RESEARCH ARTICLE

WILEY

Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support

C. J. Wilkie[1] | C. A. Miller[1] | E. M. Scott[1] | R. A. O'Donnell[1] | P. D. Hunter[2] |
E. Spyrakos[2] | A. N. Tyler[2]

[1]School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[2]Biological and Environmental Sciences, University of Stirling, Stirling, UK

Correspondence
C. J. Wilkie, School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8QQ, UK.
Email: craig.wilkie@glasgow.ac.uk

Abstract

Statistical downscaling has been developed for the fusion of data of different spatial support. However, environmental data often have different temporal support, which must also be accounted for. This paper presents a novel method of nonparametric statistical downscaling, which enables the fusion of data of different spatiotemporal support through treating the data at each location as observations of smooth functions over time. This is incorporated within a Bayesian hierarchical model with smoothly spatially varying coefficients, which provides predictions at any location or time, with associated estimates of uncertainty. The method is motivated by an application for the fusion of *in situ* and satellite remote sensing log(chlorophyll-*a*) data from Lake Balaton, in order to improve the understanding of water quality patterns over space and time.

KEYWORDS

Bayesian hierarchical modelling, change-of-support, chlorophyll-*a*, data fusion, statistical downscaling

1 | INTRODUCTION

Monitoring and understanding the changing patterns of lake health indicators has been increasingly recognised as vitally important in recent years, due to the role of lakes as sentinels of climate change and due to the potentially severe effects of climate change on the ecosystem services that lakes provide (Williamson, Saros, Vincent, & Smol, 2009). Traditionally, this has been accomplished through *in situ* monitoring of individual lakes, for which samples are collected manually from a small number of locations within a lake. Although this allows very accurate measurements to be recorded, it is costly in terms of both time and resources, meaning that very few locations are sampled, often fairly infrequently. The result of this is that these data do not provide us with a good understanding of the spatial patterns in the lake. More recently, remote sensing data have become widely available. Satellite remote sensing techniques observe larger areas of the Earth, providing maps on a temporally averaged time scale. Such data are often validated using data collected *in situ*, in order to assess and ensure their accuracy. This leads to a change-of-support problem (Cressie & Wikle, 2011), where temporally averaged grid-cell-scale remote sensing data must be fused with point-location, point-time *in situ* data.

Chlorophyll-*a* is an indicator of phytoplankton biomass and as such has been widely used as an indicator of lake health. High levels of chlorophyll-*a* may be caused by excessive nutrient levels and are often associated with phytoplankton blooms. These blooms can be poisonous to fish and other wildlife and, in the case of cyanobacteria, may form mats

covering the lake surface and reducing light levels in the water (Bláha, Babica, & Maršálek, 2009; Teta et al., 2017). Chlorophyll-*a* is a useful measure of lake water quality because it can be measured directly from *in situ* samples and retrieved by an algorithm from optical remote sensing observations (Markelin et al., 2017; Palmer, Hunter, et al., 2015; Palmer, Odermatt, et al., 2015).

Statistical downscaling has been developed to fuse data of different spatial support. For example, Berrocal, Gelfand, and Holland (2010a, 2010b) fuse grid-cell-scale numerical model output for air quality with point-location monitoring station data, using a Bayesian hierarchical model with smoothly spatially-varying coefficients. Wilkie et al. (2015) extend this to lakes by presenting a statistical downscaling model for the fusion of point-scale *in situ* data and grid-cell-scale remote sensing data for log(chlorophyll-*a*) in Lake Balaton. The statistical downscaling approach is successful in accomplishing the fusion of data of different spatial support. However, the modelling requires the assumption that the *in situ* and remote sensing data are available on the same temporal support, which may not be valid in practice. Typically, aggregation of the *in situ* data to the same time scale as the monthly averaged remote sensing data may be required, potentially resulting in an increase in uncertainty relating to the *in situ* data. This paper presents a new approach that extends statistical downscaling to enable the fusion of data of different spatiotemporal support, avoiding the need for this aggregation and allowing predictions to be made at any time, including during periods of time for which no *in situ* data are available.

This paper is organised as follows. Section 2 introduces the log(chlorophyll-*a*) data set that motivates the methodological development. Section 3 presents the novel nonparametric statistical downscaling model. Section 4 describes the model-fitting procedure and presents the results. Finally, Section 5 presents the conclusions, along with some discussion.

2 | DATA

Lake Balaton, in Hungary, is the largest lake in Central Europe. The lake is very shallow, with a mean depth of only around 3 m but a surface area of around 596 km² (Istvánovics, Osztoics, & Honti, 2004; Palmer, Odermatt, et al., 2015). The lake has suffered from poor water quality since the 1960s, with one of its four basins reaching eutrophic status in the 1970s, before beginning to recover after measures were taken to manage nutrient inflow to the lake in the 1980s (Istvánovics et al., 2004). The lake is therefore of particular interest. The main inflow to the lake is the River Zala, which flows into the southwest basin of the lake, bringing nutrient-rich water. The main outflow is the Sió Canal, which leaves the northeast basin of the lake (Palmer, Odermatt, et al., 2015).

Data for chlorophyll-*a* were collected at nine *in situ* locations (shown in Figure 1), between the start of 2002 and the end of 2012, by the Balaton Limnological Institute and by the Central Transdanubian Water and Environment Management Board. The lake was sampled approximately fortnightly. The data are assumed to be accurate within measurement error because they were derived directly from samples of water that were taken from the lake surface by boat and were then analysed in a laboratory. However, the spatial coverage of the samples does not give a good indication of important spatial patterns in chlorophyll-*a* across the lake surface. The *in situ* data used in this paper were made available by the GloboLakes project (GloboLakes, 2014), which is a research consortium project investigating the changing state of the health of lakes and their responses to environmental pressures on a global scale.

Remote sensing data are available for 7,616 grid cells (shown in Figure 1), covering the entire lake surface at up to 300 m resolution. These data were collected as Earth surface reflectance data by the MERIS (Medium Resolution Imaging Spectrometer) instrument on board the ENVISAT satellite of the European Space Agency. These measurements were
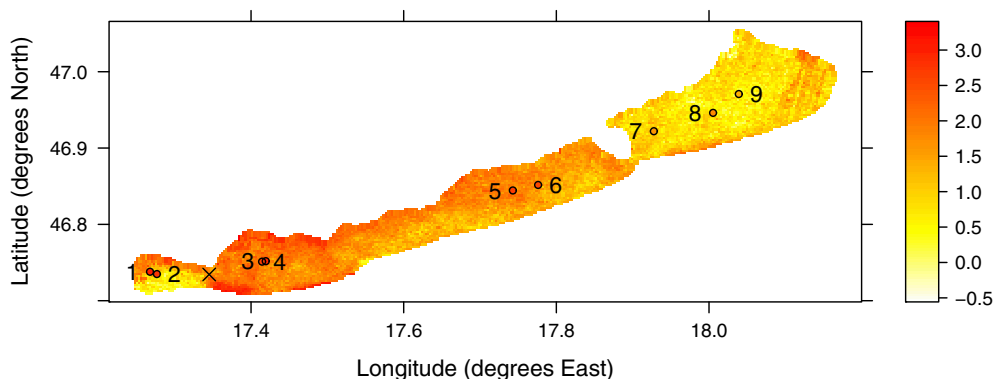

FIGURE 1 Remotely sensed log(chlorophyll-*a*) data (mg/m³) for 7,616 grid cells for March 2011, with the corresponding nine *in situ* data points overlaid (surrounded by black circles and numbered) and remote sensing grid cell P1 marked with a cross
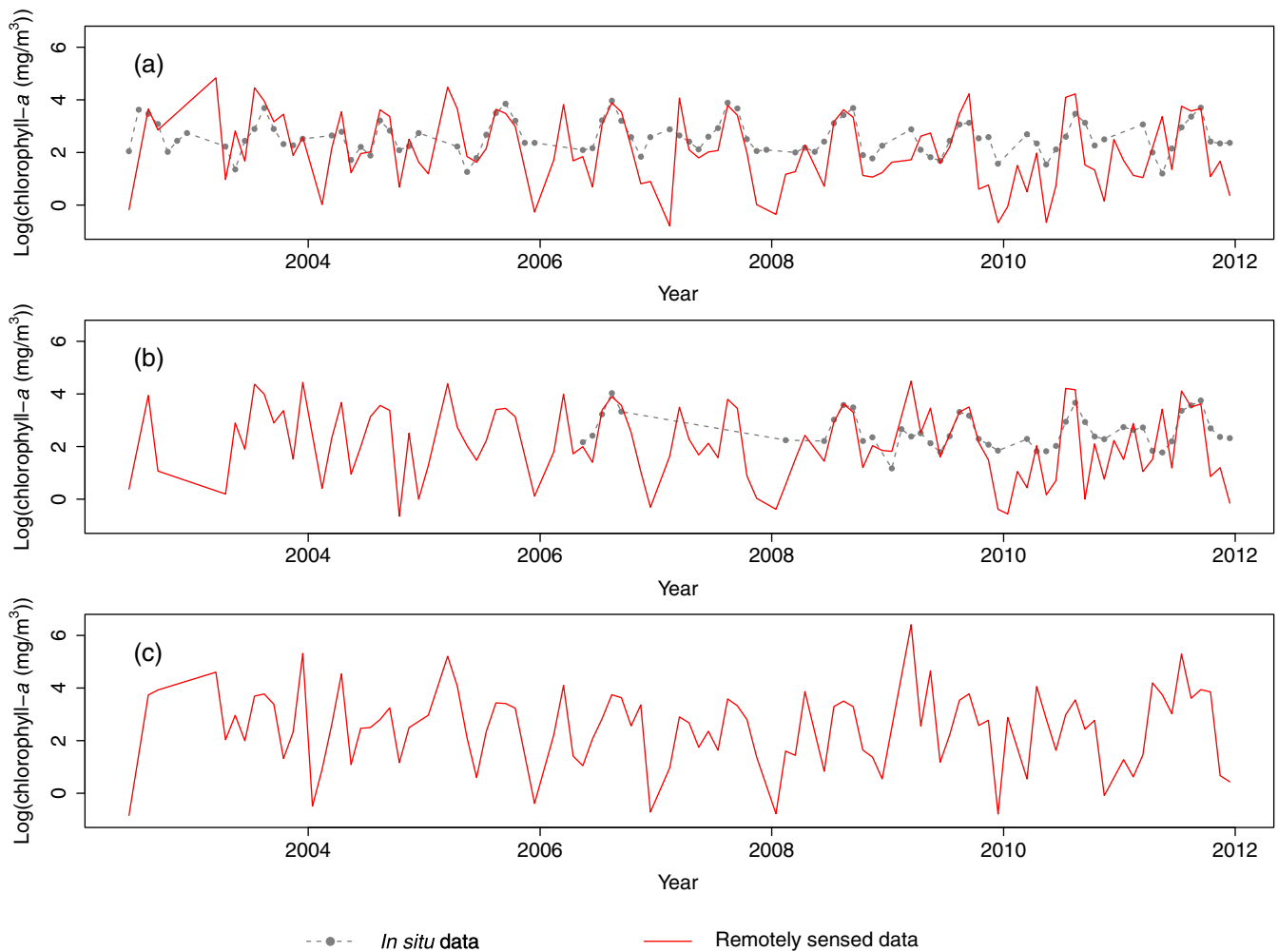
FIGURE 2 Time series of data for three example locations: (a) *in situ* data for Location 1 and remote sensing data for the corresponding grid cell, (b) *in situ* data for Location 2 and remote sensing data for the corresponding grid cell, and (c) remote sensing data for grid cell P1. (See Figure 1)

converted to chlorophyll-*a* data using algorithms as part of the Diversity II project and are available from their website (Diversity II, 2018).

The *in situ* and remote-sensing chlorophyll-*a* data on the original scale are right-skewed. Therefore, a natural logarithm-transformation was applied to the data before analysis.

Figure 2a,b shows how the *in situ* data for Locations 1 and 2 (i.e., the two furthest left locations, in the western part of the lake) relate to the remotely sensed data for the corresponding grid cells. The time series for each location follow similar patterns overall, with two peaks in log(chlorophyll-*a*) per year. However, the trace of the remotely sensed data is more variable than the trace of the *in situ* data for both locations, demonstrating the need for calibration of the remote sensing data with the *in situ* data, to ensure accuracy. Additionally, Location 2 has no *in situ* data recorded in the early part of the time period, with additional gaps around the year 2007. Figure 2c shows the remote sensing data for the grid cell marked with a cross in Figure 1.

3 | METHODS

3.1 | Spatial statistical downscaling

Spatial statistical downscaling is a method for the fusion of data of different spatial support, where inference is to be made at a finer spatial resolution than that of the original data series. The approach is motivated by the spatially varying coefficient model of Gelfand, Kim, Sirmans, and Banerjee (2003), which was developed by Berrocal et al. (2010b) into a

statistical downscaling model for the fusion of air quality data. For a vector of point-scale data $\mathbf{y} = (y_1, \ldots, y_n)^T$ collected at n locations and a vector of grid-cell-scale data $\mathbf{x} = (x_1, \ldots, x_n)^T$ (where each x_i is the remote sensing data point for the grid cell containing the point-location of y_i), the spatial statistical downscaling model is

$$y_i \sim N\left(\alpha_i + \beta_i x_i, \sigma_\varepsilon^2\right) \text{ for } i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ are the spatially varying intercept coefficients, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ are the spatially varying slope coefficients, and σ_ε^2 is the error variance. The spatially varying intercept and slope parameters are given prior distributions

$$\boldsymbol{\alpha} \sim N\left(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D}_{\text{data}})\right) \text{ and}$$
$$\boldsymbol{\beta} \sim N\left(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D}_{\text{data}})\right),$$

where $\mathbf{D}_{\text{data}}$ is an $(n \times n)$ matrix of Euclidean distances between the *in situ* data locations, with $(\mathbf{D}_{\text{data}})_{j,k} = |y_j - y_k|$ (for $j = 1, \ldots, n$ and $k = 1, \ldots, n$), σ_α^2 and σ_β^2 are the spatial variances, and ϕ_α and ϕ_β are the spatial decay parameters that control how fast the correlations in the intercept and slope parameters decrease towards zero as the distance between locations increases. The error variance parameter and the spatial variance parameters are typically given noninformative inverse-gamma prior distributions.

Model (1) was applied by Wilkie et al. (2015) to log(chlorophyll-a) data for Lake Balaton. However, this model and that of Berrocal et al. (2010b) only account for the spatial change of support and ignore the temporal change of support. In this paper, the temporal change of support is accounted for in the nonparametric statistical downscaling model, through smoothing.

3.2 | Nonparametric (spatiotemporal) statistical downscaling

In order to combine data sources with different spatial and temporal support, the novel nonparametric statistical downscaling model, Model (2), allows for the temporal change of support through treating the data at each location as observations of smooth functions over time, whereas the spatial change of support is accounted for through the spatially-varying coefficients part of the model. This hierarchical model, fitted in the Bayesian framework, allows the errors at each stage of the model to be propagated, so that all predictions have error estimates that account for the uncertainty at each stage of the model.

Model (2) fits smooth curves to the *in situ* data $\mathbf{y}_i$ for each location i ($i = 1, \ldots, n$) and to the remote sensing data $\mathbf{x}_i$ for the grid cells containing each *in situ* location i. These smooth curves are estimated through the use of basis functions of identical dimension m (Ramsay & Silverman, 2005). The same basis must be used for each $\mathbf{x}_i$ and $\mathbf{y}_i$ and must also be defined over the same time period, although the times at which it is evaluated may differ for each $\mathbf{x}_i$ and $\mathbf{y}_i$. The resulting vectors of basis coefficients ($\mathbf{c}_i$ for the *in situ* data and $\mathbf{d}_i$ for the remote sensing data) are related through a spatially-varying coefficients regression, which accommodates the spatial aspect of the data fusion.

Given a point-time data set of *in situ* data $\mathbf{y}_i$ at each point-location i ($i = 1, \ldots, n$) and a data set of remotely sensed data $\mathbf{x}_i$ for each grid cell containing each point-location i, the full description of the nonparametric statistical downscaling model is as follows:

$$
\begin{aligned}
\mathbf{y}_i | \mathbf{c}_i, \sigma_y^2 &\sim N_{q_i}\left(\Phi_i \mathbf{c}_i, \sigma_y^2 \mathbf{I}_{q_i}\right), \\
(\sigma_y^2)^{-1} &\sim Ga(a_y, b_y), \\
c_{ij} | \alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 &\sim N\left(\alpha_{ij} + \beta_{ij} d_{ij}, \sigma_c^2\right), \\
\boldsymbol{\alpha}_j | \sigma_\alpha^2 &\sim N_n\left(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D}_{\text{data}})\right), \\
\boldsymbol{\beta}_j | \sigma_\beta^2 &\sim N_n\left(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D}_{\text{data}})\right), \\
(\sigma_\alpha^2)^{-1} &\sim Ga(a_\alpha, b_\alpha), \\
(\sigma_\beta^2)^{-1} &\sim Ga(a_\beta, b_\beta), \\
(\sigma_c^2)^{-1} &\sim Ga(a_c, b_c), \\
\mathbf{x}_i | \mathbf{d}_i, \sigma_x^2 &\sim N_{p_i}\left(\Psi_i \mathbf{d}_i, \sigma_x^2 \mathbf{I}_{p_i}\right), \\
(\sigma_x^2)^{-1} &\sim Ga(a_x, b_x), \\
\mathbf{d}_i &\sim N_m(\boldsymbol{\mu}_d, \Sigma_d),
\end{aligned}
\tag{2}
$$

where:

- $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq_i})^T$ is the vector of *in situ* data at point location i ($i = 1, \ldots, n$) for times 1 to q_i.
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_i})^T$ is the vector of remote sensing data for the grid cell containing point location i, for times 1 to p_i. Note that the times of the remote sensing data may differ from those of the *in situ* data.
- n is the number of *in situ* data point locations.
- q_i is the number of *in situ* data points collected at point location i.
- p_i is the number of remotely sensed data points collected for grid cell i.
- $\mathbf{\Phi}_i$ is the ($q_i \times m$) matrix of basis functions evaluated at the q_i times of data collection for $\mathbf{y}_i$.
- $\mathbf{\Psi}_i$ is the ($p_i \times m$) matrix of basis functions evaluated at the p_i times of data collection for $\mathbf{x}_i$. Note that the basis being evaluated here is the same as for $\mathbf{\Phi}_i$.
- m is the basis dimension, that is, the number of basis functions in each $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_i$.
- $\mathbf{D}_{\text{data}}$ is the ($n \times n$) matrix of distances between the n point locations of the *in situ* data.
- ϕ_α and ϕ_β are the spatial decay parameters that are selected a priori.
- $a_y, b_y, a_\alpha, b_\alpha, a_\beta, b_\beta, a_c, b_c, a_x, b_x, \mu_d$, and Σ_d are values of hyperparameters, to be chosen a priori. Suitable values for μ_d and Σ_d are $\mathbf{0}$ and some multiple of $\mathbf{I}_m$, respectively, reflecting a lack of knowledge of the signs of the coefficients $\mathbf{d}_i$ and of their dependence structure. It is suggested by Lunn, Jackson, Best, Thomas, and Spiegelhalter (2013) that a small value for each of $a_y, b_y, a_\alpha, b_\alpha, a_\beta, b_\beta, a_c, b_c, a_x$, and b_x, such as 0.001, leads to noninformative priors. However, Gelman et al. (2014) note that there is no proper limiting distribution with these values, so that the posterior inferences are sensitive to the values of the coefficients of the prior distribution, implying that this prior is informative. Based upon Sahu, Gelfand, and Holland (2006) and Sahu, Gelfand, and Holland (2010), a suggestion is instead to set each of $a_y, a_\alpha, a_\beta, a_c$, and a_x equal to 2 and each of $b_y, b_\alpha, b_\beta, b_c$, and b_x equal to 1, so that each of their corresponding prior distributions has mean 1 and infinite variance. A sensitivity analysis showed that this choice did not affect predictions from the model, but that the posterior distributions of the variance terms displayed high skewness when Ga(0.001, 0.001) prior distributions were used. Therefore, the Ga(2, 1) prior distribution is preferred.

The model is represented as a directed acyclic graph in Figure 3, which visually displays the dependencies between the model parameters.
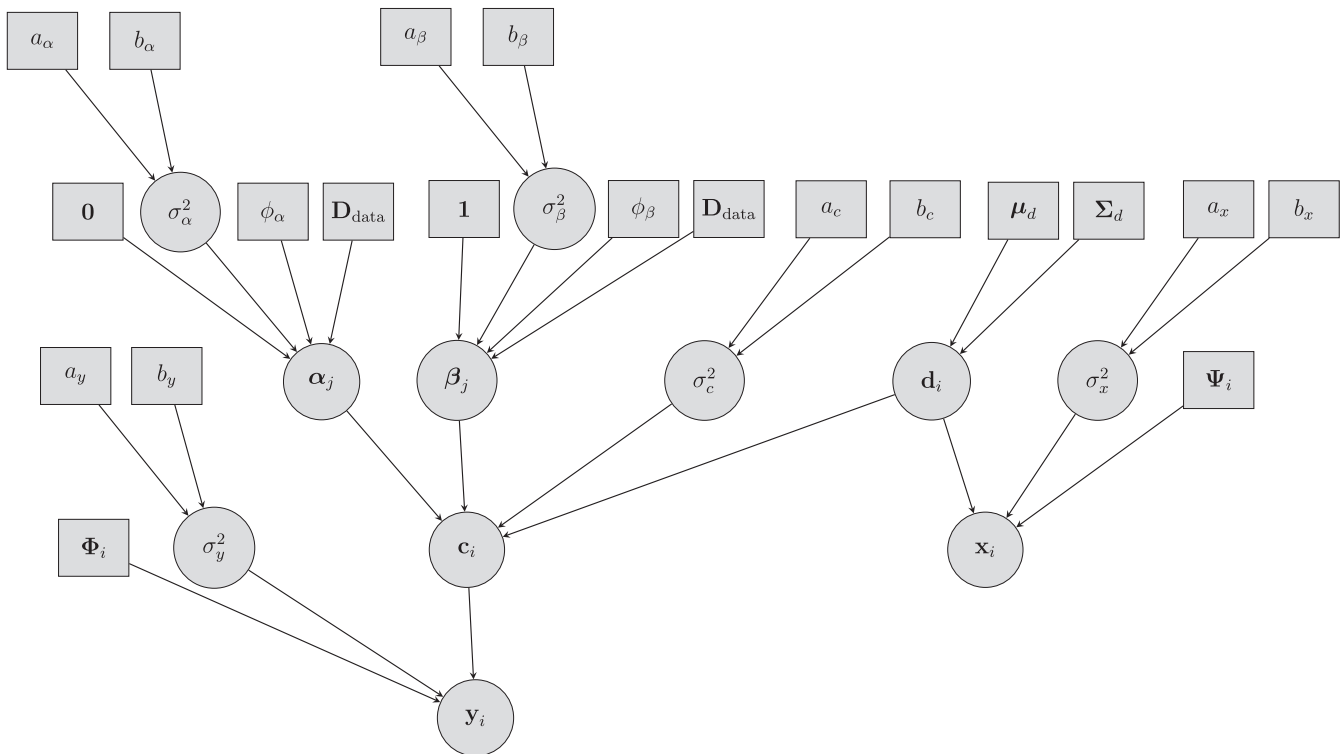

FIGURE 3 The directed acyclic graph (DAG) for Model (2). The circles are variables and the rectangles are constants. The arrows represent direct dependencies between the variables and constants

Note that the model uses shared variance parameters over time (σ_α^2 and σ_β^2) and over space and time (σ_x^2, σ_y^2, and σ_c^2). It is possible to use separate variance parameters for each spatial location and time point. However, the use of shared parameters was found to result in more accurate predictions from the model.

The full conditional posterior distributions of each parameter in Model (2) can be derived, allowing the model to be fitted using Gibbs Sampling (Gelman et al., 2014). These distributions are given in Section 1 of the supplementary material. At each iteration of the MCMC algorithm, a $\tilde{q}_i$-length vector of predictions $\tilde{\mathbf{y}}_i$ at location i ($i = 1, \ldots, \tilde{n}$, where $\tilde{n}$ is the number of prediction locations) is obtained for times $j = 1, \ldots, \tilde{q}_i$ by drawing from the posterior predictive distribution, conditional on the updated values of the other parameters, as follows:

$$\tilde{\mathbf{y}}_i | \tilde{\mathbf{c}}_i, \sigma_y^2 \sim \mathrm{N}_{\tilde{q}_i} \left(\tilde{\boldsymbol{\Phi}}_i \tilde{\mathbf{c}}_i, \sigma_y^2 \mathbf{I}_{\tilde{q}_i} \right),$$

$$\tilde{c}_{ij} | \tilde{\alpha}_{ij}, \tilde{\beta}_{ij}, \tilde{d}_{ij}, \sigma_c^2 \sim \mathrm{N} \left(\tilde{\alpha}_{ij} + \tilde{\beta}_{ij} \tilde{d}_{ij}, \sigma_c^2 \right),$$

$$\tilde{\boldsymbol{\alpha}}_j | \boldsymbol{\alpha}_j \sim \mathrm{N}_{\tilde{n}} \left(\mathbf{0} + \exp(-\phi_\alpha \mathbf{D}_{\mathrm{pred:data}}) \exp\left(-\phi_\alpha \mathbf{D}_{\mathrm{data}}\right)^{-1} (\boldsymbol{\alpha}_j - \mathbf{0}), \right.$$

$$\left. \sigma_\alpha^2 \left(\exp(-\phi_\alpha \mathbf{D}_{\mathrm{pred}}) - \exp(-\phi_\alpha \mathbf{D}_{\mathrm{pred:data}}) \exp\left(-\phi_\alpha \mathbf{D}_{\mathrm{data}}\right)^{-1} \exp(-\phi_\alpha \mathbf{D}_{\mathrm{data:pred}}) \right) \right),$$

$$\tilde{\boldsymbol{\beta}}_j | \boldsymbol{\beta}_j \sim \mathrm{N}_{\tilde{n}} \left(\mathbf{1} + \exp(-\phi_\beta \mathbf{D}_{\mathrm{pred:data}}) \exp\left(-\phi_\beta \mathbf{D}_{\mathrm{data}}\right)^{-1} (\boldsymbol{\beta}_j - \mathbf{1}), \right.$$

$$\left. \sigma_\beta^2 \left(\exp(-\phi_\beta \mathbf{D}_{\mathrm{pred}}) - \exp(-\phi_\beta \mathbf{D}_{\mathrm{pred:data}}) \exp\left(-\phi_\beta \mathbf{D}_{\mathrm{data}}\right)^{-1} \exp(-\phi_\beta \mathbf{D}_{\mathrm{data:pred}}) \right) \right),$$

$$\tilde{\mathbf{d}}_i \sim \mathrm{N}_m \left(\tilde{\Sigma}_{d_i} \tilde{\mathbf{A}}_{d_i}, \tilde{\Sigma}_{d_i} \right),$$

with

$$\tilde{\Sigma}_{d_i} = \left(\Sigma_d^{-1} + \tilde{\Psi}_i^{\mathrm{T}} \left(\sigma_x^2 \mathbf{I}_{\tilde{p}_i} \right)^{-1} \tilde{\Psi}_i \right)^{-1} \quad \text{and}$$

$$\tilde{\mathbf{A}}_{d_i} = \Sigma_d^{-1} \boldsymbol{\mu}_d + \tilde{\Psi}_i^{\mathrm{T}} \left(\sigma_x^2 \mathbf{I}_{\tilde{p}_i} \right)^{-1} \tilde{\mathbf{x}}_i,$$

where $\tilde{\mathbf{c}}_i = (\tilde{c}_{i1}, \ldots, \tilde{c}_{im})^{\mathrm{T}}$, $\tilde{\boldsymbol{\alpha}}_j = (\tilde{\alpha}_{1j}, \ldots, \tilde{\alpha}_{\tilde{n}j})^{\mathrm{T}}$, $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{1j}, \ldots, \tilde{\beta}_{\tilde{n}j})^{\mathrm{T}}$, and $\tilde{\mathbf{d}}_i = (\tilde{d}_{i1}, \ldots, \tilde{d}_{im})^{\mathrm{T}}$, and where $\tilde{\boldsymbol{\Phi}}$ is the ($\tilde{q}_i \times m$) matrix of basis functions evaluated at the $\tilde{q}_i$ prediction times for the *in situ* data at location i, $\tilde{\Psi}$ is the ($\tilde{p}_i \times m$) matrix of basis functions evaluated at the $\tilde{p}_i$ sampling times for the remotely sensed data for the grid cell containing location i, and $\tilde{\mathbf{x}}_i$ is the $\tilde{p}_i$-length vector of remotely sensed data for the grid cell containing the location i at which prediction is to be made. The matrices of distances between *in situ* data locations and prediction locations are defined as follows:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{\mathrm{pred}} & \mathbf{D}_{\mathrm{pred:data}} \\ \mathbf{D}_{\mathrm{data:pred}} & \mathbf{D}_{\mathrm{data}} \end{pmatrix},$$

where $\mathbf{D}_{\mathrm{pred}}$ is the ($\tilde{n} \times \tilde{n}$) matrix of distances between the $\tilde{n}$ prediction locations, $\mathbf{D}_{\mathrm{data}}$ is the ($n \times n$) matrix of distances between the n locations of the *in situ* data, $\mathbf{D}_{\mathrm{pred:data}}$ is the ($\tilde{n} \times n$) matrix of distances between the prediction locations and the *in situ* data locations, and $\mathbf{D}_{\mathrm{data:pred}} = \mathbf{D}_{\mathrm{pred:data}}^{\mathrm{T}}$.

3.2.1 | Choosing the basis type

There are many types of basis that could be used with the nonparametric statistical downscaling model. See, for example, the list given by Ramsay and Silverman (2005). The Fourier basis is most appropriate for periodic data and has the benefit of enabling computationally efficient calculations (Ramsay & Silverman, 2005). It is therefore focussed on for the Lake Balaton log(chlorophyll-*a*) data in this paper.

3.2.2 | Choosing the basis dimension

Choosing the number of basis functions, m, can be considered as a bias-variance trade-off problem (Ramsay & Silverman, 2005). The larger the basis dimension, the closer the fitted curve tracks the data. A large-enough basis dimension is required for the curve to capture the important features of the data, but too large a dimension means that the curve is greatly influenced by noise, with high variance (Ramsay & Silverman, 2005).

For a function fitted to a single time series, Ramsay and Silverman (2005) note that stepwise procedures are often used to select the basis dimension, whereas Wood (2017) suggests setting the basis dimension to be larger than is thought to be necessary, with excess curvature in the smooth function penalised using a penalty term in the model.

In the nonparametric statistical downscaling model, the basis coefficients are used in the spatially varying regression level of the model and the basis dimension must therefore be the same at each location, for the *in situ* data and the remotely sensed data. This means that the usual methods for selecting the basis dimension of a function fitted to a single time series cannot be used. A suggested approach to basis dimension selection is to re-fit the model using a range of basis

dimensions and make a choice using cross-validation, through comparing statistics for the resulting predictions, such as root mean squared error (RMSE), mean absolute error (MAE), mean bias, variance, and credible interval (CI) coverage and width. This method of basis dimension selection is demonstrated in the application to the Lake Balaton data in the following section.

4 | RESULTS

The nonparametric statistical downscaling model, Model (2), was fitted to the data for Lake Balaton, with the data for nine *in situ* locations as the response vectors $\mathbf{y}_i$ ($i = 1, \ldots, 9$) and the remotely sensed data for 115 months between June 2002 and December 2011, for the grid cells containing the corresponding *in situ* locations, as the explanatory data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_9)$.

4.1 | Basis dimension selection

The data follow a seasonal pattern of two peaks per year (Palmer, Odermatt, et al., 2015), as shown in Figure 2, with a smaller bloom in winter or spring and a larger bloom between late summer and early autumn (Palmer, Odermatt, et al., 2015). A simple heuristic for the Fourier basis is to use a basis dimension of at least $2r + 1$, to be able to adequately represent a pattern with at least *r* peaks per year, giving more flexibility than that for a simple temperature curve. This leads to an estimate for the basis dimension of 5 for this application.

A leave-one-out cross-validation was performed, where the *in situ* data and the corresponding remotely sensed data for each of the nine *in situ* locations were removed in turn and predicted using the nonparametric statistical downscaling model, Model (2), fitted to the data for the eight remaining *in situ* locations. This was carried out for each of a sequence of basis dimensions (3, 5, 7, 9, 11, 13, and 15). Using these predictions, RMSE, MAE, the variance of the predictions, the mean 95% credible interval coverage, and the mean 95% credible interval length were calculated for each basis dimension.

There was very little change in the values of the summary statistics between Dimensions 5 and 15. However, the RMSE, MAE, and mean 95% credible interval length reached their minima for Dimension 9, which was therefore chosen as the basis dimension for the following work. This value for the basis dimension suggests that the temporal patterns in the *in situ* and remote sensing data are more complex than a simple pattern of two peaks per year.

4.2 | Choice of ϕ_α and ϕ_β

The spatial decay parameters, ϕ_α and ϕ_β, were each set equal to 0.1 because a sensitivity analysis determined that varying these estimates between 0.001 and 0.5 made very little difference to the resulting model fit and predictions. The values of RMSE, MAE, variance of predictions, mean 95% credible interval coverage, and 95% credible interval mean width were each calculated for a range of values of ϕ_α and ϕ_β, with minima for RMSE and MAE reached at 0.1 for both ϕ_α and ϕ_β. An illustrative plot is available in the supplementary material. The value of 0.1 is therefore an appropriate choice for this data set.

This small value for each of the spatial decay parameters means that these parameters are estimated to have strong spatial correlation. At the maximum distance between data points in the lake, the spatial correlation for each of the intercept and slope parameters is around 0.9. This means that the intercept and slope of the relationship between the *in situ* and remote sensing data do vary over the lake but not by a large amount, indicating that the calibration of the satellite data to the *in situ* data is fairly stable spatially over the lake, even though the lake is large.

4.3 | Choice of prediction locations

In order to gain a good understanding of the spatial patterns in log(chlorophyll-*a*) across Lake Balaton, predictions must be made at locations that provide a comprehensive spatial coverage of the lake. In this paper, this was accomplished through a Delaunay triangulation, constrained by points on the lake boundaries that were selected manually. The Delaunay triangulation results in the optimal coverage of the lake, for the number of prediction locations chosen, through maximising the minimum angles between the points of all possible triangulations (Shewchuk, 1997). A triangulation was carried out using the R package RTriangle (Shewchuk, 1996), with the maximum allowable triangle area set to 4.53×10^{-5}units^2,

giving the optimal spatial coverage of the lake for 997 prediction locations. A plot showing the resulting spatial coverage of the lake is available in the supplementary material.

4.4 | Predictions

Having chosen the basis dimension, the nonparametric statistical downscaling model, Model (2), was fitted to the data and predictions were made. Figure 4 illustrates the predictions from the model and the measured data, taking the example of *in situ* Locations 1 and 2 and remote sensing grid cell P1, for the Fourier basis of dimension 9. Figure 4a,b illustrates that the fitted lines follow the data fairly closely, reflecting the main patterns over time in the data for each location, specifically the pattern of two peaks of log(chlorophyll-*a*) per year. Figure 4b shows that the model performs well in the presence of periods with no *in situ* data, whereas Figure 4c demonstrates that the model is able to predict over time at locations without *in situ* data. For Location 2, the model predicts using information both from neighbouring locations with *in situ* data and from time periods when data are available for Location 2. Through the sharing of information over space, predictions can be made for locations within remote sensing grid cells that contain no *in situ* data at all.

Predictions were made over space at the 997 locations that were determined by the Delaunay triangulation. Figure 5a shows the predictions over the lake for March 15, 2011, with the corresponding *in situ* data for that month overlaid. The predictions in Figure 5a are generally slightly higher than the original remotely sensed data plotted in Figure 1. The *in situ* data in Figure 5a are almost indistinguishable from their surrounding predictions, illustrating the capacity of the
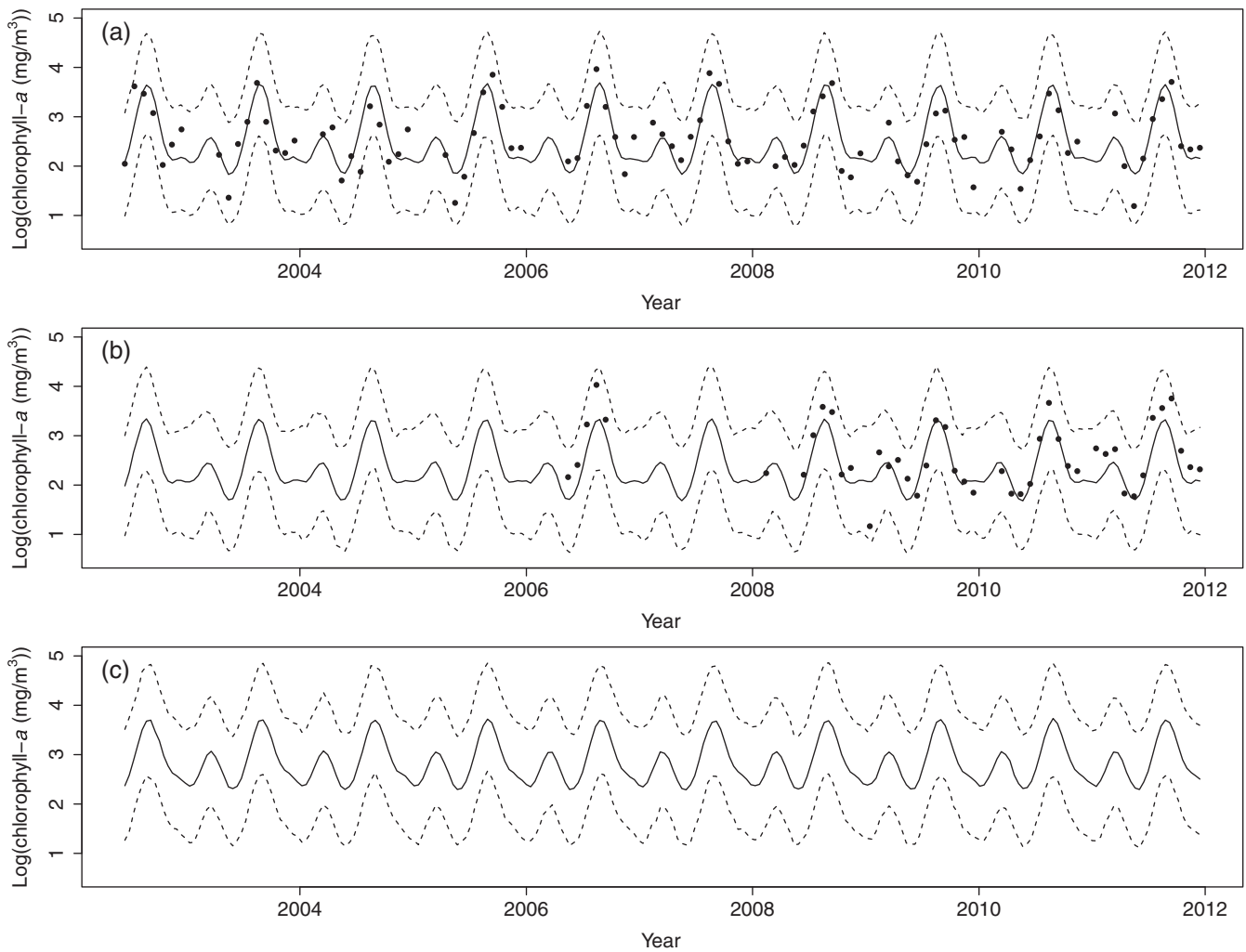

FIGURE 4 Plots of predictions from Model (2) using a Fourier basis of dimension 9, for three locations: (a) *in situ* Location 1, (b) *in situ* Location 2, and (c) centre of grid cell P1. (See Figure 1.) The *in situ* data are plotted as points, the predictions are plotted as solid lines, and 95% credible intervals for predictions are plotted as dashed lines
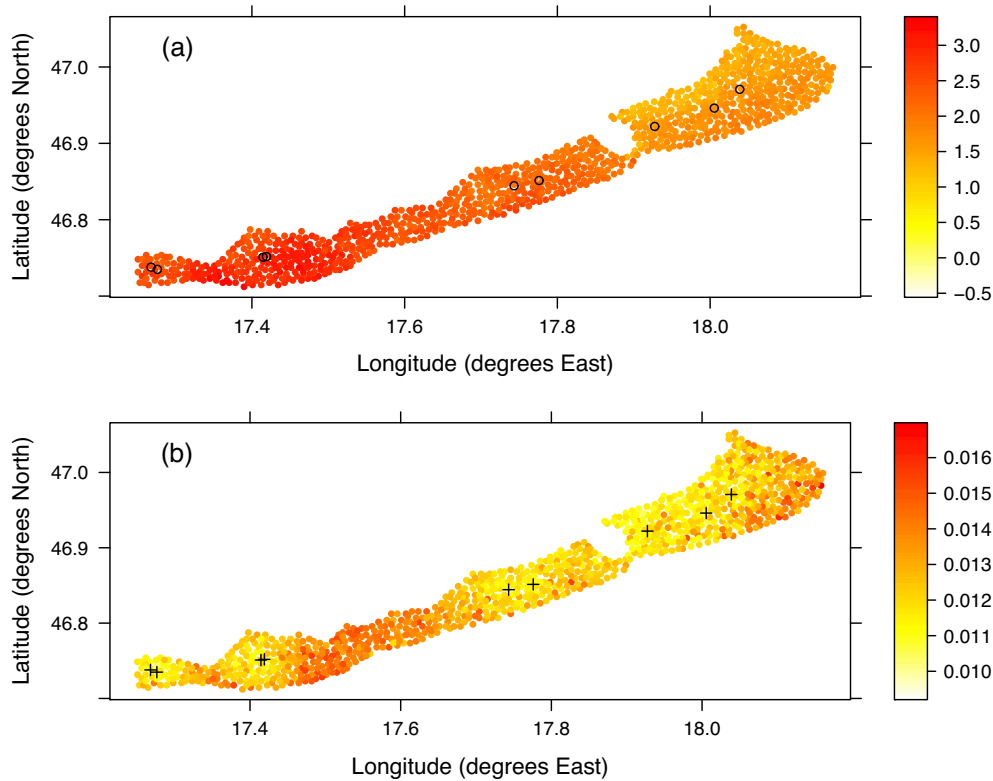
FIGURE 5 (a) Plot of the predictions from Model (2) for the middle of March 2011 at 997 locations as determined by a Delaunay triangulation, with the corresponding *in situ* data overlaid (surrounded by black circles). (b) Plot of the standard errors of the predictions from Model (2) for the middle of March 2011, with the *in situ* data locations marked with black crosses

model to calibrate the remotely sensed data. Figure 5b shows the standard errors associated with the predictions. The smallest values are located close to the *in situ* data locations, as expected. The standard errors are all small in comparison to the variation across the lake shown in Figure 5a, providing some evidence of a true pattern in log(chlorophyll-*a*) across the lake.

The plot of predictions shows higher levels of log(chlorophyll-*a*) in the southwestern part of the lake for this time point, suggesting that this part of the lake may need to be studied in more detail by water quality experts. The spatial patterns may reflect the flow of water through the lake, with the nutrient-rich inflow from the River Zala leading to high levels of log(chlorophyll-*a*) in the southwest basins of the lake, whereas the northwest basin has lower levels of nutrient inflow and, therefore, lower levels of log(chlorophyll-*a*). Validation of the model is additionally provided by the fact that the spatial and temporal patterns in the predictions appear to reflect the known spatial and temporal patterns in the ecology of the lake.

4.5 | Comparison between the models

A comparison was carried out, to investigate how the accuracy and precision of the predictions vary between the spatial statistical downscaling model, Model (1), and the nonparametric (spatiotemporal) statistical downscaling model, Model (2). Here, the *in situ* data were aggregated to the monthly scale, to enable the spatial statistical downscaling model, Model (1), to be fitted. The nonparametric statistical downscaling model was fitted using a Fourier basis with the optimal dimension chosen in the previous section (i.e., Dimension 9). A leave-one-out cross-validation was carried out, with predictions made at the *in situ* data locations. The summary statistics resulting from comparing the predictions to the values of the *in situ* data are given in Table 1.

These results show that the nonparametric statistical downscaling model results in more accurate predictions in comparison to the spatial statistical downscaling model because the RMSE and MAE are lower. The absolute value of mean bias is also lower for the nonparametric model. The predictions from the nonparametric model have lower variance, which may be explained by the fact that the predictions are constrained by the nonparametric model to lie along a smooth

TABLE 1 Table of model summary statistics for the spatial statistical downscaling model, that is, Model (1), and the nonparametric statistical downscaling model, that is, Model (2), fitted with a Fourier basis of dimension 9

Model	RMSE	MAE	Mean bias	Variance of predictions	Mean 95% CI coverage	Mean 95% CI length
(1) Spatial	0.561	0.392	−0.036	0.648	0.911	1.584
(2) Fourier Dimension 9	0.455	0.355	0.006	0.305	0.967	2.052

Note. RMSE = root mean squared error; MAE = mean absolute error; CI = credible interval.

curve at each location. The mean 95% credible intervals are wider for the nonparametric model than for the spatial model, but this apparent decrease in precision results in empirical credible interval coverage that is equal to, or greater than, the nominal 95%, in contrast to that of the spatial model.

4.6 | Computation details and assumptions

The nonparametric statistical downscaling model was fitted using Gibbs Sampling because every full conditional posterior distribution could be derived. The model was fitted in `C++`, via the `R` package `Rcpp` (Eddelbuettel, 2013), allowing efficient computation. Summaries of the MCMC chains were calculated using the `R` package `coda` (Plummer, Best, Cowles, & Vines, 2006). Each model was run for 10,000 iterations, for 2 chains, with every 10th iteration saved (to reduce the memory requirements). The derived full conditional distributions are given in Section 1 of the supplementary material.

The convergence of the model parameters was checked, using trace and density plots. Examples of a selection of these are given in Section 2 of the Supplementary Material. These provide no evidence against the validity of the assumption that the MCMC chains have converged for each parameter. Additionally, diagnostic plots provide no evidence against the assumptions of mean-zero, Normally distributed, homoscedastic residuals. Example diagnostic plots are also given in Section 2 of the Supplementary Material.

Stationary, isotropic spatial covariance functions are used in the model. It may be of interest to investigate more complex nonstationary or anisotropic spatial covariance functions for such a model, but the nature of the Lake Balaton data makes this difficult because there are only nine *in situ* data locations, which are arranged almost linearly.

Potential spatial and temporal autocorrelation in the residuals and the c_{ij} parameters was investigated. Variograms with Monte Carlo envelopes were produced, which showed no statistically significant evidence of spatial autocorrelation in the residuals or c_{ij} parameters. Because the model regresses a measure of a variable on another measure of the same variable, and because the *in situ* data are at most fortnightly at each location, it is not expected that there will be any remaining residual temporal autocorrelation. The Fourier basis should also account for the seasonal patterns over time. This was investigated through plotting the autocorrelation functions of the residuals at each *in situ* location. These plots showed no statistically significant evidence of temporal autocorrelation for all *in situ* locations except Location 5, which has data for only 19 time points and is therefore prone to being affected by unusual patterns. It does not appear that temporal autocorrelation is a problem for the model in general. Should any residual autocorrelation be encountered in a future application of the model, prior distributions that account for this can be investigated. However, this is beyond the scope of the current work.

5 | CONCLUSIONS AND DISCUSSION

Nonparametric statistical downscaling fuses data of different spatiotemporal support, such as point-location, point-time *in situ* data and grid-cell-scale, monthly averaged remotely sensed data. The method does this successfully, through a novel combination of statistical downscaling, to account for the spatial change of support, and smoothing over time, to account for the temporal change of support. Nonparametric statistical downscaling can be applied to real life data sets, without the requirement that one data set is aggregated over time, allowing predictions to be made at any location and time point, including during time periods for which no data are available. This is accomplished through the sharing of information over space and over time.

The development of the model was motivated by an application to the fusion of *in situ* data and remote sensing data for log(chlorophyll-*a*) in Lake Balaton, Hungary. The resulting predictions from the nonparametric statistical downscaling model, fitted to data for nine *in situ* point locations and 7,616 remote sensing grid cells, followed the known pattern

of two peaks in log(chlorophyll-*a*) per year. The resulting spatial maps of predictions across the lake may be used to monitor spatial patterns in water quality and to identify regions of the lake of particular concern. For example, the map of predictions in Figure 5a reflected the effect of the inflow of nutrient-rich water from the River Zala into the west of the lake, which might suggest a need for further measures to control the water quality entering the lake from the river.

In this work, the basis dimension for the nonparametric statistical downscaling model was selected through cross-validation. Future work could investigate the development of a more automatic procedure for basis dimension selection.

It is hoped that the motivational example given in this paper demonstrates the utility of the nonparametric statistical downscaling model as a tool for the fusion of environmental data of different spatiotemporal support. For the example of Lake Balaton, which motivated the development of the nonparametric statistical downscaling model, the resulting predictions and corresponding uncertainty estimates should be useful for those interested in understanding the spatial and temporal changes in the water quality of the lake, over both space and time.

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

AUTHOR CONTRIBUTIONS

CW developed the methodology, carried out the analysis of the data, and wrote this paper, under the supervision of CM and MS, with additional guidance and data provided by RO'D. PH, ES, and AT provided the *in situ* data. All authors read and approved this paper for publication.

ORCID

C. J. Wilkie https://orcid.org/0000-0003-0805-0195
E. Spyrakos https://orcid.org/0000-0001-7970-5211

REFERENCES

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2010a). A bivariate space-time downscaler under space and time misalignment. *The Annals of Applied Statistics*, *4*(4), 1942.

Berrocal, V. J., Gelfand, A. E., & Holland, D. M. (2010b). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, *15*(2), 176–197.

Bláha, L., Babica, P., & Maršálek, B. (2009). Toxins produced in cyanobacterial water blooms — Toxicity and risks. *Interdisciplinary Toxicology*, *2*(2), 36–41.

Cressie, N., & Wikle, C. K. (2011). *Wiley Series in Probability and Statistics. Statistics for spatio-temporal data*. Hoboken, NJ: John Wiley & Sons.

Diversity II. (2018). *Diversity II: Supporting the Convention on Biological Diversity*. Retrieved from http://www.diversity2.info/products/inlandwaters/

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York, NY: Springer.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, *98*, 387–396.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

GloboLakes (2014). *GloboLakes: A global observatory of lake responses to environmental change*. Retrieved from http://www.globolakes.ac.uk/

Istvánovics, V., Osztoics, A., & Honti, M. (2004). Dynamics and ecological significance of daily internal load of phosphorus in shallow Lake Balaton, Hungary. *Freshwater Biology*, *49*, 232–252.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book*. Boca Raton, FL: CRC Press.

Markelin, L., Simis, S. G. H., Hunter, P. D., Spyrakos, E., Tyler, A. N., Clewley, D., & Groom, S. (2017). Atmospheric correction performance of hyperspectral airborne imagery over a small eutrophic lake under changing cloud cover. *Remote Sensing*, *9*, 2.

Palmer, S. C. J., Hunter, P. D., Lankester, T., Hubbard, S., Spyrakos, E., Tyler, A. N., … Tóth, V. R. (2015). Validation of Envisat MERIS algorithms for chlorophyll retrieval in a large, turbid and optically-complex shallow lake. *Remote Sensing of Environment*, *157*, 158–169.

Palmer, S. C. J., Odermatt, D., Hunter, P. D., Brockmann, C., Présing, M., Balzter, H., & Tóth, V. R. (2015). Satellite remote sensing of phytoplankton phenology in Lake Balaton using 10 years of MERIS observations. *Remote Sensing of Environment*, *158*, 441–452.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*, 7–11.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York, NY: Springer.

Sahu, S. K., Gelfand, A. E., & Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*, 61.

Sahu, S. K., Gelfand, A. E., & Holland, D. M. (2010). Fusing point and areal level space-time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*, 77–103.

Shewchuk, J. R. (1996). Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator. In M. C. Lin & D. Manocha (Eds.), *Applied computational geometry: Towards geometric engineering* (Vol. 1148, pp. 203–222). Berlin, Germany: Springer.

Shewchuk, J. R. (1997). *Delaunay refinement mesh generation* (Doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA.

Teta, R., Romano, V., Della Sala, G., Picchio, S., De Sterlich, C., Mangoni, A., … Lega, M. (2017). Cyanobacteria as indicators of water quality in Campania coasts, Italy: A monitoring strategy combining remote/proximal sensing and *in situ* data. *Environmental Research Letters*, *12*, 024001.

Wilkie, C. J. (2017). *Nonparametric statistical downscaling for the fusion of in-lake and remote sensing data* (Doctoral dissertation). University of Glasgow, Glasgow, UK.

Wilkie, C. J., Scott, E. M., Miller, C., Tyler, A. N., Hunter, P. D., & Spyrakos, E. (2015). Data fusion of remote-sensing and in-lake chlorophyll$_a$ data using statistical downscaling. *Procedia Environmental Sciences*, *26*, 123–126.

Williamson, C. E., Saros, J. E., Vincent, W. F., & Smol, J. P. (2009). Lakes and reservoirs as sentinels, integrators, and regulators of climate change. *Limnology and Oceanography*, *54*, 2273–2282.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). New York, NY: Chapman and Hall/CRC.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Section 1 gives the full conditional posterior distributions of the parameters of the nonparametric statistical downscaling model. **Section 2** contains additional plots. **R package.** An R package NSD has been developed to demonstrate the use of the nonparametric statistical downscaling model that was introduced in this paper. The package is available at http://dx.doi.org/10.5525/gla.researchdata.651.

How to cite this article: Wilkie CJ, Miller CA, Scott EM, et al. Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics*. 2019;30:e2549. https://doi.org/10.1002/env.2549