

Spatial Benchmarking of Four Spatially Explicit Models for Groundwater Potential Mapping in Cheongju, South Korea

Sooyeon Han

KIGAM: Korea Institute of Geoscience and Mineral Resources

Saro Lee

leesaro@kigam.re.kr

KIGAM: Korea Institute of Geoscience and Mineral Resources <https://orcid.org/0000-0003-0409-8263>

Research Article

Keywords: Groundwater reconnaissance, Relative well-occurrence favorability, Spatial cross-validation, Model comparison, Explainable artificial intelligence, Fold-refit dispersion

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10626377/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Spatial Benchmarking of Four Spatially Explicit Models for Groundwater Potential Mapping in Cheongju, South Korea

Sooyeon Han^a, Saro Lee^{a,*}

^a Geoscience Data Research Department, Korea Institute of Geoscience and Mineral Resources (KIGAM), 124 Gwahak-ro, Yuseong-gu, Daejeon 34132, Republic of Korea

* Corresponding author: Saro Lee; e-mail: leesaro@kigam.re.kr

ORCID: Sooyeon Han, 0000-0001-7934-353X; Saro Lee, 0000-0003-0409-8263

Abstract

Groundwater reconnaissance maps can prioritize field verification where observations are sparse, but spatial clustering and unverified background samples can make small model differences appear more certain than they are. This study compared four spatial model structures for relative well-occurrence favorability in Cheongju, South Korea: geographical random forest (GRF), k-nearest-neighbor label-lag logistic regression (kNN-lag), coordinate-augmented extreme gradient boosting (Coord-XGBoost), and a spatial Gaussian process classifier (SGP). The analytical sample comprised 207 georeferenced positive well locations and 207 sampled background points. Because a common pumping-rate or specific-capacity threshold was unavailable, outputs were interpreted as conditional rankings rather than aquifer-yield probabilities. Thirteen conditioning factors on a 10 m grid were evaluated using a shared five-fold spatial-block partition and a secondary stratified hold-out set. Mean spatial-partition area under the receiver operating characteristic curve (AUC) ranged from 0.850 to 0.871, and hold-out AUC ranged from 0.858 to 0.875. Pairwise DeLong tests detected no hold-out differences (minimum $p = 0.296$), and estimated power to detect an AUC difference of 0.03 was 0.312. Fixed-configuration factors-only variants differed from full models by 0.000-0.010 in mean AUC, whereas spatial-only variants yielded 0.538-0.640. Model-specific attribution and fold-refit dispersion provided

complementary information about recurrent factor associations and within-model score stability. The outputs support staged groundwater reconnaissance and monitoring, not estimates of yield, sustainable abstraction, groundwater quality, or drilling success.

Keywords: Groundwater reconnaissance · Relative well-occurrence favorability · Spatial cross-validation · Model comparison · Explainable artificial intelligence · Fold-refit dispersion

Supplementary Information: Online Resource 1 contains Figs. S1-S4 and Tables S1-S9

1 Introduction

Groundwater assessment is a practical component of water-resource planning where direct aquifer observations are sparse and field investigation is costly. Reconnaissance maps can order candidate locations for hydrogeologic surveys, monitoring, and follow-up pumping tests, but their management value depends on whether mapped rankings remain credible when predictions are transferred across space. This requirement is especially important in heterogeneous urban and peri-urban settings, where clustered well records can reflect both hydrogeologic opportunity and the accessibility, land use, and reporting processes that determine where wells are installed and documented.

Recent studies in Water Resources Management demonstrate the range of quantitative groundwater decision support. Ensemble and optimized machine-learning models have been applied to groundwater-potential prediction (Mosavi et al. 2021; Mouhoumed et al. 2026), multi-criteria methods to recharge-priority delineation (Akbari et al. 2021), machine learning to groundwater-quality classification (Zegaar et al. 2024), and ensemble optimization to groundwater-level forecasting (Saqr et al. 2025). Distributed drought characterization (Christelis et al. 2025), collaborative local-scale monitoring (Reading et al. 2025), and sensitivity analysis of coupled groundwater-management models (Rodríguez-Flores et al. 2022) further show that accuracy alone is insufficient; validation design, monitoring context, and a clear definition of the predicted quantity are also required.

48 The term groundwater potential can be ambiguous when a response dataset does not contain a common
49 yield threshold or verified dry wells (Bradbury 2025). A binary well inventory may represent the spatial
50 occurrence of documented wells rather than aquifer productivity. Under such conditions, model outputs
51 are most appropriately treated as relative, sampling-design-dependent indicators for reconnaissance,
52 not population probabilities, drilling-success probabilities, or sustainable-abstraction estimates. This
53 distinction is central here because the analytical inventory contains positive well locations but no
54 uniform pumping-rate or specific-capacity criterion for all 207 records.

55 Two additional issues affect the credibility of spatial ranking maps. Random partitions can place nearby
56 records in training and evaluation sets, allowing spatial autocorrelation to reward interpolation rather
57 than transfer to geographically withheld areas. Spatial blocking can reduce this concern, although a
58 single unbuffered partition is a sensitivity assessment rather than a strict test of fixed-distance
59 extrapolation (Roberts et al. 2017; Schratz et al. 2019). In addition, small AUC differences should not
60 be interpreted as model superiority without paired comparison, consideration of power, and assessment
61 of design sensitivity. Explainable artificial intelligence can assist interpretation, but attributions from
62 different fitted objects, explainers, samples, and output scales are not directly interchangeable
63 (Lundberg and Lee 2017; Lundberg et al. 2020; Maier et al. 2024; Zounemat-Kermani and Kheimi
64 2026).

65 Cheongju has previously been mapped using a frequency-ratio approach (Lee 2025). A record-level
66 crosswalk with that inventory was unavailable, so the present study does not claim an independent well
67 dataset. Its contribution is a bounded comparison framework: four contrasting spatial model structures
68 are evaluated with common predictors and fold identifiers; environmental and geographic components
69 are examined using fixed-configuration variants; secondary hold-out AUCs are compared formally;
70 and fold-refit dispersion is mapped as a within-model stability diagnostic. The objectives were to
71 compare discrimination under one transparent spatial partition, evaluate the evidential strength of small

hold-out differences, describe model-specific environmental attribution, examine the relative contribution of environmental and geographic components, and combine relative favorability with refit stability for staged groundwater reconnaissance.

2 Study Area and Data

2.1 Study area and management context

Cheongju is located in North Chungcheong Province in the central interior of the Republic of Korea (Fig. 1). Broad alluvial plains, fluvial terraces, and low hills occupy much of the west, whereas the east rises into steep, dissected mountain terrain. Summer monsoon rainfall and the west-to-east physiographic gradient influence runoff, infiltration opportunity, land use, access, and the distribution of documented wells. The resulting contrast provides a useful setting for examining whether relations learned in densely sampled lowlands remain informative in geographically different parts of the municipality. The maps are intended to prioritize reconnaissance and additional observations rather than replace site-specific hydrogeologic investigation.

2.2 Target construct and well inventory

The analytical response comprised 207 georeferenced positive well locations. The inventory did not contain a uniform pumping-rate or specific-capacity threshold applicable to all records. The target was therefore defined as relative well-occurrence favorability: the degree to which a location resembles the environmental and geographic conditions of the positive inventory under the selected case-background design. It was not interpreted as direct aquifer yield, sustainable groundwater availability, or drilling success.

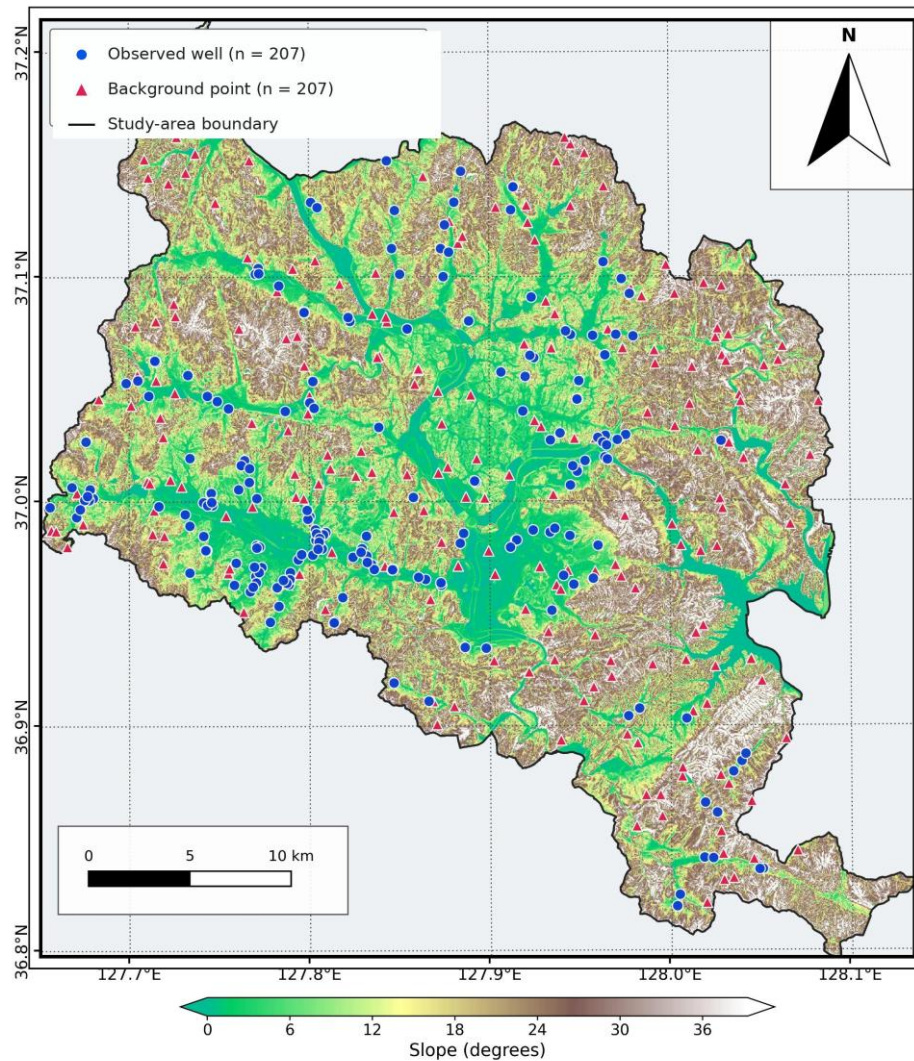


Fig. 1 Cheongju study area, slope, and analytical inventory. Blue circles show 207 positive well locations, and red triangles show 207 sampled background points. The background points are not verified dry or low-yield wells

Regional hydrogeologic and pumping-test context was checked against the Groundwater Basic Survey Report for the Cheongwon (Cheongju) Area and its appendices (Ministry of Construction and Transportation and Korea Water Resources Corporation 2006). The report documents 160 well records, including 148 pumping-rate values and 138 records from which specific capacity could be calculated; its screened summaries used 144 pumping-rate and 124 specific-capacity records. These totals do not

correspond directly to the 207 modeled locations, and a record-level crosswalk was not available. The report subset was therefore not used as a continuous validation dataset, and no productivity threshold was assigned to the analytical inventory.

2.3 Background sampling and data partitioning

An equal number of background points was generated by rejection sampling with a NumPy PCG64 generator seeded at 42. Uniform candidates were retained only when they fell inside the Cheongju administrative boundary, within the valid factor-stack extent, and more than 60 m from every positive well. Coordinates were sampled continuously. The realized minimum background-to-background distance was 66.7 m, and the minimum positive-well-to-background distance was 79.9 m. No water-body or land-use exclusion mask was applied. These locations are background points, not verified dry or low-yield wells; any threshold-dependent background-rejection metric refers only to this sampled design.

The 414 observations were divided by stratified random sampling with seed 42 into 290 development points and 124 secondary hold-out points, including 62 positive records. Because this split retained close development-to-evaluation neighbors, its formal pairwise results were treated as secondary. The development set was also divided into five geographic folds as described in Sect. 3.3. Continuous-factor centering and scaling were based on the relevant development data. None of the conditioning factors was interpolated from the labeled well network.

2.4 Conditioning factors and preprocessing

Fifteen candidate factors were compiled from co-registered terrain, hydrologic, soil, geological, land-cover, and forest layers on a common 10 m grid in the Korea 2000 Central Belt projection (EPSG:5186). The 13 retained factors are summarized in Table S1 and mapped in Fig. S1. Spearman rank correlation was used to examine redundancy. Terrain ruggedness and slope-length factor were

removed because their correlations with slope were $\rho = 0.99$ and 0.98 , respectively; no retained pair exceeded $|\rho| = 0.80$ (Fig. S2).

The retained continuous variables were standardized to zero mean and unit variance, with catchment area and flow-path length log-transformed first. Categorical variables were one-hot encoded; one reference category per factor was omitted, and classes occupying less than 2% of the study area were combined into an Other category. The resulting design matrix contained nine continuous columns and 34 indicator columns. Dataset-provider, acquisition-year, native-resolution, and licensing metadata were not available in the retained analytical record and require completion before final submission.

3 Methodology

3.1 Analytical framework

The workflow used a common factor matrix and common data partitions to compare four implemented geographic mechanisms. Full models were evaluated under the shared spatial partition and the secondary hold-out set; model-specific environmental attribution, fixed-configuration component variants, combined design sensitivity, and fold-refit dispersion were then used to describe how the fitted surfaces differed. No result was interpreted as an estimate of groundwater yield or regional well prevalence.

3.2 Model structures and component definitions

Geographical random forest (GRF). GRF was used to examine whether location-specific fitting represented spatially nonstationary associations between the 13 factors and the positive inventory. For each prediction location, a local random forest was fitted from its k nearest development points with Gaussian geographic weighting (Breiman 2001; Georganos et al. 2021). The recorded selected configuration used $k = 120$, 200 trees, and the square root of p candidate features at each split, where p is the number of input columns. The retained record did not include the Gaussian bandwidth definition

148 or a complete execution specification; these items are identified in the revision report for author
149 confirmation.

150 k-nearest-neighbor label-lag logistic regression (kNN-lag). This model augmented L2-regularized
151 logistic regression with the mean label of the k nearest geographic training neighbors. For location i ,
152 the lag was defined as

$$L_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} y_j, \quad (1)$$

153 where L_i is the dimensionless label-lag score; k is the number of neighbors (count); $\mathcal{N}_k(i)$ is the set of
154 the k nearest training locations to i , identified using projected coordinates in metres; and y_j is the binary
155 training label of neighbor j . The positive-class score was modeled as

$$\text{logit}(p_i) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \gamma L_i, \quad (2)$$

156 where p_i is the dimensionless case-background score; β_0 is the intercept; $\mathbf{x}_i$ is the environmental
157 predictor vector composed of standardized continuous variables and categorical indicators; $\boldsymbol{\beta}$ is its
158 coefficient vector; γ is the label-lag coefficient; and $\text{logit}(p_i) = \ln[p_i/(1 - p_i)]$. Within each fold, neighbor
159 searches used training points only, validation labels never entered the lag, and each training point
160 excluded its own label. The selected configuration used $C = 0.1$, $k = 8$, L2 regularization, and the lbfgs
161 optimizer.

162 Coordinate-augmented XGBoost (Coord-XGBoost). Projected easting and northing were appended to
163 the environmental design matrix before fitting a global extreme gradient boosting classifier (Chen and
164 Guestrin 2016). This structure allows nonlinear coordinate interactions but does not fit a separate local
165 model. The recorded selected configuration used 200 trees, a learning rate of 0.03, maximum depth of
166 5, row subsampling of 0.8, and column subsampling of 0.8.

167 Spatial Gaussian process classifier (SGP). The SGP represented smooth dependence in the combined
168 standardized environmental and coordinate space. Its latent formulation was

$$f(\mathbf{z}) \sim \mathcal{GP}[0, k_{\text{RBF}}], p_i = \sigma[f(\mathbf{z}_i)], \quad (3)$$

169 where $\mathbf{z}_i$ is the standardized environmental-and-coordinate vector for location i ; f is a dimensionless
170 latent function; k_{RBF} is the radial basis function covariance; σ is the logistic link; and p_i is the
171 dimensionless case-background score. Kernel parameters were estimated by marginal-likelihood
172 optimization. The recorded selected initial length scale was 0.25 with one optimizer restart. The fitted
173 kernel parameters and convergence diagnostics were not retained in the manuscript record and require
174 author confirmation. This implementation is a point-estimate classifier rather than a fully Bayesian
175 hierarchical groundwater model (Rasmussen and Williams 2006).

176

177 **Table 1** Implemented model structures, recorded tuning information, and fixed-configuration
178 component variants

Model	Implemented geographic mechanism	Recorded search or selected configuration	Factors-only variant	Spatial-only variant
GRF	Location-specific random forest with nearest-neighbor geographic weighting	$k \in \{60, 80, 120\}$; trees $\in \{100, 200\}$; selected $k = 120$ and 200 trees	Global random forest using environmental factors	Geographic/local mechanism without environmental factors, as implemented
kNN-lag	Mean label of k nearest training neighbors appended to logistic regression	$C \in \{0.1, 0.3, 1, 3, 10\}$; $k \in \{4, 8, 12\}$; selected $C = 0.1$ and $k = 8$	Logistic regression without label lag	Label-lag term without environmental factors
Coord-XGBoost	Projected coordinates appended to the environmental matrix	Recorded selected configuration: 200 trees; learning rate 0.03; depth 5; row/column subsampling 0.8	XGBoost without coordinates	XGBoost using coordinates only
SGP	RBF covariance over standardized environmental predictors and coordinates	Initial length scale $\in \{0.25, 0.5, 1.0\}$; restarts $\in \{1, 2\}$; selected initial length scale 0.25 and one restart	Gaussian process without coordinates	Gaussian process using coordinates only

3.3 Shared spatial-partition sensitivity and secondary hold-out comparison

The development-area bounding box was tessellated into a 6×6 lattice. Occupied block centroids were assigned to five folds by k-means clustering with seed 42, and identical fold identifiers were used for all models and component variants. The cells were 6,253.0 m east-west by 6,960.0 m north-south (43.5 km²); 30 of 36 cells were occupied. Folds contained 36-69 points and 14-44 positive locations. Each fold was contiguous under rook and queen adjacency, but no buffer was imposed and the minimum distance between points in different folds was 213 m (Fig. 2; Table S2).

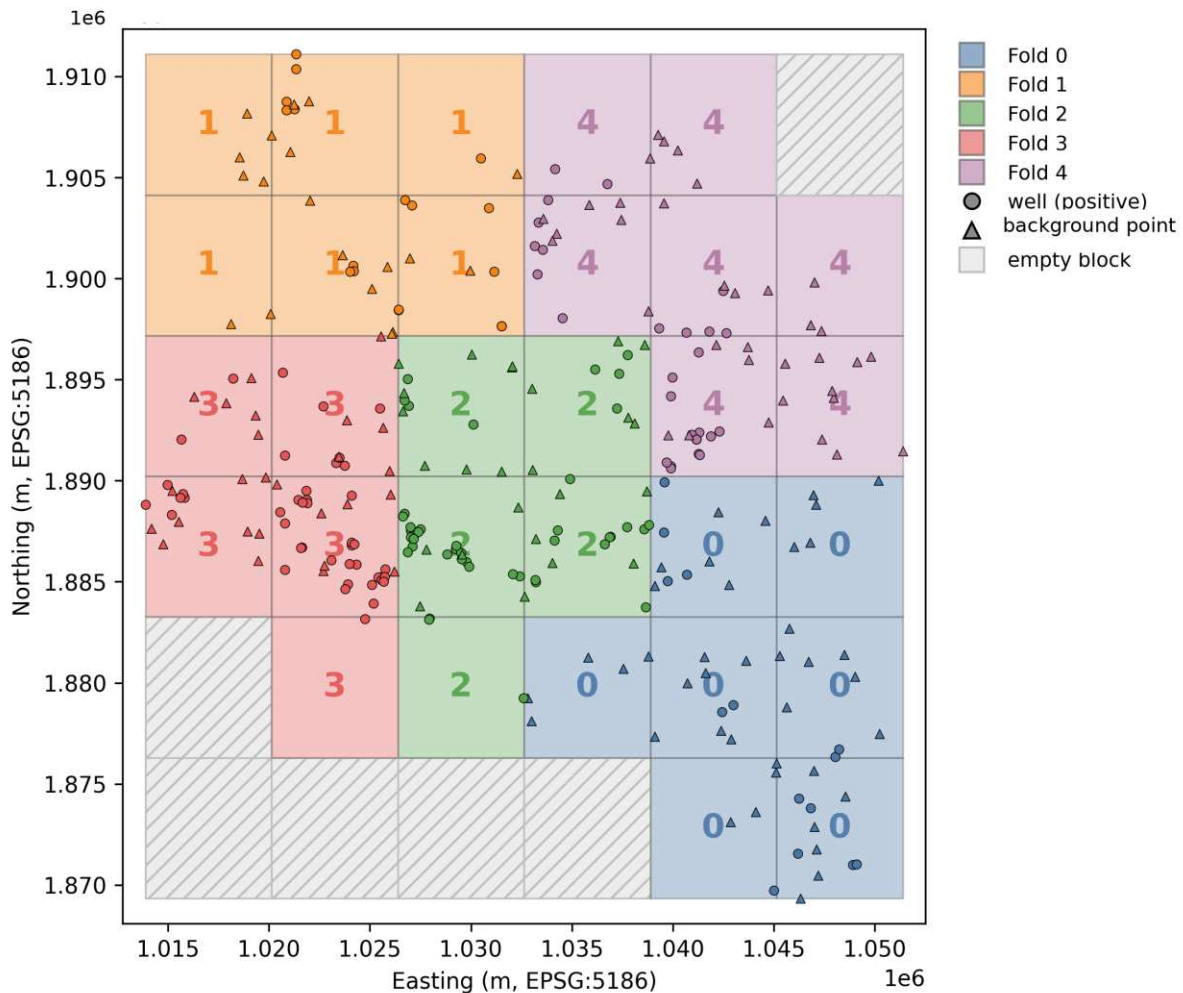


Fig. 2 Five-fold spatial-block partition of the 290 development points. Colored cells are occupied blocks labeled by fold, hatched cells contain no development point, circles show positive wells, and triangles show background points

Hyperparameters were selected using the same five-fold system later used to summarize mean AUC, sample standard deviation, and pooled out-of-fold (OOF) AUC. The resulting estimates are therefore interpreted as a common spatial-partition sensitivity assessment, not as an unbiased nested estimate or a strict fixed-distance extrapolation test (Roberts et al. 2017; Schratz et al. 2019). No formal pairwise test was applied to blocked OOF predictions.

After model selection, the 124-point secondary hold-out set was scored. Its median evaluation-to-development nearest-neighbor distance was 739 m, so it was not spatially independent. Hold-out AUC confidence intervals and pairwise comparisons used the DeLong method (DeLong et al. 1988). Two one-sided tests (TOST) and a post-hoc power calculation were applied only to the widest-separated AUC pair. Equivalence margins of 0.05 and 0.10 were exploratory sensitivity bounds, not prespecified evidence that all models were interchangeable.

3.4 Model-specific environmental attribution

Shapley additive explanations (SHAP) were used to summarize environmental contributions within each fitted explanation object (Lundberg and Lee 2017; Lundberg et al. 2020). Mean absolute SHAP values were calculated at the encoded-column level, summed within each parent categorical factor, and normalized within each model. Coordinate columns and the kNN label-lag column were excluded. The explainer, background data, evaluation sample, and seed are documented in Table S3.

The explanation objects and samples differed by model. TreeExplainer was applied to Coord-XGBoost. Because the location-specific GRF forests do not form one global fitted object, its panel represents a 300-tree global random-forest surrogate. LinearExplainer was applied to kNN-lag. KernelExplainer was applied to SGP using 40 development rows as background and the first 60 hold-out points as the evaluation subset, whereas the other panels used 124 hold-out points. The attributions were therefore interpreted qualitatively within each panel and not as causal effects or directly comparable percentages across model families.

3.5 Fixed-configuration variants, fold-refit dispersion, and combined design sensitivity

Component variants used the recorded selected configurations and the same fold identifiers. Factors-only variants removed each model's implemented geographic input or mechanism while retaining environmental predictors. Spatial-only variants retained the implemented geographic component while removing environmental predictors. Because the variants were not independently retuned and the GRF factors-only model is a global random forest rather than a local GRF with one component removed, these results are described as fixed-configuration, model-specific diagnostics rather than equivalent causal ablations.

Each model was refitted across the five folds and applied to the analysis grid. For model m and grid cell s , fold-refit dispersion was calculated as

$$D_m(s) = \sqrt{\frac{1}{K} \sum_{r=1}^K [\hat{p}_{m,r}(s) - \bar{p}_m(s)]^2}, \quad (4)$$

where $D_m(s)$ is dimensionless fold-refit dispersion; $K = 5$ is the number of fold-wise refits; $\hat{p}_{m,r}(s)$ is the dimensionless score at cell s from refit r ; and $\bar{p}_m(s)$ is the mean across refits. The population divisor was used. $D_m(s)$ is a within-model refit-instability index, not calibrated posterior uncertainty or predictive reliability. Because model families can differ in score range and calibration, raw dispersion magnitudes were not used to rank models.

A combined design-sensitivity analysis repeated background sampling, the stratified development-hold-out split, and continuous-factor standardization for 15 seeds (1000-1014) while retaining the main-analysis hyperparameters. The main realization sampled continuous coordinates, whereas the recorded sensitivity implementation sampled the 1,090,365 valid 10 m cells (Table S4). This 10 m sampling frame differs from the 10 m factor grid and should be verified by the authors before submission. The resulting spread reflects combined sensitivity to background selection, re-splitting,

235 and re-standardization; it does not isolate background uncertainty, rerun spatial-block cross-validation,
 236 or propagate hyperparameter-selection uncertainty.

237 4 Results

238 4.1 Discrimination and evidential strength

239 Under the common spatial partition, mean AUC ranged from 0.850 to 0.871 (Table 2). GRF had the
 240 highest nominal mean, but the between-model range of 0.021 was smaller than the fold-to-fold standard
 241 deviations of 0.040-0.054. Pooled OOF AUC ranged from 0.852 to 0.873, and fold-specific results are
 242 provided in Table S5. These values describe sensitivity to the selected unbuffered partition and do not
 243 establish a spatially independent performance ranking.

244
 245 **Table 2** Discrimination under the common spatial partition and secondary hold-out evaluation. AUC,
 246 area under the receiver operating characteristic curve; SD, standard deviation; OOF, out-of-fold; CI,
 247 confidence interval

Model	Spatial-partition AUC, mean ± SD	Pooled OOF AUC	Secondary hold-out AUC (95% CI)
GRF	0.871 ± 0.040	0.873	0.858 (0.791-0.925)
SGP	0.850 ± 0.054	0.855	0.875 (0.816-0.935)
kNN-lag	0.854 ± 0.051	0.858	0.858 (0.792-0.925)
Coord-XGBoost	0.850 ± 0.040	0.852	0.875 (0.816-0.934)

248
 249 On the less spatially stringent hold-out set, AUC ranged from 0.858 to 0.875 with overlapping 95%
 250 confidence intervals (Fig. S3). Table S6 reports threshold-dependent case-background diagnostics
 251 rather than population decision thresholds. Pairwise DeLong tests detected no differences ($p = 0.296$ -
 252 1.000 ; Table S7). For the widest-separated pair, SGP versus GRF, the AUC difference was 0.01717,
 253 the paired standard error was 0.02041, and power to detect a true difference of 0.03 was 0.312. For this
 254 pair only, TOST gave $p = 0.0538$ at a margin of 0.05 and $p = 2.46 \times 10^{-5}$ at a margin of 0.10 (Table

255 S8). Thus, narrow equivalence was not established, and the wider exploratory result does not imply
256 interchangeability among all models.

257 **4.2 Combined design sensitivity**

258 Across the 15 combined design realizations, mean hold-out AUC ranged from 0.853 to 0.873, with
259 standard deviations of 0.029-0.036 and coefficients of variation below 4.3% (Table S9; Fig. S4).
260 Because each realization changed the background sample, split, standardization, and sampling frame,
261 this spread is interpreted as combined design sensitivity rather than isolated background-sample
262 robustness.

263 **4.3 Model-specific environmental attribution**

264 Within-model environmental attribution placed slope among the two leading factors in all four panels
265 (Fig. 3). Forest type ranked first in the GRF surrogate, while slope ranked first for SGP, kNN-lag, and
266 Coord-XGBoost. Soil-texture index, slope height, land cover, and forest type also appeared
267 prominently.

268 The panels must be read separately. Panel (a) explains the global random-forest surrogate, panel (b)
269 the fitted SGP on its reduced evaluation subset, panel (c) kNN-lag, and panel (d) Coord-XGBoost.
270 Different explainers and output representations were used. The recurring rankings therefore support
271 qualitative identification of factor associations within the fitted objects, not numerical comparison of
272 percentages or causal hydrogeologic interpretation.

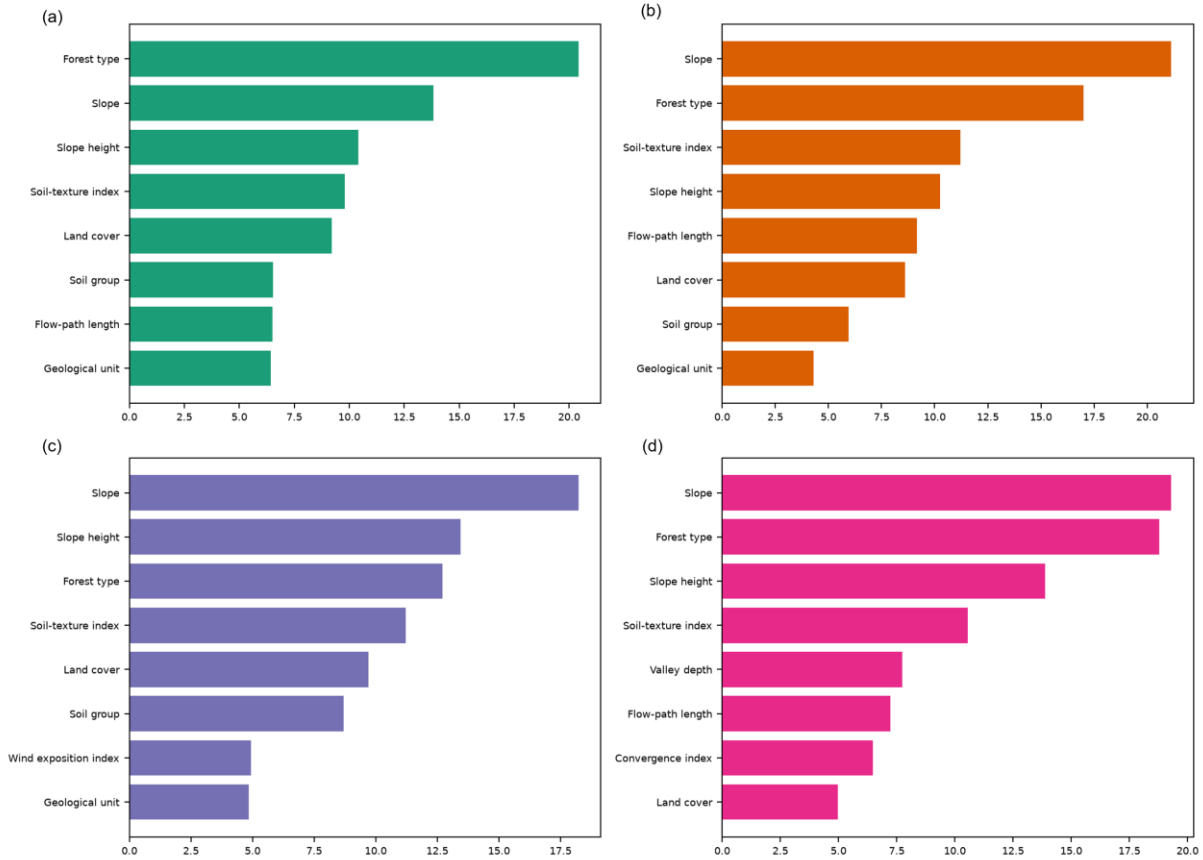


Fig. 3 Top eight within-model environmental shares of mean absolute SHAP attribution: (a) global random-forest surrogate for GRF, (b) SGP, (c) kNN-lag, and (d) Coord-XGBoost. Coordinate and label-lag inputs are excluded. Bar magnitudes are normalized within panels and are not directly comparable among model families

4.4 Fixed-configuration component variants

Environmental predictors retained nearly all blocked discrimination when the implemented geographic component was removed (Table 3). Factors-only variants changed mean AUC by 0.000-0.010 relative to their full models, whereas spatial-only variants yielded mean AUC of 0.538-0.640. Under this fold system and the recorded fixed configurations, the shared environmental factor stack accounted for most discrimination transferred among folds. The variants are descriptive and do not establish that a simpler model would outperform after independent retuning or external validation.

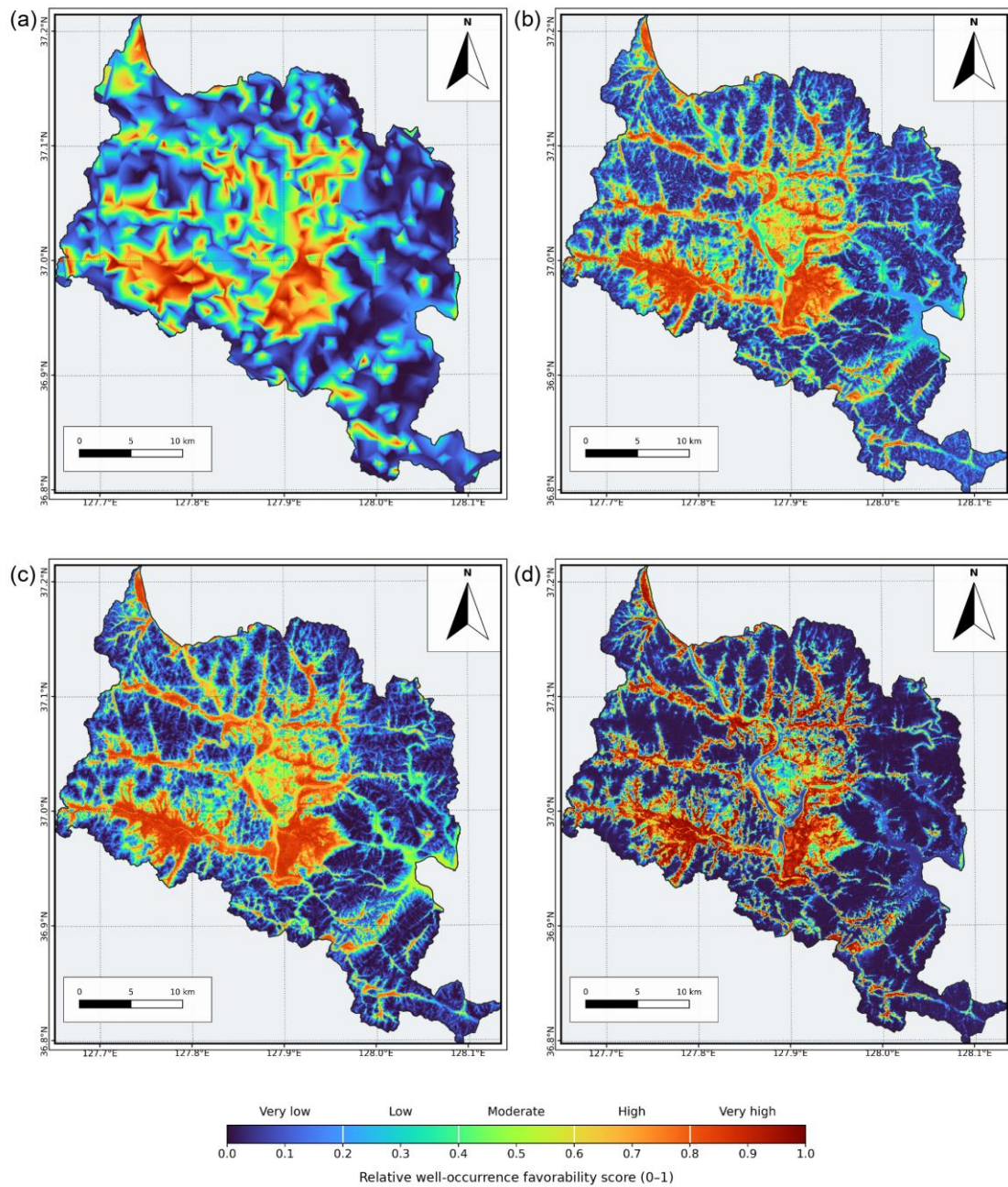
285 **Table 3** Fixed-configuration component variants under the common spatial partition. Δ AUC is variant
286 minus full-model mean AUC

Model	Variant	Mean AUC	SD	Pooled OOF AUC	Δ AUC
GRF	Full	0.871	0.040	0.873	-
GRF	Factors only	0.881	0.028	0.884	0.010
GRF	Spatial only	0.538	0.116	0.520	-0.333
SGP	Full	0.850	0.054	0.855	-
SGP	Factors only	0.852	0.051	0.860	0.002
SGP	Spatial only	0.553	0.112	0.621	-0.297
kNN-lag	Full	0.854	0.051	0.858	-
kNN-lag	Factors only	0.853	0.051	0.858	0.000
kNN-lag	Spatial only	0.640	0.111	0.537	-0.214
Coord-XGBoost	Full	0.850	0.040	0.852	-
Coord-XGBoost	Factors only	0.858	0.040	0.862	0.008
Coord-XGBoost	Spatial only	0.550	0.085	0.569	-0.300

287

288 **4.5 Favorability surfaces and fold-refit dispersion**

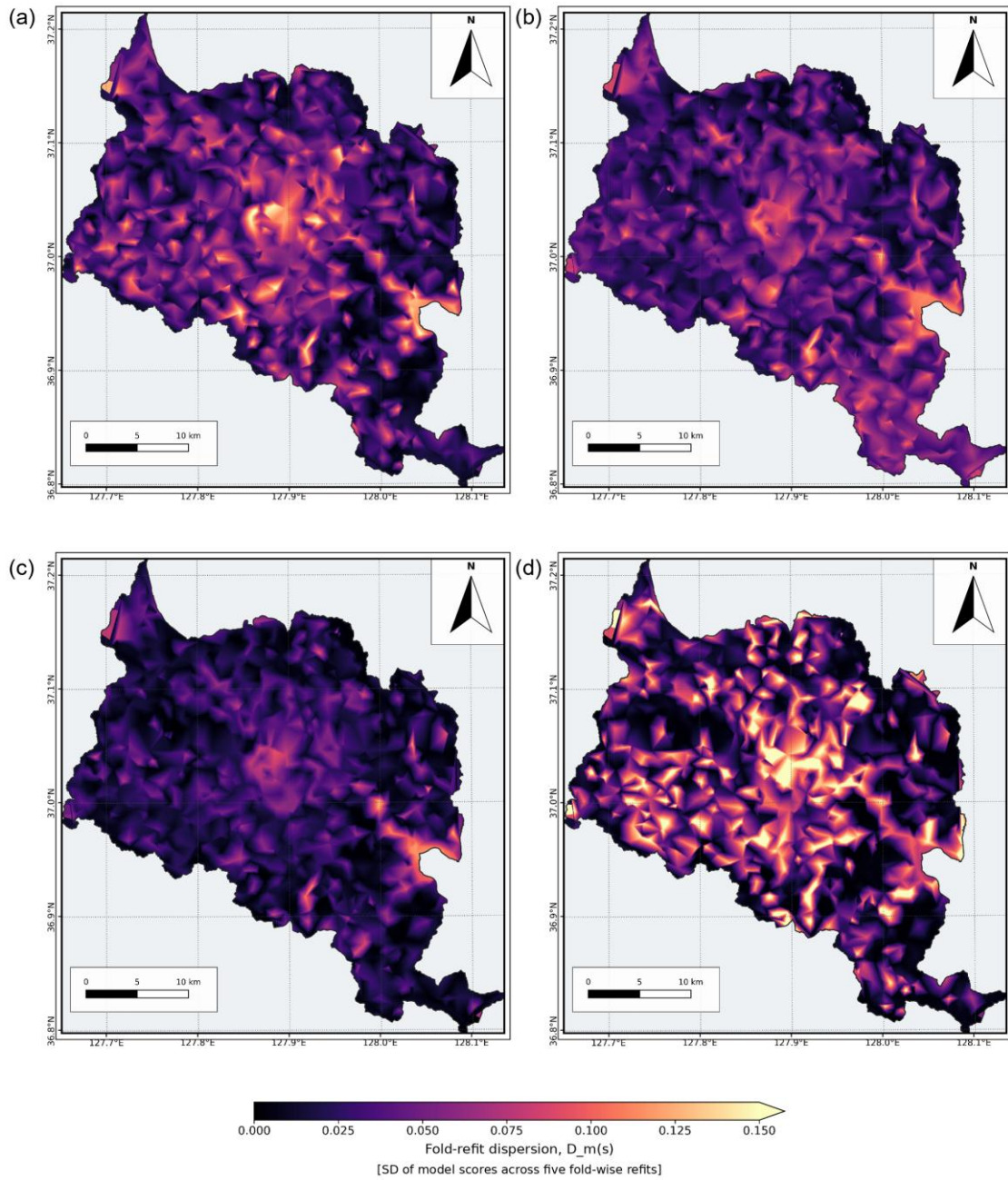
289 All four models assigned higher relative scores to the western lowlands and central valley corridors
290 and lower scores to much of the eastern mountain terrain (Fig. 4). The broad pattern is consistent with
291 the recurrent slope association and the concentration of positive records in accessible lowland settings.
292 Model families differed mainly in surface texture. Mapped values are raw case-background scores on
293 a 0-1 display scale; the same value need not have identical meaning across models, and the maps
294 support within-model ranking rather than cross-model comparison of absolute probabilities.



295

296 **Fig. 4** Relative well-occurrence favorability scores predicted by (a) GRF, (b) SGP, (c) kNN-lag, and
 297 (d) Coord-XGBoost. Scores are conditional on the balanced case-background design and are within-
 298 model rankings, not calibrated probabilities or yield estimates

299 Fold-refit dispersion across Cheongju is shown in Fig. 5.



0.000 0.025 0.050 0.075 0.100 0.125 0.150
 Fold-refit dispersion, $D_m(s)$
 [SD of model scores across five fold-wise refits]

300

301 **Fig. 5** Fold-refit dispersion for (a) GRF, (b) SGP, (c) kNN-lag, and (d) Coord-XGBoost. Dispersion is
 302 the pixel-wise standard deviation of model scores across five fold-wise refits and is a within-model
 303 instability diagnostic, not calibrated posterior uncertainty or predictive reliability

304

305 Mean raw dispersion was 0.041 for GRF, 0.040 for SGP, 0.025 for kNN-lag, and 0.048 for Coord-
 306 XGBoost. These values summarize score variability across five refits and should not be used to declare

one model globally more reliable. For staged assessment, high-score cells with lower within-model dispersion are more consistently supported by the current inventory, whereas high-score cells with higher dispersion warrant additional observations before field investment. No composite stability score was retained because dispersion alone is not calibrated predictive reliability.

5 Discussion

5.1 Evidence for model ranking

The principal comparative result is that the available evaluation does not resolve an accuracy ranking among the four spatial structures. The spatial-partition means differed by 0.021 and the secondary hold-out AUCs by 0.017. DeLong tests did not detect hold-out differences, but the evaluation sample also had limited power to resolve a 0.03 AUC gap. The TOST result applies only to SGP versus GRF and supported equivalence only at the wider margin of 0.10. It would therefore be inappropriate to select a definitive winner or describe all four models as equivalent.

The spatial evidence is deliberately bounded. Geographic blocking is more informative than a purely random split for a clustered inventory, but the same fold system was used for tuning and evaluation, adjacent folds shared boundaries, and no buffer was imposed. The reported AUCs show how scores changed under one transparent spatial partition; they do not estimate performance at a site separated by a fixed distance from all training observations. Repeated, buffered, nested, or external validation would be required for stronger transferability claims.

5.2 Environmental and geographic signal

The component variants indicate that environmental factors, rather than the explicit geographic inputs, carried most blocked discrimination under the tested partition. Geographic mechanisms still changed local texture and refit dispersion, but they did not provide a consistent mean-AUC gain over the corresponding factors-only variants. This suggests that spatial complexity should be justified by the intended decision product, computational demands, and validation evidence rather than nominal score

331 alone. Related groundwater studies similarly demonstrate effective ensemble or optimized models
332 while emphasizing the importance of data scarcity, predictor support, and management context (Mosavi
333 et al. 2021; Saqr et al. 2025; Mouhoumed et al. 2026).

334 Slope and land-surface factors were recurrent within the model-specific explanation objects. Their
335 prominence is hydrogeologically plausible because terrain influences runoff, infiltration opportunity,
336 regolith development, and access, but the analysis does not establish causality. Forest type and land
337 cover covary with terrain and human accessibility, and the positive inventory is concentrated where
338 wells are more likely to be installed and recorded. Moreover, the GRF panel is a surrogate, the SGP
339 panel uses a reduced sample, and the explainers differ. Responsible interpretation therefore requires
340 within-panel, association-based language (Maier et al. 2024; Zounemat-Kermani and Kheimi 2026).

341 **5.3 Management use of favorability and refit stability**

342 The maps are most defensible as a screening layer in a staged groundwater-assessment process. They
343 can guide limited survey resources toward locations combining relatively high within-model scores
344 with lower refit dispersion and identify high-score locations where additional observations are
345 warranted. This use aligns with the journal's focus on water-resources assessment, applied groundwater
346 hydrology, monitoring, and management-related modelling. Collaborative monitoring research also
347 shows that spatial evidence becomes more useful when linked to observation strategies and science-
348 policy translation (Reading et al. 2025).

349 The maps do not estimate water-table depth, pumping rate, specific capacity, recharge, groundwater
350 quality, sustainable abstraction, or drilling success. A management sequence should therefore use them
351 only to screen candidate areas, followed by field hydrogeology, geophysics, water-quality assessment,
352 land-access review, pumping tests, and applicable regulatory evaluation. Groundwater-quality
353 classification and recharge-zone delineation address complementary decisions and should not be
354 inferred from a well-occurrence score (Akbari et al. 2021; Zegaar et al. 2024).

355 **5.4 Scope, reproducibility, and priority extensions**

356 The scope of inference is defined by the available analytical record. The 207 positive records lack a
357 common yield threshold, and their record-level relationship to the 2006 survey and Lee (2025) was not
358 established. Background points are not verified negatives, and no water-body or land-use mask was
359 applied. The combined sensitivity analysis changes several design elements simultaneously. Spatial
360 evaluation is non-nested and unbuffered, and explanation objects differ. Complete independent
361 reproduction also requires source-layer provenance, the full tuning record, GRF weighting details,
362 fitted SGP parameters, software versions, final-map fitting details, and a persistent code-and-data
363 archive.

364 These conditions do not negate the descriptive comparison; they delimit its appropriate use. Priority
365 extensions are a record-level inventory audit, verified dry or continuous-yield observations,
366 harmonized background designs, nested and buffered repeated spatial validation, spatially aware paired
367 inference, matched-sample explanation or surrogate-fidelity assessment, complete factor metadata, and
368 external testing in a contrasting basin. Items that require new analysis or author-supplied provenance
369 are separated in the accompanying revision report and were not inferred or fabricated in this revision.

370 **6 Conclusions**

371 This study compared GRF, SGP, kNN-lag, and Coord-XGBoost for relative well-occurrence
372 favorability in Cheongju using one common spatial-block partition and a secondary stratified hold-out
373 set. Mean spatial-partition AUC ranged from 0.850 to 0.871, and hold-out AUC ranged from 0.858 to
374 0.875. Formal hold-out comparisons did not resolve differences, while limited power and spatial
375 dependence precluded a claim of model equivalence. Fixed-configuration factors-only variants
376 changed mean AUC by no more than 0.010, whereas spatial-only variants declined to 0.538-0.640.

377 The four surfaces shared a broad lowland-versus-mountain pattern. Model-specific attribution and fold-
378 refit dispersion added complementary information about recurrent factor associations and within-

379 model score stability. The outputs are conditional rankings from a positive-well/background design,
380 not estimates of aquifer yield, sustainable abstraction, groundwater quality, or drilling success. Their
381 most defensible management use is staged reconnaissance: prioritize relatively high-score, lower-
382 dispersion locations for initial investigation and direct additional observations to high-score, higher-
383 dispersion locations. Stronger transferability claims require the additional validation and provenance
384 work identified in the revision report.

385 **References**

- 386 Akbari M, Meshram SG, Krishna RS et al (2021) Identification of the groundwater potential recharge
387 zones using MCDM models: full consistency method (FUCOM), best worst method (BWM) and
388 analytic hierarchy process (AHP). *Water Resour Manag* 35:4727-4745.
389 <https://doi.org/10.1007/s11269-021-02924-1>
- 390 Bradbury KR (2025) Groundwater potential mapping: a misused and dubious concept. *Groundwater*
391 63:664-665. <https://doi.org/10.1111/gwat.70013>
- 392 Breiman L (2001) Random forests. *Mach Learn* 45:5-32. <https://doi.org/10.1023/A:1010933404324>
- 393 Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd*
394 *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New
395 York, pp 785-794. <https://doi.org/10.1145/2939672.2939785>
- 396 Christelis V, Mansour MM, Jackson CR (2025) Characterisation of groundwater drought using
397 distributed modelling, standardised indices, and principal component analysis. *Water Resour Manag*
398 39:4227-4241. <https://doi.org/10.1007/s11269-024-03997-4>
- 399 DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more
400 correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837-845.
401 <https://doi.org/10.2307/2531595>

402 Georganos S, Grippa T, Niang Gadiaga A et al (2021) Geographical random forests: a spatial extension
 403 of the random forest algorithm. *Geocarto Int* 36:121-136.
 404 <https://doi.org/10.1080/10106049.2019.1595177>
 405 Lee S (2025) Groundwater potential map data of Cheongju area using probabilistic methods. *GEO*
 406 *DATA* 7:160-171. <https://doi.org/10.22761/GD.2025.0013>
 407 Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: *Advances in*
 408 *Neural Information Processing Systems* 30, pp 4765-4774
 409 Lundberg SM, Erion G, Chen H et al (2020) From local explanations to global understanding with
 410 explainable artificial intelligence for trees. *Nat Mach Intell* 2:56-67. [https://doi.org/10.1038/s42256-](https://doi.org/10.1038/s42256-019-0138-9)
 411 [019-0138-9](https://doi.org/10.1038/s42256-019-0138-9)
 412 Maier HR, Taghikhah FR, Nabavi E et al (2024) How much X is in XAI: responsible use of
 413 “explainable” artificial intelligence in hydrology and water resources. *J Hydrol X* 25:100185.
 414 <https://doi.org/10.1016/j.hydroa.2024.100185>
 415 Ministry of Construction and Transportation, Korea Water Resources Corporation (2006) Groundwater
 416 basic survey report for the Cheongwon (Cheongju) area. Ministry of Construction and Transportation,
 417 Republic of Korea (in Korean)
 418 Mosavi A, Sajedi Hosseini F, Choubin B et al (2021) Ensemble boosting and bagging based machine
 419 learning models for groundwater potential prediction. *Water Resour Manag* 35:23-37.
 420 <https://doi.org/10.1007/s11269-020-02704-3>
 421 Mouhoumed RM, Başakın EE, Abdallah IS et al (2026) Integrating metaheuristic optimization with
 422 stochastic gradient boosting for groundwater potential prediction in data-scarce arid environments.
 423 *Water Resour Manag* 40:412. <https://doi.org/10.1007/s11269-026-04788-9>
 424 Rasmussen CE, Williams CKI (2006) *Gaussian processes for machine learning*. MIT Press, Cambridge

425 Reading L, Gurieff LB, Catania S (2025) Sustainable groundwater management through collaborative
 426 local scale monitoring. *Water Resour Manag* 39:725-740. [https://doi.org/10.1007/s11269-024-03986-](https://doi.org/10.1007/s11269-024-03986-7)
 427 7
 428 Roberts DR, Bahn V, Ciuti S et al (2017) Cross-validation strategies for data with temporal, spatial,
 429 hierarchical, or phylogenetic structure. *Ecography* 40:913-929. <https://doi.org/10.1111/ecog.02881>
 430 Rodríguez-Flores JM, Valero Fandiño JA, Cole SA et al (2022) Global sensitivity analysis of a coupled
 431 hydro-economic model and groundwater restriction assessment. *Water Resour Manag* 36:6115-6130.
 432 <https://doi.org/10.1007/s11269-022-03344-5>
 433 Saqr AM, Kartal V, Karakoyun E et al (2025) Improving the accuracy of groundwater level forecasting
 434 by coupling ensemble machine learning model and coronavirus herd immunity optimizer. *Water*
 435 *Resour Manag* 39:5415-5442. <https://doi.org/10.1007/s11269-025-04210-w>
 436 Schratz P, Muenchow J, Iturrity E, Richter J, Brenning A (2019) Hyperparameter tuning and
 437 performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol Model*
 438 406:109-120. <https://doi.org/10.1016/j.ecolmodel.2019.06.002>
 439 Zegaar A, Ounoki S, Telli A (2024) Machine learning for groundwater quality classification: a step
 440 towards economic and sustainable groundwater quality assessment process. *Water Resour Manag*
 441 38:621-637. <https://doi.org/10.1007/s11269-023-03690-y>
 442 Zounemat-Kermani M, Kheimi M (2026) Explainable artificial intelligence in hydrology: a review.
 443 *Water Resour Manag* 40:106. <https://doi.org/10.1007/s11269-025-04435-9>

444 **Statements and Declarations**

445 **Funding.** This research was supported by the Basic Research Project of the Korea Institute of
 446 Geoscience and Mineral Resources (26-3131, GP2025-014) and by a National Research Foundation of
 447 Korea grant funded by the Korean government, Ministry of Science and ICT (26-6811, NP2023-021,

448 No. 2023R1A2C1003095). The funders had no role in study design, data analysis, interpretation, or
449 manuscript preparation.

450 **Competing Interests.** The authors have no relevant financial or non-financial interests to disclose.

451 **Author Contributions.** Sooyeon Han: conceptualization, methodology, software, validation, formal
452 analysis, investigation, data curation, visualization, writing - original draft, and writing - review and
453 editing. Saro Lee: conceptualization, resources, supervision, project administration, funding
454 acquisition, and writing - review and editing. Both authors read and approved the final manuscript.

455 **Data Availability.** Derived favorability maps, tabulated model results, and code-generated outputs
456 supporting this study are available from the corresponding author upon reasonable request. The original
457 elevation, geology, soil, land-cover, forest, and groundwater-well datasets cannot be redistributed by
458 the authors because they are owned by third-party providers and are subject to their licensing and access
459 conditions. Source identifiers and a persistent location for shareable derived outputs should be added
460 before final submission.

461 **Code Availability.** The code used for modeling, spatial partitioning, statistical testing, SHAP analysis,
462 component variants, and dispersion mapping is available to editors and reviewers from the
463 corresponding author upon reasonable request during peer review and will be deposited in a versioned
464 public repository upon acceptance. The repository URL and release identifier should be inserted when
465 available.

466 **Ethics Approval.** Not applicable. This study did not involve human participants or animals.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [CheongjuGWPMWRMSupplementaryInformation.docx](#)