution (Figure 2). A meandering main river with a northern tributary is bordered by Holocene alluvium, older alluvium lies between them, and a hard-rock zone forms the southern edge. A total of 1,800 wells were sampled at random locations and depths. Arsenic exceedance (As > 10 µg/L, prevalence 16.9%) followed a known nonlinear rule summarised in Table 1, with a shallow-screen peak at 20–60 m and a sharp decline below 110 m. Fluoride exceedance (F > 1.5 mg/L, prevalence 14.6%) was generated to rise in hard rock and with depth and elevation. Rainfall and depth to water table were included as plausible but inactive controls so that the explanations could be checked.

Table 1. GIS predictor layers, real-data sources for Bihar and role in the synthetic arsenic benchmark.

Predictor	Bihar data source	Hydrogeological rationale	Role in As truth
Geomorphic unit	GSI geology / geomorphology maps	Sediment age controls As release	Strong (+ Holocene)
Distance to river	Drainage network from DEM	Young floodplain deposits	Strong (−)
Well depth	PHED / CGWB well records	Shallow screens most affected	Nonlinear hump
Clay fraction	Soil maps / borelogs	Organic, reducing layers	Threshold × Holocene
GW irrigation	Minor irrigation census	Pumping draws As downward	Interaction × Holocene
Elevation	SRTM DEM	Low-lying recharge zones	Weak (−)
Depth to water table	CGWB monitoring wells	Redox / recharge proxy	None (control)
Rainfall	IMD gridded rainfall	Recharge and dilution	None (control)

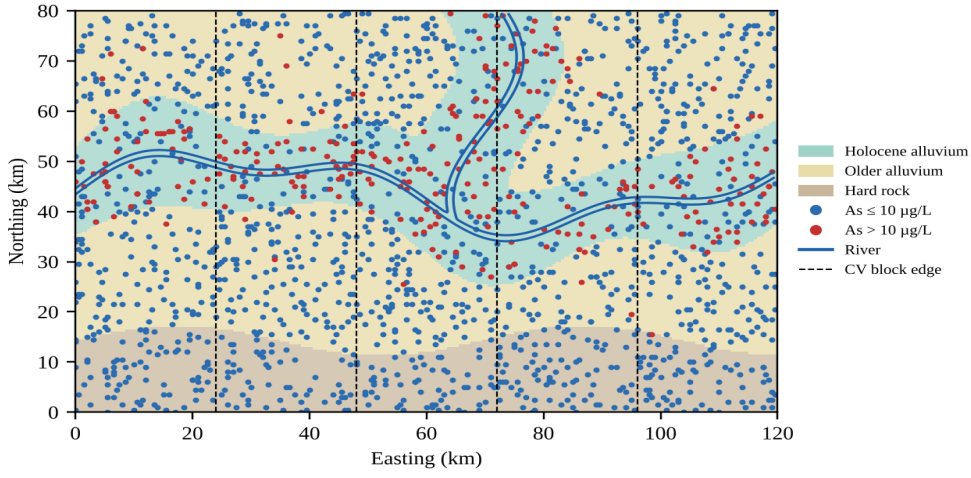


Figure 2. Synthetic Ganga-plain benchmark: geomorphic units, drainage and 1,800 sampled wells coloured by arsenic exceedance, with the five spatial cross-validation blocks.

Methodology

4.1 Framework Architecture

The framework has six stages (Figure 1): GIS data assembly, feature engineering, ML modelling, explainability, risk mapping and drinking-water security assessment. Predictors are stacked on a common grid and extracted to well locations. Collinearity is screened before modelling [18].

4.2 Models and Spatial Validation

Four classifiers were compared: logistic regression (LR), support vector machine with a radial kernel (SVM) [15], random forest (RF, 400 trees) [13] and XGBoost (350 trees, depth 4) [14]. To avoid optimistic bias from spatial autocorrelation [16], [17], performance was estimated with five-fold spatial-block cross-validation, using 24 km longitudinal blocks so that each test block is geographically separate from the training data. Evaluation used the area under the ROC curve (AUC), accuracy, F1-score, Cohen's κ, the Brier score and reliability curves.

4.3 SHAP Explanations

SHAP decomposes a model prediction for well i into a baseline and additive feature contributions [20]:

$$f(x_i) = \varphi_0 + \sum_{j=1}^{M} \varphi_{ij} \quad (1)$$

where each φ is the Shapley value of feature j, averaged over all feature subsets S that exclude j:

$$\varphi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \left[f(S \cup \{j\}) - f(S) \right] \quad (2)$$

Exact TreeSHAP values [21] were computed for the selected model to obtain global importance (mean |φ|), summary plots and dependence plots. These were then compared with the known generating rules.

4.4 Risk Mapping and Drinking-Water Security Index

Contaminant-specific probabilities were predicted on the full grid at the typical domestic hand-pump depth of 40 m. They were combined, assuming conditional independence, into the probability that water from a well exceeds any guideline:

$$P_{any} = 1 - \prod_k (1 - P_k) \quad (3)$$

Drinking-water exposure E weights this probability by the fraction of residents who depend on untreated groundwater, and the DWSI is its complement:

$$E = P_{any} \times D_{gw}, \quad DWSI = 1 - E \quad (4)$$

Cells were classed as Low (E < 0.15), Moderate (0.15–0.35), High (0.35–0.55) or Very high (> 0.55) priority and overlaid on gridded population.

Results and Discussion

5.1 Predictive Performance

Table 2. Spatial-block cross-validated performance for arsenic exceedance (n = 1,800 wells).

Model	AUC	Accuracy	F1-score	Cohen's κ	Brier
Logistic regression	0.929	0.896	0.704	0.641	0.073
SVM (RBF)	0.878	0.897	0.687	0.625	0.075
Random forest	0.925	0.897	0.690	0.628	0.074
XGBoost	0.917	0.894	0.683	0.620	0.079

Table 2 and Figure 3 summarise spatially cross-validated performance for arsenic. LR, RF and XGBoost achieved similar discrimination (AUC 0.917–0.929, κ 0.62–0.64), whereas SVM was clearly weaker (AUC 0.878) because its ROC curve flattens at high sensitivity. Brier scores (0.073–0.079) indicate reasonable calibration. The strong result for LR shows that, when dominant controls are close to monotonic and spatial transfer is tested honestly, model complexity must be justified by validated gains rather than assumed. RF was retained for explanation and mapping because it equalled the best AUC within 0.004 while representing the nonlinearities and interactions that SHAP is designed to reveal. The RF fluoride model reached an AUC of 0.870.

5.2 Explainability and Recovery of Known Processes

The SHAP summary (Figure 4a) ranks clay fraction (mean $|\varphi| = 0.078$), geomorphic unit (0.057), distance to river (0.052) and well depth (0.052) as the leading drivers, which matches the strongest imposed controls. The direction of each effect is also correct: high clay content and Holocene sediment raise risk, while greater distance from the river lowers it. The well-depth dependence plot (Figure 4b) recovers the imposed nonlinearity, with positive contributions in the 20–60 m shallow zone, a sharp decline below about 60–110 m and the largest effects in Holocene wells. This agrees with the low arsenic typically found in deeper, older aquifers of the Ganga–Brahmaputra basin [1].

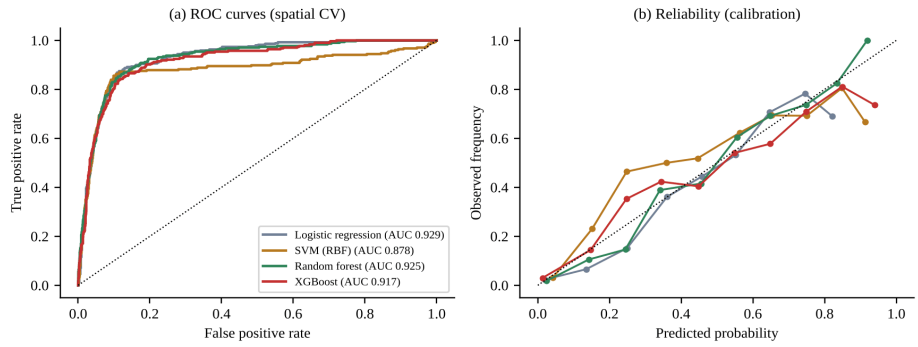


Figure 3. (a) ROC curves and (b) reliability diagrams for the four classifiers under spatial-block cross-validation.

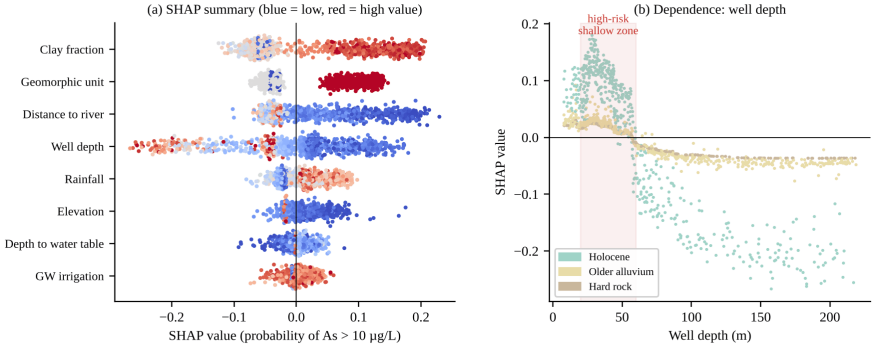


Figure 4. SHAP explanation of the random-forest arsenic model: (a) summary plot of feature contributions; (b) dependence of SHAP value on well depth, coloured by geomorphic unit.

The explanations also reveal two cautions. First, rainfall, which plays no part in the arsenic rule, received a non-zero importance (0.025) because rainfall increases northward and therefore acts as a proxy for position relative to the river. Second, groundwater irrigation, a real but interaction-only driver, received low importance (0.008) because its effect is shared with correlated Holocene variables. SHAP explains the model, not the aquifer [18], [19], so explanations must be checked against hydrogeochemical knowledge before they justify interventions.

5.3 Risk Maps and Drinking-Water Security

The predicted arsenic surface (Figure 5a) traces the Holocene corridors along both rivers, whereas fluoride risk (Figure 5b) is confined to the hard-rock south. This reproduces the north–south contrast in contaminants reported for Bihar [7].

Combining the two with groundwater dependence gives the priority classes in Figure 5c. In the 8.94 million-resident domain, the Low, Moderate, High and Very high classes cover 48.9%, 6.4%, 17.1% and 27.6% of the area and hold 5.6%, 3.6%, 24.7% and 66.1% of expected exposure, respectively. High and Very high classes thus contain 53.7% of residents and 90.8% of expected exposure on 44.7% of the area, because towns on the floodplain coincide with the highest arsenic probabilities. The population-weighted mean DWSI was 0.645. Where piped supply reduces groundwater dependence, exposure falls despite high hazard, so hazard maps alone would misallocate resources.

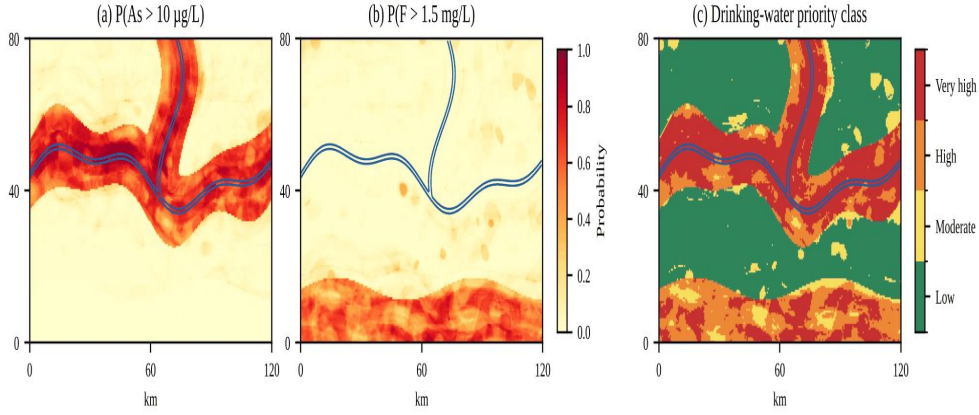


Figure 5. Predicted probability of (a) arsenic and (b) fluoride exceedance at 40 m well depth, and (c) drinking-water priority classes from the DWSI.

5.4 Implications for Bihar

Operationally, wells in High and Very high classes should be tested first. SHAP dependence plots can guide safe-depth advice, such as deeper screens in Holocene arsenic zones but not in fluoride zones where risk rises with depth, and the DWSI can rank villages for surface-water-based piped supply. Because every prediction carries an explanation, district engineers can see why a habitation is flagged, which builds trust and exposes errors [19].

Conclusion

This study presented an integrated GIS and explainable machine learning framework that links contaminant prediction to drinking-water security. On a controlled synthetic Ganga-plain benchmark, spatially cross-validated models achieved AUC values of up to 0.929 for arsenic and 0.870 for fluoride. SHAP recovered the imposed

sedimentological, proximity and depth controls, and it exposed a spurious proxy variable. The DWSI focused 90.8% of expected exposure into the High and Very high classes, which cover less than half of the area. Accuracy, honest spatial validation and verified explanations must therefore be achieved together if risk maps are to guide public-health investment. The main limitation is that the results come from a synthetic benchmark designed to verify the method. They are not estimates for any real district. The next step is to apply the framework to PHED and CGWB well-chemistry records for Bihar, add iron and nitrate, and quantify uncertainty and the area of applicability [17].

References

S. Fendorf, H. A. Michael, and A. van Geen, "Spatial and temporal variations of groundwater arsenic in South and Southeast Asia," *Science*, vol. 328, no. 5982, pp. 1123–1127, 2010, doi: 10.1126/science.1172974.

J. Podgorski and M. Berg, "Global threat of arsenic in groundwater," *Science*, vol. 368, no. 6493, pp. 845–850, 2020, doi: 10.1126/science.aba1510.

World Health Organization, *Guidelines for Drinking-Water Quality*, 4th ed. incorporating the first and second addenda. Geneva, Switzerland: WHO, 2022. [Online]. Available: https://www.who.int/publications/i/item/9789240045064

J. E. Podgorski, P. Labhasetwar, D. Saha, and M. Berg, "Prediction modeling and mapping of groundwater fluoride contamination throughout India," *Environmental Science & Technology*, vol. 52, no. 17, pp. 9889–9898, 2018, doi: 10.1021/acs.est.8b01679.

J. Podgorski and M. Berg, "Global analysis and prediction of fluoride in groundwater," *Nature Communications*, vol. 13, art. 4232, 2022, doi: 10.1038/s41467-022-31940-x.

D. Chakraborti *et al.*, "Arsenic groundwater contamination in Middle Ganga Plain, Bihar, India: A future danger?," *Environmental Health Perspectives*, vol. 111, no. 9, pp. 1194–1201, 2003, doi: 10.1289/ehp.5966.

Public Health Engineering Department, Government of Bihar, "Water quality affected districts," PHED Bihar. Accessed: Oct. 31, 2026. [Online]. Available: https://phedbihar.gov.in/WaterQuality.aspx

Government of Bihar, "Report on behalf of Government of Bihar in O.A. No. 728 of 2023," National Green Tribunal, New Delhi, India, 2024. [Online]. Available: https://www.greentribunal.gov.in/ (Report, O.A. No. 728 of 2023, PDF)

A. Kumar *et al.*, "Arsenic exposure in Indo Gangetic plains of Bihar causing increased cancer risk," *Scientific Reports*, vol. 11, art. 2376, 2021, doi: 10.1038/s41598-021-81579-9.

D. Machiwal, M. K. Jha, V. P. Singh, and C. Mohan, "Assessment and mapping of groundwater vulnerability to pollution: Current status and challenges," *Earth-Science Reviews*, vol. 185, pp. 901–927, 2018, doi: 10.1016/j.earscirev.2018.08.009.

J. Podgorski, R. Wu, B. Chakravorty, and D. A. Polya, "Groundwater arsenic distribution in India by machine learning geospatial modeling," *International Journal of Environmental Research and Public Health*, vol. 17, no. 19, art. 7119, 2020, doi: 10.3390/ijerph17197119.

V. Rodriguez-Galiano, M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo, and L. Ribeiro, "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain)," *Science of the Total Environment*, vol. 476–477, pp. 189–206, 2014, doi: 10.1016/j.scitotenv.2014.01.001.

L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.

H. Meyer and E. Pebesma, "Predicting into unknown space? Estimating the area of applicability of spatial prediction models," *Methods in Ecology and Evolution*, vol. 12, no. 9, pp. 1620–1633, 2021, doi: 10.1111/2041-210X.13650.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, art. 307, 2008, doi: 10.1186/1471-2105-9-307.

C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.

S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.

S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.