NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

06-24

Spatial-Statistical Downscaling with Uncertainty Quantification in Biodiversity Modelling

Xiaotian Zheng, Noel Cressie, David A. Clarke,
Melodie A. McGeoch, and Andrew Zammit-Mangion

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia T: +61 2 42215076. E: karink@uow.edu.au

Spatial-statistical downscaling with uncertainty quantification in biodiversity modelling

Xiaotian Zheng[1,2], **Noel Cressie**[1,2], **David A. Clarke**[3,4], **Melodie A. McGeoch**[3,4], **and Andrew Zammit-Mangion**[1,2]

[1] *School of Mathematics and Applied Statistics, University of Wollongong, Australia;* [2] *Securing Antarctica's Environmental Future, University of Wollongong, Australia;* [3] *School of Biological Sciences, Monash University, Australia;* [4] *Securing Antarctica's Environmental Future, Monash University, Australia*

June 20, 2024

Summary

1. We provide a general framework for spatial modelling of biodiversity that involves downscaling environmental covariates. Accurate downscaling with uncertainty quantification and its inclusion in estimating biodiversity models are crucial for accurate, valid inferences and predictions.
2. We derive model-based statistical downscaling with uncertainty quantification that can be used for spatial modelling of species data and diversity metrics. We then propose a two-stage protocol that propagates downscaling uncertainty to generalised linear models (GLMs), commonly used in biodiversity modelling. The protocol is readily implemented with the aid of standard software packages.
3. Simulation studies show that our two-stage protocol provides accurate downscaling uncertainty in biodiversity models such as GLMs. We show that when used in these models, our protocol improves uncertainty quantification for inferences when compared to existing methods.
4. Extensions of the two-stage protocol that include accounting for measurement errors and missing values in the covariate data, spatial misalignment, non-Gaussian data, and fusing multi-source data, are discussed.

Key-words: change-of-support; errors-in-variables; Gaussian processes; prediction uncertainty; spatially dependent errors; species distribution models

1 Introduction

As biodiversity becomes increasingly threatened by multiple global-change drivers (e.g., climate change), measuring and estimating changes in biodiversity is critical for supporting policy decision about land-use and habitat conservation (e.g., Jaureguiberry et al. 2022). Such biodiversity information is typically delivered based on the outcome of (i) single or multi-species distribution models that use empirical records of species occurrence data (e.g., Ovaskainen and Soininen 2011; Warton et al. 2015; Poggiato et al. 2021) or (ii) spatially explicit estimates of compound metrics of biodiversity, such as community composition or turnover (e.g., Ferrier et al. 2007; Latombe et al. 2017; Hoskins et al. 2020). In such models, the underlying ecological process or pattern is typically defined as a function of environmental covariates in a regression-type relationship, which is estimated through fitting a model to species or diversity-metric data. The fitted function is then used to understand and predict how biodiversity changes with environmental change. (We use the term 'covariate' instead of 'predictor' to denote an explanatory variable in regression models, and we use the term 'predictor' for results obtained after applying fitted functions to the covariates.)

Valid inference based on biodiversity responses to environmental covariates requires the covariate data to be available at the same spatial resolution as the responses (Dungan et al., 2002). Consequences of the spatial-resolution mismatch have been highlighted in previous studies; see, for example, Miller et al. (2004) and McInerny and Purves (2011). In particular, for studies that involve fine-resolution biodiversity-response data (e.g., Jarzyna and Jetz 2018; Keil et al. 2018; Lu and Jetz 2023), the availability of similarly fine-resolution environmental data is often limited due to the impractical and costly effort of collecting *in situ* data, such as soil properties or temperature. Thus, current studies commonly derive fine-resolution environmental data by downscaling from coarse-resolution data available, for example, from numerical models, general circulation models (GCMs), and climate reanalyses, such as ERA5 (Hersbach et al., 2020).

Downscaling refers to any procedure that obtains data at a spatial resolution finer than the corresponding coarse-resolution data. Throughout this article, the term 'spatial resolution' refers to the basic areal units (i.e., the smallest spatial units) at which biodiversity inference is made. To focus on the general problem of the basic-areal-unit mismatch between the biodiversity response and covariate variables, we use 'spatial resolution' rather than 'grain' or 'scale' to distinguish our purpose from others, where terms such as 'intrinsic scale' (Wu and Li, 2006) and 'response grain' (Mertes and Jetz, 2018) address the idea that organisms most strongly respond to environments at a particular grain or scale (Chase and Knight, 2013).

In this article, we focus on developing statistical methods with uncertainty quantification for downscaling of environmental data to fine resolutions, and for their use, along with downscaling uncertainty, in biodiversity models. We focus on *statistical* downscaling, in contrast to another commonly used approach for downscaling climate variables, namely, *dynamical* downscaling, that typically refers to the use of regional climate models (RCMs). RCMs generate local climate processes that use GCM output as boundary conditions and physical principles involving the exchange of energy and water (e.g., see the review by Giorgi and Gutowski Jr 2015). On the other hand, statistical downscaling employs statistical models to produce fine-resolution data. In general, the validity of a statistical-downscaling method relies on the underlying statistical assumptions; see Wilby et al. (2004) for a discussion of different categories of statistical-downscaling methods and general guidelines for using them.

Our approach to statistical downscaling is 'stand-alone', in the sense that, for the variable of interest, only its coarse-resolution data are needed for downscaling. This differs from downscaling methods that use linear or non-linear models (e.g., Berrocal et al. 2012; Ahmed et al. 2015; Majumder et al. 2021) that require both fine-resolution and coarse-resolution data for the variable of interest. Thus, our downscaling approach is widely applicable, especially for biodiversity studies where *in situ* observations of environmental variables are often unavailable. Moreover, our approach explicitly deals with data types of differing resolutions which, in the spatial-statistics literature, is referred to as 'change-of-support' (Cressie 1993, Ch.3; Gotway and Young 2002; Ma et al. 2019). Changing spatial support creates a new variable at a different resolution that, although related to the original variable, has different statistical properties. Thus, not explicitly resolving the change in resolution when downscaling may lead to inaccurate results and conclusions.

Downscaling of a variable from coarse resolution to fine resolution introduces spatial details and heterogeneities, with possibly large uncertainty. The need for including downscaling uncertainty in biodiversity modelling has been increasingly recognised (e.g., Jetz et al. 2019; Jansen et al. 2022). However, many biodiversity studies tend to simply substitute downscaled covariate values into a model without accounting for the downscaling uncertainty (e.g., Hannah et al. 2014; Maclean 2020), hereafter referred to as the 'plug-in' approach. Ignoring downscaling uncertainty in covariates present in biodiversity models can lead to inaccurate estimation of regression coefficients and of variation in diversity, which may adversely affect biodiversity prediction and result in incorrect quantification of the prediction uncertainty. Accurate prediction of biodiversity changes and quantification of uncertainty of those predictions are crucial for conservation and policy decision-making (e.g., Jetz et al. 2019).

In this article, we provide a two-stage protocol for biodiversity modelling in the presence of downscaling uncertainty in covariates, in a form that is straightforward to implement with standard statistical software. In Stage 1, the protocol produces many fine-resolution downscaled Monte Carlo (MC) samples of the covariate via conditional simulation, conditional on the known coarse-resolution data; then, in Stage 2, the protocol fits the biodiversity model to each of the downscaled MC samples. The result is a collection of fits, predictions, and prediction uncertainties, where each one corresponds to one of the downscaled MC samples.

In Stage 2 of our protocol, we account for downscaling uncertainty in covariates by using an *errors-in-variables* model. Generally, 'errors-in-variables' refers to problems where covariate data are obtained, collected, or measured imperfectly (Carroll et al., 2006). In biodiversity modelling, accounting for errors in covariates has been explored under one of two types of errors-in-variables models (see Section 3.2): the classical-error model (e.g., Elston et al. 1997; Denham et al. 2011; Stoklosa et al. 2015) and the Berkson-error model (e.g., McInerny and Purves 2011; Mourguiart et al. 2024). In contrast to existing approaches that assume independent errors for covariates, our protocol considers a spatially-correlated downscaled field, which models a natural environment that is structured and heterogeneous in space. Based on a Berkson-error model, our methodology is focused on spatial prediction of covariates that are not obtained at the same resolution and/or the same locations of the response data (e.g., Foster et al. 2012). In Section 4.2, we show through a simulation study that, when compared to the 'plug-in' approach and an 'ensemble' approach our Berkson-error-model approach improves uncertainty quantification for biodiversity inferences. Note here that 'ensemble' refers to the use of downscaled MC samples, different to 'ensemble species distribution modelling' (e.g., Hao et al. 2020).

In the remainder of the article, we present the two-stage protocol and assess it using simulation studies. We then use bryophyte (moss) occurrence data from the Bunger Hills in East Antarctica to illustrate the application of the two-stage protocol.

2 Methods: Spatial statistical downscaling for ecological data

To lighten the technical presentations in Sections 2 and 3, we use a 'square-bracket' notation throughout to denote either different types of probability distributions (e.g., normal distribution, $N(\mu, \sigma^2)$) or their probability densities: For generic random quantities V_1 and V_2, we use $[V_1, V_2]$ to denote the joint distribution or joint density of (V_1, V_2), depending on the context. Similarly, let $[V_1]$ and $[V_1 \,|\, V_2]$ denote the marginal distribution (or marginal density) of V_1 and the conditional

distribution (or conditional density) of V_1 given V_2, respectively. Other notation used in this section is summarised in Table 1.

2.1 Spatial statistical models on Basic Areal Units

Numerical-model output, for example, weather or climate variables on regular grid cells, generally has a coarse resolution. To downscale this output to a (fine) resolution of interest, we introduce spatial-statistical models within a geostatistical framework (e.g., Cressie 1993, Part I) defined over a (fine) discretised spatial domain of interest. For illustrate purposes, the spatial domain is $D \subset \mathbb{R}^2$ (where $\mathbb{R}^2$ is the Euclidean plane), which consists of N_D non-overlapping areal units $\{U_i\}$; that is, $D = \cup\{U_i \subset \mathbb{R}^2 : i = 1, \ldots, N_D\}$ with associated centroids $\mathcal{D} \equiv \{\boldsymbol{s}_i : i = 1, \ldots, N_D\} \subset D$, where $\boldsymbol{s}_i$ is the centroid of U_i, for $i = 1, \ldots, N_D$. We refer to the units $\{U_i\}$ as Basic Areal Units (BAUs; e.g., Nguyen et al. 2012).

In our downscaling framework, the BAUs define the resolution to which we downscale the coarser-resolution data. The shapes and/or sizes of BAUs might be influenced by the granularity of the response data (e.g., species occurrence records). In our spatial models, BAUs represent the finest resolution of the latent process (i.e., the resolution where biodiversity inference is made). Different types of biodiversity-response variables may correspond to different field-sampling methods that involve different spatial resolutions, which the underlying fine resolution of BAUs allows. For the case study in Section 4.3, we define the BAUs as equal-size $1 \times 1\,\text{km}^2$ grid cells, corresponding to the data-sampling units in the study's systematic survey design (Leishman et al., 2020).

Ecological sampling methods come in various forms, such as plant sampling plots, geographic grid cells in large collections of biodiversity occurrence data, and line transects, among others. The illustration in Fig. 1(a) shows how BAUs might be specified along line transects, but the approach is flexible for the broad range of biodiversity sampling protocols. In Fig. 1(a), the spatial domain (left sub-panel) is made up of non-overlapping, rectangular BAUs (central sub-panel) of size $l \times d$. The length l might be chosen for practical reasons, such as the travel distance in a fixed time interval between field-sampling locations (centroids of the BAUs) along a line transect. In this hypothetical sampling design, we set the width d of the rectangular BAU equal to, for example, the width of the line transect (red dashed lines in central sub-panel). Our goal is to downscale the data from a coarse resolution (right sub-panel) to the fine-resolution BAUs (central sub-panel) that match the resolution of the biodiversity-response data (Stage 1), and then to fit an interpretable biodiversity model (Stage 2).

Table 1: Notation used in our statistical-downscaling framework.

Spatial domain, Basic Areal Units (BAUs), and blocks	
$\boldsymbol{s}_i$	The i-th BAU centroid, $i = 1, \ldots, N_D$, in $\mathcal{D}$
B	Generic notation for a coarse-resolution blockin D
B_j	The j-th coarse-resolution block, $j = 1, \ldots, M_D$, where $\cup\{B_j : j = 1, \ldots, M_D\} = D$
$\mathcal{B}$	Generic notation, $\mathcal{B} \equiv B \cap \mathcal{D}$, for the set of centroids of BAUs that are in B
D	$D = \cup\{U_i \subset \mathbb{R}^2 : i = 1, \ldots, N_D\}$, a spatial domain made up of BAUs $\{U_i\}$
$\mathcal{D}$	$\mathcal{D} \equiv \{\boldsymbol{s}_i : i = 1, \ldots, N_D\}$, set of centroids of the BAUs $\{U_i\}$
M_D	Number of coarse-resolution blocks in D
N_D	Number of BAUs in D
U_i	The i-th BAU, for $i = 1, \ldots, N_D$
Data, spatial process, and parameters	
$\boldsymbol{C}_{\boldsymbol{\phi}}$	$N_D \times N_D$ covariance matrix where the (i, j)-th entry is $C(\boldsymbol{s}_i, \boldsymbol{s}_j \,\vert \boldsymbol{\phi})$
$X(\cdot)$	$X(\cdot) \equiv \{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$, the BAU-level Gaussian process (GPB) defined on $\mathcal{D}$
$\boldsymbol{X}_B$	$\boldsymbol{X}_B \equiv (X(B_1), \ldots, X(B_{M_D}))^\top$, the column vector of coarse-resolution data
$\boldsymbol{X}_{\boldsymbol{s}}$	$\boldsymbol{X}_{\boldsymbol{s}} \equiv (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_{N_D}))^\top$, the column vector of fine-resolution, BAU-level variables
$\boldsymbol{R}$	$N_D \times p$ matrix, where the i-th row of $\boldsymbol{R}$ is $(r_1(\boldsymbol{s}_i), \ldots, r_p(\boldsymbol{s}_i))$
$\boldsymbol{W}_B$	$M_D \times N_D$ matrix that aggregates $\boldsymbol{X}_{\boldsymbol{s}}$ to $\boldsymbol{X}_B$: $\boldsymbol{X}_B = \boldsymbol{W}_B \boldsymbol{X}_{\boldsymbol{s}}$
$\boldsymbol{x}_{\boldsymbol{s}}^{(k)}$	The k-th MC sample of $\boldsymbol{X}_{\boldsymbol{s}}$ given data $\boldsymbol{X}_B$, $k = 1, \ldots, K$
$x^{(k)}(\boldsymbol{s}_i)$	The i-th element of $\boldsymbol{x}_{\boldsymbol{s}}^{(k)}$ (i.e., the k-th MC sample on BAU U_i), $i = 1, \ldots, N_D$
$\mu(\cdot)$	$\mu(\cdot) \equiv \{\mu(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$, large-scale deterministic variability of $X(\cdot)$
$\omega(\cdot)$	$\omega(\cdot) \equiv \{\omega(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$, mean-zero Gaussian process (GP) defined on $\mathcal{D}$
$\{\gamma_0, \boldsymbol{\gamma}\}$	Intercept and regression coefficients of the BAU-level covariates
$\boldsymbol{\phi}$	Parameters of the BAU-level covariance function of $\omega(\cdot)$

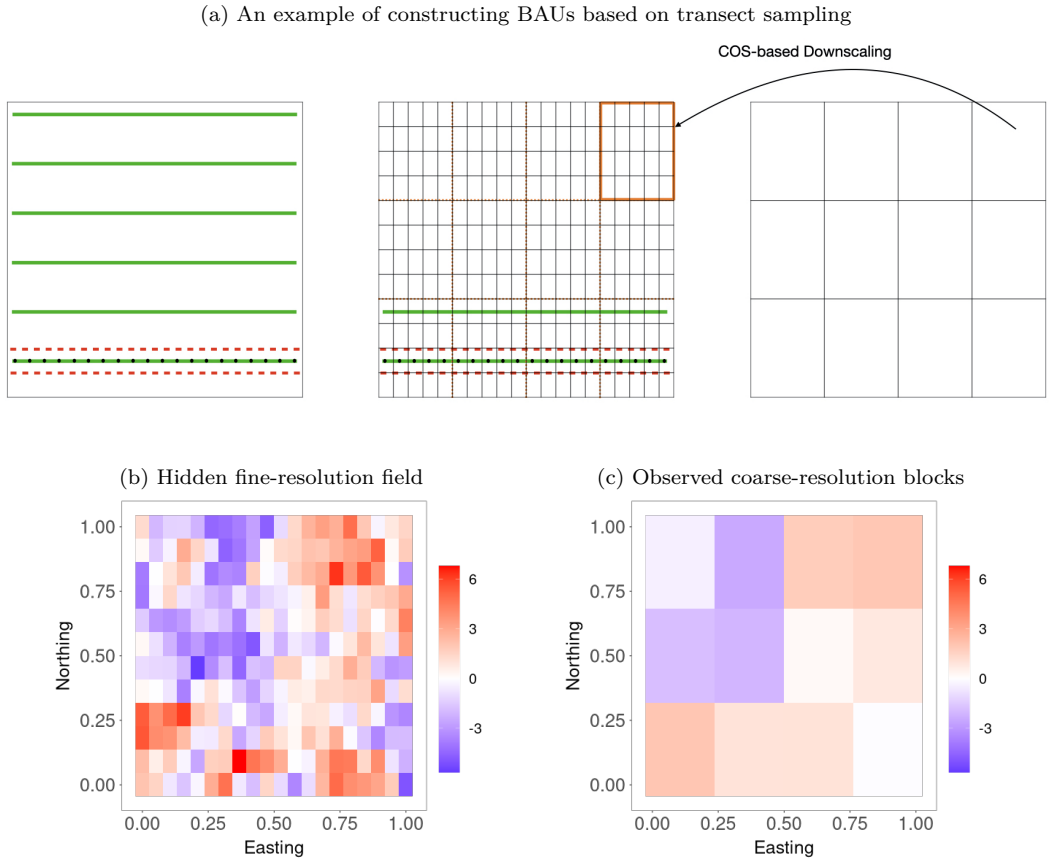


Figure 1: Specification of BAUs: An example of constructing BAUs from a transect sampling design (a): (left) line transects (green solid line), sampling points (black dots), and the sampling region (within red dashed lines); (middle) BAUs constructed along the line transects; (right) a coarse resolution at which the covariate data are available; simulated BAU-resolution field (b); coarse-resolution data (c) aggregated from the data shown in (b).

Key to our statistical-downscaling method is the assumption that the fine-resolution data are generated from a latent (hidden), univariate spatial process $X(\cdot) \equiv \{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ defined on the centroids of the BAUs, where for $\boldsymbol{s} \in \mathcal{D}$, $-\infty < X(\boldsymbol{s}) < \infty$. Consider the following model for $X(\cdot)$:

$$X(\cdot) = \mu(\cdot) + \omega(\cdot), \qquad \text{eqn 1}$$

where $\mu(\cdot) \equiv \{\mu(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ describes deterministic (i.e., non-stochastic) variability, and $\omega(\cdot) \equiv \{\omega(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is a mean-zero Gaussian process (GP) that captures short-range variability.

Coarse-resolution data are treated as an aggregation of the process $X(\cdot)$. In the synthetic example given in Fig. 1, the process $X(\cdot)$ corresponds to the fine-resolution spatial field (Fig. 1(b)) defined on the $12 \times 20 = 240$ BAUs. The coarse-resolution data are available for the much larger blocks, where each block is the union of 20 BAUs (Fig. 1(c)). Formally, let $\{B_1, \ldots, B_{M_D}\}$ be the set of M_D non-overlapping blocks, and D is the union of these blocks. Take $B \subset D$ as the

generic notation for a block. We define the coarse-resolution data $X(B)$ as the average of $X(\cdot)$ over $\mathcal{B} \equiv B \cap \mathcal{D}$; that is,

$$X(B) \equiv |\mathcal{B}|^{-1} \sum_{\boldsymbol{s} \in \mathcal{B}} X(\boldsymbol{s}), \qquad \text{eqn 2}$$

where the set $\mathcal{B}$ contains centroids of BAUs that are in B; $|\mathcal{B}| \equiv \sum_{\boldsymbol{s} \in \mathcal{D}} \mathbb{1}(\boldsymbol{s} \in \mathcal{B})$ is the number of BAUs whose centroids are in block B; and $\mathbb{1}(A)$ an indicator that equals 1 if the statement A is true and equals 0 otherwise.

In equation 2, the spatial support of $X(B)$ refers to the shape, orientation, and size of B. For $\boldsymbol{s} \in \mathcal{D}$, the spatial support of $X(\boldsymbol{s})$ refers to the BAU whose centroid is located at $\boldsymbol{s}$. Equation 2 elucidates the relationship between the spatial support and the properties of the process evaluated over that support. For example, the expectation of $X(B)$ and the covariance between $X(B)$ and $X(B')$, for $B, B' \subset D$, are

$$\begin{aligned} E(X(B)) &\equiv |\mathcal{B}|^{-1} \sum_{\boldsymbol{s} \in \mathcal{B}} E(X(\boldsymbol{s})), \\ \text{cov}(X(B), X(B')) &\equiv (|\mathcal{B}||\mathcal{B}'|)^{-1} \sum_{\boldsymbol{s} \in \mathcal{B}} \sum_{\boldsymbol{s}' \in \mathcal{B}'} \text{cov}(X(\boldsymbol{s}), X(\boldsymbol{s}')), \end{aligned} \qquad \text{eqn 3}$$

where $\mathcal{B}' \equiv B' \cap \mathcal{D}$ contains the centroids of the BAUs that are in B'. Equation 3 reveals the importance of accounting for the change-of-support (COS) from BAUs to blocks B to avoid the 'ecological fallacy' (Cressie, 1996; Gotway and Young, 2002).

Statistical downscaling uses statistical methods to infer the BAU-level components $\{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{B}\}$ of equation 2, given coarse-resolution data $X(B)$. Hereafter, we refer to $X(\cdot)$ in equation 1 as the *GPB* model to emphasise that the GP is defined over BAUs. Finally, we remark that the aggregation of $X(\cdot)$ defined in equation 2 assumes perfect alignment between the coarse-resolution blocks and BAUs, which can be a consideration when defining the BAUs. In Supporting Information S6, we illustrate a scenario of misalignment and show an extension of equation 2 to accommodate this; see also our discussion in Section 5.

2.2 Spatial-statistical downscaling procedures (Stage 1 of protocol)

We divide our statistical downscaling framework into three steps: (i) define the fine-resolution BAUs and specify the GPB $X(\cdot)$; (ii) fitting the GPB model using coarse-resolution data $\boldsymbol{X}_B$; (iii) generate samples from the predictive distribution of $X(\cdot)$ via conditional simulations.

Step 1: Specification of the BAUs and the GPB

Given a spatial domain of interest, the first step is to define the BAUs upon which a process $X(\cdot)$ is specified according to equation 1. There are two spatial components of variation in $X(\cdot)$, the first being the deterministic variation $\mu(\cdot)$. We initially consider a linear model, $\mu(\boldsymbol{s}) = \gamma_0 + \boldsymbol{\gamma}^\top \boldsymbol{r}(\boldsymbol{s})$, where γ_0 is an intercept; $\boldsymbol{\gamma} \equiv (\gamma_1, \ldots, \gamma_p)^\top$ is a p-dimensional column vector of regression coefficients; '$\top$' denotes vector transpose; and $\boldsymbol{r}(\boldsymbol{s}) \equiv (r_1(\boldsymbol{s}), \ldots, r_p(\boldsymbol{s}))^\top$ is a vector of BAU-level covariates for $X(\cdot)$, assumed available at all $\boldsymbol{s} \in \mathcal{D}$ (e.g., topographic data). We note that BAU-level covariates are not essential for downscaling; one could simply set $\mu(\boldsymbol{s}) = \gamma_0$ for all $\boldsymbol{s} \in \mathcal{D}$ and proceed with the subsequent steps.

We consider a generic GP model for the stochastic component $\omega(\cdot)$, with mean zero and covariance function $C(\boldsymbol{s}, \boldsymbol{s}' \,|\, \boldsymbol{\phi}) \equiv \text{cov}(\omega(\boldsymbol{s}), \omega(\boldsymbol{s}'))$, for $\boldsymbol{s}, \boldsymbol{s}' \in \mathcal{D}$, where $\boldsymbol{\phi}$ is a vector of unknown parameters. As an example, consider the commonly-used exponential covariance function, $C(\boldsymbol{s}, \boldsymbol{s}' \,|\, \boldsymbol{\phi}) \equiv \sigma^2 \exp(-||\boldsymbol{s} - \boldsymbol{s}'||/\psi)$, where $||\boldsymbol{s} - \boldsymbol{s}'||$ is the distance (e.g., Euclidean distance) between $\boldsymbol{s}$ and $\boldsymbol{s}'$, and the parameters are $\boldsymbol{\phi} \equiv (\psi, \sigma^2)^\top$ with ψ the spatial-range parameter and σ^2 the variance parameter; see Chapter 2 in Cressie (1993) for other spatial covariance functions.

Let $\boldsymbol{X}_{\boldsymbol{s}} \equiv (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_{N_D}))^\top$ be the vector that defines the BAU-level process. According to equation 1, $\boldsymbol{X}_{\boldsymbol{s}}$ follows a multivariate normal distribution, denoted as

$$[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}] = \text{N}(\mathbf{1}_{N_D}\gamma_0 + \boldsymbol{R}\boldsymbol{\gamma}, \boldsymbol{C}_{\boldsymbol{\phi}}), \qquad \text{eqn 4}$$

where $\mathbf{1}_{N_D}$ is the N_D-dimensional vector of ones; $\boldsymbol{R}$ is an $N_D \times p$ matrix such that the i-th column of $\boldsymbol{R}^\top$ is $\boldsymbol{r}(\boldsymbol{s}_i)$; and $\boldsymbol{C}_{\boldsymbol{\phi}}$ is the $N_D \times N_D$ covariance matrix, such that the (i,j)-th entry $C(\boldsymbol{s}_i, \boldsymbol{s}_j \,|\, \boldsymbol{\phi})$ describes the covariance between $X(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_j)$. If we consider the exponential covariance function defined above with parameters $\boldsymbol{\phi} \equiv (\psi, \sigma^2)^\top$, equation 4 implies that $[X(\boldsymbol{s}_i) \,|\, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}] = \text{N}(\gamma_0 + \boldsymbol{\gamma}^\top \boldsymbol{r}(\boldsymbol{s}_i), \sigma^2)$, for $i = 1, \ldots, N_D$.

Step 2: Fitting the GPB model using coarse-resolution data

Let $\boldsymbol{X}_B \equiv (X(B_1), \ldots, X(B_{M_D}))^\top$ be the M_D-dimensional vector of block values defined by equation 2. Define the aggregation matrix $\boldsymbol{W}_B$ such that $\boldsymbol{X}_B = \boldsymbol{W}_B \boldsymbol{X}_{\boldsymbol{s}}$, where the (j,i)-th entry of $\boldsymbol{W}_B$ is $w_{ji} = \mathbb{1}(\boldsymbol{s}_i \in \mathcal{B}_j) / \sum_{\boldsymbol{s} \in \mathcal{D}} \mathbb{1}(\boldsymbol{s} \in \mathcal{B}_j)$, for $j = 1, \ldots, M_D$, and $i = 1, \ldots, N_D$. Since aggregation preserves the normal distribution, from equation 4 we have that

$$[\boldsymbol{X}_B \,|\, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}] = \text{N}(\boldsymbol{W}_B(\mathbf{1}_{N_D}\gamma_0 + \boldsymbol{R}\boldsymbol{\gamma}), \boldsymbol{W}_B \boldsymbol{C}_{\boldsymbol{\phi}} \boldsymbol{W}_B^\top), \qquad \text{eqn 5}$$

which defines the likelihood function for parameters $\{\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$. These parameters are specified for

the process $X(\cdot)$ at the BAU level, but their likelihood is defined in terms of the coarse-resolution data $\boldsymbol{X}_B$.

Statistical downscaling predicts the unknown BAU-level vector $\boldsymbol{X}_{\boldsymbol{s}}$ given $\boldsymbol{X}_B$ and, when predicting $\boldsymbol{X}_{\boldsymbol{s}}$, we account for parameter uncertainty by integrating out the parameters. This can be interpreted as a Bayesian approach and the parameters are weighted according to a prior distribution (see Supporting Information S4). In Step 3 below, we show how Markov chain Monte Carlo (MCMC) can be used to obtain the statistically downscaled distribution $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B]$.

Step 3: Generating predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$ via conditional simulation

The aggregation relation that connects $\boldsymbol{X}_B$ and $\boldsymbol{X}_{\boldsymbol{s}}$, namely, $\boldsymbol{X}_B = \boldsymbol{W}_B \boldsymbol{X}_{\boldsymbol{s}}$, implies that $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}]$ is a multivariate normal distribution. Then, the predictive distribution of $\boldsymbol{X}_{\boldsymbol{s}}$ given $\boldsymbol{X}_B$ can be obtained by integrating out the variability in $\{\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$, namely,

$$[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B] = \int [\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}]\, [\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi} \,|\, \boldsymbol{X}_B]\, \mathrm{d}\gamma_0\, \mathrm{d}\boldsymbol{\gamma}\, \mathrm{d}\boldsymbol{\phi}. \qquad \text{eqn 6}$$

Note that in Step 2, we simulate samples from $[\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi} \,|\, \boldsymbol{X}_B]$ (the second distribution inside the integral in equation 6). Thus, we can generate samples from $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B]$ via conditional simulations as follows. Let $\{\gamma_0^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\phi}^{(k)}\}$ be the k-th posterior sample from $[\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi} \,|\, \boldsymbol{X}_B]$ obtained via MCMC (see Supporting Information S4). We generate the k-th predictive sample $\boldsymbol{x}_{\boldsymbol{s}}^{(k)}$ from $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\phi}^{(k)}]$, where $\boldsymbol{x}_{\boldsymbol{s}}^{(k)} \equiv (x^{(k)}(\boldsymbol{s}_1), \ldots, x^{(k)}(\boldsymbol{s}_{N_D}))^\top$. Repeating this step for $k = 1, \ldots, K$, we obtain a set of K downscaled samples $\{\boldsymbol{x}_{\boldsymbol{s}}^{(k)} : k = 1, \ldots, K\}$ generated from $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B]$, hereafter referred to as the predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$ given $\boldsymbol{X}_B$. Details of the distribution $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}]$ and the conditional simulation procedure can be found in Supporting Information S1.

Our downscaling procedure is summarised in Fig. 2. Using the predictive samples $\{\boldsymbol{x}_{\boldsymbol{s}}^{(k)} : k = 1, \ldots, K\}$, we can obtain a downscaled value with its predictive mean and quantify the downscaling uncertainty with its predictive variance or with its $100(1-\alpha)\%$ prediction intervals, for $0 < \alpha < 1$. In Section 4.1, we provide an example of applying this method and measure downscaling uncertainty through prediction intervals from the predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$.

3 Incorporating statistical downscaling into biodiversity models

We first examine the impact of neglecting downscaling uncertainty when making inferences on biodiversity through an analysis of variance (Section 3.1). Then, in Section 3.2, we develop a

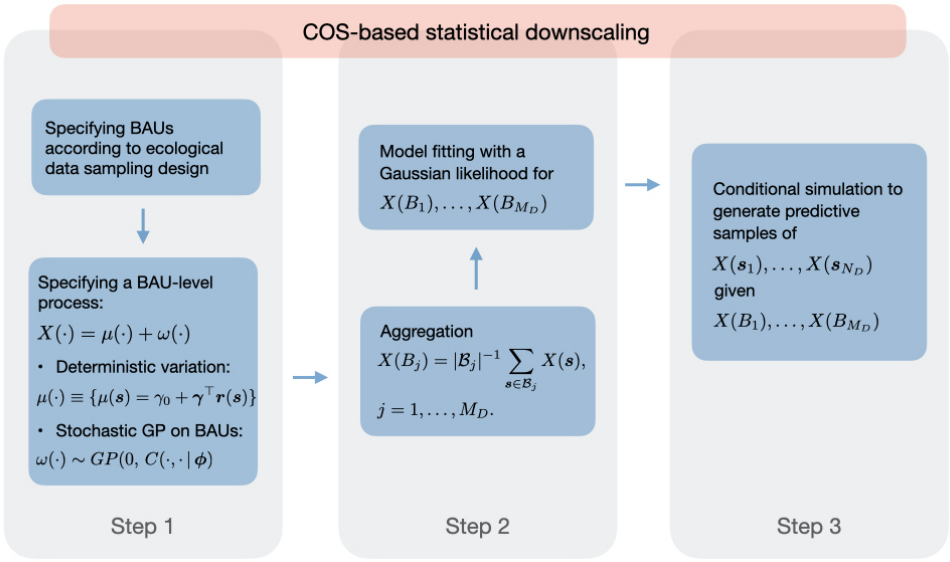


Figure 2: Schematic diagram for the change-of-support (COS)-based statistical downscaling.

protocol that incorporates downscaling uncertainty when fitting biodiversity models. Notation used in this section is summarised in Table 2.

3.1 Analysis of variance attributed to downscaling

Let $[\boldsymbol{Z}_{\boldsymbol{s}} \mid \boldsymbol{X}_{\boldsymbol{s}}, \boldsymbol{\theta}]$ represent a biodiversity model, where $\boldsymbol{Z}_{\boldsymbol{s}} \equiv (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_{N_D}))^\top$ is a vector of univariate responses and the vector $\boldsymbol{\theta}$ collects all unknown parameters. In biodiversity studies that involve downscaling, $\boldsymbol{X}_{\boldsymbol{s}}$ is not observable, and only data at a coarser resolution, namely, $\boldsymbol{X}_B$, is available. Given $\boldsymbol{X}_B$, the biodiversity mode is $[\boldsymbol{Z}_{\boldsymbol{s}} \mid \boldsymbol{X}_B, \boldsymbol{\theta}]$. Through conditioning and marginalisation, we see that $[\boldsymbol{Z}_{\boldsymbol{s}} \mid \boldsymbol{X}_B, \boldsymbol{\theta}] = \int [\boldsymbol{Z}_{\boldsymbol{s}} \mid \boldsymbol{X}_{\boldsymbol{s}}, \boldsymbol{\theta}] [\boldsymbol{X}_{\boldsymbol{s}} \mid \boldsymbol{X}_B] \, \mathrm{d}\boldsymbol{X}_{\boldsymbol{s}}$, which motivates the two-stage protocol referred to in Section 1: statistical-downscaling using $[\boldsymbol{X}_{\boldsymbol{s}} \mid \boldsymbol{X}_B]$ in the first stage, and using the biodiversity model $[\boldsymbol{Z}_{\boldsymbol{s}} \mid \boldsymbol{X}_{\boldsymbol{s}}, \boldsymbol{\theta}]$ in the second stage.

Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$ in $[\boldsymbol{Z}_{\boldsymbol{s}} \mid \boldsymbol{X}_{\boldsymbol{s}}, \boldsymbol{\theta}]$. Using the law of total variance, we decompose the variance of $\hat{\boldsymbol{\theta}}$ as $\mathrm{var}(\hat{\boldsymbol{\theta}}) = E[\mathrm{var}(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}_{\boldsymbol{s}})] + \mathrm{var}[E(\hat{\boldsymbol{\theta}} \mid \boldsymbol{X}_{\boldsymbol{s}})]$, where the first term can be interpreted as a weighted average of the variances of $\hat{\boldsymbol{\theta}}$ conditional on different values of $\boldsymbol{X}_{\boldsymbol{s}}$. It measures within-group variability, where in practice the 'group' refers to a specific data set defined by $\boldsymbol{X}_{\boldsymbol{s}}$. On the other hand, the second term measures between-group variability as it captures the variability caused by the changes in $\boldsymbol{X}_{\boldsymbol{s}}$ that result in different conditional means of $\hat{\boldsymbol{\theta}}$.

Ignoring downscaling uncertainty has consequences. The plug-in approach referred to in Section 1 replaces the unobservable $\boldsymbol{X}_{\boldsymbol{s}}$ with its prediction (e.g., $E(\boldsymbol{X}_{\boldsymbol{s}} \mid \boldsymbol{X}_B)$ or a smoother such as from

Table 2: Notation used in Section 3.

$\hat{\boldsymbol{X}}_{\boldsymbol{s}}$	$\hat{\boldsymbol{X}}_{\boldsymbol{s}} \equiv (\hat{X}(\boldsymbol{s}_1), \ldots, \hat{X}(\boldsymbol{s}_{N_D}))^\top$, the vector of GPB predictive means
$\boldsymbol{Z}_{\boldsymbol{s}}$	$\boldsymbol{Z}_{\boldsymbol{s}} \equiv (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_{N_D}))^\top$, the vector of biodiversity response data
$\boldsymbol{\eta}_{\boldsymbol{s}}$	$\boldsymbol{\eta}_{\boldsymbol{s}} \equiv (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_{N_D}))^\top$, the vector of downscaling errors
$\hat{x}(\boldsymbol{s}_i)$	$\hat{x}(\boldsymbol{s}_i) \equiv K^{-1} \sum_{k=1}^{K} x^{(k)}(\boldsymbol{s}_i)$, a simulation-based estimate of $\hat{\boldsymbol{X}}(\boldsymbol{s}_i)$, $i = 1, \ldots, N_D$
$\hat{\boldsymbol{x}}_{\boldsymbol{s}}$	$\hat{\boldsymbol{x}}_{\boldsymbol{s}} \equiv (\hat{x}(\boldsymbol{s}_1), \ldots, \hat{x}(\boldsymbol{s}_{N_D}))^\top$
$\hat{\eta}^{(k)}(\boldsymbol{s}_i)$	$\hat{\eta}^{(k)}(\boldsymbol{s}_i) \equiv x^{(k)}(\boldsymbol{s}_i) - \hat{x}(\boldsymbol{s}_i)$, based on the k-th sample $\boldsymbol{x}_{\boldsymbol{s}}^{(k)}$, $i = 1, \ldots, N_D$
$\hat{\boldsymbol{\eta}}_{\boldsymbol{s}}^{(k)}$	$\hat{\boldsymbol{\eta}}_{\boldsymbol{s}}^{(k)} \equiv (\hat{\eta}^{(k)}(\boldsymbol{s}_1), \ldots, \hat{\eta}^{(k)}(\boldsymbol{s}_{N_D}))^\top$
$\boldsymbol{\theta}$	Vector of biodiversity-model parameters
$\hat{\boldsymbol{\theta}}$	Estimator of parameter $\boldsymbol{\theta}$

a family of splines), and thus it considers only the first term in the variance decomposition above for $\hat{\boldsymbol{\theta}}$. It does not account for the second term, potentially resulting in over-confident scientific conclusions. Extending the decomposition of variance from estimating parameters to predicting biodiversity (see Supporting Information S2), demonstrates that ignoring downscaling uncertainty may lead to underestimation of the prediction variance, resulting in an uncertainty-quantification procedure that fails to account for all possible outcomes.

3.2 Biodiversity modelling with downscaled covariate samples (Stage 2 of protocol

Section 3.1 details the importance of accounting for downscaling uncertainty in biodiversity modelling. Here, we outline how downscaling uncertainty from the first stage of our protocol can be propagated to biodiversity models in the second stage (Fig. 3). Without loss of generality and for illustrative purposes, we assume that response data $\{Z(\boldsymbol{s}_i)\}$ are available on all BAUs. Supporting Information S3 provides a more general protocol that considers situations with response data limited to certain BAU locations.

Recall that we generate predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$ from $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B]$ in equation 6. Let $\hat{\boldsymbol{X}}_{\boldsymbol{s}} \equiv (\hat{X}(\boldsymbol{s}_1), \ldots, \hat{X}(\boldsymbol{s}_{N_D}))^\top \equiv E(\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B)$ be the predictive mean. We consider the following representation: Conditional on $\boldsymbol{X}_B$,

$$\boldsymbol{X}_{\boldsymbol{s}} = \hat{\boldsymbol{X}}_{\boldsymbol{s}} + \boldsymbol{\eta}_{\boldsymbol{s}}, \qquad \text{eqn 7}$$

in distribution. The vector $\boldsymbol{\eta}_{\boldsymbol{s}} \equiv (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_{N_D}))^\top$ in equation 7 represents the downscaling

error associated with $\hat{\boldsymbol{X}}_{\boldsymbol{s}}$, where $E(\eta(\boldsymbol{s}_i) \,|\, \boldsymbol{X}_B) = 0$, for $i = 1, \ldots, N_D$. The plug-in approach that simply uses $\hat{\boldsymbol{X}}_{\boldsymbol{s}}$ ignores uncertainty that arises from the spatially-correlated downscaling error $\boldsymbol{\eta}_{\boldsymbol{s}}$.

Equation 7 induces a Berkson-error relationship (e.g., Carroll et al. 2006) in which the unobserved $\boldsymbol{X}_{\boldsymbol{s}}$, conditional on $\boldsymbol{X}_B$, is centred around $\hat{\boldsymbol{X}}_{\boldsymbol{s}}$ with error $\boldsymbol{\eta}_{\boldsymbol{s}}$. In this relationship, $\boldsymbol{X}_{\boldsymbol{s}}$ is more variable than $\hat{\boldsymbol{X}}_{\boldsymbol{s}}$, which is in contrast with the classical-error model in which the predicted covariate is more variable than the unobserved covariate; see Stoklosa et al. (2015) for a discussion of the two types of error models in the context of building species distribution models (SDMs).

To incorporate the downscaling uncertainty into biodiversity modelling, we focus on the class of generalised linear models (GLMs) that are commonly used in biodiversity studies, including in SDMs. For illustration, we assume a simple linear regression on the covariate $X(\cdot)$. Specifically, conditional on $\boldsymbol{X}_B$,

$$\begin{aligned} [Z(\boldsymbol{s}_i) \,|\, Y(\boldsymbol{s}_i)] &= \mathrm{EF}(Y(\boldsymbol{s}_i)), \\ g(Y(\boldsymbol{s}_i)) &= \beta_0 + \beta_1 \hat{X}(\boldsymbol{s}_i) + \delta(\boldsymbol{s}_i), \end{aligned} \qquad \text{eqn 8}$$

in distribution, for some link function $g(\cdot)$, where EF is a distribution that belongs to an exponential family with mean parameter $Y(\cdot)$ such that $E(Z(\boldsymbol{s}_i) \,|\, Y(\boldsymbol{s}_i)) = Y(\boldsymbol{s}_i)$, and the vector of unknown parameters $\boldsymbol{\theta} \equiv (\beta_0, \beta_1)^\top$. The downscaling uncertainty is incorporated through the component $\delta(\boldsymbol{s}_i) \equiv \beta_1 \eta(\boldsymbol{s}_i)$, which is the downscaling error $\eta(\boldsymbol{s}_i)$ scaled by β_1. Since the model in equation 8 utilises the Berkson-error relationship given in equation 7, we refer it as *GLM-Berkson*. Without the component $\delta(\boldsymbol{s}_i)$, the model in equation 8 reduces to using the plug-in $\hat{X}(\boldsymbol{s}_i)$ with link function, $g(Y(\boldsymbol{s}_i)) = \beta_0 + \beta_1 \hat{X}(\boldsymbol{s}_i)$, hereafter referred to as *GLM-plugin*.

Stage 2 of the protocol involves fitting the GLM-Berkson model for each of the K predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$. Our primary goal is a protocol that can be implemented with the aid of standard statistical software. In this regard, when fitting the GLM-Berkson model for the k-th iteration, we replace $\delta(\boldsymbol{s}_i)$ with $\hat{\delta}^{(k)}(\boldsymbol{s}_i) \equiv \tilde{\beta}_1 \hat{\eta}^{(k)}(\boldsymbol{s}_i)$, where $\tilde{\beta}_1$ is obtained from fitting the GLM-plugin model; $\hat{\eta}^{(k)}(\boldsymbol{s}_i) \equiv x^{(k)}(\boldsymbol{s}_i) - \hat{x}(\boldsymbol{s}_i)$, for $i = 1, \ldots, N_D$; and $\hat{x}(\boldsymbol{s}_i) \equiv K^{-1} \sum_{k=1}^{K} x^{(k)}(\boldsymbol{s}_i)$. Thus, the GLM-Berkson model can be fitted by taking $\hat{X}(\boldsymbol{s}_i) = \hat{x}(\boldsymbol{s}_i)$, treating $\hat{\delta}^{(k)}(\boldsymbol{s}_i)$ as an offset (Zuur et al., 2009), and using standard statistical software that supports fitting GLMs. Uncertainty associated with estimating the GLM's parameters is quantified through a parametric bootstrap, which is a data-resampling method commonly used to estimate the distribution of a parameter estimate obtained from fitting a parametric model (e.g., Davison and Hinkley 1997). The final output of the protocol is a collection of estimates of parameter vector $\boldsymbol{\theta}$ that can be used to quantify uncertainty associated with any inferences and/or predictions related to the model. Further details are given in Supporting Information S3.

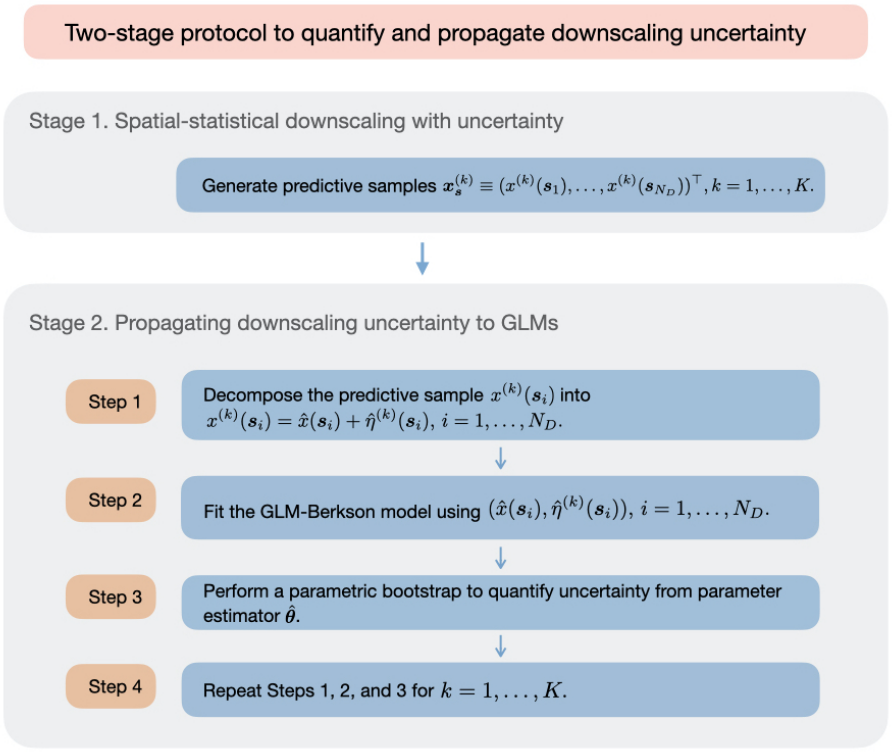


Figure 3: Schematic diagram for the two-stage protocol for downscaling environmental covariates and for their use with downscaling uncertainty in biodiversity studies based on generalised linear models (GLMs).

There is another approach that on the surface might seem appropriate. One might fit a separate GLM for each of the K samples: For $k = 1, \ldots, K$, replace the linear relation in equation 8 with $g(Y(\boldsymbol{s}_i)) = \beta_0 + \beta_1 x^{(k)}(\boldsymbol{s}_i)$, and estimate β_0 and β_1 based on this model. However, this approach, which we call *GLM-ensemble*, may produce a large bias in the estimate of β_1. Recall that the downscaled sample value, $x^{(k)}(\boldsymbol{s}_i)$, can be decomposed into $\hat{x}(\boldsymbol{s}_i)$ and $\hat{\eta}^{(k)}(\boldsymbol{s}_i)$ according to equation 7. Compared with GLM-plugin that uses $\hat{x}(\boldsymbol{s}_i)$, GLM-ensemble uses the downscaled sample value, $x^{(k)}(\boldsymbol{s}_i)$, which contains the error $\hat{\eta}^{(k)}(\boldsymbol{s}_i)$. It is well known that adding an error to a covariate (e.g., adding $\hat{\eta}^{(k)}(\boldsymbol{s}_i)$ to $\hat{x}(\boldsymbol{s}_i)$) can cause a biased estimate of the regression coefficient (Carroll et al., 2006) and, thus, when compared with the estimate of β_1 by GLM-plugin, the estimate by GLM-ensemble is biased. In Section 4.2, our simulation study that compares GLM-Berkson to GLM-plugin, also includes GLM-ensemble as a warning that this intuitive approach can result in biased inference.

4 Simulation and case study

We conducted two simulation studies: The first one focuses on Stage 1 of the protocol, demonstrating the importance of using COS to accurately quantify uncertainty when downscaling. The second one uses a Poisson GLM in Stage 2 and investigates the effects of downscaling uncertainty on inferences and predictions; we show that our GLM-Berkson model leads to improved estimation and predictive inferences when compared to GLM-plugin and GLM-ensemble. Finally, we analyse an Antarctic bryophyte (moss) data set using our two-stage protocol. Here, we focus on reporting the main results, with the details of the implementation given in Supporting Information S5.

4.1 Simulation study: Statistical downscaling with uncertainty (Stage 1)

Our 'ground truth' is a covariate spatial field $\boldsymbol{X}_{\boldsymbol{s}} \equiv (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_{N_D}))^\top$ simulated from a GP with an exponential covariance function (see Section 2.2) on a unit square domain with $50 \times 60 = 3000$ rectangular BAUs (i.e., $N_D = 3000$); see Fig. 4(a). The covariate data are obtained by aggregating the fine-resolution field in Fig. 4(a) to a coarse-resolution grid of 5×6 blocks (i.e., $M_D = 30$) as shown in Fig. 4(b). In the study, it is assumed that we only know the covariate values on the 30 blocks; our aim is to recover the 3000 covariate values at the fine-resolution BAU field.

We considered two downscaling methods: (i) the COS-based downscaling based on GPB (Section 2.2); (ii) the thin-plate spline (TPS; e.g., Wood 2003) smoother that is commonly used for downscaling (see, e.g., Werner et al. 2019; Maclean 2020). The TPS is a generalisation of the natural cubic spline in one dimension, producing a smooth surface. In the context of downscaling, the TPS assumes the M_D coarse-resolution values and their coordinates are in fact M_D fine-resolution values at those coordinates, and there are N_D fine-resolution values to predict at the BAU centroids. Thus, TPS does not consider change-of-support when it is used for downscaling. We implemented TPS using the R package **mgcv** (Wood, 2017), which generated a prediction with a standard-error estimate for each of the N_D fine-resolution BAUs. In contrast, the GPB produced, for each BAU, not just one value but a set of MC samples from the predictive distribution.

Fig. 4(d) and 4(e) show the GPB's predictive means and TPS's predictions on the fine-resolution BAUs, respectively. Although the two fields look visually similar, there is a distinct difference between the two methods when it comes to uncertainty quantification. We constructed two nominal 95% prediction intervals. For GPB, we used the 2.5-th and 97.5-th percentiles of the predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$. For TPS, we defined the prediction interval by adding and subtracting two standard errors from the prediction. We then examined the empirical coverage probability, namely, the

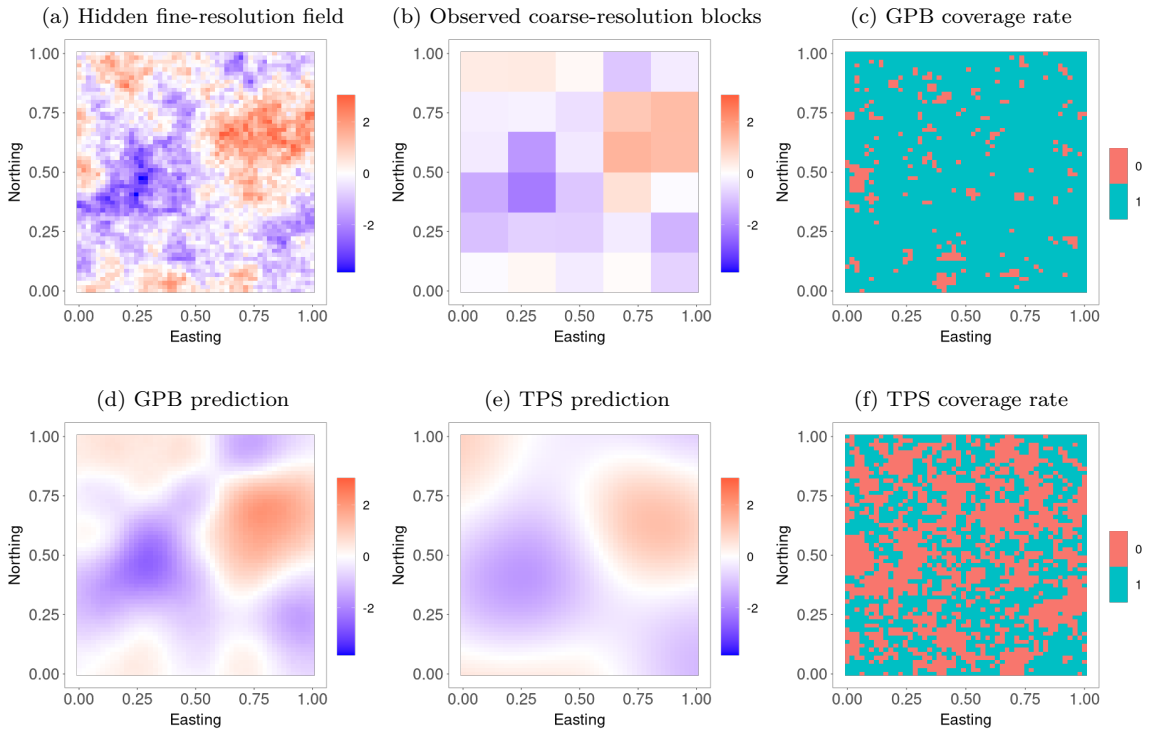


Figure 4: First simulation study (Section 4.1): unobserved fine-resolution covariate field (a); observed coarse-resolution covariate (b) obtained by aggregating (a); predictions of the fine-resolution covariate field given by GPB (d) and TPS (e). In (c) and (f), each pixel indicates whether or not the corresponding 95% prediction interval covers the true value.

percentage of prediction intervals that covered the true values on the BAUs, which are known in this simulation study. Fig. 4(c) and 4(f) reveal the benefit of accounting for change-of-support in downscaling. There, each pixel indicates whether or not the corresponding prediction interval covers the true value. The COS-based downscaling based on GPB achieved a coverage probability of 92%, which is close to the nominal value of 95%, while the TPS smoother's coverage probability was only 50%. The results show that, under suitable modelling assumptions, our proposed GPB method for downscaling provides sound quantification of uncertainty.

4.2 Simulation study: Model fitting and uncertainty propagation (Stage 2)

As in Section 4.1, we simulated covariate fields on the unit square with 3000 BAUs, using a GP with an exponential covariance function with range ψ_0 and variance σ_0^2. Each simulation of a covariate field is a 'ground truth' for the study and corresponds to different values of ψ_0 (from 0.1 to 0.8 in increments of 0.1) and σ_0^2 (from 1 to 5 in increments of 1). This resulted in 40 scenarios, each corresponding to a different simulated covariate field, allowing us to investigate

how biodiversity inferences are affected by the degree of spatial dependence and variability in the covariate that requires downscaling. Note that each simulated covariate field has 3000 covariate values corresponding to the 3000 BAUs.

For each of the 40 scenarios, we first simulated BAU-level response data from a Poisson GLM using the true simulated covariate field; the intercept and regression coefficient were set to $(\beta_0, \beta_1)^\top = (1, 0.5)^\top$. We then aggregated the simulated covariate field to 30 blocks, each of which consists of 100 BAUs. This produced 3000 BAU-level response data and 30 coarse-resolution covariate data, illustrating the mismatch in resolution between response and covariate data. For training data to which the Poisson GLM is fitted, a common set of 600 randomly-selected locations from a total of 3000 were used to provide the 600-dimensional vector of response data for each of the 40 scenarios. The remaining 2400 locations were used to evaluate model performances.

To fit models, for each of the 40 scenarios, we first downscaled the coarse-resolution covariate data using the GPB method and obtained 500 downscaled MCMC samples (Stage 1); each downscaled sample of the covariate has 3000 values. Then in Stage 2, we fitted the three GLMs described in Section 3.2 (GLM-Berkson, GLM-plugin, and GLM-ensemble) to the 600-dimensional response data with downscaled covariate data at the same locations. We used the mean of the downscaled covariate samples when fitting GLM-plugin. We also fitted a GLM that uses the true simulated covariate field as a 'gold standard', which we call *GLM-oracle*. Note that GLM-oracle can only be done in the simulation study where we know 'the truth'. We fitted all four models using the **stats** package in the R computing environment (R Core Team, 2021).

We used the following two metrics to evaluate model inference on the coefficient: (i) bias of $\hat{\beta}_1$ and (ii) nominal 95% coverage probability (CP) for β_1. Each model's predictive performance was evaluated based on test data at the remaining 2400 locations, using metrics (iii) out-of-sample root-mean-square-predictive-error (RMSPE) for the linear predictor and (iv) out-of-sample nominal 95% CP for the linear predictor. We report results below for the 40 scenarios; for each of the 40 scenarios, given the simulated covariate field, we repeated the process of simulating response data and fitting models 100 times in order to obtain precise estimates of the metrics (see Supporting Information S5 for more details).

The bias of $\hat{\beta}_1$ and the 95% CP for β_1 are plotted against increasing values of the spatial range ψ_0 in Fig. 5(a) and Fig. 5(c) and of the variance σ_0^2 in Fig. 5(b) and Fig. 5(d). The GLM-ensemble produced significantly biased estimates, leading to poor coverage of β_1. On the other hand, GLM-plugin and GLM-Berkson had much less bias in their estimates, with accuracy comparable to those from GLM-oracle. Furthermore, GLM-Berkson that properly incorporates downscaling uncertainty,

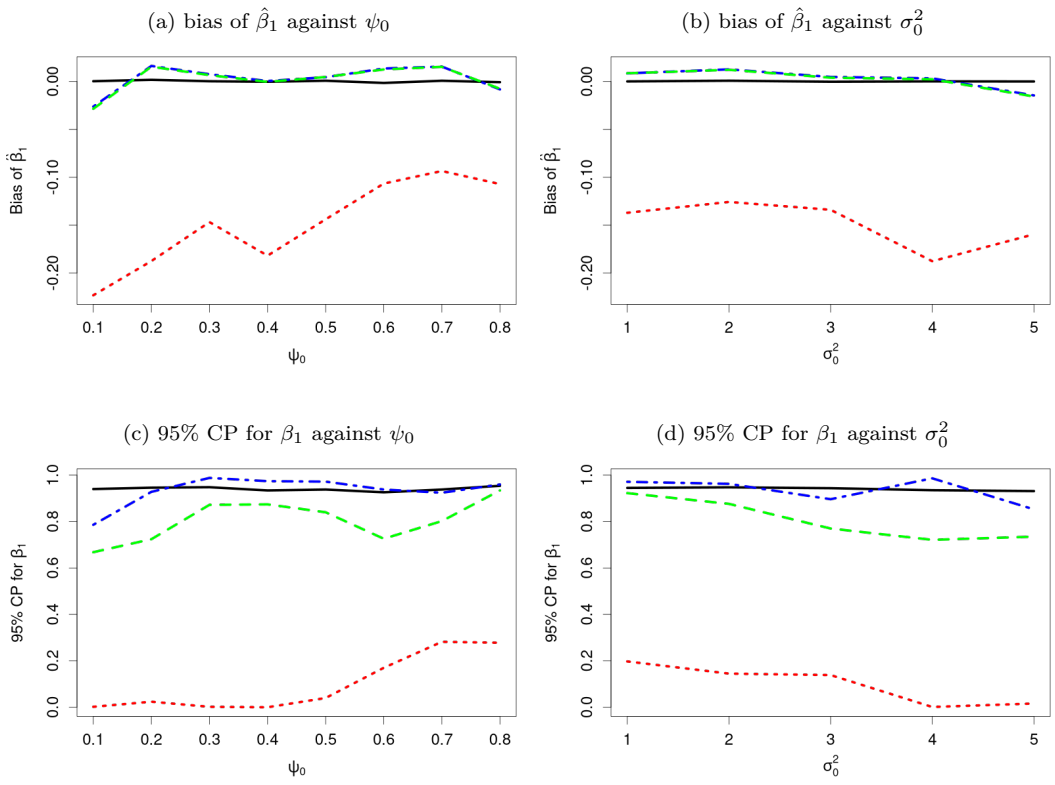


Figure 5: Second simulation study (Section 4.2): plots of biases of $\hat{\beta}_1$ averaged across spatial range ψ_0 (a) and across variance σ_0^2 (b); plots of 95% coverage probabilities (CPs) for β_1 averaged across ψ_0 (c) and across σ_0^2 (d). The lines correspond to the following models: GLM-oracle (black solid), GLM-Berkson (blue dot-dashes), GLM-plugin (green dashes), GLM-ensemble (red dots).

was superior to GLM-plugin in terms of 95% CPs, providing interval estimates of β_1 similar to those from GLM-oracle.

Regarding out-of-sample predictive performance (Fig. 6), GLM-Berkson slightly outperformed GLM-plugin for out-of-sample prediction based on RMSPE, and both models had smaller RMSPEs than GLM-ensemble. GLM-Berkson's predictive 95% CPs are close to those from GLM-oracle. Probably due to its biased estimates of β_1, GLM-ensemble produced worse predictive CPs compared to GLM-Berkson, but still better than those for GLM-plugin, which does not consider downscaling uncertainty at all.

Overall, our simulations demonstrate that biodiversity model inference and prediction based on a GLM that involves coarse-resolution covariates can be highly susceptible to downscaling errors. Thus, accounting for uncertainty from downscaling when fitting a biodiversity model is crucial. Our two-stage protocol of spatial-statistical downscaling followed by GLM-Berkson, gave precise, valid inferences. Importantly, we used readily available software to make those inferences.

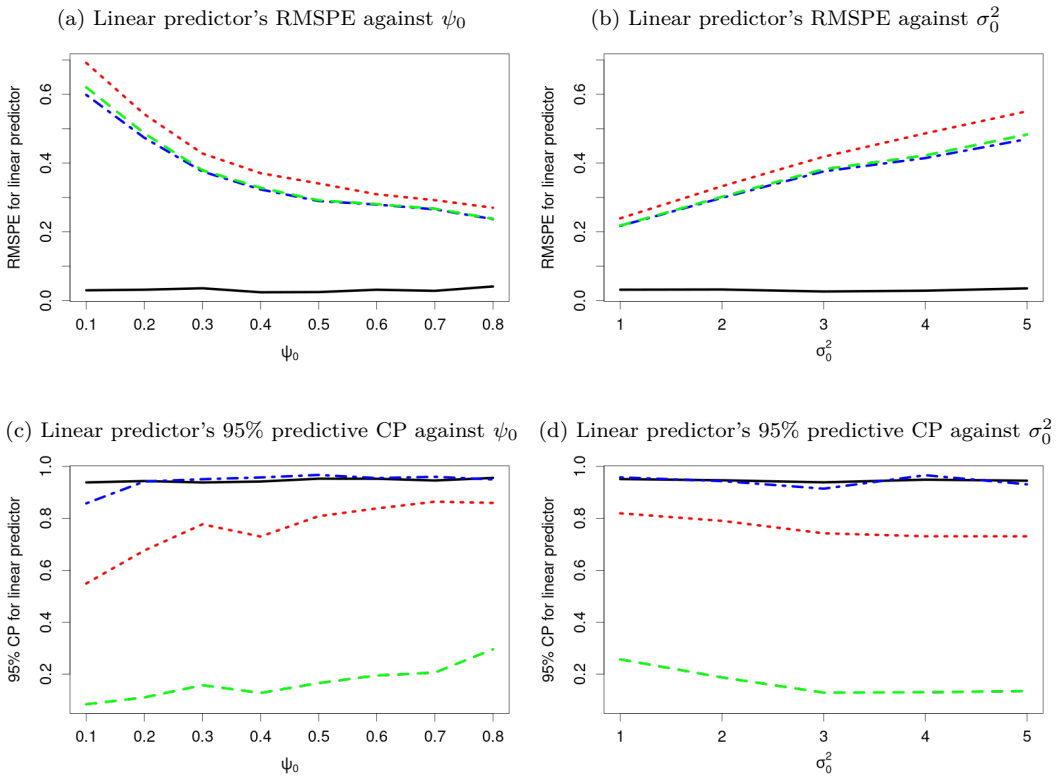
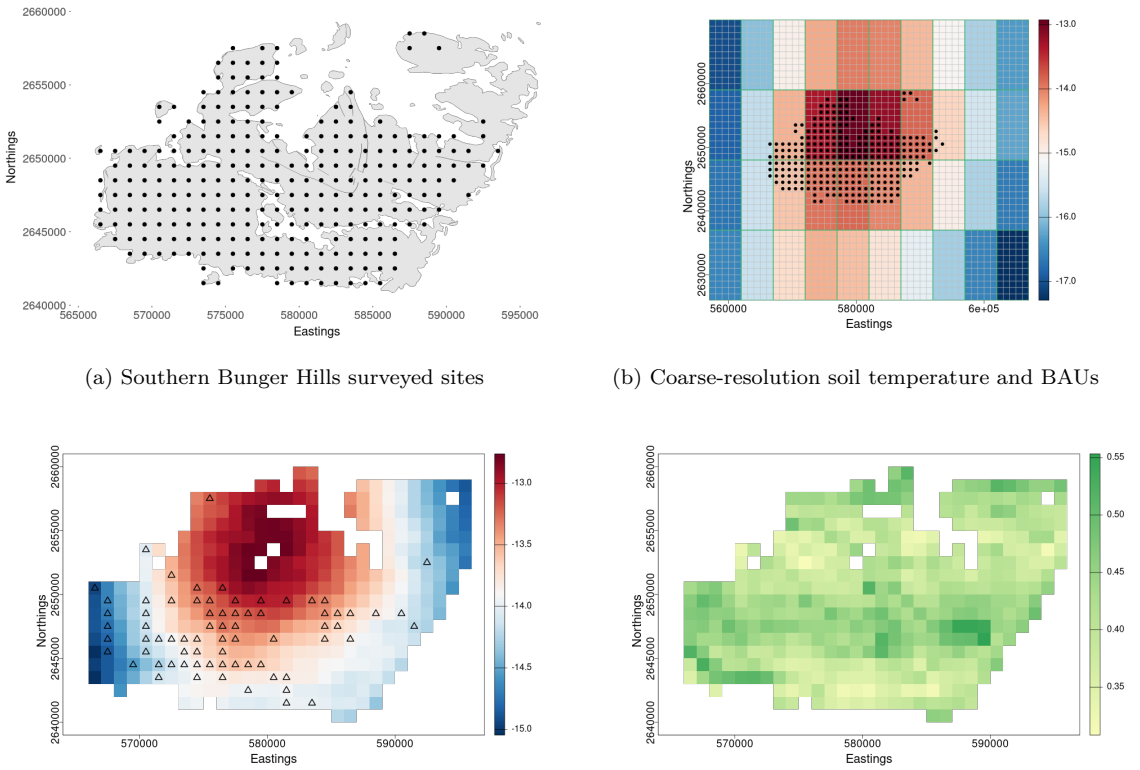


Figure 6: Second simulation study (Section 4.2): plots of RMSPEs for the linear predictors averaged across spatial range ψ_0 (a) and across variance σ_0^2 (b); plots of 95% coverage probabilities (CPs) for the linear predictors averaged across ψ_0 (c) and across σ_0^2 (d). The lines correspond to the following models: GLM-oracle (black solid), GLM-Berkson (blue dot-dashes), GLM-plugin (green dashes), GLM-ensemble (red dots).

4.3 Case study: Incorporating downscaling uncertainty when predicting moss occurrence in East Antarctica

Ecosystems of Antarctica and the sub-Antarctic islands are fragile. Cannone et al. (2022) show Antarctic flora undergoing accelerated responses to climate warming. The vast majority of known terrestrial biodiversity in Antarctica is restricted to ice-free areas, which occupy a small fraction of the entire landmass. As such, modelling of Antarctic biodiversity in response to environmental change typically requires environmental data at a resolution finer than currently available. Here, we demonstrate the two-stage protocol for modelling and predicting moss occurrence. Mosses are a dominant component of the vegetation on the Antarctic continent, making their distribution a crucial aspect of the structure and functioning of Antarctic terrestrial ecosystems, and of how these ecosystems will respond to climate change (Robinson et al., 2018).

Specifically, we consider the southern Bunger Hills (Fig. 7) in East Antarctica, omitting the



(a) Southern Bunger Hills surveyed sites

(b) Coarse-resolution soil temperature and BAUs

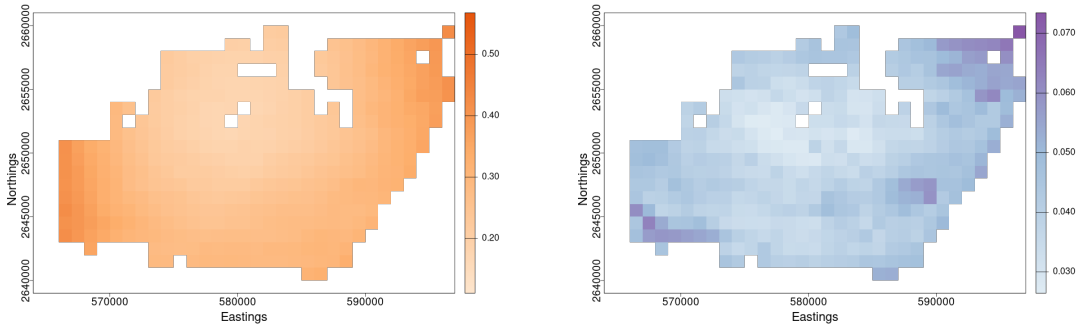
(c) Predicted soil temperature

(d) Standard deviations of the predicted soil temperature

Figure 7: Bunger Hills data analysis: (a) surveyed sites in southern Bunger Hills between 1995 and 2000; (b) coarse-resolution temperature in blocks, overlaid with survey sites, and BAUs are drawn with white borders; (c) predictive mean of the BAU-level soil-temperature, overlaid with sites where moss was present (triangles); (d) standard deviation of the soil-temperature predictive distribution at the fine-resolution BAU level.

islands and rock areas north of Fuller Island. As an ice-free area , Bunger Hills is a small coastal region with hilly topography and many lakes. Among the data available from the Australian Antarctic Data Centre (Gore et al., 2019) and Leishman et al. (2020), are the moss occurrences recorded between 1995 and 2000. The presence of moss was found in 72 out of a total of 277 grid cells (Fig. 7(a)); each cell is of size $1 \times 1\,\text{km}^2$. However, covariate data are not available at this resolution (Leishman et al., 2020). The climate reanalysis product we will use, ERA5-Land, is available at a 0.1-degree resolution, which corresponds to blocks of size around $5 \times 11\,\text{km}^2$ in the Bunger Hills. Consequently, studies for this region require downscaling from coarse-resolution blocks to fine-resolution grid cells of size $1 \times 1\,\text{km}^2$, to align with the occurrence data collected according to the sampling design.

We considered a Poisson GLM, and for illustrative purposes, modelled moss occurrence is a function of one covariate, soil temperature. Since the area of southern Bunger Hills is small, we



(a) Mean presence probabilities

(b) Standard deviations of the predictions

Figure 8: Modelling of the moss occurrence across the southern Bunger Hills of East Antarctica: mean (a) and standard deviation (b) of the predictive distribution of the moss presence probability.

downscaled the soil-temperature data from a larger region that consists of 40 ERA5-Land blocks, which allowed for borrowing information from the surrounding area. We set the BAUs as the grid cells of the field-sampling design, so each BAU has size $1 \times 1\,\text{km}^2$. This results in 2200 BAUs within the 40 coarse-resolution blocks (Fig. 7(b)).

We carried out the first stage of the two-stage protocol (Section 2.2) to downscale soil temperature, using a GPB with an exponential covariance function, and topographic data are used as BAU-level covariates to model the soil temperature: Elevation data from the Reference Elevation Model of Antarctica (Howat et al., 2019), terrain slope (the rate of change in elevation across the surface of the terrain), and aspect (compass direction that the slope of the terrain faces). Maps at the BAU-level of the mean and standard deviation of soil-temperature predictive distribution are given in Fig. 7(c) and 7(d), respectively. Fig. 7(c) is overlaid with moss-presence sites, which shows qualitatively that mosses tends to be present at relatively low temperatures in the region.

Stage 2 of our protocol involved fitting the GLM-Berkson model to the moss occurrence data with the predictive samples of the soil temperature obtained from Stage 1. The soil-temperature-coefficient estimate was $\hat{\beta}_1 = -0.524$, with its 95% interval estimates: $(-0.681, -0.369)$. The result suggests that, on average, lower soil temperatures increased moss-presence probabilities in the southern Bunger Hills. Fig. 8(a) and 8(b) maps each BAU's predictive mean and predictive standard deviation of the moss-presence probability, respectively, from the GLM-Berkson model. It can be seen that larger prediction uncertainties appear in BAUs where mosses were not found or not surveyed and soil temperatures were relatively low, as well as in BAUs where soil-temperature downscaling uncertainties were large. These uncertainty estimates arise from uncertainties in both downscaling (Stage 1) and model-fitting (Stage 2).

5 Discussion

We have introduced and demonstrated a framework for statistically downscaling coarse-resolution data, whereby spatial models defined at fine-resolution BAUs account for spatial change-of-support. The downscaling method is the first stage of a two-stage protocol in the second stage that incorporates downscaling uncertainty in GLM-based biodiversity models. Clearly, the second stage could be used in non-biodiversity applications where the response is at a resolution different to that of the covariate(s). The protocol's accuracy and its valid quantification of uncertainty has been demonstrated through simulation studies, and its practicality using standard software has been shown in a case study. The downscaling method involves fitting an 'aggregated-GP' model; see Supporting Information S7 for a brief discussion of the computational cost. We remark below on extensions to accommodate possibly more complicated situations than those presented in Sections 2 and 3.

Numerical models that generate, or instruments that measure covariates at coarse resolutions may include errors. As such, equation 2 may be replaced with $X(B) \equiv |\mathcal{B}|^{-1} \sum_{\boldsymbol{s} \in B} X(\boldsymbol{s}) + \epsilon(B)$, where the error $\epsilon(B)$ is independent of $\{X(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ and captures instrument bias and/or measurement-error variance (e.g., Nguyen et al. 2012). This modification leads to a distribution of $\boldsymbol{X}_B$ different to that given by equation 5. Suppose $\epsilon(B) \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau^2)$ for non-overlapping blocks. Then we have that $[\boldsymbol{X}_B \,|\, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}, \tau^2] = \mathrm{N}(\boldsymbol{W}_B(\boldsymbol{1}_{N_D}\gamma_0 + \boldsymbol{R}\boldsymbol{\gamma}), \boldsymbol{W}_B\boldsymbol{C}_{\boldsymbol{\phi}}\boldsymbol{W}_B^\top + \tau^2\boldsymbol{I}_{M_D})$, where $\boldsymbol{I}_{M_D}$ is an $M_D \times M_D$ diagonal matrix.

In practice, there may exist spatial misalignment: The coarse-resolution blocks aggregated over the BAUs may not completely align with the coarse-resolution covariate data, which could happen if the centroids of the BAU-aggregated blocks and the centroids of the covariate-data blocks are not exactly co-located. For the case study in Section 4.3, the spatial misalignment was mild, and thus we simply assigned the soil-temperature covariate data from ERA5-Land to the BAU-aggregated blocks. In Supporting Information S6, we provide a brief discussion of an extension of equation 2 to deal with spatial misalignment.

Missing values at the coarse resolution are possible. Our approach to downscaling naturally resolves this issue, as the whole process $X(\cdot)$ is specified at the fine-resolution BAU level, and parameters are estimated based on the (possibly incomplete) coarse-resolution data. Along the same lines, if the coarse-resolution data come from multiple sources, fusing them is straightforward, since they are aggregations from the same process $X(\cdot)$ (see, e.g., Nguyen et al. 2012).

This article focuses on variables that are Gaussian-distributed. Environmental and climate variables (e.g., wind speed) can be non-Gaussian. The most straightforward way to work with

them is to apply an appropriate transformation to the GPB, say $a(\cdot)$ such that $a(X(\boldsymbol{s})) = \mu(\boldsymbol{s}) + \omega(\boldsymbol{s})$, but care must be taken to aggregate/disaggregate appropriately. Alternatively, one could consider a generalised linear mixed model (GLMM) for $X(\cdot)$ with a non-linear link function; see, for example, Sainsbury-Dale et al. (2024) where a class of GLMMs using spatial basis-function models is implemented for for $\omega(\cdot)$.

Finally, we remark that biodiversity modelling routinely demands environmental covariate data that are often available at a resolution coarser than the response data. The two-stage protocol offers a sound solution, with additional strengths including its widespread relevance to biodiversity models and its relative ease of implementation. Given its implementation within a GLM framework, there is scope for advances in the approach, such as species distribution models (SDMs) with multiple covariates and/or non-linear functions of covariates, joint SDMs, among other possibilities. Our results support the notion that there are multiple sources of uncertainty in biodiversity modelling that can be captured with careful, principled statistical modelling.

References

Ahmed, K., Shahid, S., Haroon, S. B., and Xiao-Jun, W. (2015), "Multilayer perceptron neural network for downscaling rainfall in arid region: a case study of Baluchistan, Pakistan," *Journal of Earth System Science*, 124, 1325–1341. https://doi.org/10.1007/s12040-015-0602-9.

Berrocal, V. J., Craigmile, P. F., and Guttorp, P. (2012), "Regional climate model assessment using statistical upscaling and downscaling techniques," *Environmetrics*, 23, 482–492. https://doi.org/10.1002/env.2145.

Cannone, N., Malfasi, F., Favero-Longo, S. E., Convey, P., and Guglielmin, M. (2022), "Acceleration of climate warming and plant dynamics in Antarctica," *Current Biology*, 32, 1599–1606. https://doi.org/10.1016/j.cub.2022.01.074.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman and Hall/CRC, Boca Raton, FL. https://doi.org/10.1201/9781420010138.

Chase, J. M. and Knight, T. M. (2013), "Scale-dependent effect sizes of ecological drivers on biodiversity: why standardised sampling is not enough," *Ecology Letters*, 16, 17–26. https://doi.org/10.1111/ele.12112.

Cressie, N. (1993), *Statistics for Spatial Data (rev. ed.)*, Wiley-Interscience, New York, NY. https://doi.org/10.1002/9781119115151.

— (1996), "Change of support and the modifiable areal unit problem," *Geographical Systems*, 3, 159–180.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511802843.

Denham, R. J., Falk, M. G., and Mengersen, K. L. (2011), "The Bayesian conditional independence model for measurement error: applications in ecology," *Environmental and Ecological Statistics*, 18, 239–255. https://doi.org/10.1007/s10651-009-0130-3.

Dungan, J. L., Perry, J., Dale, M., Legendre, P., Citron-Pousty, S., Fortin, M.-J., Jakomulska, A., Miriti, M., and Rosenberg, M. (2002), "A balanced view of scale in spatial statistical analysis," *Ecography*, 25, 626–640. https://doi.org/10.1034/j.1600-0587.2002.250510.x.

Elston, D. A., Jayasinghe, G., Buckland, S. T., MacMillan, D. C., and Aspinall, R. J. (1997), "Adapting regression equations to minimize the mean squared error of predictions made using covariate data from a GIS," *International Journal of Geographical Information Science*, 11, 265–280. https://doi.org/10.1080/136588197242392.

Ferrier, S., Manion, G., Elith, J., and Richardson, K. (2007), "Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment," *Diversity and Distributions*, 13, 252–264. https://doi.org/10.1111/j.1472-4642.2007.00341.x.

Foster, S. D., Shimadzu, H., and Darnell, R. (2012), "Uncertainty in spatially predicted covariates: is it ignorable?" *Journal of the Royal Statistical Society Series C: Applied Statistics*, 61, 637–652. https://doi.org/10.1111/j.1467-9876.2011.01030.x.

Giorgi, F. and Gutowski Jr, W. J. (2015), "Regional dynamical downscaling and the CORDEX initiative," *Annual Review of Environment and Resources*, 40, 467–490. https://doi.org/10.1146/annurev-environ-102014-021217.

Gore, D., Leishman, M., and Gibson, J. A. (2019), "Biogeography of the Bunger Hills - plant and bird locations 1995-2000," Australian Antarctic Data Centre, Hobart, Australia. https://doi.org/10.26179/5e015f5468952.

Gotway, C. A. and Young, L. J. (2002), "Combining incompatible spatial data," *Journal of the American Statistical Association*, 97, 632–648. https://doi.org/10.1198/016214502760047140.

Hannah, L., Flint, L., Syphard, A. D., Moritz, M. A., Buckley, L. B., and McCullough, I. M. (2014), "Fine-grain modeling of species' response to climate change: holdouts, stepping-stones, and microrefugia," *Trends in Ecology & Evolution*, 29, 390–397. https://doi.org/10.1016/j.tree.2014.04.006.

Hao, T., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2020), "Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models," *Ecography*, 43, 549–558. https://doi.org/10.1111/ecog.04890.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Jean, B., Bonavita, M., De Chiara, G., et al. (2020), "The ERA5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Hoskins, A. J., Harwood, T. D., Ware, C., Williams, K. J., Perry, J. J., Ota, N., Croft, J. R., Yeates, D. K., Jetz, W., Golebiewski, M., Purvis, A., Robertson, T., and Ferrier, S. (2020), "BILBI: Supporting global biodiversity assessment through high-resolution macroecological modelling," *Environmental Modelling & Software*, 132, 104806. https://doi.org/10.1016/j.envsoft.2020.104806.

Howat, I. M., Porter, C., Smith, B. E., Noh, M.-J., and Morin, P. (2019), "The reference elevation model of Antarctica," *The Cryosphere*, 13, 665–674. https://doi.org/10.5194/tc-13-665-2019.

Jansen, J., Woolley, S. N., Dunstan, P. K., Foster, S. D., Hill, N. A., Haward, M., and Johnson, C. R. (2022), "Stop ignoring map uncertainty in biodiversity science and conservation policy," *Nature Ecology & Evolution*, 6, 828–829. https://doi.org/10.1038/s41559-022-01778-z.

Jarzyna, M. A. and Jetz, W. (2018), "Taxonomic and functional diversity change is scale dependent," *Nature Communications*, 9, 2565. https://doi.org/10.1038/s41467-018-04889-z.

Jaureguiberry, P., Titeux, N., Wiemers, M., Bowler, D. E., Coscieme, L., Golden, A. S., Guerra, C. A., Jacob, U., Takahashi, Y., Settele, J., Díaz, S., Molnár, Z., and Purvis, A. (2022), "The

direct drivers of recent global anthropogenic biodiversity loss," *Science Advances*, 8, eabm9982. https://doi.org/10.1126/sciadv.abm9982.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F. E., Pereira, H. M., Regan, E. C., Schmeller, D. S., and Turak, E. (2019), "Essential biodiversity variables for mapping and monitoring species populations," *Nature Ecology & Evolution*, 3, 539–551. https://doi.org/10.1038/s41559-019-0826-1.

Keil, P., Pereira, H. M., Cabral, J. S., Chase, J. M., May, F., Martins, I. S., and Winter, M. (2018), "Spatial scaling of extinction rates: Theory and data reveal nonlinearity and a major upscaling and downscaling challenge," *Global Ecology and Biogeography*, 27, 2–13. https://doi.org/10.1111/geb.12669.

Latombe, G., Hui, C., and McGeoch, M. A. (2017), "Multi-site generalised dissimilarity modelling: Using zeta diversity to differentiate drivers of turnover in rare and widespread species," *Methods in Ecology and Evolution*, 8, 431–442. https://doi.org/10.1111/2041-210X.12756.

Leishman, M. R., Gibson, J. A., and Gore, D. B. (2020), "Spatial distribution of birds and terrestrial plants in Bunger Hills," *Antarctic Science*, 32, 153–166. https://doi.org/10.1017/S0954102020000012.

Lu, M. and Jetz, W. (2023), "Scale-sensitivity in the measurement and interpretation of environmental niches," *Trends in Ecology & Evolution*, 38, 554–567. https://doi.org/10.1016/j.tree.2023.01.003.

Ma, P., Kang, E. L., Braverman, A. J., and Nguyen, H. M. (2019), "Spatial statistical downscaling for constructing high-resolution nature runs in global observing system simulation experiments," *Technometrics*, 61, 322–340. https://doi.org/10.1080/00401706.2018.1524791.

Maclean, I. M. (2020), "Predicting future climate at high spatial and temporal resolution," *Global Change Biology*, 26, 1003–1011. https://doi.org/10.1111/gcb.14876.

Majumder, S., Guan, Y., Reich, B. J., O'Neill, S., and Rappold, A. G. (2021), "Statistical downscaling with spatial misalignment: Application to wildland fire PM2.5 concentration forecasting," *Journal of Agricultural, Biological and Environmental Statistics*, 26, 23–44. https://doi.org/10.1007/s13253-020-00420-4.

McInerny, G. J. and Purves, D. W. (2011), "Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice," *Methods in Ecology and Evolution*, 2, 248–257. https://doi.org/10.1111/j.2041-210X.2010.00077.x.

Mertes, K. and Jetz, W. (2018), "Disentangling scale dependencies in species environmental niches and distributions," *Ecography*, 41, 1604–1615. https://doi.org/10.1111/ecog.02871.

Miller, J. R., Turner, M. G., Smithwick, E. A., Dent, C. L., and Stanley, E. H. (2004), "Spatial extrapolation: the science of predicting ecological patterns and processes," *BioScience*, 54, 310–320. https://doi.org/10.1641/0006-3568(2004)054[0310:SETSOP]2.0.CO;2.

Mourguiart, B., Chevalier, M., Marzloff, M., Caill-Milly, N., Mengersen, K., and Liquet, B. (2024), "Dealing with area-to-point spatial misalignment in species distribution models," *Ecography*, e07104. https://doi.org/10.1111/ecog.07104.

Nguyen, H., Cressie, N., and Braverman, A. (2012), "Spatial statistical data fusion for remote sensing applications," *Journal of the American Statistical Association*, 107, 1004–1018. https://doi.org/10.1080/01621459.2012.694717.

Ovaskainen, O. and Soininen, J. (2011), "Making more out of sparse data: hierarchical modeling of species communities," *Ecology*, 92, 289–295. https://doi.org/10.1890/10-1251.1.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., and Thuiller, W. (2021), "On the interpretations of joint modeling in community ecology," *Trends in Ecology & Evolution*, 36, 391–401. https://doi.org/10.1016/j.tree.2021.01.002.

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

Robinson, S. A., King, D. H., Bramley-Alves, J., Waterman, M. J., Ashcroft, M. B., Wasley, J., Turnbull, J. D., Miller, R. E., Ryan-Colton, E., Benny, T., Mullany, K., Clarke, L. J., Barry, L. A., and Hua, Q. (2018), "Rapid change in East Antarctic terrestrial vegetation in response to regional drying," *Nature Climate Change*, 8, 879–884. https://doi.org/10.1038/s41558-018-0280-0.

Sainsbury-Dale, M., Zammit-Mangion, A., and Cressie, N. (2024), "Modeling big, heterogeneous, non-Gaussian spatial and spatio-temporal data using FRK," *Journal of Statistical Software*, 108, 1–39. https://doi.org/10.18637/jss.v108.i10.

Stoklosa, J., Daly, C., Foster, S. D., Ashcroft, M. B., and Warton, D. I. (2015), "A climate of uncertainty: accounting for error in climate variables for species distribution models," *Methods in Ecology and Evolution*, 6, 412–423. https://doi.org/10.1111/2041-210X.12217.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. (2015), "So many variables: joint modeling in community ecology," *Trends in Ecology & Evolution*, 30, 766–779. https://doi.org/10.1016/j.tree.2015.09.007.

Werner, A., Schnorbus, M., Shrestha, R., Cannon, A., Zwiers, F., Dayon, G., and Anslow, F. (2019), "A long-term, temporally consistent, gridded daily meteorological dataset for northwestern North America," *Scientific Data*, 6, 1–16. https://doi.org/10.1038/sdata.2018.299.

Wilby, R. L., Charles, S. P., Zorita, E., Timbal, B., Whetton, P., and Mearns, L. O. (2004), "Guidelines for use of climate scenarios developed from statistical downscaling methods," In Supporting Material of the Intergovernmental Panel on Climate Change, available from the DDC of IPCC TGCIA 27, URL: https://www.ipcc-data.org/guidelines/.

Wood, S. (2017), *Generalized Additive Models: An Introduction with R (2nd ed.)*, Chapman and Hall/CRC, New York, NY. https://doi.org/10.1201/9781315370279.

Wood, S. N. (2003), "Thin plate regression splines," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65, 95–114. https://doi.org/10.1111/1467-9868.00374.

Wu, J. and Li, H. (2006), "Concepts of scale and scaling," in Wu, J., Jones, K. B., Li, H., and Loucks, O. L. (eds), *Scaling and Uncertainty Analysis in Ecology: Methods and Applications* (pp. 3–15), Springer, Dordrecht, Netherlands. https://doi.org/10.1007/1-4020-4663-4.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M. (2009), *Mixed Effects Models and Extensions in Ecology with R*, Springer, New York, NY. https://doi.org/10.1007/978-0-387-87458-6.

Supporting Information for "Spatial-statistical downscaling with uncertainty quantification in biodiversity modelling"

Xiaotian Zheng[1,2], **Noel Cressie**[1,2], **David A. Clarke**[3,4], **Melodie A. McGeoch**[3,4], **and Andrew Zammit-Mangion**[1,2]

[1] *School of Mathematics and Applied Statistics, University of Wollongong, Australia;*

[2] *Securing Antarctica's Environmental Future, University of Wollongong, Australia;*

[3] *School of Biological Sciences, Monash University, Australia;*

[4] *Securing Antarctica's Environmental Future, Monash University, Australia*

In the Supporting Information, we follow the notation and definitions in Sections 2 and 3 of the main text.

S1 Conditional simulation

Let $[\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi} \,|\, \boldsymbol{X}_B]$ be the posterior distribution of the downscaling parameters (Stage 1) $\{\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}\}$. We generate samples from the distribution of $\boldsymbol{X}_{\boldsymbol{s}}$ given the coarse-resolution data $\boldsymbol{X}_B$, using the following relationship,

$$[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B] = \int [\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}]\, [\gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi} \,|\, \boldsymbol{X}_B]\, \mathrm{d}\gamma_0\, \mathrm{d}\boldsymbol{\gamma}\, \mathrm{d}\boldsymbol{\phi}, \tag{S1.1}$$

where the conditional distribution, $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}]$, is a multivariate Gaussian distribution,

$$N(\boldsymbol{\mu}_{1|2},\, \boldsymbol{\Sigma}_{1|2}), \tag{S1.2}$$

where $\boldsymbol{\mu}_{1|2} \equiv E(\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi})$ and $\boldsymbol{\Sigma}_{1|2} \equiv \mathrm{var}(\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi})$ are, respectively, the conditional mean and conditional covariance.

From a set of K posterior samples $\{\gamma_0^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\phi}^{(k)} : k = 1, \ldots, K\}$, we obtain a collection of samples of size K from $[\boldsymbol{X}_s \,|\, \boldsymbol{X}_B]$, the predictive distribution of $\boldsymbol{X}_{\boldsymbol{s}}$ given $\boldsymbol{X}_B$, as follows: The k-th predictive sample of $\boldsymbol{X}_{\boldsymbol{s}}$, denoted by $\boldsymbol{x}_{\boldsymbol{s}}^{(k)} \equiv (x^{(k)}(\boldsymbol{s}_1), \ldots, x^{(k)}(\boldsymbol{s}_{N_D}))^\top$ is generated from (S1.2), where posterior samples $\{\gamma_0^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\phi}^{(k)}\}$ replace $\{\gamma_0, \boldsymbol{\gamma}, \phi\}$ in (S1.2), for $k = 1, \ldots, K$. The k-th sample's conditional mean $\boldsymbol{\mu}_{1|2}^{(k)}$ and the conditional covariance $\boldsymbol{\Sigma}_{1|2}^{(k)}$, which depend on

Algorithm 1: Spatial statistical downscaling

Input: Coarse-resolution covariate data $\boldsymbol{X}_B \equiv (X(B_1), \ldots, X(B_M))^\top$

1. Fit the GPB model with likelihood $[\boldsymbol{X}_B \,|\, \gamma_0, \boldsymbol{\gamma}, \boldsymbol{\phi}]$ and priors $[\gamma_0]\,[\boldsymbol{\gamma}]\,[\boldsymbol{\phi}]$, and draw posterior samples $\{\gamma_0^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\phi}^{(k)} : k = 1, \ldots, K\}$ using MCMC (see Section S4).

2. Simulate predictive samples from $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \boldsymbol{X}_B]$.

for $k \leftarrow 1$ ***to*** K **do**

$$\boldsymbol{\mu}_{\boldsymbol{s}}^{(k)} = \mathbf{1}_{N_D}\gamma_0^{(k)} + \boldsymbol{R}\boldsymbol{\gamma}^{(k)},$$
$$\boldsymbol{\mu}_{1|2}^{(k)} = \boldsymbol{\mu}_{\boldsymbol{s}}^{(k)} + \boldsymbol{C}_{\boldsymbol{\phi}^{(k)}}\boldsymbol{W}_B^\top(\boldsymbol{W}_B\boldsymbol{C}_{\boldsymbol{\phi}^{(k)}}\boldsymbol{W}_B^\top)^{-1}(\boldsymbol{X}_B - \boldsymbol{W}_B\boldsymbol{\mu}_{\boldsymbol{s}}^{(k)}),$$
$$\boldsymbol{\Sigma}_{1|2}^{(k)} = \boldsymbol{C}_{\boldsymbol{\phi}^{(k)}} - \boldsymbol{C}_{\boldsymbol{\phi}^{(k)}}\boldsymbol{W}_B^\top(\boldsymbol{W}_B\boldsymbol{C}_{\boldsymbol{\phi}^{(k)}}\boldsymbol{W}_B^\top)^{-1}\boldsymbol{W}_B\boldsymbol{C}_{\boldsymbol{\phi}^{(k)}},$$
Draw $\boldsymbol{x}_{\boldsymbol{s}}^{(k)}$ from $\mathrm{N}(\boldsymbol{\mu}_{1|2}^{(k)}, \boldsymbol{\Sigma}_{1|2}^{(k)})$.

end

Return $\{\boldsymbol{x}_{\boldsymbol{s}}^{(k)} \equiv (x^{(k)}(\boldsymbol{s}_1), \ldots, x^{(k)}(\boldsymbol{s}_{N_D}) : k = 1, \ldots, K\}$.

$\{\gamma_0^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\phi}^{(k)}\}$, are defined in Algorithm 1 below, where the whole procedure of downscaling is described in detail.

S2 Analysis of variance for prediction

For illustrative purposes, let $h_{\boldsymbol{\theta}}(\boldsymbol{s}) \equiv h_{\boldsymbol{\theta}}(X(\boldsymbol{s}))$ be a functional of a single covariate $X(\boldsymbol{s})$ used in prediction of biodiversity patterns, where $\boldsymbol{s} \in \mathcal{D}$. As an example, consider a generalised linear model (GLM) commonly used in species distribution modelling. The linear predictor of the GLM at site $\boldsymbol{s}$ can be written as $h_{\boldsymbol{\theta}}(\boldsymbol{s}) = \beta_0 + \beta_1 X(\boldsymbol{s})$, where $\boldsymbol{\theta} \equiv (\beta_0, \beta_1)^\top$. The uncertainty in predicting biodiversity using $h_{\boldsymbol{\theta}}(\cdot)$ comes from both the uncertainty in estimating the GLM parameters $\boldsymbol{\theta}$ (Stage 2) and the uncertainty in predicting $X(\cdot)$ from the coarse-resolution data $\boldsymbol{X}_B$ (Stage 1).

Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}$. Conditional on $X(\cdot)$, the variance of $h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})$ is $\mathrm{var}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, X(\cdot))$, where the variance is taken with respect to $[\hat{\boldsymbol{\theta}} \,|\, X(\cdot)]$, the sampling distribution of $\hat{\boldsymbol{\theta}}$ conditional on $X(\cdot)$. The plug-in approach estimates the prediction variance of $h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})$ using $\mathrm{var}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, X(\cdot))$ by replacing $X(\cdot)$ with its prediction $\hat{X}(\cdot)$. However, this approach ignores the uncertainty of downscaling to $X(\cdot)$ at the BAU level from the given $\boldsymbol{X}_B$ defined on coarse-resolution blocks. To see this, we marginalise out $X(\cdot)$ and obtain the predictive distribution

of $h_{\hat{\boldsymbol{\theta}}}(\cdot)$ conditional on $\boldsymbol{X}_B$, namely,

$$[h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, \boldsymbol{X}_B] = \int [h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, X(\cdot)]\, [X(\cdot) \,|\, \boldsymbol{X}_B]\, \mathrm{d}X(\cdot), \tag{S2.1}$$

where the distribution on the left-hand side in (S2.1) is over the (conditional) uncertainty in $\hat{\boldsymbol{\theta}}$, and $[X(\cdot) \,|\, \boldsymbol{X}_B]$ is the predictive distribution of $X(\cdot)$ conditional on $\boldsymbol{X}_B$. Therefore, taking into consideration the downscaling uncertainty, yields the predictive variance of $h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})$:

$$\mathrm{var}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, \boldsymbol{X}_B) = E[\mathrm{var}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, \boldsymbol{X}(\boldsymbol{s})) \,|\, \boldsymbol{X}_B] + \mathrm{var}[E(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}) \,|\, \boldsymbol{X}(\boldsymbol{s})) \,|\, \boldsymbol{X}_B]. \tag{S2.2}$$

Since the plug-in approach ignores downscaling uncertainty in predicting $\hat{X}(\cdot)$, it does not account for the second term in (S2.2), and thus it may lead to underestimation of the prediction variance of $h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})$.

S3 The two-stage protocol

In Section S3.1, we first provide details of the two-stage protocol that considers the general case where response data are limited to certain BAU locations. In Section S3.2, we illustrate how bootstrap samples obtained from the protocol are used for making inferences on regression coefficients and for biodiversity prediction.

S3.1 Details of the protocol

Recall that the resolution of the BAUs corresponds to the finest resolution of scientific interest. In practice, biodiversity data may not be collected at all BAUs in the spatial domain. We order the BAUs such that $\{\boldsymbol{v}_i \in \mathcal{D} : i = 1, \ldots, n \leq N_D\}$ denotes the set of BAU locations at which biodiversity data were collected. When $n = N_D$, we have the two-stage protocol presented in Section 3.2 of the main text.

Associated with these n BAU locations we have an n-dimensional vector of univariate responses $\boldsymbol{Z}_{\boldsymbol{v}} \equiv (Z(\boldsymbol{v}_1), \ldots, Z(\boldsymbol{v}_n))^\top$, and define the n-dimensional vector of unavailable univariate covariate data, $\boldsymbol{X}_{\boldsymbol{v}} \equiv (X(\boldsymbol{v}_1), \ldots, X(\boldsymbol{v}_n))^\top$. Conditional on $\boldsymbol{X}_B$, we have the following decomposition in distribution: $\boldsymbol{X}_{\boldsymbol{v}} = \hat{\boldsymbol{X}}_{\boldsymbol{v}} + \boldsymbol{\eta}_{\boldsymbol{v}}$, where $\hat{\boldsymbol{X}}_{\boldsymbol{v}} \equiv E(\boldsymbol{X}_{\boldsymbol{v}} \,|\, \boldsymbol{X}_B)$ and $\boldsymbol{\eta}_{\boldsymbol{v}} \equiv (\eta(\boldsymbol{v}_1), \ldots, \eta(\boldsymbol{v}_n))^\top$.

We consider now the *GLM-Berkson* model, where $n \leq N_D$. Conditional on $\boldsymbol{X}_B$,

$$\begin{aligned}[\boldsymbol{Z}_{\boldsymbol{v}} \,|\, \boldsymbol{Y}_{\boldsymbol{v}}] &= \prod_{i=1}^{n} f(Z(\boldsymbol{v}_i) \,|\, Y(\boldsymbol{v}_i)), \\ g(\boldsymbol{Y}_{\boldsymbol{v}}) &= \beta_0 \mathbf{1}_n + \beta_1 \hat{\boldsymbol{X}}_{\boldsymbol{v}} + \boldsymbol{\delta}_{\boldsymbol{v}},\end{aligned} \tag{S3.1}$$

in distribution, for some link function $g(\cdot)$; $\boldsymbol{Y}_{\boldsymbol{v}} \equiv (Y(\boldsymbol{v}_1), \ldots, Y(\boldsymbol{v}_n))^\top$; $\boldsymbol{\delta}_{\boldsymbol{v}} \equiv (\delta(\boldsymbol{v}_1), \ldots, \delta(\boldsymbol{v}_n))^\top$; and $\boldsymbol{\theta} \equiv (\beta_0, \beta_1)^\top$, where $\delta(\boldsymbol{v}_i) \equiv \beta_1 \eta(\boldsymbol{v}_i)$, $i = 1, \ldots, n$. The density $f(\cdot \,|\, \cdot)$ in (S3.1) belongs to an exponential family with mean parameter $Y(\cdot)$ such that $E(Z(\boldsymbol{v}_i) \,|\, Y(\boldsymbol{v}_i)) = Y(\boldsymbol{v}_i)$, for $i = 1, \ldots, n$.

GLM-plugin is simpler than (S3.1): Conditional on $\boldsymbol{X}_B$, it is given by

$$\begin{aligned}[\boldsymbol{Z}_{\boldsymbol{v}} \,|\, \boldsymbol{Y}_{\boldsymbol{v}}] &= \prod_{i=1}^{n} f(Z(\boldsymbol{v}_i) \,|\, Y(\boldsymbol{v}_i)), \\ g(\boldsymbol{Y}_{\boldsymbol{v}}) &= \beta_0 \mathbf{1}_n + \beta_1 \hat{\boldsymbol{X}}_{\boldsymbol{v}},\end{aligned} \tag{S3.2}$$

where $\hat{\boldsymbol{X}}_{\boldsymbol{v}}$ is the plug-in estimate of the unobservable $\boldsymbol{X}_{\boldsymbol{v}}$. If $\boldsymbol{X}_{\boldsymbol{v}}$ were known, it would be used in the following GLM,

$$\begin{aligned}[\boldsymbol{Z}_{\boldsymbol{v}} \,|\, \boldsymbol{Y}_{\boldsymbol{v}}] &= \prod_{i=1}^{n} f(Z(\boldsymbol{v}_i) \,|\, Y(\boldsymbol{v}_i)), \\ g(\boldsymbol{Y}_{\boldsymbol{v}}) &= \beta_0 \mathbf{1}_n + \beta_1 \boldsymbol{X}_{\boldsymbol{v}},\end{aligned} \tag{S3.3}$$

which is what we refer to as *GLM-oracle*, the 'gold standard', to which all downscaled results should be compared in simulation studies.

The two-stage protocol for response-data available over $\{\boldsymbol{v}_i \in \mathcal{D} : i = 1, \ldots, n \leq N_D\}$ and covariate data $\boldsymbol{X}_B$ available over blocks of BUAs is given as follows.

Stage 1. Spatial-statistical downscaling with uncertainty

Given coarse-resolution covariate data $\boldsymbol{X}_B$, use the downscaling algorithm (Algorithm 1) to generate predictive samples $\{\boldsymbol{x}_{\boldsymbol{s}}^{(k)} : k = 1, \ldots, K\}$ of $\boldsymbol{X}_{\boldsymbol{s}}$ at the BAU resolution. Any downscaling parameters are integrated out leaving only the downscaled covariate.

Stage 2. Propagating downscaling uncertainty to GLMs

(i) Compute the average, $\hat{\boldsymbol{x}}_{\boldsymbol{s}} \equiv K^{-1} \sum_{k=1}^{K} \boldsymbol{x}_{\boldsymbol{s}}^{(k)}$. Given the locations $\{\boldsymbol{v}_i \in \mathcal{D} : i = 1, \ldots, n\}$ at which responses are observed, extract the corresponding downscaling samples $\{\boldsymbol{x}_{\boldsymbol{v}}^{(k)} : k = 1, \ldots, K\}$, where $\boldsymbol{x}_{\boldsymbol{v}}^{(k)} \equiv (x^{(k)}(\boldsymbol{v}_1), \ldots, x^{(k)}(\boldsymbol{v}_n))^\top$, as well as the average $\hat{\boldsymbol{x}}_{\boldsymbol{v}} \equiv (\hat{x}(\boldsymbol{v}_1), \ldots, \hat{x}(\boldsymbol{v}_n))^\top$. Compute $\{\hat{\boldsymbol{\eta}}_{\boldsymbol{v}}^{(k)} \equiv \boldsymbol{x}_{\boldsymbol{v}}^{(k)} - \hat{\boldsymbol{x}}_{\boldsymbol{v}} : k = 1, \ldots, K\}$, where

$\hat{\boldsymbol{\eta}}_{\boldsymbol{v}}^{(k)} \equiv (\hat{\eta}^{(k)}(\boldsymbol{v}_1), \ldots, \hat{\eta}^{(k)}(\boldsymbol{v}_n))^\top$.

(ii) Fit the GLM-plugin model (S3.2) to $\boldsymbol{Z}_{\boldsymbol{v}}$ with $\hat{\boldsymbol{X}}_{\boldsymbol{v}} = \hat{\boldsymbol{x}}_{\boldsymbol{v}}$, and obtain an estimate of β_1, denoted as $\tilde{\beta}_1$.

(iii) Fit the GLM-Berkson model (S3.1) to $\boldsymbol{Z}_{\boldsymbol{v}}$ with $\hat{\boldsymbol{X}}_{\boldsymbol{v}} = \hat{\boldsymbol{x}}_{\boldsymbol{v}}$ and $\boldsymbol{\delta}_{\boldsymbol{v}}$ replaced by $\hat{\boldsymbol{\delta}}_{\boldsymbol{v}}^{(k)} = \tilde{\beta}_1 \hat{\boldsymbol{\eta}}_{\boldsymbol{v}}^{(k)}$, and obtain an estimate of $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}^{(k)}$.

(iv) For $l = 1, \ldots, L$, simulate bootstrap samples $\boldsymbol{Z}_{\boldsymbol{v}}^{(k,l)}$ from the GLM-Berkson model by taking $\hat{\boldsymbol{X}}_{\boldsymbol{v}} = \hat{\boldsymbol{x}}_{\boldsymbol{v}}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, replacing $\boldsymbol{\delta}_{\boldsymbol{v}}$ with $\hat{\boldsymbol{\delta}}_{\boldsymbol{v}}^{(k)} = \tilde{\beta}_1 \hat{\boldsymbol{\eta}}_{\boldsymbol{v}}^{(k)}$, and then fitting the GLM-Berkson model to $\boldsymbol{Z}_{\boldsymbol{v}}^{(k,l)}$ with $\hat{\boldsymbol{X}}_{\boldsymbol{v}} = \hat{\boldsymbol{x}}_{\boldsymbol{v}}$ and $\boldsymbol{\delta}_{\boldsymbol{v}}$ replaced by $\hat{\boldsymbol{\delta}}_{\boldsymbol{v}}^{(k)} = \tilde{\beta}_1 \hat{\boldsymbol{\eta}}_{\boldsymbol{v}}^{(k)}$. From the fit, obtain an estimate of $\boldsymbol{\theta}$, denoted as $\hat{\boldsymbol{\theta}}^{(k,l)}$.

(v) Repeat steps (iii) and (iv) for $k = 1, \ldots, K$, resulting in $\{\hat{\boldsymbol{\theta}}^{(k,l)} : k = 1, \ldots, K, l = 1, \ldots, L\}$.

In the next section, we discuss how to use this collection of parameter estimates for inference and prediction.

S3.2 Sample-based inference and prediction

Let $\hat{\boldsymbol{\theta}}$ be the estimator of $\boldsymbol{\theta}$ of the GLM-Berkson model (S3.1). Using the collection $\{\hat{\boldsymbol{\theta}}^{(k,l)} : k = 1, \ldots, K, l = 1, \ldots, L\}$ obtained from the protocol and the bootstrap, we estimate $E(\hat{\boldsymbol{\theta}})$ and $\mathrm{var}(\hat{\boldsymbol{\theta}})$ as

$$\widehat{E}(\hat{\boldsymbol{\theta}}) \equiv (KL)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{L} \hat{\boldsymbol{\theta}}^{(k,l)},$$

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\theta}}) \equiv (KL-1)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{L} (\hat{\boldsymbol{\theta}}^{(k,l)} - \hat{E}(\hat{\boldsymbol{\theta}}))^2.$$

Sorting the samples $\{\hat{\boldsymbol{\theta}}^{(k,l)} : k = 1, \ldots, K, l = 1, \ldots, L\}$ in ascending order, we can estimate any quantile (or percentile) of $\hat{\boldsymbol{\theta}}$ with respect to its sampling distribution (approximated by $\{\hat{\boldsymbol{\theta}}^{(k,l)}\}$) to construct its interval estimates.

Regarding prediction, for illustration purposes, we consider a simple linear regression given by $h_{\boldsymbol{\theta}}(\boldsymbol{s}) = \beta_0 + \beta_1 X(\boldsymbol{s})$, for $\boldsymbol{s} \in \mathcal{D}$, where $\boldsymbol{\theta} \equiv (\beta_0, \beta_1)^\top$. Let $\{x^{(k)}(\boldsymbol{s}) : k = 1, \ldots, K\}$ be the collection of predictive samples of $X(\boldsymbol{s})$. We obtain samples of $h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})$ by calculating

$$h_{\hat{\boldsymbol{\theta}}}^{(k,l)}(\boldsymbol{s}) \equiv \hat{\beta}_0^{(k,l)} + \hat{\beta}_1^{(k,l)} \hat{x}(\boldsymbol{s}) + \tilde{\beta}_1 \hat{\eta}^{(k)}(\boldsymbol{s}),$$

where $\hat{x}(\boldsymbol{s}) \equiv K^{-1}\sum_{k=1}^{K} x^{(k)}(\boldsymbol{s})$, $\hat{\eta}^{(k)}(\boldsymbol{s}) \equiv x^{(k)}(\boldsymbol{s}) - \hat{x}(\boldsymbol{s})$, and $\hat{\boldsymbol{\theta}}^{(k,l)} \equiv (\hat{\beta}_0^{(k,l)}, \hat{\beta}_1^{(k,l)})^\top$, for $k = 1, \ldots, K$ and $l = 1, \ldots, L$.

Using the collection $\{h_{\hat{\boldsymbol{\theta}}}^{(k,l)}(\boldsymbol{s}) : k = 1, \ldots, K,\, l = 1, \ldots, L\}$, we can estimate $E(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}))$ and $\text{var}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}))$ as

$$\widehat{E}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})) \equiv (KL)^{-1}\sum_{k=1}^{K}\sum_{l=1}^{L} h_{\hat{\boldsymbol{\theta}}}^{(k,l)}(\boldsymbol{s}),$$

$$\widehat{\text{var}}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})) \equiv (KL-1)^{-1}\sum_{k=1}^{K}\sum_{l=1}^{L}(h_{\hat{\boldsymbol{\theta}}}^{(k,l)}\left(\boldsymbol{s}) - \hat{E}(h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s}))\right)^2.$$

Similarly, we can use the collection $\{h_{\hat{\boldsymbol{\theta}}}^{(k,l)}(\boldsymbol{s}) : k = 1, \ldots, K,\, l = 1, \ldots, L\}$ to estimate any quantile (or percentile) of $h_{\hat{\boldsymbol{\theta}}}(\boldsymbol{s})$ and construct its interval estimates.

S4 MCMC algorithm

We introduce the MCMC algorithm for the GPB model (Section 2 of the main text) when used in our COS-based downscaling. As an illustration, we used an exponential spatial covariance function, that is, $\text{cov}(\omega(\boldsymbol{s}), \omega(\boldsymbol{s}')) \equiv C(\boldsymbol{s}, \boldsymbol{s}' \,|\, \boldsymbol{\phi}) = \sigma^2 \exp(-||\boldsymbol{s} - \boldsymbol{s}'||/\psi)$, where σ^2 is the variance parameter, ψ is the range parameter, and $\boldsymbol{\phi} \equiv (\psi, \sigma^2)^\top$. Note that we used exponential covariance functions for all the GPB models implemented in Section 4 of the main text.

Let $\boldsymbol{X}_{\boldsymbol{s}} \equiv (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_{N_D}))^\top$ be the vector that defines the random process at the BAU level. For notational convenience, let $\tilde{\boldsymbol{\gamma}} \equiv (\gamma_0, \boldsymbol{\gamma}^\top)^\top$ be a $(p+1)$-dimensional vector, where γ_0 is the intercept, and $\boldsymbol{\gamma}$ is a p-dimensional vector of regression coefficients for BAU-level covariates. Take $\tilde{\boldsymbol{R}} \equiv [\boldsymbol{1}_{N_D}, \boldsymbol{R}]$, where $\boldsymbol{R}$ is the $N_D \times p$ matrix of covariates. Then according to equation 1 in the main text, we have that $[\boldsymbol{X}_{\boldsymbol{s}} \,|\, \tilde{\boldsymbol{\gamma}}, \boldsymbol{\phi}] = \text{N}(\tilde{\boldsymbol{R}}\tilde{\boldsymbol{\gamma}}, \boldsymbol{C}_{\boldsymbol{\phi}})$, where $\boldsymbol{C}_{\boldsymbol{\phi}} \equiv \sigma^2 \boldsymbol{K}_\psi$ is an $N_D \times N_D$ covariance matrix, and $\boldsymbol{K}_\psi$ is an $N_D \times N_D$ correlation matrix whose (i, i')-th element is $\exp(-||\boldsymbol{s}_i - \boldsymbol{s}_{i'}||/\psi)$, for $i = 1, \ldots, N_D$, $i' = 1, \ldots, N_D$.

Let $\boldsymbol{X}_B \equiv (X(B_1), \ldots, X(B_{M_D}))^\top$ be the M_D-dimensional vector of block values defined by equation 2 in the main text. Let $\boldsymbol{W}_B$ be the aggregation matrix such that $\boldsymbol{X}_B = \boldsymbol{W}_B \boldsymbol{X}_{\boldsymbol{s}}$, where the (j, i)-th entry of $\boldsymbol{W}_B$ is $w_{ji} = \mathbb{1}(\boldsymbol{s}_i \in \mathcal{B}_j)/\sum_{\boldsymbol{s}\in\mathcal{D}} \mathbb{1}(\boldsymbol{s} \in \mathcal{B}_j)$, for $j = 1, \ldots, M_D$, and $i = 1, \ldots, N_D$. It follows that $[\boldsymbol{X}_B \,|\, \tilde{\boldsymbol{\gamma}}, \sigma^2, \psi] = \text{N}(\boldsymbol{W}_B\tilde{\boldsymbol{R}}\tilde{\boldsymbol{\gamma}}, \boldsymbol{W}_B\boldsymbol{C}_{\boldsymbol{\phi}}\boldsymbol{W}_B^\top)$. Take $\boldsymbol{R}_B \equiv \boldsymbol{W}_B\tilde{\boldsymbol{R}}$ and $\boldsymbol{K}_B(\psi) \equiv \boldsymbol{W}_B\boldsymbol{K}_\psi\boldsymbol{W}_B^\top$. Then we can re-write the distribution as,

$$[\boldsymbol{X}_B \,|\, \tilde{\boldsymbol{\gamma}}, \sigma^2, \psi] = \text{N}(\boldsymbol{R}_B\tilde{\boldsymbol{\gamma}}, \sigma^2\boldsymbol{K}_B(\psi)). \tag{S4.1}$$

Then, according to Section 2.2, the posterior distribution of the unknown parameters $\{\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi\}$

is given by $[\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi \,|\, \boldsymbol{X}_B] \propto [\boldsymbol{X}_B \,|\, \tilde{\boldsymbol{\gamma}}, \sigma^2, \psi]\,[\tilde{\boldsymbol{\gamma}}]\,[\sigma^2]\,[\psi]$. For the priors, we let $[\tilde{\boldsymbol{\gamma}}] = \mathrm{N}(\boldsymbol{\mu}_{\tilde{\boldsymbol{\gamma}}}, \boldsymbol{V}_{\tilde{\boldsymbol{\gamma}}})$, $[\sigma^2] = \mathrm{IG}(u_{\sigma^2}, v_{\sigma^2})$, and $[\psi] = \mathrm{Unif}(u_\psi, v_\psi)$, where $\mathrm{IG}(u, v)$ denotes the inverse gamma distribution with mean $v/(u-1)$, and $\mathrm{Unif}(u, v)$ denotes the continuous uniform distribution on the interval $[u, v]$.

Note that the posterior distribution of the unknown parameters $\{\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi\}$ can be factorised as $[\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi \,|\, \boldsymbol{X}_B] = [\tilde{\boldsymbol{\gamma}} \,|\, \sigma^2, \psi, \boldsymbol{X}_B]\,[\sigma^2 \,|\, \psi, \boldsymbol{X}_B]\,[\psi \,|\, \boldsymbol{X}_B]$ (see, e.g., Berger et al. 2001); that is,

$$\begin{aligned} [\tilde{\boldsymbol{\gamma}} \,|\, \sigma^2, \psi, \boldsymbol{X}_B] &= \mathrm{N}(\tilde{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\gamma}}}, \tilde{\boldsymbol{V}}_{\tilde{\boldsymbol{\gamma}}}), \\ [\sigma^2 \,|\, \psi, \boldsymbol{X}_B] &= \mathrm{IG}(\tilde{u}_{\sigma^2}, \tilde{v}_{\sigma^2}), \\ [\psi \,|\, \boldsymbol{X}_B] &\propto |\boldsymbol{K}_B(\psi)|^{-1/2} \cdot |\boldsymbol{R}_B^\top (\boldsymbol{K}_B(\psi))^{-1} \boldsymbol{R}_B|^{-1/2} \cdot \tilde{v}_{\sigma^2}^{-\tilde{u}_{\sigma^2}}, \end{aligned} \tag{S4.2}$$

where $\tilde{\boldsymbol{V}}_{\tilde{\boldsymbol{\gamma}}} \equiv (\boldsymbol{R}_B^\top (\boldsymbol{C}_B(\psi))^{-1} \boldsymbol{R}_B + \boldsymbol{V}_{\tilde{\boldsymbol{\gamma}}}^{-1})^{-1}$, $\boldsymbol{C}_B(\psi) \equiv \sigma^2 \boldsymbol{K}_B(\psi)$, $\tilde{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\gamma}}} \equiv \tilde{\boldsymbol{V}}_{\tilde{\boldsymbol{\gamma}}} (\boldsymbol{R}_B^\top (\boldsymbol{C}_B(\psi))^{-1} \boldsymbol{X}_B + \boldsymbol{V}_{\tilde{\boldsymbol{\gamma}}}^{-1} \boldsymbol{\mu}_{\tilde{\boldsymbol{\gamma}}})$, $S^2(\psi) \equiv (\boldsymbol{X}_B - \boldsymbol{R}_B \tilde{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\gamma}}})^\top (\boldsymbol{K}_B(\psi))^{-1} (\boldsymbol{X}_B - \boldsymbol{R}_B \tilde{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\gamma}}})$, $\tilde{u}_{\sigma^2} \equiv u_{\sigma^2} + (M_D - p - 1)/2$, and $\tilde{v}_{\sigma^2} \equiv v_{\sigma^2} + S^2(\psi)/2$.

We sample the unknown parameters as a block using a Metropolis-Hasting (MH) algorithm, which is a commonly-used MCMC method. We denote the proposal distribution of the MH algorithm as $[\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi]_P = [\tilde{\boldsymbol{\gamma}} \,|\, \sigma^2, \psi]_P\,[\sigma^2 \,|\, \psi]_P\,[\psi]_P$, where $[\tilde{\boldsymbol{\gamma}} \,|\, \sigma^2, \psi]_P = \mathrm{N}(\tilde{\boldsymbol{\mu}}_{\tilde{\boldsymbol{\gamma}}}, \tilde{\boldsymbol{V}}_{\tilde{\boldsymbol{\gamma}}})$ and $[\sigma^2 \,|\, \psi]_P = \mathrm{IG}(\tilde{u}_{\sigma^2}, \tilde{v}_{\sigma^2})$ according to (S4.2). It follows that the probability that the sample $\{\tilde{\boldsymbol{\gamma}}', (\sigma^2)', \psi'\}$ drawn from the proposal distribution $[\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi]_P$ is accepted in the t-th MCMC iteration is given by

$$\begin{aligned} \alpha_t &= \min\left\{1, \frac{[\tilde{\boldsymbol{\gamma}}', (\sigma^2)', \psi' \,|\, \boldsymbol{X}_B]\,[\tilde{\boldsymbol{\gamma}}^{(t-1)}, (\sigma^2)^{(t-1)}, \psi^{(t-1)}]_P}{[\tilde{\boldsymbol{\gamma}}^{(t-1)}, (\sigma^2)^{(t-1)}, \psi^{(t-1)} \,|\, \boldsymbol{X}_B]\,[\tilde{\boldsymbol{\gamma}}', (\sigma^2)', \psi']_P}\right\} \\ &= \min\left\{1, \frac{[\psi' \,|\, \boldsymbol{X}_B]\,[\psi^{(t-1)}]_P}{[\psi^{(t-1)} \,|\, \boldsymbol{X}_B]\,[\psi']_P}\}\right\}, \end{aligned} \tag{S4.3}$$

where $\{\tilde{\boldsymbol{\gamma}}^{(t)}, (\sigma^2)^{(t)}, \psi^{(t)}\}$ denotes a sample of $\{\tilde{\boldsymbol{\gamma}}, \sigma^2, \psi\}$ at the t-th MCMC iteration.

In summary, we sample the unknown parameters as follows. Let T be the total number of MCMC iterations. For $t = 1, \ldots, T$,

(i) Generate a proposal ψ' from $[\psi]_P$.

(ii) Calculate the acceptance probability α_t according to (S4.3).

(iii) Generate a random number u from $\mathrm{Unif}(0, 1)$.
If $u \leq \alpha_t$, set $\psi^{(t)} = \psi'$; then according to (S4.2), sample $(\sigma^2)^{(t)}$ from $[\sigma^2 \,|\, \psi^{(t)}, \boldsymbol{X}_B]$, followed by $\tilde{\boldsymbol{\gamma}}^{(t)}$ from $[\tilde{\boldsymbol{\gamma}} \,|\, (\sigma^2)^{(t)}, \psi^{(t)}, \boldsymbol{X}_B]$.

If $u > \alpha_t$, set the t-th iterate to be unchanged: $\{\tilde{\boldsymbol{\gamma}}^{(t)}, (\sigma^2)^{(t)}, \psi^{(t)}\} = \{\tilde{\boldsymbol{\gamma}}^{(t-1)}, (\sigma^2)^{(t-1)}, \psi^{(t-1)}\}$,

where $\{\tilde{\boldsymbol{\gamma}}^{(0)}, (\sigma^2)^{(0)}, \psi^{(0)}\}$ are specified initial values for the first iteration $t = 1$. As a result of the MCMC iterations, we obtain posterior samples $\{\tilde{\boldsymbol{\gamma}}^{(t)}, (\sigma^2)^{(t)}, \psi^{(t)} : t = 1, \ldots, T\}$, which will be thinned and used to generate K predictive samples of $\boldsymbol{X}_{\boldsymbol{s}}$ given $\boldsymbol{X}_B$ based on (S1.1) via conditional simulation.

S5 Implementation details

Here, we provide important implementation details for Section 4 of the main text. We used an isotropic exponential covariance function, $\sigma^2 \exp(-d/\psi)$, for $d \geq 0$, with variance parameter σ^2 and range parameter ψ for all GPB models, with priors $[\sigma^2, \psi] = [\sigma^2][\psi]$, where $[\sigma^2] = \text{IG}(u_{\sigma^2}, v_{\sigma^2})$ and $[\psi] = \text{Unif}(u_\psi, v_\psi)$. To update ψ, in Step (i) of the MCMC algorithm in Section S4, we consider a random-walk Metropolis step with a normal proposal distribution on a generalised-logit (glogit) scale (u_ψ, v_ψ) where, for $y \in (u_\psi, v_\psi)$, $\text{glogit}(y) = \log((y-u_\psi)/(v_\psi - y))$. Results were based on 500 posterior samples, obtained as follows: We discarded the first 3000 samples as burn-in and then retained one in every 10 iterations to account for correlated iterates.

S5.1 First simulation study

We generated data from a mean-zero Gaussian process (GP) with an exponential covariance function such that the covariance between two sites $\boldsymbol{s}$ and $\boldsymbol{s}'$ is $\sigma_0^2 \exp(-||\boldsymbol{s} - \boldsymbol{s}'||/\phi_0)$, where $\sigma_0^2 = 1$ and $\phi_0 = 1$. When downscaling the covariate data using the GPB, we used the priors $[\gamma_0, \boldsymbol{\gamma}] \propto 1$, $[\sigma^2] = \text{IG}(3, 2)$, $[\psi] = \text{Unif}(0.03, 1)$, and the likelihood given by equation 5 in the main text. For TPS, the basis dimension was 12; it was chosen such that the model achieved the best empirical predictive coverage probability, which was used to quantify its uncertainty and compared to the GPB.

S5.2 Second simulation study

Simulation and model-fitting

We simulated the fine-resolution covariate field $\boldsymbol{X}_{\boldsymbol{s}} \equiv (X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_{N_D}))^\top$ over a regular grid of $N_D = 3000$ rectangular BAUs, using a mean-zero GP with an exponential covariance function with parameters $\boldsymbol{\phi} = (\psi_0, \sigma_0^2)^\top$, where ψ_0 is the spatial range and σ_0^2 is the variance. We considered eight different values of the spatial range $\psi_0 \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$

and five different values of the variance $\sigma_0^2 \in \{1, 2, 3, 4, 5\}$, resulting in $8 \times 5 = 40$ different covariate fields (i.e., 40 scenarios).

For each of the 40 covariate fields, we repeatedly simulated response data 100 times from the following Poisson GLM,

$$\begin{aligned} [Z(\boldsymbol{s}_i) \,|\, Y(\boldsymbol{s}_i)] &= \text{Poisson}(Y(\boldsymbol{s}_i)), \\ \log(Y(\boldsymbol{s}_i)) &= \beta_0 + \beta_1 X(\boldsymbol{s}_i), \end{aligned} \tag{S5.1}$$

for $i = 1, \ldots, N_D$, where $(\beta_0, \beta_1)^\top = (1, 0.5)^\top$. After we simulated the response data, we aggregated the covariate field to 30 blocks at coarse resolutions, where each block consists of 100 BAUs. Thus, for each scenario, we had $C = 100$ replicates of data sets: $\{\boldsymbol{Z}_{\boldsymbol{s}}^{(c)}, \boldsymbol{X}_B\}$, where $\boldsymbol{Z}_{\boldsymbol{s}}^{(c)} \equiv (Z^{(c)}(\boldsymbol{s}_1), \ldots, Z^{(c)}(\boldsymbol{s}_{N_D}))^\top$ is the c-th replicate of response data, for $c = 1, \ldots, C$.

Next, for each of the 40 scenarios, we downscaled the block-valued covariate data $\boldsymbol{X}_B$ using the GPB with priors $[\gamma_0, \boldsymbol{\gamma}] \propto 1$, $[\sigma^2] = \text{IG}(3, 2)$, $[\psi] = \text{Unif}(0.03, 1)$, and obtained downscaled samples $\{\boldsymbol{x}_{\boldsymbol{s}}^{(k)} : k = 1, \ldots, K\}$, where $K = 500$. Then, for each scenario, we had 100 replicates of data sets at the fine resolution: $\{\boldsymbol{Z}_{\boldsymbol{s}}^{(c)}, \{\boldsymbol{x}_{\boldsymbol{s}}^{(k)} : k = 1, \ldots, K\}\}$, for $c = 1, \ldots, C$.

We then randomly selected 600 BAU locations to associate with the training data; the responses on the remaining 2400 locations were used as test data. Note that the same set of 600 BAU locations were used for all 40 scenarios.

For each of the 40 scenarios, let $\boldsymbol{Z}_{\boldsymbol{v}}^{(c)} \equiv (Z^{(c)}(\boldsymbol{v}_1), \ldots, Z^{(c)}(\boldsymbol{v}_n))^\top$ be c-th replicate of the response data corresponding to the randomly-selected locations $\{\boldsymbol{v}_i \in \mathcal{D} : i = 1, \ldots, n\}$, where $n = 600$. Denote by $\{\boldsymbol{x}_{\boldsymbol{v}}^{(k)} : k = 1, \ldots, K\}$ the downscaled samples corresponding to those locations, $\hat{\boldsymbol{x}}_{\boldsymbol{v}} \equiv (\hat{x}(\boldsymbol{v}_1), \ldots, \hat{x}(\boldsymbol{v}_n))^\top$ the average, and $\{\hat{\boldsymbol{\eta}}_{\boldsymbol{v}}^{(k)} \equiv \boldsymbol{x}_{\boldsymbol{v}}^{(k)} - \hat{\boldsymbol{x}}_{\boldsymbol{v}} : k = 1, \ldots, K\}$ the downscaling-error samples. Thus , for each scenario, we had 100 replicates of training data $\{\boldsymbol{Z}_{\boldsymbol{v}}^{(c)}, \{\boldsymbol{x}_{\boldsymbol{v}}^{(k)} : k = 1, \ldots, K\}\}$, for $c = 1, \ldots, C$.

Then, for each of the 40 scenarios and for each of the $C = 100$ replicates, we used $K = 500$ downscaled samples of the covariate to fit the GLM-Berkson and the GLM-ensemble models, used the mean of the downscaled samples of the covariate to fit the GLM-plugin model, and used the true simulated covariate to fit the GLM-oracle model. When fitting the models, we performed parametric bootstraps to quantify parameter uncertainty.

- GLM-Berkson: The model in (S3.1) with $f(\cdot \,|\, \cdot)$ in the Poisson family, following the two-stage protocol in Section S3.1. We obtained a collection of parameter estimates $\left\{\hat{\beta}_0^{(c,k,l),Berkson}, \hat{\beta}_1^{(c,k,l),Berkson}\right\}$, $c = 1, \ldots, C$, $k = 1, \ldots, K$, $l = 1, \ldots, L$.
- GLM-plugin: The model in (S3.2) with $f(\cdot \,|\, \cdot)$ in the Poisson family and $\hat{\boldsymbol{X}}_{\boldsymbol{v}} = \hat{\boldsymbol{x}}_{\boldsymbol{v}}$.

We obtained a collection of parameter estimates $\left\{\hat{\beta}_0^{(c,l),plugin}, \hat{\beta}_1^{(c,l),plugin}\right\}$, $c = 1, \ldots, C$, $l = 1, \ldots, L$.

- GLM-ensemble: The model in (S3.3) with $f(\cdot \,|\, \cdot)$ in the Poisson family and $\boldsymbol{X_v} = \boldsymbol{x_v}^{(k)}$, fitted $k = 1, \ldots, K$ times. We obtained a collection of parameter estimates $\left\{\hat{\beta}_0^{(c,k,l),ensemble}, \hat{\beta}_1^{(c,k,l),ensemble}\right\}$, $c = 1, \ldots, C$, $k = 1, \ldots, K$, $l = 1, \ldots, L$.
- GLM-oracle: The model in (S3.3) with $f(\cdot \,|\, \cdot)$ in the Poisson family and $\boldsymbol{X_v}$ obtained from the true simulated fine-resolution covariate field $\boldsymbol{X_s}$. We obtained a collection of parameter estimates $\left\{\hat{\beta}_0^{(c,l),oracle}, \hat{\beta}_1^{(c,l),oracle}\right\}$, $c = 1, \ldots, C$, $l = 1, \ldots, L$.

Model evaluation

Recall that we considered four metrics in the main text: (i) bias, $E(\hat{\beta}_1) - \beta_1$; (ii) 95% coverage probability (CP) for β_1; (iii) out-of-sample root mean square predictive error (RMSPE) for the linear predictor; (iv) out-of-sample 95% CP for the linear predictor.

Bias: We estimated the bias as follows:

$$\begin{aligned}
\widehat{\text{bias}}^{Berkson} &\equiv (CKL)^{-1} \sum_{c=1}^{C} \sum_{k=1}^{K} \sum_{l=1}^{L} \hat{\beta}_1^{(c,k,l),Berkson} - \beta_1, \\
\widehat{\text{bias}}^{plugin} &\equiv (CL)^{-1} \sum_{c=1}^{C} \sum_{l=1}^{L} \hat{\beta}_1^{(c,l),plugin} - \beta_1, \\
\widehat{\text{bias}}^{ensemble} &\equiv (CKL)^{-1} \sum_{c=1}^{C} \sum_{k=1}^{K} \sum_{l=1}^{L} \hat{\beta}_1^{(c,k,l),ensemble} - \beta_1, \\
\widehat{\text{bias}}^{oracle} &\equiv (CL)^{-1} \sum_{c=1}^{C} \sum_{l=1}^{L} \hat{\beta}_1^{(c,l),oracle} - \beta_1,
\end{aligned}$$

where $\beta_1 = 0.5$.

95% coverage probability for $\boldsymbol{\beta_1}$: For each replicate and each model, we first used the collection of estimates of β_1 to calculate 2.5-th and 97.5-th percentiles; for example, for the c-th replicate, the collection of estimates of β_1 from the GLM-Berkson model is $\{\hat{\beta}_1^{(c,k,l),Berkson} : k = 1, \ldots, K, l = 1, \ldots, L\}$, which can used to calculate its 2.5-th and 97.5-th percentiles. Then, for each replicate and each model, we obtained an indicator equal to 1 if the interval formed by the two percentiles covered β_1, and equal to 0 otherwise.

Denote by $a_{est}^{(c),Berkson}$, $a_{est}^{(c),plugin}$, $a_{est}^{(c),ensemble}$, and $a_{est}^{(c),oracle}$ the indicators for the four models and for the c-th replicate, $c = 1, \ldots, C$. We estimated the 95% coverage probability for β_1 as follows:

$$\widehat{\mathrm{CP}}_{est,95}^{Berkson} \equiv C^{-1} \sum_{c=1}^{C} a_{est}^{(c),Berkson}, \quad \widehat{\mathrm{CP}}_{est,95}^{plugin} \equiv C^{-1} \sum_{c=1}^{C} a_{est}^{(c),plugin},$$
$$\widehat{\mathrm{CP}}_{est,95}^{ensemble} \equiv C^{-1} \sum_{c=1}^{C} a_{est}^{(c),ensemble}, \quad \widehat{\mathrm{CP}}_{est,95}^{oracle} \equiv C^{-1} \sum_{c=1}^{C} a_{est}^{(c),oracle}.$$

RMSPE for the linear predictor: Let $\{\boldsymbol{u}_i \in \mathcal{D} : i = 1, \ldots, n_{\text{test}}\}$ be the set of testing locations; in our simulation study, $n_{\text{test}} = 2400$. For each testing location, we have that

$$\begin{aligned}
h^{(c,k,l),Berkson}(\boldsymbol{u}_i) &\equiv \hat{\beta}_0^{(c,k,l),Berkson} + \hat{\beta}_1^{(c,k,l),Berkson} \hat{x}(\boldsymbol{u}_i) + \tilde{\beta}_1^{(c)} \hat{\eta}^{(k)}(\boldsymbol{u}_i), \\
h^{(c,l),plugin}(\boldsymbol{u}_i) &\equiv \hat{\beta}_0^{(c,l),plugin} + \hat{\beta}_1^{(c,l),plugin} \hat{x}(\boldsymbol{u}_i), \\
h^{(c,k,l),ensemble}(\boldsymbol{u}_i) &\equiv \hat{\beta}_0^{(c,k,l),ensemble} + \hat{\beta}_1^{(c,k,l),ensemble} x^{(k)}(\boldsymbol{u}_i), \\
h^{(c,l),oracle}(\boldsymbol{u}_i) &\equiv \hat{\beta}_0^{(c,l),oracle} + \hat{\beta}_1^{(c,l),oracle} X(\boldsymbol{u}_i),
\end{aligned} \tag{S5.2}$$

for $c = 1, \ldots, C$, $K = 1, \ldots, K$, $l = 1, \ldots, L$, where $\tilde{\beta}_1^{(c)}$ is an estimate of β_1 defined in Section S3.1, and $x^{(k)}(\boldsymbol{u}_i)$ and $\hat{\eta}^{(k)}(\boldsymbol{u}_i) \equiv x^{(k)}(\boldsymbol{u}_i) - \hat{x}(\boldsymbol{u}_i)$ are, respectively, the k-th downscaled sample and the k-th downscaling-error sample at location $\boldsymbol{u}_i$, $i = 1, \ldots, n_{\text{test}}$.

Then, for each replicate and each model, we estimated the linear predictor as

$$\hat{g}^{(c),Berkson}(\boldsymbol{u}_i) \equiv (KL)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{L} h^{(c,k,l),Berkson}(\boldsymbol{u}_i), \quad \hat{g}^{(c),plugin}(\boldsymbol{u}_i) \equiv L^{-1} \sum_{l=1}^{L} h^{(c,l),plugin}(\boldsymbol{u}_i),$$
$$\hat{g}^{(c),ensemble}(\boldsymbol{u}_i) \equiv (KL)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{L} h^{(c,k,l),ensemble}(\boldsymbol{u}_i), \quad \hat{g}^{(c),oracle}(\boldsymbol{u}_i) \equiv L^{-1} \sum_{l=1}^{L} h^{(c,l),oracle}(\boldsymbol{u}_i),$$

and we calculated the RMSPE as follows:

$$\text{RMSPE}^{(c),Berkson} \equiv \sqrt{n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \left[\hat{g}^{(c),Berkson}(\boldsymbol{u}_i) - \log(Y(\boldsymbol{u}_i))\right]^2},$$

$$\text{RMSPE}^{(c),plugin} \equiv \sqrt{n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \left[\hat{g}^{(c),plugin}(\boldsymbol{u}_i) - \log(Y(\boldsymbol{u}_i))\right]^2},$$

$$\text{RMSPE}^{(c),ensemble} \equiv \sqrt{n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \left[\hat{g}^{(c),ensemble}(\boldsymbol{u}_i) - \log(Y(\boldsymbol{u}_i))\right]^2},$$

$$\text{RMSPE}^{(c),oracle} \equiv \sqrt{n_{\text{test}}^{-1} \sum_{i=1}^{n_{\text{test}}} \left[\hat{g}^{(c),oracle}(\boldsymbol{u}_i) - \log(Y(\boldsymbol{u}_i))\right]^2},$$

where $\{Y(\boldsymbol{u}_i) : i = 1, \ldots, n_{\text{test}}\}$ are the true responses held aside for testing.

Finally, we obtained the RMSPE for each model as follows:

$$\text{RMSPE}^{Berkson} \equiv C^{-1} \sum_{c=1}^{C} \text{RMSPE}^{(c),Berkson}, \quad \text{RMSPE}^{plugin} \equiv C^{-1} \sum_{c=1}^{C} \text{RMSPE}^{(c),plugin},$$

$$\text{RMSPE}^{ensemble} \equiv C^{-1} \sum_{c=1}^{C} \text{RMSPE}^{(c),ensemble}, \quad \text{RMSPE}^{oracle} \equiv C^{-1} \sum_{c=1}^{C} \text{RMSPE}^{(c),oracle}.$$

95% coverage probability for the linear predictor: Recall that for each replicate and each model, we have a collection of predictions from (S5.2) for each testing location. For example, from the GLM-Berkson model, we have $\{h^{(c,k,l),Berkson}(\boldsymbol{u}_i) : k = 1, \ldots, K, l = 1, \ldots, L\}$, for the c-th replicate, $c = 1, \ldots, C$ and $i = 1, \ldots, n_{\text{test}}$. For each c and each i, we used the collection of predictions to calculate its 2.5-th and 97.5-th percentiles, and obtained an indicator equal to 1 if the interval formed by the two percentiles covered $\log(Y(\boldsymbol{u}_i))$, and equal to zero otherwise. Denote by $a_{pred}^{(c),Berkson}(\boldsymbol{u}_i)$ the indicator for the GLM-Berkson model, for the c-th replicate and for location $\boldsymbol{u}_i$, $c = 1, \ldots, C$, $i = 1, \ldots, n_{\text{test}}$. Similarly, we obtained indicators for the other three models, denoted as $a_{pred}^{(c),plugin}(\boldsymbol{u}_i)$, $a_{pred}^{(c),ensemble}(\boldsymbol{u}_i)$, and $a_{pred}^{(c),oracle}(\boldsymbol{u}_i)$, $c = 1, \ldots, C$, $i = 1, \ldots, n_{\text{test}}$.

Then, for each model, we estimated the 95% coverage probability for the linear predictor,

averaged over all locations, as follows:

$$\widehat{\mathrm{CP}}_{pred,95}^{Berkson} \equiv (Cn_{\text{test}})^{-1} \sum_{c=1}^{C} \sum_{i=1}^{n_{\text{test}}} a_{pred}^{(c),Berkson}(\boldsymbol{u}_i), \quad \widehat{\mathrm{CP}}_{pred,95}^{plugin} \equiv (Cn_{\text{test}})^{-1} \sum_{c=1}^{C} \sum_{i=1}^{n_{\text{test}}} a_{pred}^{(c),plugin}(\boldsymbol{u}_i),$$

$$\widehat{\mathrm{CP}}_{pred,95}^{ensemble} \equiv (Cn_{\text{test}})^{-1} \sum_{c=1}^{C} \sum_{i=1}^{n_{\text{test}}} a_{pred}^{(c),ensemble}(\boldsymbol{u}_i), \quad \widehat{\mathrm{CP}}_{pred,95}^{oracle} \equiv (Cn_{\text{test}})^{-1} \sum_{c=1}^{C} \sum_{i=1}^{n_{\text{test}}} a_{pred}^{(c),oracle}(\boldsymbol{u}_i).$$

S5.3 Case study

We followed the two-stage protocol in Section S3.1. When downscaling, we used priors $[\gamma_0, \boldsymbol{\gamma}] \propto 1$, $[\psi] = \text{Unif}(1000, 21730.67)$, and $[\sigma^2] = \text{IG}(3, 2)$ for the GPB model.

S6 Extensions to account for spatial misalignment

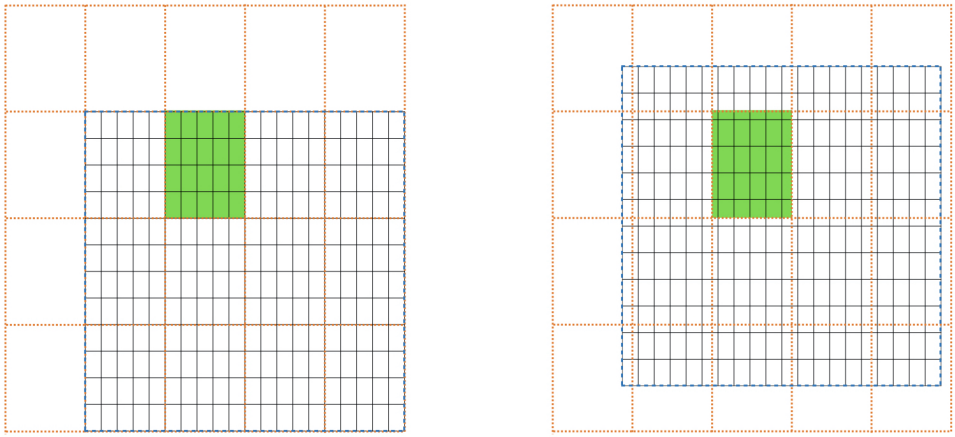
Let $\mathcal{I}_B \equiv \{i : B \cap U_i \neq \emptyset\}$ be the indices of the BAUs that overlap with B, and let $\tilde{B} \equiv \cup\{U_i : i \in \mathcal{I}_B\}$ be the associated set of BAUs. We say that the BAUs $\{U_1, \ldots, U_{N_D}\}$ and the blocks $\{B_1, \ldots, B_{M_D}\}$ are perfectly aligned if and only if the sets $\mathcal{I}_{B_j}, j = 1, \ldots, M_D$, are disjoint; otherwise the BAUs and the blocks are spatially misaligned; see Fig. S1

We then define the coarse-resolution covariate $X(B)$ as the weighted average of $X(\cdot)$ over $\tilde{\mathcal{B}} \equiv \tilde{B} \cap \mathcal{D}$, namely

$$X(B) = \sum_{\boldsymbol{s} \in \tilde{\mathcal{B}}} w_{\tilde{\mathcal{B}}}(\boldsymbol{s}) X(\boldsymbol{s}), \tag{S6.1}$$

where $\tilde{\mathcal{B}}$ consists of centroids of BAUs that are in $\tilde{B}$, and the weights $\{w_{\tilde{\mathcal{B}}}(\boldsymbol{s})\}$ satisfy $w_{\tilde{\mathcal{B}}}(\boldsymbol{s}) > 0$ for every $\boldsymbol{s} \in \tilde{\mathcal{B}}$ and $\sum_{\boldsymbol{s} \in \tilde{B}} w_{\tilde{\mathcal{B}}}(\boldsymbol{s}) = 1$. Each weight measures the contribution of the associated BAU to the block. Specification of the weights depends on the relevant scientific problem, but in our context where we assume that the block is an average of $X(\cdot)$ over its support, we can set the weights to be proportional to the area of overlap between the BAU and the block (see, e.g., Sainsbury-Dale et al. 2024).

Let $\boldsymbol{X}_B \equiv (X(B_1), \ldots, X(B_{M_D}))^\top$ be the M_D-dimensional vector of block values, where $X(B_j)$ is defined by (S6.1), for $j = 1, \ldots, M_D$. We then define the aggregation matrix $\boldsymbol{W}_B$ such that $\boldsymbol{X}_B = \boldsymbol{W}_B \boldsymbol{X}_{\boldsymbol{s}}$, where the (j, i)-th entry of $\boldsymbol{W}_B$ is $w_{ji} \equiv w_{\tilde{\mathcal{B}}_j}(\boldsymbol{s}_i)$ if $\boldsymbol{s}_i \in \tilde{\mathcal{B}}_j$ and $w_{ji} \equiv 0$ otherwise, for $j = 1, \ldots, M_D$, and $i = 1, \ldots, N_D$. The remaining steps for downscaling follow Section 2.2 of the main text.



(a) Perfect alignment (b) Misalignment

Figure S1: Example of perfect alignment (a) and misalignment (b). The region within the blue dashed lines is the spatial domain of interest. The black solid lines define the BAUs, while the dotted orange lines form the blocks. The green shading shows how the BAUs are related to the coarse-resolution blocks in the two cases.

S7 Computation cost of fitting the GPB model

Our downscaling method specifies a GPB (GP at the BAU level), but the model parameters are inferred from coarse-resolution data. The computational complexity of evaluating the likelihood (equation 5 in the main text) is of order $M_D N_D^2$. When $M_D \ll N_D$ (i.e., there is a large-fold difference between the coarse and fine resolutions), the complexity approximately grows as N_D^2, and the aggregation matrix is sparse. Specialised algorithms that work efficiently with sparse matrices can be employed to economise the computing time. When N_D is very large, scalable GP models may be used (e.g., Zammit-Mangion and Rougier 2018; Heaton et al. 2019; Cressie et al. 2022) to specify the BAU process.

References

Berger, J. O., De Oliveira, V., and Sansó, B. (2001), "Objective Bayesian analysis of spatially correlated data," *Journal of the American Statistical Association*, 96, 1361–1374. https://doi.org/10.1198/016214501753382282.

Cressie, N., Sainsbury-Dale, M., and Zammit-Mangion, A. (2022), "Basis-function models in spatial statistics," *Annual Review of Statistics and Its Application*, 9, 373–400. https://doi.org/10.1146/annurev-statistics-040120-020733.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F.,

Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. (2019), "A case study competition among methods for analyzing large spatial data," *Journal of Agricultural, Biological and Environmental Statistics*, 24, 398–425. https://doi.org/10.1007/s13253-018-00348-w.

Sainsbury-Dale, M., Zammit-Mangion, A., and Cressie, N. (2024), "Modeling big, heterogeneous, non-Gaussian spatial and spatio-temporal data using FRK," *Journal of Statistical Software*, 108, 1–39. https://doi.org/10.18637/jss.v108.i10.

Zammit-Mangion, A. and Rougier, J. (2018), "A sparse linear algebra algorithm for fast computation of prediction variances with Gaussian Markov random fields," *Computational Statistics & Data Analysis*, 123, 116–130. https://doi.org/10.1016/j.csda.2018.02.001.