Bayesian adaptive design for citizen science data collection: Exploring tensions between data and design

Max Savery[1] and Stijn Luca[1]

[1]Ghent University, Department of Data Analysis and Mathematical Modelling
Coupure links 653, B-9000 Ghent, Belgium

1 Abstract

1. Bayesian adaptive design can be applied in spatial settings where future survey locations need to be selected based on already available data. An important use case of adaptive design is the recommendation of locations for opportunistic, citizen science collection of species observation data, where some areas are already overrepresented and others are severely undersampled.

2. This work conducts an extensive comparative study of adaptive Bayesian design approaches, specifically for recommending survey locations for citizen science data collection. We explore compromises between design-based, model-based, and exchange-based approaches in order to better understand the statistical implications of using adaptive design for citizen science projects.

3. To evaluate the adaptive design approaches effectively, we work in a simulated environment, allowing us to compare the performance between fourteen different design

strategies. For each method, we vary design size and the initial data conditions, and in a separate analysis show the effect of weighting the site recommendations by varying degrees of citizen scientist preference. Additionally, we conduct a sensitivity experiment on the approximation used to integrate out the uncertainty associated with the future data collection.

4. Our results highlight the tension between the data and the design methodology: We show that the model- and exchange-based approaches perform better when more prior data is available, particularly when estimation or prediction accuracy is preferred. In contrast, design-based methods are more stable when prior information is limited. However, it remains a challenge to balance all design priorities and data scenarios in a single approach. Additionally, we show that including citizen scientist preference in the utility function can impact the design in surprising ways. These outcomes are not uniform and depend on the method used to optimize the design, the metric of interest for optimization, as well as the amount of prior data and the desired size of the future survey. Taken together, our results show that effective survey design must be carefully matched to both the available baseline information and the study objectives.

Keywords

bayesian adaptive design, adaptive sampling, spatial design, opportunistic data collection, citizen science, biodiversity monitoring, species distribution modelling

2 Introduction

Traditional surveys for estimating species distributions are designed and carried out according to an imposed structure, ensuring that a statistical model can be fit without bias in the parameter estimation. However, unstructured citizen science data has become a useful and alternative source of species observation data [Zipkin et al., 2021]. For example, citizen science presence-only (PO) data has been shown to be a valuable supplement to presence-absence (PA) survey data [Dorazio, 2014, Koshkina et al., 2017, Simmonds et al., 2020] when both data sources are combined in an integrated model. Though such citizen science data is ubiquitous and has demonstrated benefit for ecological modelling, it exhibits highly opportunistic data collection patterns [Johnston et al., 2023]. It is therefore desirable to direct citizen scientists to locations that maximize information complementary to the data already collected.

To recommend locations, the framework of Bayesian adaptive design can be applied to survey design for species monitoring. Generally, Bayesian design is used to find a configuration of design points that optimize a utility or loss function of interest. Adaptive design can be considered a useful extension when previously collected data is available [Mateu and Muller, 2013]. In the species monitoring setting, this is framed as the construction of a survey conditioned on data previously collected at a limited set of locations. The design itself is characterized by the number of sites to visit, the locations of the sites, and the number of visits to make at each site [Reich, 2020]. However, there is a limited body of research and no standard guidance regarding design strategies for citizen scientists–Bayesian, adaptive, or otherwise. Based on the few works that do consider survey design, there appears to be substantial advantages to including adaptive design

methodology in citizen science data collection. Callaghan et al. [2019] use statistical leverage to assign a marginal value to each sampling event, in order to predict leverage values for future sampling events, and they highlight the importance of sampling both hotspots and unsampled areas, depending on desired design resolution. Mondain-Monval et al. [2024] compare six design- and model-based approaches for adaptive design. They target uncertainty and rare species in their designs and consider varying levels of citizen science uptake of the recommended locations. They find that even a small proportion of adaptively sampled data improved species distribution model performance. Flint et al. [2024], using citizen science roadkill as a case study, study the utility of additional sampling by comparing relative intensities between models with and without predictions made at an additional design point. Importantly, they incorporate organizational value and financial cost of additional sampling into their utility function. They highlight that their framework can avoid resource waste and select high-utility locations that are particularly valuable for future sampling. While these works provide excellent practical approaches to the design of adaptive surveys for citizen scientists and consider many factors important for citizen science surveying, they do not, however, embed their methods within the broader framework of Bayesian design. Furthermore, they do not offer multifaceted comparisons of algorithms for adaptive design, nor do they provide a general framework for including citizen scientist preference in the design optimization.

Beyond citizen scientist data collection, there are a few other notable ecological applications of Bayesian adaptive design. Williams et al. [2018] use the Bayesian adaptive design framework for creating designs that measure dynamic populations. In their work, future data is simulated from the posterior predictive distribution of prior data, in order to

minimize abundance uncertainty and determine the optimal next site to visit. Similarly, Leach et al. [2022] formalize the recursive nature of Bayesian adaptive design for ecological applications, specifically when using the Prior-Proposal Recursive Bayes approach of Hooten et al. [2021]. Thilan et al. [2023] and Thilan et al. [2024] evaluate the Bayesian adaptive design framework in its use for tracking changes and trends in ecological monitoring, using a Laplace approximation to estimate the posterior distribution and assess the change in information when adding new sites to the already collected data with Kullback-Leibler (KL) divergence.

When planning and managing citizen science projects, not only are the statistical properties of the design important, it is also essential to consider citizen scientist behaviour. Callaghan et al. [2023] study the extent to which citizen scientists are willing to follow such adaptive sampling schemes and show that "behavioural nudges" do in fact incentivize citizen scientist to survey recommended locations. Pocock et al. [2023] study the effect of different types of citizen scientist recorder behaviour on the estimates of species trend and occupancy across different butterfly species in the UK. August et al. [2020] characterize the behaviour of citizen scientists using a variety of derived metrics based on aspects of the data collection in space, time, and content. These scores are then clustered and the principal components analysed in order to describe the behaviour via continuous axes. In addition, Callaghan and Gawlik [2015] compare estimates of species abundance derived from eBird data and from structured survey data, finding citizen science surveys can be used as an effective substitute for the structured surveys. Callaghan et al. [2022] focus on sampling effort, highlighting the relationship between the amount of collected citizen science data, sample completeness, and species diversity. Thompson et al.

[2023] survey participants of a citizen science project and find that some people may be willing to change their surveying behaviour if they knew the change would lead to data more valuable for the project. However, while the works mentioned here do address citizen scientist behaviour and the impact of citizen science data collection on downstream modelling outcomes, they do not fully address the statistical issues related to adaptive design for citizen scientists. Henrys et al. [2024] review at a broad level the issues with creating adaptive designs for citizen scientists, suggesting that clearer guidance regarding the statistical methodology and additional examples of its application are required for the uptake of adaptive design methodology.

The literature described above indicates that there are certainly advantages of using the framework of Bayesian adaptive design to recommend site locations for citizen science data collection. It is also evident that while there are a number of potential approaches for Bayesian adaptive designs for citizen scientists, the advantages, disadvantages, and practicalities of implementing Bayesian design algorithms are not well-studied. This leaves us with the question as to what works best for citizen science applications, what refinements are needed in the Bayesian design framework, and how can we include the most information from both a statistical perspective and a practical one when generating designs and recommending potential survey locations that can be provided either directly to an interested citizen scientist or via an organization tasked with developing a citizen science survey project. Building on the adaptive design work of Callaghan et al. [2019], Flint et al. [2024], and Mondain-Monval et al. [2024], and considering the gaps in the literature discussed by Henrys et al. [2024], in this study we explore the associated statistical advantages and complications that come with using Bayesian adaptive design for

citizen science applications. We expand specifically upon Mondain-Monval et al. [2024] by conducting a more extensive comparison of design- and model-based methods, as well as the exchange family of algorithms, and we consider the practical challenges of implementing model-based and exchanged-based approaches in a Bayesian design framework. An additional novel contribution of our work is to show the effect of considering citizen scientist accessibility or preference on the utility functions used with the design optimization. Importantly, we also perform a sensitivity analysis on the effect of drawing future data from the posterior predictive distribution of the model fit to the initial observations, a necessary step in Bayesian adaptive design but one that is rarely assessed. Finally, to conclude our comparative study, we provide a list of five high-level recommendations for the adaptive design of citizen science surveys. These recommendations take into account the the compromise between the computation efficiency of design-based approaches and the flexible optimization of the more computationally intensive Bayesian model- or exchange based algorithms, as well as the effect of incorporating citizen scientist preference in the optimization of the designs themselves.

By putting these aspects of design and citizen science data collection together in a single work, we seek to make Bayesian adaptive design more transparent and accessible to organizations with interest in designing surveys for citizen scientists.

3 Methods

In the this section we introduce the framework of Bayesian adaptive design and describe the design approaches we use for creating Bayesian adaptive designs for citizen scientists.

3.1 Citizen science data

We refer to a PO (presence-only) dataset as a set of points $\mathbf{y_0} = \{s_1, s_2, ..., s_n\}$ collected in region $\mathcal{D}$, with n total observations in the dataset. The region D is divided into c cells of equal area $|A_i| = |D|/c$, where A_i refers to each cell or site. Following Mondain-Monval et al. [2024] we use a virtual ecologist approach [Zurell et al., 2010] to study the effect of different adaptive design methodologies on citizen science data collection and species distribution model (SDM) estimation. We use the entire domain of the country of Belgium as a case-study for survey design, simulating species observations via covariates derived from remote-sensing data within the country. We briefly summarize the initial data generation workflow as follows. The data generation and model estimation details are more fully described in the Supplementary material.

1. We use a Log-Gaussian Cox Process (LGCP) as the true data-generating SDM. This data-generating model is used to generate the initial dataset $\mathbf{y_0}$, which we treat as the initial, fixed set of citizen science observations. In practice, the $\mathbf{y_0}$ dataset will be the previously collected data upon which the adaptive design is built.

2. We then fit a working model to the simulated $\mathbf{y_0}$ data. At this step, we again use the LGCP, and we induce a misspecification of the bias function in order to mimic situations where the data generating process is not fully known or model specification may be difficult.

3. Using the posterior from the working model, we simulate R posterior predictive datasets, referred to as $\tilde{\mathbf{y}}_\mathbf{1}^{(\mathbf{r})}$. These R datasets represent possible future data outcomes and are used within the Bayesian design framework to approximate the

expectation of the utility function.

In practical applications of Bayesian design, it will be necessary to use the posterior resulting from $\mathbf{y_0}$ to generate the $\tilde{\mathbf{y}}_\mathbf{1}$ datasets, and the working procedure that we use here reflects that reality. This is explained in more detail in the next section.

3.2 Bayesian adaptive design

The purpose of Bayesian adaptive design is to find the configuration of design points $\mathbf{d}^*$ within region $\mathcal{D}$ that maximizes the expected utility or minimizes the expected loss taken over the joint distribution $p(\boldsymbol{\theta}, \mathbf{y_1}|\mathbf{y_0}, \mathbf{d})$ of parameters $\boldsymbol{\theta}$ and future data $\mathbf{y_1}$, conditional on the prior data $\mathbf{y_0}$. In this work we primarily refer to maximization of the expected utility function $\mathrm{E}[U(\mathbf{d})|\mathbf{y_0}]$:

$$\mathbf{d}^* = \mathrm{argmax}_{\mathbf{d}\in\mathcal{D}}\mathrm{E}[U(\mathbf{d})|\mathbf{y_0}]. \tag{1}$$

The Bayesian adaptive design framework accounts for the uncertainty associated with both future data collection and parameter estimation associated with the joint distribution. However, a primary challenge is the approximation of the expectation of the utility for every proposed design. The expected utility is defined as

$$\mathrm{E}[U(\mathbf{d})|\mathbf{y_0}] = \int_{\mathbf{y_1}}\int_{\boldsymbol{\theta}} u(\mathbf{y_1}, \boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}|\mathbf{y_1}, \mathbf{y_0}, \mathbf{d}) p(\mathbf{y_1}|\mathbf{y_0}, \mathbf{d}) d\boldsymbol{\theta} d\mathbf{y_1}, \tag{2}$$

where $p(\boldsymbol{\theta}|\mathbf{y_1}, \mathbf{y_0}, \mathbf{d})$ is the posterior of model parameters, and $p(\mathbf{y_1}|\mathbf{y_0}, \mathbf{d})$ is the posterior predictive distribution (PPD) induced by the historical data $\mathbf{y_0}$, where $p(\mathbf{y_1}|\mathbf{y_0}, \mathbf{d}) = \int_{\boldsymbol{\theta}} p(\mathbf{y_1}|\boldsymbol{\theta}, \mathbf{d}) p(\boldsymbol{\theta}|\mathbf{y_0}) d\boldsymbol{\theta}$. This distinguishes adaptive design from non-adaptive

design where the prior predictive distribution $p(\mathbf{y}|\mathbf{d}) = \int_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{d})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is used instead, with no distinction between $\mathbf{y_1}$ and $\mathbf{y_0}$ because there is no historical data. The use of recursive posterior predictive distributions in ecology for the integration over the utility is further discussed in Williams et al. [2018] and Leach et al. [2022].

To approximate the expected utility, it is common to use Monte Carlo integration:

$$\mathrm{E}[U(\mathbf{d})] \approx \frac{1}{R}\sum_{r=1}^{R} u(\boldsymbol{\theta}^{(\boldsymbol{r})}, \tilde{\mathbf{y}}_{\mathbf{1}}^{(\mathbf{r})}, \mathbf{d}). \tag{3}$$

As it is necessary to draw R datasets and fit R posteriors to approximate the expected utility for each proposed design, this step is the bottleneck of Bayesian design. This problem is exacerbated when using an iterative search algorithm such as exchange algorithm [Royle, 2002]. And, due to the necessity of fitting multiple posteriors for every update of the design, approximating the posterior with a faster approach such as that of the Laplace approximation does not necessary alleviate the core computational issues. This bottleneck will motivate the design strategies discussed below.

3.3 Design strategies

We now describe the design strategies compared in this work. We use three general approaches: Design-based, model-based, and exchange-based. Design-based strategies are also referred to as empirical in Henrys et al. [2024], meaning designs are developed based only on observed properties of the available data or design region. This includes, for example, the classic space-filling approach. Model-based designs involve algorithmic

optimization of a model-dependent quantity obtained after a statistical model is fit to the available data. Exchange-based designs use an exchange algorithm or one of its variants to optimize the design locations iteratively. Though the designs generated via exchange algorithm inherently use a model to optimize the utility or loss function, we make a distinction between these and the "model-based designs" in order to distinguish between designs that perform an iterative search and model-based methods that optimize via a greedy rank-based selection after a single iteration of model-fitting. All the approaches are summarized in Table 1, and more elaborate descriptions are provided in the Supplementary material. In this work, whenever we refer to a model or the resulting posterior, we use a LGCP and its estimation of the latent intensity of the point process. This model is fit using the INLA and SPDE method [Rue et al., 2009, Lindgren et al., 2011] and the inlabru package [Bachl et al., 2019]. All further details regarding implementation are described in the Supplementary materials.

Table 1: Design strategies considered in the simulation study, separated by design-based, model-based, and exchange-based. Full algorithmic details are provided in the Supplement.

Design-based	
Random + space-filling	Randomly select m sites subject to a minimum inter-site distance of 5 km.
Unsampled regions	Randomly select m sites from locations with the fewest existing observations.
Stratified sampling	Randomly select m sites, stratified between sampled and unsampled regions.

Model-based	
Maximum predictive variance	Select the m sites with the largest posterior predictive variance under the model fit to $\tilde{\mathbf{y}}_{\mathbf{1}}^{(\mathbf{r})}$, targeting sites of high potential uncertainty [Mondain-Monval et al., 2024].
Weighted predictive variance	Weight predictive variance by the inverse number of samples in each site, balancing the selection of uncertain locations with poorly sampled locations [Mondain-Monval et al., 2024]
Rejection sampling	Randomly propose sites and accept with probability proportional to scaled predicted intensity, including threshold to control concentration in high-intensity areas [Liu and Vanhatalo, 2020].
Additive KL	For each candidate site, augment $\mathbf{y_0}$ with site-specific observations from $\tilde{\mathbf{y}}_{\mathbf{1}}^{(\mathbf{r})}$ and rank sites by the KL divergence between updated and baseline posteriors.
LOO KL	For each candidate site, start with the full $\tilde{\mathbf{y}}_{\mathbf{1}}^{(\mathbf{r})}$ dataset and remove site-specific observations and rank sites by the KL divergence between the full-data and leave-one-out posteriors.
Weighted LOO	Apply an intensity-based weight during the LOO optimization to favor lower-intensity regions.
Utility emulation (GP)	Use a Gaussian process emulator of the utility surface to identify sites with high predicted utility [Overstall and Woods, 2017]. We use both an Additive and LOO version of this approach.

Exchange-based	
Exchange algorithm	Iteratively replace selected sites with nearby candidates whenever utility improves [Royle, 2002].
Approximate Coordinate Exchange (ACE)	Construct the design using the Gaussian emulation of the utility surface and improve on this initial configuration with ACE stochastic updates [Overstall and Woods, 2017, Buchhorn et al., 2024b]. We implement both LOO and Additive versions.

3.4 Citizen scientist preference

We also evaluate the effect of weighting the site selection by citizen science preference. To weight the utility, we use the function $\tilde{U}(\mathbf{d}) = \alpha P(\mathbf{d}) + (1-\alpha)U(\mathbf{d})$, where α balances the unweighted utility and a preference function. We use two different features as a proxy for preference: the distance of each site from the nearest primary road in Belgium and the distance from the user's current location. To smooth the preference scores we use a Gaussian decay function to transform raw distances $d(s)$ from a given feature of interest: $P(\mathbf{d}) = \exp\left(-\frac{d(s)^2}{2\gamma^2}\right)$, where γ is the median of the distances. Both utilities and preference scores are standardized before combination. Other useful metrics of accessibility and preference are possible, such as classifying different types of landcover as more preferable to others, the cost of each visit for the user, the personal user collection history, or the desired sampling duration.

3.5 Evaluation strategy

To evaluate the optimized designs, we again use the virtual ecologist approach to mimic the situation in which additional data is collected in the real world. However, the evaluation differs from the design optimization procedure in that we use the known fixed effect parameters, as well as a new realization of the spatial field, to generate uncollected $\mathbf{y_1}$ data. Then, using the selected sites of each design $\mathbf{d}$ resulting from the respective design optimization approach, data from $\mathbf{y_1}$ is collected and the working model fit to the updated dataset, $(\mathbf{y_0}, \mathbf{y_{1,d}})$. This is repeated for 50 simulations to account for uncertainty in the collection of $\mathbf{y_1}$. To compare each design approach we use four metrics: Mean Absolute

Error (MAE) of the predicted intensity, predictive variance averaged over all sites, Root Mean Square Error (RMSE) over all fixed effect parameter estimates, and the posterior standard deviation (SD) over all fixed effect parameters. These metrics balance predictive power and model uncertainty, giving a relatively broad assessment of performance of the design methodology. We provide additional information regarding the evaluation process in the Supplemental material.

4 Evaluation of design methodology

4.1 Comparisons of strategies

We now conduct a comparative analysis of the design methodologies described above. We evaluate the designs in conditions when the underlying data generating process leads to both sparse and abundant $\mathbf{y_0}$ data, and we compare designs of size $m \in \{5, 10, 20, 30, 50\}$. Figure 1 shows the full comparison over strategies, methods, and survey sizes. Figure 2 summarizes the boxplot spread per method, and Figure 3 ranks the performance of each method over all metrics.

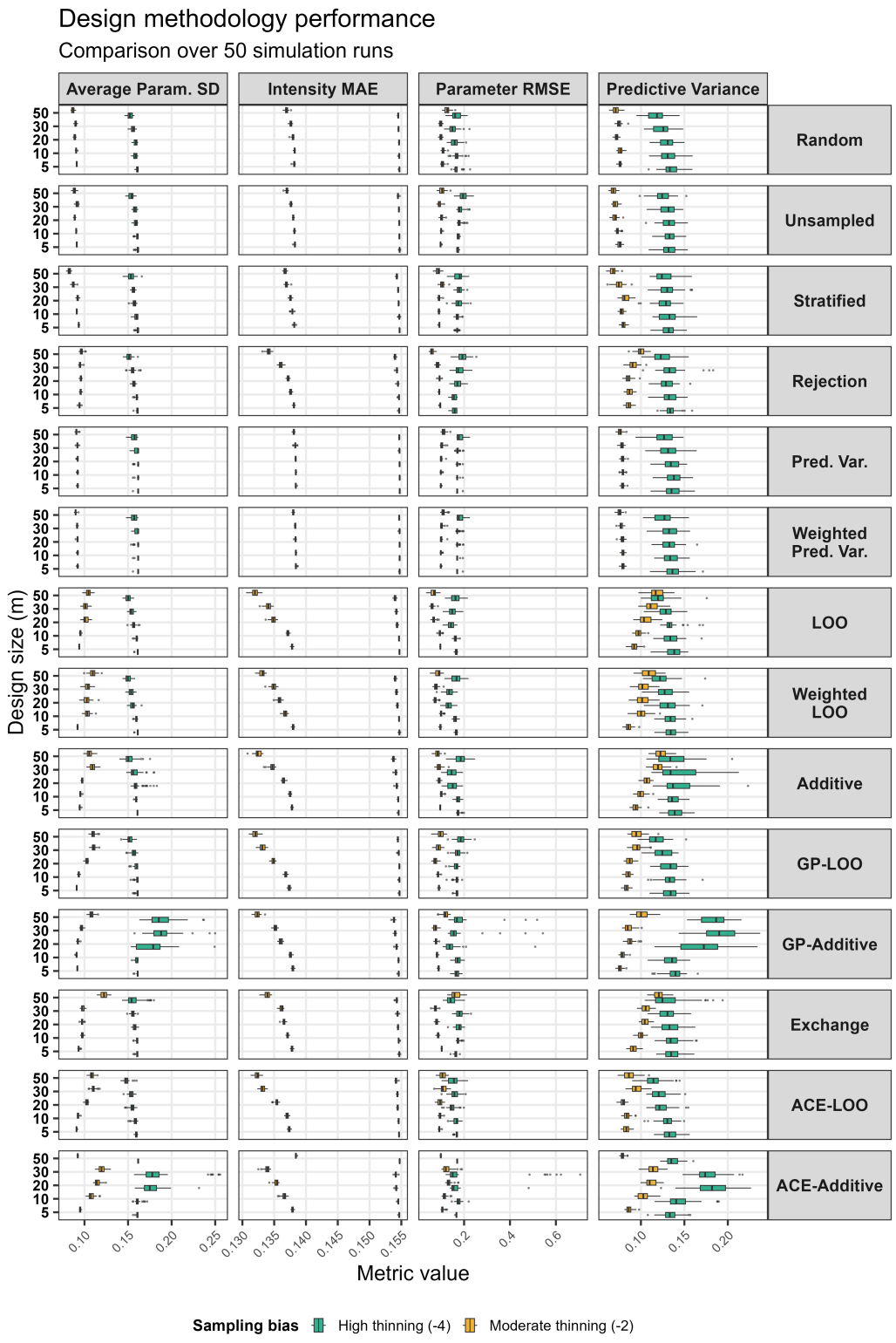


Figure 1: Comparison over methods, metrics, and design survey size. The boxplots show the distribution of performance over 50 simulations of $\mathbf{y_1}$ data collection. Method type is shown on the right y axis; metric type is shown on the top x axis; design size is shown on the left y axis; scores are shown on the bottom x axis. The method types are organized by design-based first, then model-based, then exchange-based. The sparse (high thinning, $\gamma_0 = -4$) and abundant (moderate thinning, $\gamma_0 = -2$) data collection settings are shown within each plot in green and yellow respectively.

Figure 1 makes a few design behaviours clear. Across all methods, the performance trends differ quite dramatically between the two levels of thinning, i.e., the richness of available $\mathbf{y_0}$ data. As described in the Supplementary material, to generate $\mathbf{y_0}$ we thin the data generating process using respective intercepts of $\gamma_0 = -2$ (moderate thinning) and $\gamma_0 = -4$ (high thinning), in combination with a road-distance covariate used as a proxy for sampling bias, with the thinning probability parameterized via a logit link function. This construction mirrors collection scenarios with differing degrees of baseline sampling probability for the citizen scientist, and, from the perspective of the design optimization, changes the information available in the starting conditions for the design methods. Consequently, the initial data conditions can lead to very different outcomes when comparing the design-based approaches to the model- and exchange-based approaches. When data is plentiful (thinning= -2), as is often the case with opportunistic data in the big-data paradigm, all approaches improve in metric performance. But the variability between the simulation runs does not necessarily stabilize when more data is available. When considering the spread of the boxplots (the variability of performance across the 50 simulations), as more clearly visualized in Figure 2, the design-based methods often exhibit a narrower range of behaviour in the thinning= -2 setting than in the -4 setting. And the design-based methods are much less variable across the simulations when compared to the Additive family (Additive, GP-Additive, and ACE-Additive), especially in the thinning= -4 setting. The LOO family shows less spread than the Additive family in general, but it depends on the method and metric.

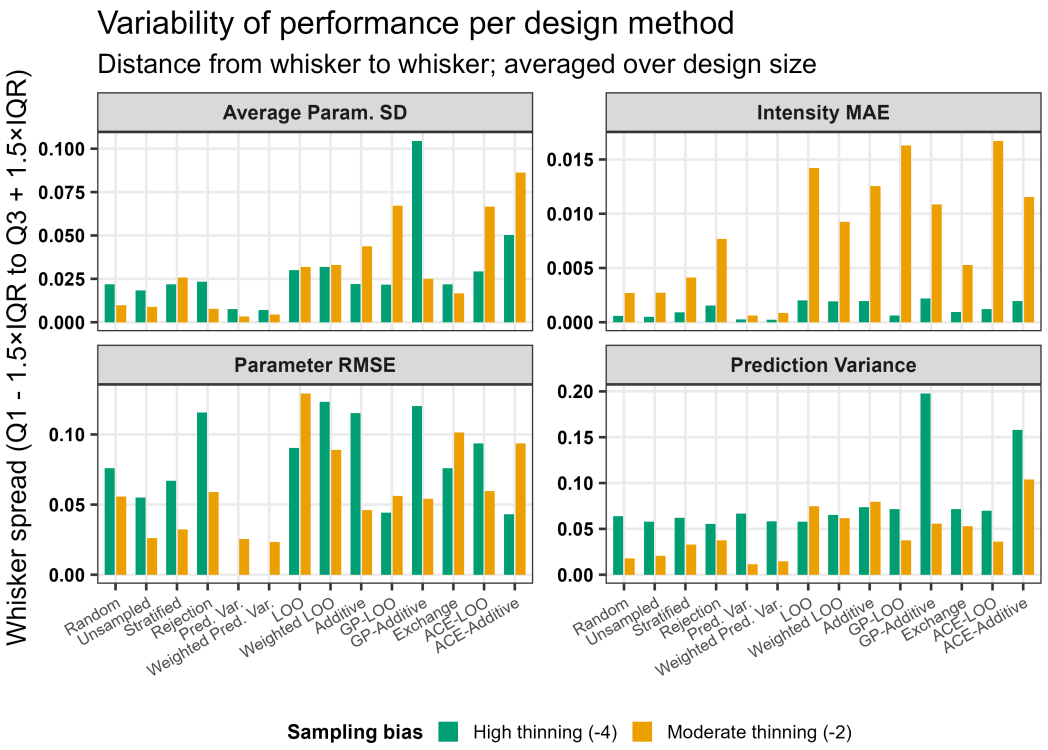


Figure 2: Variability of performance shown in Figure 1 per design method. Bar heights represent the total whisker span, calculated as the distance between $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$. The distances are averaged over all design sizes (m) for each metric and method. The thinning conditions are shown for each method.

Figure 1 also shows that the Rejection method and the LOO and Additive families improve in performance of intensity MAE and parameter RMSE as the size of the design increases, primarily in the abundant thinning= -2 data setting. There are also some improvements as design size increases in certain thinning= -4 scenarios (e.g., the predictive variance of GP-LOO and ACE-LOO, and the parameter SD of the entire LOO family). It can also be seen in Figure 1 that as the design size increases, the median parameter SD and predictive variance both trend lower in the thinning=-4 setting, but trend higher when thinning=-2. This is likely due to conflict between the prior $\mathbf{y_0}$ data and the imperfectly simulated $\tilde{\mathbf{y}}_\mathbf{1}$ datasets during the design optimization in the thinning=-2 setting. We will address this further issue in the discussion.

The overall performance of each method is further highlighted in Figure 3, where the methods are ranked by the total number of "wins", i.e., the best performance per metric and per simulation of $\tilde{\mathbf{y}}_\mathbf{1}$ data collection. In terms of MAE and Parameter RMSE (for thinning$= -2$), LOO and GP-LOO outperform the other approaches. All three design-based approaches (Random, Unsampled, and Stratified) dominate in terms of parameter SD and predictive variance. Then, in the sparse (thinning$= -4$) setting, we see a very different ranking. ACE-LOO, Additive, Weighted LOO, GP-Additive are all top performers, each with somewhat different performance profiles. Interestingly, the ACE-LOO performance is composed of parameter SD, RMSE, and predictive variance, suggesting that it may offer the most balanced performance under sparse-data conditions. We see that the Additive family performs well in terms of MAE in the sparse setting, but these methods have very large boxplot spread (Figure 2). However, the strong thinning effect leads to a low-density intensity surface and a compressed error scale, and we therefore don't observe much absolute MAE difference between the methods (see the Intensity MAE column in Figure 1 for thinning$= -4$).

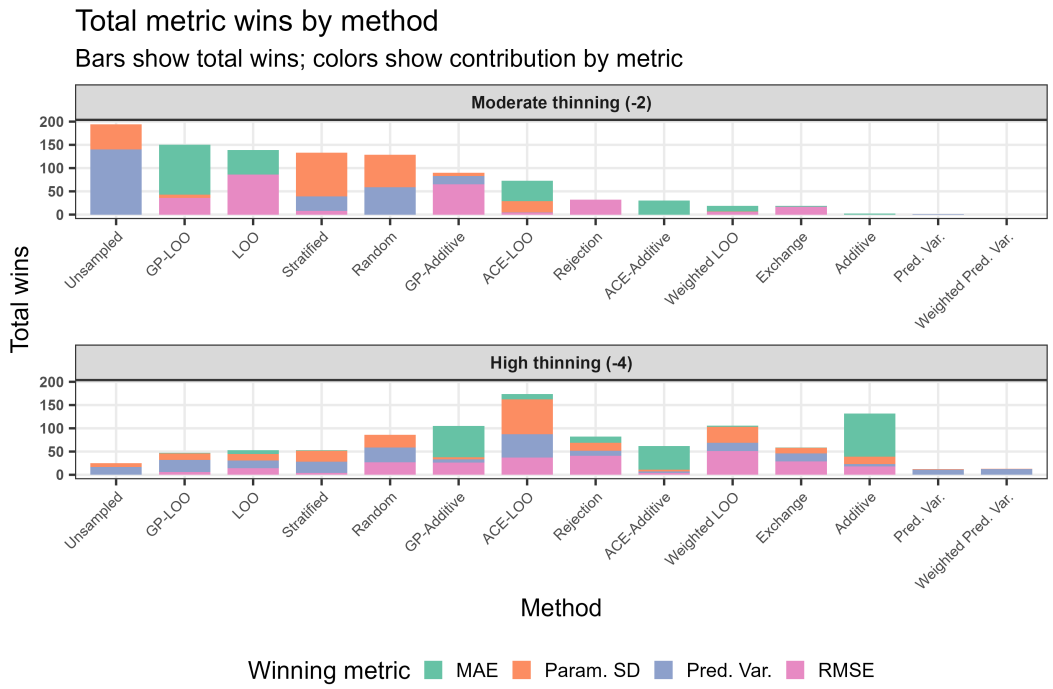


Figure 3: Cumulative performance ranking of design methods, per metric wins. Method order is set by the ranking in the moderate thinning setting. Each bar represents the frequency with which a method achieved the best (lowest) performance for a specific metric per each simulation runs and design size (m). The stacked segments illustrate the contribution of four performance dimensions: Intensity MAE, parameter RMSE, predictive variance, and average parameter SD. Higher total bar heights indicate design methods that consistently outperformed competitors across the evaluation criteria.

Finally, in the Supplementary material we include an evaluation of the sensitivity of the design methods to the number of PPD samples used in the approximation of the expected utility. The methods tested there showed robustness to the number of PPD samples used, and between 8-16 samples for most methods was sufficient.

4.2 Citizen science preference

We next repeat the analysis, now weighting the utilities by citizen science preference. In these runs we use only the thinning$= -2$ environment, and we include only a subset of the methods from the previous analysis: the Additive and LOO families, as well as the

predictive and weighted predictive variance approach. We do not include the design-based approaches because they do not use a utility function to select sites.

The primary takeaway is that the strength of observer preference or convenience can introduce unexpected bias. We can see this when weighting heavily towards the user's preference (high α), where there is a distinct difference between weighting for distance from current location and weighting for proximity from primary roads. Given Belgium's dense road network, surveyors are never far from main roads at the 10 km resolution we use in our study, and therefore weighting heavily towards road accessibility maintains some space-filling properties of the design. This means that even when heavily weighting towards roads, there is often not a dramatic change in performance of the method. Weighting by location behaves otherwise. As the user's location we use the city of Ghent, an area in which the covariates are homogeneous and the intensity for the simulated process is relatively low, compared to the hot spots in the Southeast of the country. Therefore, when weighting heavily ($\alpha = 0.8$) for location, we observe an increase in MAE, because the model is receiving less information from $\mathbf{y_1}$ about high intensity areas. Conversely, other metrics exhibit inconsistent, method-dependent responses to intense preference weighting. For example, LOO and Additive both exhibit decreases in parameter SD and predictive variance when $\alpha = 0.8$, relative to the other α levels; and the parameter RMSE doesn't dramatically change, indicating that by including the lower-intensity areas in the design, the model uncertainty has been reduced. In contrast, GP-LOO and GP-additive show increases in parameter RMSE for both distance and roads preference weighting at $\alpha = 0.8$. Furthermore, we don't see large changes in the predictive variance methods because the weighted utilities still strongly favor the border regions of the domain.

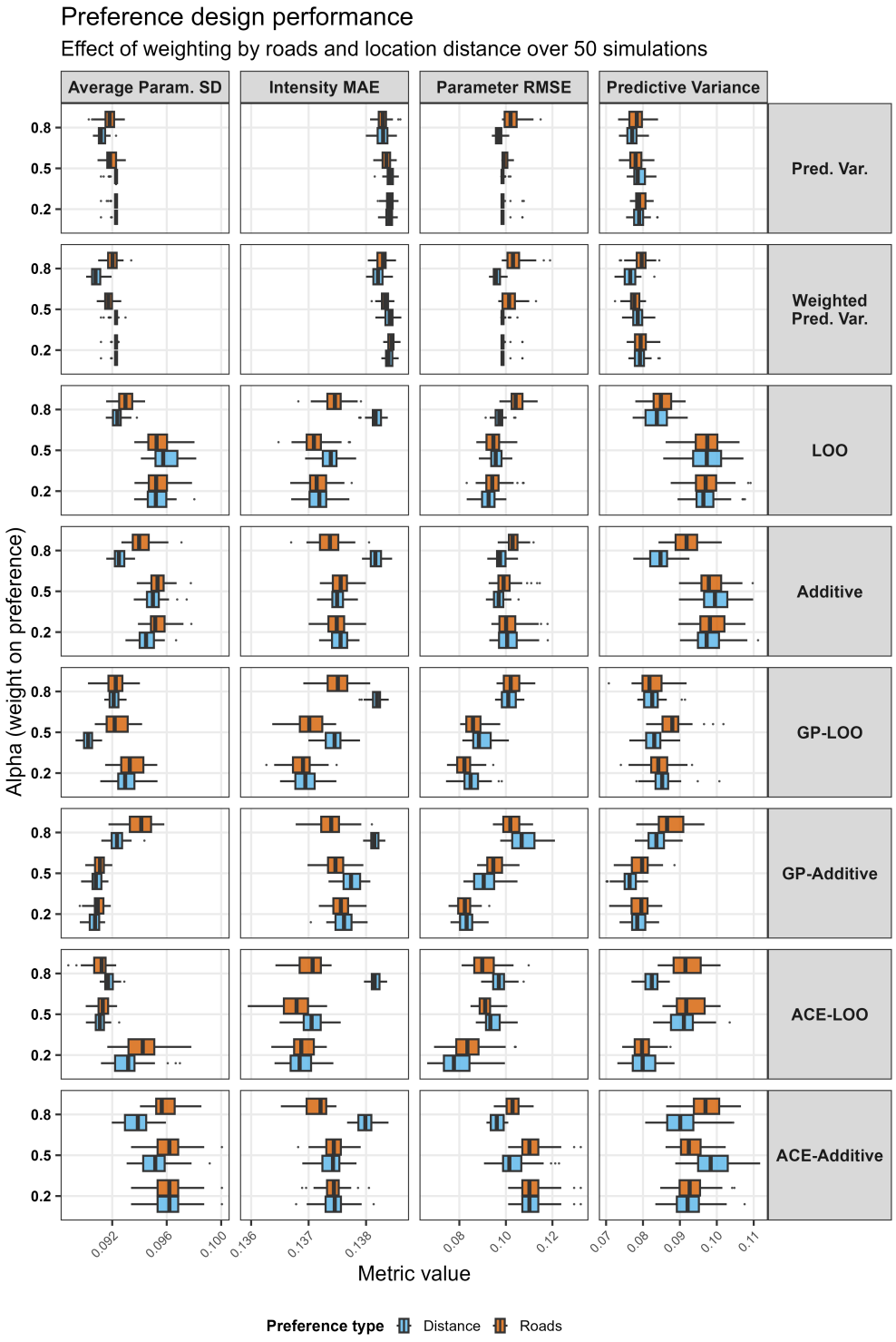


Figure 4: Comparison over methods, metrics, citizen science preference weighting. We only include runs for thinning= -2. The boxplots show the distribution of performance over 50 simulations of $\mathbf{y_1}$ data collection. Method type is shown on the right y axis; metric type is shown on the top x axis; preference weight (α) is shown on the left y axis; scores are shown on the bottom x axis. The method types are listed by model-based first and then exchange-based. User distance from Ghent (label: Distance) and distance from primary roads (label: Roads) used within the weighting function are shown in blue and orange respectively.

As in the analysis without preference weighting, the best method will depend on the optimization goal. Figure 5 shows the ranks of the methods by metric wins. ACE-LOO generally outperforms the other approaches, now in terms of MAE and parameter RMSE. GP-Additive and the Predictive Variance methods outperform the others in terms of parameter SD and predictive variance (which were previously led by the design-based approaches in the unweighted analysis). But because an increase in one metric can lead to a decrease in another, for a real-world application it is important to consider the costs and benefits associated with weighting in the context of the specific design method being used. The core message is then that weighting by preference can have a counterbalancing effect to the purely statistical optimization: Weighting can reduce uncertainty in the model or introduce overconfidence, at the same time driving a directional shift in estimation and prediction accuracy depending on the preferred location and the strength of the weights.

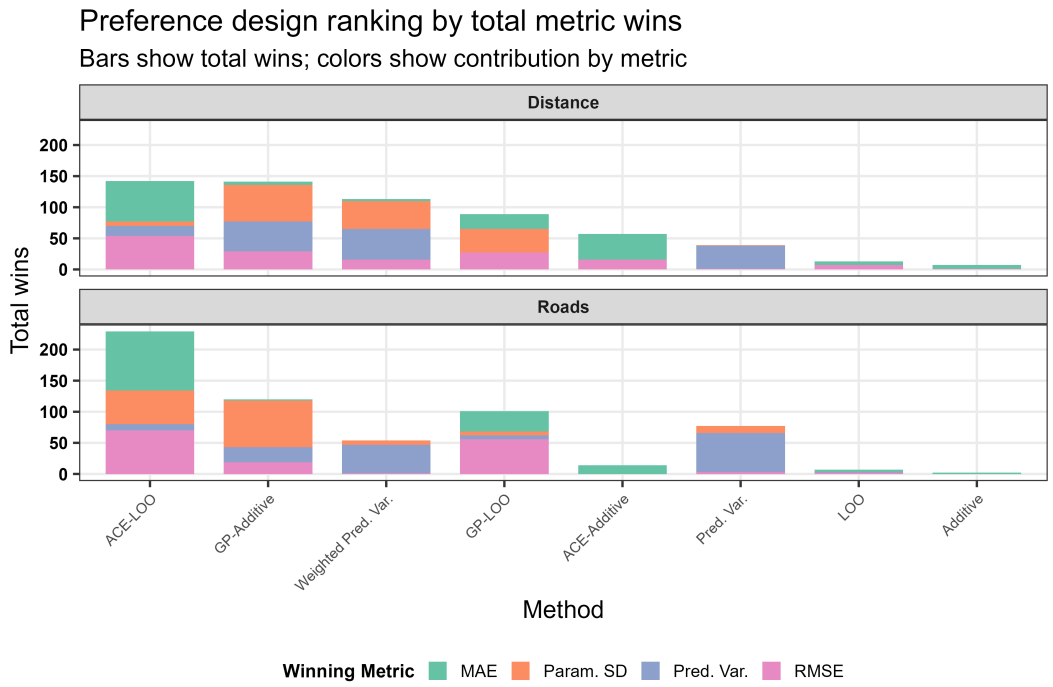


Figure 5: Cumulative performance ranking of design methods, per metric wins. Method order is set by the ranking in the distance from location (Distance) weighting setting. Each bar represents the frequency with which a method achieved the best (lowest) performance per metric across all simulation runs and all α weights. The stacked segments illustrate the contribution of the four performance dimensions: Intensity MAE, parameter RMSE, predictive variance, and average parameter SD. Higher total bar heights indicate design methods that consistently outperformed competitors across the evaluation criteria.

5 Discussion

We have performed a comprehensive evaluation and methodological benchmark of Bayesian adaptive design when applied to ecological data collection. We focused specifically on approaches for opportunistic data collection and the associated technical issues when developing adaptive designs. While we have focused on presence-only citizen science data, the framework we use is very flexible and can be applied to many other types of ecological and environmental data collection. Such data includes presence-absence or detection-nondetection data, capture-recapture data, animal movement data, phenology

data, or continuous data for environmental gradients such as air pollution or water monitoring networks. The issues we consider (design-based vs model-based, sample size, density of prior data, sensitivity of the expected utility approximation, weighting by preference) are general issues anywhere in adaptive sampling of ecological, environmental, and spatial data. We hope that this work encourages further statistical applications for site selection and Bayesian survey design. However, while the evaluation we have performed is extensive, there still remain quite a number of interesting avenues of research as well as a number of limitations to our work. We first offer five high-level key takeaways from this article and then address the limitations.

Takeaway 1: **Amount of prior data**. The design approach and the amount of data that has already been collected cannot be disentangled. Though this may appear obvious, it has broad implications for adaptive design. In situations where detection is low or the presence of species is rare, design-based methods can be considered generally safer but less optimal overall. In situations when strong prior knowledge regarding the species can be incorporated into the design, i.e., when previously collected data is abundant, there is more to be gained by using an model-based or algorithmic approach, though at some risk of variability of performance.

Takeaway 2: **Metric of optimization**. Project and design decisions are completely conditional on the quantity used for optimization. The more focused the design algorithm is on optimizing a specific quantity, the better it will perform when evaluated in terms of that quantity, at the risk of becoming overly specialized. From this perspective, the exact design method or algorithm is less important than the metric of optimization and evaluation, and the selection of these must match the goals of the project and the species

and region of interest.

Takeaway 3: **LOO for performance, design-based for stability**. We observed the LOO-family to be consistently well-performing across the evaluation dimensions, and the design-based methods to be generally more consistent across all experimental conditions (prior data availability, design size, and metric). Though we don't endorse the LOO-family as the "best strategy", these methods appear able to balance the tensions between the data and optimization process, even when weighting for citizen science preference. If the project scope is well defined statistically, then the extra computational effort appears worthwhile. Additionally, our results reinforce the idea often found in design literature that bespoke *hybrid* approaches are the most likely to fill specific organizational needs.

Takeaway 4: **The computational bottleneck**. The computational bottleneck of Bayesian design should not be intimidating and in fact we observed good stability of the approximation of the expected utility. We find the main complication of model-based design to be the development of the working model, as this will be used to simulate the future $\mathbf{y_1}$ datasets and must match the data generating process of $\mathbf{y_0}$. If the model specification poorly matches the data generating process of $\mathbf{y_0}$, it will likely lead to even worse conflicts between the two datasets than we observed during our experiments.

Takeaway 5: **Preference weighting**: If taking citizen science preference into account, consider what sort of unintended consequences this might have for the quantity of optimization and how this will interact with the design method of choice. For example, we observed that heavy preference weighting can have a regularizing effect on model parameters: Reducing parameter SD and predictive variance, while simultaneously increasing predictive error and parameter RMSE. When selecting design locations that

fulfill stakeholder requirements or qualitative organizational priorities, there will be implicit bias built in to the downstream data collection, even if that bias is not necessarily harmful in terms of model performance.

Having provided the high-level messages we seek to communicate with this work, we next discuss a few technical limitations. First, though our working model includes spatial correlation structure, in the model-based design optimization we consider each site independently, in order to avoid the computational complexity of exchange algorithms that compare utilities of m-size configurations. As m and the size of the design space increase (or the resolution of the designs increase), and the complexity of the model increases, there are few available scalable approaches. The work of Gray and Evangelou [2023] considers the site locations jointly, where the authors model the entire sampling design as a random variable with probability proportional to its utility, incorporating experimenter preference into the utility.

Moreover, efficient approximations to the expectation of the utility remains the key issue for Bayesian design. Specifically, efficient approaches to the outer loop bottleneck of the expected utility are required, such as via the Sequential Monte Carlo approach of Senarathne et al. [2020], though the application of their method to ecological survey design is not straightforward. Other sequential updating approaches, such as that of Leach et al. [2022], can also improve efficiency by avoiding the full recomputation of the posterior at every additional design point.

As mentioned previously, we observed an increase in model uncertainty as the design size increased in the thinning$= -2$ environment. This may reflect noise in the $\mathbf{y_0}$ dataset and the increasing contribution from $\mathbf{y_{1,d}}$ to to the updated data $(\mathbf{y_0}, \mathbf{y_{1,d}})$ as the design size

grows. More generally, simulation of $\tilde{\mathbf{y}}_\mathbf{1}$ is a standard step of Bayesian adaptive design when the expected utility is approximated by averaging over posterior predictive draws. As a result, uncertainty in the posterior distribution based on $\mathbf{y_0}$ propagates into the simulated $\tilde{\mathbf{y}}_\mathbf{1}$ values and, consequently, into the utility estimates. Care must therefore be taken to develop a reliable working model, including providing as much relevant prior information as possible, just as in any other application of Bayesian statistical analysis.

There are additional ecological limitations of our work. While Mondain-Monval et al. [2024] considers multiple species in their modelling, we use a single simulated species here for the purpose of focusing on the design behaviour itself, unconditional on the species of interest. Tailoring adaptive design approaches to specific species is essential to bring our work closer to ecological reality, as no method offers a one-size-fits-all solution. The effective information content of the prior data depends on species ecology, detectability, spatial structure, and sampling bias inherent to the species. Considering such factors in real-world data collection—in combination with the tensions between data and design methodology that we have explored here—remains a required area of study. Moreover, testing the methods in settings where occupancy probabilities over multiple collection periods have been estimated is essential, in order to be able to compare the design performance to baseline ecological knowledge.

Finally, to incorporate Bayesian adaptive design into real-world site selection and sampling projects, the design approaches need to be considered under the full scope of organizational and stakeholder objectives. In this paper we have focused primarily on the statistical aspects of the designs, but we welcome any future collaboration that works towards embedding Bayesian design methods within citizen science projects. We hope to achieve

the goal of real-time design described by Callaghan et al. [2021]: "If there are further developments of semi-automated or automated pipelines to interact with citizen scientists in near real time then the collective effort of citizen scientists will be able to reduce redundancies and gaps in the data collected." Integrating the lessons from our work into the Routine-Opportunistic Adaptive Monitoring framework of Pollock et al. [2026] appears to be a clear path forward. Bayesian adaptive design naturally fills the role of their inner loop of opportunistic burst sampling, and potentially can directly communicate with the long-term objectives of the outer monitoring loop. From this perspective, Bayesian adaptive design has the potential to support rapid biodiversity monitoring and environmental assessment under climate and anthropogenic change, while contributing to a more unified ecological framework that balances both statistical ecological priorities.

6 Acknowledgements

This paper received funding from the Flemish Government under the research program: Artificiele Intelligentie (AI) Vlaanderen. Additional thanks are given to the members of the Biostat group in the Faculty of Bioscience Engineering at Ghent University for their helpful conversations related to the statistical topics in this work.

7 Conflict of Interest statement

The authors declare no conflict of interest.

8 Author Contributions

MS and SL conceived the initial ideas for the project. MS implemented the design algorithms and associated code. MS managed the experiments and analyzed the outcomes, and SL provided guidance during this process; MS drafted and wrote the manuscript. SL provided feedback and editing of the manuscript drafts. All authors contributed critically to the drafts and gave final approval for publication.

9 Data Availability

All data used in this work can be found in the anonymous figshare repository at `https://figshare.com/s/61fe2afc676a2a725531`. Following publication of this article in a peer-reviewed journal, the figshare repository will be made public at DOI 10.6084/m9.figshare.31915401, along with a GitHub repository for public use of the code and data.

References

Tom August, Richard Fox, David B. Roy, and Michael J. O. Pocock. Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. *Scientific Reports*, 10(1):11009, July 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-67658-3.

Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian. Inlabru: An R

package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766, 2019. ISSN 2041-210X. doi: 10.1111/2041-210X.13168.

Katie Buchhorn, Kerrie Mengersen, Edgar Santos-Fernandez, Erin E Peterson, and James M McGree. Bayesian design with sampling windows for complex spatial processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(2):378–397, March 2024a. ISSN 0035-9254. doi: 10.1093/jrsssc/qlad099. URL `https://doi.org/10.1093/jrsssc/qlad099`.

Katie Buchhorn, Kerrie Mengersen, Edgar Santos-Fernandez, Erin E Peterson, and James M McGree. Bayesian design with sampling windows for complex spatial processes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(2):378–397, March 2024b. ISSN 0035-9254. doi: 10.1093/jrsssc/qlad099.

Corey T. Callaghan and Dale E. Gawlik. Efficacy of eBird data as an aid in conservation planning and monitoring. *Journal of Field Ornithology*, 86(4):298–304, 2015. ISSN 1557-9263. doi: 10.1111/jofo.12121.

Corey T. Callaghan, Alistair G. B. Poore, Richard E. Major, Jodi J. L. Rowley, and William K. Cornwell. Optimizing future biodiversity sampling by citizen scientists. *Proceedings of the Royal Society B: Biological Sciences*, 286(1912):20191487, October 2019. doi: 10.1098/rspb.2019.1487.

Corey T Callaghan, Alistair G B Poore, Thomas Mesaglio, Angela T Moles, Shinichi Nakagawa, Christopher Roberts, Jodi J L Rowley, Adriana VergÉs, John H Wilshire, and William K Cornwell. Three Frontiers for the Future of Biodiversity Research Using

Citizen Science Data. *BioScience*, 71(1):55–63, January 2021. ISSN 0006-3568. doi: 10.1093/biosci/biaa131.

Corey T. Callaghan, Diana E. Bowler, Shane A. Blowes, Jonathan M. Chase, Mitchell B. Lyons, and Henrique M. Pereira. Quantifying effort needed to estimate species diversity from citizen science data. *Ecosphere*, 13(4):e3966, 2022. ISSN 2150-8925. doi: 10.1002/ecs2.3966.

Corey T Callaghan, Maureen Thompson, Adam Woods, Alistair G B Poore, Diana E Bowler, Fabrice Samonte, Jodi J L Rowley, Nadiah Roslan, Richard T Kingsford, William K Cornwell, and Richard E Major. Experimental evidence that behavioral nudges in citizen science projects can improve biodiversity data. *BioScience*, 73(4): 302–313, April 2023. ISSN 0006-3568. doi: 10.1093/biosci/biad012.

Robert M. Dorazio. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, 23(12):1472–1484, 2014. ISSN 1466-8238. doi: 10.1111/geb.12216.

Ian Flint, Chung-Huey Wu, Roozbeh Valavi, Wan-Jyun Chen, and Te-En Lin. Maximising the informativeness of new records in spatial sampling design. *Methods in Ecology and Evolution*, 15(1):178–190, 2024. ISSN 2041-210X. doi: 10.1111/2041-210X.14260.

Elizabeth J Gray and Evangelos Evangelou. A design utility approach for preferentially sampled spatial data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 72(4):1041–1063, September 2023. ISSN 0035-9254, 1467-9876. doi: 10.1093/jrsssc/qlad040.

Peter A. Henrys, Thomas O. Mondain-Monval, and Susan G. Jarvis. Adaptive sampling in

ecology: Key challenges and future opportunities. *Methods in Ecology and Evolution*, 15 (9):1483–1496, 2024. ISSN 2041-210X. doi: 10.1111/2041-210X.14393.

Mevin B. Hooten, Johnson , Devin S., and Brian M. and Brost. Making Recursive Bayesian Inference Accessible. *The American Statistician*, 75(2):185–194, May 2021. ISSN 0003-1305. doi: 10.1080/00031305.2019.1665584.

Alison Johnston, Eleni Matechou, and Emily B. Dennis. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 14(1):103–116, 2023. ISSN 2041-210X. doi: 10.1111/2041-210X.13834.

Vira Koshkina, Yan Wang, Ascelin Gordon, Robert M. Dorazio, Matt White, and Lewi Stone. Integrated species distribution models: Combining presence-background data and site-occupancy data with imperfect detection. *Methods in Ecology and Evolution*, 8(4): 420–430, 2017. ISSN 2041-210X. doi: 10.1111/2041-210X.12738.

Clinton B. Leach, Perry J. Williams, Joseph M. Eisaguirre, Jamie N. Womble, Michael R. Bower, and Mevin B. Hooten. Recursive Bayesian computation facilitates adaptive optimal design in ecological studies. *Ecology*, 103(2):e03573, 2022. ISSN 1939-9170. doi: 10.1002/ecy.3573.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (4):423–498, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.00777.x.

Jia Liu and Jarno Vanhatalo. Bayesian model based spatiotemporal survey designs and

partially observed log Gaussian Cox process. *Spatial Statistics*, 35:100392, March 2020. ISSN 2211-6753. doi: 10.1016/j.spasta.2019.100392.

Quan Long, Marco Scavino, Raúl Tempone, and Suojin Wang. Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Computer Methods in Applied Mechanics and Engineering*, 259:24–39, June 2013. ISSN 0045-7825. doi: 10.1016/j.cma.2013.02.017. URL https://www.sciencedirect.com/science/article/pii/S0045782513000492.

Jorge Mateu and Werner G. Muller. *Spatio-Temporal Design: Advances in Efficient Data Acquisition*. John Wiley & Sons, Ltd, 2013.

Thomas Mondain-Monval, Michael Pocock, Simon Rolph, Tom August, Emma Wright, and Susan Jarvis. Adaptive sampling by citizen scientists improves species distribution model performance: A simulation study. *Methods in Ecology and Evolution*, 15(7):1206–1220, 2024. ISSN 2041-210X. doi: 10.1111/2041-210X.14355.

Antony M. Overstall and David C. Woods. Bayesian Design of Experiments Using Approximate Coordinate Exchange. *Technometrics*, 59(4):458–470, October 2017. ISSN 0040-1706. doi: 10.1080/00401706.2016.1251495.

Antony M. Overstall, James M. McGree, and Christopher C. Drovandi. An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Statistics and Computing*, 28(2):343–358, March 2018. ISSN 1573-1375. doi: 10.1007/s11222-017-9734-x. URL https://doi.org/10.1007/s11222-017-9734-x.

Michael J. O. Pocock, Mark Logie, Nick J. B. Isaac, Richard Fox, and Tom August. The recording behaviour of field-based citizen scientists and its impact on biodiversity trend

analysis. *Ecological Indicators*, 151:110276, July 2023. ISSN 1470-160X. doi: 10.1016/j.ecolind.2023.110276.

Laura J. Pollock, Pedro Henrique Pereira Braga, Christopher R. Florian, Katherine Hébert, Jenna Kline, R. Patrick Lyon, John T. Van Stan, Sara Beery, Michael E. Dillon, Diego Ellis Soto, Brooke Goodman, Niall Hanan, Marta A. Jarzyna, Justin Kitzes, Anke Kügler, Daniel Mosse, Yiluan Song, and Jeff Larkin. Putting the 'Adaptive' in Adaptive Monitoring: From Fast Data to Meaningful Ecological Change. *EcoEvoRxiv*, February 2026.

Henry T. Reich. Optimal sampling design and the accuracy of occupancy models. *Biometrics*, 76(3):1017–1027, 2020. ISSN 1541-0420. doi: 10.1111/biom.13203.

J.A Royle. Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, 100(2):121–134, February 2002. ISSN 03783758. doi: 10.1016/S0378-3758(01)00127-6.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00700.x.

Elizabeth G. Ryan, Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, 2016. ISSN 1751-5823. doi: 10.1111/insr.12107. URL

https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12107. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12107.

S. G. J. Senarathne, C. C. Drovandi, and J. M. McGree. A Laplace-based algorithm for Bayesian adaptive design. *Statistics and Computing*, 30(5):1183–1208, September 2020. ISSN 1573-1375. doi: 10.1007/s11222-020-09938-6.

S. G. J. Senarathne, Werner G. Müller, and James M. McGree. Bayesian design for minimizing prediction uncertainty in bivariate spatial responses with applications to air quality monitoring. *Biometrical Journal*, 65(4):2100386, 2023. ISSN 1521-4036. doi: 10.1002/bimj.202100386. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202100386. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.202100386.

Emily G. Simmonds, Susan G. Jarvis, Peter A. Henrys, Nick J. B. Isaac, and Robert B. O'Hara. Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, 43(10):1413–1422, 2020. ISSN 1600-0587. doi: 10.1111/ecog.05146.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1–28, February 2017. ISSN 0883-4237, 2168-8745. doi: 10.1214/16-STS576. URL https://projecteuclid.org/journals/statistical-science/volume-32/issue-1/Penalising-Model-Component-Complexity--A-Principled-Practical-Approach-to/10.1214/16-STS576.full. Publisher: Institute of Mathematical Statistics.

A. W. L. Pubudu Thilan, Erin Peterson, Patricia Menéndez, Julian Caley, Christopher Drovandi, Camille Mellin, and James McGree. Bayesian design methods for improving the effectiveness of ecosystem monitoring. *Environmental and Ecological Statistics*, 31(4): 893–919, December 2024. ISSN 1573-3009. doi: 10.1007/s10651-024-00623-9.

Awlp Thilan, P Menéndez, and Jm McGree. Assessing the ability of adaptive designs to capture trends in hard coral cover. *Environmetrics*, 34(6):e2802, 2023. ISSN 1099-095X. doi: 10.1002/env.2802.

Maureen M. Thompson, Katie Moon, Adam Woods, Jodi J. L. Rowley, Alistair G. B. Poore, Richard T. Kingsford, and Corey T. Callaghan. Citizen science participant motivations and behaviour: Implications for biodiversity data coverage. *Biological Conservation*, 282:110079, June 2023. ISSN 0006-3207. doi: 10.1016/j.biocon.2023.110079.

Perry J. Williams, Mevin B. Hooten, Jamie N. Womble, George G. Esslinger, and Michael R. Bower. Monitoring dynamic spatio-temporal ecological processes optimally. *Ecology*, 99(3):524–535, 2018. ISSN 1939-9170. doi: 10.1002/ecy.2120.

Elise F Zipkin, Erin R Zylstra, Alexander D Wright, Sarah P Saunders, Andrew O Finley, Michael C Dietze, Malcolm S Itter, and Morgan W Tingley. Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, 19(1):30–38, 2021. ISSN 1540-9309. doi: 10.1002/fee.2290.

Damaris Zurell, Uta Berger, Juliano S. Cabral, Florian Jeltsch, Christine N. Meynard, Tamara Münkemüller, Nana Nehrbass, Jörn Pagel, Björn Reineking, Boris Schröder, and

632 Volker Grimm. The virtual ecologist approach: Simulating data and observers. *Oikos*,
633 119(4):622–635, 2010. ISSN 1600-0706. doi: 10.1111/j.1600-0706.2009.18284.x.

Supplementary Material

S1 Virtual environment and data simulation

Study region

To evaluate the adaptive design framework, we construct a virtual environment using the geographic extent of Belgium as a case-study. The study domain D was defined using official Belgian administrative boundaries (Lambert 72 projection, EPSG:3137). To ensure a clean simulation window, we processed the national border by removing small slivers or internal exclaves within the Netherlands and Germany, as well as applying a small concave hull adjustment to smooth boundary irregularities. We also scaled the coordinate system from meters to kilometers to improve the numerical stability of the LGCP parameter estimation. The continuous region D was discretized into a regular grid of c cells, with a user-defined resolution (10 km in the experiments shown in the paper). Each cell is referred to as site A_i. This grid serves as the basis for covariate aggregation and the subsequent intensity surface calculation.

Covariate processing

We incorporated four primary environmental drivers to represent the true process underlying species distribution and sampling bias:

- Vegetation: We used the Enhanced Vegetation Index (EVI) MODIS product

(MCD12Q2 Version 6.1).

- Topography: Elevation data was obtained from the NASADEM Digital Elevation Model.

- Land cover: Categorical land cover data (CORINE Land Cover 2018) was processed to calculate the fractional cover of specific classes within each grid cell. We use fractional covariates of the classes "2 - Discontinuous urban fabric", "12 - Non-irrigated arable land", and "24 - Coniferous forest". For the purpose of our simulation study these classes serve as representative proxies for the primary environmental and anthropogenic drivers of species intensity within the study region.

- Accessibility features: To simulate accessibility bias, we calculated the Euclidean distance from the centroid of each grid cell to the nearest primary highway. The highway data was obtained from the Top10Vector (v7.0, 2023) dataset maintained by the National Geographic Institute of Belgium.

All covariates were standardized prior to simulation.

Point process simulation

The true latent species distribution is represented by a log-intensity surface $\log(\lambda(s))$. This data generating model is defined as a linear combination of environmental fixed effects and a latent spatial random field:

$$log(\lambda(s)) = \beta_0 + \sum_{k=1}^{K} \beta_k X_k(s) + \omega(s) \tag{S1}$$

where β_k are the true regression parameters for the k-th environmental covariate X_k. $\omega(s)$ is a Gaussian Random Field (GRF) with Matérn covariance structure simulated via the package gstat. The GRF represents latent environmental drivers or inherent properties of the population process that influence species distribution beyond the remote-sensing covariates. The Matérn covariance was configured with smoothness parameter $\nu = 1$, variance $\sigma^2 = 0.5$, and spatial decay parameter $\phi = 5$ km.

The resulting intensity surface was then sampled to produce the set of points $\mathbf{y_0} = \{s_1, s_2, ..., s_n\}$, representing the observed historical data upon which the adaptive sampling designs are optimized. The generation of $\mathbf{y_0}$ follows a two-step process. First, a realization of the Log-Gaussian Cox Process (LGCP) is generated based on the latent intensity $\lambda(s)$: $\mathbf{y}_{\text{init}} \sim$ Poisson Process$(\lambda(s))$ using the spatstat package. Next, the probability that a true individual at location s is observed, with probability $p(s_i)$, is constructed via a logit link function using bias-inducing covariates (e.g., proximity to roads): $\text{logit}(p(s_i)) = \gamma_0 + \sum_{j=1}^{J} \gamma_j B_j(s)$ where $B_j(s)$ represents the j-th bias covariate and γ_j is its associated effect. Each raw point is subjected to a Bernoulli trial with probability $p(s_i)$. Points that fail the trial are discarded (thinned), resulting in the final observed dataset $\mathbf{y_0}$. This two-stage data generation procedure decouples the biological process (fixed effects βs) from the sampling process (bias effects γ) in the data generating procedure. In this work we use the distance from primary highways in Belgium as the bias covariate, in addition to an intercept, meaning that as the distance increases, the probability of observation decreases when the parameter associated with the covariate is negative. In our simulations we use two different starting conditions for the bias: $\gamma_0 = -2$ and $\gamma_0 = -4$. The -4 setting generates a much higher rate of thinning ("high" thinning in

the main text), i.e., an overall lower baseline probability of making a positive observation. This could be attributed to a number of reasons: For example, a difficult to detect species, a rare species, or a species of less interest to volunteer observers. We include a map of this thinning probability in Supplemental figure S1, and the realizations of the data generating process in Supplemental figure S2.

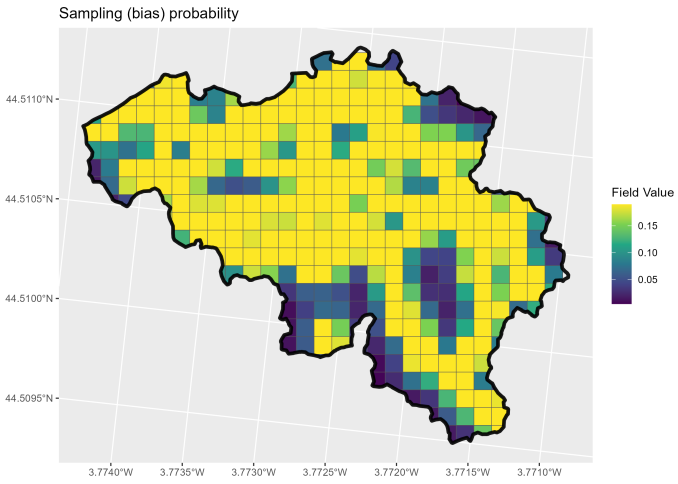
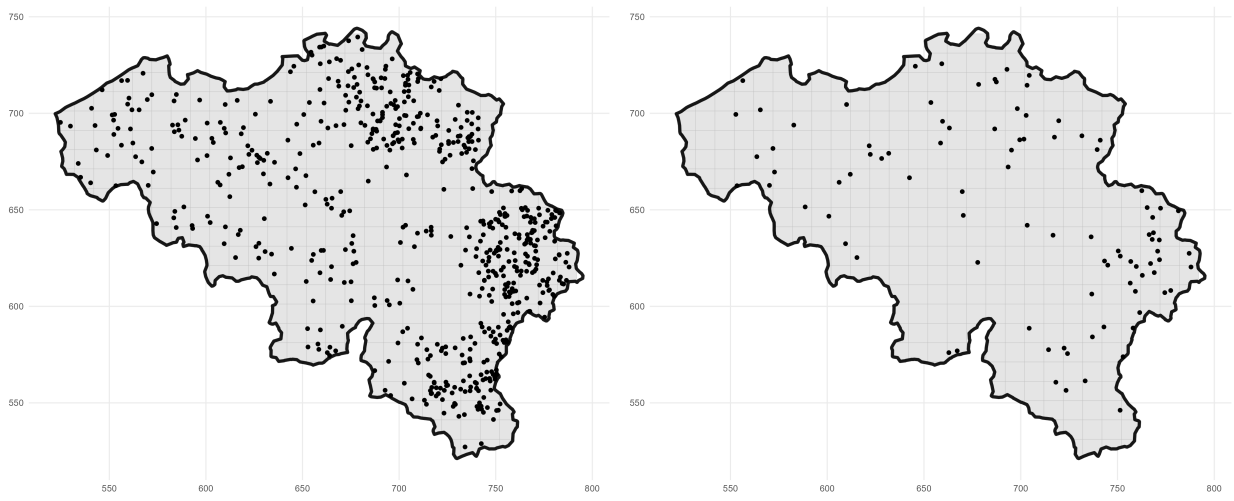


Figure S1: Probability of making a positive observation of the simulated species within Belgium, $\gamma_0 = -2$, based on the parameterized thinning process using the logit link. The overall probability is relatively low which reflects the inherent sparsity and incomplete coverage that is characteristic of opportunistic datasets.

S2 Modelling

To estimate the species distribution and recover the environmental parameters and latent spatial field, we apply the LGCP model of Supplemental equation S1. However, in the working model the probability of observation is parameterized via a log-linear relationship, intentionally introducing a source of misspecification. We fit the model using the INLA (Integrated Nested Laplace Approximation) and Stochastic Partial Differential Equation



(a) $\gamma_0 = -2$ (b) $\gamma_0 = -4$

Figure S2: Simulated $\mathbf{y_0}$ PO datasets in Belgium resulting from data generating process, after thinning has been performed using the parameterized thinning link. The black points represent locations at which an opportunistic observation has been collected. (a) shows the data from which fewer points have been thinned ($\gamma_0 = -2$ or "moderate" thinning), and (b) shows the data where the thinning effect is much stronger ($\gamma_0 = -4$ or "high" thinning).

(SPDE) approach [Rue et al., 2009, Lindgren et al., 2011]. For the Matérn covariance, we use Penalized Complexity (PC) priors [Simpson et al., 2017]: Spatial range: $P(\rho < 5 \text{ km}) = 0.01$. Marginal standard deviation: $P(\sigma > 0.5) = 0.01$. The data generating model uses the gstat range ϕ, which is related to the SPDE range by $\rho = \phi \cdot \sqrt{8\nu}$ [Lindgren et al., 2011]. The choice of range, $\phi = 5$, implies a source of misspecification between the data generating and working models. Because we work with aggregated 10 km resolution data, the recovery of fine-scale variation is limited, a common issue in ecological modelling. We induce this misspecification to mimic real-world data collection and modelling scenarios.

Inference was performed using the Laplace approximation strategy of INLA. We use the inlabru package for model fitting. Because the adaptive design requires fitting models iteratively or in parallel across many posterior predictive datasets, we optimized the

computation in the following ways: Integration strategy: We used the Empirical Bayes integration strategy to reduce computational overhead. Priors: The intercept was given a Gaussian prior with a mean of -3 and a precision of 1, reflecting our prior knowledge of the expected point density across the study area. For iterative model updates, we used the control.mode option in inlabru to restart the optimization from the previous model's configuration.

S3 Utility functions

Within the design optimization, there are a number of interesting utility and loss functions that can be considered. In this paper we use two general types: Kullback-Leibler (KL) divergence and predictive variance, though we compare weighted variants as well, described in the design algorithm section in the main text and in the section below. A more elaborate discussion of utility functions and their properties can be found in Ryan et al. [2016]. Interested readers may consider work such as Overstall and Woods [2017], Overstall et al. [2018], or Liu and Vanhatalo [2020] to see additional examples of the utilities used within Bayesian design.

For a design d, the KL utility function can be written as follows, for example when comparing posteriors resulting from $\mathbf{y_1}$ and $\mathbf{y_0}$:

$$u_{\mathrm{KL}}(\mathbf{y_1}, \mathbf{d}) = D_{KL}(p(\boldsymbol{\theta}|\mathbf{y_1}, \mathbf{y_0}, \mathbf{d})\| \, p(\boldsymbol{\theta}|\mathbf{y_0})) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbf{y}_1, \mathbf{y}_0, \mathbf{d}) \log \left\{ \frac{p(\boldsymbol{\theta} \mid \mathbf{y}_1, \mathbf{y}_0, \mathbf{d})}{p(\boldsymbol{\theta} \mid \mathbf{y}_0)} \right\} d\boldsymbol{\theta}. \tag{S2}$$

where, in our case, $\boldsymbol{\theta}$ denotes the full set of unknown model quantities including fixed

regression effects and latent spatial effects. $\mathbf{y_0}$ may also be conditional on a previous design $\mathbf{d_0}$ but we do not consider such prior collection structure here. When estimating the expected utility with MC integration, we must first approximate the posterior of the model and then the expectation of the utility. If using MCMC to estimate the posterior, this means we rely on a computationally expensive double approximation. To avoid the inner expensive approximation, in this work we use the Laplace approximation as implemented via the INLA software. The advantage of using the Laplace approximation, in addition to speed, is that when two multivariate normal distributions are compared, KL divergence can be expressed in the analytic form:

$$u_{\mathrm{KL}}(\mathbf{y_1}, \mathbf{d}) = \frac{1}{2}\left[-\ln\left(\frac{|\Sigma_{y_1}|}{|\Sigma_{y_0}|}\right) + \mathrm{tr}((\Sigma_{y_0})^{-1}\Sigma_{y_1}) + (\theta_{y_0} - \theta_{y_1})^T(\Sigma_{y_0}^{-1})(\theta_{y_0} - \theta_{y_1}) - k\right]. \quad \text{(S3)}$$

This is similar or the same as the design formulation as used in Long et al. [2013], Liu and Vanhatalo [2020], Thilan et al. [2024], Senarathne et al. [2023], Buchhorn et al. [2024a] and Overstall et al. [2018], all of whom use the normal distribution and normal form of the KL divergence to approximate the expected loss.

The expectation of the utility over the posterior predictive distribution will be

$$\mathrm{E}[U_{\mathrm{KL}}(\mathbf{d}) \mid \mathbf{y}_0] = \int_{\mathbf{y}_1} u_{\mathrm{KL}}(\mathbf{y}_1, \mathbf{d})\, p(\mathbf{y}_1 \mid \mathbf{y}_0, \mathbf{d})\, d\mathbf{y}_1. \quad \text{(S4)}$$

We also consider predictive variance in our design optimization. For each candidate site, we compute the expected posterior variance of the intensity, conditional on $\mathbf{y_1}$ and $\mathbf{y_0}$ and

report the average over all sites:

$$\mathrm{E}[\mathrm{Var}(\tilde{\lambda}(s_i)|\mathbf{y_1}, \mathbf{y_0})|\mathbf{y_0}] = \int_{\mathbf{y}_1} \mathrm{Var}(\tilde{\lambda}(s_i)|\mathbf{y_1}, \mathbf{y_0}) p(\mathbf{y_1}|\mathbf{y_0}) d\mathbf{y_1}. \tag{S5}$$

S4 Design methods

Here we provide more detailed descriptions of the design approaches we use in the paper.

Design-based

- Random + space-filling design: Randomly sample m sites. Then for each site, sample j new locations and check if any of these locations satisfy a minimum distance criteria so that no points are within 5 kilometers of the new site.
- Unsampled regions: Randomly sample m sites from locations with the fewest observations taken.
- Stratified sampling: Random sample m locations, stratified between sampled and unsampled regions.

Model-based

- Maximum predictive variance: Based on Supplemental equation S5, select the m sites that lead to the largest predictive variance, in order to target locations of high uncertainty for future sampling. This is a Bayesian version of the the "uncertainty only" method of Mondain-Monval et al. [2024]

- Weighted predictive variance: Weight the predictive variances by the inverse of the number of samples in the site. This prioritizes sites where there are few samples and the predictive variance is relatively high. This approach is also used by Mondain-Monval et al. [2024].

- Rejection sampling design: Sample sites at random and accept these candidates into the design with probability p. The design is grown until m sites have been added. We use a model-based quantity, $p = \min(p_{\max}, \frac{\tilde{\lambda}}{\max(\tilde{\lambda})})$. $p_{\max}$ allows the tuning of the design away from solely high-intensity areas. This is based on Liu and Vanhatalo [2020].

- Additive design: This method evaluates the marginal information gain of each candidate site A_i, $i \in 1, ...n$ by iteratively augmenting $\mathbf{y_0}$ with the observations $s \in A_i$ from the predictive dataset $\mathbf{y_1^{(r)}}$. The model is refit using this augmented $(\mathbf{y_0}, \mathbf{y_{1,i}})$ dataset, and the KL utility is computed between the original posterior and the updated posterior. The m sites that result in the highest KL divergence are selected. This is inspired by Flint et al. [2024] but adapted for Bayesian design.

- Leave-one-out (LOO): A removal-style method based on the predictive data $\mathbf{y_1}$. For each site A_i, $i \in 1, ...n$, a LOO dataset $\mathbf{y_{1-i}^{(r)}}$ is generated from the predictive dataset $\mathbf{y_1}^{(r)}$, where the observations $s \in A_i$ are removed. The KL divergence between the posterior of the full predictive dataset and the LOO dataset is computed and the m sites that result in the highest utility are selected. This can be considered as a sort of Bayesian analogue to the statistical leverage approach of Callaghan et al. [2019].

- Weighted LOO: As the above method disregards the importance of unsampled sites, this approach weights the LOO utility towards lower intensity areas. This weight is

based on the inverse of the predicted intensity with tuning parameter β: $w(s) = (1 + \tilde{\lambda}(s))^{-\beta}$ for location s. The product of the KL divergence and $w(s)$ is taken on a site-by-site basis. As β increases, the contribution of high-intensity sites is increasingly penalized.

- Utility emulation using Gaussian Process (GP): Based on the first phase of the ACE algorithm of Overstall and Woods [2017] and [Buchhorn et al., 2024a], this Bayesian optimization approach selects the m best sites in terms of utility predicted by a GP emulator fit to the utility surface. It requires first computing the full utility for a subset of sites and then fitting the GP to these utilities. The utility itself is flexible, and here we use both the additive and LOO versions.

Exchange-based

- Exchange algorithm [Royle, 2002]. For a given design $\mathbf{d}$, exchange each site A_i in the design with each candidate site A_i' in its local neighborhood. Update the design with the candidate if the exchange leads to improved utility. Converge if after a full iteration, no changes are made to the design. Within the exchange we use the additive version of the KL utility to compare candidate designs.

- ACE (Stochastic exchange): We use a stochastic exchange algorithm similar to Overstall and Woods [2017] and [Buchhorn et al., 2024a], where the utility surface is estimated using a Gaussian process and the design is then created by selecting the sites that best maximize the estimated utility. From this design the points are then exchanged and accepted with probability p as in Phase II of the Overstall and Woods [2017] algorithm. These designs are selected via the KL utility, using both the LOO

and additive strategies mentioned above.

S5 Design evaluation strategy

To evaluate each candidate design, we again use a simulation-based framework, in order to compare the designs against a known, ground truth intensity surface and known fixed effect parameters. Mirroring the data generating procedure described above, we first simulate a virtual ecological environment and then sample data from this environment according to the candidate designs. All evaluations were conducted using the same LGCP working model and the designs were compared under 50 repeated realizations of future sampling data.

We generate a known spatially varying log-intensity surface across the study domain, using the same combination of spatial covariates and a new realization of the spatial random field described above. We then simulated new realizations of hypothetical species observations based on that underlying intensity. These data were "collected" based on the candidate design, and the newly sampled points were appended to the baseline dataset $\mathbf{y_0}$ to create an updated dataset $(\mathbf{y_0}, \mathbf{y_{1,d}})$.

We then fit the LGCP to this updated dataset. To measure the performance of each design strategy, we extracted the posterior estimates from the fitted LGCP models and calculated several performance metrics against the known simulated truth. The metrics are described below:

- Parameter recovery: We evaluated how well the designs recovered the true underlying

relationships between the intensity and the spatial covariates. We calculated the Root Mean Squared Error (RMSE):

$$\text{RMSE}_\beta = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{\beta}_k - \beta_k^{\text{true}})^2},$$

where $\hat{\beta}_k$ is the posterior mean.

- Parameter precision: The posterior variance, $\text{Var}(\hat{\beta}_k)$, was extracted to assess estimation precision and reported as the average variance over all parameters.

- Predictive error: We assessed the model's ability to map the true species distribution by predicting the expected intensity λ across the spatial grid. The predictions were evaluated using the Mean Absolute Error (MAE):

$$\text{MAE}_\lambda = \frac{1}{C}\sum_{i=1}^{C}|\hat{\lambda}_i - \lambda_i^{\text{true}}|,$$

where C is the total number of grid cells, $\hat{\lambda}_i = \exp(\hat{\eta}_i + \sigma_i^2/2)$ is the expected intensity for a log-normal variable with predictive variance σ_i^2, λ_i^{true} is the ground truth intensity at cell i, and $\hat{\eta}_i$ is the log-linear predictor of the intensity.

- Predictive variance: To evaluate overall predictive confidence, we extracted the predictive variance $\hat{\sigma}_i^2$, averaged across all grid cells:

$$\overline{\sigma}^2_{\text{pred}} = \frac{1}{C}\sum_{i=1}^{C}\hat{\sigma}_i^2.$$

S6 Sensitivity analysis of Monte Carlo integration

In this section we show the sensitivity of the expected utility to the Monte Carlo approximation. The Monte Carlo approximation of the expected utility is

$$\mathrm{E}[U(\mathbf{d})|\mathbf{y_0}] \approx \frac{1}{R}\sum_{r=1}^{R} u(\boldsymbol{\theta}^{(\mathbf{r})}, \tilde{\mathbf{y}}^{(\mathbf{r})}, \mathbf{d}). \tag{S6}$$

We compare design results for increasing number of samples (increasing R), i.e., draws $\tilde{\mathbf{y}}_{\mathbf{1}}^{(\mathbf{r})}$ from the Posterior Predictive Distribution and respective posterior fits $\boldsymbol{\theta}^{(\boldsymbol{r})}$. We focus on only the LOO family, including also predictive and weighted predictive variance, and the exchange algorithm. From Supplemental figure S3, we can see that the performance of the designs begins to converge as R increases, particularly for the LOO family of methods. The predictive and weighted predictive variance methods remains constant across all values of R, as these designs target the boundaries of the design region and are not sensitive to the realizations of $\mathbf{y_1}$. The Exchange algorithm shows higher sensitivity in these runs

For the LOO family, at the lowest number of integration samples ($R = 1$ to $R = 4$) we observe higher variability across all metrics the Average Parameter SD. This suggests that using a single PPD draw provides an unstable approximation of the expected utility, leading the algorithm to select sub-optimal sites. As R increases, the most dramatic improvements in performance occur from $R = 4$ to $R = 16$. Beyond $R = 24$ the performance begins to plateau, and in general, using approximately 8 to 24 draws appears sufficient to capture the utility signal required for site selection.

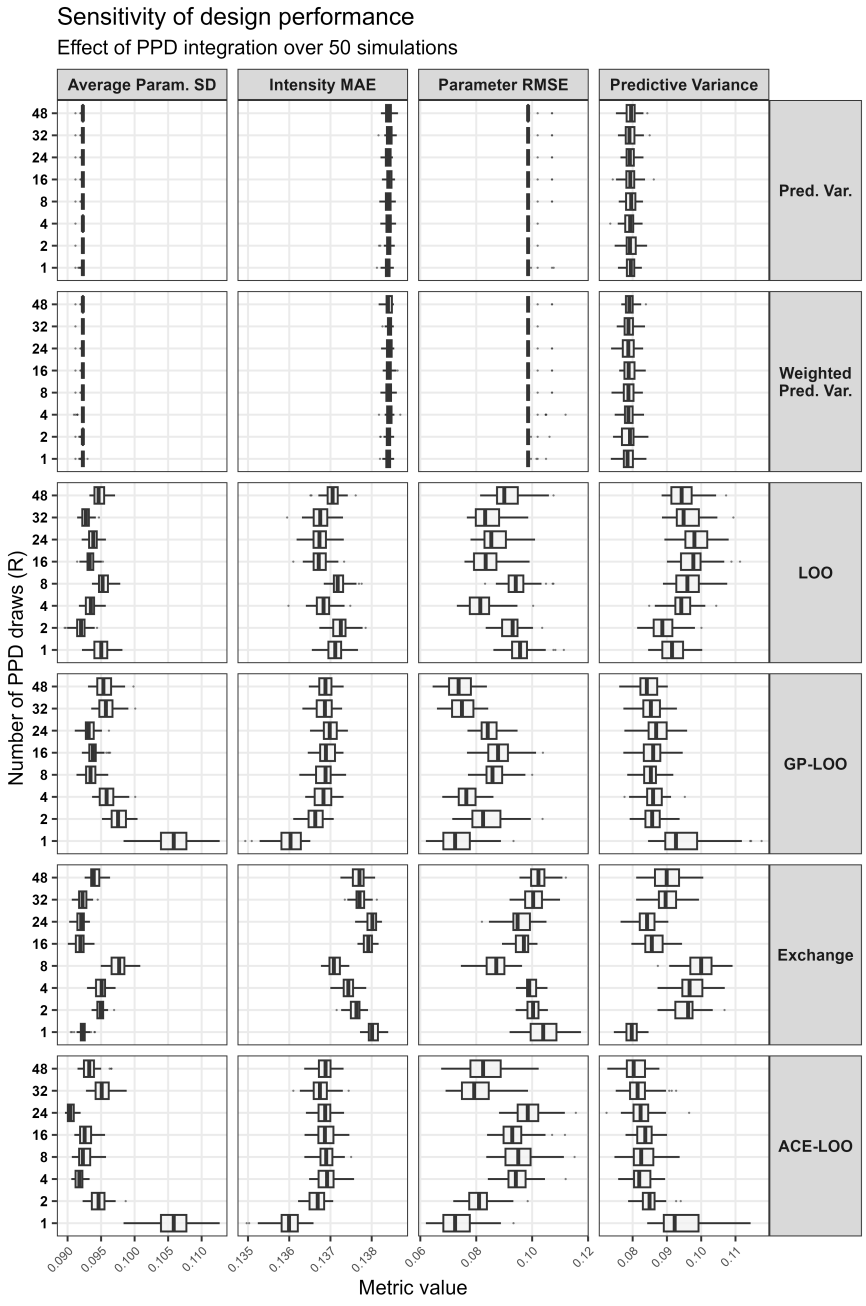


Figure S3: Sensitivity of design performance to the number of Monte Carlo draws (R) from the Posterior Predictive Distribution of $\mathbf{y_0}$. The boxplots show the distribution design performance under each metric: Average Parameter SD, Intensity MAE, Parameter RMSE, and Predictive Variance, across 50 simulation runs under the moderate thinning condition ($\gamma_0 = -2$). Rows show the design methods. The y-axis on the left denotes the number of PPD datasets (R) used to approximate the expected utility $U(\mathbf{d})$. Results indicate convergence in performance as R increases, with model- and exchange-based methods exhibiting higher volatility at fewer PPD draws, and the predictive variance methods showing stability across R.