UPPSALA
UNIVERSITET

RANDOM FOREST WITH GENERALIZED LEAST SQUARES FOR SPATIAL MODELING: A CASE STUDY ON LAKE PH IN SWEDEN

Submitted by

Georgios Pnigouras

A thesis submitted to the Department of Statistics in partial fulfillment of the requirements for the two-year degree of Master of Arts in Statistics in the Faculty of Social Sciences

Supervisor: *Tatjana Pavlenko*

Co-supervisor: *Claudia von Brömssen*

Spring, 2025

ABSTRACT

Environmental monitoring increasingly relies on flexible, data-driven methods to model complex spatial processes. Random Forest (RF) algorithms are widely used for capturing non-linear relationships among numerous predictors, but they assume independent observations and thus fail to account for spatial autocorrelation. Random Forest–Generalized Least Squares (RF–GLS) addresses this limitation by modeling a Gaussian-process covariance structure into each fit. This thesis evaluates the performance of Random Forest Generalized Least Squares (RF–GLS) in comparison to standard Random Forest (RF) through simulation experiments and a real-world case study of lake-water pH across Sweden. The study aims to identify the order of importance of various variables influencing pH levels, providing insights into the most significant environmental factors affecting lake-water acidity. Simulation results show that RF–GLS consistently outperforms standard RF when spatial autocorrelation is strong, but this advantage diminishes as the nugget effect (non-spatial error) becomes comparable in magnitude to the spatial variance component. In the case study, both methods perform similarly in predictive accuracy across four environmentally distinct subregions. Permutation-based importance analysis identifies calcium concentration as the most influential predictor of pH, while geographic variables contribute minimally. These findings indicate that, in practice, RF and RF–GLS yield broadly similar results in both prediction and variable importance, even when moderate spatial dependence is present.

Contents

1 Introduction

Environmental data are a domain in which statistical methods are widely applied. As computational power has grown, new techniques—particularly machine learning algorithms like Random Forest—have shown great promise across many areas of environmental science. Spatial data, which record measurements at geographic locations, form a special category because nearby observations tend to be similar: patterns often arise from factors such as distance, elevation, or land cover. At the same time, advances in sensor technology have made it possible to collect environmental measurements with ever higher precision and resolution. These developments have driven renewed interest in spatial statistics and hybrid approaches that combine traditional geostatistics with machine learning, offering more accurate and robust analyses of complex environmental processes.

Spatial datasets often feature complex, nonlinear relationships and sudden changes in how covariates influence the response variable. Traditional methods that rely on fixed, smooth basis functions (e.g., splines or polynomials) struggle to capture these discontinuities or to scale gracefully when dozens of predictors are involved. In contrast, Random Forests build an ensemble of decision trees that use data-driven, piecewise-constant basis functions. This adaptive approach can flexibly approximate highly irregular and interactive effects far more expressively than fixed bases, making RF a leading choice for nonlinear regression in many fields. However, standard RF assumes that all observations are independent. In environmental settings—where measurements taken close together tend to be correlated—this independence assumption is violated. As a result, naïvely applying RF can overlook or misattribute spatial structure, motivating extensions like RF–GLS that explicitly model residual spatial autocorrelation.

Over the last decade, researchers have developed methods that retain the strengths of random forests while accounting for spatial correlation. These methods fall into two main categories. The first and simplest embeds spatial structure directly into the forest by adding spatial proxies—such as coordinates or distance-to-feature fields—to the predictor set. Although flexible and easy to implement, these proxy-based approaches often lack a formal theoretical foundation and do not explicitly model the underlying spatial process ([11],[1],[9]). The second approach fits a standard random forest to estimate the covariate effect without any adjustments for spatial dependence, then computes the residuals from that fit and applies a Gaussian-process model to them in order to perform kriging ([7]).

In 2023, Saha et al. introduced a novel method—Random Forest Generalized Least Squares

(RF-GLS)—that unifies the two prevailing strategies for handling spatially correlated data in random forests ([17]). Traditional ("vanilla") random forests choose splits and make node predictions by minimizing ordinary least-squares error, which assumes independent residuals and thus becomes inefficient when spatial autocorrelation is present. RF-GLS replaces each local OLS step with a generalized least-squares (GLS) step that explicitly incorporates the Gaussian-process covariance structure among observations. By doing so, RF-GLS makes both the split decisions and the node predictions optimal under spatial dependence, effectively combining the flexibility of random forests with the rigor of spatial mixed-effects modeling.

Because RF-GLS is such a recent development, its use to date has been mostly in simulation studies rather than real-world applications. One notable case study is by [14] who applied RF-GLS alongside several other methods to analyze air temperature and particulate air pollution data across Spain, reporting promising improvements in predictive accuracy.

In this thesis, the analysis of the new RF-GLS method will be examined, tested and extended. The main point of comparison is the standard ("vanilla") random forest. The aim is to determine whether—and under what conditions—RF-GLS outperforms the simple random forest. To address this question, it will be conducted a series of simulation studies by generating domains with varying spatial-autocorrelation characteristics. These simulated fields will include both idealized scenarios and patterns that mimic real-world cases, allowing us to evaluate the strengths and limitations of each method.

In the second half of this thesis, it will be carried out a case study using lake-water data collected across Sweden from 2019 to 2024. First, it will be quantified the degree of spatial autocorrelation present in the response variable—lake pH, measured on a log-transformed acidity/basicity scale—throughout the entire country. Next, Sweden will be partitioned into four environmentally meaningful subregions and assess spatial autocorrelation separately within each region. Finally, both the standard random forest and the RF-GLS methods will be applied to these real-world data in two ways.

1. **Predictive accuracy across regions.** For each region, each model will be trained on 80% of the observations and its predictive performance will be evaluated on the remaining 20%. This will allow to compare how well each algorithm generalizes to unseen data under differing spatial patterns.

2. **Variable-importance analysis.** Both methods will be used to identify the most influential environmental covariates in determining lake water quality (pH). Then it will be ex-

amined whether the two algorithms agree on which predictors matter most, and whether these key drivers vary across the four regions.

By combining simulation-based insights with this real-data application, the aim is to understand not only when and why RF-GLS can outperform a simple random forest, but also what practical benefits it may offer for environmental modelling and inference in spatially correlated settings.

2 Spatial Correlation

Before selecting an appropriate modeling approach, we conducted a formal test for spatial autocorrelation using Moran's I ([1],[2]). This test evaluates whether observed spatial patterns are consistent with random spatial distribution or if significant clustering exists. The procedure is outlined below:

1. Hypotheses:

   The global Moran's I test evaluates whether a variable is spatially autocorrelated. The null hypothesis is that the observed values are spatially independent (complete spatial randomness), i.e. no spatial autocorrelation. Equivalently, one tests,

   $$H_0 : I = \mathbb{E}[I] \quad \text{(no spatial autocorrelation)},$$
   $$H_1 : I \neq \mathbb{E}[I] \quad \text{(presence of spatial autocorrelation)}.$$

   Here, $\mathbb{E}[I]$ is the expected Moran's I under H_0. Under H_0, the spatial pattern is assumed random, whereas under H_1 the pattern exhibits either clustering (positive autocorrelation) or dispersion (negative autocorrelation) beyond chance.

2. Moran's I Statistic:

   Moran's I is defined as

   $$I \;=\; \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \, (Y_i - \bar{Y}) \, (Y_j - \bar{Y})}{\left(\sum_{i \neq j} w_{ij} \right) \, \sum_{i=1}^{n} (Y_i - \bar{Y})^2},$$

   where

   - n is the total number of observations,
   - Y_i is the value at location i,
   - $\bar{Y}$ is the overall mean of the Y_i,
   - w_{ij} are spatial weights describing the neighborhood relation (by construction $w_{ii} = 0$).

   In our analysis the weight matrix $W = [w_{ij}]$ was built using the k nearest–neighbors rule, adopting the k value according to the dense in observations of the region we work on:

   $$w_{ij} = \begin{cases} 1, & \text{if location } j \text{ is among the } k \text{ closest neighbors of } i, \\ 0, & \text{otherwise.} \end{cases}$$

Thus each observation has exactly k neighbors, so $\sum_j w_{ij} = k$ for all i (and $\sum_{i,j} w_{ij} = S_0$, the total number of neighbor-pairs). Intuitively, Moran's I is a ratio of weighted cross-products of deviations; positive (negative) values of I suggest similar (dissimilar) values cluster in space.

3. Expectation and Variance under H_0:

   Under the null hypothesis of no spatial autocorrelation, observations Y_i are independent and identically distributed, and I is asymptotically normally distributed with mean and variance equal to

$$\mathbb{E}[I] \;=\; -\frac{1}{n-1}$$

   and

$$\mathrm{Var}(I) \;=\; \frac{n^2(n-1)\,S_1 \;-\; n(n-1)\,S_2 \;-\; 2\,S_0^2}{(n+1)\,(n-1)^2\,S_0^2}$$

   where,

$$S_0 \;=\; \sum_{i\neq j} w_{ij}, \quad S_1 \;=\; \tfrac{1}{2}\sum_{i\neq j}(w_{ij}+w_{ji})^2, \quad S_2 \;=\; \sum_{k=1}^{n}\Big(\sum_j w_{kj} \;+\; \sum_i w_{ik}\Big)^2.$$

   Moran's I typically ranges between -1 and 1. Values of Moran's I that lie well above $\mathbb{E}[I] = -\frac{1}{n-1}$ reflect positive spatial autocorrelation (clustering), meaning nearby areas tend to have similar values. Conversely, values substantially below $\mathbb{E}[I]$ indicate negative spatial autocorrelation (dispersion), where neighboring regions display dissimilar values. Finally, Moran's I values close to $\mathbb{E}[I]$ are consistent with a random spatial pattern, showing no clear clustering or dispersion.

4. Decision Rule (Asymptotic Normal Test)

   For large n, Moran's I is asymptotically normally distributed under H_0. We therefore standardize I to a z-score:

$$z = \frac{I - \mathbb{E}[I]}{\sqrt{Var(I)}}$$

   which approximately follows a standard normal distribution under H_0. A two-sided p-value is obtained from this normal distribution. At a chosen significance level (e.g. $\alpha = 0.05$), we reject H_0 if the p-value is below α. Equivalently, we compare z to the critical normal quantiles. Importantly, this inference uses the asymptotic normal approximation (a z-test).

3 Model

The non-linear mixed spatial model for lakes data in the Swedish region is given by

$$Y_i = m(X_i) + w(l_i) + \varepsilon_i$$

where the triplet (Y_i, X_i, l_i) is the standard geospatial data unit with Y_i the univariate response, the pH value, the X_i is the D-dimensional covariate vector (feature), and l_i is the spatial location in coordinates $(latitude_i, longitude_i)$ for $lake_i$. The triplet (Y_i, X_i, l_i) is known, and all values are given from the entire data set.

The mean function $m(X_i)$ expresses the non linear relationship between the lake-level covariates X_i and the pH for the lake in the location l_i. The form of the mean function is unknown, and its estimation is one of the project's main goals. The method has been decided to apply for the estimation of it is the random forest generalized least square(RFGLS) ([17]), and more information about the reason and details of the method will be available later in the method section.

$w(l_i)$ expresses the spatial random effect for the lake in the location l_i and is modeled as a centered Gaussian Process, $\boldsymbol{w}(\cdot) \sim GP(\mathbf{0}, \boldsymbol{C})$ with spatial covariance matrix $\boldsymbol{C}$. Because of its flexibility in tuning the smoothness level of the spatial surface, as well as its popularity in the spatial field, the Matérn covariance family is the prevalent choice that will be used in the specific project, for the specification of the spatial covariance matrix. ([20]). Moreover, for the finite collection of locations $l_1, l_2, ..., l_n$, such as the observation lakes across Sweden in this specific example, the vector, $\boldsymbol{w} = w(l_1), w(l_2), ..., w(l_n)$ can be modeled as a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{C})$ where $\boldsymbol{C}$ is an $n \times n$ covariance matrix with entries defined as $cov(w(l_i), w(l_j)) = C(l_i, l_j|\boldsymbol{\theta})$. $\boldsymbol{\theta}$ denotes the parameters of the covariance function (e.g. range, smoothness) and needs to be estimated.

ε_i denotes the random noise modeled as $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$ and it is independent from the spatial effect $w(l_i)$. The τ^2 is unknown and its estimation takes place during the estimation of the rest of the unknown spatial variables.

Having specified the components of the model, the joint distribution can be expressed as

$$P(\boldsymbol{Y}, \boldsymbol{w}|m, \boldsymbol{\theta}, \tau^2) = P(Y|m, \boldsymbol{w}, \tau^2)P(\boldsymbol{w}|\boldsymbol{\theta}).$$

Since the spatial effect w is never directly observed, we cannot condition the likelihood on its realized values. Instead, to obtain a valid likelihood for the observed data $\boldsymbol{Y}$, we marginalize

over the distribution of $\boldsymbol{w}$ ([8]). This integration accounts for all possible latent surfaces $\boldsymbol{w}$ according to their prior distribution, yielding the marginal likelihood $P(Y|m, \boldsymbol{\theta}, \tau^2)$ without requiring direct data on $\boldsymbol{w}$. Specifically, this is done through the integral

$$P(\boldsymbol{Y}|m, \boldsymbol{\theta}, \tau^2) = \int_{\mathbb{R}^n} P(\boldsymbol{Y}|m, \boldsymbol{w}, \tau^2) P(\boldsymbol{w}|\boldsymbol{\theta}) d\boldsymbol{w}.$$

Here, the integration is performed over the entire n-dimensional real space $\mathbb{R}^n$, reflecting all possible realizations of the spatial random effects vector $\boldsymbol{w} = (w(l_1), w(l_2), \ldots, w(l_n))^\top$. This integral effectively averages the likelihood over the distribution of $\boldsymbol{w}$, incorporating the uncertainty associated with the unobserved spatial effects.

Because both $\boldsymbol{Y}|m, \boldsymbol{\theta}, \tau^2$ and $\boldsymbol{w}|\boldsymbol{\theta}$ follow Gaussian distributions, with

$$\boldsymbol{Y}|m, \boldsymbol{\theta}, \tau^2 \sim N(m(\boldsymbol{X}) + \boldsymbol{w}, \tau^2 \boldsymbol{I}), \boldsymbol{w}|\boldsymbol{\theta} \sim N(\boldsymbol{0}, \boldsymbol{C}),$$

the integral has a closed-form solution $\boldsymbol{Y} \sim N(m(\boldsymbol{X}), \boldsymbol{C} + \tau^2 \boldsymbol{I})$.

There are several reasons for choosing a marginal model over a hierarchical one. First, integrating out $\boldsymbol{w}$ reduces the dimensionality of the model, eliminating the need to estimate the n-dimensional latent vector and allowing inference to focus solely on the parameters of interest. Second, since both the observation model and the prior on $\boldsymbol{w}$ are Gaussian, this approach ensures a well-specified distribution for $\boldsymbol{Y}$. Finally, the resulting covariance matrix $\boldsymbol{C} + \tau^2 \boldsymbol{I}$ is strictly positive definite, which improves numerical stability. This matrix is better conditioned for Cholesky decomposition, matrix inversion, and log-determinant computations, avoiding issues related to near-singularity ([8]).

4 Method

4.1 Simple Random Forest

The Random Forest (RF)([3]) algorithm builds an ensemble of n_{tree} fully grown regression trees, each using a bootstrap (or subsample) of the original data. For each tree and at each internal node, randomly select m_{try} out of D features. For each chosen feature d, search over all possible cut-points c , the "gaps" between sorted values of $X^{(d)}$ and compute the CART split criterion ([4])

$$V_n^{CART}((d,c)) = \frac{1}{n_P}[\sum_{i=1}^{n_P}(Y_i^P - \bar{Y}^P)^2 - \sum_{i_r=1}^{n_R}(Y_i^R - \bar{Y}^R)^2 - \sum_{i_l=1}^{n_L}(Y_i^L - \bar{Y}^L)^2],$$

where $n_P = n_R + n_L$ is the parent node-size, n_R, n_L are the sizes of the right and left child nodes, and $\bar{Y}^P$, $\bar{Y}^R$, $\bar{Y}^L$ are their respective means.

The decision of the "best" pair (d, c) is according to the one that maximizes the V_n^{CART}. The method continues until a prespecified stopping rule is met, either when each leaf has one observation or fewer than a prespecified minimum.

Finally, given a new feature vector $x \in \mathbb{R}^D$, each tree t in the forest routes x down its splitting rules until it reaches a leaf node $C_k^{(t)}$. That tree's prediction is simply the average of all training responses that fell into that leaf:

$$\hat{m}^{(t)}(x) = \frac{1}{\left|C_k^{(t)} \cap \{X_i\}\right|} \sum_{X_i \in C_k^{(t)}} Y_i.$$

The Random Forest then *averages* these n_{tree} individual-tree predictions to obtain the final estimate:

$$\hat{m}_{RF}(x) = \frac{1}{n_{\text{tree}}} \sum_{t=1}^{n_{\text{tree}}} \hat{m}^{(t)}(x).$$

In a standard Random Forest, both the CART-based splitting and the computation of each node's prediction depend solely on the responses and covariates of the observations that fall into that node—any spatial coordinates l_i are completely ignored. When the data are truly independent, this localization causes no harm, since observations in different nodes carry no mutual information. But if the data exhibit spatial autocorrelation, then treating every node's subset as independent throws away the very information encoded by nearby locations, and can lead to poorer performance.

4.2 RF-GLS

The standard Random Forest ignores spatial dependence in the residuals, which can lead to biased splits when observations are spatially autocorrelated. To address this, the RF-GLS (Random Forest Generalized Least Squares) is proposed, which embeds a GLS-style variance criterion into the split-finding step ([17]). The RF-GLS split criterion follows the same fundamental philosophy as the standard Random Forest: the goal is to select the best split by maximizing the reduction in variance within each node. However, in this case, variance is measured using an alternative criterion, $V_{n,Q}^{\mathrm{DART}}$, which incorporates spatial correlation through the estimated covariance matrix $\tilde{\boldsymbol{\Sigma}}$.

The modified variance reduction criterion is given by:

$$V_{n,\boldsymbol{Q}}^{\mathrm{DART}}(d,c) = \frac{1}{n}\Big[\underbrace{(\boldsymbol{Y} - \boldsymbol{Z}^{(0)}\hat{\boldsymbol{\beta}}_{GLS}(\boldsymbol{Z}^{(0)}))^T\boldsymbol{Q}(\boldsymbol{Y} - \boldsymbol{Z}^{(0)}\hat{\boldsymbol{\beta}}_{GLS}(\boldsymbol{Z}^{(0)}))}_{\text{parent loss}}$$

$$-\underbrace{(\boldsymbol{Y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{GLS}(\boldsymbol{Z}))^T\boldsymbol{Q}(Y - Z\hat{\boldsymbol{\beta}}_{GLS}(\boldsymbol{Z}))}_{\text{children loss}}\Big],$$

where, $\boldsymbol{Q} = \tilde{\boldsymbol{\Sigma}}^{-1}$ the inverse of the estimated covariance matrix. (d, c) are defined as in the simple random forest the feature and the cutoff value, respectively of the split. $\mathbf{Z}^{(0)}$ is the $n \times R$ membership matrix, n number of locations, R number of nodes, that tracks node assignments before the split, while $\mathbf{Z}$ is the $n \times (R+1)$ membership matrix that contains the two new nodes.

Here $\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}(\boldsymbol{Z})$ are the node-means adjusted by the precision matrix $\boldsymbol{Q}$, so that correlated residuals receive less weight,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}(\boldsymbol{Z}) = \hat{\boldsymbol{\beta}} = (\boldsymbol{Z^T Q Z})^{-1}(\boldsymbol{Z^T Q Y}).$$

Because the criterion uses $\boldsymbol{Q}$, which itself is estimated from the entire dataset, every candidate split takes into account the global spatial correlation structure, rather than only the local variance within a node. This leads to partitions that respect the spatial smoothness of the underlying process.

Just like in simple RF, each tree casts one prediction (now a GLS-leaf mean), and the forest output is the arithmetic mean of those predictions.

$$\hat{m}_{\mathrm{GLS}}^{(t)}(x) = \hat{\beta}_{\mathrm{GLS},k}^{(t)} \quad \text{whenever } x \in C_k^{(t)},$$

where $\hat{\beta}^{(t)}_{\mathrm{GLS},k}$ is the GLS-estimated mean in leaf $C^{(t)}_k$. The forest-level estimate is then

$$\hat{m}_{RF\text{-}GLS}(x) = \frac{1}{n_{\mathrm{tree}}} \sum_{t=1}^{n_{\mathrm{tree}}} \hat{m}^{(t)}_{\mathrm{GLS}}(x).$$

4.3 Estimation of the covariance matrix

In real-world applications, the true covariance matrix $\boldsymbol{\Sigma}$ is unknown. Therefore, an essential preliminary step is to estimate the covariance matrix $\tilde{\boldsymbol{\Sigma}}$ before applying the method. In practice, however, the estimation of covariance matrices arising from Gaussian processes (GPs)—except in special cases such as the exponential kernel—are dense. Consequently, for large n, computing the precision $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ via a full Cholesky factorization costs $O(n^3)$ operations and $O(n^2)$ storage, making RF-GLS infeasible for large datasets.

NNGP: a sparse approximation. To overcome this, we approximate the inverse of the covariance matrix $\boldsymbol{Q}$ by using the nearest-neighbor Gaussian process (NNGP) built on the Matérn covariance function ([6]). This replaces the dense inverse with a *sparse* precision whose construction scales linearly in n. In particular, likelihood and quadratic-form computations drop from $O(n^3)$ to $O(n\,m^2)$ (and often effectively $O(n\,m)$), where $m \ll n$ is the neighbor-set size.

From full GP to Vecchia's product of conditionals. Let

$$\boldsymbol{w}(\boldsymbol{l}) \;\sim\; GP\big(\mathbf{0},\; \boldsymbol{C}(\cdot,\cdot \mid \boldsymbol{\theta})\big),$$

where $\boldsymbol{l} = \{l_1, \ldots, l_n\}$ is our ordered set of locations and $\boldsymbol{\theta}$ the Matérn parameters. For a finite set of locations $\boldsymbol{l}$ by definition of a GP,

$$\boldsymbol{w}_{\boldsymbol{l}} = (\boldsymbol{w}(l_1), \ldots, \boldsymbol{w}(l_n))^{\top} \;\sim\; \mathcal{N}\big(\mathbf{0},\; \boldsymbol{C}_{\boldsymbol{l}}(\boldsymbol{\theta})\big),$$

with $[\boldsymbol{C}_{\boldsymbol{l}}]_{ij} = \boldsymbol{C}(l_i, l_j \mid \boldsymbol{\theta})$. Its joint density factorizes exactly as

$$p(\boldsymbol{w}_{\boldsymbol{l}}) = p\big(\boldsymbol{w}(l_1)\big)\, p\big(\boldsymbol{w}(l_2) \,|\, \boldsymbol{w}(l_1)\big) \,\cdots\, p\big(\boldsymbol{w}(l_n) \,|\, \boldsymbol{w}(l_1), \ldots, \boldsymbol{w}(l_{n-1})\big).$$

Vecchia's approximation via m neighbors. According to [22], for each i choose a small neighbor set

$$N(l_i) \subset \{l_1, \ldots, l_{i-1}\}, \qquad |N(l_i)| = m \ll n,$$

consisting of the m closest locations according to Euclidean distance. Then the $p(\boldsymbol{w}_{\boldsymbol{l}})$ can be approximated from

$$\tilde{p}(\boldsymbol{w}_{\boldsymbol{l}}) = \prod_{i=1}^{n} p\big(\boldsymbol{w}(l_i) \,|\, \boldsymbol{w}_{N(l_i)}\big).$$

Graphical interpretation. The pair $\{\boldsymbol{l}, N_{\boldsymbol{l}}\}$, with $\boldsymbol{l}$ being the set of nodes and $N_{\boldsymbol{l}} = \{N(l_i) : i = 1, 2, \ldots, n\}$ the collection of all conditioning sets over $\boldsymbol{l}$, can be viewed as a directed graph, with nodes corresponding to locations in $\boldsymbol{l}$ and directed edges determined by the $N_{\boldsymbol{l}}$. Specifically, for two nodes l_i and l_j, we say that l_j is a directed neighbor of l_i if $l_j \in N(l_i)$. That is, l_j belongs to the conditioning set of l_i.

A directed cycle in a graph is a sequence of nodes $l_{i_1}, l_{i_2}, \ldots, l_{i_b}$ such that $l_{i_1} = l_{i_b}$ and each consecutive pair has a directed edge. If the directed graph $\{\boldsymbol{l}, N_{\boldsymbol{l}}\}$ has no directed cycles, it is called a directed acyclic graph (DAG).

According to Vecchia and Stoud ([22], [21]), by specifying $N(l_i)$ to be the m nearest neighbors of location l_i with respect to the Euclidean distance among previously ordered locations, we ensure that $\{\boldsymbol{l}, N_{\boldsymbol{l}}\}$ is a DAG. This guarantees that $\tilde{p}(\boldsymbol{w}_{\boldsymbol{l}})$ defines a valid joint density ([12]), which approximates the original multivariate Gaussian.

Explicit form of each conditional. Each factor in the product is

$$p\big(\boldsymbol{w}(l_i) \,|\, \boldsymbol{w}_{N(l_i)}\big) = \mathcal{N}\Big(\boldsymbol{w}(l_i) \;\Big|\; \boldsymbol{B}_{l_i}\, \boldsymbol{w}_{N(l_i)},\ \boldsymbol{F}_{l_i}\Big),$$

where

$$\boldsymbol{B}_{l_i} = \boldsymbol{C}_{l_i, N(l_i)}\, \boldsymbol{C}^{-1}_{N(l_i), N(l_i)}, \qquad \boldsymbol{F}_{l_i} = \boldsymbol{C}(l_i, l_i) \;-\; \boldsymbol{C}_{l_i, N(l_i)}\, \boldsymbol{C}^{-1}_{N(l_i), N(l_i)}\, \boldsymbol{C}_{N(l_i), l_i}$$

with $\boldsymbol{C}_{N(l_i)}$ being the covariance matrix of $\boldsymbol{w}_{N(l_i)}$ and $\boldsymbol{C}_{l_i, N(l_i)}$ the cross-covariance between $\boldsymbol{w}(l_i)$ and $\boldsymbol{w}_{N(l_i)}$.

Connection to RF-GLS. In RF-GLS, we need a *working precision* $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ to whiten residuals in the Random Forest loss. Rather than using the dense inverse of the full Matérn matrix, we set

$$\widetilde{\boldsymbol{\Sigma}}^{-1} \;=\; \boldsymbol{L}^{\top} \boldsymbol{L}$$

where $\boldsymbol{L}$ is the sparse Cholesky factor obtained from the NNGP conditionals above. This substitution reduces the per-iteration cost of evaluating the GLS criterion from $O(n^3)$ to $O(n\, m^2)$ (often effectively $O(n\, m)$), enabling RF-GLS to scale to very large spatial datasets.

Thus, the NNGP precision $\widetilde{\boldsymbol{\Sigma}}^{-1}$ is exactly the covariance input you use in RF-GLS to achieve both statistical rigor and computational efficiency.

Parameters estimation via L-BFGS-B. An important aspect of the method is estimating the unknown spatial parameters $\boldsymbol{\theta} = \{\sigma, \phi, \nu\}$ and the random-noise parameter τ. Because there is no closed-form solution for these estimates, the values are obtained numerically by maximizing the Vecchia-approximated log-likelihood.

The Vecchia approximation simplifies the full Gaussian process likelihood by decomposing it into a product of conditional densities, thereby reducing computational complexity. Specifically, the approximated log-likelihood is expressed as:

$$log p_{Vecchia}(\boldsymbol{Y}|\sigma, \phi, \nu, \tau) = \sum_{i=1}^{n} log\ p(Y_i|\boldsymbol{Y}_{N(l_i)}; \sigma, \phi, \nu, \tau)$$

where $\boldsymbol{Y}$ denoted the observed data, and $N(l_i)$ represents a selected subset of indices corresponding to the neighbors of the i-th observation, chosen from the set $1, 2, \ldots, i-1$. This formulation allows for efficient computation by considering only a limited number of conditioning variables for each observation.

Following the approach of Arkajyoti Saha, Sumanta Basu and Abhirup Datta [17] and their R package implementation of RF-GLS [18], parameter estimation is performed using the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. L-BFGS is a quasi-Newton optimization method that approximates the inverse Hessian matrix using a limited amount of memory, making it particularly suitable for large-scale computations where storing the full Hessian is impractical. This optimization technique iteratively updates parameter estimates to maximize the approximated log-likelihood efficiently ([5], [13]).

5 Simulation Study

Before applying the algorithms to the lake data, a simulation study was conducted to evaluate and compare the estimation and prediction performance of RF-GLS and standard Random Forest under varying levels of spatial correlation.

5.1 Simulation Model

Data were generated under the following spatial regression model:

$$Y_i \;=\; m\big(X_{1i}, X_{2i}, X_{3i}\big) \;+\; w(\ell_i) \;+\; \epsilon_i,$$

where:

- $m(X_{1i}, X_{2i}, X_{3i})$ denotes the mean function, capturing the nonlinear relationship between the the three covariates X_1, X_2, X_3 and the response variable for the i location.
- $w(\ell_i)$ represents a zero-mean spatial random effect over the domain $D = [0, 100] \times [0, 100]$ for the location i. The vector $\boldsymbol{w} = \big(w(\ell_1), \ldots, w(\ell_n)\big)^\top$ follows

  $$\boldsymbol{w} \sim \mathcal{N}_n\big(\mathbf{0},\, \boldsymbol{C}\big),$$

  where the covariance matrix $\boldsymbol{C}$ is induced by the Matérn covariance function with smoothness $\nu = 3/2$:

  $$C(\ell_i, \ell_j \mid \boldsymbol{\theta}) = \sigma^2\Big(1 + \tfrac{\sqrt{3}\, d_{ij}}{\rho}\Big) \exp\Big(-\tfrac{\sqrt{3}\, d_{ij}}{\rho}\Big), \quad d_{ij} = \|\ell_i - \ell_j\|, \quad \boldsymbol{\theta} = (\sigma^2, \rho, \nu).$$

- ϵ_i is an independent measurement-error term for the location i,

  $$\epsilon_i \;\sim\; \mathcal{N}(0, \tau^2),$$

  independent of $\boldsymbol{w}$.

Defining $C_{ij} = C(\ell_i, \ell_j \mid \boldsymbol{\theta})$ yields the $n \times n$ matrix $\boldsymbol{C}$. Adding the nugget variance τ^2 produces the full covariance of the observations:

$$\boldsymbol{\Sigma} \;=\; \boldsymbol{C} \;+\; \tau^2 \boldsymbol{I}_n.$$

5.2 Hyperparameters tuning

Before conducting the core simulation experiments, it is essential to examine the key hyper-parameters that define the scenarios under which model performance will be evaluated. In particular, the Matérn covariance function depends on three parameters: the spatial variance σ^2, the range ρ, and the smoothness ν. For this study, ν is fixed at 1.5 to impose a moderate degree of smoothness. Additionally, the variability of the response surface over the simulated domain (of size 100×100) is governed by the nugget variance τ^2, which represents independent measurement noise and is the only variance component not determined by the Matérn function.

To assess how each covariance parameter affects the level of spatial autocorrelation, a total of 27 combinations of $\{\rho, \sigma^2, \tau^2\}$ were examined. This complete factor design allows for the systematic evaluation of the independent and joint impacts of the range (ρ), spatial variance (σ^2) and nugget variance (τ^2) on autocorrelation strength.

Firstly, three values of the range parameter ρ were selected to represent different distances over which spatial correlation decays. As illustrated in Figure 1, the first step of the simulation is to define the study domain. In this analysis, each domain is a square of size 100×100. The square is divided into 200 equally spaced columns and 200 equally spaced rows, yielding a regular grid of $200 \times 200 = 40,000$ points. In a square domain, the maximum distance between any two grid points is the diagonal, $D_{\max} = \sqrt{100^2 + 100^2} \approx 141.4$. The three values of ρ correspond to the distances at which spatial correlation falls to near zero, namely

$$\rho \in \left\{ D_{\max}/20,\ D_{\max}/10,\ D_{\max}/2 \right\}.$$

These choices ensure that the models are tested under short, moderate, and long correlation ranges prior to reaching the point, beyond which residual variation is effectively random.

The spatial autocorrelation structure in the simulation study is governed by the Matérn covariance function, which, for a smoothness parameter $\nu = 3/2$, simplifies to ([10]) :

$$C(l_i, l_j \mid \boldsymbol{\theta}) = \sigma^2 \left(1 + \frac{\sqrt{3} d_{ij}}{\rho}\right) \exp\left(-\frac{\sqrt{3} d_{ij}}{\rho}\right),$$

where $d_{ij} = \|l_i - l_j\|$ denotes the Euclidean distance between locations l_i and l_j, σ^2 is the spatial variance, and ρ is the range parameter. This formulation captures how spatial correlation decays with distance, with the exponential term dominating the decay at moderate to large distances due to its super-linear decrease.

To determine appropriate values for ρ that correspond to specific distances at which the spatial correlation becomes negligible (e.g., 0.05), we solve:

$$\exp\left(-\frac{\sqrt{3}d_{ij}}{\rho}\right) = 0.05.$$

Taking natural logarithms:

$$-\frac{\sqrt{3}d_{ij}}{\rho} = \ln(0.05) \approx -3,$$

which leads to:

$$\rho = \frac{\sqrt{3}d_{ij}}{3}.$$

By applying this relationship to various distances where the correlation is intended to diminish significantly, specifically at $D_{\max}/20$, $D_{\max}/10$, and $D_{\max}/2$, we obtain the corresponding ρ values:

$$\rho \in \{\ 4.08,\ 8.16,\ 40.8\}.$$

Here, $D_{\max} = \sqrt{100^2 + 100^2} \approx 141.4$ represents the maximum distance within the 100×100 domain.

In addition to varying ρ, the simulation study explores different levels of spatial variability and measurement noise. The spatial variance σ^2 is set to values $\{0.215,\ 0.86,\ 2.15\}$, representing low, moderate, and high variability, respectively. The nugget effect τ^2, accounting for micro-scale variation or measurement error, is defined as a proportion of σ^2, specifically at 1%, 10%, and 25%. The smoothness parameter ν remains fixed at 1.5 throughout the study to maintain a consistent moderate degree of smoothness in the spatial process.

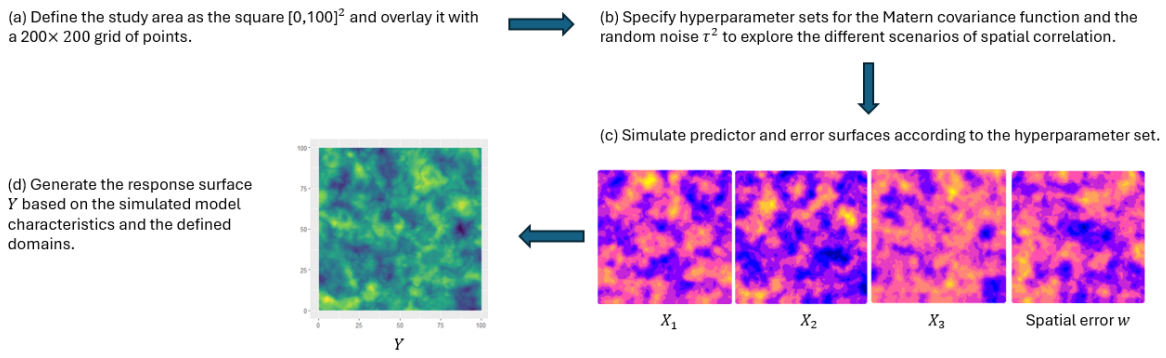

Figure 1: Workflow for generating the response surface. The example fields shown are illustrative and correspond to the low-spatial-autocorrelation hyperparameter set.

5.3 Simulation cases

Driven by the objective to evaluate the performance of the two random forest approaches across diverse spatial fields with varying characteristics, the simulation study is structured into three distinct scenario sets. The first scenario set comprises four different simulated cases, while the second and third sets each consist of three simulated cases. The primary differentiation among these scenario sets lies in the magnitude of the random error component, denoted as τ^2, which represents the nugget effect. Specifically, each scenario set is designed to reflect a different proportion of random error relative to the total error in the simulation model, where the total error is defined as the sum of spatial error and random error.

Within each scenario set, the simulated fields vary according to different values of the range parameter ρ, which influences the extent of spatial autocorrelation. By systematically altering ρ within each set, the study examines how varying degrees of spatial dependence affect model performance under consistent levels of random error. Notably, the first scenario set includes an additional case characterized by negligible spatial autocorrelation, effectively representing a completely random spatial field. This specific case was incorporated solely within the first scenario set to assess model performance in the absence of spatial structure, and was not extended to the other scenario sets, as its inclusion in the initial set sufficiently addressed the evaluation of models under conditions of minimal spatial autocorrelation.

In total, the simulation study encompasses $4 + 3 + 3 = 10$ distinct simulation cases. To ensure robustness and generalizability of the results, each simulation case is replicated 100 times, resulting in a comprehensive dataset of 1,000 simulated models. This extensive simula-

tion framework facilitates a thorough assessment of the random forest approaches under a wide range of spatial conditions.

As outlined in Figure 1, the initial phase of the simulation study involves generating various predictor and spatial error fields, which serve as the foundation for constructing the response domain. Specifically, unconditional sequential Gaussian simulation is employed to create three independent predictor fields, corresponding to the covariates $\{X_1, X_2, X_3\}$, along with a spatial random effect field denoted as $w(l)$. These fields are generated on a regular 100×100 grid, utilizing a Matérn covariance model to define the spatial structure.

Figure 2 presents representative images from four different cases within the first scenario set. In each case, the first four images illustrate the generated covariates and the spatial error field. These components are integrated into the simulation model, as detailed in Section 5.1, to produce the corresponding response field. The four generated fields exemplify varying levels of spatial autocorrelation, thereby enabling the simulation study to assess model performance under diverse spatial conditions.

Case (a) depicts a field characterized by negligible spatial autocorrelation. In this instance, the Matérn covariance function parameters are configured to achieve minimal spatial dependence, with a spatial variance of $\sigma^2 = 0.1$ and a nugget effect of $\tau^2 = 1$. The effective range of spatial autocorrelation in this case is approximately 0.1, which is minimal relative to the maximum distance of approximately 141 units within the domain. Conversely, cases (b), (c), and (d) exhibit progressively increasing levels of spatial autocorrelation. For these cases, the spatial variance is set to $\sigma^2 = 1.72$, and the nugget effect is $\tau^2 = 0.1$. The effective ranges of spatial autocorrelation for these cases are 7.03, 14.1, and 70.36, respectively, reflecting short, moderate, and long-range spatial dependencies.

In Scenario Sets 2 and 3, the range parameter ρ remains consistent with that employed in Scenario Set 1, ensuring that the spatial autocorrelation structures across all scenarios are directly comparable. The primary distinction among these scenario sets lies in the configurations of the nugget parameter τ^2 and the spatial variance σ^2, which modulate the relative contributions of random noise and spatial structure to the total variability in the simulated fields.

In Scenario Set 2, the nugget effect is increased to $\tau^2 = 0.8$, while the spatial variance is maintained at $\sigma^2 = 1.72$, as in Scenario Set 1. This adjustment elevates the proportion of random noise relative to spatially structured variability. Consequently, this scenario set allows for the assessment of model performance under conditions where random noise plays a more

prominent role. Scenario Set 3 is characterized by a further shift in the balance between spatial structure and random noise. Here, the spatial variance is reduced to $\sigma^2 = 0.7$, and the nugget effect is increased to $\tau^2 = 0.6$. By systematically varying τ^2 and σ^2 across these scenario sets, the simulation study comprehensively examines the performance of the random forest methodologies under a spectrum of spatial dependence conditions, ranging from strong spatial structure to dominance by random noise.

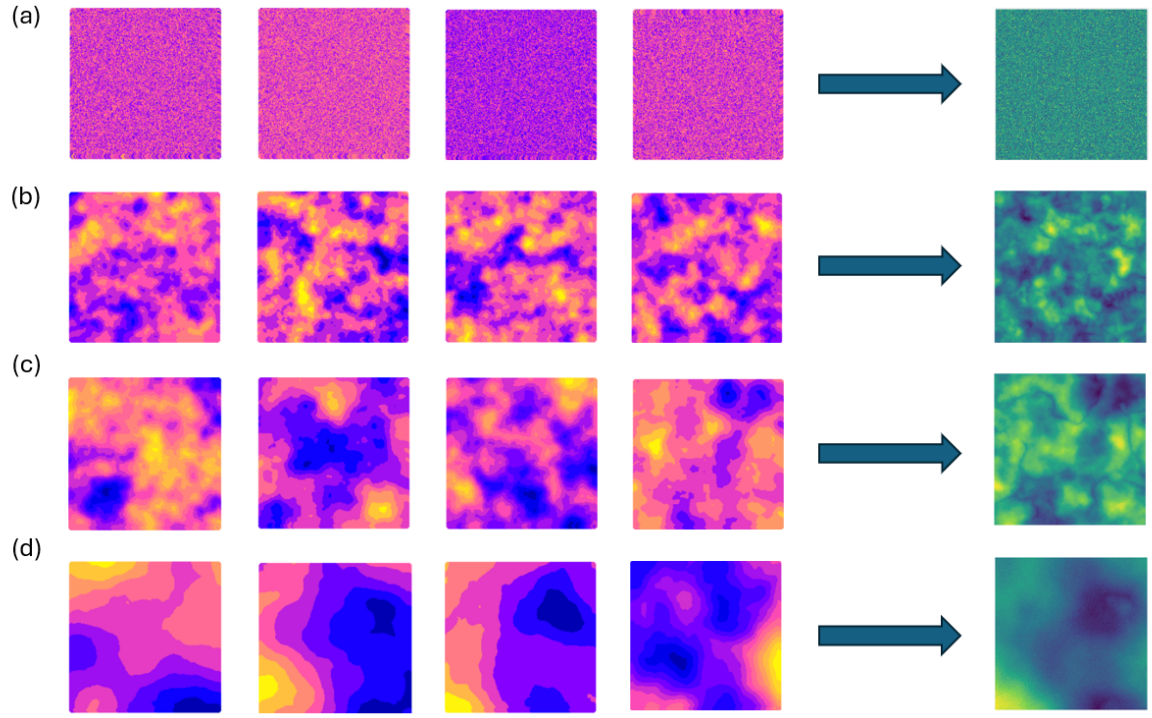

Figure 2: Case (a) of the first simulation scenario consists of fields drawn from a Gaussian Matérn model with effectively zero spatial correlation. Scenarios (b), (c), and (d) all use the same Matérn covariance—partial sill $\sigma^2 = 1.7$, smoothness $\nu = 1.5$, and nugget error zero, differ only in the range parameter ρ, which was set 4, 8 and 40, respectively. The panels in the figure show one representative realization from each scenario; to obtain robust summary statistics, each scenario was simulated 100 times.

5.4 Evaluation of Models' Performance

We evaluated both random-forest methods on two fronts: *prediction* and *estimation*.

Estimation performance was quantified by how accurately each model recovered the underlying mean function $m(\mathbf{x})$, excluding the spatial random effect $w(\ell)$. To do this, we computed

the *Mean Integrated Squared Error (MISE)*:

$$\text{MISE} \;=\; \frac{1}{N}\sum_{i=1}^{N}\big[m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)\big]^2,$$

where $N = 40,000$ is the total number of grid cells and $\hat{m}(\mathbf{x}_i)$ is the model's estimate of the true mean at $\mathbf{x}_i$.

Predictive performance was assessed by how accurately each model forecasted the full response $Y = m(\mathbf{x}) + w(\ell) + \varepsilon$, including the spatial random effect and nugget noise. For each scenario we drew a *training set* of 400 points and an independent *test set* of 100 points , all sampled at random from the 100×100 response domain Figure 3. By ensuring these two sets do not overlap, we evaluated generalization to entirely unseen locations.

Prediction accuracy was then quantified via the *Mean Squared Error (MSE)* on the test set:

$$\text{MSE} \;=\; \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\big(Y_i - \hat{Y}_i\big)^2,$$

where $n_{\text{test}} = 100$, Y_i is the observed response at the ith test point, and $\hat{Y}_i$ is its model prediction.

To stabilize our results, each of the 4 hyperparameter sets was simulated and scored over 100 independent replicates, and we report the average MSE across those runs.

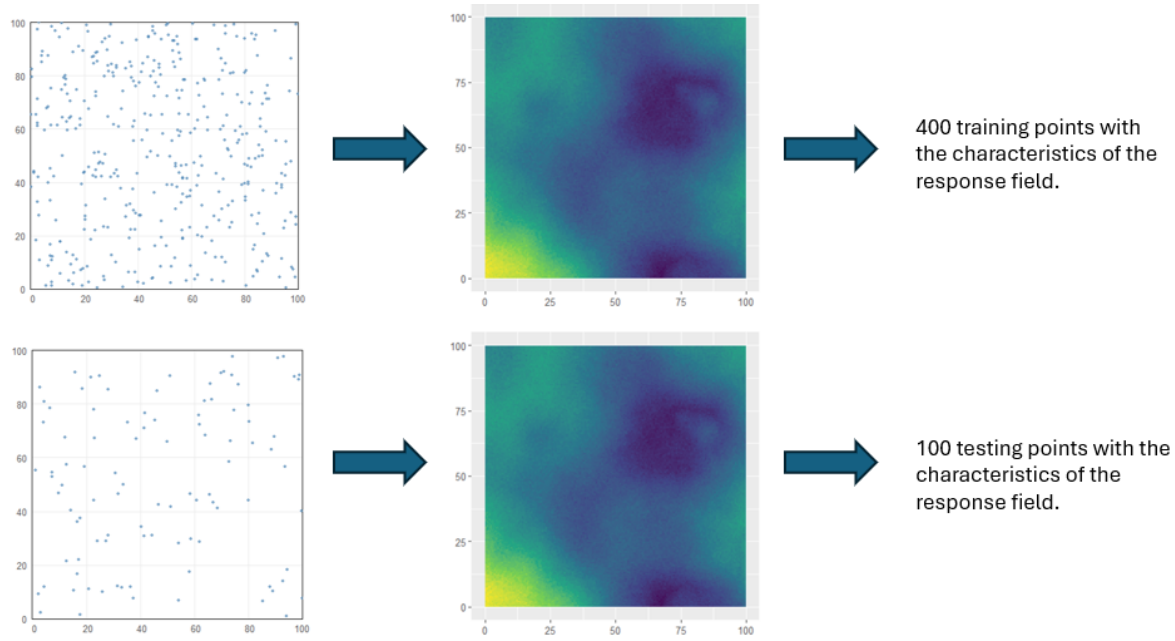

Figure 3: The visualization illustrates the generation of training and test sets used for mean squared error evaluation. The points shown on the left overlap with the response domain depicted on the right.

6 Case study

The Swedish Lake Survey (SLS) is a national monitoring program that, since its inception, has sampled mid-lake surface waters in approximately 4,800 lakes larger than 0.01 km², selected from a stratified national lake registry covering different regions and lake-size classes. Over time, additional lakes were added regionally, bringing the total to about 6,230. Each lake is revisited once every six years, so roughly one-sixth of the lakes are sampled each year. Although the yearly subset of lakes is spread evenly across Sweden, sampling density tends to be higher in the south. Sampling takes place during the autumn circulation period—typically September through December—because conditions then best represent whole-lake water quality. Occasionally, measurements taken as late as January are backdated to the preceding autumn. In autumn 2018 (and January 2019), fewer lakes were surveyed due to a helicopter failure; those missed in 2018 were included in the 2019 sampling. Early data from the program's first year (2007) showed quality issues. For this study we only use measurements collected between 2019 and 2024.

The analysis draws on two types of variables for each lake sampled between 2019 and 2024. First, we have dynamic water-chemistry measurements—parameters whose values change with each visit—of which there is exactly one observation per lake in our timeframe. These include acidity and major ion concentrations such as pH; aluminum (Al, $\mathrm{mg\,L^{-1}}$); arsenic (As, $\mathrm{mg\,L^{-1}}$); calcium (Ca, $\mathrm{mg\,L^{-1}}$); cadmium (Cd, $\mathrm{mg\,L^{-1}}$); chloride (Cl, $\mathrm{mg\,L^{-1}}$); cobalt (Co, $\mathrm{mg\,L^{-1}}$); chromium (Cr, $\mathrm{mg\,L^{-1}}$); copper (Cu, $\mathrm{mg\,L^{-1}}$); and others. Second, static landscape and catchment descriptors are incorporated, assuming they remain unchanged throughout the study period. These consist of lake-specific identifiers and geographic attributes (elevation, unique water-body), catchment area ($\mathrm{m^2}$ and $\mathrm{km^2}$), land-cover proportions (mixed forest, coniferous/broadleaf mix, heathland), and the lake's precise coordinates (latitude and longitude given at the official projected coordinate system for Sweden, defined under EPSG:3006). Together, these two categories let us model both how in-situ chemistry relates to spatial context and how fixed landscape features influence water quality across Sweden.

Our analysis encompasses all lakes monitored by the Swedish Lake Survey (SLS) between 2019 and 2024, covering the entire mainland of Sweden. Due to missing values in some covariate categories, the final dataset for the analysis includes 4641 observations. To explore whether the predictor importance for lake water pH varies geographically, we partitioned the data into four subregions: Norra Inlandet (NI), Östra Mellansverige (ÖM), Norra Kusten (NK)

and Södra Västergötland (SV). NI contains 1,632 observations over 176,399.75 km^2, ÖM has 996 observations in 72,747.00 km^2, NK comprises 238 observations across 29,789.07 km^2, and SV includes 1,775 observations over 87,160.54 km^2. These regions differ not only in their physical extent and catchment characteristics but also markedly in sampling density. Because model performance and variable importance can be sensitive to both environmental heterogeneity and the number of observations available, it will be particularly interesting to investigate whether differences in lake density—on top of each region's unique landscape and chemical profile—lead to systematic shifts in the ranking and strength of our predictors when models are fitted separately within NI, ÖM, NK and SV.

Initially, the standard Random Forest (RF) and the spatial RF-GLS methods are evaluated separately in each subregion using Monte Carlo cross-validation. Twenty independent random splits are performed per region, with 80% of the observations allocated to the training set and 20% to the test set. Three covariate models are fitted in each region—one containing only water-chemistry predictors, one containing only environmental predictors, and one combining both sets (Appendix C). In each split, RF and RF-GLS models are trained on the 80% training data and their predictive performance is assessed on the 20% hold-out data. Accuracy is quantified by the Root-Mean-Squared Error (RMSE) and the coefficient of determination (R^2):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \left(y_i - \hat{y}_i\right)^2} \qquad R^2 = 1 \; - \; \frac{\sum_{i=1}^{N_{\mathrm{test}}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{N_{\mathrm{test}}} \left(y_i - \overline{y}\right)^2},$$

where y_i is the observed pH in the test set, $\hat{y}_i$ the corresponding prediction, $\overline{y}$ the mean of the test-set pH values, and N_{test} the number of test observations. The RMSE and R^2 values reported for each model and region are the averages computed over all twenty random splits.

After completing the evaluation of predictive performance, an extended analysis is conducted to identify which predictors most strongly influence the response variable pH. This analysis is applied separately to each of the four subregions and across all three covariate models.

For each subregion, observations are again split into training (80%) and test (20%) sets. Both Random Forest (RF) and spatial RF–GLS models are fitted on the training data. Once training is complete, the baseline RMSE on the test set is computed using all available predic-

tors:

$$\mathrm{RMSE}_{\mathrm{baseline}} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \big(y_i - \hat{y}_i\big)^2}.$$

Next, for each predictor j, its association with the response is broken by randomly shuffling its values in the test set, and the RMSE is recomputed. This permutation is repeated 100 times per predictor. The increase in RMSE for the ith permutation is

$$\Delta\mathrm{RMSE}_j^{(i)} = \mathrm{RMSE}\big(y_{\mathrm{test}},\, \hat{y}_{perm,j}\big) - \mathrm{RMSE}_{\mathrm{baseline}}.$$

Predictor importance is defined as the average of these 100 increases:

$$\overline{\Delta\mathrm{RMSE}}_j = \frac{1}{100} \sum_{i=1}^{100} \Delta\mathrm{RMSE}_j^{(i)}.$$

A larger value of $\overline{\Delta\mathrm{RMSE}}_j$ indicates that permuting predictor j causes a greater degradation in predictive accuracy, and thus that j is more important. To quantify uncertainty, the 95% confidence interval of the $\{\Delta\mathrm{RMSE}_j^{(i)}\}_{i=1}^{100}$ distribution is also reported for each predictor.

7 Results

7.1 Simulation study

7.1.1 Hyperparameters influence

The primary objective of the simulation study is to compare the performance of the standard Random Forest algorithm and the RF-GLS (Random Forest with Generalized Least Squares) method across fields exhibiting varying degrees of spatial autocorrelation and differing proportions of spatial random error. Prior to applying these methods, a sensitivity analysis of the hyperparameters is conducted to determine how their values affect spatial correlation, which is quantified using Moran's I statistic. All simulated fields are generated by varying key Matérn covariance parameters—specifically the range ρ, the spatial variance σ^2, and the nugget variance τ^2—while holding the smoothness parameter ν constant. Three levels are tested for each parameter:

$$\rho \in \{4.08,\ 8.16,\ 40.8\}, \quad \sigma^2 \in \{0.215,\ 0.86,\ 2.15\}, \quad \tau^2 \in \{0.01\,\sigma^2,\ 0.10\,\sigma^2,\ 0.25\,\sigma^2\}.$$

Each choice of ρ corresponds to distances $D_{\max}/20$, $D_{\max}/10$, and $D_{\max}/2$ at which spatial correlation decays to near zero. By combining all possible triples (ρ, σ^2, τ^2), the study produces $3 \times 3 \times 3 = 27$ distinct response fields, each characterized by a unique level of spatial autocorrelation.

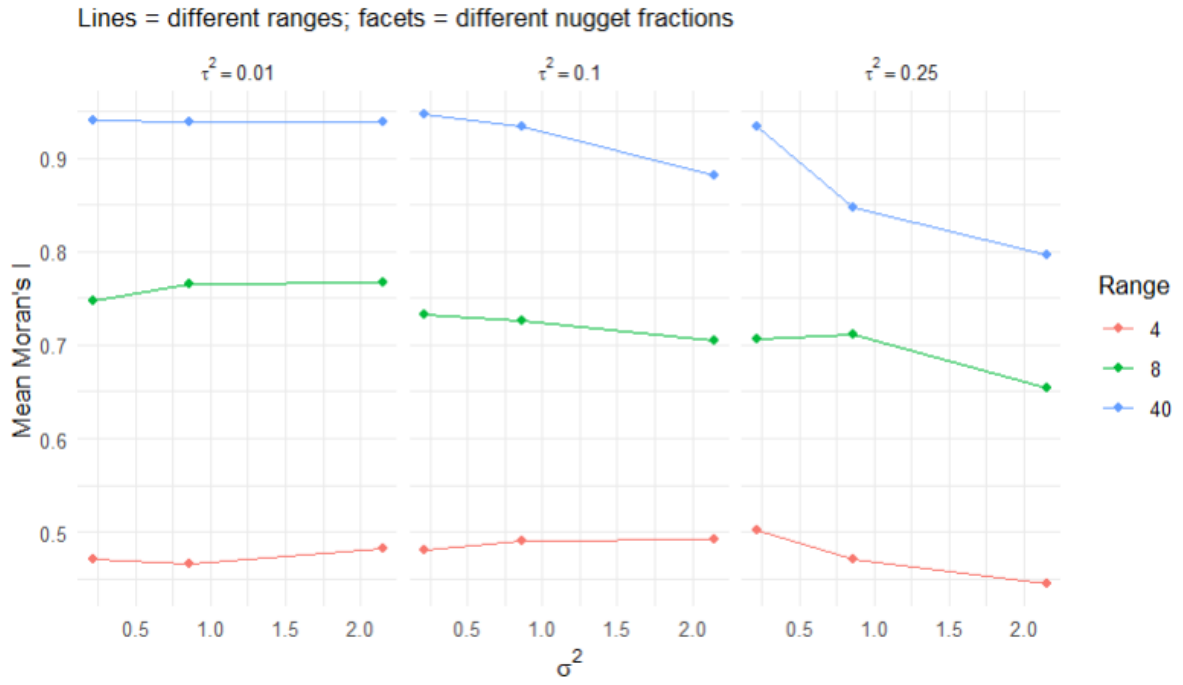

Figure 4: Spatial autocorrelation levels across the 27 different parameter sets.

The analysis revealed a clear increase in spatial autocorrelation as the effective correlation distance grows. In a domain with maximum interpoint distance $D_{\max} \approx 141$, Moran's I remains below 0.5 when the correlation range is $D_{\max}/20$, regardless of the other parameter values. When the range increases to $D_{\max}/10$, Moran's I exceeds 0.7 in most settings, except when the nugget variance τ^2 attains its highest value ($0.25 \times 2.15 = 0.54$). At the largest range ($D_{\max}/2$), Moran's I exceeds 0.9 in the majority of the nine combinations.

By contrast, the effects of σ^2 and τ^2 on spatial autocorrelation are less straightforward. Variations in σ^2 have little impact when τ^2 is only 1% of σ^2; any change in spatial variance under these conditions produces negligible change in Moran's I. However, when τ^2 represents 10% or 25% of σ^2, increasing σ^2 shifts a greater portion of the total variance into the nugget component, thereby reducing spatial autocorrelation.

To ensure robust conclusions, each of the 27 parameter combinations was replicated 100 times, yielding $2,700$ simulated response fields. Figure 4 displays the average Moran's I over these replications for each case.

7.1.2 Comparison of the two methods

Having established how the Matérn hyperparameters influence spatial autocorrelation in the simulated fields, the final phase of the study evaluates the predictive and estimation perfor-

mance of the standard Random Forest and the RF-GLS method across these fields. Figure 2 illustrates the four cases in the first scenario set. Case (a) corresponds to one of the 100 replicated domains with near–zero spatial autocorrelation (mean Moran's $I = -0.001$). Cases (b), (c), and (d) likewise represent individual replications, with average Moran's I values of 0.283, 0.554, and 0.864, respectively. In each of these cases, the spatial variance σ^2 and nugget variance τ^2 were held constant, while the range parameter ρ was set to 4.08, 8.16, and 40.8, respectively, to achieve the desired autocorrelation levels.

The two methods are compared using two evaluation metrics. First, the Mean Integrated Squared Error (MISE) assesses how accurately each method estimates the underlying mean function $m(\cdot)$. Second, the Mean Squared Error (MSE) quantifies each method's predictive accuracy on unseen data.

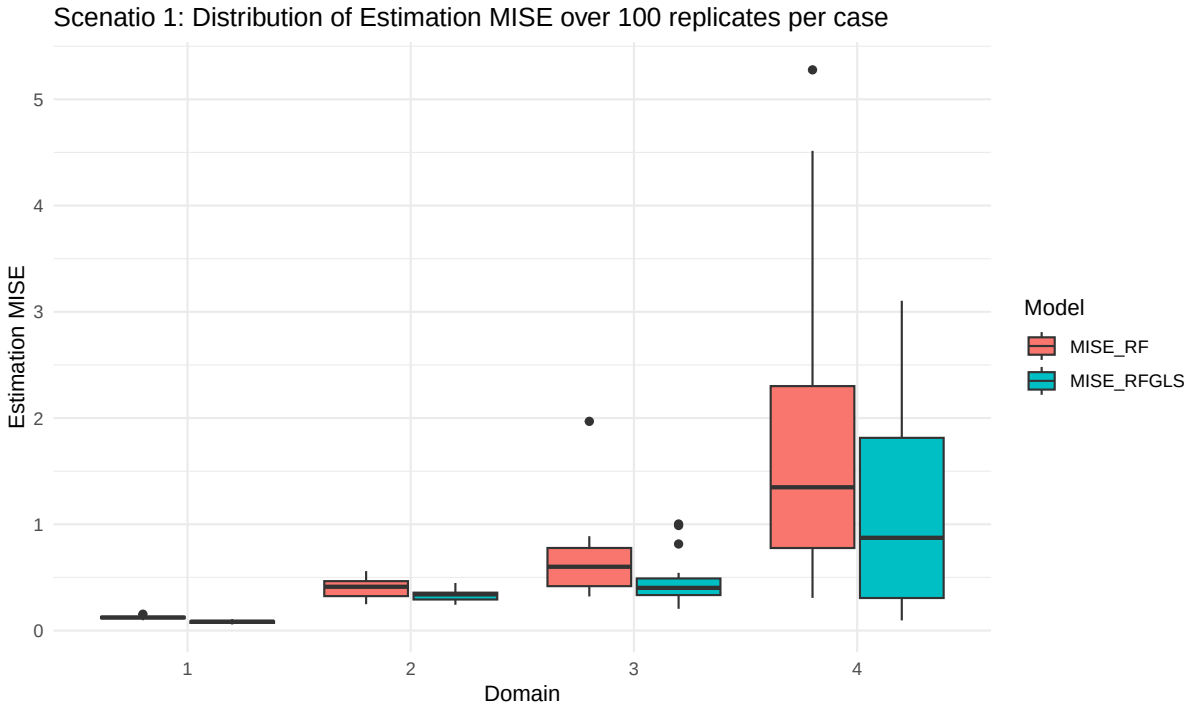

Figure 5: Estimation performance comparison of the two models based on mean squared error (MISE) across the four cases.

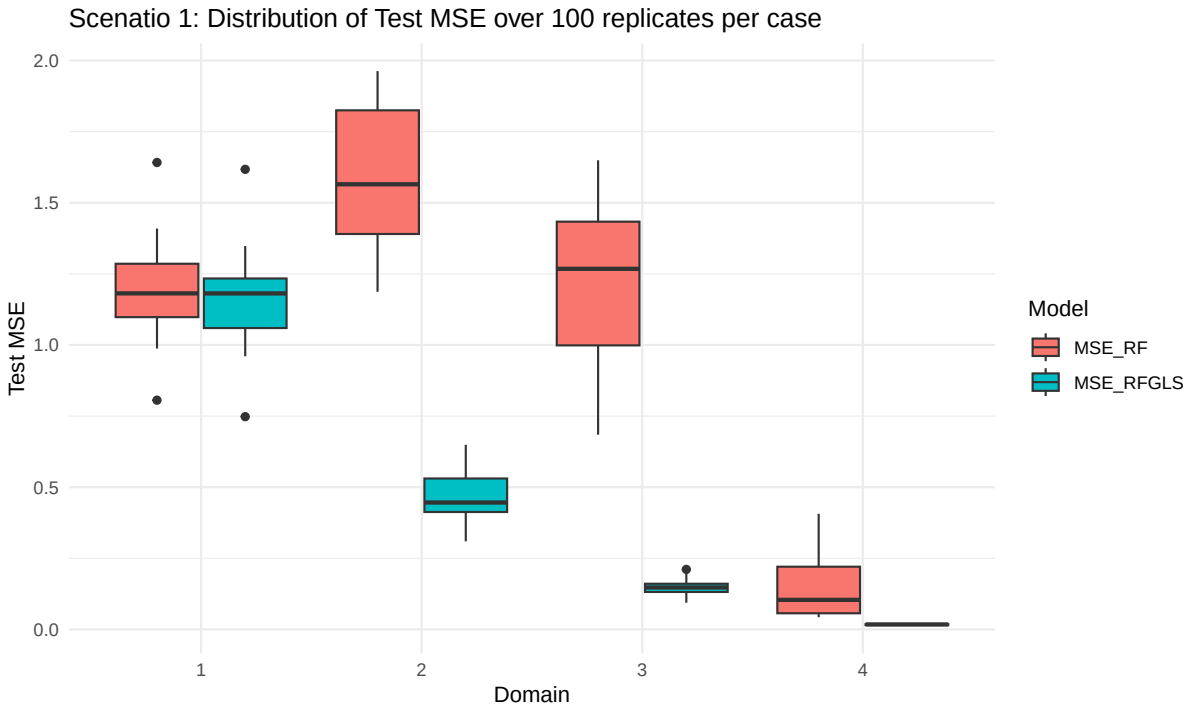

Figure 6: Predictive performance comparison of the two models based on mean squared error (MSE) across the four cases.

The results for estimation performance, shown in Figure 5, indicate that RF-GLS outperforms the standard Random Forest in all four cases, including the field with negligible spatial autocorrelation. Specifically, both methods achieve their best performance when Moran's $I \approx 0$, with average MISE values of 0.12 for Random Forest and 0.08 for RF-GLS. In this first case, both methods also exhibit minimal variance in their estimates. As spatial autocorrelation increases, the MISE and its variability grow for both models. However, the gap between methods widens: RF-GLS consistently produces lower error than Random Forest, demonstrating its advantage in estimating the mean function $m(\mathbf{X})$ when the covariates $\mathbf{X}$ exhibit strong spatial dependence.

Prediction performance, measured by MSE, further confirms RF-GLS's superiority. In the moderately autocorrelated case (Moran's $I \approx 0.28$), Random Forest and RF-GLS perform similarly, with average MSEs of 1.22 and 1.19, respectively. As autocorrelation intensifies, RF-GLS's prediction error decreases and its variability narrows, while Random Forest's error remains high. For example, at Moran's $I \approx 0.28$, Random Forest yields MSE $= 1.63$ (SD=0.10), versus RF-GLS's MSE $= 0.48$ (SD=0.02). At $I \approx 0.55$, the gap is even larger: Random Forest MSE $= 1.24$ (SD=0.15) versus RF-GLS MSE $= 0.14$ (SD=0.009). Interest-

ingly, in the highest-autocorrelation case ($I \approx 0.86$), Random Forest's prediction error drops to MSE=0.13 (SD=0.009), matching RF-GLS. This convergence occurs because both the covariate and spatial error surfaces become nearly smooth "hills," making the domain easy to predict for either method.

All cases except (a), which represents a random field with zero spatial autocorrelation, differ only in the range parameter ρ. In cases (b), (c), and (d), the ratio $\tau^2/\sigma^2 = 0.1/1.72 \approx 0.06$ implies that the nugget effect has virtually no impact on the generated domains. This negligible influence of random noise yields highly smooth, almost deterministic fields—an idealized scenario that underscores the need to evaluate model performance under less favorable conditions. Accordingly, the second scenario set introduces three additional cases in which ρ and σ^2 remain as before, but the random variance is increased to $\tau^2 = 0.8$, raising the ratio to $\tau^2/\sigma^2 \approx 0.47$. These fields exhibit substantial microscale variability, breaking the near-deterministic smoothness observed previously and providing a more challenging test of both Random Forest and RF-GLS.

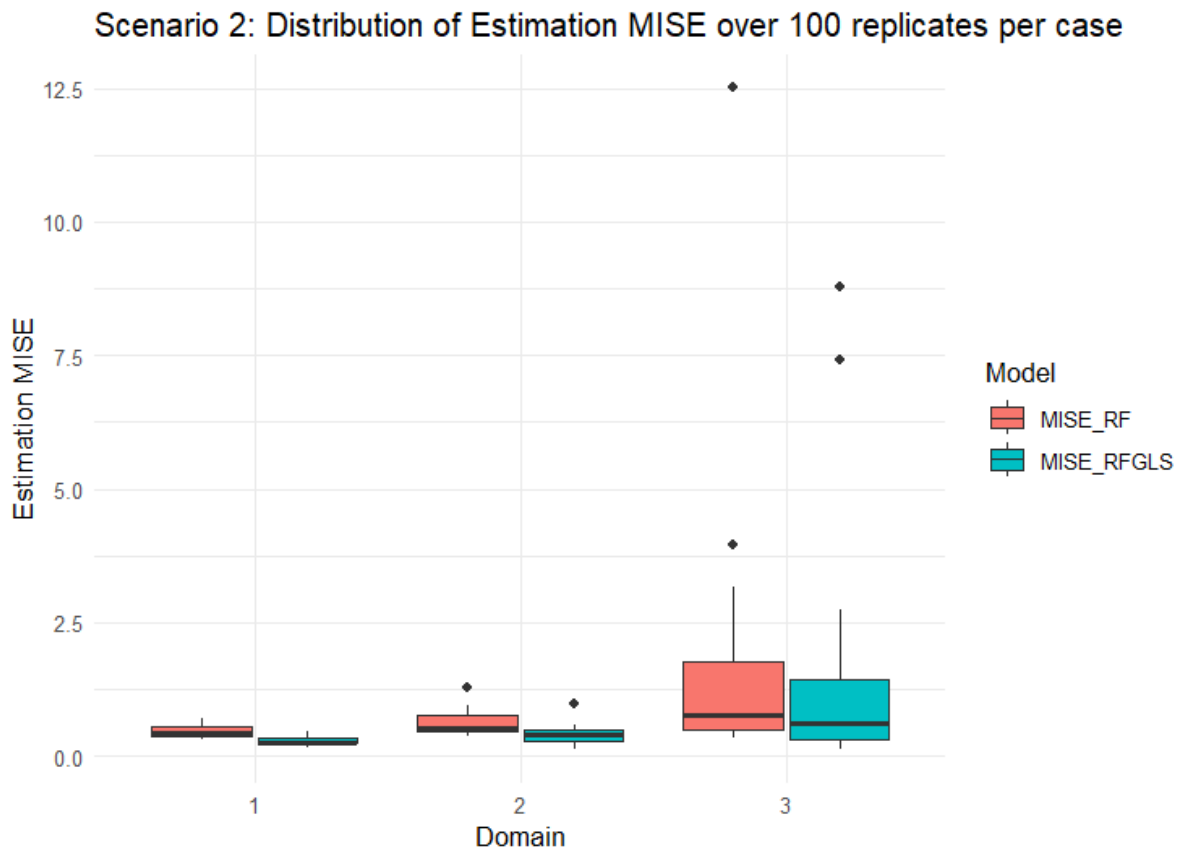

Figure 7: Estimation performance comparison of the two models based on mean squared error (MISE) across the four cases.

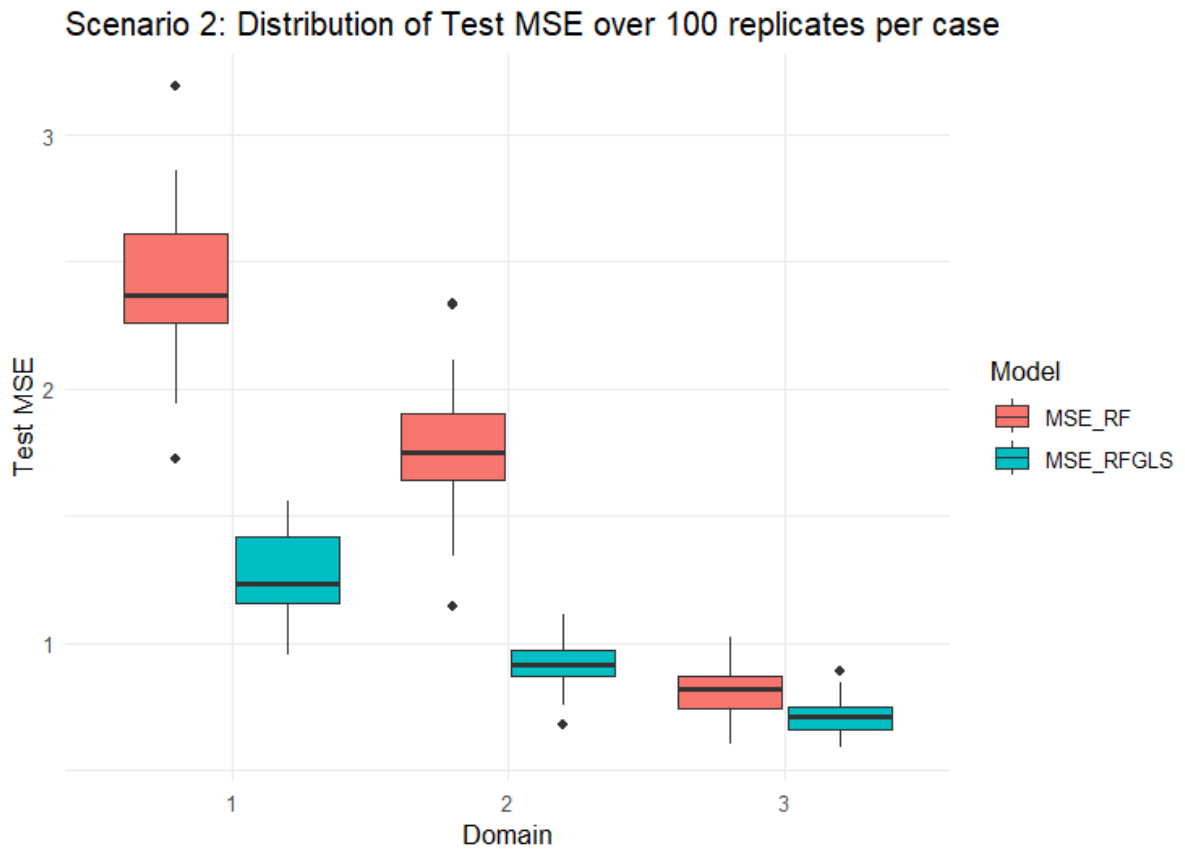

Figure 8: Predictive performance comparison of the two models based on mean squared error (MSE) across the four cases.

The estimation and prediction results for Scenario Set 2 (Figures 7 and 8) mirror those from Scenario Set 1, reaffirming that RF-GLS outperforms the standard Random Forest across all simulated fields. By increasing the nugget variance τ^2, the predictive performance gap narrows—particularly in the cases with mean Moran's $I \approx 0.532$ and 0.764. In these two settings, RF-GLS exhibits slightly higher mean MSE and greater variability compared to the corresponding cases in Scenario Set 1. This indicates that when random noise becomes a larger proportion of total variability—even in fields with moderate to strong spatial autocorrelation—RF-GLS retains its advantage but at the cost of some loss in accuracy and consistency.

Scenario Set 3 (Figures9 and 10) further adjusts the balance between spatial and nugget variance by reducing σ^2 to 0.7 and setting τ^2 to 0.6. Comparison with Scenario Set 2 reveals negligible differences in both estimation and prediction metrics. Notably, for the case with the largest range ($D_{\max}/2$), RF-GLS's MISE variability increases slightly relative to the analogous case in Scenario Set 2, suggesting that narrowing the gap between σ^2 and τ^2 modestly undermines estimation stability. Prediction trends, however, remain effectively unchanged from

Scenario Set 2.

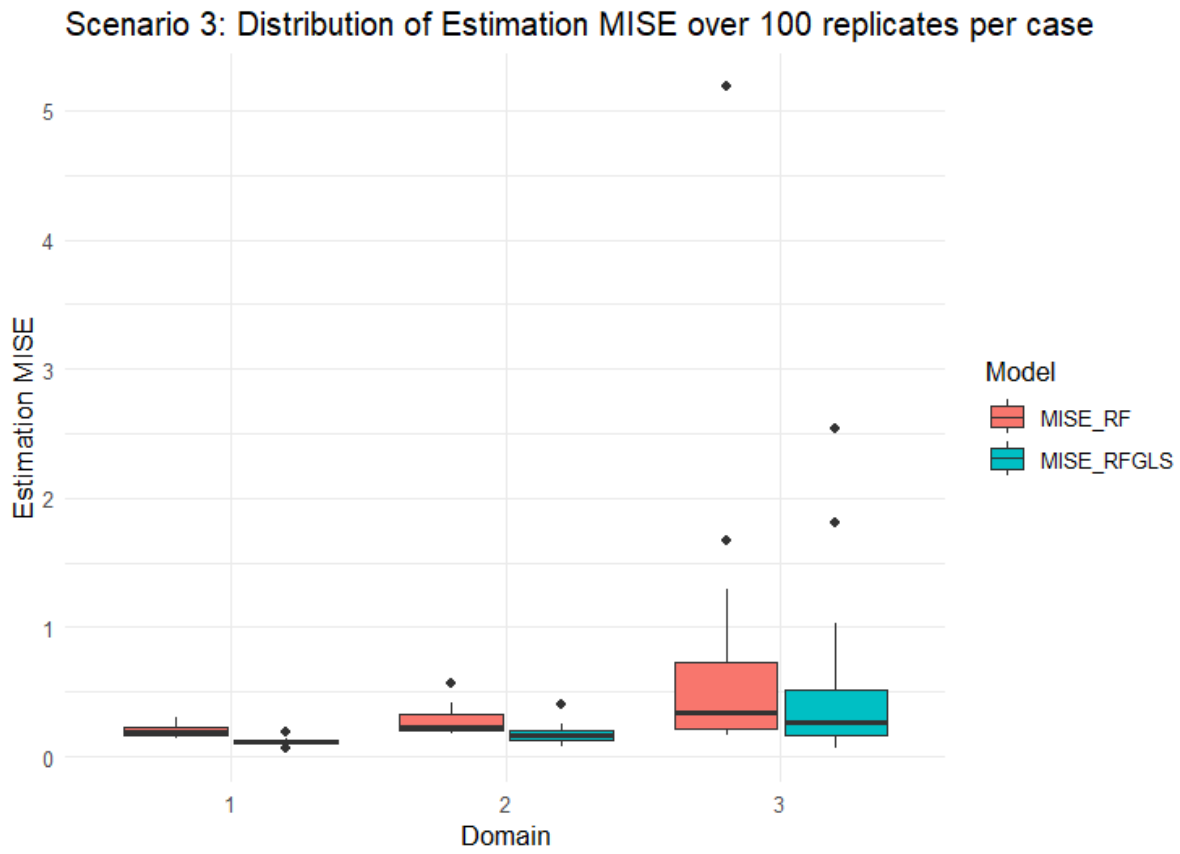

Figure 9: Estimation performance comparison of the two models based on mean squared error (MISE) across the four cases.

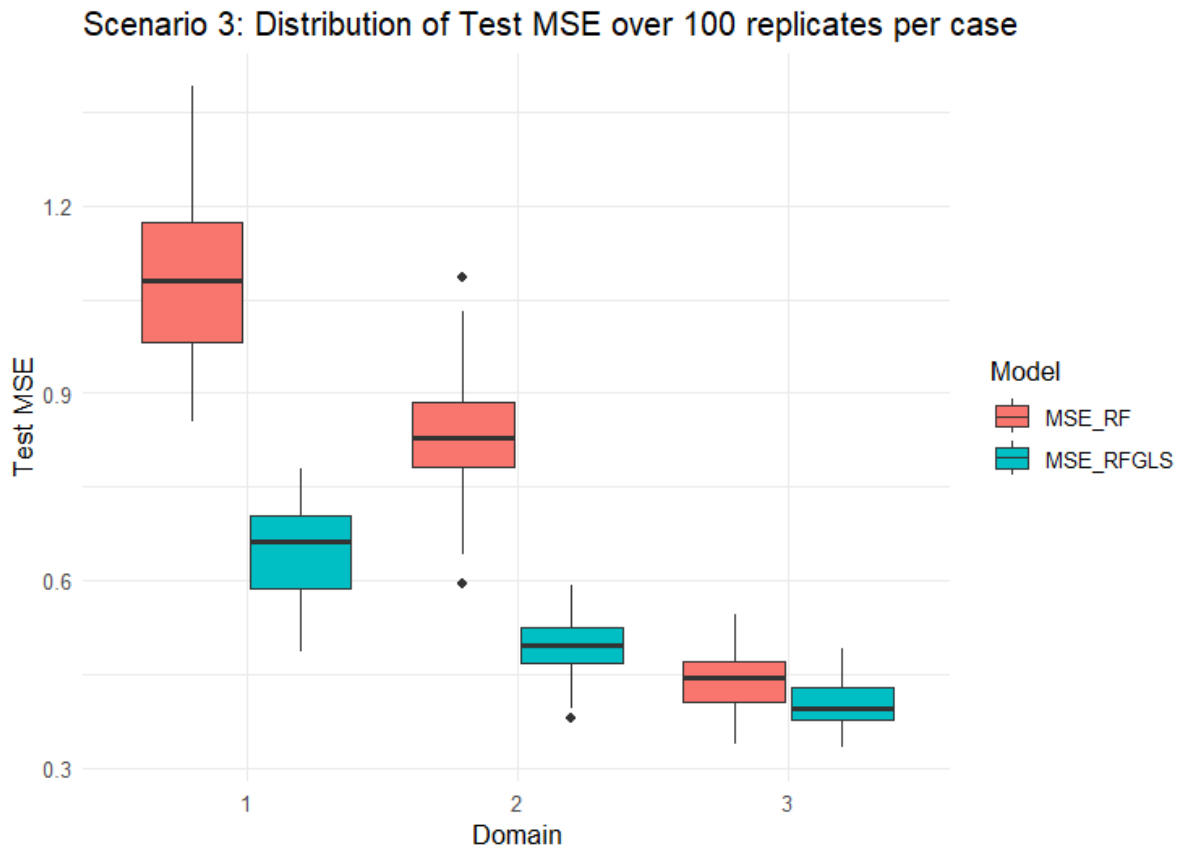

Figure 10: Predictive performance comparison of the two models based on mean squared error (MSE) across the four cases.

7.2 Case study

7.2.1 Spatial Autocorrelation levels

By applying Moran's I test first to the entire dataset for Sweden—all $4{,}641$ lakes with pH measurements—the following results were obtained (see Figure 11):

The Moran's I test, under the randomization assumption, yielded an observed statistic

$$I_{\text{obs}} = 0.360466,$$

whereas under H_0 (complete spatial randomness)

$$\mathbb{E}[I] = -2.155 \times 10^{-4}, \quad \text{Var}(I) = 3.884 \times 10^{-5}.$$

Standardizing gives

$$z = \frac{0.360466 - \left(-2.155 \times 10^{-4}\right)}{\sqrt{3.884 \times 10^{-5}}} = 57.872,$$

with a one-sided p-value reported as $< 2.2 \times 10^{-16}$. Because this p-value is far below any conventional significance level (e.g. $\alpha = 0.05$), H_0 is rejected. The observed Moran's $I \approx 0.36$ indicates moderate positive spatial autocorrelation: lakes that are geographically close tend to have more similar pH values than those farther apart.

Applying the same test to each of the four subregions of environmental interest also yields significant autocorrelation, but at different levels (see Figure 12). Östra Mellansverige (ÖM) exhibits the highest Moran's I at 0.39, while Norra Kusten (NK) has the lowest at 0.10. Norra Inlandet (NI) and Södra Västergötland (SV) fall in between, with Moran's I values of 0.29 and 0.24, respectively. These differences in spatial autocorrelation, together with variations in sample density, make these regions an interesting subject for evaluating the two Random Forest methods.

7.2.2 Predictive performance

	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
		NI			NK			SV			ÖM	
Metric	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
RF												
RMSE	0.2,(0.02)	0.503, (0.015)	0.218, (0.03)	0.234, (0.03)	0.457, (0.07)	0.243, (0.036)	0.21, (0.007)	0.587, (0.02)	0.213, (0.01)	0.252, (0.013)	0.45, (0.021)	0.252, (0.018)
R^2	0.894, (0.027)	0.269, (0.021)	0.872, (0.026)	0.801, (0.05)	0.218, (0.12)	0.809, (0.051)	0.935, (0.004)	0.47, (0.034)	0.936, (0.006)	0.89, (0.012)	0.599, (0.048)	0.877, (0.014)
RF-GLS												
RMSE	0.212, (0.021)	0.505, (0.015)	0.228, (0.028)	0.232, (0.03)	0.464, (0.069)	0.236, (0.041)	0.215, (0.01)	0.599, (0.016)	0.218, (0.01)	0.248, (0.013)	0.439, (0.024)	0.246, (0.016)
R^2	0.888, (0.027)	0.262, (0.021)	0.858, (0.025)	0.812, (0.03)	0.203, (0.102)	0.81, (0.051)	0.929, (0.007)	0.446, (0.032)	0.928, (0.006)	0.896, (0.014)	0.619, (0.05)	0.885, (0.015)

Table 1: Prediction Performance of RF and RF-GLS Across Regions (Values in Parentheses Indicate Standard Deviations).

Table 1 shows the mean out-of-sample RMSE and R^2 (with standard deviations in parentheses), averaged over twenty Monte Carlo splits, for both the standard Random Forest (RF) and the Random Forest–Generalized Least Squares (RF-GLS) models in each of the four study regions (NI, NK, SV, and ÖM) and for each of the three model covariate specifications. Across all regions and covariate configurations, RF and RF-GLS deliver very similar scores, with differences in RMSE and R^2 typically low. Neither method consistently dominates the other by

a wide margin. Both algorithms agree strongly that Model 1—comprising water-chemistry predictors—outperforms Model 2, which relies on geographic and identifier covariates, in every region. This pattern holds for both RMSE and R^2, suggesting that variables related to water chemistry carry substantially more predictive information for lake pH than do purely spatial or administrative attributes. Interestingly, Model 3—which combines both sets of predictors—yields performance nearly to that of Model 1 in all the different regions. This finding reinforces the primacy of the water-chemistry covariates: adding geographic or identifier variables to the water-chemistry set does not materially improve prediction.

7.2.3 Variables Importance

Appendix D shows the top covariates—those most affecting pH—for each region and model. RF and RF-GLS largely agree on key predictors. Discrepancies arise mostly among variables with near-zero importance, where small numerical differences can change lower rankings. In a few cases, even top predictors differ slightly in order, though both methods identify the same high-impact variables.

Following our evaluation of predictive performance, it was evident that Model 2 predictors consistently yielded poorer test-set accuracy than those of Model 1 across all four subregions. The permutation-importance analysis further supports this result. In Model 3—which includes all 40 covariates (Figures 15, 18, 21, and 24)—the geographic and identifier variables invariably occupy the lowest ranks: permuting them produces almost no increase in RMSE.

Moreover, in Model 2 (Figures 14, 17, 20, and 23), several predictors such as semi-urban cover, impervious surface, and heathland have negative or partially negative 95% confidence intervals for $\Delta\mathrm{RMSE}$. This indicates that, when these covariates are shuffled, predictive performance often improves rather than deteriorates.

Lastly, it examined whether different regions exhibit distinct key drivers of pH. Having established that geographic and identifier covariates have negligible influence, this analysis is restricted to the water-chemistry models. Across all regions, the same eight predictors dominate—with only minor reordering. Notably, in Norra Inlandet and Östra Mellansverige, specific conductance at 25 celsius degrees is the second most important variable after calcium, whereas in Norra Kusten and Södra Västergötland it falls slightly lower. Similarly, cobalt ranks second only in Norra Kusten, but is less important elsewhere. Overall, calcium concentration overwhelmingly governs pH, and Model 2 predictors remain of low relevance throughout.

8 Discussion

Our simulation study demonstrated that RF–GLS is a strong candidate for spatially correlated data: it consistently outperformed the standard Random Forest under ideal hyperparameter settings, and although its advantage diminished as noise levels increased, it still matched or exceeded RF performance across all scenarios.

In the real-world case study, RF–GLS and RF achieved almost identical predictive accuracy on the four Swedish subregions. The smaller gap—compared to the simulations—likely reflects two factors: (1) the simulation used only three covariates, whereas the case study models employed up to 40 predictors, and (2) real field data exhibit more complex variability than can be captured by simple hyperparameter perturbations.

Both methods agreed closely on variable importance. In every model and region, calcium concentration (`ca_mg_l`) emerged as the dominant predictor of pH. RF and RF–GLS also concurred that geographic and identifier covariates have negligible influence. A single exception occurred in the Norra Kusten region under RF, where cobalt concentration (`co_mg_l`) ranked second when used the whole set of covariates, model 3.

The limited importance of geographic variables such as land use can be explained by the scale at which they were defined—on a catchment level, which covers a relatively large area. Because these variables represent broad, aggregated characteristics, they may be too coarse or general to capture the local variations that directly affect lake pH, reducing their usefulness in predicting pH accurately.

Overall, the case study confirms that RF–GLS produces sensible and robust results in practice. Although the four subregions exhibit only low to moderate spatial autocorrelation in pH, RF–GLS matched RF performance despite the extra modeling complexity. In applications with stronger spatial structure, RF–GLS may yield even greater gains.

The primary drawback of RF–GLS remains its computational cost. While methods such as the Nearest Neighbor Gaussian Process (NNGP) approximation can mitigate this burden (see Section 4.3), RF–GLS still runs substantially slower than a standard Random Forest implementation.

References

[1] Thorsten Behrens et al. “Spatial modelling with Euclidean distance fields and machine learning”. In: *European journal of soil science* 69.5 (2018), pp. 757–770.

[2] Roger S Bivand and David WS Wong. “Comparing implementations of global and local indicators of spatial association”. In: *Test* 27.3 (2018), pp. 716–748.

[3] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[4] Leo Breiman et al. *Classification and regression trees*. Routledge, 2017.

[5] Richard H Byrd et al. “A limited memory algorithm for bound constrained optimization”. In: *SIAM Journal on scientific computing* 16.5 (1995), pp. 1190–1208.

[6] Abhirup Datta et al. “Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets”. In: *Journal of the American Statistical Association* 111.514 (2016), pp. 800–812.

[7] Ibrahim Fayad et al. “Regional scale rain-forest height mapping using regression-kriging of spaceborne and airborne LiDAR data: Application on French Guiana”. In: *Remote Sensing* 8.3 (2016), p. 240.

[8] Alan E. Gelfand and Erin M. Schliep. “Spatial statistics and Gaussian processes: A beautiful marriage”. In: *Spatial Statistics* 18 (2016), pp. 86–104. DOI: 10.1016/j.spasta.2016.03.006. URL: https://doi.org/10.1016/j.spasta.2016.03.006.

[9] Stefanos Georganos et al. “Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling”. In: *Geocarto International* 36.2 (2021), pp. 121–136.

[10] Peter Guttorp and Tilmann Gneiting. “Studies in the history of probability and statistics XLIX on the Matérn correlation family”. In: *Biometrika* 93.4 (2006), pp. 989–995.

[11] Tomislav Hengl et al. “Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables”. In: *PeerJ* 6 (2018), e5518.

[12] Steffen L Lauritzen. *Graphical models*. Vol. 17. Clarendon Press, 1996.

[13] Dong C Liu and Jorge Nocedal. “On the limited memory BFGS method for large scale optimization”. In: *Mathematical programming* 45.1 (1989), pp. 503–528.

[14] Carles Milà et al. “Random forests with spatial proxies for environmental modelling: opportunities and pitfalls”. In: *Geoscientific Model Development* 17.15 (2024), pp. 6007–6033.

[15] Patrick AP Moran. “Notes on continuous stochastic phenomena”. In: *Biometrika* 37.1/2 (1950), pp. 17–23.

[16] Emilio Porcu et al. “The Matérn model: A journey through statistics, numerical analysis and machine learning”. In: *Statistical Science* 39.3 (2024), pp. 469–492.

[17] Arkajyoti Saha, Sumanta Basu, and Abhirup Datta. “Random Forests for Spatially Dependent Data”. In: *Journal of the American Statistical Association* 118.541 (2023), pp. 665–683. DOI: 10.1080/01621459.2021.1950003. URL: https://doi.org/10.1080/01621459.2021.1950003.

[18] Arkajyoti Saha, Sumanta Basu, and Abhirup Datta. “RandomForestsGLS: An R Package for Random Forests With Spatially Correlated Errors”. In: *Journal of Open Source Software* 7.72 (2022), p. 3780. DOI: 10.21105/joss.03780. URL: https://doi.org/10.21105/joss.03780.

[19] Arkajyoti Saha and Abhirup Datta. “BRISC: Fast inference for large spatial datasets using BRISC”. In: *R package version* 1.5 (2022).

[20] Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

[21] Jonathan R Stroud, Michael L Stein, and Shaun Lysen. “Bayesian and maximum likelihood estimation for Gaussian processes on an incomplete lattice”. In: *Journal of computational and Graphical Statistics* 26.1 (2017), pp. 108–120.

[22] Aldo V Vecchia. “Estimation and model identification for continuous spatial processes”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 50.2 (1988), pp. 297–312.

9 Appendix

A pH observations across Sweden

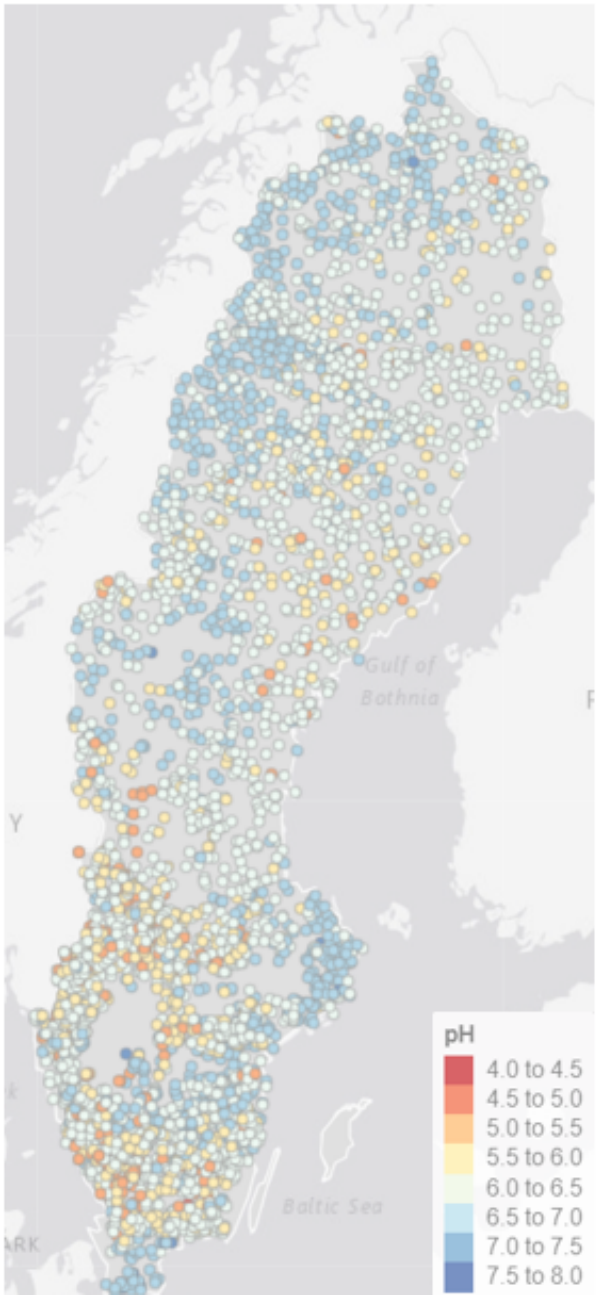

Figure 11: Map of pH values

B Spatial autocorrelation for the four subregions.

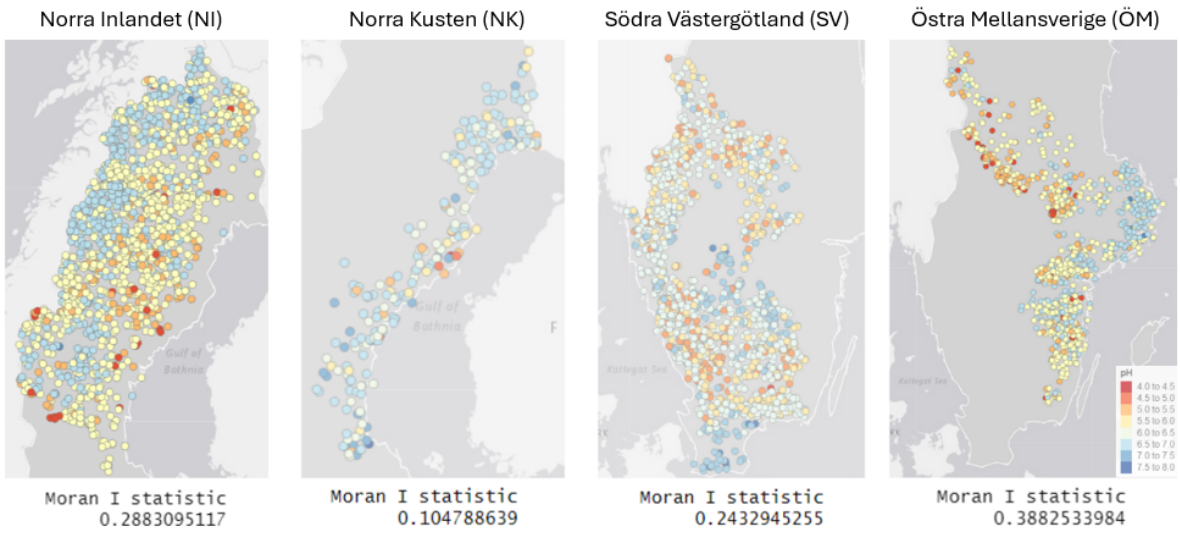

Figure 12: Moran's I test results per region

C Covariates list.

List of response and predictors used in each of the three models. All data are from the Cyclically low-intensity monitored lakes program.

Covariate	Category	Model1	Model2	Model3
pH	Response	Response	Response	Response
Aluminium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Arsenic (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Calcium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Cadmium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Chloride (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Cobalt (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Chromium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Copper (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Fluoride (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Iron (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Potassium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Conductivity 25 °C (mS m^{-1})	Water chemistry	Predictor	—	Predictor
Magnesium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Manganese (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Sodium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Ammonium (mg L^{-1} N)	Water chemistry	Predictor	—	Predictor
Nickel (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Nitrate + Nitrite (mg L^{-1} N)	Water chemistry	Predictor	—	Predictor
Lead (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Silica (mg L^{-1})	Water chemistry	Predictor	—	Predictor
TOC (mg L^{-1} C)	Water chemistry	Predictor	—	Predictor
Total N (mg L^{-1} N)	Water chemistry	Predictor	—	Predictor
Vanadium (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Zinc (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Sulfate (mg L^{-1})	Water chemistry	Predictor	—	Predictor
Elevation (m)	Landscape & catchment	—	Predictor	Predictor
Catchment area (km^2)	Landscape & catchment	—	Predictor	Predictor
Mixed forest	Landscape & catchment	—	Predictor	Predictor
Conifer-broadleaf forest	Landscape & catchment	—	Predictor	Predictor
Heathland	Landscape & catchment	—	Predictor	Predictor
Broadleaf forest	Landscape & catchment	—	Predictor	Predictor
Impervious surface	Landscape & catchment	—	Predictor	Predictor
Broadleaf deciduous forest	Landscape & catchment	—	Predictor	Predictor
Semi-urban	Landscape & catchment	—	Predictor	Predictor
Lakes & waterways	Landscape & catchment	—	Predictor	Predictor
Wetland forest	Landscape & catchment	—	Predictor	Predictor
Pine forest	Landscape & catchment	—	Predictor	Predictor
Young forest	Landscape & catchment	—	Predictor	Predictor
Arable land	Landscape & catchment	—	Predictor	Predictor
Open wetland	Landscape & catchment	—	Predictor	Predictor

Table 2: Covariates list for each model.

D Variables importance results.

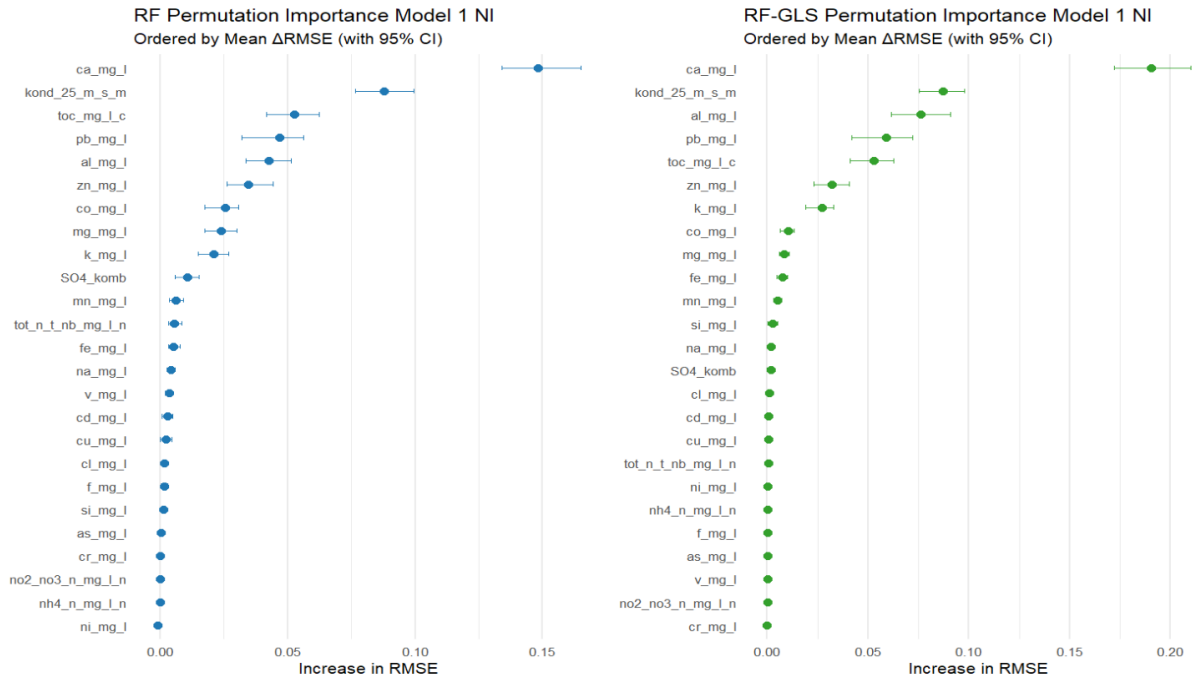

Figure 13: Variable importace for model 1 in region Norra Inlandet.

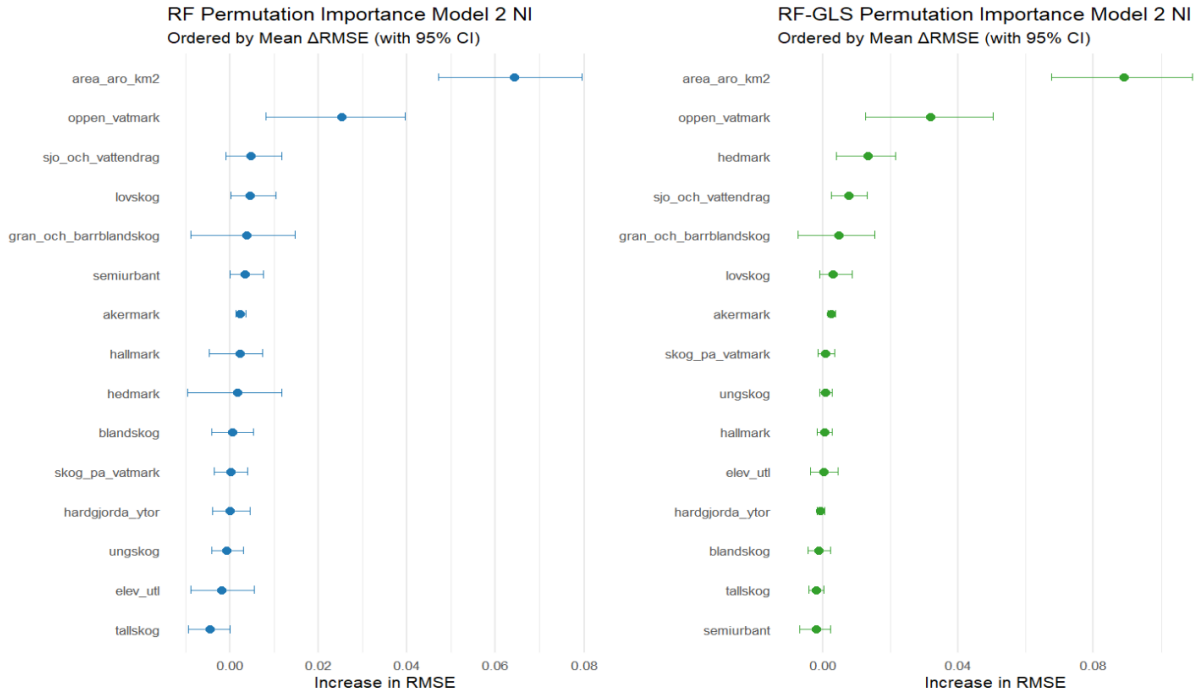

Figure 14: Variable importace for model 2 in region Norra Inlandet.

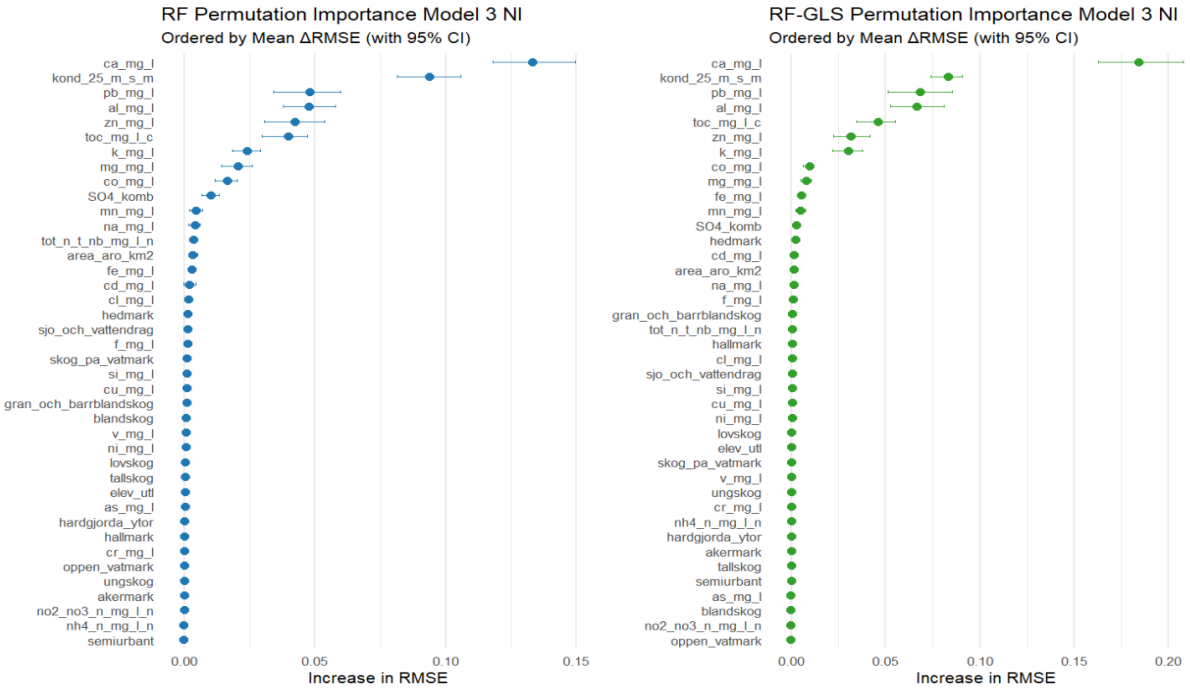

Figure 15: Variable importace for model 3 in region Norra Inlandet.

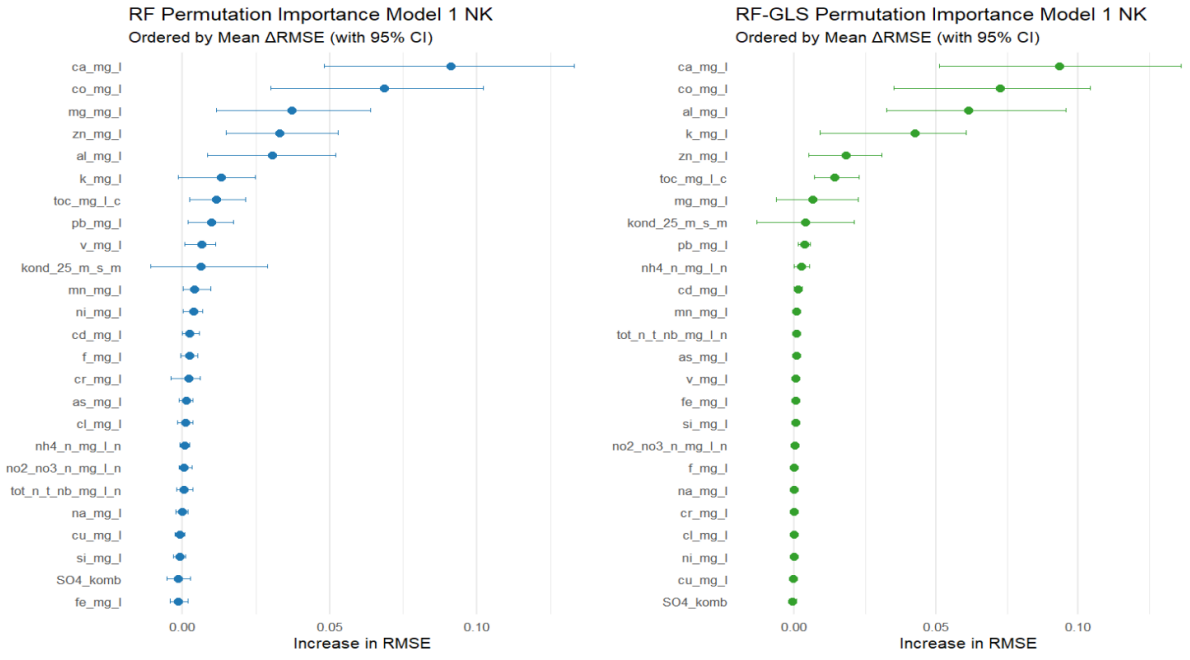

Figure 16: Variable importace for model 1 in region Norra Kusten.

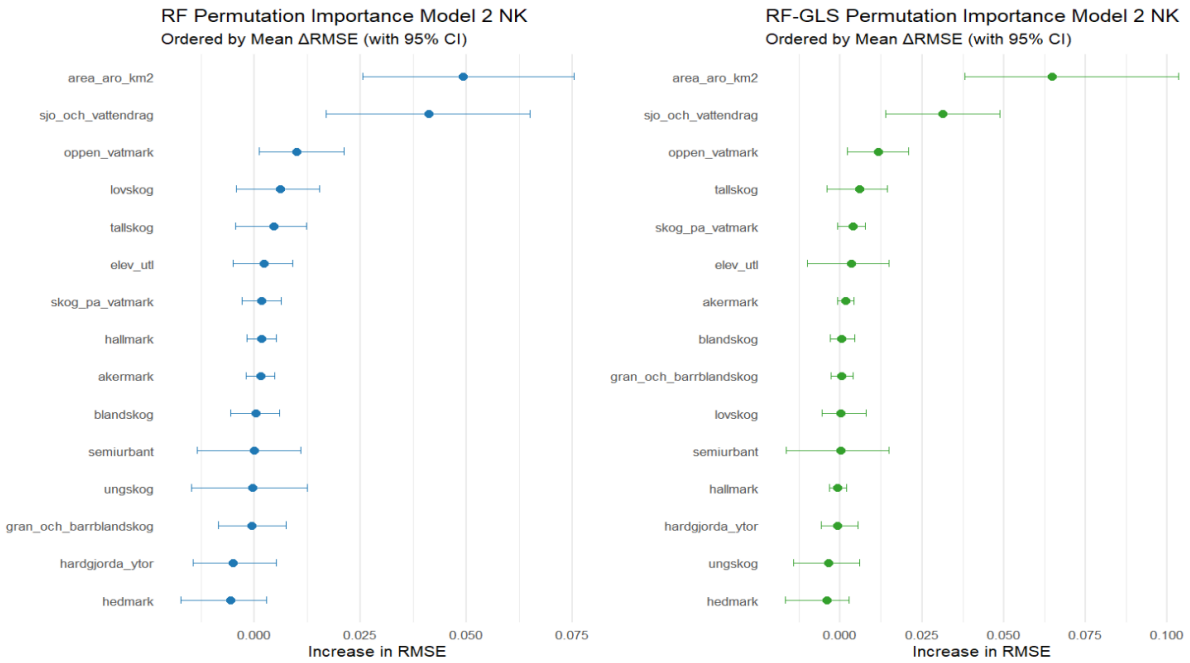

Figure 17: Variable importace for model 2 in region Norra Kusten.

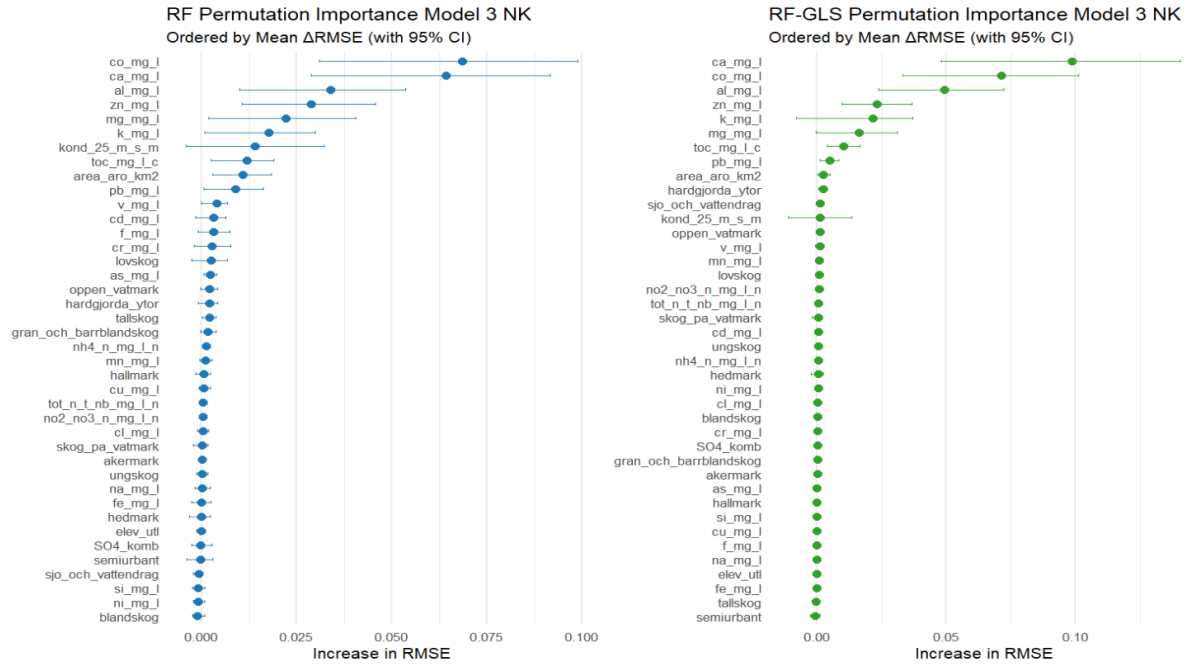

Figure 18: Variable importace for model 3 in region Norra Kusten.

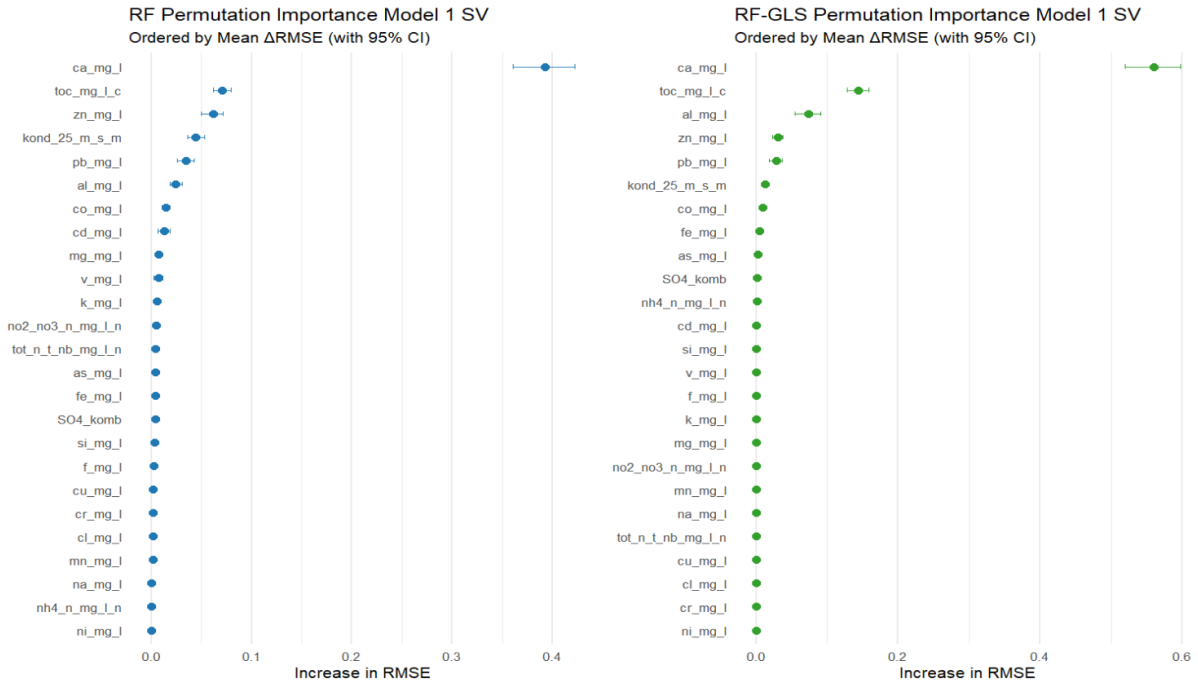

Figure 19: Variable importace for model 1 in region Södra Västergötlan.

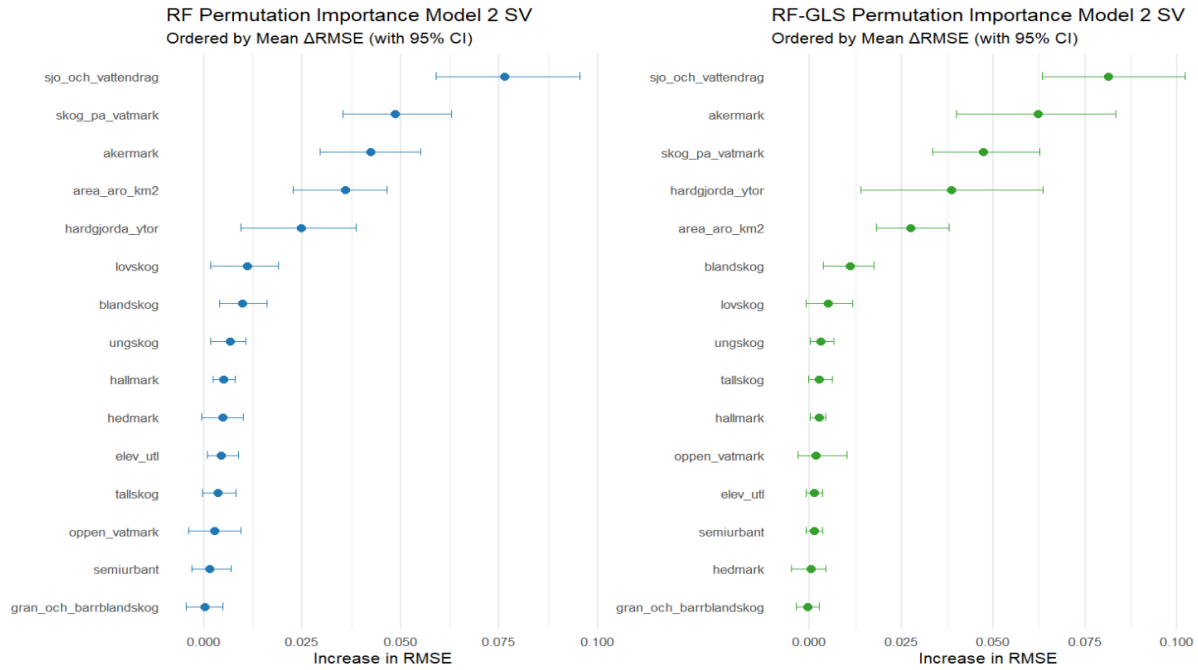

Figure 20: Variable importace for model 2 in region Södra Västergötlan.

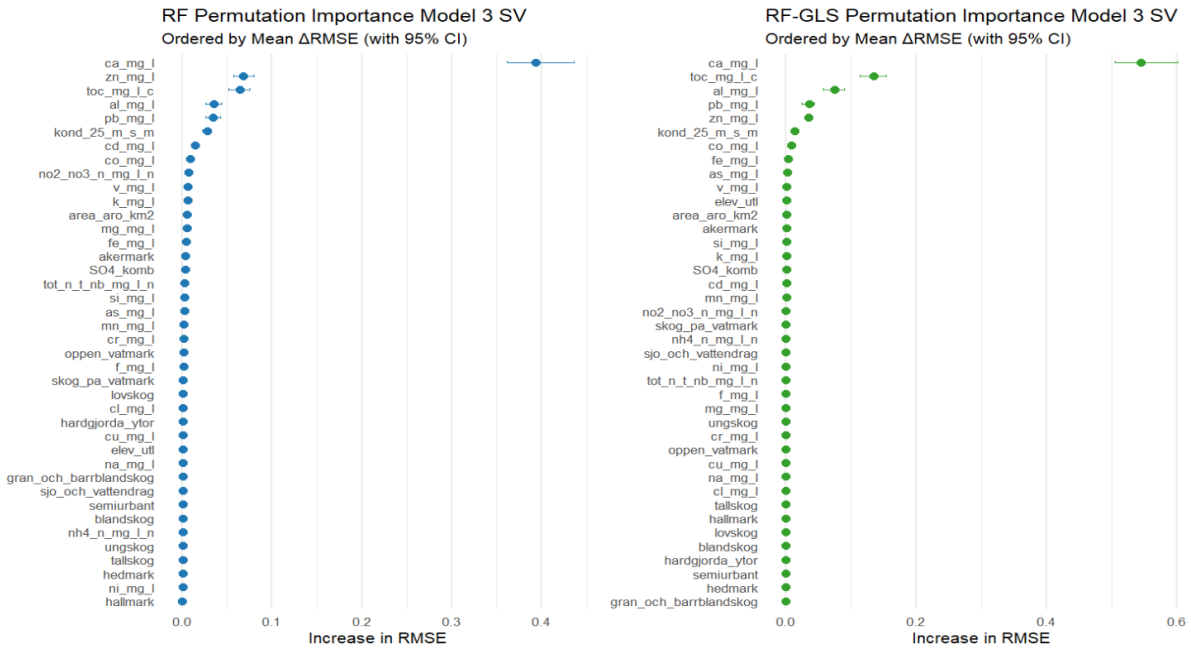

Figure 21: Variable importace for model 3 in region Södra Västergötlan.

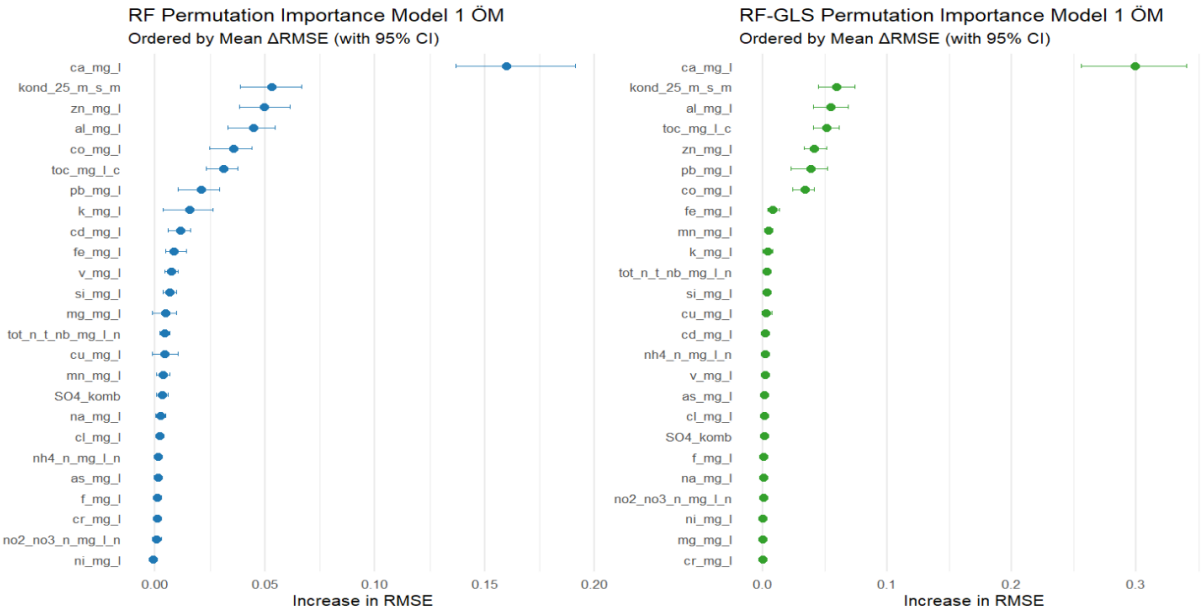

Figure 22: Variable importace for model 1 in region Östra Mellansverige.

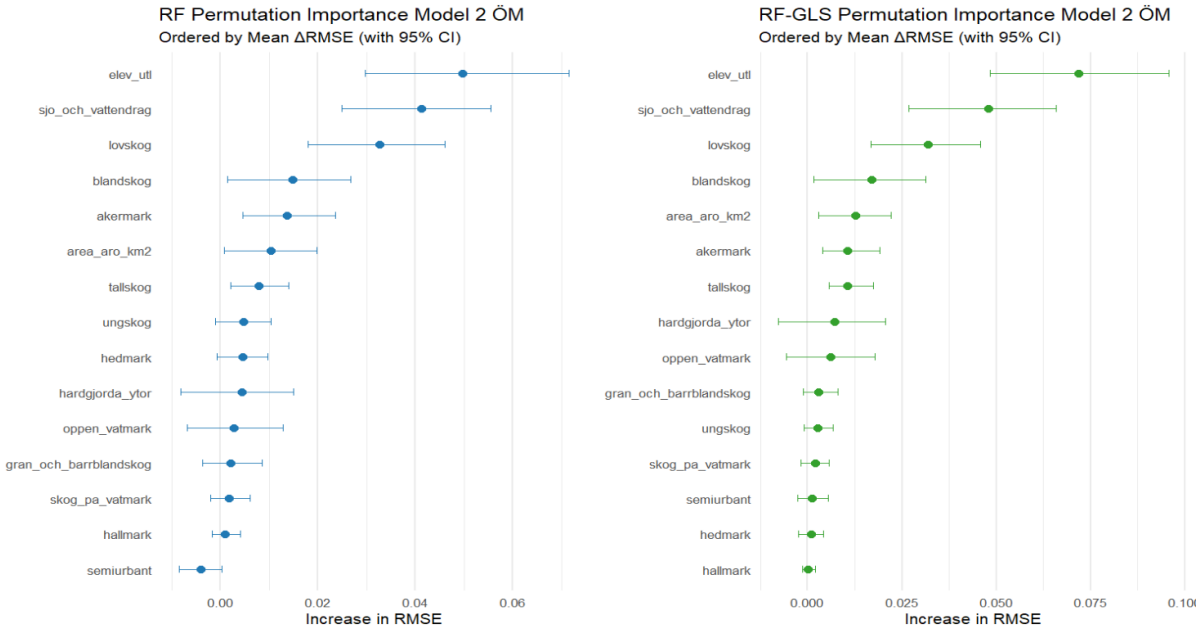

Figure 23: Variable importace for model 2 in region Östra Mellansverige.

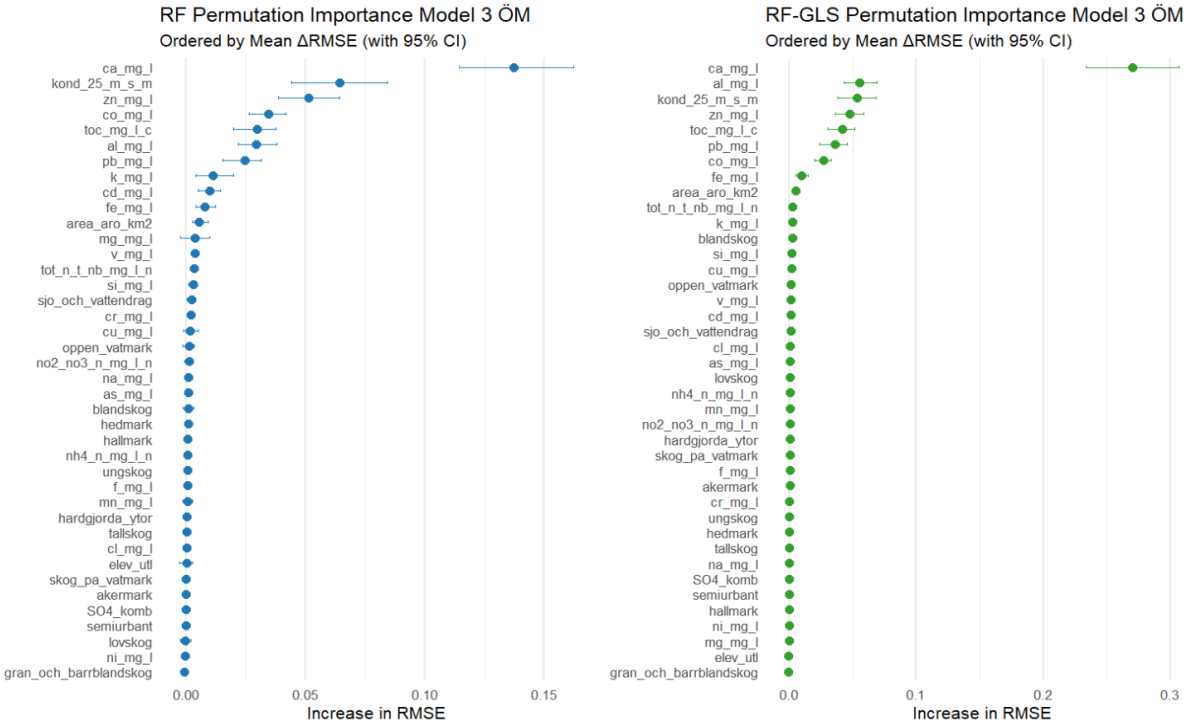

Figure 24: Variable importace for model 3 in region Östra Mellansverige.