Received: 20 March 2023 | Revised: 9 November 2023 | Accepted: 18 January 2024

DOI: 10.1002/env.2842

RESEARCH ARTICLE

WILEY

Sampling design methods for making improved lake management decisions

Vilja Koski[1] | **Jo Eidsvik**[2]

[1] Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

[2] Department of Mathematical Sciences, NTNU, Trondheim, Norway

Correspondence
Vilja Koski, Department of Mathematics and Statistics, University of Jyväskylä, P.O Box 35, Jyväskylä 40014, Finland.
Email: vilja.a.koski@jyu.fi

Funding information
Koneen Säätiö; Norges Forskningsråd, Grant/Award Numbers: 305445, 309960; Emil Aaltosen Säätiö

Abstract

The ecological status of lakes is important for understanding an ecosystem's biodiversity as well as for service water quality and policies related to land use and agricultural run-off. If the status is weak, then decisions about management alternatives need to be made. We assess the value of information of lake monitoring in Finland, where lakes are abundant. With reasonable ecological values and restoration costs, the value of information analysis can be compared with the survey's costs. Data are worth gathering if the expected value from the data exceeds the costs. From existing data, we specify a hierarchical Bayesian spatial logistic regression model for the ecological status of lakes. We then rely on functional approximations and Laplace approximations to get closed-form expressions for the value of information of a sampling design. The case study contains thousands of lakes. The combinatorially difficult design problem is to wisely pick the right subset of lakes for data gathering. To solve this optimization problem, we study the performance of various heuristics: greedy forward algorithms, exchange algorithms and Bayesian optimization approaches. The value of information increases quickly when adding lakes to a small design but then flattens out. Good designs are usually composed of lakes that are difficult to manage, while also balancing a variety of covariates and geographic coverage. The designs achieved by forward selection are reasonably good, but we can outperform them with the more nuanced search algorithms. Statistical designs clearly outperform other designs selected according to simpler criteria.

KEYWORDS

data collection, decision-making, environmental monitoring, optimal design, value of information

1 | INTRODUCTION

We consider a survey design problem connected to environmental monitoring. The inspiration for this study comes from the real-life challenge of lake monitoring in Finland, where lakes are abundant. Inland waters and freshwater biodiversity constitute a valuable natural resource in economic, cultural, aesthetic, scientific and educational terms and need to be protected (Dudgeon et al., 2006). As a result of the Water Framework Directive (WFD) of the European Union (European Parliament, 2000), Finland has implemented a water monitoring program for improving and securing the quality of its inland waters. In the current program, lakes are classified into five ecological status classes (high, good, moderate, poor and bad) according to several variables representing biotic structure, supported by the physical and chemical properties of water, and hydrological as well as morphological features. Existing data on these variables are used to determine the reference conditions for each status class. In addition, according to the directive, some management alternatives must be implemented to improve the ecological status if the status of the water system is classified as moderate or lower. Though biologically principled, the current monitoring program has been considered to be very expensive, and the question is if the efforts are worth it. How should decision-makers wisely allocate monitoring resources at a subset of lakes to significantly aid the decisions about the management alternatives?

A critical question is then to find the optimal sampling design under some information criterion. Regarding the ecological status of Finnish lakes, there are relatively clear management alternatives and rather specific monetary values associated with the various alternatives. Hence, it makes sense to phrase the design criterion according to the notion of decision theory (Abbas & Howard, 2015). In particular, we value information that can improve management decisions via the expected posterior value (PoV) as compared with the prior value (PV) using only the currently available knowledge (Eidsvik et al., 2015). An integral part of this criterion is reduced uncertainty in the statistical model for lake status because it enters into the expected values used in the decision rule. The goal is to find the sampling design which gives the largest value of information (VOI) compared with the cost of data acquisition and processing.

In this paper, the VOI is calculated assuming a Bayesian spatial logistic regression model for the ecological status data. Statistical model parameters are specified from existing data gathered in Finnish lakes. Our large-scale VOI calculations rely on closed-form approximations for hierarchical general linear models (Evangelou & Eidsvik, 2017), which enable fast evaluation of the VOI for each design.

Generally, the problem of selecting an optimal design under some criterion is a central research question in the planning of survey data. However, there are several thousand lakes in Finland, and to find a truly optimal design one would have to evaluate all the available designs. This becomes a combinatorial challenge which is infeasible for our case. One can only evaluate a subset of the designs and we need heuristic algorithms to search for promising subsets. A straightforward heuristic which is easy to implement is the greedy method. It is well-known by mathematicians and computer scientists, and in statistics it is often referred to as a sequential search method (Dykstra, 1971). More nuanced heuristics can naturally build on the result obtained by this approach.

Fortunately, due to the traditional role of statistics in environmental planning, there already exists a significant amount of literature on effective data designs. Other design studies include Jauslin et al. (2022), who consider sequential balanced designs with inclusion probabilities and illustrate this on a data set of species of amphibians; Prentius and Grafström (2022), who compare efficiencies of two-phase methods for adaptive cluster sampling in environmental settings; Foss et al. (2022), who construct dynamic monitoring designs for characterizing the concentration of mine tailings using a spatio-temporal model; and Thilan et al. (2023), who propose adaptive spatio-temporal designs for evaluating trends in coral cover. Nguyen et al. (2018) provide a review of adaptive sampling designs in environmental monitoring. Recent studies concerning the evaluation of information in ecology include the VOI tutorial by Canessa et al. (2015) and its applications in species management, and Williams and Brown (2020), who use scenarios to split settings of pre-selected designs and alternatives that adapt to information. Reich et al. (2018) suggest minimizing the expected misclassification rate of occupancy maps in an ecological application with citizen science data. Our study is different in how we approach the spatial decision situations and in the methods connecting this to a logistic regression model with spatially correlated latent variables.

Several statistical researchers have focused on common optimality criteria of experimental spatial designs, such as D- and A-optimality. Woods et al. (2017) present several approaches for Bayesian design of experiments in logistic regression models with non-spatial applications. Hays et al. (2021) propose a method that links linear integer programming to optimality measures of covariance matrices resulting from mixed models, and as in this work, results are presented on data from freshwater sites. Integer programming has been a popular method to solve the subset selection problem (see, e.g., Arthur et al., 1997). More similar to our work, Paglia et al. (2022) study the VOI computation tasks and propose a Bayesian optimization technique to find approximately optimal spatial designs. We test this method for our case which is of much larger size and involves a different model concerning the hierarchical logistic regression model.

The article is organized as follows: Section 2 provides the background for the case on lake monitoring and the associated sampling design problems, along with a suggested workflow. Section 3 presents the decision situation and the Bayesian spatial logistic regression model for lake status variability, as well as the computational approaches for conducting VOI analysis and heuristic search algorithms used to find good designs. Section 4 explores the results of implementing the algorithms on the lake example from Finland along with sensitivity analysis. Section 5 contains interpretations of the results. Section 6 concludes and presents future work.

2 | BACKGROUND

2.1 | Monitoring the ecological status of lakes

The aim of the WFD is to prevent the deterioration of the ecological status of water systems, with the aim of having at least good ecological and chemical status class. In order to put the legislation into practice in Finland (Figure 1), River Basin Management Planning (RBMP) is implemented in six-years cycles (Aroviita et al., 2019). In brief, the essential parts of one RBMP period are (Higgins et al., 2021; Stankey et al., 2005)

1. the monitoring of the water systems,
2. the assessment and classification of the water systems into status classes,
3. the planning of management alternatives based on the classification, and
4. the implementation of the alternatives.

In the first step, monitoring includes data acquisition of several parameters indicative of the water quality. Biological factors such as phytoplankton, chlorophyll-*a* content in algae, benthic fauna and aquatic plants are monitored at observation sites every 1 to 6 years, depending on the factors. Physical and chemical parameters, such as temperature, phosphorus, nitrogen and oxygen content are gathered from water samples at the lakes at regular intervals, either annually or every few years. Within a year, samples are usually taken about 2 to 12 times. Here, we are mainly interested in the biological quality of chlorophyll-*a* samples, which are collected from the observation sites in summertime (approximately from May to August). Chlorophyll-*a* content is indicative of water body productivity and therefore generally correlates well with the ecological status of lentic water bodies suffering from human-induced eutrophication.

In the second step, one defines the status class of each lake. The conditions are assessed on the basis of the intensity of the ecological changes caused by human activity (Nõges et al., 2009). Thus, the classification is based on several indicators

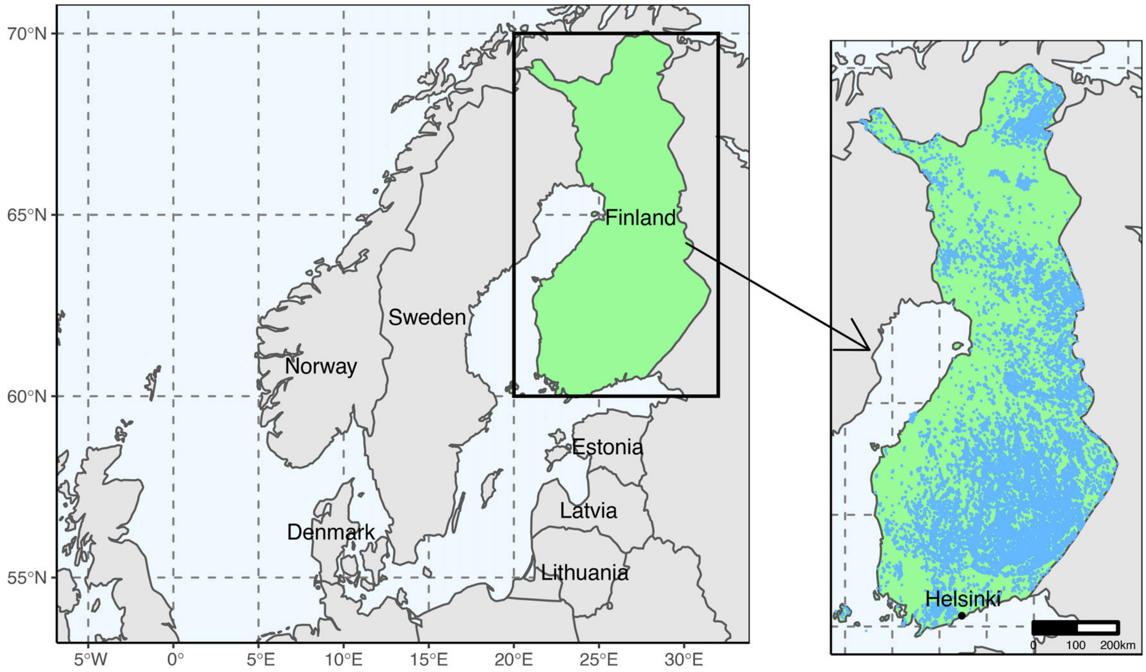

FIGURE 1 Lakes are abundant in Finland. The map shows the lakes which are defined as water bodies in Finland according to the Water Framework Directive (European Parliament, 2000). For the monitoring of ecological status, a subset of lakes must be chosen for sampling.

mentioned above. The classification provides information on the water systems that need measures to achieve or maintain good status.

In the third step, the decisions about the restoration and management alternatives are made based on the classification. These alternatives vary depending on the problems a lake might have. For instance, if the problem is eutrophication, then the management alternative should start with preventing nutrient discharge to the water system. As this is not always possible, the next steps may be dredging to decrease the amounts of aquatic plants and fishing (Søndergaard et al., 2007). Other attempts of lake restoration include raising the water level of the lake or biomanipulation (Jeppesen et al., 2017).

The fourth step is implementing the management alternatives. After the restoration, the effect of the alternatives must again be evaluated via monitoring, and this produces the new status assessment, which returns to the first step.

The basic unit in water management is a water body. It is a separate and significant part of surface water, such as a lake, a creek or a river. In this study, we are only interested in the status of lakes. A lake may form a single water body or it may be divided into several water bodies if it is justified from an ecological point of view. Each lake has at least one sampling site, but the largest lakes might have several sites since they have various habitats and thus several water bodies. Currently, not all smaller lakes (area less than 1 km^2) are defined as water bodies, and they are hence not included. However, smaller water bodies may also be included in the classification at a later stage if they are considered to be significant.

The classification of waters has been conducted three times in Finland. In this study, we use the third ecological status classification, and it is based on the monitoring data gathered during the third RBMP period from 2012 to 2017. The classification is available via open source data maintained by the Finnish Environment Institute (http://www.syke.fi/en-US/Open_information). Since the demand of management alternatives is our main interest, we have narrowed our inspection to the binary ecological classification of lakes, based on whether a lake needs management alternatives (bad, poor or moderate) or not (good or high).

The aim is to predict the ecological status of lakes, and then to use these predictions to make decisions about management alternatives. For the purposes of prediction, we use publicly available information on Finnish lakes. The basic features of 58,707 lakes in Finland can be found from the database maintained by the Finnish Environment Institute (https://www.syke.fi/en-US/Open_information/Open_web_services/Environmental_data_API). Each lake has characteristics such as location (the municipality, drainage basin, center latitude and longitude coordinates, altitude), waterbed area (hectares), length of shoreline (kilometers), average and maximum depth (meters) and volume of water mass (1000 cubic meters). There are also covariates for the agricultural area of municipalities where each lake is located (Official Statistics of Finland, 2020b) and the number of free-time residences in the municipality where each lake is located (Official Statistics of Finland, 2020a). To remove the effect of the municipality area, we divided the agricultural area of municipalities by the area of the municipality to obtain the percentage of agricultural area in each municipality (Figure 2, left).

Status classification is already available for 4360 of the 58,707 lakes. For the remaining lakes, we can predict the status class using the model trained on data from lakes with both status and covariates. Using a logistic regression model, we get the probabilities of ecological status displayed in Figure 2 (right). Here, the most important covariate appears to be agricultural land (left display). We point out that lakes that have a very high probability (near 1) of being in the ecological target status need no remediation. Further, lakes that have a very low probability (near 0) of being in the target status, clearly need to be addressed. Lakes in these two groups are hence not important or worthwhile to monitor because one already knows what to decide. However, there are plenty of lakes for which it is very difficult to make a management decision, and for these it can be very valuable to get information about the status class. But this kind of information comes with a cost, and the dilemma is which lakes should be sampled to make informed decisions on management alternatives for all the lakes.

2.2 | Workflow

In this section, we present the steps that sum up the process of sampling design selection. After framing the decision situation as described in Section 3.1, the following steps are performed:

1. Model fitting from existing data: Construct a statistical model for ecological status based on the 4360 lakes having both status classification and covariates (see Section 4.1).
2. Limit the scope to relevant lakes: Identify lakes with large uncertainty about ecological status classification that could be important to sample (see Section 4.1).

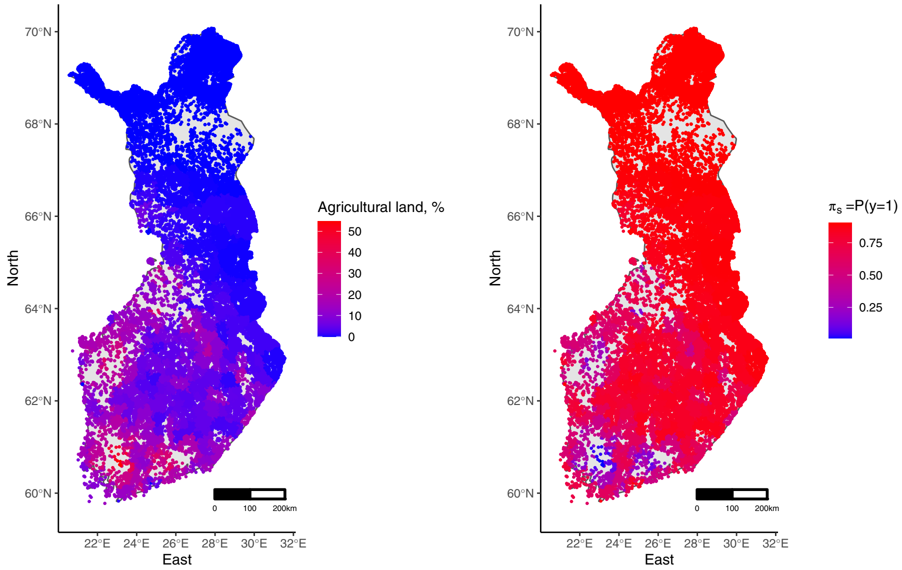

FIGURE 2 There are 54,347 lakes with all covariates available but with missing ecological status. Left: The lakes are color-coded according to the amount of agricultural land (in percentage) in the municipality where the lake is located. Right: The probability of a lake being in the target ecological status ($y_s = 1$), that is, in high or good status, based on a logistic regression model.

3. Sequential selection: Use a greedy forward selection algorithm to find sample designs with large VOI (see Algorithm 1 in Section 3.3.2).
4. Heuristic design search: Conduct nuanced exchange algorithms or Bayesian optimization to search for designs with larger VOI (see Section 3.3.2).

High-quality designs are characterized by large VOI. The results are compared with the cost of gathering data according to the respective sampling designs. In the search algorithms the goal is to optimize the VOI, but one could of course also be motivated by other criteria when selecting designs. We compare and discuss the value of other designs based on various criteria in Section 5.

3 | STATISTICAL FRAMEWORK

3.1 | Framing the decision and sampling problems

The notation connected to the sampling design problem is presented in Table 1. One can choose to leave a lake s untreated ($a_s = 0$) or act to bring the lake to a satisfying condition ($a_s = 1$). We adopt the monetary units associated with these alternatives from Koski et al. (2020). The value of a lake in good condition is set to $R =$ EUR 1000 per hectare, while a lake in poor condition is valued at EUR 0 per hectare. Under the management alternative, it costs EUR 200 per hectare to bring the lake to a sufficiently good condition. No matter the final condition of the lake, the resulting value is $C =$ EUR 1000 − EUR 200 = EUR 800 per hectare. Because of the uncertainty in determining the ecological condition of a lake, it is difficult for managers to make decisions about lake management. For a risk neutral decision maker, the PV is the maximum expected value over the two management alternatives. For a particular lake s with area A_s hectare, this can be written as

$$\mathrm{PV}_s = \max\{R_s \cdot \mathbb{E}(\pi_s),\ C_s\} = C_s + \max\{R_s \cdot \mathbb{E}(\pi_s) - C_s,\ 0\},$$

where for alternative $a_s = 1$ the monetary amount is $C_s = CA_s$, while for alternative $a_s = 0$ there is revenue $R_s = RA_s$ and $\mathbb{E}(\pi_s)$ denotes the expected condition of lake $s = 1, \dots, N$. We will later model this via a latent logistic regression model for variable x_s, where $\pi_s = \frac{e^{x_s}}{1+e^{x_s}}$.

TABLE 1 Summary of the notation used in the article.

Notation	Definition
$s \in \{1, \dots, N\}$	Index for lakes
$\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_N\}$	Locations of all lakes
$D \in \mathcal{D}$	Design in all possible design sets
π_s	Probability for lake s not needing management alternatives
x_s	Latent random variable at lake s
$\boldsymbol{f}_s$	Covariates at lake s
$\boldsymbol{x}_D$	Latent length-$\lvert D\rvert$ random vector of design D
$\boldsymbol{y}_D$	Prospective data vector of length-$\lvert D\rvert$, gathered in design D
$\boldsymbol{F}_D$	Covariate matrix of design D
$a_s \in \{0, 1\}$	Management alternative to choose for lake s
A_s	Area of lake s
R_s	Revenue of alternative for lake s
C_s	Cost of alternative for lake s
PV	Prior value
$\text{PoV}(D)$	Posterior value of design D
$\text{VOI}(D)$	Value of information of design D
$\text{VOI}_s(D)$	Lake s effect value of information of design D
$P(D)$	Price or cost of data of design D

We assume that managers are free to select the best alternative for every lake (Eidsvik et al., 2015). This means that the total PV decouples to a sum over all lakes and we have

$$\text{PV} = \sum_{s=1}^{N} \text{PV}_s = \sum_{s=1}^{N} \left[C_s + \max\{R_s \cdot \mathbb{E}(\pi_s) - C_s,\ 0\} \right]. \tag{1}$$

Additional data can assist the decision-makers in choosing among the difficult lake management alternatives. In particular, the VOI is positive when various data outcomes lead to different alternatives being chosen, because this gives added value from that of the PV. Still, this gain must be compared with the cost of collecting and processing the data. Moreover, there are several possibilities for the design of gathering spatial data used in determining the ecological status of lakes.

Assume that one wants to select a subset of lakes to observe their ecological status. We denote such a subset by design D of size $|D|$. Collecting data for all lakes would be too expensive, and one can only afford to measure a subset. We denote the (latitude, longitude) positions of the N lakes of interest by $\boldsymbol{u}_1, \dots, \boldsymbol{u}_N$. Possible spatial survey designs contain no sites, single sites, couples, triplets, and so on, up to the design where all N sites are included, and the entire set of designs is denoted $\mathcal{D} = \bigcup_{i=0}^{N} \mathcal{D}_i$, where

$$\begin{aligned}
\mathcal{D}_0 &= \emptyset, \\
\mathcal{D}_1 &= \{(\boldsymbol{u}_1), (\boldsymbol{u}_2), \dots, (\boldsymbol{u}_N)\}, \\
\mathcal{D}_2 &= \{(\boldsymbol{u}_1, \boldsymbol{u}_2), (\boldsymbol{u}_1, \boldsymbol{u}_3), \dots, (\boldsymbol{u}_{N-1}, \boldsymbol{u}_N)\}, \\
&\ \vdots \\
\mathcal{D}_N &= \{(\boldsymbol{u}_1, \boldsymbol{u}_2, \dots, \boldsymbol{u}_N)\}.
\end{aligned} \tag{2}$$

Finding the optimal design is extremely difficult because there are 2^N possible designs. One often resorts to heuristics approaches to find several useful designs that provide a basis for decision support about information gathering.

Denote the prospective data to be measured in a design D by $\boldsymbol{y}_D = (y_{D,1}, \dots, y_{D,|D|})$. The associated covariates are denoted by matrix $\boldsymbol{F}_D$, which has one row for each design location. For this decision situation, the PoV of data $\boldsymbol{y}_D$ is defined by

$$\text{PoV}(D) = \sum_{\boldsymbol{y}_D} \sum_{s=1}^{N} \left[C_s + \max\{R_s \cdot \mathbb{E}(\pi_s | \boldsymbol{y}_D) - C_s,\ 0\} \right] p(\boldsymbol{y}_D), \tag{3}$$

where $\mathbb{E}(\pi_s|\boldsymbol{y}_D)$ is the conditional expected lake status, given observations $\boldsymbol{y}_D$ distributed according to the probability mass function $p(\boldsymbol{y}_D)$. Here, the sums over the data outcomes and lakes can be interchanged, so that $\text{PoV}(D) = \sum_{s=1}^{N} \text{PoV}_s(D)$, where the entries in the sum are the expected value contribution from the data $\boldsymbol{y}_D$ to the decision at lake s.

Under the assumptions of a risk neutral decision-maker (Eidsvik et al., 2015), the VOI equals the difference in PoV and PV, so that

$$\text{VOI}(D) = \text{PoV}(D) - \text{PV}. \tag{4}$$

Note that the fixed C_s part is the same for both Equations (1) and (3). Further, the decoupling over lake decisions means that the VOI is the additive contributions from the VOI at each lake s, that is,

$$\begin{aligned} \text{VOI}_s(D) &= \sum_{\boldsymbol{y}_D} \max\{R_s \cdot \mathbb{E}(\pi_s|\boldsymbol{y}_D) - C_s,\ 0\} p(\boldsymbol{y}_D) - \max\{R_s \cdot \mathbb{E}(\pi_s) - C_s,\ 0\}, \\ \text{VOI}(D) &= \sum_{s=1}^{N} \text{VOI}_s(D). \end{aligned} \tag{5}$$

The goal is to choose a sampling design D that is expected to provide data that substantially affect the decisions made, especially at those lakes where it is difficult to choose between the management alternatives. It is common to choose the design with the largest VOI(D) compared with the data gathering cost $P(D)$, as managers are interested in making the best out of their data acquisition and processing expenses. Alternatively, one can also have a budget for the data gathering, and the goal is to find the largest VOI among all designs D having costs not exceeding this budget. Overall, the VOI results for various designs will support difficult decisions related to data gathering.

3.2 | Binary regression

3.2.1 | Logistic model

Assume that a binary response $y_s \in \{0, 1\}$ at lake $s = 1, \dots, N$ is distributed as

$$\begin{aligned} P(y_s = 1|x_s) = \pi_s, \quad & P(y_s = 0|x_s) = 1 - \pi_s, \\ \text{logit}(\pi_s) = x_s = \boldsymbol{f}_s' \boldsymbol{\beta} + w_s, \quad & \pi_s = \frac{e^{x_s}}{1 + e^{x_s}}, \end{aligned} \tag{6}$$

where the linear predictor x_s at lake s includes covariates $\boldsymbol{f}_s = (f_1(s), \dots, f_J(s))'$ in combination with regression parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$. It further has a lake-specific effect w_s that is spatially correlated. For short, we denote effects $\boldsymbol{w} = (w_1, \dots, w_N)'$.

Assuming conditional independence in Equation (6), the log-likelihood of data $\boldsymbol{y}_D = (y_{D,1}, \dots, y_{D,|D|})$ is obtained by

$$\sum_{\boldsymbol{u}_s \in D} \log(p(y_s|\boldsymbol{\beta}, \boldsymbol{w})) = \sum_{\boldsymbol{u}_s \in D} y_s(\boldsymbol{f}_s' \boldsymbol{\beta} + w_s) - \log(1 + \exp(\boldsymbol{f}_s' \boldsymbol{\beta} + w_s)). \tag{7}$$

3.2.2 | Bayesian latent spatial logistic model

The regression parameter $\boldsymbol{\beta}$ is unknown and has a prior probability density function (pdf) $p(\boldsymbol{\beta})$. This pdf is here assumed to be Gaussian with mean vector $\boldsymbol{\mu}_\beta^0$ and covariance matrix $\boldsymbol{\Sigma}_\beta^0$. The spatial effects $\boldsymbol{w}$ are represented by a zero mean

Gaussian process model. We specify a fixed variance $\mathrm{Var}(w_s) = \sigma^2$ and impose a Matern correlation function such that $\mathrm{Corr}(w_s, w_t) = (1 + \phi h_{st})\exp(-\phi h_{st})$, where h_{st} is the great-circle distance between two lakes centered at locations $\boldsymbol{u}_s$ and $\boldsymbol{u}_t$. Given the currently available lake data, we specify model parameters $\boldsymbol{\mu}^0_\beta$, $\boldsymbol{\Sigma}^0_\beta$, σ and ϕ from an approximate marginal likelihood expression.

Keeping the model parameters fixed in the following, $(\boldsymbol{\beta}', \boldsymbol{w}')'$ are Gaussian distributed. In particular, the linear predictor $x_s = \boldsymbol{f}'_s\boldsymbol{\beta} + w_s$ has mean $\mu_s = \boldsymbol{f}'_s\boldsymbol{\mu}^0_\beta$ and variance $\sigma_s^2 = \boldsymbol{f}'_s\boldsymbol{\Sigma}^0_\beta\boldsymbol{f}_s + \sigma^2$. Let $\boldsymbol{x}_D = \{x_s; \boldsymbol{u}_s \in D\}$ denote the vector of linear predictor variables at the design locations defined via set D. We similarly define the vector $\boldsymbol{w}_D$ of latent effects, $\boldsymbol{\mu}_D$ for the prior mean vector and $\boldsymbol{F}_D$ for the size $|D| \times J$ matrix of covariates at these design locations. Building on properties of Gaussian processes, the joint distribution of $(x_s, \boldsymbol{x}'_D)' = (\boldsymbol{f}'_s\boldsymbol{\beta}, \boldsymbol{F}_D\boldsymbol{\beta})' + (w_s, \boldsymbol{w}'_D)'$ is Gaussian with mean $(\mu_s, \boldsymbol{\mu}'_D)'$ and covariance matrix

$$\mathrm{Var}[(x_s, \boldsymbol{x}'_D)'] = \begin{bmatrix} \sigma_s^2 & \boldsymbol{\Sigma}_{s,D} \\ \boldsymbol{\Sigma}_{D,s} & \boldsymbol{\Sigma}_D \end{bmatrix}, \tag{8}$$

with $\boldsymbol{\Sigma}_{D,s}$ being a length $|D|$ vector holding all the cross-covariance terms between variable x_s and the linear predictor variables in the design D, that is, $\boldsymbol{\Sigma}_{s,D} = \boldsymbol{f}_s\boldsymbol{\Sigma}^0_\beta\boldsymbol{f}'_D + \sigma^2\mathrm{Corr}(w_s, \boldsymbol{w}_D)$, while $\boldsymbol{\Sigma}_D$ is a $|D| \times |D|$ matrix with variance-covariance terms within all the design location variables.

By standard Gaussian expressions, the conditional distribution of x_s given $\boldsymbol{x}_D$ is then Gaussian with mean and variance

$$m_s = \mu_s + \boldsymbol{\Sigma}_{s,D}\boldsymbol{\Sigma}_D^{-1}(\boldsymbol{x}_D - \boldsymbol{\mu}_D), \quad \xi_s^2 = \sigma_s^2 - \boldsymbol{\Sigma}_{s,D}\boldsymbol{\Sigma}_D^{-1}\boldsymbol{\Sigma}_{D,s}. \tag{9}$$

In our setting the data are binary, and there is no closed form like Equation (9) for the conditional mean and variance. One can however derive approximate expressions for the expected variance reduction from binomial data. In the VOI approximation below we rely on the following expression from Evangelou and Eidsvik (2017) for the variance reduction associated with binomial measurements

$$\begin{aligned} \chi_s^2 &= \boldsymbol{\Sigma}_{s,D}[\boldsymbol{\Sigma}_D + \boldsymbol{K}_D]^{-1}\boldsymbol{\Sigma}_{D,s}, \\ \boldsymbol{K}_D &= \mathrm{diag}\left\{ 2 + \exp\left(-\mu_s + \frac{\sigma_s^2}{2}\right) + \exp\left(\mu_s + \frac{\sigma_s^2}{2}\right); \boldsymbol{u}_s \in D \right\}. \end{aligned} \tag{10}$$

Comparing with the variance reduction in Equation (9), we notice an additional $\boldsymbol{K}_D$ for the center matrix that is inverted to get χ_s^2 in Equation (10). This means that the variance reduction is smaller than when observing the linear predictors directly. Moreover, the magnitudes of this matrix $\boldsymbol{K}_D$ depend on the mean μ_s and variance σ_s^2 at the design locations.

3.3 | The value of information for spatial binary data

By using the logistic model formulation, the VOI contribution at lake s in Equation (5) equals

$$\mathrm{VOI}_s(D) = \sum_{\boldsymbol{y}_D \in \{0,1\}^{|D|}} \max\left\{ R_s \cdot \mathbb{E}\left(\frac{e^{x_s}}{1 + e^{x_s}} | \boldsymbol{y}_D\right) - C_s,\ 0\right\} p(\boldsymbol{y}_D) - \max\left\{ R_s \cdot \mathbb{E}\left(\frac{e^{x_s}}{1 + e^{x_s}}\right) - C_s,\ 0\right\}. \tag{11}$$

There is no closed-form expression for Equation (11), and we next outline an approximate solution building on the results in Section 3.2.2.

3.3.1 | Approximating the VOI

We rely on an analytical approximation of the VOI developed by Evangelou and Eidsvik (2017). The VOI is computed using the Laplace approximation based on Gaussian approximations in Equations (9) and (10), in combination with normal cumulative distribution function (cdf) fitting of the logistic function. We discuss these in some more detail next.

First, the conditional expectation of $e^{x_s}/(1+e^{x_s})$ in Equation (11) is approximated by linearizing the logistic likelihood and quadratic fitting of the curvature giving Equation (10). In doing so, the integral depends on the unknown conditional mode (approximate Gaussian distributed with variance in Equation (10)) rather than the discrete data. Next, we build upon the idea of approximating the logistic function $g(x_s)=e^{x_s}/(1+e^{x_s})$ by the normal cdf $\Phi(\alpha x_s)$ for an appropriately selected scaling parameter α. Depending on the criterion one uses to minimize the mismatch between the two functions, one gets a different α. We choose $\alpha = 0.59$, which is one of the scaling parameters mentioned in Demidenko (2013). The two functions are then very close in a large span of x_s values. Finally, we compute the complete and incomplete logistic-normal integrals by

$$\begin{aligned}
\Lambda(\mu,\sigma^2) &= \int_{-\infty}^{\infty}\frac{e^x}{1+e^x}\varphi(x;\mu,\sigma^2)dx \approx \int_{-\infty}^{\infty}\Phi(\alpha x)\varphi(x;\mu,\sigma^2)dx = \Phi\left(\frac{\alpha\mu}{\sqrt{1+\alpha^2\sigma^2}}\right)\\
\Lambda_a(\mu,\sigma^2) &= \int_{a}^{\infty}\frac{e^x}{1+e^x}\varphi(x;\mu,\sigma^2)dx \approx \int_{a}^{\infty}\Phi(\alpha x)\varphi(x;\mu,\sigma^2)dx\\
&= \Phi\left(\frac{\mu-a}{\sigma}\right) - \Phi_2\left(\frac{\mu-a}{\sigma}, -\frac{\alpha\mu}{\sqrt{1+\alpha^2\sigma^2}}; \frac{\alpha\sigma}{\sqrt{1+\alpha^2\sigma^2}}\right), \qquad (12)
\end{aligned}$$

where $\varphi(x;\mu,\sigma^2)$ denotes the normal probability density function evaluated at x with mean μ and variance σ^2, and $\Phi_2(z_1,z_2;r)$ is the bivariate standard normal cdf with correlation r, evaluated at (z_1,z_2).

The VOI_s in Equation (11) is then approximated by

$$\begin{aligned}
\mathrm{VOI}_s(D) \approx{}& R_s\Lambda_a\left(\frac{\mu_s}{\sqrt{1+\alpha^2\xi_s^2}}, \frac{\chi_s^2}{1+\alpha^2\xi_s^2}\right) - R_s g(a)\Phi\left(\frac{\mu_s - a\sqrt{1+\alpha^2\xi_s^2}}{\chi_s}\right)\\
&- R_s\max\{\Lambda(\mu_s, \xi_s^2+\chi_s^2) - g(a),\ 0\}, \qquad (13)
\end{aligned}$$

where $a=\log([C_s/R_s]/(1-[C_s/R_s]))$ and $g(a)=1/(1+e^{-a})$. Evangelou and Eidsvik (2017) use extensive Monte Carlo simulations to study the properties of this approximation for binomial data and Poisson distributed data. Similar expressions have been used to approximate the expected Bernoulli variance in logistic models (Anyosa et al., 2023).

3.3.2 | Search algorithms for optimal designs

Our aim is to find designs with large VOI. Ideally, this entails solving an optimization problem as follows:

$$D^{\dagger} = \arg\max_{D}\{\mathrm{VOI}(D)\}, \quad \mathrm{VOI}(D) > P(D), \qquad (14)$$

where $P(D)$ is the cost of gathering the monitoring data from the design D. There may also be interest in maximizing the gap between the information value and the design cost, that is, $\mathrm{VOI}(D) - P(D)$.

In general, the optimal design problem in Equation (14) is NP-hard because of the enormous number of combinations. With no constraints on $|D|$, there are 2^N possible designs. Even if we limit the scope to fixed size designs, there are $N!/[(N-|D|)!|D|!]$ possible designs. With $N = 4748$ and $|D| = 50$ this number of combinations is enormous (about 10^{100}). It is hence infeasible to analyze all available combinations and heuristics are needed.

We next describe a forward selection algorithm aimed to maximize the VOI up to a certain size of designs D. In Algorithm 1, the heuristic approach sequentially adds observation locations $j = 1, 2, \dots$ to the design. This continues until the maximum size is reached. In the extreme event one continues until size N, but in practice it stops for $|D| << N$, when the VOI increase is negligible from j to $j+1$, or when the VOI is clearly too small to justify purchasing all that data. Instead of choosing just one extra lake in the design at each stage, one can choose more sites at a time. If two lakes are equally good in the forward evaluation, the selection between them is performed randomly.

The forward selection algorithm presented here often gives reasonable designs, but it is only a heuristic search, which has no guarantee of returning the optimal design. More complex search methods for efficient sampling designs include variants of the randomized exchange algorithm (see, e.g., Harman et al., 2020). This defines an iterative search among new (random) combinations of designs. In one of its forms, which we use for our data below, each iteration includes an

Algorithm 1. Forward selection of design

```
$j = 1$,
$D = \emptyset$                                                ▷ set of already selected sites
while $j \leq N$ do
    for $i = 1, \dots, N$ and $\boldsymbol{u}_i \notin D$ do
        $D^{(i)} = D \cup \{\boldsymbol{u}_i\}$                ▷ Candidate design
        $\mathrm{VOI}(D^{(i)}) = \sum_{s=1}^{N} \mathrm{VOI}_s(D^{(i)})$    ▷ VOI of candidate design
    end for
    $i^* = \arg\max_i \{\mathrm{VOI}(D^{(i)}); i = 1, \dots, N \text{ and } \boldsymbol{u}_i \notin D\}$    ▷ optimal new design site
    $D = D \cup \{\boldsymbol{u}_{i^*}\}$
    $j = j + 1$,
end while
```

exchange where one lake is removed from the design and another is added to the design. The exchange probability is in our case guided by single-location results: $\mathrm{VOI}(\{\boldsymbol{u}_s\})$, that is, assuming $N = |D| = 1$ for all $s = 1, \dots, N$. Lakes with a large single-lake VOI are hence more likely to be added to the design, while the ones with small single-lake VOI are more likely to be removed from the design. Still, the probabilities are positive for including or excluding any lake to the design, and this ensures some randomness helping the optimization approach from getting stuck in a local optimum. For our dataset we also test the approach of Paglia et al. (2022), who used Bayesian optimization and expected improvement to search for promising designs. The Hausdorff distances between designs D are used to form a covariance matrix in a Gaussian process surrogate model for $\mathrm{VOI}(D)$, taking D as input. One learns this Gaussian process from previous VOI evaluations. At each iteration, a batch of promising designs are selected as the ones having high expected improvement in VOI according to the surrogate model. This is a fast calculation. After this selection, all designs in the batch go to the much more costly VOI evaluation.

4 | RESULTS

The modelling, preliminary steps, and greedy algorithm were implemented in R (R Core Team, 2021). The randomized exchange algorithm and the Bayesian optimization algorithm were implemented with Matlab (MATLAB, 2021).

4.1 | Modelling and preliminary steps

We used a standard logistic regression model to determine the important covariates. Candidate covariates were center latitude and longitude coordinates, waterbed area (1000 square kilometers), length of shoreline (kilometers), and by municipality, the agricultural area, population and number of summer residences scaled with municipality area. Additional explanatory variables that are challenging to measure, such as drainage basin, average depth, maximum depth and volume of water mass, were not included in the analysis due to the high number of missing values.

We fitted models that contained each of the seven covariates one at a time. The ones that seemed important on their own based on $-2\hat{l}$, where $\hat{l}$ is the log-likelihood of the logistic regression model, were then analyzed more closely. The covariates that had a significant effect at this point were the latitude and longitude coordinates and the agricultural area scaled with municipality area. We computed the change in the value of $-2\hat{l}$ when each variable on its own was omitted. Only those that lead to a significant increase in the value of $-2\hat{l}$ were retained in the model. As the result of the selection, we chose the latitude coordinate and the agricultural land in the municipality where the lake was located as covariates. We specified $\boldsymbol{\mu}_{\beta}^{0}$ and $\boldsymbol{\Sigma}_{\beta}^{0}$ as the approximate mean and covariance of $\boldsymbol{\beta}$, given the initial data. Using scaled agricultural land and latitude coordinate as covariates, the model has $\boldsymbol{\mu}_{\beta}^{0} = \hat{\boldsymbol{\beta}} = (1.6, -13.8, 0.02)$, $\mathrm{diag}(\boldsymbol{\Sigma}_{\beta}^{0}) = (3.9, 0.5, 0.001)$ and a substantial correlation of -0.6 between intercept and slope with scaled agricultural type. Goodness of fit measures show that the model fits reasonably well to the existing data (deviance statistics).

We then used the Laplace approximation (see, e.g., Shun & McCullagh, 1995) to estimate the covariance parameters of the spatial random effect. For the variability of the spatially structured variables we get an estimate of $\sigma = 1.11$. For the correlation decay we get $\phi = 0.06$, meaning that the spatial correlation is reduced to 0.05 at a distance of 80 km. We were also interested in seeing whether the selection methods are sensitive to changes in the spatial covariance parameter values. Thus, we varied (σ, ϕ) values. Based on the second derivatives of the marginal log-likelihood function of the parameters, 1 standard deviation up and down (in the log-space) gives $\sigma = 0.99$ as a low value and $\sigma = 1.26$ as a high value for the scale. Similarly, this gives $\phi = 0.047$ as a low value and $\phi = 0.077$ as a high value of the correlation decay. In what follows, we tried five different sets of the spatial covariance parameter values: $(\sigma, \phi) = (1.11, 0.06)$ as benchmark parameter values and in addition, $(\sigma, \phi) = \{(0.99, 0.06), (1.26, 0.06), (1.11, 0.047), (1.11, 0.077)\}$.

Regarding the design, as indicated in the workflow outlined in Section 2.2, we first reduced the set of possible designs. This was done by reducing the set of 54,347 lakes to 4748 lakes. First, from the mean value $\boldsymbol{\mu}_\beta^0$ and the covariates for those lakes, we computed predictive probabilities of a lake being in the target status, as in Equation (6) (see Figure 2). From the revenue R_s and the cost C_s of lake s, we calculated the prior value PV_s, as in Equation (1), for each lake. We selected 1000 lakes which have the first term in the maximum in Equation (1), that is, $R_s \cdot \mathbb{E}(\pi_s) - C_s$, close to zero. We assumed these lakes are among the most interesting in the sense of the VOI evaluations. Second, we calculated the VOI with only single sites in the design, that is, $\mathrm{VOI}_s(\{\boldsymbol{u}_s\})$ (referred to as the self-effect). When calculating self-effect, we assumed that $N = |D| = 1$ for all s. We assumed that if this self-effect is minuscule, then that lake is unlikely to have a significant effect on the total VOI with all N lakes included. There were 3798 lakes that have a self-effect $\mathrm{VOI}_s(\{\boldsymbol{u}_s\}) >$ EUR 138. Partly, these lakes overlap with the first 1000 lakes selected. We combined the first and second selected lakes and limited our optimal design selection into the resulting $N = 4748$ lakes.

4.2 | Selection of data sites

The greedy approach in Algorithm 1 was used to construct a relatively small-size design from the 4748 lakes. In doing so, we conducted the subset selection with greedy forward selection. We ran this approach until $|D| = 300$ lakes were included in the design. The VOI of that forward-selected subsample is shown in Figure 3 (left) with varying spatial covariance parameter values (σ, ϕ). Naturally, the VOI of the design increases as the number of lakes in the design grows.

Our aim was to compare the VOI of a design to the cost of gathering data in that design. We assumed that each selected lake implies one sampling site. Currently, collecting and analyzing one chlorophyll-a sample costs EUR 138

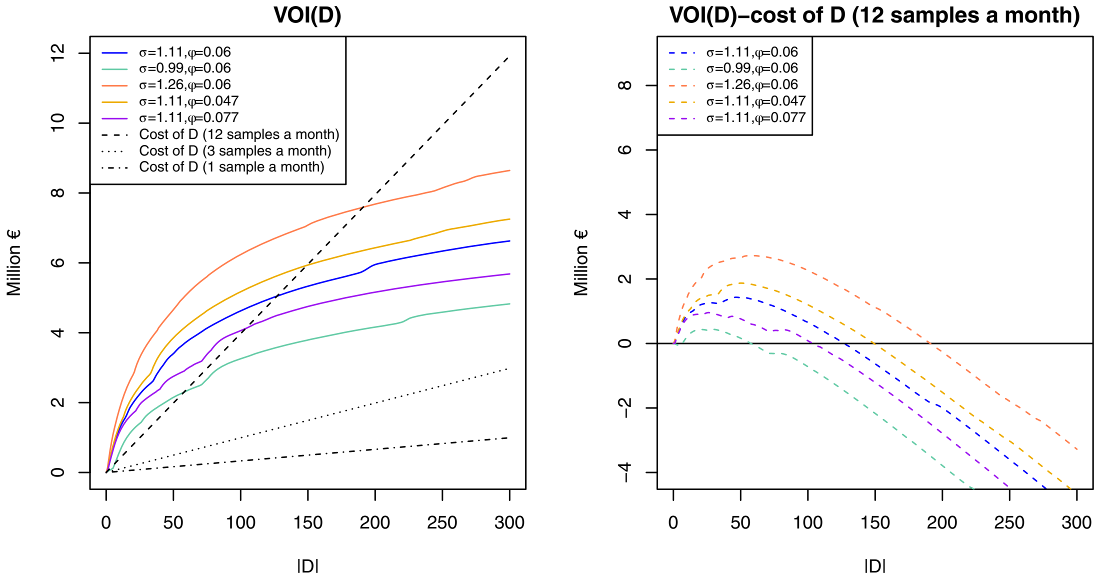

FIGURE 3 Left: The VOI (solid colored curve) and the cost of data gathering (dashed curves) of design of size $|D|$ in million euro plotted with respect to the $|D|$. The color-coded curves show the calculation for different spatial covariance parameter values (σ, ϕ). We assume three alternatives for data gathering: a large amount of data (12 samples a month), an average amount of data (3 samples a month) or a small amount of data (1 sample a month) gathered per site. Right: The difference between the VOI for different spatial covariance parameter values (σ, ϕ) and the cost of gathering a large amount of data in million euro.

(Koski et al., 2020). The total cost of chlorophyll-a data gathering from a design of size $|D|$ was assumed to consist of the samples gathered from four months per year and during six years, as it is the length of the RBMP period. We assumed three data gathering alternatives: either a large amount of data (12 samples a month), an average amount of data (3 samples a month) or a small amount of data (1 sample a month) per site is gathered during that time. Processing of these data gives the ecological classification $y_s = 0$ or $y_s = 1$ for each lake s. Note that we are using the same model for ecological status class y_s, no matter what acquisition and processing is required for the chlorophyll-a samples.

The three dashed curves in Figure 3 (left) illustrate costs of sampling plans. Generally, the studied VOI results of the selected subsamples seem to exceed the cost of gathering that subsample, meaning that the data acquisition is worth doing. When assuming twelve samples in a month, the cost exceed the VOI (calculated with benchmark parameter values (σ, ϕ)) after selecting $|D| = 126$ lakes in the design. Then, VOI(D) = EUR 5.03 million and $P(D)$ = EUR 5.01 million. An even higher VOI value is reached when using $(\sigma, \phi) = (1.26, 0.06)$. In that case, the 12 samples a month cost is reached after selecting $|D| = 192$ lakes, when VOI(D) = EUR 7.60 million and $P(D)$ = EUR 7.63 million.

The gap between the VOI with varying spatial covariance parameter values (σ, ϕ) and the cost of gathering 12 samples a month from the design is shown in Figure 3 (right). This illustrates the excess information value over the cost of the data for different sample sizes. When assuming 12 samples a month and benchmark parameter values (σ, ϕ), the most benefit is achieved when 47 lakes are measured. Then, VOI(D) = EUR 3.26 million and $P(D)$ = EUR 1.83 million, which gives a gap of EUR 1.43 million. When using parameter values $(\sigma, \phi) = (1.26, 0.06)$, the most benefit is achieved when 59 lakes are measured. Then, VOI(D) = EUR 5.03 million and $P(D)$ = EUR 2.31 million, which gives a gap of EUR 2.72 million.

Figure 4 illustrates the selection on the map of Finland when varying the spatial covariate parameter values (σ, ϕ). The circle colors illustrate the order of the 300 selected lakes. The lakes that the algorithm did not include in the design are

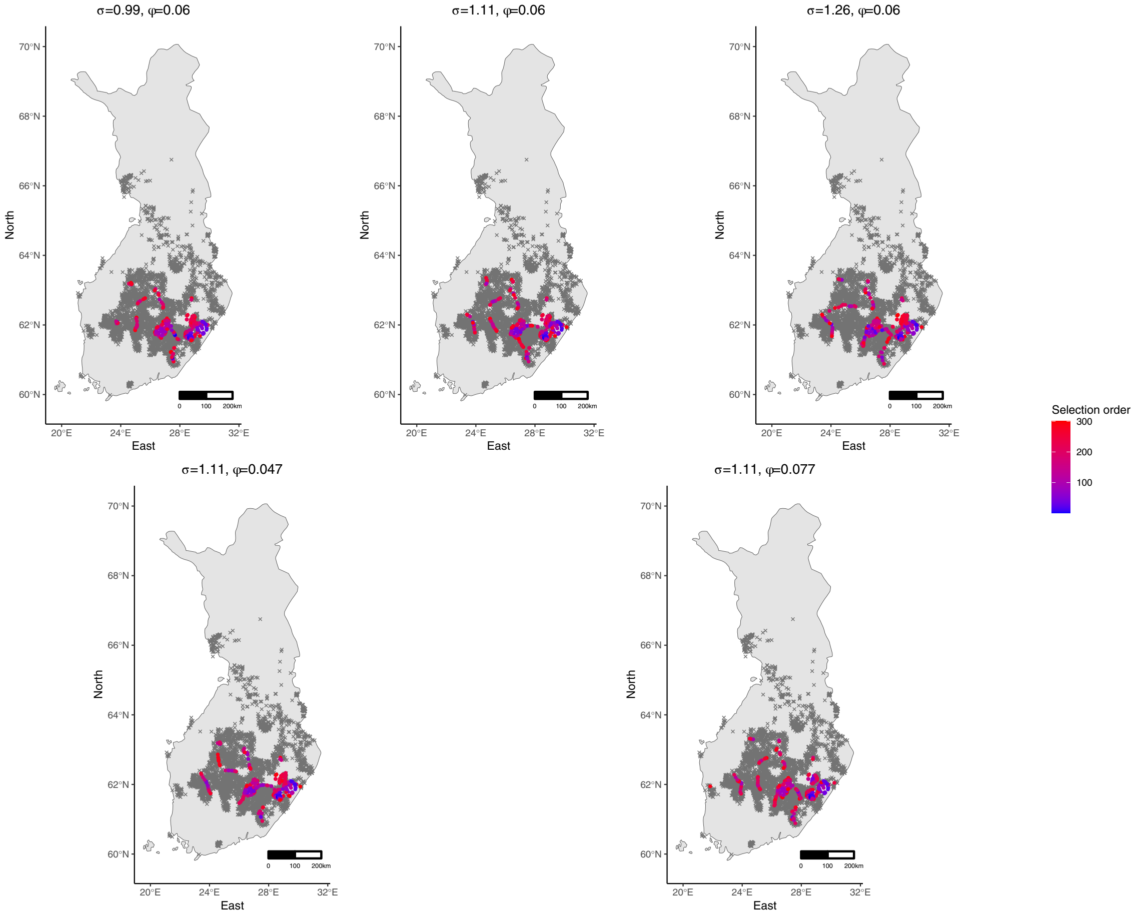

FIGURE 4 Map view indicating the order of 300 selected lakes from the 4748 interesting lakes. The sequential selection is computed using different spatial covariance parameters (σ, ϕ). The red circles are the selected lakes by algorithm in order and the crosses are the lakes excluded in the design.

depicted with crosses. The design selection appears to be similar with very few differences, regardless of the parameter values. All the selected designs are clustered in the southeast region of Finland, which is known to be rich in water areas.

As observed, there are no selected lakes in the northern region. To examine this further, we focused on this region when the 300th lake is selected, as an anecdotal example. When we have selected $|D| = 299$ lakes in the design, there are 4449 potential lakes to be the 300th selected lake. All potential lakes are ranked in descending order based on the VOI when a single lake is added to the current design. From the Northern area, the highest ranked lake is Lake Kattilajärvi (66.16°N, 24.40°E), which is only the 2292nd ranked. Initially, this lake has $\pi_s = \mathrm{P}(y_s = 1) = 0.80$ and a self-effect of $\mathrm{VOI}_s(\{\boldsymbol{u}_s\}) = \mathrm{EUR}\ 59$. If Lake Kattilajärvi is selected, the total VOI with $|D| = 300$ is $\mathrm{VOI}(D) = \mathrm{EUR}\ 6,623,243$. According to the sequential selection algorithm, the 300th selected lake is Lake Vääräjärvi (63.30°N, 26.43°E) and the total VOI after that selection is $\mathrm{VOI}(D) = \mathrm{EUR}\ 6,626,647$.

Figure 5 illustrates what kind of lakes are the most important to measure from the VOI point of view, along with their relationship to the agricultural land (first axis) and the waterbed area (second axis). The color-coded circles show the selection order of the $|D| = 300$ lakes, when using the benchmark parameter values of (σ, ϕ). The crosses are the lakes not included in design. We have denoted the relevant $N = 4748$ lakes with color-coded crosses. The selection order of lakes indicates that the selection covers most lake types but not the ones with very large waterbed and low agricultural land covariates. There is a tendency to choose big lakes early in the design order, but small and average size lakes are also chosen. A closer inspection shows that 48% of the $|D| = 300$ selected lakes are selected from the top 10% largest lakes (19.18–71258.50 ha), while 38% are selected from the next smallest decile (10.25–19.18 ha). The selection further

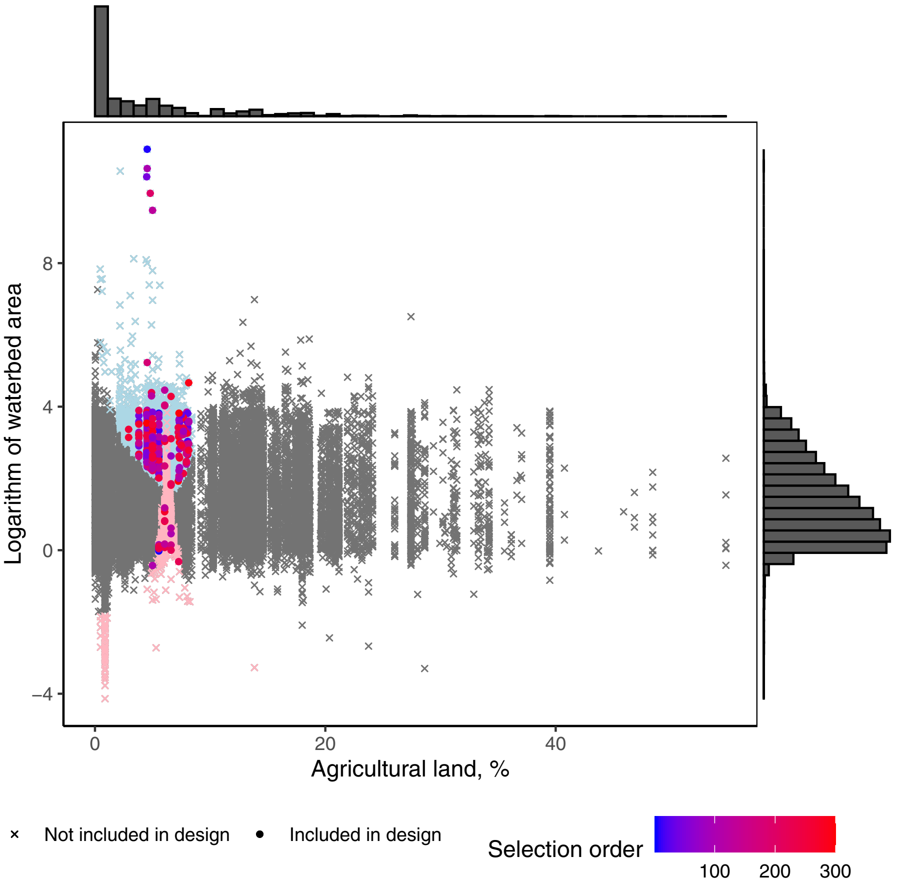

FIGURE 5 The relation between agricultural land (first axis) and waterbed area (second axis) of the full data of 54,347 lakes with the marginal distributions. The circles are the selected lakes by algorithm and the crosses are the lakes excluded in the design. The color-code shows the selection order of the 300 selected lakes. In addition, the color-coded crosses show the 4748 interesting lakes selected based on two criteria: PV_s (light pink) and self-effect VOI (light blue).

seems to prefer high to average agricultural land covariate values. In fact, 18% of the $|D| = 300$ selected lakes are selected from those areas with the top 20% amount of agricultural land (7%–54% agricultural land of the municipal area), 81% are selected from the next 20% (3%–7% agricultural land of the municipal area) and only 1% are selected from municipalities with a lower amount of agricultural land (0%–0.8% agricultural land of the municipal area). This makes sense because there is more ambiguity in the management decision for average agricultural land covariates, leading to high VOI values. From a purely statistical perspective, one would expect very high and low covariates to provide more information about the regression parameter, and in doing so reduce the uncertainty going into the VOI calculations. Here, there is a large amount of data at the initial step, and this element of regression fitting appears to be less relevant in the design.

5 | DISCUSSION

The results in Figures 3–5 show the performance of a forward selection strategy to find useful designs. We now compare different designs of size $|D| = 50$, with the goal of searching for the optimal design. This will tell us if the sequential method performs reasonably or if this way of greedy augmentation of designs overlooks high-value sampling designs. The size of $|D| = 50$ is chosen because the gap between the VOI and the curve for the cost of a scenario with 12 chlorophyll-*a* samples in Figure 3 appears to be at its largest for this design size.

We search for more optimal designs based on the exchange algorithm and the Bayesian optimization approach of Paglia et al. (2022). For the exchange algorithm, we start the iterative routines with the optimal set of size $|D| = 50$ from the forward evaluation. The exchange of two lakes (one removed from the design and the other added to the design) is based on probabilities vaguely honoring high marginal self-effect of VOI. The Bayesian optimization algorithm starts with 1000 evaluations of the exchange algorithm, and continues with batches of 100 VOI evaluations selected from the expected improvement over 1000 designs.

Figure 6 shows the percentage increase in the VOI as a function of iterations of the two algorithms. This is illustrated for VOI evaluation number on the first axis, and over ten independent runs of the exchange algorithm (solid, red) and

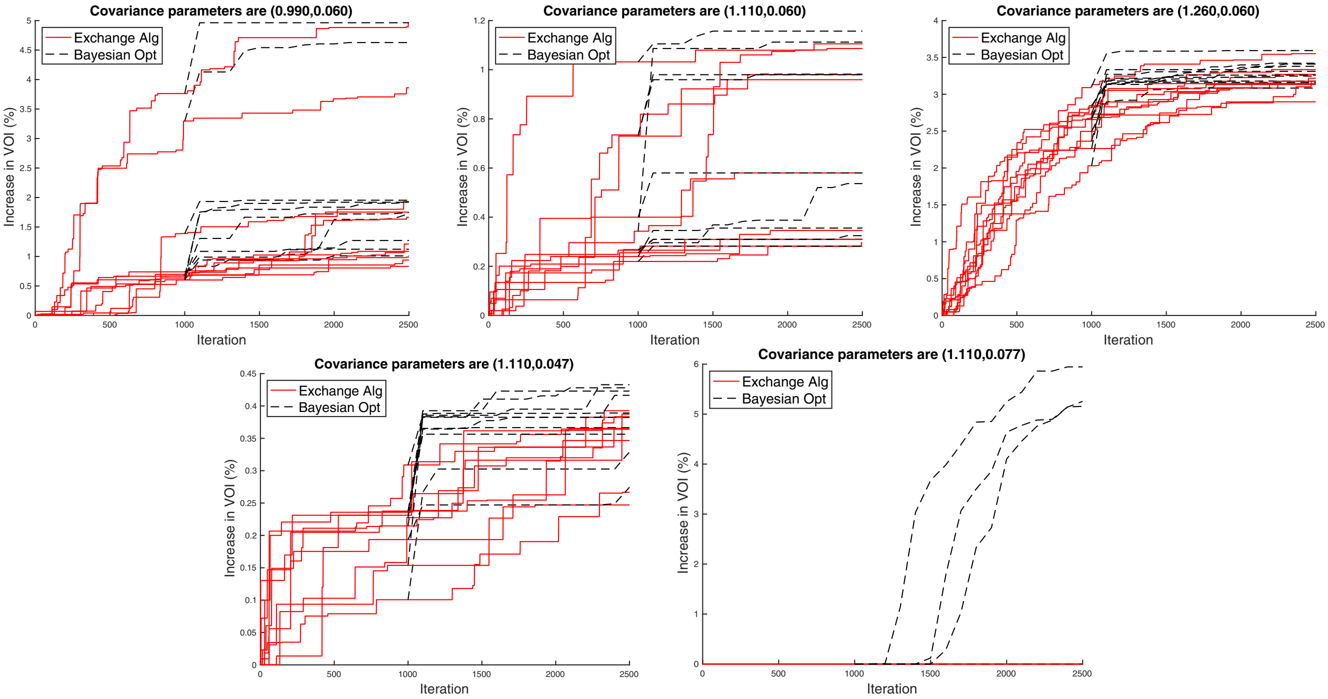

FIGURE 6 VOI results of the exchange algorithm (solid, red) and the Bayesian optimization approach (dashed, black) for 10 independent runs. The five displays reflect different spatial covariance parameters: low variance (top left), benchmark inputs (top center), high variance (top right), low correlation decay (bottom left), high correlation decay (bottom right). Results are shown as percentage VOI increase over evaluations, relative to the sequential forward selection results for a design size of 50. The first 1000 iterations are common. After that the Bayesian optimization algorithm runs batches with a size of 100 using expected improvement of designs.

TABLE 2 VOI(D) computed for different designs of size $|D| = 50$ using different spatial covariance parameters (σ, ϕ). Here, M refers to million euro and K refers to thousand euro.

	VOI(D)				
	$\sigma = 1.11$	$\sigma = 0.99$	$\sigma = 1.26$	$\sigma = 1.11$	$\sigma = 1.11$
Design	$\phi = 0.06$	$\phi = 0.06$	$\phi = 0.06$	$\phi = 0.047$	$\phi = 0.077$
Sequential best	3.40 M	2.15 M	4.65 M	3.86 M	2.75 M
Exchange	3.43 M	2.25 M	4.81 M	3.87 M	2.75 M
Bayes optimization	3.44 M	2.26 M	4.82 M	3.88 M	2.91 M
Highest self-effect	860 K	511 K	1.29 M	1.32 M	546 K
Largest lakes	548 K	-	-	-	-
Geo-spreading	381 K	-	-	-	-
High agriculture	390	-	-	-	-

the Bayesian optimization scheme (dashed, black). The reference case (center, top row) has parameters $\sigma = 1.11$ and $\phi = 0.06$. For the reference case, both the exchange algorithm and Bayesian optimization obtain better designs than the greedy result. The largest VOI improvement for the exchange algorithm is about 1.1% while it is 1.2% using Bayesian optimization. The other displays show VOI increase in cases where the model parameters indicate high/low variance or high/low spatial correlation. Similar to what we see in the reference case, there are designs with higher VOI than that achieved by the sequential forward selection, which indicates that more nuanced algorithms could further improve this. Nevertheless, from what we see here, it is not straightforward to get a much higher value than that obtained by the sequential forward selection (it is only about 0%–5%). For the case with the fast spatial correlation decay parameter, none of the ten exchange algorithm runs, and only three of the ten Bayesian optimization runs, managed to improve the design. This case represents less dependence, and intuitively the sequential method performs better. Still, some of the new designs detected by Bayesian optimization are significantly better, but the search seems more difficult with these parameter settings.

We will now compare designs with a size of 50 based on other criteria. Going beyond the statistical models and decision analytic views, policy makers could have other elements they must consider, and it is insightful to show the VOI results of designs based on a variety of principles. Again, we focus this discussion on designs of size $|D| = 50$. First, we select 50 lakes with the highest self-effect $\mathrm{VOI}_s(\{\boldsymbol{u}_s\})$ from the whole lake data. Second, we calculate the VOI of the 50 lakes with the largest waterbed area from the whole data. Third, we test the set of 50 lakes which are spread out as much as possible on the map of Finland. The spreading was set by selecting the lakes from each county of Finland. There are 18 counties in our data, meaning we randomly selected 2 or 3 lakes from each county. Fourth, we also formed a design with the 50 lakes having highest covariate value (agricultural land in Figure 2, left). Table 2 summarizes the VOI of the greedy selection, the exchange and Bayesian optimization selection (maximum of 10 runs doing 2500 evaluations), as well as the other designs with simpler selection criteria listed above, when varying the statistical covariance parameter values (σ, ϕ). It seems that the VOI of these designs remain very small compared to the results we achieve with the statistical algorithms. Large values of σ seem to produce larger values of VOI.

We therefore recommend that policy makers use statistical methods in the design construction. Making monitoring plans on the highest self-effect alone misses out on the correlations in the statistical model and the interactions of having very similar lakes in a design. For the designs with a size of 50, it gets less than one-fourth of the VOI compared with the more nuanced search approaches. Designs that either focus on geographical coverage, large lake area or high agricultural covariates do not necessarily capture the interesting lakes for ecological purposes. The greedy algorithm succeeds in finding a reasonably good design at moderate computation costs here, and provides a reference for the additional search approaches.

6 | CONCLUSION

We have demonstrated approximate optimal design selection methods that aim to maximize the VOI of the design compared with the cost of the data acquisition of the design. The VOI selection criterion assesses the profitability of designs when taking into account the costs and benefits of the decisions as well as the associated uncertainty. This approach is exemplified in the context of lake management in Finland. Similar design questions occur in a range of applications, such as the environmental studies brought up in Section 1 related to coral monitoring, animal habitat conservation or the mapping of mine tailings. Decision-makers must plan wisely where to conduct environmental sampling so as to obtain valuable information and to maintain budget limitations.

We calculated the VOI assuming a Bayesian spatial logistic regression model for the ecological status data. Statistical model parameters were obtained from existing data gathered in Finnish lakes. Our VOI calculations relied on approximations of functions and integrals for hierarchical general linear models, which we coupled with the large-size design selection procedures required for the lake monitoring case.

In addition to the heuristic greedy forward selection method, which sequentially adds units into the design, we tested two other heuristics for improved selection result: an exchange algorithm based on randomness and enlightened exchange based on the single-lake VOI, and a selection algorithm based on Bayesian optimization. The VOIs achieved with these statistical approaches were much higher than that of other design criteria based on the initial marginal values, geographical spread, high model covariate values or large lake areas.

We are aware that many considerations must be made in order to calculate the VOI of lake monitoring design in practice, and we are limiting our results. For example, we chose to use chlorophyll-*a* as our ecological indicator of interest, among many, and we focused on the costs of collecting that one indicator. In reality, the lake monitoring process is a more complex exercise. Furthermore, analyzing the monitoring data of one lake produces the ecological status of that lake, and it does not take into account how much monitoring data was used for the classification. In addition, we thought that associating the costs and revenues to the lake areas would have an impact on the results. However, the area seems to have less effect on the selection than we assumed.

Since the problem of optimal design has been widely examined in statistics, there exist many other heuristic methods to solve this problem. We believe it is possible to obtain better designs than we did here. Our purpose was to highlight the possibility of forming a statistically based design for these large-size spatial logistic regression models, and in doing so we see that they clearly outperform designs made from basic principles.

This study does not consider any temporal variation in the ecological status of lakes. Spatio-temporal variation in lake status would thus be interesting to address in future work. In this paper, we relied on earlier studies considering the management decision space and associated costs. In the future it would be relevant to expand the space of management alternatives to a more detailed level, and see how this influences the selection of lakes for the design.

ACKNOWLEDGMENTS

VK acknowledges support from the Emil Aaltonen Foundation and the Kone Foundation. JE acknowledges support from the Norwegian Research Council via relevant grants 305445 and 309960.

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data supporting the analyses are downloadable from https://nextcloud.jyu.fi/index.php/s/2BjQjkzWsaBYRr8. The data that support the findings of this study are openly available in Koski&Eidsvik2022 at https://nextcloud.jyu.fi/index.php/s/PmqmRDTp4F4sC3Z.

ORCID

Vilja Koski https://orcid.org/0000-0002-5970-3582

REFERENCES

Abbas, A. E., & Howard, R. A. (2015). *Foundations of Decision Analysis*. Pearson Higher Ed.

Anyosa, S., Eidsvik, J., & Pizarro, O. (2023). Adaptive spatial designs minimizing the integrated bernoulli variance in spatial logistic regression models-with an application to benthic habitat mapping. *Computational Statistics & Data Analysis*, *179*, 107643.

Aroviita, J., Mitikka, S., & Vienonen, S. (2019). *Pintavesien tilan Luokittelu ja Arviointiperusteet Vesienhoidon Kolmannella Kaudella*. Finnish Environment Institute (SYKE) (In Finnish). https://helda.helsinki.fi/handle/10138/306745Placeholder Text.

Arthur, J., Hachey, M., Sahr, K., Huso, M., & Kiester, A. (1997). Finding all optimal solutions to the reserve site selection problem: formulation and computational analysis. *Environmental and Ecological Statistics*, *4*, 153–165.

Canessa, S., Guillera-Arroita, G., Lahoz-Monfort, J. J., Southwell, D. M., Armstrong, D. P., Chadès, I., Lacy, R. C., & Converse, S. J. (2015). When do we need more data? A primer on calculating the value of information for applied ecologists. *Methods in Ecology and Evolution*, *6*(10), 1219–1228.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. John Wiley & Sons.

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A. H., Soto, D., Stiassny, M. L., & Sullivan, C. A. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews of the Cambridge Philosophical Society*, *81*(2), 163–182.

Dykstra, O. (1971). The augmentation of experimental data to maximize [X'X]. *Technometrics*, *13*(3), 682–688.

Eidsvik, J., Mukerji, T., & Bhattacharjya, D. (2015). *Value of Information in the Earth Sciences: Integrating Spatial Modeling and Decision Analysis*. Cambridge University Press.

European Parliament (2000). Directive 2000/60/EC, of the European Parliament and Council of 23 October 2000 establishing a framework for Community action in the field of water policy.

Evangelou, E., & Eidsvik, J. (2017). The value of information for correlated GLMs. *Journal of Statistical Planning and Inference*, *180*, 30–48.

Foss, K. H., Berget, G. E., & Eidsvik, J. (2022). Using an autonomous underwater vehicle with onboard stochastic advection-diffusion models to map excursion sets of environmental variables. *Environmetrics*, *33*(1), e2702.

Harman, R., Filová, L., & Richtárik, P. (2020). A randomized exchange algorithm for computing optimal approximate designs of experiments. *Journal of the American Statistical Association*, *115*(529), 348–361.

Hays, S., Kumari, B., Stewart-Koster, B., Boone, E., & Sheldon, F. (2021). Site reduction in redundant ecosystem sampling schemes. *Environmental and Ecological Statistics*, *28*, 1–20.

Higgins, J., Zablocki, J., Newsock, A., Krolopp, A., Tabas, P., & Salama, M. (2021). Durable freshwater protection: A framework for establishing and maintaining long-term protection for freshwater ecosystems and the values they sustain. *Sustainability*, *13*(4), 1950.

Jauslin, R., Panahbehagh, B., & Tillé, Y. (2022). Sequential spatially balanced sampling. *Environmetrics*, *33*(8), e2776.

Jeppesen, E., Søndergaard, M., & Liu, Z. (2017). Lake restoration and management in a climate change perspective: An introduction. *Water*, *9*(2), 122.

Koski, V., Kotamäki, N., Hämäläinen, H., Meissner, K., Karvanen, J., & Kärkkäinen, S. (2020). The value of perfect and imperfect information in lake monitoring and management. *Science of the Total Environment*, *726*, 138396.

MATLAB. (2021). *Version 9.11.0 (R2021b)*. The MathWorks Inc.

Nguyen, L., Ulapane, N., & Miro, J. V. (2018). *Adaptive sampling for spatial prediction in environmental monitoring using wireless sensor networks: A review*. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)* (pp. 346–351). IEEE.

Nõges, P., van de Bund, W., Cardoso, A. C., Solimini, A. G., & Heiskanen, A.-S. (2009). Assessment of the ecological status of European surface waters: A work in progress. *Hydrobiologia*, *633*, 197–211.

Official Statistics of Finland. (2020a). *Buildings and free-time residences [e-publication]*. Statistics Finland http://www.stat.fi/til/rakke/index·en.html

Official Statistics of Finland. (2020b). *Utilised Agricultural Area [e-publication]*. Natural Resources Institute Finland http://www.stat.fi/til/kaoma/index·en.html

Paglia, J., Eidsvik, J., & Karvanen, J. (2022). Efficient spatial designs using Hausdorff distances and Bayesian optimization. *Scandinavian Journal of Statistics*, *49*(3), 1060–1084.

Prentius, W., & Grafström, A. (2022). Two-phase adaptive cluster sampling with circular field plots. *Environmetrics*, *33*(5), e2729.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Reich, B. J., Pacifici, K., & Stallings, J. W. (2018). Integrating auxiliary data in optimal spatial design for species distribution modelling. *Methods in Ecology and Evolution*, *9*(6), 1626–1637.

Shun, Z., & McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(4), 749–760.

Søndergaard, M., Jeppesen, E., Lauridsen, T. L., Skov, C., Van Nes, E. H., Roijackers, R., Lammens, E., & Portielje, R. (2007). Lake restoration: Successes, failures and long-term effects. *Journal of Applied Ecology*, *44*(6), 1095–1105.

Stankey, G., Clark, R., & Bormann, B. (2005). *Adaptive Management of Natural Resources: Theory, Concepts, and Management Institutions*. USDA Forest Service General Technical Report PNW.

Thilan, A., Menéndez, P., & McGree, J. (2023). Assessing the ability of adaptive designs to capture trends in hard coral cover. *Environmetrics*, *34*, e2802.

Williams, B. K., & Brown, E. D. (2020). Scenarios for valuing sample information in natural resources. *Methods in Ecology and Evolution*, *11*(12), 1534–1549.

Woods, D. C., Overstall, A. M., Adamou, M., & Waite, T. W. (2017). Bayesian design of experiments for generalized linear models and dimensional analysis with industrial and scientific application. *Quality Engineering*, *29*(1), 91–103.

How to cite this article: Koski, V., & Eidsvik, J. (2024). Sampling design methods for making improved lake management decisions. *Environmetrics*, e2842. https://doi.org/10.1002/env.2842