Nonparametric statistical downscaling for the fusion of in-lake and remote sensing data

Craig John Wilkie

This thesis is submitted in fulfilment of the requirements
of the degree of Doctor of Philosophy.

School of Mathematics & Statistics
College of Science and Engineering
University of Glasgow

December 2017

Abstract

Lakes are vital components of the global biosphere, supporting complex ecosystems and playing important roles in the global biogeochemical cycle. However, they are vulnerable to the threat from climate change and their responses to climate forcing, eutrophication and other pressures, and their possibly confounding interactions, are not yet well understood. Monitoring lake health is therefore essential, in order to understand the changing patterns over space and time.

Traditionally, *in situ* data, which are collected directly from within lakes and analysed in laboratories, have been available for analysis. However, although these data are assumed to be accurate within measurement error, they are expensive to collect, so that few, if any, *in situ* sampling locations are available for each lake, often with infrequent sampling at each location. On the other hand, remotely-sensed data, which are derived from reflectance measurements of the Earth's surface, obtained from satellites, have recently become widely available. These data have good spatial coverage of up to 300 metre resolution, covering entire lakes, often with a monthly-average time-scale, but they must firstly be calibrated with the *in situ* data to ensure accuracy, before inferences are made.

The data for this research were provided by the GloboLakes project (www.globolakes.ac.uk), which is a consortium research project that is investigating the state of lakes and their responses to environmental drivers on a global scale. The research primarily focusses on $\log(\text{chlorophyll}_a)$ data for Lake Balaton, in Hungary, and for the Great Lakes of North America.

The key question of interest for this research is: "How can data fusion be performed for *in situ* and remotely-sensed lake water quality data, accounting for the spatiotemporal change of support between the point-location, point-time *in situ* data and the grid-cell-scale, monthly-averaged remotely-sensed data, producing a fused dataset that takes accuracy from the *in situ* data and spatial and temporal information from the remotely-sensed data?"

In order to answer this question, this thesis presents the following work:

- An initial analysis of the data for Lake Balaton motivates the following work, by demonstrating the spatial and temporal patterns in the data, using mixed-effects models, generalised additive models, kriging and principal components analysis.

- Following the identification of statistical downscaling as an appropriate method for fusion of the data, statistical downscaling models are developed, specifically in the framework of Bayesian hierarchical models with spatially-varying coefficients, for the novel application to data for $\log(\text{chlorophyll}_a)$, producing fully calibrated maps of fused data across lake surfaces, with associated comprehensive uncertainty measures.

- Bivariate and multiple-lakes statistical downscaling models are developed and applied, motivated by the assumption that sharing information between variables and between lakes can improve the accuracy of model predictions.

- The statistically novel method of nonparametric statistical downscaling is developed, to account for both the spatial and temporal aspects of the change of support between the *in situ* and remotely-sensed data. Using methodology from both functional data analysis and statistical downscaling, the model treats *in situ* and remotely-sensed data at each location as observations of smooth functions over time, estimated using bases, with the basis coefficients related via a spatially-varying

coefficient regression. This is computed within a Bayesian hierarchical model, enabling the calculation of comprehensive uncertainties.

This thesis presents the background, motivation, model development and application of the novel method of nonparametric statistical downscaling, filling the gap in the literature of accounting for changing temporal support in statistical downscaling modelling. Results are presented throughout this thesis, to demonstrate the utility of the method for real lake water quality data.

Acknowledgements

First and foremost, I would like to thank my supervisors, Claire Miller and Marian Scott, who have given me immense help and support throughout my time here. Whenever I have lost my confidence, they have been there to encourage me to carry on and be positive. I cannot thank them enough.

Thanks also go to the GloboLakes team, especially Andrew Tyler, Peter Hinter and Evangelos Spyrakos at the University of Stirling, for supplying and helping to understand the data, and Ruth O'Donnell at the University of Glasgow.

I would also like to thank my office mates throughout my PhD: Irene, Edith, Katie, Mengyi, Kelly, Amira, Guowen, Cunyi, Aisyah, George, Daniel, Alan, Maryam, Francesca and Sam. They have all put up with my silly questions over the years and I have always enjoyed our office chats, even those that were actually about statistics.

Thanks should also go to all of the postgraduate students in Statistics at the University of Glasgow, for the drinks, various brewery tours and creating a great community, and to the staff, especially those whose inspiring teaching of the subject led me to study Statistics at postgraduate level.

Thank you to my annual reviewers, Duncan Lee and Nema Dean, for their helpful suggestions, which have helped to improve the work in this thesis.

I would like to thank my dad and the rest of my family, for always being there when I've needed them. I would not be writing this, if it were not for the support and encouragement of my mum throughout my school years.

Finally, I would like to thank all of my friends, who have always been

there to remind me that there is more to life than work and to help me to take a much-needed break from thinking about my research.

Thanks to all of you!

Declaration

I declare that I have composed this thesis by myself and that all work described herein was carried out by myself, except as otherwise clearly stated and referenced in the text. I confirm that this work has not been previously submitted for any degree or professional qualification.

Some of the material in Chapter 3 of this thesis was published in Procedia Environmental Sciences (Volume 26, 2015), with the title "Data fusion of remote sensing and in-lake chlorophyll$_a$ data using statistical downscaling", and was presented as a poster at the Spatial Statistics conference in 2015. Further material from Chapter 3 was presented as a talk at the Research Students' Conference in Probability and Statistics in 2015, while some material in Chapter 4 was presented as a talk at The International Environmetrics Society conference in 2016. The material in Chapter 5 was presented as a talk at the Royal Statistical Society conference in 2017. A manuscript based upon the material in Chapter 5 is currently in preparation.

Contents

1 Introduction and background **1**

1.1 Background to the research 1

1.2 Research aims and objectives 4

1.3 Structure of this thesis 5

1.4 Introduction to the data 6

1.4.1 *In situ* data 7

1.4.2 Remotely-sensed data 10

1.5 Literature review 12

1.5.1 Data fusion 12

1.5.2 Downscaling literature 16

1.5.3 Conclusions from this literature review 19

1.6 Spatial and temporal modelling 20

1.6.1 Geostatistics 20

1.6.2 Nonparametric smoothing 25

1.7 Bayesian modelling 38

1.7.1 Gibbs sampling 39

1.7.2 Metropolis algorithm 40

1.7.3 Alternative methods 41

1.7.4 Convergence diagnostics 42

1.8 Conclusions 42

2 Initial spatial and temporal analysis of data **44**

2.1 Exploring the *in situ* data 45

2.1.1 Exploratory plots 46
2.1.2 Mixed-effects models for the *in situ* data 49
2.2 Exploring the remotely-sensed data 55
2.2.1 Kriging the ARC-Lake temperature data 55
2.2.2 Principal component analysis (PCA) of the remotely-sensed temperature data 60
2.3 Investigating the relationship between *in situ* and remote sensing data through additive modelling 69
2.3.1 Application to the Lake Balaton data 69
2.4 Conclusions 73

3 Statistical downscaling **75**
3.1 Background and motivation 75
3.2 Spatial statistical downscaling: model development 76
3.2.1 Spatially-varying coefficient modelling and statistical downscaling 77
3.2.2 Application of spatial statistical downscaling model 3.1 to log(chlorophyll$_a$) data for Lake Balaton 81
3.2.3 The Berrocal et al. (2010*b*) spatial downscaling model . 95
3.2.4 Simulation study 97
3.3 Spatiotemporal statistical downscaling model development . . 100
3.3.1 Spatiotemporal development of model 3.1 101
3.3.2 Spatiotemporal models including smoothing over time . 104
3.4 Applications to the Lake Erie data 111
3.5 Conclusions and discussion 119

4 Bivariate and multiple lakes downscaling **121**
4.1 Bivariate statistical downscaling 122
4.1.1 Motivational exploratory analysis 122
4.1.2 Spatial bivariate downscaling model 125
4.1.3 Spatiotemporal bivariate downscaling model 129

4.1.4 Application of spatiotemporal bivariate downscaling model to data for Lake Balaton 132
4.2 Multiple-lakes statistical downscaling 133
4.2.1 Model development 134
4.2.2 Model fitting and results 141
4.2.3 Conclusions and discussion 145
4.3 Overall conclusions and discussion 146

5 Nonparametric statistical downscaling 148
5.1 Background and motivation 148
5.2 Preliminary application to the data 149
5.2.1 Preliminary application of frequentist model 150
5.2.2 Preliminary application of Bayesian model 152
5.3 Developing a model for nonparametric statistical downscaling 155
5.3.1 Examining the correspondence of *in situ* and remotely-sensed basis coefficients 156
5.3.2 Combining a linear model and functional data analysis methodology 157
5.3.3 Nonparametric statistical downscaling: a fully Bayesian model for data fusion 160
5.4 Model fitting 161
5.4.1 Application to data for Lake Balaton 161
5.4.2 Application to data for Lake Erie 171
5.5 Conclusions and further work 178

6 Conclusions 183
6.1 Chapter 1: Introduction and background 184
6.2 Chapter 2: Initial spatial and temporal analysis of data 184
6.3 Chapter 3: Statistical downscaling 186
6.4 Chapter 4: Bivariate and multiple lakes downscaling 187
6.4.1 Bivariate statistical downscaling 188

6.4.2 Multiple lakes statistical downscaling 188
6.5 Chapter 5: Nonparametric statistical downscaling 189
6.6 Discussion, limitations and future work 191

A Derivation of full conditional posterior distributions 195
A.1 Spatiotemporal statistical downscaling model 3.3 196
A.2 Spatiotemporal statistical downscaling model 3.3a 201
A.3 Bivariate spatial model 4.1 204
A.4 Nonparametric downscaling model 5.8 211

B Diagnostic plots for statistical downscaling models 217

Bibliography 252

List of Tables

2.1 Numbers of *in situ* data available for Lake Balaton. 45

2.2 Table of number of grid cells with available data by number of months. 56

3.1 Table of summary statistics for leave-one-out cross-validation for models 3.1 (with $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$) and 3.2 (with $\phi_0 = 0.01$ and $\phi_1 = 0.1$). 96

3.2 Summary table for estimates of variance parameters of model 3.3 for each month j. 102

3.3 Table of summary statistics of leave-one-out cross-validations for models 3.3 and 3.3a. 103

3.4 Table of summary statistics for leave-one-out cross-validation for model 3.5, with $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$. 107

3.5 Table of summary statistics for leave-one-out cross-validation for model 3.5a, with $\psi_\mu = 20$ and $\psi_\nu = 15$. 110

3.6 Table of summary statistics for leave-one-out cross-validation for models 3.1 and 3.3a, with $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$, for Lake Erie data. 115

3.7 Table of summary statistics for leave-one-out cross-validation for models 3.5 and 3.5a, with $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$, and with ψ_μ and ψ_ν set equal to their chosen values, for the Lake Erie data. 119

4.1 Table of summary statistics of leave-one-out cross-validations for models 3.1, 4.1 and 4.1a, fitted to Lake Balaton log(total suspended matter) and log(chlorophyll$_a$) data. 127
4.2 Table of summary statistics for leave-one-out cross-validation for models 3.3a, 4.2 and 4.2a. 132
4.3 Performance statistics for several models for four Great Lakes, for log(chlorophyll$_a$). 142
4.4 Individual performance statistics for several models for four Great Lakes, for log(chlorophyll$_a$). 142

5.1 Summary statistics for cross validation assessing nonparametric downscaling model performance, for the Lake Balaton data. 167
5.2 Summary statistics for cross validation comparing traditional and nonparametric downscaling models, for the Lake Erie data. 175

List of Figures

1.1 Map of Lake Balaton, showing the nine *in situ* sampling locations. 7

1.2 Map of the Great Lakes, showing the *in situ* sampling locations. 9

1.3 Example of a fitted variogram, showing the partial sill, nugget and range. 22

2.1 Plot of log(chlorophyll$_a$ concentration) over time, with separate lines for each location. 46

2.2 Plot of log(total suspended matter concentration) over time, with separate lines for each location. 47

2.3 Plot of temperature over time, with separate lines for each location. 47

2.4 Plots of log(chlorophyll), log(total suspended matter) and temperature, coloured by location. 48

2.5 Plot of *in situ* log(chlorophyll$_a$) by year, showing predictions from model 2.1. 50

2.6 Plot of *in situ* log(total suspended matter) by year, showing predictions from model 2.3. 52

2.7 Plots of the autocorrelation functions and partial autocorrelation functions for models 2.1 and 2.3. 53

2.8 Plots of the autocorrelation functions and partial autocorrelation functions for models 2.1a and 2.3a, with AR(1) error structure. 54

2.9 Map of Lake Balaton, showing the nine *in situ* sampling locations and the forty-one ARC-Lake remote sensing data grid cell centres. 56
2.10 Remotely-sensed lake surface water temperature data for Lake Balaton, for a selection of months in 2006. 57
2.11 Variograms for linear model residuals for Lake Balaton 2006 temperature data. 59
2.12 Universal kriging predictions for temperature for 2006. 60
2.13 Universal kriging standard errors for temperature for 2006. . . 61
2.14 Scree plot and biplot for S-mode PCA on ARC-Lake data. . . 65
2.15 Plot of PC1 loadings and scores for S-mode PCA. 66
2.16 Scree plot and biplot for T-mode PCA on ARC-Lake data. . . 67
2.17 Plot of PC1 scores and loadings for T-mode PCA. 68
2.18 Plot of smoothing terms for Model 2.8. 71
2.19 Plot of predictions from model 2.8 at *in situ* locations, showing remotely-sensed data and predictions. 72

3.1 Plot of 997 nodes of Delaunay triangulation for Lake Balaton data, constrained by the input boundary points. 84
3.2 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.1, fitted to log(chlorophyll$_a$) data for October 2008, for Lake Balaton. . 85
3.3 Example trace and density plots for parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\epsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$, for model 3.1. 86
3.4 Example trace and density plots for parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\epsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$, for model 3.1. 87
3.5 Plots of cross-validation summary statistics for model 3.1, for each combination of ϕ_α and ϕ_β. 89
3.6 Plots of cross-validation summary statistics for model 3.1, for each combination of ϕ_α and ϕ_β. 90
3.7 Data for March 2011 and resulting predictions from model 3.1. 92

3.8 Predicted $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from model 3.1, for March 2011. 93
3.9 Upper and lower bounds for 95% credible interval for March 2011 predictions from model 3.1. 94
3.10 Plot of simulated *in situ* and remotely-sensed data. 98
3.11 Plots of performance statistics from the simulation study for model 3.1 and the model of Berrocal et al. (2010*b*). 99
3.12 Plots of summary statistics for a leave-one-out cross-validation for model 3.5 for sequences of values of ψ_μ and ψ_ν. 107
3.13 Plots of summary statistics for a leave-one-out cross-validation for model 3.5a for sequences of values of ψ_μ and ψ_ν. 109
3.14 *In situ* and remotely sensed log(chlorophyll$_a$) data versus time for one location in Lake Erie. 112
3.15 Plots of cross-validation summary statistics for model 3.1, for Lake Erie data, for each combination of ϕ_α and ϕ_β. 114
3.16 Data and predicted $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ for model 3.3a fitted to the log(chlorophyll$_a$) data for Lake Erie. 116
3.17 Plots of cross-validation summary statistics for models 3.5 and 3.5a, for Lake Erie data, for each combination of ϕ_μ and ϕ_ν. . 118

4.1 Plots showing relationships between *in situ* lake surface water temperature, log(chlorophyll$_a$) and log(total suspended matter). 123
4.2 Plots showing relationships between remotely-sensed lake surface water temperature, log(chlorophyll$_a$) and log(total suspended matter). 123
4.3 Plot of residuals for model 3.1 fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 124
4.4 Plots of remotely-sensed log(chlorophyll$_a$) data for March 2011 and resulting predictions from model 4.1a, with *in situ* data overlaid and circled in white. 129

4.5 Plots of remotely-sensed log(total suspended matter) data for March 2011 and resulting predictions from model 4.1a, with *in situ* data overlaid and circled in white. 130
4.6 Remotely-sensed log(chlorophyll$_a$) and downscaled surface from model 4.7-ST for August 2003, for 1005 locations in four Great Lakes. 144

5.1 Basis functions, smooth function fitted to *in situ* data and smooth function fitted to remotely-sensed data, for Lake Balaton location 9. 151
5.2 Basis functions and smooth function fitted to *in situ* data, for Lake Balaton location 1. 152
5.3 Smooth function fitted to the Lake Balaton location 9 *in situ* data using the Bayesian model, along with the corresponding 95% credible intervals. 154
5.4 Smooth function fitted to the Lake Balaton location 1 *in situ* data using the Bayesian model, along with the corresponding 95% credible intervals. 155
5.5 Scatterplot of *in situ* and remote sensing data basis coefficients, from fitting curves using a cubic B-spline basis of dimension 41. 156
5.6 Predicted smooth curve using basis coefficients estimated using a linear model, for Lake Balaton location 9 data. 157
5.7 Directed acyclic graph (DAG) for model 5.8. 162
5.8 Plots of summary statistics versus basis dimension for a leave-one-out cross-validation for Lake Balaton, for B-spline basis. . 164
5.9 Plots of various summary statistics versus basis dimension for a leave-one-out cross-validation. 166

5.10 Predictions for Lake Balaton location 1 log(chlorophyll$_a$) data from a nonparametric downscaling model fitted to data for locations 2 to 9, using a B-spline basis of dimension 49 and a Fourier basis of dimension 9. 168
5.11 Plots of data and predictions from statistical downscaling model 3.4 and nonparametric statistical downscaling model 5.8 with Fourier and B-spline bases, for March 2003 to May 2003, for Lake Balaton. 170
5.12 Plots of standard errors for predictions from statistical downscaling model 3.4 and nonparametric statistical downscaling model 5.8 with Fourier and B-spline bases, for March 2003 to May 2003, for Lake Balaton. 172
5.13 Plots of log(chlorophyll$_a$) over time for two locations in Lake Erie. 173
5.14 Summary statistics for Erie data from cross-validation using B-spline and Fourier basis. 174
5.15 Plots of predictions from model 5.8 at Erie locations 1 and 21, for B-spline dimension 14 and Fourier dimension 5 bases. . . . 176
5.16 Plots of data and predictions from statistical downscaling model 3.4 and nonparametric statistical downscaling model 5.8 with Fourier and B-spline bases, for March 2003 to May 2003, for Lake Erie. 177
5.17 Plots of standard errors for predictions from statistical downscaling model 3.4 and nonparametric statistical downscaling model 5.8 with Fourier and B-spline bases, for March 2003 to May 2003, for Lake Erie. 179

B.1 Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\epsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$, of model 3.1, fitted to the log(chlorophyll$_a$) data for Lake Balaton for October 2008. . . . 218

B.2 Trace and density plots for the parameters a_{11}, a_{21}, $w_{0,1}$, $w_{1,1}$, γ, α_1, δ and β_1 of model 3.2, fitted to the log(chlorophyll$_a$) data for Lake Balaton for October 2008. 219
B.3 Trace and density plots for the parameters $(\sigma^2_{\alpha 1})^{-1}$, $(\sigma^2_{\beta 1})^{-1}$, $(\sigma^2_{\varepsilon 1})^{-1}$, α_{11} and β_{11} of model 3.3, fitted to the log(chlorophyll$_a$) data for Lake Balaton. 220
B.4 Trace and density plots for the parameters $(\sigma^2_{\alpha})^{-1}$, $(\sigma^2_{\beta})^{-1}$, $(\sigma^2_{\varepsilon})^{-1}$, α_{11} and β_{11} of model 3.3a, fitted to the log(chlorophyll$_a$) data for Lake Balaton. 221
B.5 Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_{\varepsilon})^{-1}$, $(\sigma^2_{\alpha})^{-1}$ and $(\sigma^2_{\beta})^{-1}$ of model 3.5, fitted to the log(chlorophyll$_a$) data for Lake Balaton. 222
B.6 Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_{\varepsilon})^{-1}$, $(\sigma^2_{\alpha})^{-1}$ and $(\sigma^2_{\beta})^{-1}$ of model 3.5a, fitted to the log(chlorophyll$_a$) data for Lake Balaton. 223
B.7 Trace and density plots for the parameters $(\sigma^2_{\alpha})^{-1}$, $(\sigma^2_{\beta})^{-1}$, $(\sigma^2_{\varepsilon})^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$ of model 3.1, fitted to the log(chlorophyll$_a$) data for Lake Erie. 224
B.8 Trace and density plots for the parameters $(\sigma^2_{\alpha})^{-1}$, $(\sigma^2_{\beta})^{-1}$, $(\sigma^2_{\varepsilon})^{-1}$, α_{11}, β_{11}, $\tilde{\alpha}_{11}$, $\tilde{\beta}_{11}$ and $\tilde{y}_{11}$ of model 3.3a, fitted to the log(chlorophyll$_a$) data for Lake Erie. 225
B.9 Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_{\varepsilon})^{-1}$, $(\sigma^2_{\alpha})^{-1}$ and $(\sigma^2_{\beta})^{-1}$ of model 3.5, fitted to the log(chlorophyll$_a$) data for Lake Erie. 226
B.10 Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_{\varepsilon})^{-1}$, $(\sigma^2_{\alpha})^{-1}$ and $(\sigma^2_{\beta})^{-1}$ of model 3.5a, fitted to the log(chlorophyll$_a$) data for Lake Erie. 227
B.11 Trace and density plots for the parameters α_{111}, β_{111}, ρ_1, $\sigma_{\alpha_{11}}$, $\sigma_{\beta_{11}}$ and $\sigma^2_{\varepsilon_1}$ of model 4.1, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 228

B.12 Trace and density plots for the parameters α_{111}, β_{111}, ρ_1, $\sigma_{\alpha_{11}}$, $\sigma_{\beta_{11}}$ and $\sigma^2_{\varepsilon_1}$ of model 4.1a, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 229

B.13 Trace and density plots for the parameters $\alpha_{1,1,1}$, $\alpha_{1,1,2}$, $\beta_{1,1,1}$, $\beta_{1,1,2}$, ρ, $\sigma^{-1}_{\alpha^2_1}$ and $(\sigma^2_{\varepsilon,1})^{-1}$ of model 4.2, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 230

B.14 Trace and density plots for the parameters $\alpha_{1,1,1}$, $\alpha_{1,1,2}$, $\beta_{1,1,1}$, $\beta_{1,1,2}$, ρ, $\sigma^{-1}_{\alpha^2_1}$ and $(\sigma^2_{\varepsilon,1})^{-1}$ of model 4.2a, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 231

B.15 Trace and density plots for the parameters $y_{1,1(1)}$, $\gamma_{1,1(1)}$, β_1, α, η_1, δ, $(\sigma^2_\varepsilon)^{-1}$ and $(\sigma^2_{\gamma_1})^{-1}$ of model 4.4a-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes. 232

B.16 Trace and density plots for the parameters $\tilde{y}_{1,1}$, $\gamma_{1,1}$, $\beta_{1,1}$, α, η_1, δ and $(\sigma^2_\varepsilon)^{-1}$ of model 4.4b-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes. 233

B.17 Trace and density plots for the parameters $y_{1,1(1)}$, $\gamma_{1,1(1)}$, α, δ, $(\sigma^2_\varepsilon)^{-1}$ and $(\sigma^2_{\gamma_1})^{-1}$ of model 4.5a-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes. 234

B.18 Trace and density plots for the parameters $y_{1,1(1)}$, $\gamma_{1,1(1)}$, β_1, η_1, $(\sigma^2_\varepsilon)^{-1}$ and $(\sigma^2_{\gamma_1})^{-1}$ of model 4.6-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes. 235

B.19 Trace and density plots for the parameters $\tilde{y}_{1,1}$, $\gamma_{1,1}$, β, η, $(\sigma^2_\varepsilon)^{-1}$ and σ^2_γ of model 4.7-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes. 236

B.20 Trace and density plots for the parameters $(\sigma^2_\alpha)^{-1}$, $(\sigma^2_\beta)^{-1}$, $(\sigma^2_y)^{-1}$, $(\sigma^2_c)^{-1}$, $\alpha_{1,1}$, $\beta_{1,1}$, $c_{1,1}$ and $(\sigma^2_x)^{-1}$ of model 5.8, fitted to the log(chlorophyll$_a$) data for Lake Balaton. 237

B.21 Trace and density plots for the parameters $d_{1,1}$, $\tilde{\alpha}_{1,1}$, $\tilde{\beta_{1,1}}$, $\tilde{d}_{1,1}$, $\tilde{c}_{1,1}$ and $\tilde{y}_{1,1}$ of model 5.8, fitted to the log(chlorophyll$_a$) data for Lake Balaton. 238

B.22 Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_y^2)^{-1}$, $(\sigma_c^2)^{-1}$, $\alpha_{1,1}$, $\beta_{1,1}$, $c_{1,1}$ and $(\sigma_x^2)^{-1}$ of model 5.8, fitted to the log(chlorophyll$_a$) data for Lake Erie. 239
B.23 Trace and density plots for the parameters $d_{1,1}$, $\tilde{\alpha}_{1,1}$, $\tilde{\beta_{1,1}}$, $\tilde{d}_{1,1}$, $\tilde{c}_{1,1}$ and $\tilde{y}_{1,1}$ of model 5.8, fitted to the log(chlorophyll$_a$) data for Lake Erie. 240
B.24 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.1, fitted to log(chlorophyll$_a$) data, for Lake Balaton. 241
B.25 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.2, fitted to log(chlorophyll$_a$) data for Lake Balaton. 241
B.26 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.3, fitted to log(chlorophyll$_a$) data for Lake Balaton. 242
B.27 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.3a, fitted to log(chlorophyll$_a$) data for Lake Balaton. 242
B.28 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.5, fitted to log(chlorophyll$_a$) data for Lake Balaton. 243
B.29 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.5a, fitted to log(chlorophyll$_a$) data for Lake Balaton. 243
B.30 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.1, fitted to log(chlorophyll$_a$) data for Lake Erie. 244
B.31 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.3a, fitted to log(chlorophyll$_a$) data for Lake Erie. 244

B.32 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.5, fitted to log(chlorophyll$_a$) data for Lake Erie. 245
B.33 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 3.5a, fitted to log(chlorophyll$_a$) data for Lake Erie. 245
B.34 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.1, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 246
B.35 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.1a, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 246
B.36 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.2, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 247
B.37 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.2a, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. 247
B.38 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.4a-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes. 248
B.39 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.4b-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes. 248

B.40 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.5b-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes. 249
B.41 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.6-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes. 249
B.42 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 4.7-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes. 250
B.43 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 5.8, fitted to log(chlorophyll$_a$) data for Lake Balaton. 250
B.44 Residuals versus fitted values and theoretical versus sample quantiles of the distribution of the residuals of model 5.8, fitted to log(chlorophyll$_a$) data for Lake Erie. 251

Chapter 1

Introduction and background

This introductory chapter presents the motivation for the research described in this thesis. The research is associated with the GloboLakes project, which is a consortium project investigating the state of lakes and their responses to environmental change, on a global scale (GloboLakes 2016). The project provided ecological data from several lakes for analysis, which helped to motivate the main questions of interest for this research. The background to the research is presented, followed by a brief review of the current relevant literature. Finally, the thesis structure is presented.

1.1 Background to the research

Lakes are complex ecosystems, playing vital roles in both the global hydrological and biogeochemical cycles and acting as important parts of the global biosphere (Williamson et al. 2009). However, they are vulnerable to climate change and their responses to climate forcing, eutrophication and other pressures, and their possibly confounding interactions, are not yet fully understood (Ormerod et al. 2010). Many lakes have been studied in-depth individually, but a global picture has not yet been built up of how the global patterns of lake health are changing over space and time.

This research is associated with GloboLakes (www.globolakes.ac.uk), which

is a five-year long consortium project investigating the state of lake health and their response to changing environmental drivers on a global scale. The aim of GloboLakes is to gain a more in-depth understanding of lake-health through the production of a 20 year database of observed ecological parameters (GloboLakes 2016). The project facilitated interdisciplinary work between environmental scientists and statisticians.

GloboLakes provided data for various water quality variables, including both *in situ* and remotely-sensed data. *In situ* data are sampled directly from the lake surface, usually taken in a sampling tube from a boat at various predefined locations within a lake. These samples are taken to a laboratory and analysed to give data that are assumed accurate within measurement error. However, the cost involved, in terms of resources and in monetary terms, means that few sampling locations are available in each lake, if any at all.

In contrast, remotely-sensed data are available for a large number of lakes worldwide, with good spatial coverage and resolution. These data are obtained indirectly, often from satellites or aircraft. The remotely-sensed data used in this research are satellite data. These include temperature data, which were obtained from the ARC-Lake project (MacCallum & Merchant 2012, 2013). These data were acquired from the medium resolution imaging spectrometer (MERIS) on board the European Space Agency's ENVISAT satellite (ESA n.d. *a*). This Earth-observing satellite was launched on 1st March 2002, with contact lost on 8th April 2012, recording ten years of data (ESA n.d. *b*). The main functions of ENVISAT have begun to be replaced by satellites of the Sentinel programme. Remotely-sensed chlorophyll$_a$ and total suspended matter data are available from the Diversity II project, having been output from the Calimnos processing chain (Brockmann Consult et al. 2015). The benefit of these data is that the continuous monitoring by the satellite allows the entire lake to be monitored each month, with grid cells of up to 300m. resolution covering entire lakes. However, the fact that these data are indirectly obtained means that calibration is required to en-

sure accuracy, before these data can be used by environmental scientists to assess water quality.

Chlorophyll$_a$ is a variable of great importance to environmental scientists in their understanding of lake health. Chlorophyll$_a$ is produced by plants in order to absorb their required energy from sunlight. It can be understood as a proxy for phytoplankton biomass (Kasprzak et al. 2008), which is an indicator of lake health. Higher levels of chlorophyll$_a$ can be caused by higher levels of cyanobacteria within the lake water, often caused by high nutrient levels, so it is of importance to water quality investigators to know where and when high levels of chlorophyll$_a$ are occurring (Büttner et al. 1987).

Total suspended matter is a water quality variable, measuring the weight of material left in a filter after pouring a sample of water through it. The variable is a measure of how dense suspended particles are in a sampled water body, with low values generally meaning clearer or cleaner water. High phytoplankton levels may cause high levels of total suspended matter and so high levels of this variable can mean poorer water quality. In addition, suspended matter may be regarded as a pollutant, reducing drinking water quality and damaging fish habitats (Paul et al. 1982). Understanding total suspended matter levels is important for controlling sedimentation (Büttner et al. 1987), which is the process of deposition of sediment on the lake floor.

Lake surface water temperature is not a direct measure of water quality, but it does affect water quality indirectly, with phytoplankton blooms occurring at certain times of the year, mostly due to favourable temperatures. Lake surface water temperature has much less variation over space, within each lake, than the other measured variables, but it may still be a useful indicator for understanding lake water quality.

1.2 Research aims and objectives

Given the fact that two data sources are available, one with good spatial coverage, but requiring calibration, and the other available only for several point locations, but assumed accurate within measurement error, the key question of interest that this research will attempt to answer, is:

- How can data fusion be performed, to make use of the high resolution spatial and temporal information from the remotely-sensed data, calibrated over space and time with the *in situ* data?

An exploratory analysis of the available data was carried out, in order to gain an understanding of the spatiotemporal support of the data, to investigate patterns in variables over space and time, and to investigate the relationship between data for related variables. Based upon the understanding gained from this analysis, the following objectives were defined:

1. Statistical downscaling: Investigate the application and development of statistical downscaling techniques for the calibration of remotely-sensed data using *in situ* data.

2. Bivariate statistical downscaling: Develop a new framework for bivariate statistical downscaling, allowing the sharing of information between variables, to increase the understanding of the relationship between *in situ* and remotely-sensed data, and increasing calibration accuracy.

3. Multiple-lakes downscaling: Develop methodology for a novel framework of multiple-lakes downscaling, where data are downscaled for multiple lakes simultaneously, sharing information across lakes.

4. Spatiotemporal support: Develop novel methodology to allow the fusion of data with different spatial and temporal supports, specifically temporally-averaged, grid-scale remotely-sensed data and point-time, point-location *in situ* data.

Each of these objectives will be achieved through statistical model development and assessment, and data analyses, to gain an understanding of the performance and suitability of the models and modelling frameworks.

While $\log(\text{chlorophyll}_a)$ is a novel application for statistical downscaling, the statistical novelty in this work is the development of nonparametric statistical downscaling, which answers the question of how statistical downscaling can be applied to data of different spatiotemporal support. The method can be applied to a wide range of lake data, which could not otherwise be analysed in this way. The novel methodology developed in this thesis could be applied in a wide range of areas of study beyond that of lake water quality, allowing the fusion of information from diverse sources of data of different spatiotemporal support.

1.3 Structure of this thesis

This thesis is divided into the following chapters:

1. Introduction and background: This chapter contains the background information to the research project, giving the aims and objectives, an introduction to the data and a review of the relevant literature.

2. Exploratory analysis: This chapter contains exploratory analyses of the *in situ* and remotely-sensed data for Lake Balaton, investigating each dataset separately to gain an understanding of the spatial and temporal patterns within and between variables. The *in situ* data are investigated using mixed-effects models, since there are only a small number of data locations, while remotely-sensed data are investigated using kriging and additive models, since these data have good spatial coverage and so a model that takes spatial information into account is appropriate. A discussion of common spatial and temporal patterns in the remotely-sensed data, understood from S- and T-mode principal component analysis, is also included.

3. Statistical downscaling: This chapter introduces statistical downscaling using a preliminary spatial-only model, before going on to deal with spatiotemporal models, allowing the sharing of information on the spatial relationship between *in situ* and remotely-sensed data over time.

4. Bivariate and multiple-lakes statistical downscaling: This chapter discusses bivariate downscaling, where information is shared between variables through simultaneous downscaling, and multiple-lakes downscaling, where information is shared between nearby lakes, to discover whether these can improve the estimation of the model parameters and therefore improve the accuracy of the model predictions.

5. Nonparametric statistical downscaling: This chapter introduces the statistically novel framework of nonparametric statistical downscaling, which takes account of both the spatial and temporal changes of support, through the use of methodology from functional data analysis, allowing the calibration of remotely-sensed data using *in situ* data that are collected at different times at each point location.

6. Discussion and conclusions: This chapter sums up what has been achieved by the research, any limitations, further work required and important findings to highlight.

Chapters 3 to 5 include both methodological developments and data analyses.

1.4 Introduction to the data

Lake water quality data are collected frequently in sampling programs by water management agencies, for example the Balaton Limnological Institute and the United States Environmental Protection Agency, while remotely-sensed data are acquired from Earth-observing satellites operated by various space agencies, such as the European Space Agency. Some of these data

are available publicly, but the data used in this research have been made available by the GloboLakes project. This section will describe in more detail the data used in this research, focussing on the spatiotemporal support and data quality.

1.4.1 *In situ* data

This research makes use of *in situ* data for Lake Balaton, in Hungary, and the Great Lakes of North America.

Lake Balaton data

Lake Balaton, Hungary, is the largest lake in Central Europe (Büttner et al. 1987). The lake is long and shallow (see Figure 1.1), with a surface

Figure 1.1: Map of Lake Balaton, showing the nine *in situ* sampling locations. Map ©OpenStreetMap contributors (www.openstreetmap.org).

area of around 596km^2 and a mean depth of just 3.3m (Palmer et al. 2015). The lake has four basins along its length, which behave somewhat separately in terms of their hydrology (Tátrai et al. 2000). The main inflow is the River

Zala, which flows into the westernmost basin, and the main outflow is the Sió Canal, which leaves the lake partway along the easternmost basin (Tátrai et al. 2000). The Tihany peninsula juts into the lake between the third and fourth basins.

The lake has suffered greatly from poor water quality over the years, due to overly high levels of nutrients. These high nutrient levels cause blooms of cyanobacteria. These cyanobacteria are of concern, since a high proportion of them are toxic to humans, farm animals, birds and fish (Bláha et al. 2009). Additionally, the bacteria can form mats over the water surface for certain levels of light, nutrients and temperature, leading to fish kills (Teta et al. 2017). These cyanobacteria are used as indicators of degraded water quality and can be monitored from satellite imagery (Teta et al. 2017). Nutrient levels in Lake Balaton increased in the 1960s and 1970s, following developments in farming and a population increase in the surrounding region, causing the lake to become eutrophic (Padisák & Reynolds 1998). In recent years, steps have been taken to improve the water quality, including the Kis-Balaton wetland project at the mouth of the River Zala, which aims to reduce the levels of nutrients reaching the lake (Tátrai et al. 2000). The highest nutrient levels are found in the westernmost basin, with the easternmost basin having improved water quality (Tátrai et al. 2000).

The Balaton Limnological Institute (BLI) was founded with the aim of studying and understanding water quality in Lake Balaton (BLI n.d.). The institute has collected data on chlorophyll$_a$ concentration, total suspended matter and temperature, for five locations, between the start of 2006 and the end of 2011. Data were also collected by the Central Transdanubian Water and Environment Management Board (KDKVI) for four locations between the start of 2002 and the start of 2012, although temperature and total suspended matter data are only available until the end of 2006. This gives a total of 9 available *in situ* locations for this lake (see Figure 1.1). These data were provided for this research by the University of Stirling and GloboLakes

(www.globolakes.ac.uk).

Great Lakes data

The Great Lakes system of North America contains some of the largest freshwater lakes in the world (see Figure 1.2). The Great Lakes themselves

Figure 1.2: Map of the Great Lakes, showing the *in situ* sampling locations. Map data ©2016 Google (www.google.co.uk/maps).

are (from the highest altitude to the lowest altitude) Lakes Superior, Michigan, Huron, Erie and Ontario. Their mean surface areas are 82,100km^2, 57,800km^2, 58,600km^2, 25,700km^2 and 18,960km^2, while their mean depths are 147m, 85m, 59m, 19m and 86m, respectively (Botts & Krushelnicki 1995). Lakes Michigan and Huron may be considered as a single lake, since the flow between them, through the straits of Mackinac, can be in either direction (Sellinger et al. 2008). Water from Lake Superior flows into Lake Huron via the St. Mary's River, while water from Lake Huron flows into Lake Erie via

the St. Clair River, Lake St. Clair and the Detroit River. Water in turn leaves Lake Erie via the Niagara River and the Welland Canal, to flow into Lake Ontario. This means that any upstream changes in water quality are carried down to Lakes Erie and Ontario, with Erie being especially affected due to its shallow depth (Botts & Krushelnicki 1995).

Like Lake Balaton, the health of the Great Lakes system has been at risk in recent years. The lakes have been affected by increases in human settlements, farming and industrialisation along the lake shores. Since water outflow from the lakes is slow, pollutants remain within each lake for long periods of time. Lake Superior has little pollution entering it, but Lakes Michigan, Erie and Ontario have large urban areas on their shores, along with areas of intensive agriculture and industry. Runoff from mills and shoreline erosion continue to affect water quality (Botts & Krushelnicki 1995).

The United States Environmental Protection Agency (EPA) carries out a sampling program within the Great Lakes, twice each year (usually in April and October). Data are available for 19, 11, 14, 20 and 8 sampling locations respectively (see the red circles in Figure 1.2), for Lakes Superior, Michigan, Huron, Erie and Ontario, between August 2002 and April 2012. These data are available for chlorophyll$_a$ concentration. Additional chlorophyll$_a$ data are available from the Lake Erie Committee of the Great Lakes Fishery Commission (LEC), for 23 locations within Lake Erie (see the blue triangles in Figure 1.2), between April 1999 and October 2011, sampled throughout the year between Spring and Autumn. These data are all available from the Great Lakes Monitoring website (greatlakesmonitoring.org).

1.4.2 Remotely-sensed data

Remotely-sensed data have been provided by the Diversity II project (www.diversity2.info) and ARC-Lake (www.laketemp.net). These data are converted using various algorithms from reflectance data from Earth surface-facing satellites (Duan et al. 2012, Simis et al. 2005, Matthews et al. 2012,

Brockmann et al. 2004, Brockmann Consult et al. 2015), which leads to the loss of uncertainty information from the original data. Each satellite makes frequent passes over each lake. A monthly-average (or a fortnightly-average, for the ARC-Lake temperature data) is then calculated and made available. The data resolution is fairly good for the ARC-Lake temperature data, with 41 pixels covering Lake Balaton at a resolution of 0.05°, comparing favourably, in terms of spatial coverage, to the 9 *in situ* locations. Data are available for chlorophyll and total suspended matter, from Diversity II. These data have a much higher resolution than the temperature data, with 7616 grid cells covering Lake Balaton, at a resolution of approximately 300 metres. These data provide precise information on the spatial patterns in the variables of interest to environmental scientists. However, they must first be calibrated to ensure their accuracy.

The data for chlorophyll$_a$ concentration and total suspended matter were found to be positively skewed, so the natural log-transformed data will be used in this research. Temperature data displayed no such skewness, so will be left untransformed.

It was noted that some extreme values appeared in the Diversity II data around the lake edges, which could indicate land contamination. This is where the satellite picks up areas of land within a grid cell containing water, meaning that the reflectance measurements are unreliable for these grid cells (Parkinson 1997, pages 50 to 51). This issue can be dealt with simply by removing some data around the lake edges. For Lakes Balaton and Erie, 2 grid cells are removed from the lake edges, while for the other Great Lakes, 6 cells are instead removed. I made these decisions, based upon evidence from exploratory plots, which show some grid cells around the lake edges with very high or low data values in comparison to their neighbouring cells, with these cells appearing slightly further into the lake for Lakes Superior, Michigan and Huron. Removing these grid cells still leaves a large number of grid cells covering each lake. This procedure was not carried out for the

temperature data, since these data have a much larger size of grid cell, so that any land contamination within each cell should be small compared to the amount of water in each cell. Additionally, removing grid cells around the lake edges would remove the data for most of the 41 grid cells for Lake Balaton, due to its thin shape, leaving few data for analysis.

1.5 Literature review

In this section, the relevant literature is summarised and discussed. Since the main aim of this thesis is to present methodology for the fusion of data of different spatiotemporal support, this review focusses on the literature that is relevant to this topic. A brief discussion of data fusion is presented, followed by noting methodology that may be of use.

1.5.1 Data fusion

Data fusion has been described as "a process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve refined position and identity estimates, and complete and timely assessments of situations and threats as well as their significance" (White 1991) and "the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making" (Khaleghi et al. 2013). Hall & Llinas (1997) note that combining same-source data can provide a statistical advantage, since improved estimates of physical phenomena can be obtained via redundant observations, while the use of multiple types of sensors can improve the accuracy with which a quantity can be observed and characterised. Data fusion is used in many sectors, such as equipment monitoring, medical, military and remote sensing applications (Hall & Llinas 1997). Data fusion can therefore be considered a general term that covers a number of different

approaches.

Data fusion for chlorophyll

For remote sensing, data fusion is becoming increasingly important, with improving sensor technologies leading to greater numbers of available data (Schmitt & Zhu 2016). For chlorophyll$_a$ data specifically, there have been many applications of data fusion in recent years:

- Kneubühler et al. (2007) use a linear model to fuse reflectance band data from MERIS, for Lake Kivu, with very limited in-lake data.

- Doña et al. (2015) focus on the fusion of Landsat data with MERIS and MODIS data for the highly eutrophic lake Albufera de Valencia. The Landsat data have high spatial resolution, but low temporal and spectral resolution, while the MERIS and MODIS data have higher temporal and spectral resolution, but low spatial resolution. After converting the Landsat data onto the surface reflectance scale and converting the MERIS/MODIS data onto the 30 metre scale of Landsat observations (through nearest neighbour resampling, where each small-scale pixel is assigned the value of its nearest large-scale pixel), they fuse the two datasets through a pixel-by-pixel algorithm. A further process is the calibration of the data using in-lake observations, using a genetic programming model, which fits various sets of nonlinear modelling equations to a training dataset to produce a good fit to the data and automatically keeps the best set (based upon predictions made and compared to a validation dataset) (Doña et al. 2015).

- Sakuno (2013) fuses daily frequency MODIS data with hourly frequency GOCI data, for Tachibana Bay. After some data preprocessing to convert the GOCI and MODIS data onto the chlorophyll$_a$ observation scale, the author fits a linear model to the data, with MODIS as response, so that predictions can be made from GOCI data on the same

scale. The results suggest that a linear model is not enough to obtain good predictions across the whole study region, so that some kind of spatially-varying method may be more appropriate.

- Kwiatkowska & Fargion (2002) outline issues in data fusion for sea chlorophyll$_a$ data from SeaWIFS and MOS datasets. These data have different spatial resolutions and must first be projected onto a rectangular grid from their original longitude and latitude grid, with the MOS data being subsampled onto the SeaWIFS grid. The fusion takes place through wavelet multiresolution analysis, where wavelet transforms are applied to the satellite images.

Data fusion for air quality data

In recent years, much work has gone into developing methodology for the fusion of air quality data from multiple sources, where measured and modelled data are commonly available (with the measured data being *in situ*, ground truth data, and the modelled data are available on a grid scale, so that the modelled data are equivalent to the remotely-sensed lake water data in this application):

- Wikle & Berliner (2005) present a Bayesian model for data of different resolutions and fit it to wind data.

- Fuentes & Raftery (2005) present a Bayesian melding model for fusing modelled weekly sulphur dioxide concentrations with observed data from the CASTNet network, where neither dataset is assumed measured without error, but where there is an assumed true process on the point-scale. The change-of-support between the data types is dealt with through fitting integrals, which can be computationally intensive.

- Cressie & Johannesson (2008) develop fixed rank kriging, which represents large-scale spatial data using basis functions, such as smoothing

splines, wavelets or radial basis functions, enabling efficient computation. This methodology is updated by Nguyen et al. (2012), in order to fuse MODIS and MISR data for southern Africa.

- Berrocal et al. (2010*b*) present a statistical downscaling model, based upon Gelfand et al. (2003), which fuses grid-scale modelled ozone data from the Community Multi-scale Air Quality Model with point-scale *in situ* observed ozone data, for the eastern USA. They fit a Bayesian regression model, with spatially-varying coefficients, in order to relate the two sets of data, treating the *in situ* data as accurate. This methodology has been adapted to allow simultaneous fusion of two related variables (Berrocal et al. 2010*a*), to allow for mismatch between the assumed and true locations of data (Berrocal et al. 2012) and to model exceedances of ozone level standards (Berrocal et al. 2014). Additionally, a version of the model was applied by Paci et al. (2013) and Bruno & Paci (2014) and was incorporated within the model of Rundel et al. (2015). Wilkie et al. (2015) applied a version of the model to $\log(\text{chlorophyll}_a)$ data. These models do not include structured temporal dependencies, so do not explicitly take the temporal structure of the data into account.

- Sahu et al. (2010) present an alternative statistical downscaling model, which combines a conditional autoregressive model for gridded air quality modelled output with a space-time process model for observed point level data. They link the model components using latent space-time processes within a Bayesian hierarchical model, with prediction using the point-level model. This model explicitly incorporates a space-time component.

Despite the fact that this thesis focusses on $\log(\text{chlorophyll}_a)$ data, the methodology from the air quality data proves useful for this application. An important difference between the datasets for air quality and for lake water quality

data is that obtaining the *in situ* measurements for lake water quality variables is particularly expensive and so the available *in situ* data may be sparse in comparison to those for air quality. For example, there may be very few *in situ* sampling sites available across each lake, giving sparse spatial coverage of the lake surface. Additionally, data may not be collected regularly at all locations, since samples can only be taken when a boat can safely access sampling sites, leading to irregular *in situ* data. Another point to note is that, although remotely-sensed log(chlorophyll$_a$) data can be considered in the same way as modelled air quality data, they are based upon observed data, from Earth-surface reflectance measurements, so that the *in situ* and remotely-sensed data are assumed to have a positive relationship. The following chapters will make use of the statistical downscaling methodology from the air quality literature, but these differences in the data properties will be taken into account.

1.5.2 Downscaling literature

The previous subsection leads to a discussion of methodology for statistical downscaling. Originally developed in the climate modelling literature, statistical downscaling has begun to be developed for the data fusion of grid-scale modelled and observed *in situ* data (Berrocal et al. 2010*b*). The methodology was originally developed for adapting the coarse resolution of global climate models, of hundreds of kilometres, to the much smaller scale, needed for impact assessment and understanding of the processes (Wilby & Wigley 1997, Maraun et al. 2010). Two main approaches are dynamical downscaling and statistical downscaling (Maraun et al. 2010). Dynamical downscaling involves the nesting of a smaller-scale regional climate model into the global climate model, with output on resolutions as high as several kilometres (Maraun et al. 2010). On the other hand, Maraun et al. (2010) state that statistical downscaling involves the establishment of statistical links between large and local-scale weather patterns, although a more

general definition could refer simply to a process, rather than specifically weather.

Dynamical downscaling

Dynamical downscaling is where a numerical model with small-scale output is nested within a larger-scale numerical model. Taking the example of regional climate models, these have a higher resolution than global climate models (50km. or less compared to hundreds of kilometres) (Maraun et al. 2010).

Schmidli et al. (2007) show that dynamical downscaling may be preferred to statistical downscaling in some climate modelling circumstances. However, dynamical downscaling can only produce output on a grid-cell (areal) scale and so statistical downscaling will be focussed on here (in the context of data fusion of satellite and *in situ* lake data).

Statistical downscaling

Statistical downscaling maps a large-scale predictor $\mathbf{X}$ to a local-scale predictand $\mathbf{Y}$, in the form:

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = f(\mathbf{X}, \boldsymbol{\beta}), \tag{1.1}$$

where $\boldsymbol{\beta}$ is a vector of unknown parameters, which is estimated to calibrate the downscaling scheme (Maraun et al. 2010), $\mathrm{E}(\mathbf{Y}|\mathbf{X})$ represents the expected value of $\mathbf{Y}$, given $\mathbf{X}$ (where "|" means "given"), and $f(\mathbf{X}, \boldsymbol{\beta})$ is a function of $\mathbf{X}$ and $\boldsymbol{\beta}$. Unexplained variability may also be explicitly modelled as a random variable (Maraun et al. 2010).

The three types of statistical downscaling defined by Maraun et al. (2010) are perfect prognosis (PP), model output statistics (MOS) and weather generators (WGs). Wilby & Wigley (1997) present an alternative classification of downscaling methods in their review paper, categorising them into regres-

sion methods, weather-pattern-based approaches, stochastic weather generators and limited area modelling. However, the classification scheme used in the more recent review paper of Maraun et al. (2010) will be described here.

PP approaches are classical downscaling approaches, including regression models and weather pattern-based approaches, which establish a relationship between observed large-scale predictors and observed local-scale predictands. These relationships can be applied to numerical model predictors, if they are realistically simulated. Weather sequences of predictors and predictands can be related to each other event by event, instead of only relating distributions of predictors and predictands to each other (Maraun et al. 2010). Model selection should use statistical criteria, to avoid over- or underfitting. PP approaches ignore physical processes on scales between the large and local scales (Maraun et al. 2010).

MOS approaches develop relationships between output from a medium-scale numerical model (e.g. an RCM) and local-scale observed variables (Maraun et al. 2010).

WGs are statistical models generating local-scale weather time series, resembling the statistical properties of observed weather. WGs can condition parameters on large-scale weather, meaning that they are hybrids between unconditional weather generators and PP methods, or they can be calibrated only against local-scale observations, in which case they are not true downscaling methods (Maraun et al. 2010).

For PP downscaling, informative predictors should be selected, which have high predictive power. They can be identified by correlating possible predictors with the predictands. Predictors must be well simulated by the driving dynamical models and the relationship between predictors and predictands must be stationary (temporally stable) (Maraun et al. 2010).

Raw predictors are usually high dimensional fields of grid-based values. Information at neighbouring grid points is not independent. The predictor field can be decomposed into modes of variability and its dimensionality can

be reduced, e.g. by principal component analysis (PCA) (Maraun et al. 2010). PCA gives a set of orthogonal basis vectors, allowing a large part of variability in the original predictor field to be represented in lower dimensions. The predictands are not taken account of, so that the correlation between predictors and predictands may not be optimal. Sari et al. (2017), for example, use functional PCA to reduce dimensionality, before using quantile regression to estimate extreme monthly rainfall. Canonical correlation analysis and maximum covariance analysis are alternative methods, which take into account the predictand field, such that the temporal correlation between the predictor and predictand fields is maximal (Maraun et al. 2010). This dimensionality reduction approach is most appropriate for applications with many variables.

Possible statistical models for PP downscaling include linear regression, generalised linear and additive models, vector generalised linear models, weather-type based downscaling, nonlinear regression and the analog method (Maraun et al. 2010).

1.5.3 Conclusions from this literature review

This literature review presented background to the topic of data fusion, focussing on relevant ideas from air quality and chlorophyll$_a$ data fusion studies. This thesis focusses on fusing point-scale *in situ* data and grid scale remotely-sensed data, which are already on the same measurement scale (after conversion from Earth surface reflectance measurements, for the remotely-sensed data), so that only the actual fusion process is to be carried out. The method of statistical downscaling, through a spatially-varying coefficient regression, is particularly relevant and is focussed on. However, concerns specific to lake water quality data, related to the different spatiotemporal support of the *in situ* and remotely-sensed data, are taken into account in order to develop the relevant novel methodology.

In the following chapters, additional literature is presented at appropriate points.

1.6 Spatial and temporal modelling

This section explores techniques for spatial and temporal modelling, interpolation and prediction that are commonly used in environmental data analysis. The methodology is divided into two sections, geostatistics and nonparametric smoothing. Both of these groups of methods allow prediction at new locations, given data indexed on a spatial scale, and both allow prediction at new times, given spatiotemporally-indexed data. Nonparametric smoothing also enables the modelling of non-linear relationships between variables, which may or may not be spatially-indexed. The following subsections describe the main features of both geostatistics and nonparametric smoothing.

1.6.1 Geostatistics

This subsection introduces geostatistics, which is a branch of spatial statistics, where the data are a finite sample of measured values that relate to a spatially continuous process (Diggle & Ribeiro 2007). The spatial structure of the data is assessed through variogram modelling, with spatial prediction carried out through the parametric method of kriging.

Variogram modelling

Let $Z(\mathbf{s})$ be a random field, indexed by spatial location $\mathbf{s}$. Fitting the variogram, which is a "model-based measure of spatial statistical dependence in a geostatistical process" (Cressie & Wikle 2011), enables the modelling of spatial correlation. Assuming that only a single data point is available at each spatial location, the usual method of calculating correlation cannot be carried out. Instead, some assumptions must be made. Intrinsic stationarity assumes that variance over the spatial surface is constant and that correlation

does not depend on spatial location (Bivand et al. 2013), so that:

$$Z(\mathbf{s}) = \mathbf{m} + \mathbf{e}(\mathbf{s}),$$

where $\mathbf{m} = \mathrm{E}\big(Z(\mathbf{s})\big)$ is a constant mean, with random errors $\mathbf{e}(\mathbf{s})$. The variogram is therefore written as:

$$\boldsymbol{\gamma}(\mathbf{h}) = \frac{1}{2}\mathrm{E}\big(Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\big)^2,$$

(Bivand et al. 2013). Isotropy assumes that correlation does not depend on direction (Bivand et al. 2013), so that $\mathbf{h}$ is replaced by distance $d = ||\mathbf{h}||$. With these assumptions, there are now multiple pairs of data, with almost identical separation distances, from which correlation is estimated as a function of distance.

Distances $d_1, \ldots, d_n$ are grouped, or "binned", into sets of similar distances $\tilde{\mathbf{d}}_j$, for $j = 1, \ldots, m$, where m is the number of bins. For each bin $\tilde{\mathbf{d}}_j$, the sample variogram is then calculated from n_j pairs of data $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_i + d)$ $(i = 1, \ldots, n_j)$ that are approximately distance interval $\tilde{\mathbf{d}}_j$ apart. The value of the sample variogram for bin $\tilde{\mathbf{d}}_j$ is:

$$\hat{\boldsymbol{\gamma}}(\tilde{\mathbf{d}}_j) = \frac{1}{2n_j}\sum_{i=1}^{n_j}\big(Z(\mathbf{s}_i) - Z(\mathbf{s}_i + d)\big)^2 \tag{1.2}$$

$\forall d \in \tilde{\mathbf{d}}_j$ (Bivand et al. 2013).

The nugget, sill and range of the variogram, which are displayed in Figure 1.3 for an example fitted variogram, are related to certain important characteristics of the spatial distribution of the data. The nugget is the value of the semivariance at distance zero and represents the measurement error or small-scale variability (Bivand et al. 2013). Semivariance increases with distance, until it reaches the sill, which is the variance of the observation process $Z(\mathbf{s})$ (Diggle & Ribeiro 2007). The partial sill is the difference between the sill and the nugget, representing the variance of the signal (Diggle & Ribeiro

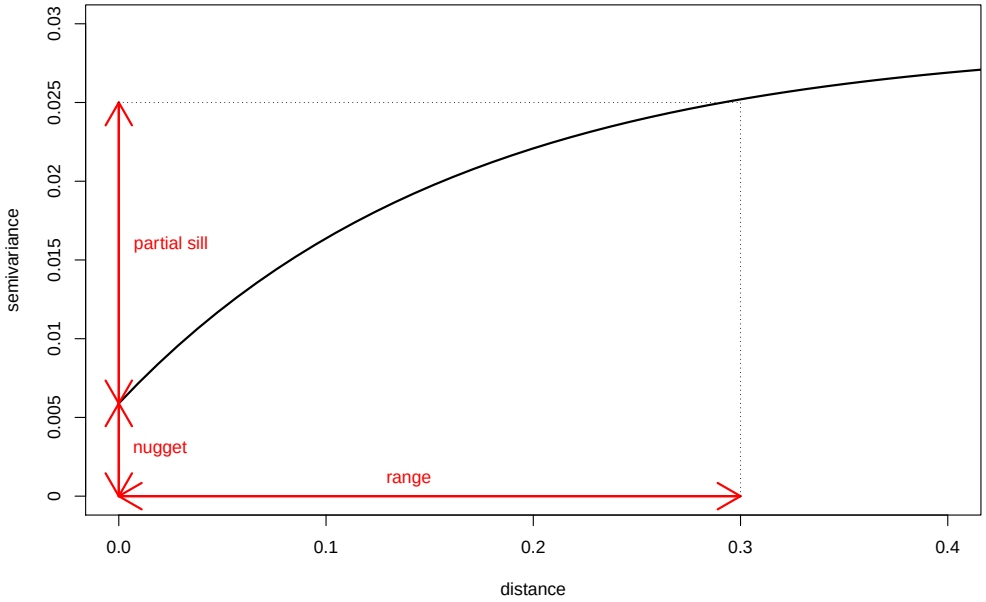

Figure 1.3: Example of a fitted variogram, showing the partial sill, nugget and range. The sill has value equal to that of the partial sill and nugget added together.

2007). The range is the distance at which the semivariance reaches the sill (Bivand et al. 2013). The correlation between the data decreases to zero for locations at least as far apart as the range.

In order to make use of the variogram for prediction and interpolation, a variogram model is fitted to the sample variogram. This ensures that the matrix of semivariances between observation points and possible prediction points is nonnegative definite, so that the prediction variances are guaranteed to be nonnegative (Bivand et al. 2013). Various parametric variogram models can be fitted, including those in the Matérn family and the powered exponential family (Cressie & Wikle 2011). The Matérn family of semivariance functions $\gamma(\mathbf{h};\boldsymbol{\theta})$ is defined as:

$$\boldsymbol{\gamma}(\mathbf{h};\boldsymbol{\theta}) = C(\mathbf{0};\boldsymbol{\theta}) - C(\mathbf{h};\boldsymbol{\theta}), \tag{1.3}$$

for $\mathbf{h} \in \mathbb{R}^d$, where:

$$C(\mathbf{h};\boldsymbol{\theta}) = \sigma_0^2 \mathrm{I}(d = 0) + \sigma_1^2 \big(2^{\kappa-1}\Gamma(\kappa)\big)^{-1} (d/\phi)^{\kappa} \mathrm{K}_{\kappa}(d/\phi),$$

where $\mathrm{I}(d = 0)$ is the indicator function that equals 1 if $d = 0$ and 0 otherwise, $\kappa > 0$ is the order parameter, determining the smoothness of the underlying spatial process, $\phi > 0$ is the scale parameter, determining how fast correlation decreases to zero as distance between data locations increases, and $\mathrm{K}_\kappa(d/\phi)$ is a modified Bessel function of the second kind, of order κ (Diggle & Ribeiro 2007, Cressie & Wikle 2011). $C(\mathbf{0}; \boldsymbol{\theta}) = \sigma_0^2 + \sigma_1^2$ (Cressie & Wikle 2011). Members of the Matérn family include the exponential variogram for $\kappa = 0.5$, where:

$$C(\mathbf{h}; \boldsymbol{\theta}) = \sigma_0^2 \mathrm{I}(d = 0) + \sigma_1^2 \exp(-d/\phi),$$

and the Gaussian variogram as $\kappa \to \infty$, where:

$$C(\mathbf{h}; \boldsymbol{\theta}) \to \sigma_0^2 \mathrm{I}(d = 0) + \sigma_1^2 \exp\big(-(d/\phi)^2\big),$$

(Diggle & Ribeiro 2007). The powered exponential family of semivariance functions $\gamma(\mathbf{h}; \boldsymbol{\theta})$ is defined as:

$$\boldsymbol{\gamma}(\mathbf{h}; \boldsymbol{\theta}) = C(\mathbf{0}; \boldsymbol{\theta}) - C(\mathbf{h}; \boldsymbol{\theta}), \tag{1.4}$$

for $\mathbf{h} \in \mathbb{R}^d$, where:

$$C(\mathbf{h}; \boldsymbol{\theta}) = \sigma_0^2 \mathrm{I}(d = 0) + \sigma_1^2 \exp\big(-(d/\phi)^\kappa\big).$$

It can be seen that this family also includes both the exponential (for $\kappa = 1$) and Gaussian (for $\kappa = 2$) variograms.

The variogram model can be fitted by weighted least squares (WLS), maximum likelihood (ML) or restricted maximum likelihood (REML). Cressie (1985) shows that weighted least squares is an appropriate way to fit variogram models, although Diggle & Ribeiro (2007) recommend using ML or REML. WLS fits the variogram model using the sample variogram, whereas ML and REML fit an explicit model directly to the data. As ML and REML

are likelihood-based, the distribution of the data should be checked for these methods (Diggle & Ribeiro 2007). The best method to use can depend on the form of the variogram, with those with a small nugget variance and a large correlation range found to be better fitted by least squares methods (Lark 2000).

Kriging

Kriging is a method for the interpolation of spatial data, from the Geostatistics literature. Given a spatial process $Z(\mathbf{s})$ for which realisations $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ are available for n spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, kriging aims to predict the value of the process at a new location $\mathbf{s}_0$, i.e. $Z(\mathbf{s}_0)$, along with an associated measure of uncertainty. The "best" predictor $\hat{Z}(\mathbf{s}_0)$ should be the best linear unbiased predictor (BLUP). The predictor $\hat{Z}(\mathbf{s}_0)$ is the BLUP of $Z(\mathbf{s}_0)$, if (Piegorsch & Bailer 2005):

1. It is linear, i.e. it takes the form $a_0 + \sum_{i=1}^{n} a_i Z(\mathbf{s}_i)$, for some known coefficients a_0, a_1, ..., a_n.

2. It is unbiased, i.e. $\mathrm{E}\big(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\big) = 0$.

3. $\mathrm{var}\big(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)\big)$ is the minimum among all linear unbiased estimators of $Z(\mathbf{s}_0)$.

The three types of kriging are simple (where it is assumed that $Z(\mathbf{s}_i) = \mu + \varepsilon(\mathbf{s}_i)$, with μ known and stationary), ordinary (where $Z(\mathbf{s}_i) = \mu + \varepsilon(\mathbf{s}_i)$, with μ unknown and stationary) and universal (where μ is allowed to vary with $\mathbf{s}$) (Piegorsch & Bailer 2005, Cressie 1993). Of these three methods, universal kriging is focussed on here, since it can be applied in a wider range of circumstances, including where there is a trend in the mean level μ.

The linear predictor is written as:

$$\hat{Z}(\mathbf{s}_0) = a_0 + \sum_{i=1}^{n} a_i Z(\mathbf{s}_i), \tag{1.5}$$

where the coefficients $a_0, ..., a_n$ are chosen to minimise the mean squared prediction error (MSPE), $\mathrm{E}\left((\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2\right)$. For universal kriging, let $\mu_Z(\mathbf{s}_0) = \mathrm{E}(\mathbf{s}_0)$ be the spatially-varying mean and let $C_Z(\mathbf{s}, \mathbf{t})$ be the covariance between $Z(\mathbf{s})$ and $Z(\mathbf{t})$. Then the universal kriging operator that minimises the MSPE is:

$$\hat{Z}(\mathbf{s}_0) = \mu_Z(\mathbf{s}_0) + \sum_{k=1}^{n} a_k\big(Z(\mathbf{s}_k) - \mu_Z(\mathbf{s}_k)\big), \tag{1.6}$$

where $\mathbf{a} = (a_1, ..., a_n)^{\mathrm{T}} = \sum_Z^{-1} C_Z(\mathbf{s}_0)$, $\sum_Z = \mathrm{cov}(\mathbf{Z}) = [C_Z(\mathbf{s}_i, \mathbf{s}_j)]_{i,j}$ and $C_Z(\mathbf{s}_0) = \big(C_Z(\mathbf{s}_0, \mathbf{s}_k) : k = 1, ..., n\big)^{\mathrm{T}}$. The universal kriging operator is the BLUP. This predictor attains a minimum value of MSPE of:

$$C_Z(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{a}^{\mathrm{T}} C_Z(\mathbf{s}_0), \tag{1.7}$$

which is the kriging variance.

1.6.2 Nonparametric smoothing

This subsection describes nonparametric smoothing. Smoothing methods from this framework can be fitted in both spatial and temporal contexts, so are particularly relevant for spatiotemporal data that have smooth patterns over space and time. Within this framework lie additive modelling and functional data analysis, two sets of methodology for nonparametric modelling of smooth functions. Additive modelling estimates the relationship between a response variable $\mathbf{y}$, i.e. $(y_1, \ldots, y_n)^{\mathrm{T}}$, which is not necessarily assumed to be smooth, and explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_p$, where $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})^{\mathrm{T}}$ $(j = 1, \ldots, p)$, where the relationship between $\mathbf{y}$ and each of the response variables is assumed to be smooth and non-linear. An additive model can be written as:

$$y_i = \beta_0 + f_1(x_{1i}) + ... + f_p(x_{pi}) + \varepsilon_i, \tag{1.8}$$

for $i = 1, \ldots, n$, where the f_j $(j = 1, \ldots, p)$ are smooth functions of the response variables $x_1, \ldots, x_p$, $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ are the independent random errors and β_0 is the intercept term (Wood 2006). Functional data analysis, on the other hand, is concerned with expressing a response variable $\mathbf{y}$ as a smooth function, where it is assumed that $y_1, \ldots, y_n$ are in fact observations of a smooth function f that takes values over an infinite-dimensional space (Ferraty & Vieu 2006), such that y_i is somewhat related to y_{i+1} $(i = 1, \ldots, n-1)$ (Ramsay & Silverman 2006). For the common example, where y_i is f evaluated at time t_i $(i = 1, \ldots, n)$, then:

$$y_i = f(t_i) + \varepsilon_i, \tag{1.9}$$

for $i = 1, \ldots, n$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ are the independent random errors (Ramsay & Silverman 2006).

Representing smooth functions

In order to fit either an additive model or a model in the functional data analysis framework, a method must be found to enable the representation of the smooth functions. Simple types of smoothing include bin smoothing, where the data for the explanatory variable $\mathbf{x}$ are partitioned into sets of values called "bins" and the mean value over each bin is calculated, in order to obtain smoothed estimates (Clarke et al. 2009). This method can be improved by changing the bin width to have equal numbers of observations per bin (moving average smoothing), or by fitting a straight line per bin (running line smoothing) (Clarke et al. 2009). A further improved method is LOESS (locally-weighted regression), which fits a weighted polynomial to data in each bin, providing smoother estimates (Clarke et al. 2009).

This section will focus, however, on smoothing via basis functions, and specifically regression splines. In order to ensure that the model is linear in the parameters, each smooth function f can be written in terms of basis

functions, defining the space of functions that includes an approximation to f as an element (Wood 2006). Let $\phi_k(x)$ be the kth basis function for variable x. Then the smooth function f is:

$$f(x) = \sum_{k=1}^{m} \phi_k(x) c_k, \tag{1.10}$$

where $\mathbf{c}$, i.e. $(c_1, \ldots, c_m)^{\mathrm{T}}$, is a vector of coefficients (Wood 2006). There are many possible basis types, with a list of commonly used types given by Ramsay & Silverman (2006):

- Monomials: $1, t, t^2, \ldots, t^m$.
- Fourier basis: suitable for periodic data.
- Splines (including B-splines, thin plate regression splines and cyclic cubic regression splines (Wood 2006)): computationally efficient and used for non-periodic data.
- Wavelets, exponential, power, polynomial, polygonal, step-function and constant bases.

The two types of basis functions that will be focussed on in this section are the Fourier basis and the spline basis, since both are computationally efficient, but splines are preferred for non-periodic data and the Fourier basis is preferred for periodic data.

A spline is a curve, made up of sections of polynomials. Specifically, a cubic spline is made up of cubic polynomials, joined so that they are continuous up to their second derivatives (Wood 2006). This leads to a fitted curve that appears smooth to the human eye (Hastie et al. 2001). The following paragraphs summarise different types of splines. The Fourier basis system is then described.

B-spline basis A spline is a smooth function, made up of polynomials of some specified degree, joined together at "breakpoints", or "knots" (Ramsay

& Silverman 2006). B-splines, short for "basis splines" (de Boor 1978), are defined, using a recurrence relation, as:

$$\mathbf{B}_{j,1}(x) = \begin{cases} 1 & \text{for } t_j \leq x < t_{j+1} \\ 0 & \text{otherwise} \end{cases}, \tag{1.11}$$
$$\mathbf{B}_{j,k}(x) = \frac{x - t_j}{t_{j+k-1} - t_j}\mathbf{B}_{j,k-1}(x) + \frac{t_{j+k} - x}{t_{j+k} - t_{j+1}}\mathbf{B}_{j+1,k-1}(x),$$

where $j = 1, \ldots, m$ indexes over the knot sequence $\mathbf{t} = t_1, \ldots, t_m$, with m being the basis dimension (i.e. the number of basis functions), where k is the order of the B-splines and where x is some value on the x-axis (de Boor 1978). A complete derivation of B-splines is available in de Boor (1978).

B-splines have several important properties, of which several are particularly important to note for this application, namely:

- (i) $\mathbf{B}_{j,k}$ has small support, i.e. $\mathbf{B}_{j,k}(x) = 0$ for $x \notin [t_j, t_{j+k}]$. Each B-spline is defined only over a small region and is zero outwith this region, so that only at most k B-splines are non-zero over each interval $[t_j, t_{j+1}]$ (for $j = 1, \ldots, m$) (de Boor 1978).

- (ii) $\sum_j \mathbf{B}_{j,k}(x) = \sum_{j=r+1-k}^{s-1} \mathbf{B}_{j,k}(x) = 1 \ \forall \ t_r < x < t_s$, i.e. the basis functions at any x sum to 1 (de Boor 1978). This is the general compact support property.

- (iii) $\mathbf{B}_{j,k}(x) > 0$ for $t_j < x < t_{j+k}$, i.e. $\mathbf{B}_{j,k}$ is positive on its support. Since $\mathbf{B}_{j,k}$ is made up of nonnegative functions that sum to 1 (from property (ii)), it is a "partition of unity" (de Boor 1978).

The first of these properties ensures computational efficiency, since the matrix of inner products of these basis functions is band-structured (with non-zero values only on the $(k-1)$ subdiagonals on either side of the main diagonal) (Ramsay & Silverman 2006).

Cubic B-splines are B-splines of order $k = 4$. As B-splines are continuous up to their $(k-2)$th derivative, cubic B-splines have a continuous first and

second derivative (Ramsay & Silverman 2006), so that they appear smooth to the human eye. In order to cause a discontinuity at each endpoint, an additional $(k-1)$ knots are placed at the locations of the two endpoint knots (Ramsay & Silverman 2006). This ensures that inferences are not made outwith the region of available data. Discontinuities can also be created at other locations along the x-axis, by placing additional knots at a location to cause discontinuity up to the desired level of derivative (Ramsay & Silverman 2006). This flexibility is an additional benefit of B-splines and is useful, for example, in the presence of a known changepoint.

Thin plate regression splines In order to fit splines to multiple variables at once, one approach is to use thin plate regression splines. An example where splines would be fitted to multiple variables is that of a spatial surface, where the value of the surface depends upon a combination of both longitude and latitude. The method of thin plate regression splines is based upon thin plate splines, but avoids the high computational cost of fitting thin plate splines (Wood 2003, 2006). In order to estimate the smooth function $g(\mathbf{x})$ from n observations $(y_i, \mathbf{x}_i)$ such that $y_i = g(\mathbf{x}_i) + \varepsilon_i$, where ε_i are random errors and the dimension of $\mathbf{x}$ is $d \leq n$, the function:

$$||\mathbf{y} - \mathbf{f}||^2 + \lambda J_{md}(f) \tag{1.12}$$

is minimised with respect to f, where $\mathbf{f} = \begin{pmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_n) \end{pmatrix}$ and $J_{md} = \int \cdots \int_{\mathbb{R}^d}$ $\sum_m \frac{m!}{v_1! \ldots v_d!} \left(\frac{\delta^m f}{\delta x_1^{v_1} \ldots \delta x_d^{v_d}} \right)^2 dx_1 \ldots dx_d$, where $m = \sum_{i=1}^{d} v_i$, is the penalty that controls excess curvature (Wood 2006). Letting $2m > d$, the function that minimises 1.12 is of the form:

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} \delta_i \eta_{md}\big(||\mathbf{x} - \mathbf{x}_i||\big) + \sum_{j=1}^{M} \alpha_j \phi_j(\mathbf{x}), \tag{1.13}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are estimated such that $\mathbf{T}^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0}$, where $\mathbf{T}_{ij} = \phi_j(\mathbf{x}_i)$, $M = \binom{m+d-1}{d}$ and:

$$\eta_{md}(r) = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}\, r^{2m-d}\log(r) & \text{for } d \text{ even}, \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & \text{for } d \text{ odd}. \end{cases}$$

Letting $\mathbf{E}$ be such that $\mathbf{E}_{ij} = \eta_{md}\big(||\mathbf{x}_i - \mathbf{x}_j||\big)$, thin plate splines are fitted by minimising:

$$||\mathbf{y} - \mathbf{E}\boldsymbol{\delta} - \mathbf{T}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\delta}^{\mathrm{T}}\mathbf{E}\boldsymbol{\delta},$$

such that $\mathbf{T}^{\mathrm{T}}\boldsymbol{\delta} = \mathbf{0}$, with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ (Wood 2006).

For thin plate regression splines, let $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}$ be the eigendecomposition of $\mathbf{E}$, such that $\mathbf{D}$ is the diagonal matrix of eigenvalues, ordered from smallest to largest, and $\mathbf{U}$ has the corresponding eigenvectors as its columns. Let $\mathbf{U}_k$ be the first k columns of $\mathbf{U}$ and let $\mathbf{D}_k$ be the first k rows and columns of $\mathbf{D}$. Then, letting $\delta = \mathbf{U}_k\boldsymbol{\delta}_k$, thin plate regression splines are fitted by minimising:

$$||\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\boldsymbol{\delta}_k - \mathbf{T}\boldsymbol{\alpha}||^2 + \lambda\boldsymbol{\delta}_k^{\mathrm{T}}\mathbf{D}_k\boldsymbol{\delta}_k,$$

such that $\mathbf{T}^{\mathrm{T}}\mathbf{U}_k\boldsymbol{\delta}_k = \mathbf{0}$, with respect to $\boldsymbol{\delta}_k$ and $\boldsymbol{\alpha}$ (Wood 2006). Finally, after finding an orthonormal column basis $\mathbf{Z}_k$ such that $\mathbf{T}^{\mathrm{T}}\mathbf{U}_k\mathbf{Z}_k = \mathbf{0}$ and letting $\boldsymbol{\delta}_k = \mathbf{Z}_k\tilde{\boldsymbol{\delta}}$, the problem reduces to minimising:

$$||\mathbf{y} - \mathbf{U}_k\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}} - \mathbf{T}\boldsymbol{\alpha}||^2 + \lambda\tilde{\boldsymbol{\delta}}^{\mathrm{T}}\mathbf{Z}_k^{\mathrm{T}}\mathbf{D}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}},$$

with respect to $\tilde{\boldsymbol{\delta}}$ and $\boldsymbol{\alpha}$, with 1.13 used to evaluate the spline, after first evaluating $\boldsymbol{\delta} = \mathbf{U}_k\mathbf{Z}_k\tilde{\boldsymbol{\delta}}$ (Wood 2006).

Cyclic cubic regression splines Cyclic cubic regression splines ensure that the spline function reaches the same value at each endpoint, which is necessary for cyclical data, such as daily temperature data that are measured

over several years. A cubic regression spline is defined by Wood (2006) as:

$$\begin{aligned} f(x) =& \frac{x_{j+1} - x}{xj + 1 - x_j} f(x_j) + \frac{x - x_j}{x_{j+1} - x_j} f(x_{j+1}) \\ &+ \left(\frac{(x_{j+1} - x)^3}{x_{j+1} - x_j} - (x_{j+1} - x_j)(x_{j+1} - x) \right) \times \frac{f''(x_j)}{6} \\ &+ \left(\frac{(x - x_j)^3}{x_{j+1} - x_j} - (x_{j+1} - x_j)(x - x_j) \right) \times \frac{f''(x_{j+1})}{6}, \end{aligned} \tag{1.14}$$

where $x_1, \ldots, x_k$ are the knots for which the cubic spline function $f(x)$ is defined. The cyclic cubic regression spline has an identical definition, except for the additional constraints that $f(x_1) = f(x_k)$ and $f''(x_1) = f''(x_k)$ (Wood 2006).

Fourier basis The Fourier basis is suitable for periodic data. Periodic data repeat the same pattern after a certain length of time, called the "period". An example of a periodic function is the sine curve, which has a period of 2π. The fitted smooth function is defined as:

$$\hat{f}(t) = c_1 + c_2 \sin(wt) + c_3 \cos(wt) + c_4 \sin(2wt) + c_5 \cos(2wt) + \ldots,$$

where the basis is defined as $\phi_1(t) = 1$, $\phi_{2r}(t) = \sin(rwt)$ and $\phi_{2r+1}(t) = \cos(rwt)$ (Ramsay & Silverman 2006) ($r = 1, \ldots, (m-1)/2$, where m is the basis dimension, i.e. the number of basis functions to use) and $c_1, \ldots, c_m$ are the basis coefficients. The parameter w is determined by the period $p = 2\pi/w$ (Ramsay & Silverman 2006), so that if the period is, for example, $p = 365$, then $w = 2\pi/365$.

If the values t_j are spaced equally over an interval τ, with the period p equal to the length of τ, then the basis is orthogonal, since the cross-product matrix of basis functions $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ is diagonal (Ramsay & Silverman 2006). Therefore, the Fourier basis system allows computationally efficient calculations. In any case, data over a long time period can be approximated using a Fourier basis with relatively few basis functions, so that calculations

should always be relatively efficient in comparison to those for a basis system requiring a higher basis dimension.

A Fourier basis is defined completely by its basis dimension (i.e. the number of basis functions m to use) and by its period p, after which it repeats (Ramsay et al. 2009).

Fitting in the frequentist framework

In the frequentist framework, the basis coefficients can simply be estimated using ordinary least squares. The least squares criterion:

$$(\mathbf{y} - \mathbf{\Phi c})^{\mathrm{T}}(\mathbf{y} - \mathbf{\Phi c}) \tag{1.15}$$

is minimised, by taking the derivative with respect to $\mathbf{c}$ and setting equal to zero, giving:

$$2\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{c} - 2\mathbf{\Phi}^{\mathrm{T}}\mathbf{y} = 0. \tag{1.16}$$

Solving for $\mathbf{c}$ gives:

$$\hat{\mathbf{c}} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}, \tag{1.17}$$

so that the fitted values are:

$$\hat{\mathbf{y}} = \mathbf{\Phi}\hat{\mathbf{c}} = \mathbf{\Phi}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y} = \mathbf{P}\mathbf{y}, \tag{1.18}$$

where $\mathbf{P}$ is the projection, or "hat", matrix (Ramsay & Silverman 2006).

This method makes the assumption that the "standard model for error" applies, i.e. that the residuals ε_j $(j = 1, \ldots, n)$ are independently and identically distributed, with mean zero and constant variance (Ramsay & Silverman 2006). However, this may be inappropriate if the errors are non-stationary or autocorrelated, in which case weighted least squares should be used instead. This involves minimising the weighted least squares criterion:

$$(\mathbf{y} - \mathbf{\Phi c})^{\mathrm{T}}\mathbf{W}(\mathbf{y} - \mathbf{\Phi c}), \tag{1.19}$$

where $\mathbf{W}$ is a symmetric positive definite matrix allowing for unequal weighting of squares and products of residuals (Ramsay & Silverman 2006).

Using the standard model for error, the variance-covariance matrix for the estimated basis coefficients is:

$$\text{var}(\mathbf{c}) = \sigma_\varepsilon^2(\mathbf{\Phi}^\text{T}\mathbf{\Phi})^{-1}, \tag{1.20}$$

so that:

$$\text{var}(\hat{\mathbf{y}}) = \sigma_\varepsilon^2\mathbf{\Phi}(\mathbf{\Phi}^\text{T}\mathbf{\Phi})^{-1}\mathbf{\Phi}^\text{T} = \sigma_\varepsilon^2\mathbf{P}. \tag{1.21}$$

The 95% confidence intervals for $\hat{\mathbf{y}}$ are:

$$\hat{\mathbf{y}} \pm 1.96 \times \sqrt{\text{SE}(\hat{\mathbf{y}})}, \tag{1.22}$$

where $\text{SE}(\hat{\mathbf{y}})$ are the standard errors of $\hat{\mathbf{y}}$, which equal the diagonal of $\sigma_\varepsilon^2\mathbf{P}$.

Fitting in the Bayesian framework

In the Bayesian framework, the basis coefficients are estimated through a hierarchical model, with prior distributions placed on the basis coefficient vector $\mathbf{c}$ and on the error variance parameter σ_ε^2. The model is:

$$\mathbf{y}|\mathbf{c}, \sigma_\varepsilon^2 \sim \text{N}_n(\mathbf{\Phi c}, \sigma_\varepsilon^2\mathbf{I}_n), \tag{1.23}$$

where $\text{N}_n(\mathbf{\Phi c}, \sigma_\varepsilon^2\mathbf{I}_n)$ is the multivariate Normal distribution with n-length mean vector $\mathbf{\Phi c}$ and $n \times n$ covariance matrix $\sigma_\varepsilon^2\mathbf{I}_n$, $\mathbf{y}$ is the n-length vector of data, $\mathbf{\Phi}$ is the $(n \times m)$ matrix of m basis functions, evaluated at the n times of data collection, $\mathbf{c}$ is the m-length vector of basis coefficient values, σ_ε^2 is the variance associated with estimating the curve, and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix (Abraham & Khadraoui 2015, Gelman et al. 2014).

Prior distributions for the parameters $(\sigma_\varepsilon^2)^{-1}$ and $\mathbf{c}$ are:

$$\begin{aligned}(\sigma_\varepsilon^2)^{-1} &\sim \text{Ga}(a, b) \text{ and} \\ \mathbf{c} &\sim \text{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\end{aligned} \tag{1.24}$$

where $\text{Ga}(a, b)$ is the Gamma distribution with shape parameter a and rate parameter b. A multivariate Normal prior distribution has been chosen for the basis coefficients, as also used in the literature by authors including Abraham & Khadraoui (2015), Denison et al. (2002) and Gelman et al. (2014). The choice of a, b, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be discussed in subsection 5.2.2.

This model can be fitted through Gibbs sampling, since the posterior distributions are easy to obtain. The disadvantage to the Bayesian approach, compared to the frequentist framework, is that the computation time and memory requirements are much greater. However, one important benefit to fitting the model in the Bayesian framework is that the prior distribution assigned to the basis coefficients ensures that there is no problem caused by gaps in the data, or by unequal start and endpoints in the data for different locations, since the prior distribution provides information that otherwise would be missing. The frequentist model does not have this additional information and so can suffer from near-singularity in the matrix $\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}$, so that model fitting is not possible for higher basis dimensions with few available data. Another benefit to fitting the model in the Bayesian framework is that the statistical downscaling models developed in forthcoming chapters are hierarchical Bayesian models, providing an opportunity for combining statistical downscaling and functional data analysis in the same framework.

Choice of basis type and dimension

The first *a priori* choice that must be made, before fitting a nonparametric smoothing model, is to select the type of basis function to use in the model. For additive modelling, thin plate regression splines are preferred for smooth surfaces fitted to multiple variables at once, while cyclic cubic regression

splines are fitted to cyclical data that repeat after one cycle (Wood 2006). For functional data analysis, the Fourier basis is preferred for periodic data, while the B-spline basis is an efficient basis for non-periodic data (Ramsay & Silverman 2006).

For additive modelling, the model-fitting method includes a term that penalises excess curvature, so that the basis dimension does not have a strong effect on any results. However, in the case where the choice of the basis dimension must be made *a priori*, this choice is of great importance, since the fitted smooth function depends on having a large enough basis dimension in order to adequately reflect patterns in the data. The dimension choice should also take into account the sparseness of the available data and the number of data points available, so that it is not too large. There are, however, ways to quantify the model fit over a changing basis dimension, in order to obtain an optimal basis dimension. These include generalised cross validation, Akaike's information criterion, the deviance information criterion and leave-one-out cross-validation.

Generalised cross validation Generalised cross validation (GCV) is a method developed by Craven & Wahba (1979) to identify the optimal value of the smoothing parameter in smoothing splines. It has been used by Pya & Wood (2016) and Ruppert (2002) to select the optimal basis dimension in a splines context. The general form of the GCV equation is:

$$\text{GCV} = \frac{n \ \text{RSS}}{\big(n - \text{trace}(\mathbf{P})\big)^2},$$

where n is the number of data, RSS is the residual sum of squares and $\mathbf{P}$ is the projection matrix, such that $\hat{\boldsymbol{y}} = \mathbf{P}\mathbf{y}$. This can be calculated easily, where the goal is to fit a smooth function to some data, for a single location

(see equation 1.10 on page 27):

$$\text{GCV} = \frac{n\ \left(\mathbf{y}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P})^{\mathrm{T}}(\mathbf{I}_n - \mathbf{P})\mathbf{y}\right)}{\left(n - \text{trace}(\mathbf{P})\right)^2},$$

where $\mathbf{P} = \mathbf{\Phi}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}$, with $\mathbf{\Phi}$ being the matrix of basis functions evaluated at sampling times of the data $\mathbf{y}$.

Akaike's information criterion and the deviance information criterion Akaike's information criterion and the deviance information criterion measure the predictive accuracy of the model, corrected for the fact that the model is fitted to the observed data. These criteria mostly make use of the deviance, the log predictive density of the data given a point estimate from the model, i.e. $\log\big(f(\mathbf{y}|\hat{\boldsymbol{\theta}})\big)$ (Gelman et al. 2014).

Akaike's information criterion (AIC) is defined as:

$$\text{AIC} = -2\ \log\big(p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{MLE}})\big) + 2k,$$

where k is the number of parameters in the model and $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is the vector of maximum likelihood estimates for the parameters (approximated by the posterior mean or median for a Bayesian model). k may be difficult to define in a Bayesian hierarchical model, with some at least partially informative prior information, since the effective number of parameters may in fact be less than the total number of parameters (Gelman et al. 2014).

An alternative to AIC is the deviance information criterion (DIC), developed for the Bayesian framework. The DIC is defined as:

$$\text{DIC} = -2\ \log\ p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) + 2\ p_{\text{DIC}},$$

where p_{DIC} is a measure of the effective number of parameters and $\hat{\boldsymbol{\theta}}_{\text{Bayes}}$ is a vector of the posterior means or medians of the parameters (Gelman et al.

2014). p_{DIC} can be calculated as either:

$$p_{\text{DIC}_1} = 2\Big(\log\, p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) - \text{E}_{\text{post}}\big(\log\, p(\boldsymbol{y}|\boldsymbol{\theta})\big)\Big),$$

where $\text{E}_{\text{post}}\big(\log\, p(\boldsymbol{y}|\boldsymbol{\theta})\big)$ is the expectation of $\log\, p(\boldsymbol{y}|\boldsymbol{\theta})$ over its posterior distribution, or:

$$p_{\text{DIC}_2} = 2\ \text{var}_{\text{post}}\big(\log\, p(\boldsymbol{y}|\boldsymbol{\theta})\big),$$

which makes use of the variance of $\log\, p(\boldsymbol{y}|\boldsymbol{\theta})$ over the posterior distribution (Gelman et al. 2014).

Leave-one-out cross-validation An alternative method to assess model performance is leave-one-out cross-validation, which is the method of removing a single data point in turn and predicting its value using the model of interest fitted to the remaining data. After predictions have been made for all data points, the accuracy and precision of these predictions is assessed and compared to the observed data, in order to compare the performances of different models. An absolute, rather than comparative, measure of model performance is the empirical interval coverage probability, i.e. the proportion of intervals of a certain nominal coverage (e.g. 95%) that contain the true data value, which should be close to the nominal value, for a model that is appropriate for the data.

Alternative methods An alternative information criterion for choosing the basis dimension is the WAIC, i.e. the Watanabe-Akaike information criterion, which is a fully Bayesian approach for estimating the out of sample expectation. This approach takes the computed log pointwise posterior predictive density and adjusts for overfitting by correcting for the effective number of parameters through a measure derived from simulations (Gelman et al. 2014). Like the AIC and DIC, the basis dimension that results in the minimum WAIC value is preferred.

A further alternative is the continuous ranked probability score (CRPS), which was developed in the weather forecasting literature. It is a generalisation of the mean absolute error and is a scoring rule that provides a measure for the evaluation of forecasts, allowing the ranking of competing forecast procedures. The CRPS evaluates probability forecasts in the form of cumulative distribution functions and assigns a numerical score, based upon the predictive distribution and on the event or value that is observed (Gneiting & Raftery 2007, Bröcker 2012).

1.7 Bayesian modelling

This section briefly presents the main features of Bayesian modelling that are relevant to the work in this thesis. Although earlier work in this thesis is presented in the frequentist context, later models are fitted that require Bayesian methodology. For these models, there is not enough information from the data to estimate all parameters in the frequentist framework, but the provision of prior distributions in the Bayesian framework allows these models to be fitted.

Given a vector of parameters $\boldsymbol{\theta}$ and data $\mathbf{y}$, it is of interest to make inference on the values of parameters $\boldsymbol{\theta}$. Since it is of interest to make probabilistic statements about $\boldsymbol{\theta}|\mathbf{y}$, a model is needed for the joint probability distribution of $\boldsymbol{\theta}$ and $\mathbf{y}$ (Gelman et al. 2014). The joint probability mass function or probability density function is $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$, where $p(\boldsymbol{\theta})$ is the prior distribution and $p(\mathbf{y}|\boldsymbol{\theta})$ is the sampling distribution, or data distribution. Using Bayes' theorem, the posterior distribution of $\boldsymbol{\theta}$, given the data $\mathbf{y}$, is:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{1.25}$$

where $p(\mathbf{y}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$ for discrete $\boldsymbol{\theta}$ and $p(\mathbf{y}) = \int p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}$ for

continuous $\boldsymbol{\theta}$. The unnormalised version of this equation is:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}), \tag{1.26}$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood, which is treated as a function of $\boldsymbol{\theta}$, for fixed $\mathbf{y}$ (Gelman et al. 2014).

The remainder of this section presents methods for obtaining samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

1.7.1 Gibbs sampling

Gibbs sampling, developed by Geman & Geman (1984), is a method of sampling from the posterior distributions of parameters in a Bayesian model. This takes the form of a Markov chain, a random walk through the parameter space, starting from an arbitrary starting point, with the next step in the chain depending only on the current position, not on previous positions. Assuming there are k unknown parameters in the model, the method cycles through drawing from the full conditional distributions for each parameter in turn, conditional on the previous values of draws for these parameters (Kruschke 2014):

1. Draw a value from the full conditional distribution of $\theta_1^{(i)}|\theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)}$.
2. Draw a value from the full conditional distribution of $\theta_2^{(i)}|\theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)}$.

   ...

k. Draw a value from the full conditional distribution of $\theta_k^{(i)}|\theta_1^{(i)}, \ldots, \theta_{k-1}^{(i)}$.

In each case above, superscripts (i) and $(i-1)$ represent the ith and $(i-1)$th draws from the full conditional distributions, respectively. After step k, the method returns to step 1, with the value of i now increased by one. This continues until i has reached a large number, say 10,000. The recorded draws for each parameter eventually come from the posterior distribution for

that parameter, but the first few draws should be discarded (Kruschke 2014, Gelman et al. 2014).

1.7.2 Metropolis algorithm

If the full conditional distributions of a parameter cannot be calculated, for example if the normalising constant cannot be evaluated, then an alternative is to use the Metropolis algorithm, which was originally developed by Metropolis et al. (1953) and was generalised by Hastings (1970). As with the Gibbs sampler, this is a random walk, with an acceptance rule to allow convergence to the target distribution (Gelman et al. 2014). This proceeds as follows (Gelman et al. 2014):

1. Draw an arbitrary value $\theta^{(0)}$ from a starting distribution $p_0(\theta)$ (Gelman et al. 2014). This starting distribution may be the unnormalised posterior distribution of θ, i.e. the product of the prior and likelihood without the normalisation constant (Kruschke 2014).

2. Sample proposal θ^* from a proposal distribution $J_t(\theta^*|\theta^{(t-1)})$, where t indexes the iteration.

3. To decide whether to accept this proposal, calculate the ratio of densities $r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$.

4. Set the current value equal to the proposal $\theta^{(t)} = \theta^*$ with probability $\min(r, 1)$, keeping the current value equal to the previous value $\theta^{(t)} = \theta^{(t-1)}$ otherwise.

5. Repeat steps 2 to 4 many times, until the values of $\theta^{(t)}$ come from a stationary distribution.

The Metropolis algorithm must be tuned, to ensure that the acceptance rate is not too high or too low. Tuning is the process of changing the variance parameter of the proposal distribution, in order to obtain an acceptance rate

close to an ideal value. A very high rate means that autocorrelation in the chain is very high, so that the chain is slow to explore the parameter space. On the other hand, a very low acceptance rate means that the chain stays in one place for long periods of time, so that a very long run is needed. The tuning procedure can be based upon an initial number of iterations (Gelman et al. 2014).

1.7.3 Alternative methods

It is possible to carry out Metropolis algorithm calculations within a Gibbs sampler, for example when some conditional posterior distributions can be sampled from directly and some cannot. This can be done with Gibbs steps for all possible parameters and one dimensional Metropolis for all other parameters. Alternatively, parameters can be updated in blocks, with either a Gibbs or a Metropolis step for each block. Gibbs sampling is generally preferred, where possible, since it is the simplest MCMC algorithm and the fact that the Metropolis algorithm can reject moves means that it can be slower to explore the parameter space (Gelman et al. 2014).

Both Gibbs and Metropolis are types of the more general Metropolis-Hastings algorithm, which allows for non-symmetrical proposal distributions and so has a different form of the acceptance ratio (Gelman et al. 2014).

Another type of MCMC algorithm is Hamiltonian Monte-Carlo, which is recommended by Gelman et al. (2014) and is used by the authors' program STAN.

A recently developed alternative to MCMC is integrated nested Laplace approximation (INLA), which is based upon a deterministic algorithm that directly approximates the posterior marginal densities (Rue et al. 2009, Blangiardo & Cameletti 2015). It is specifically designed for latent Gaussian models and is stated to be a faster, yet still valid, alternative to MCMC (Rue et al. 2009, Blangiardo & Cameletti 2015).

1.7.4 Convergence diagnostics

A practical issue for Bayesian modelling, that must be accounted for, is that of convergence diagnostics. The Markov chains must be run for many iterations, until eventually the estimated values for each parameter are drawn from a stationary posterior distribution, which closely approximates the true distribution of the parameter (Gelman et al. 2014). Convergence diagnostics are used to ensure that the stationary distribution has been reached. Two graphical methods for assessment of convergence are trace and density plots. The trace plot is simply a plot of estimated parameter values by iteration number. If convergence has been reached, this plot shows values centred around a stationary median, so that the plot looks like a "hairy caterpillar" (Gelman et al. 2014). If multiple chains are run, the distributions of these values for each chain will be centred on the same median value, with the same variance, if convergence has occurred. Plotting the density of the values of converged chains produces a shape that is similar to that expected for the posterior distribution of the parameter of interest (Gelman et al. 2014), e.g. a bell-shaped curve should be produced, for a Normal posterior distribution.

1.8 Conclusions

This chapter presented the motivation for the research in this thesis, namely the importance of developing a global understanding of lake health through the use of a database of observed ecological parameters.

The main research question to be answered is whether a limited quantity of accurate *in situ* lake data can be fused with remotely-sensed data, taking accuracy from the *in situ* data and spatial and temporal information from the remotely-sensed data.

One way to accomplish this fusion, taking into account the different spatial and temporal supports of the two datasets, is through statistical downscaling, a method that is focussed on in this thesis.

This chapter also presented methodology for the spatial and temporal modelling of the data, through geostatistics and nonparametric smoothing. These methods are applied in an initial analysis of the data in the following chapter.

Chapter 2

Initial spatial and temporal analysis of data

In this chapter, the patterns over space and through time in the *in situ* and remotely-sensed data are investigated, through the use of standard statistical approaches that are widely used for the analysis of data. This leads to an understanding of how the data relate to each other and indicates which techniques for data fusion are suitable for this application. The *in situ* and remotely-sensed data are explored separately, to investigate patterns over space and time. The chapter focusses on the data available for Lake Balaton, Hungary, discussed in the previous chapter, since Lake Balaton is of interest due to its historically poor water quality and also since the *in situ* and remotely-sensed data are available for several variables over a long period of time. The chapters following this one build upon this work, going beyond the standard approaches, through the development and application of more complex statistical models that allow the fusion of information from the two types of data investigated, rather than investigating them separately.

Firstly, an exploratory analysis is carried out on the *in situ* data for temperature, chlorophyll$_a$ and total suspended matter. Patterns over space and time are explored through the application of mixed-effects models, in order to understand how the water quality variables relate to one another,

while taking into account the effect of location.

Spatial patterns evident in the remotely-sensed temperature data are explored through the application of kriging. Temperature data are investigated in this initial analysis, since there are fewer grid cells of temperature data covering the lake, in comparison to the number of cells of log(chlorophyll$_a$) data or of log(total suspended matter) data, so that the analysis is less computationally intensive. Common spatial and temporal patterns in the remotely-sensed data are also investigated, through S- and T-mode principal component analysis (PCA), where PCA is carried out on a matrix of time versus locations and on a matrix of locations versus time, respectively.

Finally, additive models are used to investigate how well the remotely-sensed temperature data relate to the *in situ* data.

These analyses provide an understanding of patterns in the data, which are taken account of in subsequent analyses.

2.1 Exploring the *in situ* data

In situ data are available from two sources for Lake Balaton, BLI and KDKVI, for the variables log(chlorophyll$_a$ concentration (mg/m^3)), log(total suspended matter (g/m^3)) and temperature (°C) (see Table 2.1). Data from

Source	Var.	Location					Dates
		Keszthely	Szigliget	Szemes	Tihany	Siófok	
BLI	All	90	27	27	92	24	Feb 2006 – Dec 2011
KDKVI	Chl	204	203	204		205	Apr 2002 – Mar 2012
KDKVI	Temp	82	81	82		82	Apr 2002 – Dec 2006
KDKVI	TSM	45	44	45		45	Apr 2002 – Dec 2006

Table 2.1: Numbers of *in situ* data available for Lake Balaton, for the variables chlorophyll$_a$ (Chl), lake surface water temperature (Temp) and total suspended matter (TSM).

the BLI are available from the start of 2006 until the end of 2011, while data from KDKVI are available from the start of 2002 until the start of 2012. Before 2007, the KDKVI data are available approximately fortnightly for the three variables, with log(total suspended matter) only sampled every

second sampling trip. From 2007 onwards, only log(chlorophyll$_a$) is available from KDKVI, sampled weekly. Since data are sampled irregularly, there are only 17 months for which data are available for all nine *in situ* locations. All calculations and plots are produced using `R` (R Core Team 2017).

2.1.1 Exploratory plots

Data for each of the three variables of interest are plotted over time, with separate lines for each location (see Figure 2.1). This plot shows that

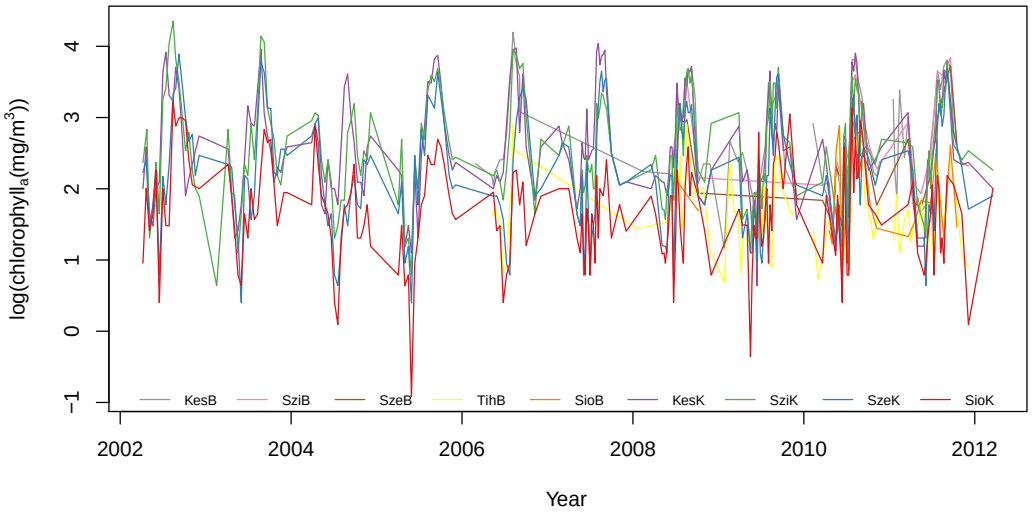

Figure 2.1: Plot of log(chlorophyll$_a$ concentration) over time, with separate lines for each location.

log(chlorophyll$_a$) follows similar patterns over time for each location, with one main peak and one smaller peak per year. It can also be seen that chlorophyll levels vary spatially along the lake. For example, levels at Keszthely in the western basin of the lake (KesB and KesK in Figure 2.1) are generally higher than levels at Siófok, in the eastern basin of the lake (SioB and SioK in Figure 2.1), which is explained by the presence of the main inflow, the river Zala, at the western end of the lake, bringing with it increased nutrients. A similar plot is produced for log(total suspended matter) (see Figure 2.2). As can be seen from this plot, the pattern in total suspended matter concentration is much more variable than that for log(chlorophyll$_a$). Unlike chlorophyll,

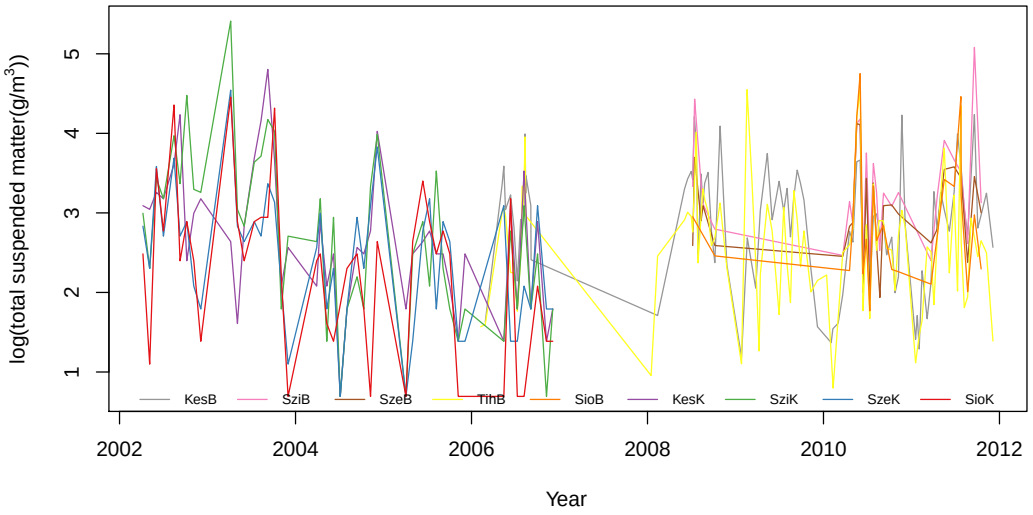

Figure 2.2: Plot of log(total suspended matter concentration) over time, with separate lines for each location.

there appears to be a change in the mean of total suspended matter over the longer term, with decreasing levels between 2002 and 2004 and then slightly increasing levels after 2008. For the data in later years, there appears to be a single peak in total suspended matter per year. Finally, a plot of temperature against time, with separate lines for each location, is produced (see Figure 2.3). This plot shows that temperature does not vary much over the lake,

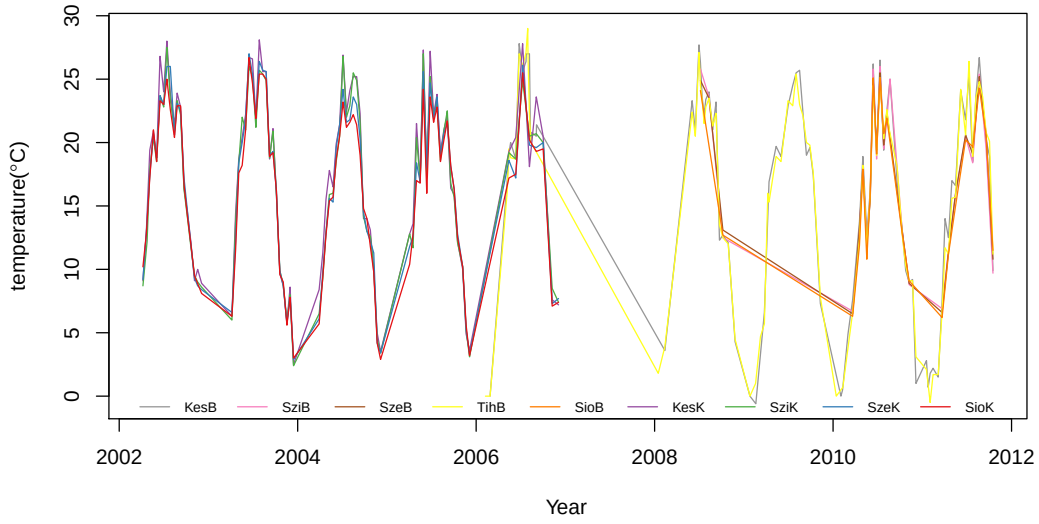

Figure 2.3: Plot of temperature over time, with separate lines for each location.

but has a strong seasonal pattern, with high values during summer and low

values during winter. To understand how the three variables relate to each other, scatterplots are produced (see Figure 2.4). These plots show that there

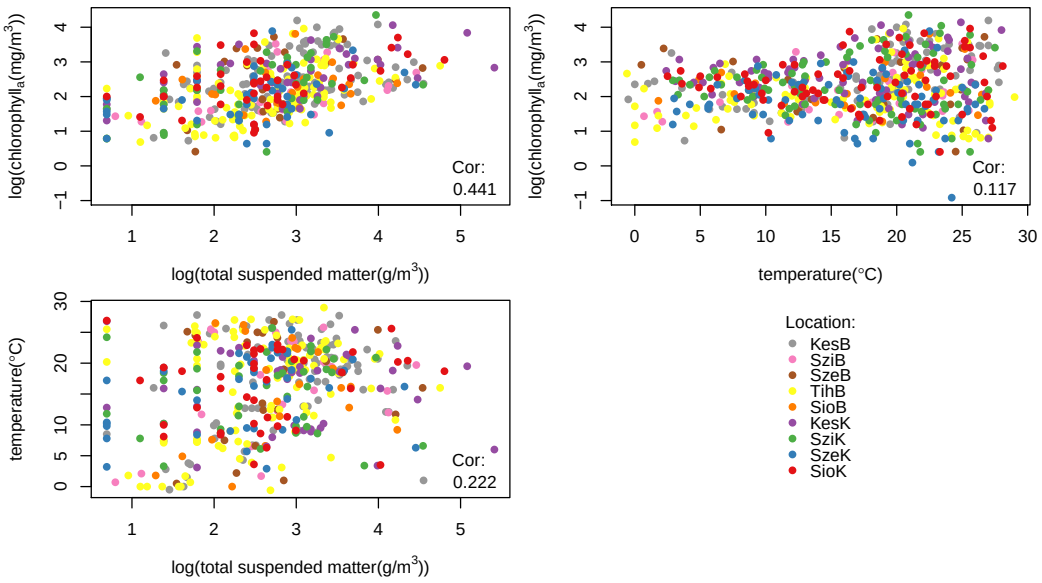

Figure 2.4: Plots of log(chlorophyll), log(total suspended matter) and temperature, coloured by location. Pearson's correlation coefficients are also shown.

is a moderately strong, positive, linear relationship between log(chlorophyll$_a$) and log(total suspended matter), with a correlation of 0.441. This positive relationship makes sense, since the same processes leading to increases in suspended matter in the lake may also lead to increases in chlorophyll, e.g. increased soil runoff from nearby farms, or additional floating vegetation entering the lake. Increases in certain bacteria and algae may directly increase both measures. There also appears to be a weak positive linear relationship between log(total suspended matter) and temperature, with a correlation of 0.222. The relationship between log(chlorophyll$_a$) and temperature is more complex, with little, if any, positive relationship. It appears that variability in log(chlorophyll$_a$) increases as temperature increases above around 15°C. This more complex relationship may be explained by the life cycle of cyanobacteria blooms, with multiple blooms per year, with timings influenced by water temperature amongst other environmental factors (Teta et al. 2017).

2.1.2 Mixed-effects models for the *in situ* data

In order to understand how the *in situ* data vary over space and time, models are fitted in a mixed-effects framework. Firstly, a model is fitted, predicting $\log(\text{chlorophyll}_a)$ from log(total suspended matter), temperature, year (to account for long-term trend), longitude, latitude and harmonic terms accounting for seasonal patterns of two peaks per year, with location fitted as a random effect. It is found that longitude does not have a significant effect on $\log(\text{chlorophyll}_a)$, so it is removed from the model, followed by year, giving:

$$\begin{aligned}\log(\text{chlorophyll}_a)_i &= \beta_0 + \beta_1 \log(\text{total suspended matter})_i + \beta_2\ (\text{temperature})_i \\ &+ \beta_3\ (\text{latitude})_i + \beta_4 \cos\left(\frac{2\pi(\text{day of year})_i}{365}\right) + \beta_5 \sin\left(\frac{2\pi(\text{day of year})_i}{365}\right) \\ &+ \beta_6 \cos\left(\frac{4\pi(\text{day of year})_i}{365}\right) + \beta_7 \sin\left(\frac{4\pi(\text{day of year})_i}{365}\right) + b\ (\text{location})_i + \varepsilon_i.\end{aligned} \tag{2.1}$$

The model showing the estimated values of the coefficients is:

$$\begin{aligned}\mathrm{E}\big(\log(\text{chlorophyll}_a)\big)_i &= 169.7 + 0.302 \log(\text{total suspended matter})_i - 0.030\ (\text{temperature})_i \\ &- 3.585\ (\text{latitude})_i - 0.289 \cos\left(\frac{2\pi(\text{day of year})_i}{365}\right) - 0.389 \sin\left(\frac{2\pi(\text{day of year})_i}{365}\right) \\ &- 0.075 \cos\left(\frac{4\pi(\text{day of year})_i}{365}\right) + 0.399 \sin\left(\frac{4\pi(\text{day of year})_i}{365}\right).\end{aligned} \tag{2.2}$$

The random errors are $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ and location is fitted as a random effect, with $b \sim \mathrm{N}(0, \sigma_b^2)$. Estimated values of the variance parameters are 0.213 and 0.018, for σ_ε^2 and σ_b^2, respectively, giving an estimate of intraclass correlation (i.e. of $\sigma_b^2/(\sigma_b^2 + \sigma_\varepsilon^2)$) of 0.078, meaning that only a small proportion of the total variability in the data is estimated to be due to the random effect. This small value is possibly due to latitude being included as a term in the model, causing a lack of identifiability between latitude and the random effect of

location. Even with this possible identifiability issue, the model is still useful for understanding the main patterns in the data. Four harmonic terms are included, the first two to fit a sine curve with a period of 365 days and the second two to fit a sine curve with a period of (365/2) days, allowing both the large and small peaks to be accounted for each year. All terms have p-values of less than 0.05. A plot of the data and predicted values is produced (see Figure 2.5). These predicted values are only calculated when data for

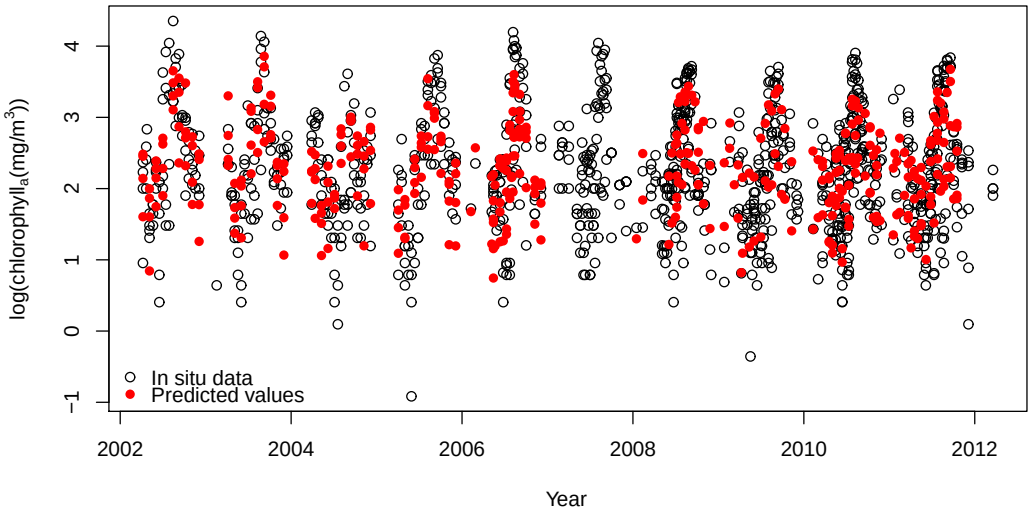

Figure 2.5: Plot of *in situ* log(chlorophyll$_a$) by year, showing predictions from model 2.1.

all predictor variables are available, so some small gaps appear on the plot. The plot suggests that the predicted values from the model follow the *in situ* data fairly closely, so the model appears to be a good fit to the data. This model estimates that log(total suspended matter) has a small positive effect on the value of log(chlorophyll$_a$), while temperature has a slightly negative effect. The model also estimates that log(chlorophyll$_a$) levels decrease by 3.6 units for every increase in latitude of one degree North. It makes sense that only one of longitude and latitude is included in the model, as the data locations are located fairly close to a straight line. The fact that all sine and cosine terms are statistically significant indicates that the model does fit two separate peaks per year, agreeing with the exploratory analysis.

Similarly, a model is fitted for log(total suspended matter). After removing both longitude and latitude, this model is:

$$\begin{aligned}\log(\text{total suspended matter})_i &= \alpha + \beta_1 \log(\text{chlorophyll}_a)_i + \beta_2 \ (\text{temperature})_i \\ &+ \beta_3 \ (\text{year})_i + \beta_4 \cos\left(\frac{2\pi(\text{day of year})_i}{365}\right) + \beta_5 \sin\left(\frac{2\pi(\text{day of year})_i}{365}\right) \\ &+ \beta_6 \cos\left(\frac{4\pi(\text{day of year})_i}{365}\right) + \beta_7 \sin\left(\frac{4\pi(\text{day of year})_i}{365}\right) + b \ (\text{location})_i + \varepsilon_i.\end{aligned} \tag{2.3}$$

The model with the estimated coefficients shown is as follows:

$$\begin{aligned}\mathrm{E}\big(\log(\text{total suspended matter})\big)_i &= 173.7 + 0.610 \log(\text{chlorophyll}_a)_i - 0.059 \ (\text{temperature})_i \\ &- 0.086 \ (\text{year})_i - 1.075 \cos\left(\frac{2\pi(\text{day of year})_i}{365}\right) - 0.396 \sin\left(\frac{2\pi(\text{day of year})_i}{365}\right) \\ &- 0.149 \cos\left(\frac{4\pi(\text{day of year})_i}{365}\right) - 0.395 \sin\left(\frac{4\pi(\text{day of year})_i}{365}\right).\end{aligned} \tag{2.4}$$

Location is again included as a random effect, with $b \sim \mathrm{N}(0, \sigma_b^2)$, and the errors are $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. The estimates of the variance parameters are 0.258 and 0.424, for σ_b^2 and σ_ε^2 respectively, giving an estimate of the intraclass correlation coefficient of 0.378. This means that a substantial proportion of variability in the *in situ* data is estimated as being due to the random effect of location. The model estimates a positive effect of $\log(\text{chlorophyll}_a)$ on log(total suspended matter), with a slightly negative effect of temperature. A slightly negative trend is estimated over time, while again two peaks are fitted per year. All p-values are statistically significant. A plot of the data and model predictions is produced (see Figure 2.6), which illustrates this fit to the data. It appears that this model does not capture the peaks in the data well. Possibly these peaks are caused by another variable that has not been measured.

In conclusion, both $\log(\text{chlorophyll}_a)$ and log(total suspended matter)

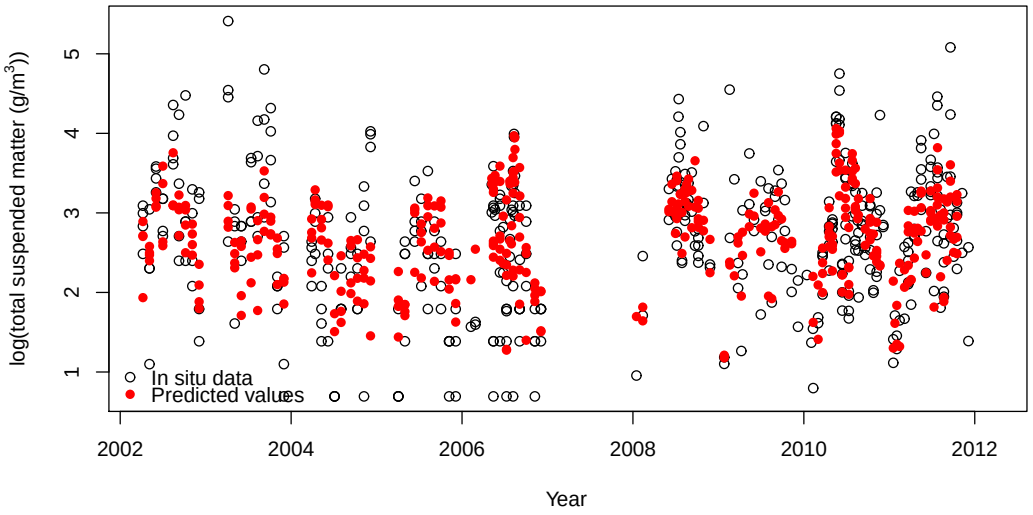

Figure 2.6: Plot of *in situ* log(total suspended matter) by year, showing predictions from model 2.3.

have strong cyclical patterns, with two peaks per year estimated for each variable. The two variables are positively related and both have a slightly negative relationship with temperature. Both variables do appear to vary across the lake. Latitude is estimated to have a significant effect on log(chlorophyll$_a$), while the random effect of location is estimated to explain a fairly large amount of variability in the data, for log(total suspended matter). There is no estimated trend over time for log(chlorophyll$_a$), while log(total suspended matter) is estimated to have a very slightly negative trend over time.

Autocorrelation in the model residuals

For each of the two fitted models, (2.1) and (2.3), plots of the autocorrelation function and partial autocorrelation function for the residuals are produced (see Figure 2.7). For both models, these plots identify positive autocorrelation at small lags, meaning that the models should take account of this autocorrelation within the fitted error structure. The aim is not to model the autocorrelation structure explicitly, but to account for it as much as possible, to ensure that the precisions of parameter estimates in the model are correct. For each model, a simple AR(1) error structure is incorporated.

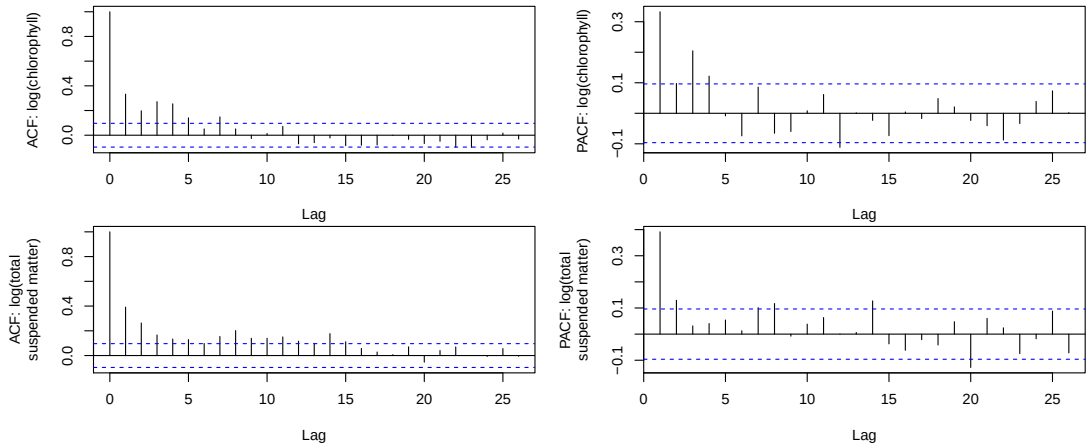

Figure 2.7: Plots of the autocorrelation functions and partial autocorrelation functions for models 2.1 and 2.3.

The models are the same as previous models, with the exception that the errors are now $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathbf{V}\sigma^2$. $\mathbf{V}$ is the correlation matrix of the form:

$$V = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \cdots \\ \phi & 1 & \phi & \cdots & \cdots \\ \phi^2 & \phi & 1 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \\ \vdots & \vdots & \vdots & & 1 \end{pmatrix}, \tag{2.5}$$

where ϕ is the AR(1) coefficient, representing how strong autocorrelation is over time. Estimated AR(1) coefficients are 0.333 and 0.391 for the models for log(chlorophyll$_a$) and log(total suspended matter), respectively. These models are named (2.1a) and (2.3a), respectively. After fitting these models, plots of the autocorrelation and partial autocorrelation functions for the residuals are produced (see Figure 2.8). These plots show that autocorrelation is very small at the first few lags, for both models. It appears that fitting an AR(1) error structure is appropriate here, as it captures much of the autocorrelation in the residuals. After fitting this error structure, there is no change in which variables are statistically significant and the estimated coefficients of variables in each model change very little.

The AR(1) error structure is further investigated in Chapter 3, where,

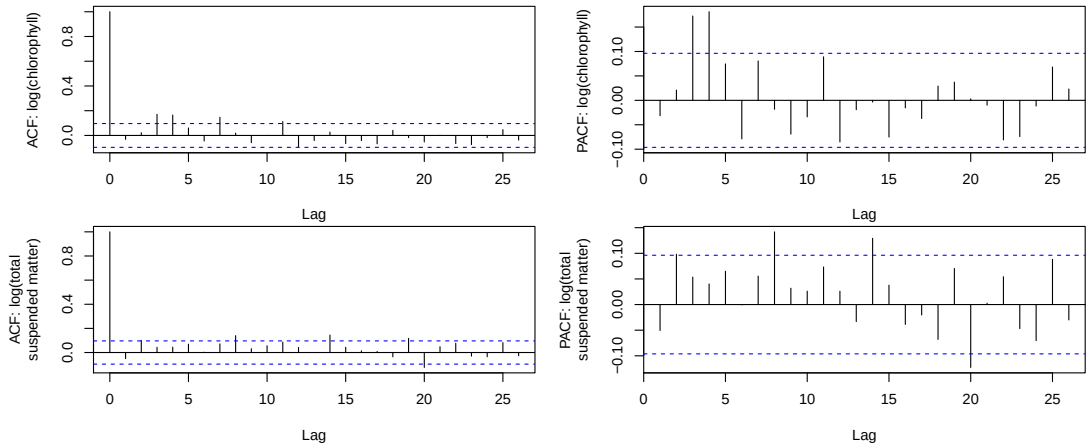

Figure 2.8: Plots of the autocorrelation functions and partial autocorrelation functions for models 2.1a and 2.3a, with AR(1) error structure.

motivated by its use here, the error structure is incorporated within a statistical downscaling model (see model 3.5 on page 105).

Conclusions

Through fitting mixed-effects models, both log(chlorophyll$_a$) and log(total suspended matter) were found to depend on temperature and each other. Both variables were estimated to have two peaks per year, in spring and summer. Exploratory analyses showed differences in log(chlorophyll$_a$) and log(total suspended matter) between locations. Therefore, location was fitted as a random effect, meaning that the effect of other variables could be investigated, while the effect of location was accounted for. Log(chlorophyll$_a$) was estimated to have a decreasing trend as latitude increased, while log(total suspended matter) was estimated to have a very slight negative trend over time. The random effect of location was estimated to explain a fairly high proportion of variability in the data, compared to the errors, for log(total suspended matter), but the effect on log(chlorophyll$_a$) was difficult to understand, due to possible lack of identifiability with the fixed effect of latitude.

2.2 Exploring the remotely-sensed data

Remotely-sensed data are available from ARC-Lake (www.laketemp.net), for the variable lake surface water temperature, and from the Diversity II inland waters project (www.diversity2.info/products/inlandwaters), for the variables log(chlorophyll$_a$) and log(total suspended matter).

2.2.1 Kriging the ARC-Lake temperature data

Data are available from the ARC-Lake project for lake surface water temperature. These data are obtained from the (Advanced) Along-Track Scanning Radiometers ((A)ATSR), instruments designed to measure sea surface temperature, on board the European Space Agency's ERS-2 and ENVISAT satellites. The data are available for 41 grid cells, covering Lake Balaton, in the form of monthly averages. They are derived from measurements of reflected and emitted radiation, taken at various wavelengths, for individual pixels, and are then averaged over space and time, to form data resolved over an approximately 0.05° grid and monthly-averaged. Each dataset is available for either day or night values. The daytime observed spatially-resolved monthly average time series is used here. These data are available between July 1995 and March 2012, covering the sampling periods of both the BLI and the KDKVI data. The locations of grid cell centres are superimposed upon a map showing the *in situ* sampling locations for Lake Balaton (see Figure 2.9). From this map, it is seen that the 41 grid cells have a reasonably good spatial coverage of the lake, in comparison to the *in situ* data. However, there are some missing cells for certain months. It is rare for data to be available for all 41 grid cells for each month, but many months do have a very high proportion of data available. Table 2.2 shows that only 10 months have data available for all 41 grid cells, but 66 months have data available for at least 31 cells and 125 months have data available for at least 21 cells.

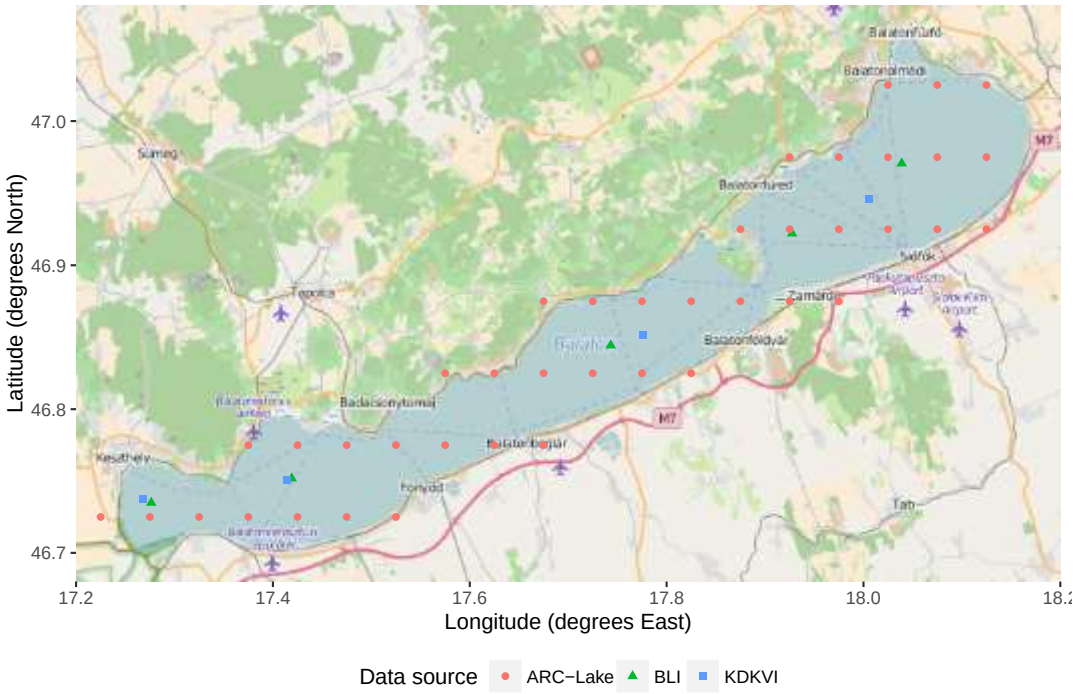

Figure 2.9: Map of Lake Balaton, showing the nine *in situ* sampling locations and the forty-one ARC-Lake remote sensing data grid cell centres. Map ©OpenStreetMap contributors (www.openstreetmap.org).

Number of grid cells with available data	Number of months
0	24
1 – 10	24
11 – 20	28
21 – 30	59
31 – 40	56
41	10

Table 2.2: Table of number of grid cells with available data by number of months.

Kriging is a method for spatial prediction, from the geostatistics literature. It enables the prediction of a variable, based upon a variogram model that models semivariance as a function of distance (Bivand et al. 2013).

Application to the remotely-sensed temperature data for 2006

An exploratory analysis is carried out for the year 2006, to both explore the spatial patterns across the lake for each month and to explore differences in those patterns over time. Plots are produced of lake surface water tem-

perature for each month (see Figure 2.10), using the `R` package `sp` (Bivand et al. 2013), with the initial data read-in aided by `ncdf4` (Pierce 2017). The

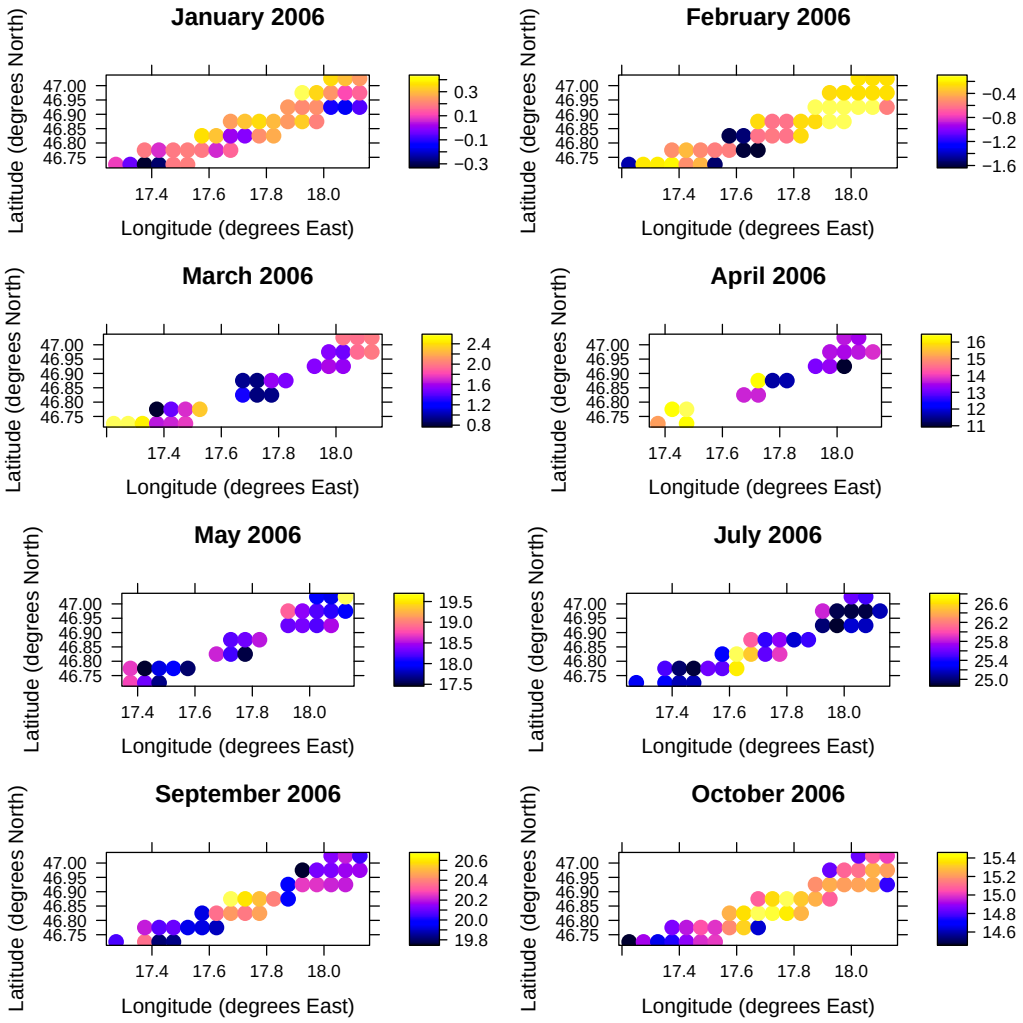

Figure 2.10: Remotely-sensed lake surface water temperature data (°C) for Lake Balaton, for a selection of months in 2006.

plots for June and August are missing, since only a few grid cells have available data for those months. Data for each month are plotted on a different scale, since otherwise spatial patterns are obscured. This is due to the much larger variation in temperature over time compared to space. There are generally smooth changes in temperature across the lake, within each month. Although 41 grid cells provide a reasonable spatial coverage of the lake, the spatial patterns can be more easily understood if interpolation is carried out,

which can be accomplished through kriging.

Before kriging is carried out, a linear trend is fitted to the data, so that this trend does not obscure the remaining spatial variation. The model fitted is $Z_i = \alpha + \beta_1(\text{longitude})_i + \beta_2(\text{latitude})_i + \varepsilon_i$, for $i = 1, \ldots, n$, where Z_i is the temperature at location i and $\varepsilon_i \sim \text{N}(0, \sigma^2_{LM})$ are the random errors. There is no temporal component to this model. Longitude has a significant effect for April, October and December, while latitude has a significant effect for January and December. In order to keep consistency between modelling for all 12 months, both longitude and latitude are left in each model.

Variograms are fitted to the residuals of these models (see Figure 2.11), using the `R` package `geoR` (Ribeiro & Diggle 2001), using the Matérn covariance model. The plots display the sample variograms and the resulting fitted variogram models, fitted by maximum likelihood (ML) and restricted maximum likelihood (REML). These models fit the sample variograms fairly well, except for April and May. This poor estimation for these months may be due to the fairly widely-spaced remotely-sensed temperature data, which are on a regular grid, meaning that there are no data located closely enough to get a good estimation of the variogram at small distances. The estimation of the nugget effect is therefore difficult. The scale of spatial variance is different for different months, with semivariance reaching around 0.03 for January, but close to 2 for April.

Universal kriging is carried out and plots are produced for the estimates and the corresponding standard errors from this method (see Figures 2.12 and 2.13). It is unclear whether there are common spatial patterns in temperature along the lake. For January to March, there is a colder part of the lake around the centre, but the pattern changes from April onwards. Figure 2.13 shows that uncertainty is lowest closest to the grid cell centres, as can be expected. April and May have the largest universal kriging standard errors of the months investigated, which may be due to the poor estimation of the variogram models for these two months, which in turn may be due

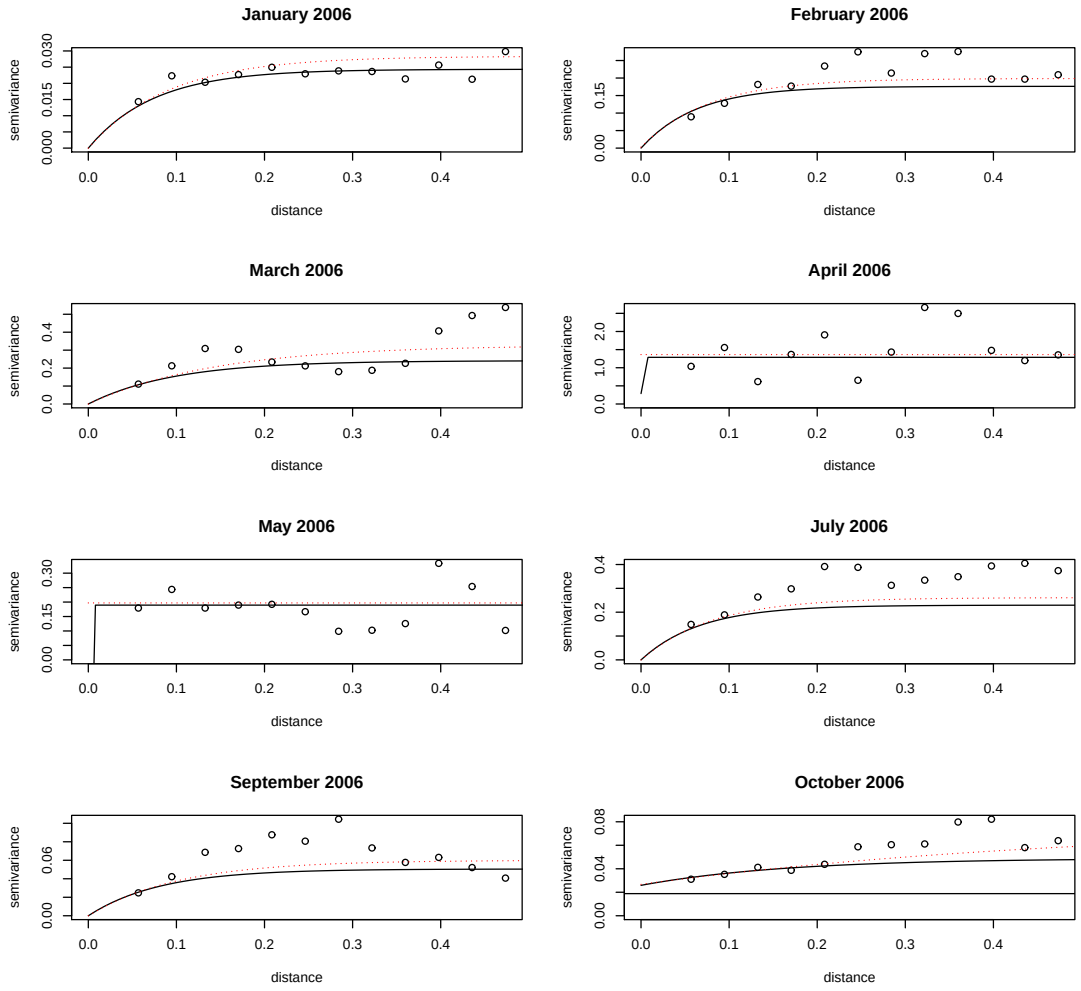

Figure 2.11: Variograms for linear model residuals for Lake Balaton 2006 temperature data. Circles are sample variograms, solid line is model fitted by ML and dotted line is model fitted by REML.

to a lack of data over space for these months. There is a sudden change between two neighbouring grid cells for the April 2006 data, meaning that the variogram has high nugget variance in comparison to those for other months, which may explain why this month has the highest standard errors of all months investigated.

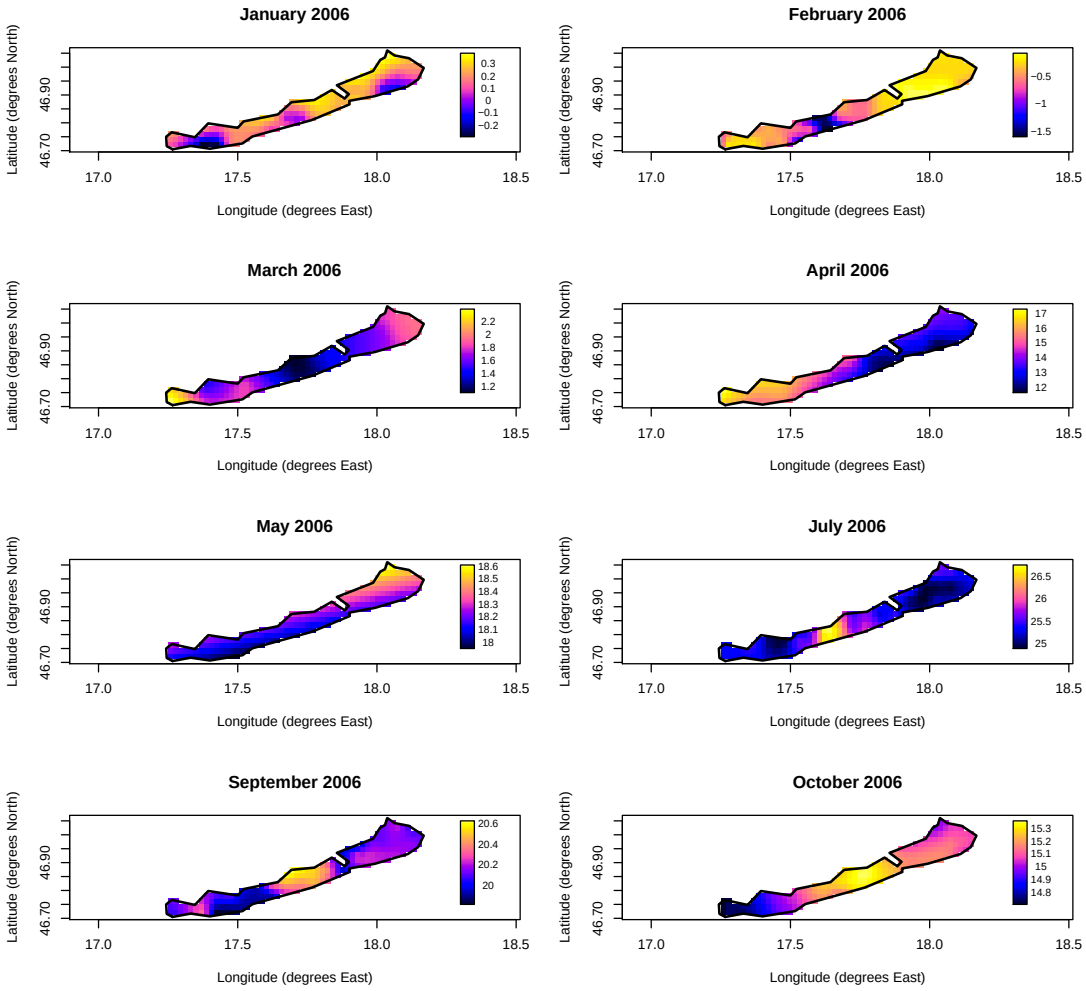

Figure 2.12: Universal kriging predictions for temperature for 2006.

2.2.2 Principal component analysis (PCA) of the remotely-sensed temperature data

Since the previous kriging analysis suggests that there are common patterns over time at each location (with higher values of temperature in summer) and potentially some common spatial patterns for different times, these patterns are more formally analysed using principal component analysis (PCA). This method reduces the dimensionality of a dataset, through transforming to a new set of uncorrelated variables, which retain as much variation as possible (Jolliffe 2002). The new set of variables is ordered from largest to smallest in terms of the amount of variation explained, so that the first few

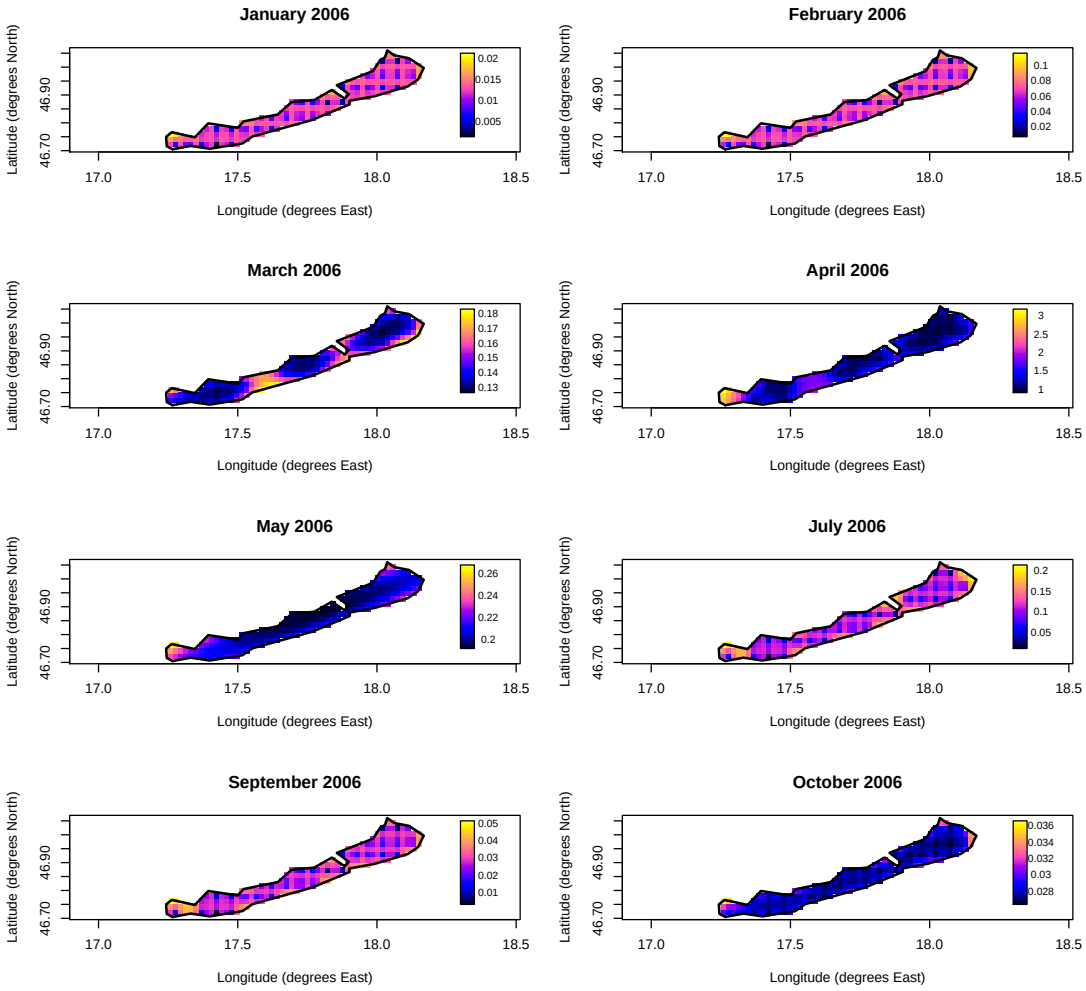

Figure 2.13: Universal kriging standard errors for temperature for 2006.

variables explain most of the variation in the original data set (Jolliffe 2002).

PCA works by carrying out a decomposition of an $m \times n$ matrix of data $\mathbf{X}$, with m measurements on each of n variables, so that $\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}}$, where $\mathbf{P}$ is an orthonormal projection matrix (i.e. $\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix) and $\mathbf{T}$ is the projection of n-dimensional $\mathbf{X}$ onto the new r-dimensional space, defined by $\mathbf{P}$ (i.e. $\mathbf{T} = \mathbf{X}\mathbf{P}$) (Demšar et al. 2013).

$\mathbf{P} \in \Re^{n \times r}$ is the loading matrix and $\mathbf{T} \in \Re^{m \times r}$ is the score matrix, where r is the number of independent columns in $\mathbf{X}$, i.e. the rank of $\mathbf{X}$, and where r is bounded by $\min(m, n)$ (Demšar et al. 2013).

The columns of $\mathbf{P}$ are the directions with the maximum variance in the

data, so the first column represents the direction with the maximum variance of all directions and is called the first principal component (PC). The second column represents the direction with the second largest amount of variance in the data and is called the second PC. Further PCs are defined similarly (Demšar et al. 2013).

Columns of $\mathbf{P}$ are eigenvectors of the covariance or correlation matrix of the data, $\mathbf{\Sigma}$, where $\mathbf{\Sigma} = \frac{\breve{\mathbf{X}}^{\mathrm{T}}\breve{\mathbf{X}}}{m-1}$, where $\breve{\mathbf{X}}$ is $\mathbf{X}$, with the mean subtracted from each column. Where the correlation matrix is used, each column is also scaled to have variance 1 (Demšar et al. 2013).

$\mathbf{\Sigma}$ is defined as being positive semidefinite, so that its eigenvalues are all ≥ 0. This means that ordering the decomposition of $\mathbf{\Sigma}$, so that the eigenvalues are in descending amplitude, gives $\mathbf{P\Lambda P}^{\mathrm{T}} = \mathbf{\Sigma}$, where $\mathbf{P}$ is the score matrix and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_r, 0, ..., 0)$ is the diagonal matrix of eigenvalues, with $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r \geq 0$ (Demšar et al. 2013).

$\mathbf{X}$ is usually approximated by a small number of PCs k, where $k << r \leq n$, which explain most of the variance in the data, i.e. $\mathbf{\Lambda}$ has a small number of large eigenvalues and many small eigenvalues.

Denoting $\mathbf{P}_k$ as the matrix of the first k columns of $\mathbf{P}$, the corresponding scores matrix is $\mathbf{T}_k = \mathbf{XP}_k$ and the total proportion of variance explained is $\mathbf{T}_k = \frac{v_k}{v_r} \times 100$, where $v_k = \sum_{i=1}^{k} \lambda_i$ and $v_r = \sum_{i=1}^{r} \lambda_i = \mathrm{trace}(\mathbf{\Lambda}) = \mathrm{trace}(\mathbf{\Sigma})$ (Demšar et al. 2013).

In practice, PCA is usually carried out by singular value decomposition (SVD). The SVD of an $n \times p$ matrix $\mathbf{X}$ is:

$$\mathbf{X} = \mathbf{ULA}^{\mathrm{T}} \tag{2.6}$$

where $\mathbf{U}$ and $\mathbf{A}$ are $n \times r$ and $p \times r$ matrices, respectively, with orthonormal columns such that $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_r$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r$ (where $\mathbf{I}_r$ is the $r \times r$ identity matrix), $\mathbf{L}$ is an $r \times r$ diagonal matrix and $r = \mathrm{rank}(\mathbf{X})$ (Jolliffe 2002).

$\mathbf{A}$ and $\mathbf{L}$ give the eigenvectors and the square roots of eigenvalues of

$\mathbf{X}^{\mathrm{T}}\mathbf{X}$, i.e. the coefficients and standard deviations of the PCs for sample covariance matrix $\mathbf{S}$. $\mathbf{U}$ contains the PC scores, scaled to have variance $\frac{1}{n-1}$. The columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$, corresponding to non-zero eigenvalues, which are of interest if the roles of observations and variables are reversed (Jolliffe 2002).

Spatiotemporal datasets, which are made up of the observed values of variables, collected at different locations and at different times, are considered to be made up of three subspaces. These are geographic space, temporal space and attribute space. Six modes of PCA can then be potentially carried out, by defining the data matrix using two of the three spaces. These are defined by Richman (1986) as:

- O-mode — attributes versus time
- P-mode — time versus attributes
- Q-mode — attributes versus locations
- R-mode — locations versus attributes
- S-mode — time versus locations
- T-mode — locations versus time

PCA is commonly applied in the atmospheric science literature, where only one variable is measured at each location, over a period of time (Demšar et al. 2013). A similar application of PCA is useful for the Lake Balaton remotely-sensed temperature data, since it reveals common patterns in the data over space and time, for the single variable. The most common mode of PCA used in the atmospheric science literature is S-mode, where the data matrix has location as the columns (variables) and time as the rows (Demšar et al. 2013), but T-mode, which uses a transposed version of the S-mode data matrix, is also used (Jolliffe 2002). The main difference between the two modes is that S-mode aims to isolate groups of stations that co-vary

similarly, useful for regionalisation or observing spatial patterns of interest, whereas T-mode aims to isolate subgroups of observations with similar spatial patterns, thereby allowing a simplification of the time series (Richman 1986).

Application to the Lake Balaton Data

Both S-mode and T-mode PCA are carried out, in order to understand the common patterns in the remotely-sensed temperature data for Lake Balaton over space and over time.

S-mode PCA S-mode PCA aims to find spatial locations that have similar temporal patterns (Richman 1986), which enables the reduction of the spatial dimensionality of the data. The Lake Balaton dataset is formed into a matrix, with the 41 locations in the lake as the columns and the 405 timepoints as the rows. The members of the matrix $x_{i,j}$ are the observations of lake surface water temperature (°C). The matrix is:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,41} \\ \vdots & & \vdots \\ x_{405,1} & \cdots & x_{405,41} \end{pmatrix}. \tag{2.7}$$

Performing an S-mode PCA on this dataset allows the investigation of the variability in the time series of temperatures across different locations in the lake, to investigate any patterns. Depending on the results, a further investigation into reducing the spatial dimensions of the data could be appropriate.

Firstly, the variances of the measurements at each of the 41 locations are investigated. These are all fairly similar, at around 80 for all locations. Due to the similarity of the variances within each location, either the correlation or the covariance matrix of the data can be used for computation of the PCs. It is decided to centre each column of the data matrix to have mean zero and then scale each column to have variance one, similar to using a correlation

matrix in the calculations. Let the scaled matrix be $\check{\mathbf{X}}$.

The PCA procedure is now carried out. The first PC is found to explain 99.92% of variance in the data, with the second largest PC only explaining 0.043% of variance in the data. This is illustrated by a scree plot, which shows

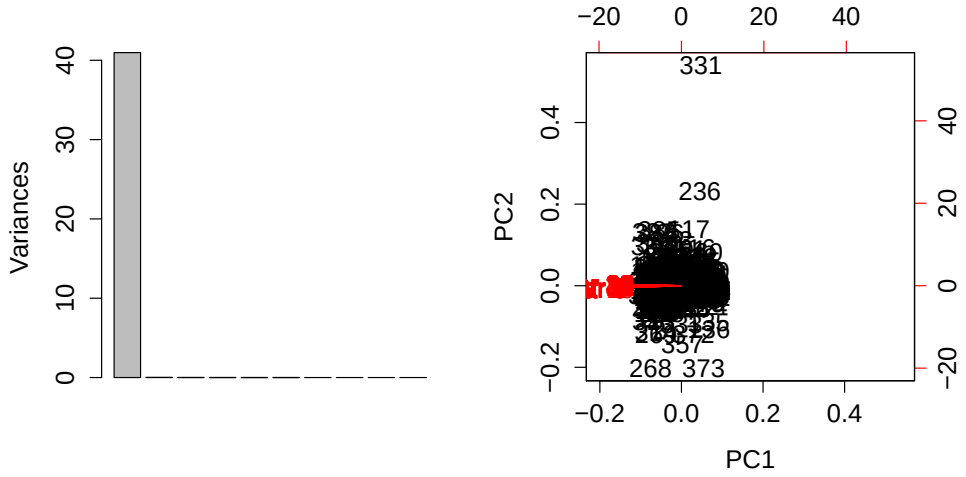

Figure 2.14: Scree plot (left) and biplot (right) for S-mode PCA on ARC-Lake data.

the proportions of variance explained by the first ten PCs, and a biplot, which shows the first and second PCs and the directions of the greatest variance in the variables (see Figure 2.14). The scree plot shows that only the first component really explains any variance in the data, with all of the other PCs, from 2 onwards, explaining almost zero variance. In addition, the biplot shows that the directions of the greatest variances in the variables are all very similar, all in the direction of PC1, with almost no variance in the direction of PC2. The biplot shows that the score for time 331 on component 2 is fairly high, but plots of the data over time for each location do not show anything unusual about the data for this time point. This may be due to the fact that the main patterns of variance in the data are already explained by the first component, so that the remaining components reflect very small changes in patterns in the data.

The variable loadings for PC1, $\alpha_{1,1}, ..., \alpha_{41,1}$, are all approximately -0.156,

so that the first principal component represents an average of the temporal patterns for each location. Since this first component explains such a high proportion of the variance of the data (99.92%), any contrasts of temporal patterns between different locations explain very little variance, compared to the average pattern.

These results may be due to the patterns of temperature over time being very similar for all locations in Lake Balaton, so that the differences in temporal patterns in temperature between locations are very small. The PCA has not identified any distinct separate groups of remotely-sensed locations that share similar temporal patterns in temperature, with the exception of the case where all locations are put into a single group.

A plot of the loadings for the first PC is produced, along with their

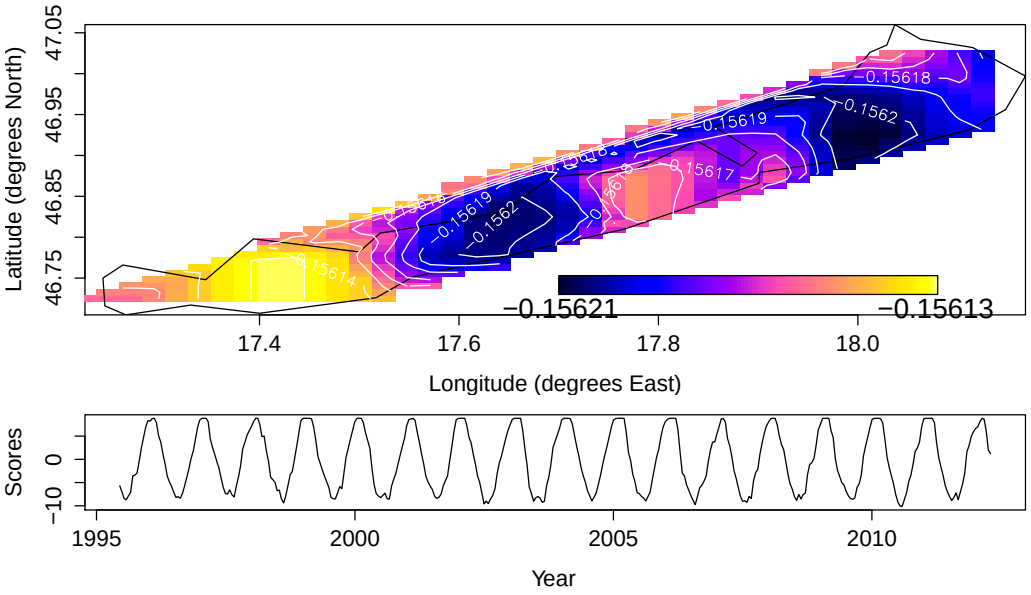

Figure 2.15: Plot of PC1 loadings (top) and scores (bottom) for S-mode PCA.

associated scores (see Figure 2.15). The plot legend makes it clear that there is very little variation in PC loading values across the lake. The scores have their highest values in winter, so they may represent the negative of average LSWT over time (i.e. highest values in summer and lowest values in winter). Loadings have been arbitrarily assigned negative values, so that the scores reflect this. The map of S-mode PC loadings shows very little variation in

the PC loading values, so that the variation in LSWT over time seems to be similar over all locations in the lake.

T-mode PCA T-mode PCA aims to find timepoints with similar spatial patterns (Richman 1986), which could enable the reduction of the temporal dimensionality of the data. A T-mode PCA, where observations of LSWT are formed into a matrix with timepoints as columns and locations as rows, can be carried out. Here, $\mathbf{X}$ is a (41×405) matrix. The scores matrix $\mathbf{T} = \check{\mathbf{X}}\mathbf{P}$ is a (41×41) matrix, where $\mathbf{P}$ is the (405×41) loadings matrix and $\check{\mathbf{X}}$ contains the centred and scaled columns of $\mathbf{X}$. Here, scores for each PC are a linear combination of scaled LSWT over time at each location. The first seven PCs cumulatively explain 39, 59, 71, 78, 84, 88 and 91 % of variance in the data, respectively, as shown in the scree plot (see Figure

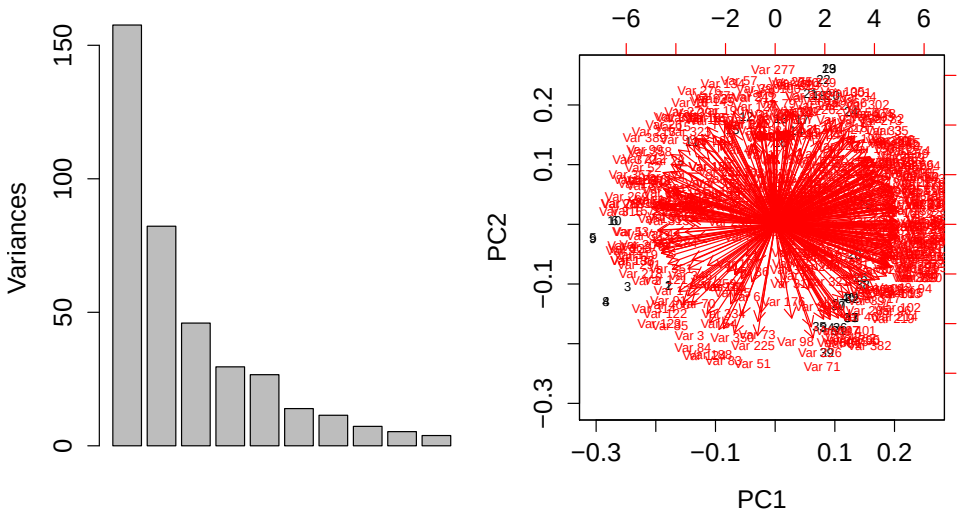

Figure 2.16: Scree plot (left) and biplot (right) for T-mode PCA on ARC-Lake data.

2.16). The appropriate number of PCs could be between 3 and 7, based on the percentages of variance explained. The biplot looks like a dandelion head, with variables pointing in all directions. Since the PCs are weighted averages of LSWT, with different weights at different times of the year, this pattern makes sense. LSWT increases and decreases fairly constantly throughout the

year, so while weights for one timepoint could be increasing with respect to one PC, they could be increasing or decreasing with respect to another.

A plot of the first T-mode PC is produced (see Figure 2.17). Here, the

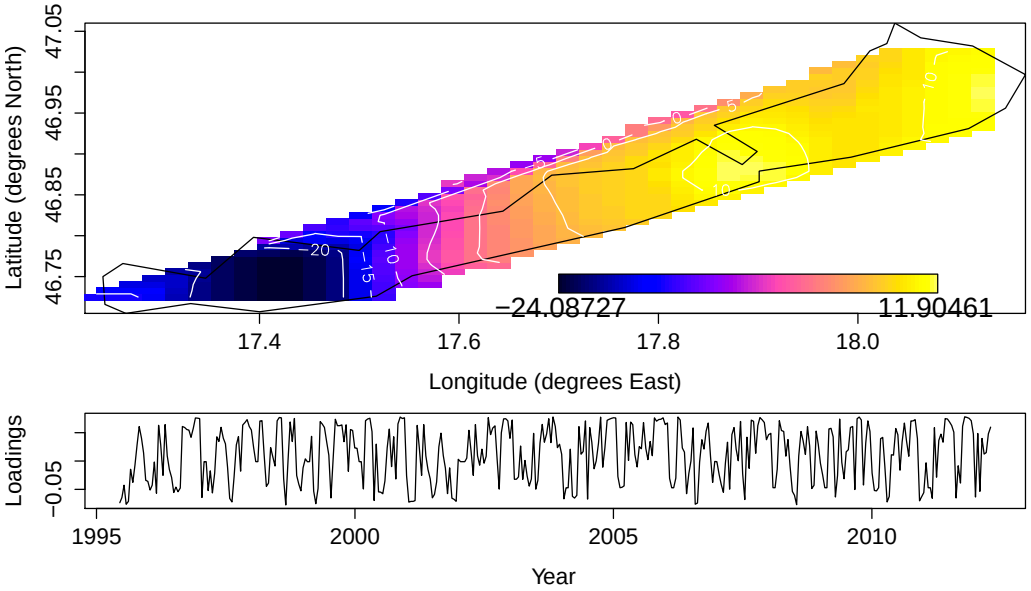

Figure 2.17: Plot of PC1 scores (top) and loadings (bottom) for T-mode PCA.

map represents the scores, rather than the loadings, and the time series plot represents the loadings, rather than the scores. The first T-mode PC seems to be an average of LSWT over the years, or at least represents the most common pattern of LSWT over the years. The loadings represent the time-points when this pattern is clearest in the data, i.e. highest values are when the observed pattern of LSWT is closest to the average over the years. Since several PCs are needed, it appears that there are several different patterns in LSWT over the lake, throughout the years of study.

Conclusions The conclusion from the S-mode PCA is that all locations in the lake have fairly similar temporal patterns of remotely-sensed temperature. The dimensionality of the data could possibly be reduced, by taking the average of the temperatures across the lake, at each time point. The results from the T-mode PCA are much less clear. A possible interpretation is that the spatial patterns of remotely-sensed temperature vary over time, with

few groups of times that share similar spatial patterns in remotely-sensed temperature.

PCA provides a useful method for dimensionality reduction and for providing an understanding of the main patterns in the data. In some applications, PCA may be the main piece of formal statistical analysis. However, the method cannot be used to make predictions at new locations, except in combination with another method, such as principal component regression. There is therefore a need to move onto more complex methods, while taking into account the understanding of the spatial and temporal patterns in the data that is gained from the PCA.

2.3 Investigating the relationship between *in situ* and remote sensing data through additive modelling

As noted in the introductory chapter, additive modelling provides a method for the smoothing of data over space and over time. Additive modelling is applied here, to predict temperature at *in situ* data locations, from remotely-sensed temperature data. This demonstrates whether smoothing the data spatially improves the estimates of *in situ* temperature, compared to simply taking the remote sensing data value at the nearest grid cell to the location of each *in situ* data point.

2.3.1 Application to the Lake Balaton data

In order to investigate how well the *in situ* and remotely-sensed data for $\log(\text{chlorophyll}_a)$ relate to each other, the datasets are matched. This could be accomplished by simply comparing the *in situ* data points to the remotely-sensed data points with the nearest cell centres in Euclidean distance. Alternatively, a smooth surface could be fitted to the remotely-sensed

data for each time point, with predictions made at the *in situ* data locations. This could be accomplished by kriging, as carried out in a previous section, but this section focusses on additive modelling, due to the ease of adding temporal components to the model.

Since several months have too few available remotely-sensed temperature data to fit a good spatial surface, the reconstructed fortnightly daytime remotely-sensed temperature data are used here. Firstly, the *in situ* temperature data are matched by their sampling dates to their nearest set of remotely-sensed temperature data in time. The data for the four *in situ* sampling locations that are monitored by the KDKVI are focussed on here, due to their fairly regular fortnightly sampling. Data are available between 2002 and 2006. The analysis is carried out using the `R` package `mgcv` (Wood 2006).

A model is fitted to the remotely-sensed temperature data:

$$y_{ij} = f(x_{1i}, x_{2i}) + f(w_j) + f(z_j) + \varepsilon_{ij}, \tag{2.8}$$

where y_{ij} is the remotely-sensed temperature for location i at time j, $f(x_{1i}, x_{2i})$ is a smooth function of longitude x_1 (degrees East) and latitude x_2 (degrees North), $f(w_j)$ is a smooth function of year and $f(z_j)$ is a smooth function of day of the year. $f(x_{1i}, x_{2i})$ and $f(w_j)$ are fitted by thin plate regression splines, while $f(z_j)$ is fitted by cyclic cubic splines, which ensures that the fitted value at day 1 is similar to the fitted value at day 365. The possible presence of spatial autocorrelation is assessed through the fitting of variograms to the model residuals, with only $f(x_{1i}, x_{2i})$ as a predictor, ensuring that the temporal autocorrelation does not enter the variogram calculations. The resulting fitted sample variograms, fitted to residuals for model 2.8 for each month, lie within the Monte-Carlo envelope (which consists of 100 simulated variograms with no spatial autocorrelation assumed (Diggle & Ribeiro 2007)), indicating that there is no evidence of spatial autocorrelation remain-

ing in the residuals. Temporal autocorrelation is assessed through fitting the model with only $f(w_j)$ and $f(z_j)$ as predictors and plotting autocorrelation and partial autocorrelation functions for each location. It is found that an autoregressive process of order 1 is appropriate, so that the errors are changed from $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ to $\varepsilon_i \sim \mathrm{N}_t(\mathbf{0}, \sigma^2\mathbf{V})$, where:

$$\mathbf{V} = \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{t-1} \\ \phi & 1 & \phi & \cdots & \phi^{t-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{t-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{t-1} & \phi^{t-2} & \phi^{t-3} & \cdots & 1 \end{pmatrix} \tag{2.9}$$

is the matrix of autocorrelation parameters. P-values for all three terms are less than 0.05, indicating that they should all remain in the model. These terms are plotted in Figure 2.18, to show their fitted effects. The estimated

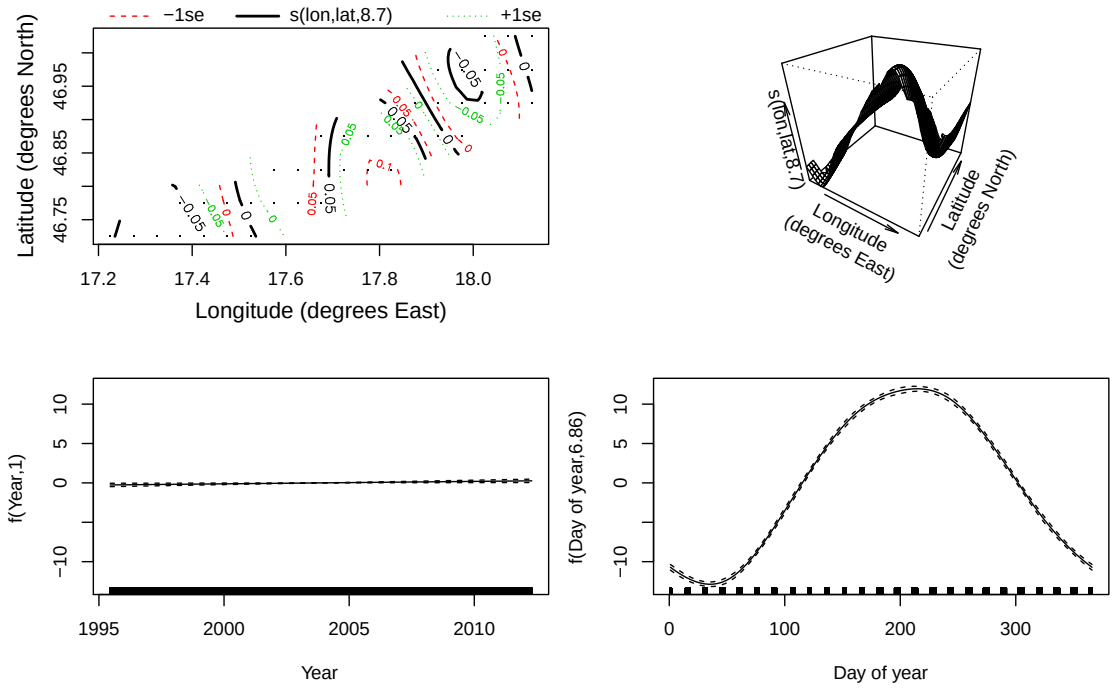

Figure 2.18: Plot of smoothing terms for Model 2.8a (left to right): contour and perspective plots of $f(x_{1i}, x_{2i})$ and plots of $f(w_j)$ and $f(z_j)$.

effect of longitude and latitude is much greater along the lake than across

the lake. It is estimated that remotely-sensed temperatures are greater in the centre of the lake and close to both ends of the lake. Year is estimated as being a linear term, so that it can be refitted as such, giving an estimated increase of 0.031°C for each one year increase, with a 95% confidence interval of 0.006°C to 0.056°C. The effect of day is fairly strong, with high values in summer, compared to winter. The model explains approximately 97.8% of the variability in the data, as assessed from adjusted R^2, so that it appears to be a very good fit to the data. A plot of predictions from model 2.8 at each of the four *in situ* locations (see Figure 2.19) shows that the variation

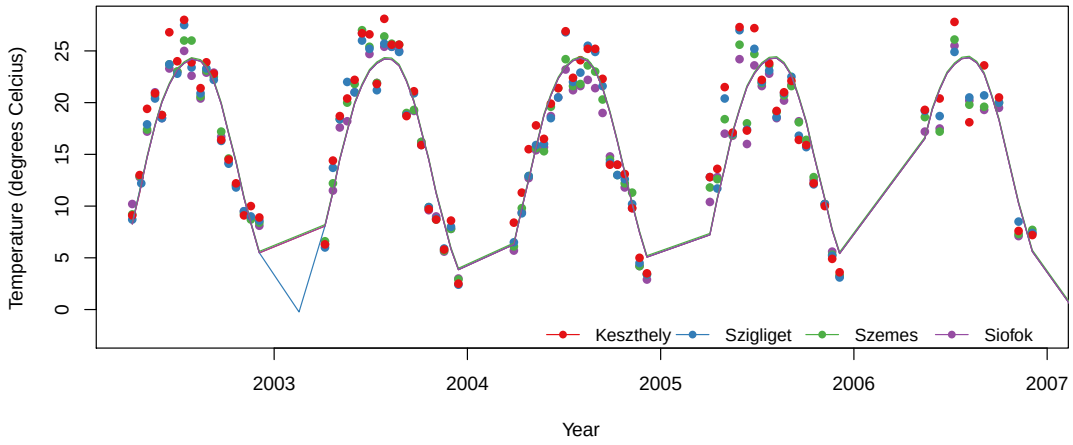

Figure 2.19: Plot of predictions from model 2.8 at *in situ* locations, showing remotely-sensed data (points) and predictions (solid lines).

between the locations is estimated as being very small in comparison to the variation over time, which seems to agree with the data.

Predictions from model 2.8 are made at the *in situ* data locations, for each fortnight. These are compared to the results from simply matching the *in situ* data to their nearest remote sensing grid cell centre, for each fortnight. Root mean squared errors (RMSE) are then calculated, to compare the model estimates, with smaller values representing a smaller difference between the predictions and the observed *in situ* data. The root mean squared errors for each set of predictions are 3.022 and 2.681, respectively, indicating that using the additive model including smooth terms for both space and time results

in improved predictions compared to simply matching *in situ* data to their nearest grid cell and fortnight centres.

2.4 Conclusions

This chapter has detailed exploratory analysis of both the *in situ* data and the remotely-sensed temperature data, in order to understand the patterns in the data and to highlight issues that might arise in subsequent analyses.

Strong cyclical patterns were observed in the *in situ* log(chlorophyll$_a$) and temperature data, with two peaks per year for the log(chlorophyll$_a$) data and one peak per year for temperature data. Patterns are less clear for log(total suspended matter), with much higher variability over time. Clear spatial patterns in log(chlorophyll$_a$) were also identified for Lake Balaton, with higher values in the southwest of the lake, near the in-flowing water from the River Zala. Some spatial variation in temperature was identified, although the variation over time was much stronger than that over space.

Mixed-effects models were fitted to the *in situ* data for log(chlorophyll$_a$) and log(total suspended matter). The model for log(chlorophyll$_a$) included terms for log(total suspended matter), longitude, latitude and cyclical terms for day of the year, suggesting that log(chlorophyll$_a$) is affected by the amount of sediment in the water and has changes in its levels over space and also over time. Similarly, the model for log(total suspended matter) suggested that the variable was affected by log(chlorophyll$_a$) levels and temperature, with a slightly positive trend over time and cyclical patterns within each year. These models and exploratory plots demonstrated the positive relationship between log(chlorophyll$_a$) and log(total suspended matter), and the less clear relationship between each of these variables and temperature.

Examining the remotely-sensed temperature data revealed that these data had better spatial coverage than the *in situ* data. An application of kriging demonstrated the smooth spatial patterns present in the data, but showed

that these differed between months. Principal component analysis was performed, in both S-mode and T-mode, which aim to find locations with similar temporal patterns, and to find times with similar spatial patterns, respectively. The results suggested that all locations have similar temporal patterns, with the first component explaining almost all of the variance in the data. The scores show that the pattern of high temperatures in summer and low temperatures in winter is common to all locations. The results of the T-mode PCA are less clear, with no single principal component explaining a large proportion of the variance in the data.

Additive modelling was performed, to investigate how smoothing the remotely-sensed data over space could improve the relationship between the *in situ* and remotely-sensed data. The work suggests that there is some variation over space, but that temporal variation is much stronger. There is a suggestion that accounting for different spatial and temporal sampling locations and times helps to bring the remotely-sensed estimates closer to the *in situ* data values.

Chapter 3

Statistical downscaling

This chapter discusses statistical downscaling methodology, focussing on its development and application in the context of lake water quality data, specifically log(chlorophyll$_a$) data. The chapter begins with a discussion of the background and motivation for the investigation of the technique, followed by the applications and model developments that these motivate. Finally, the conclusions that may be drawn from the work in this chapter are detailed, along with the requirements for further developments.

3.1 Background and motivation

The aim of this thesis is to develop methodology for the fusion of data with support that differs over space, over time, or over both space and time. This chapter deals with the spatial aspect of this, with the temporal development detailed in chapter 5. The *in situ* log(chlorophyll$_a$) data for Lake Balaton were obtained from water samples, taken at point locations directly from the lake surface and then analysed in a laboratory. Consequently, these samples are assumed to be accurate within measurement error. However, these data tell the investigator very little about the spatial patterns of log(chlorophyll$_a$) over the lake surface, since they cover only 9 locations within a fairly large lake, of approximately 596km^2 in surface area

(Palmer et al. 2015). Remotely-sensed data, captured by instruments aboard Earth-facing satellites that measure surface reflectance data and converted to log(chlorophyll$_a$) data via an algorithm, are also available for Lake Balaton. However, these data are available for averages over grid cells and over months, meaning that they have different spatial and temporal support from the point scale *in situ* data. Additionally, these data require calibration, due to the indirect nature of the data capture (via an algorithm, which results in the loss of information on uncertainty). The previous chapter contains work on additive modelling, which suggests that spatial smoothing of the remotely-sensed data leads to improved estimation over the lake. In this chapter, this idea will be taken further, so that the relationship between the *in situ* and remotely-sensed data will be modelled, with smoothly spatially-varying parameters. This chapter aims to develop methodology to fuse the *in situ* and remotely-sensed data, in order to take spatial information from the remotely-sensed data, but calibrate them using the assumed-accurate *in situ* data, thus providing improved estimates over the whole lake.

3.2 Spatial statistical downscaling: model development

This section introduces spatially-varying coefficient modelling. From this base, a spatial statistical downscaling model for data fusion of *in situ* and remotely-sensed log(chlorophyll$_a$) data is developed, followed by an application to data for Lake Balaton, which demonstrates the utility of the model for such data.

3.2.1 Spatially-varying coefficient modelling and statistical downscaling

In order to fuse data of different spatial support, a statistical model must be developed that can relate the point-scale *in situ* and grid-cell-scale remotely-sensed data. This can be accomplished through a model in which coefficients are allowed to vary smoothly over space, in a similar way to the model of Gelfand et al. (2003). In this section, the temporal change-of-support aspect is ignored, to be investigated in later sections. It is assumed here that the *in situ* and remotely-sensed data are collected at the same time, for each month. Let $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be an n-length vector of *in situ* data for a single time, where y_i is the *in situ* data value for location i (where $i = 1, \ldots, n$) and n is the number of *in situ* data locations, and let $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ be the vector of remotely-sensed data for the same time, for the grid cells that contain these *in situ* locations, so that x_i is the remotely-sensed data value for the grid cell that contains *in situ* location i. A simple model relating the two variables is the linear regression model:

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

for $i = 1, \ldots, n$, where $\varepsilon_i \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ are the random errors, with residual error variance σ_ε^2, α is the intercept coefficient and β is the slope coefficient, which together control the slope and intercept of the estimated line representing the relationship between $\mathbf{x}$ and $\mathbf{y}$. This model assumes that the errors have a Normal distribution and that the errors are independent and Normally distributed around zero, which must be checked after fitting the model to data. A more natural way of writing this model in the Bayesian framework makes explicitly clear the distribution of the data:

$$y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma_\varepsilon^2),$$

for $i = 1, \ldots, n$. Here, the mean of the Normal distribution for y_i is $\alpha + \beta x_i$ and the variance is σ_ε^2. The model is written in vector form as:

$$\mathbf{y} \sim \mathrm{N}_n(\alpha\mathbf{1} + \beta\mathbf{x}, \sigma_\varepsilon^2\mathbf{I}_n),$$

where $\mathbf{1}$ is an n-length vector of ones, $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathrm{N}_n()$ represents the multivariate Normal distribution for a vector of length n. Writing the model is this form is helpful for making clear the differences between this model and the more complex models developed in this chapter.

A model with spatially-varying coefficients is:

$$\mathbf{y} \sim \mathrm{N}_n(\boldsymbol{\alpha} + \boldsymbol{\beta} \odot \mathbf{x}, \sigma_\varepsilon^2\mathbf{I}_n), \tag{3.1}$$

where $\odot$ represents the Schur product, or Hadamard product, operation. The errors of model 3.1 are assumed to be independent, since the spatial dependence structure is accounted for through the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Here, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are now n-length vectors of coefficients, rather than scalars, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$. The model requires one intercept coefficient and one slope coefficient for each *in situ* data location, so some additional information is required in order to be able to fit the model. It is assumed that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ each vary smoothly over space, with a correlation structure that is both valid and fits decreasing correlation with increasing distance. The Matérn family of correlation functions does this and also allows for different degrees of smoothness in the underlying spatial processes (Diggle & Ribeiro 2007). The Matérn family was introduced in subsection 1.6.1 on page 20 and contains the exponential and Gaussian correlation functions. Although any member of the Matérn family would be a sensible choice, the exponential correlation function ($\rho(d) = \exp(-\phi d)$, where $\rho(d)$ is the correlation between data at distance d apart and ϕ is a spatial decay parameter, that controls how quickly the correlation decays towards zero as d increases)

is chosen. Compared to the Gaussian function, the exponential function is simpler, since choosing ϕ is more difficult on the squared scale of the Gaussian function, since changes in the spatial decay parameter cause larger changes in the resulting speed of decay to zero for the Gaussian function.

Model 3.1 is fitted in the Bayesian framework. This allows the incorporation of prior distributions, which provide information in the absence of information from the data. Fitting the model in the Bayesian framework allows the fitting of the model to data for a small number of locations, which is the case for the lake water quality data for Lake Balaton.

Prior distributions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are as follows:

$$\begin{aligned}\boldsymbol{\alpha} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big) \text{ and} \\ \boldsymbol{\beta} &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big),\end{aligned}$$

where the correlation structure chosen is the exponential spatial correlation structure, with $\mathbf{D}$ being the symmetrical $n \times n$ matrix of distances between *in situ* locations, so that:

$$\mathbf{D} = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & & \vdots \\ d_{n,1} & \cdots & d_{n,n} \end{pmatrix},$$

where $d_{i,j}$ is the distance between *in situ* locations i and j, for $i = 1, \ldots, n$ and $j = 1, \ldots, n$. ϕ_α and ϕ_β are the spatial decay parameters, representing how fast correlation decays towards zero as distance increases, while σ_α^2 and σ_β^2 are the spatial variance parameters, representing how variable estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are over space. The prior mean vectors are chosen to be $\mathbf{0}$ and $\mathbf{1}$, since the *in situ* and remotely-sensed data are both measures of the same variable, $\log(\text{chlorophyll}_a)$, so that the relationship between these data is expected to be relatively close to the line of equality. Variance parameters are given non-informative prior distributions, since nothing is known about

the values of these parameters *a priori*. These prior distributions are inverse-gamma distributions:

$$(\sigma_\alpha^2)^{-1} \sim \text{Ga}(a_\alpha, b_\alpha),$$
$$(\sigma_\beta^2)^{-1} \sim \text{Ga}(a_\beta, b_\beta) \text{ and}$$
$$(\sigma_\varepsilon^2)^{-1} \sim \text{Ga}(a_\varepsilon, b_\varepsilon),$$

where small values of a and b, such as 0.001 and 0.001, give noninformative prior distributions (Lunn et al. 2013). Small values of a and b have been used extensively in the literature, including in Clark & Gelfand (2006) and Waller & Carlin (2010). Gelman et al. (2014), however, note that there is no proper limiting distribution with $a = 0.001$ and $b = 0.001$ and state that posterior inferences are in fact sensitive to the choice of a and b. Sahu et al. (2006) and Sahu et al. (2010) instead recommend choosing the values 2 and 1 for a and b, respectively, to lead to a prior distribution with mean 1 and infinite variance. In order to investigate whether using the values $a = 2$ and $b = 1$ in place of $a = 0.001$ and $b = 0.001$ affects the results, model 3.1 is fitted using both of these sets of values for the prior distributions of the variance parameters. The spatial decay parameters, ϕ_α and ϕ_β are given uniform prior distributions, with endpoints chosen based upon the range of distances in the matrix $\mathbf{D}$. However, Sahu et al. (2006) note that the spatial decay parameters often suffer from identifiability problems when estimated along with spatial variance parameters. They also note that the computational complexity of the model calculations can be reduced, if spatial decay parameters are set to optimal values, rather than being estimated within the model. For example, in model 3.1, computational complexity is reduced through estimating $\exp(-\phi_\alpha \mathbf{D})$ and $\exp(-\phi_\beta \mathbf{D})$ only once, instead of recalculating these values at each iteration of the Markov Chain Monte Carlo sampler. Sahu et al. (2006) suggest instead that spatial decay parameters should be estimated using a grid search over a range of plausible values and then set equal to the values that result in best model fit. Upon application to

the log(chlorophyll$_a$) data for Lake Balaton, it is confirmed that convergence is poor for the spatial decay parameters, so a cross-validation is carried out in order to choose the best values for ϕ_α and ϕ_β. A full model has now been developed, which regresses the *in situ* data on the remotely-sensed data, with smoothly-varying coefficients. This model is a spatial statistical downscaling model, since it takes remotely-sensed data on a grid scale and calibrates it using point-scale *in situ* data, allowing prediction at any point location within a remotely-sensed grid cell. The model addresses the spatial change of support problem of the two types of log(chlorophyll$_a$) data.

3.2.2 Application of spatial statistical downscaling model 3.1 to log(chlorophyll$_a$) data for Lake Balaton

In order to better understand the model, it is applied to the data for log(chlorophyll$_a$), for Lake Balaton. There are remotely-sensed data available for 115 months, for 7616 remotely-sensed grid cells, but *in situ* data are only available for 9 locations. Of the 115 months in the remotely-sensed data, only 17 contain data for all 9 *in situ* locations, so this application makes use of a dataset of *in situ* data for 17 months, for 9 locations, and also the corresponding remotely-sensed data for the 9 grid cells containing the *in situ* data locations, for these 17 months. Since model 3.1 is suitable for modelling data for only a single time, it is fitted to data for each of the 17 months separately. In order to get the *in situ* and remotely-sensed data on the same temporal scale, all *in situ* data collected within a single month for a single location are averaged and assumed to have been sampled on the 15th day of that month. Similarly, the monthly-averaged remotely-sensed data for each grid cell are assumed to have been sampled on the 15th day of each month. This choice is arbitrary and a choice of any other day, for example, the first day of each month, would make no difference to the values of the predictions resulting from the model. The 15th day is simply chosen, since it is close to

the centre of each month. Chapter 5 presents methodology that removes the need for this assumption. However, an assumption such as this is required, in order to fit the initial statistical downscaling models, which deal with the spatial change-of-support only.

Taking the data for March 2011 as an example, the model is fitted to data for the 9 *in situ* sampling locations and their 9 corresponding remotely-sensed grid cells, with predictions made at grid cell centres using up to 7616 remotely-sensed data. The model-fitting procedure is carried out using `C++` and `R`, with the model itself fitted in `C++`, with help from the `Rcpp` (Eddelbuettel 2013, Eddelbuettel & François 2011) and `RcppArmadillo` (Eddelbuettel & Sanderson 2014) `R` libraries, and analysis carried out in `R` using the `coda` (Plummer et al. 2006) and `sp` (Bivand et al. 2013) libraries. The model is fitted through Markov Chain Monte Carlo, using Gibbs sampling, since all full conditional posterior distributions are of the forms of known distributions. The model is run for two chains, for 100,000 iterations each, after a burn-in period of 10,000, thinned to save parameter values for every 20th iteration. Two versions of the model are fitted, using the two different versions of prior distributions for variance parameters discussed in the previous subsection, namely Inv-Gamma(a, b), with $a = 0.001$ and $b = 0.001$, or $a = 2$ and $b = 1$.

Predictions are made from the fitted model at any location within the lake. However, since predicting within each of the 7616 grid cells would be computationally intensive, predicting instead for far fewer locations is preferable. In order to ensure that these locations allow a good understanding of spatial patterns in log(chlorophyll$_a$) throughout the lake, Delaunay refinement triangulation is carried out to choose prediction locations with optimal coverage of the lake surface. A Delaunay triangulation provides the optimal triangulation of an area (Shewchuk 1997). Delaunay refinement was used, for example, by Wan & Hu (2013) to determine the optimal locations of boreholes, so that an area under geological examination could be covered with the minimum number of boreholes to leave no gaps larger than 1.5

kilometres. In the case of Lake Balaton, it allows the optimal selection of prediction locations across the lake, to ensure a fairly even coverage with no large gaps. Given a set of points $\mathbf{P}$, an unconstrained triangulation is a set of disjoint triangles that have vertices forming $\mathbf{P}$ and that fill the convex hull of $\mathbf{P}$ (Shewchuk 1997). An unconstrained Delaunay triangulation of $\mathbf{P}$ is a triangulation such that any line in the triangulation of $\mathbf{P}$ is such that there is a circle passing through the line endpoints (which are in $\mathbf{P}$) that does not contain any other points in $\mathbf{P}$ (Shewchuk 1997). This ensures that the Delaunay triangulation provides the optimal coverage of a surface, in the sense of maximising the minimum angles out of all possible triangulations (Shewchuk 1997). The triangulation is carried out using the `R` package `RTriangle` (Shewchuk 1996), using a constrained Delaunay triangulation, where the algorithm is constrained to only insert new nodes within the defined boundaries (Shewchuk 1997), which in this case are a set of points that have been manually selected with the help of a plot of the 7616 available grid cells. With the aim of obtaining around 1000 prediction locations in the lake, after some trial and error, the algorithm is run with a maximum allowable triangle area of 4.5311×10^{-5} units^2, giving a set of Delaunay nodes of length 997 and good spatial coverage of the lake. These nodes are shown in Figure 3.1, with the original boundary nodes shown in red and the inserted Delaunay nodes coloured in blue. The plot shows that the resulting set of 997 prediction locations have good spatial coverage.

Model checking for a Bayesian model includes both checking of the model assumptions and checking that the Markov chains have converged. The assumptions of the model, using each of the variance prior distributions discussed previously, are checked using a plot of residuals versus fitted values and a plot of observed quantiles of the distribution of the residuals versus quantiles from the standard Normal distribution (see Figure 3.2). Since the residuals show a random scatter around zero in both plots, without changing patterns over the fitted values, there is no evidence against the assumptions

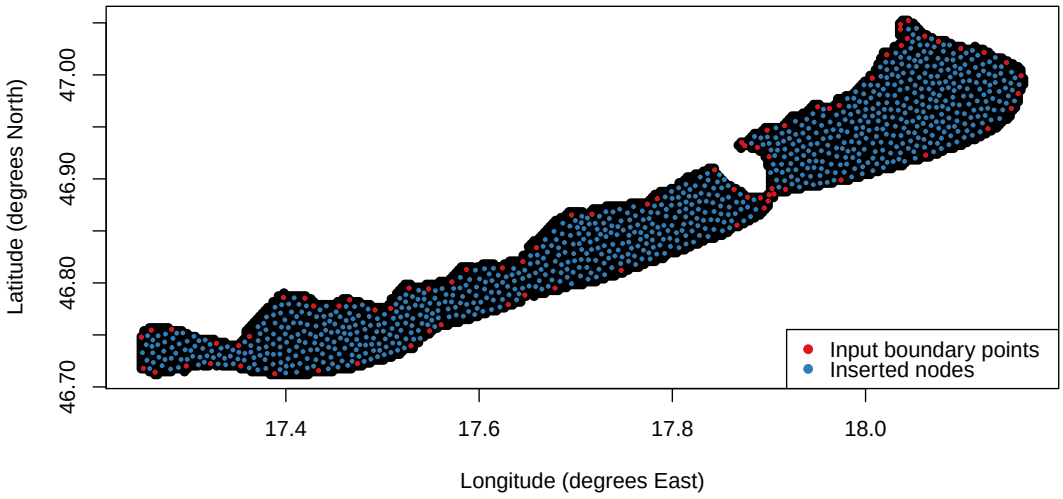

Figure 3.1: Plot of 997 nodes of Delaunay triangulation for Lake Balaton data, constrained by the input boundary points (red).

that the residuals have mean zero and that the variance of the residuals is homoscedastic. The points on each Q-Q plot lie close to a straight line, providing no evidence against the assumption that the residuals have a Normal distribution. Plots of parameter values for each iteration (trace plots) and their density plots provide no evidence against the assumption that the chains for each parameter have converged to their stationary distributions, so that inferences can be made using the resulting estimates.

Example trace and density plots are shown in Figures 3.3 and 3.4, for the Lake Balaton $\log(\text{chlorophyll}_a)$ data for October 2008, for the inverse spatial variances for the intercept and slope coefficients ($(\sigma_\alpha^2)^{-1}$ and $(\sigma_\beta^2)^{-1}$), for the inverse error variance ($(\sigma_\epsilon^2)^{-1}$), for the slope and intercept parameter for *in situ* data location 1 (α_1 and β_1) and for the predicted intercept, slope and *in situ* data value for prediction location 1 ($\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$). The trace plots show the characteristic shapes of converged chains, with values for both chains (i.e. the black line for chain 1 and the red line for chain 2) varying around a median that does not change over iteration number and with constant variance over iterations. The density plots also suggest that convergence has been reached, with right-skewed distributions for the inverse

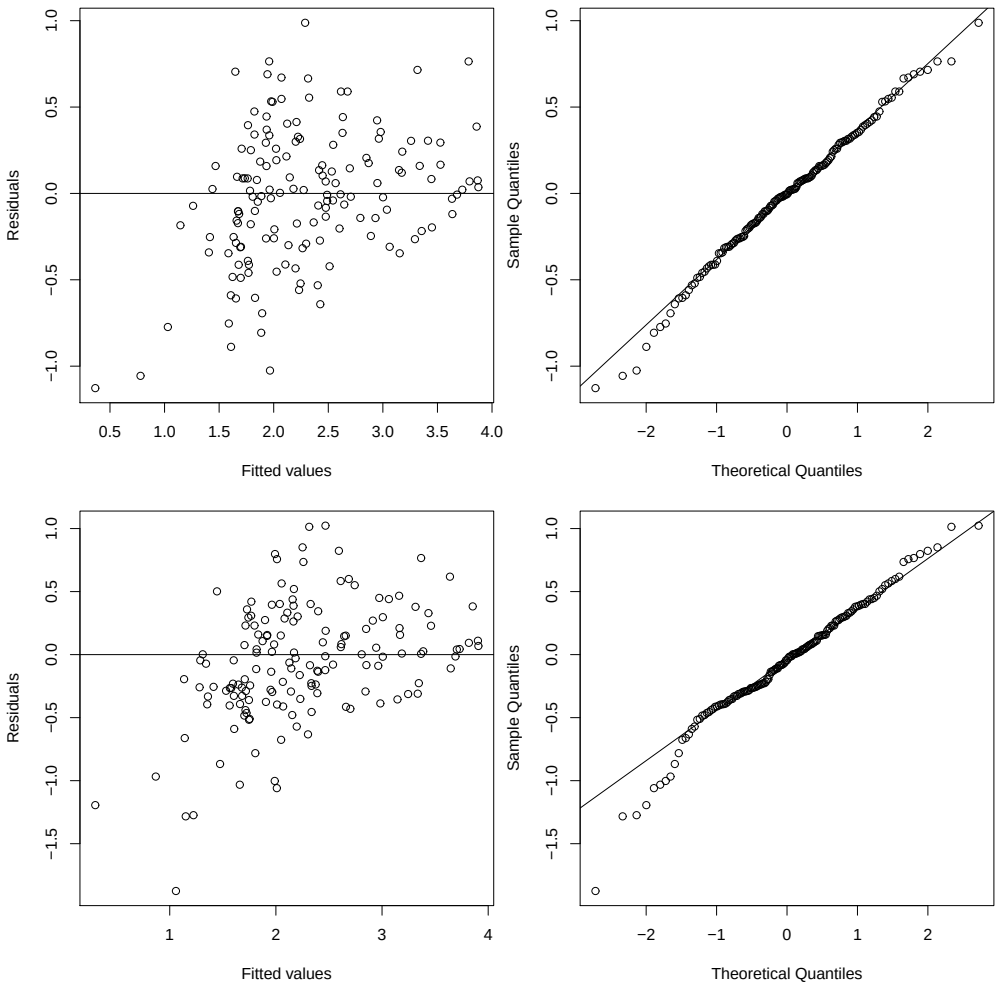

Figure 3.2: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.1, fitted to log(chlorophyll$_a$) data for October 2008, for Lake Balaton. The top plots are for model 3.1 with Inv-Ga$(0.001, 0.001)$ prior distributions for the variance parameters, while the bottom plots are for the model with Inv-Ga$(2, 1)$ prior distributions.

variance parameters ($(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$ and $(\sigma_\epsilon^2)^{-1}$) and unimodal bell-shaped curves for the remaining parameters, which have Normal posterior distributions. Figures 3.3 and 3.4 are compared to investigate the differences between the posterior distributions for the same parameters, when the variance parameters have prior distributions Inv-Ga$(0.001, 0.001)$ or Inv-Ga$(2, 1)$. The main difference between the two sets of plots is that those for the inverse-variance parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$ and $(\sigma_\epsilon^2)^{-1}$ have fewer spikes when the Inv-Ga$(2, 1)$ prior distribution is used, so that the maximum values for all

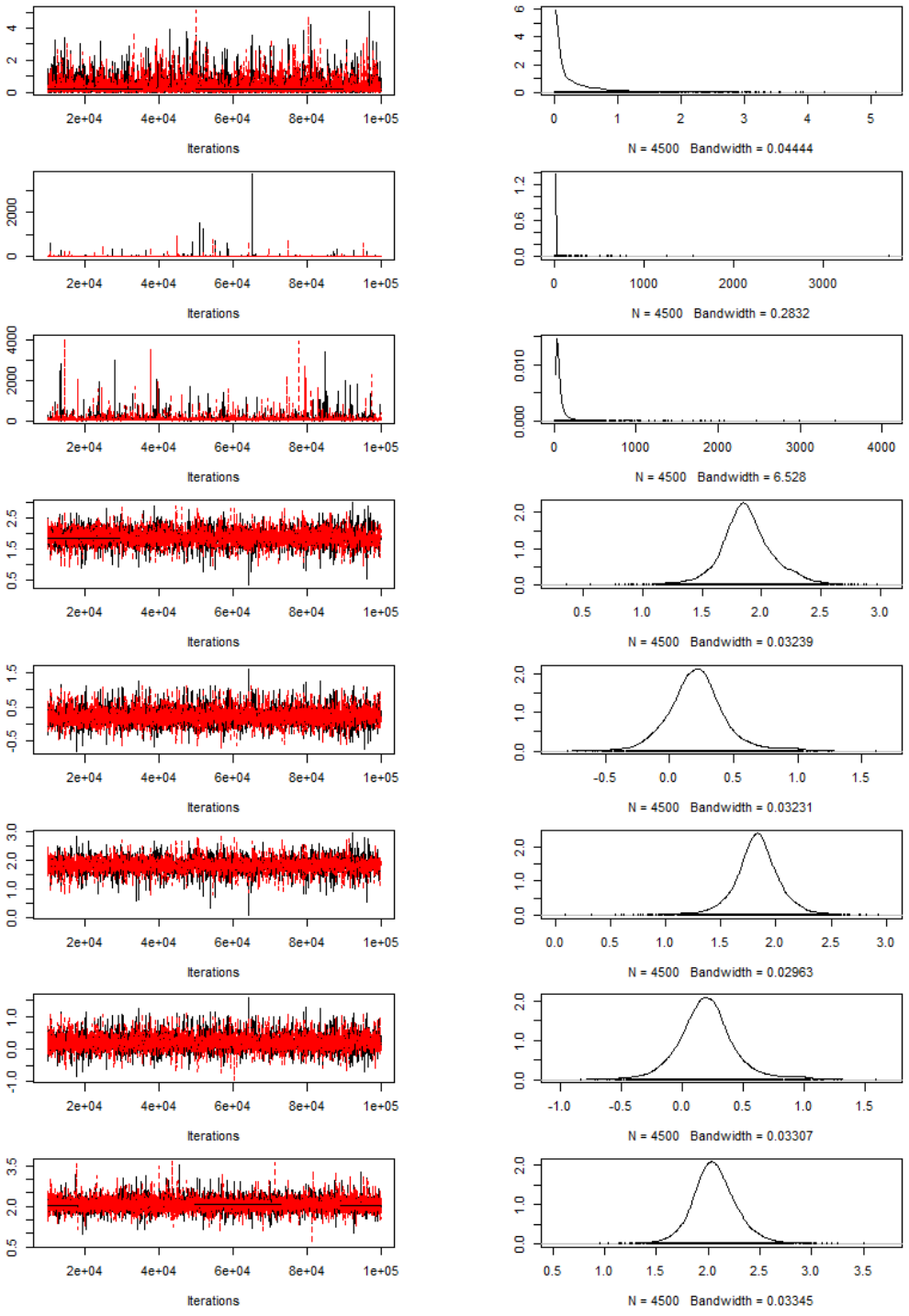

Figure 3.3: Example trace (left) and density (right) plots for parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\epsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$ (top to bottom), for model 3.1 with Inv-Ga$(0.001, 0.001)$ prior distributions for the variance parameters, fitted to the Lake Balaton data for October 2008.

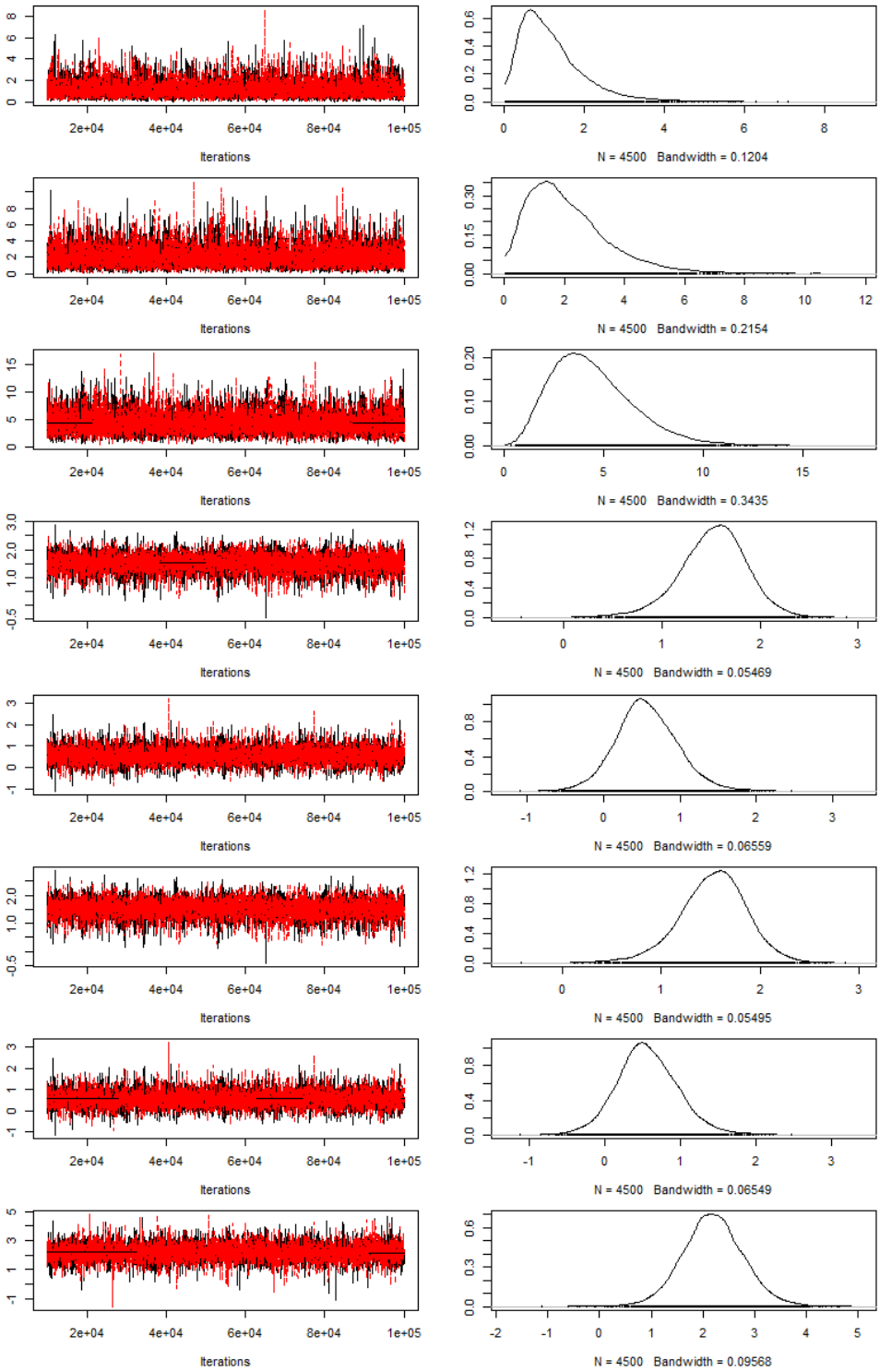

Figure 3.4: Example trace (left) and density (right) plots for parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\epsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$ (top to bottom), for model 3.1 with Inv-Ga$(2, 1)$ prior distributions for the variance parameters, fitted to the Lake Balaton data for October 2008.

iterations are less than 20, compared to over 3000 for $(\sigma^2_\beta)^{-1}$ and $(\sigma^2_\epsilon)^{-1}$, when the Inv-Ga$(0.001, 0.001)$ prior distribution is used.

As mentioned in the previous section, the spatial decay parameters exhibit poor convergence and must be set to appropriate values, rather than fitted as part of the model. These are chosen through a leave-one-out cross-validation, where data for each location are removed in turn and predicted using the model fitted to the remaining data. Once predictions have been made at each location, the predictions are compared to the observed *in situ* data, with the accuracy and precision of the predictions assessed through various summary statistics. Examples of these are root mean squared error (RMSE), mean absolute error (MAE), variance of predictions, 95% empirical credible interval coverage and mean 95% credible interval length:

- Mean squared error (MSE) is a measure of prediction accuracy and is defined as $\text{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}$, where n is the length of *in situ* data vector $\mathbf{y}$ and $\hat{\mathbf{y}}$ is the vector of predictions. MSE is a measure of the variance-bias trade-off, since $\text{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \text{var}(\hat{\mathbf{y}}) + \big(\text{bias}(\hat{\mathbf{y}}, \mathbf{y})\big)^2$. RMSE is simply the square root of MSE, i.e. MSE transformed back to the data scale. Smaller values are preferred.

- MAE is defined as $\text{MAE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n}$ and also provides a measure of prediction accuracy.

- Smaller variance of predictions $\hat{\mathbf{y}}$ is preferred.

- 95% credible interval empirical coverage provides an absolute measure of model performance, since the empirical coverage (i.e. what proportion of 95% credible intervals include the observed *in situ* data value) should be close to the nominal 95% coverage for any model.

- Smaller values of mean 95% credible interval length are preferred, since this indicates higher precision of model predictions.

This leave-one-out cross-validation is carried out for each combination of the sequence of values $0.001, 0.01, 0.1, 0.5, 5$ and 10 for ϕ_α and ϕ_β, giving a total of 36 cross-validation runs. This sequence of values is chosen to cover a wide range of possible parameter values, with 0.001 leading to high correlation across the entire lake and 10 leading to a small effective range of correlation. Summary statistics are calculated and plotted against values of ϕ_α and ϕ_β in Figure 3.5 (for the model using Inv-Ga$(0.001, 0.001)$ prior distributions for the variance parameters) and in Figure 3.6 (for the model using Inv-Ga$(2, 1)$). These plots show that smaller values of ϕ_α and

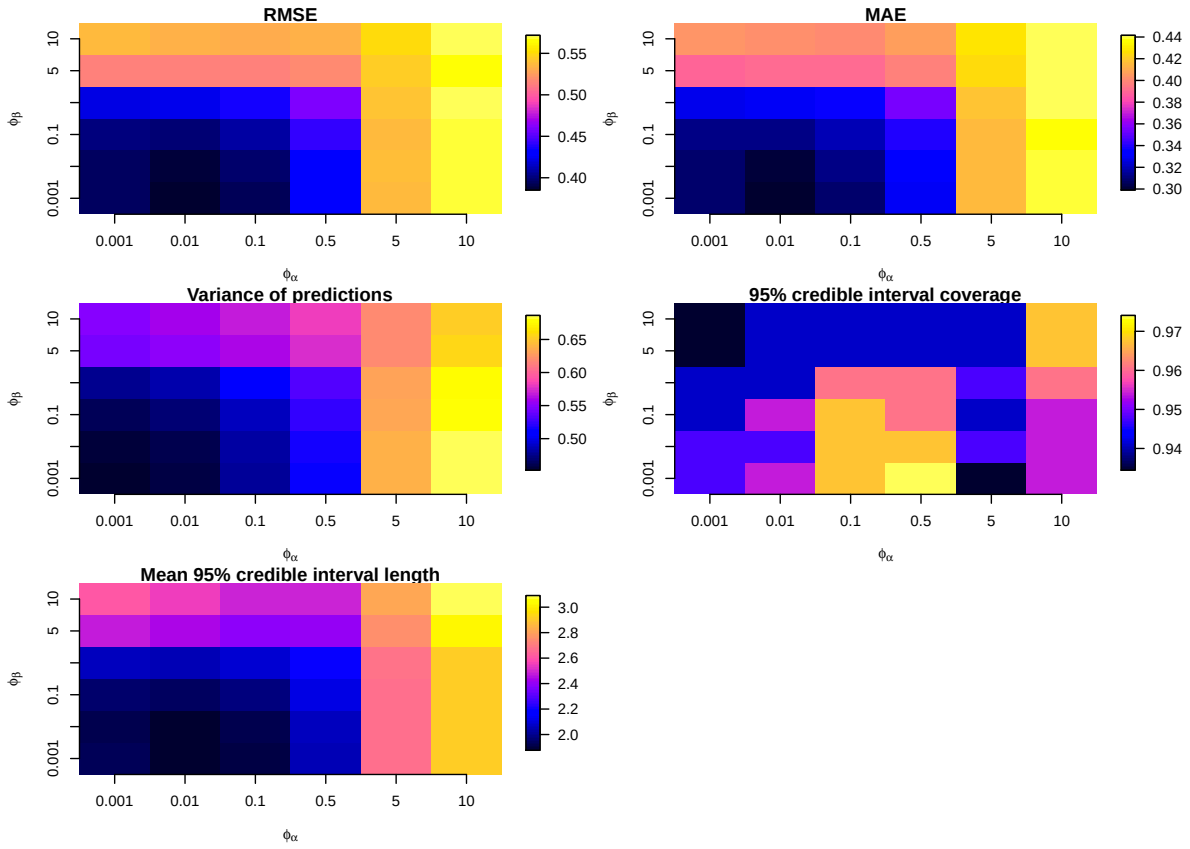

Figure 3.5: Plots of cross-validation summary statistics for model 3.1, for each combination of $\phi_\alpha = 0.001, 0.01, 0.1, 0.5, 5, 10$ and $\phi_\beta = 0.001, 0.01, 0.1, 0.5, 5, 10$.

ϕ_β are preferred for the Lake Balaton dataset. For the model using Inv-Ga$(0.001, 0.001)$ prior distributions, RMSE, MAE and 95% credible interval length reach their minima at $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$, while the variance

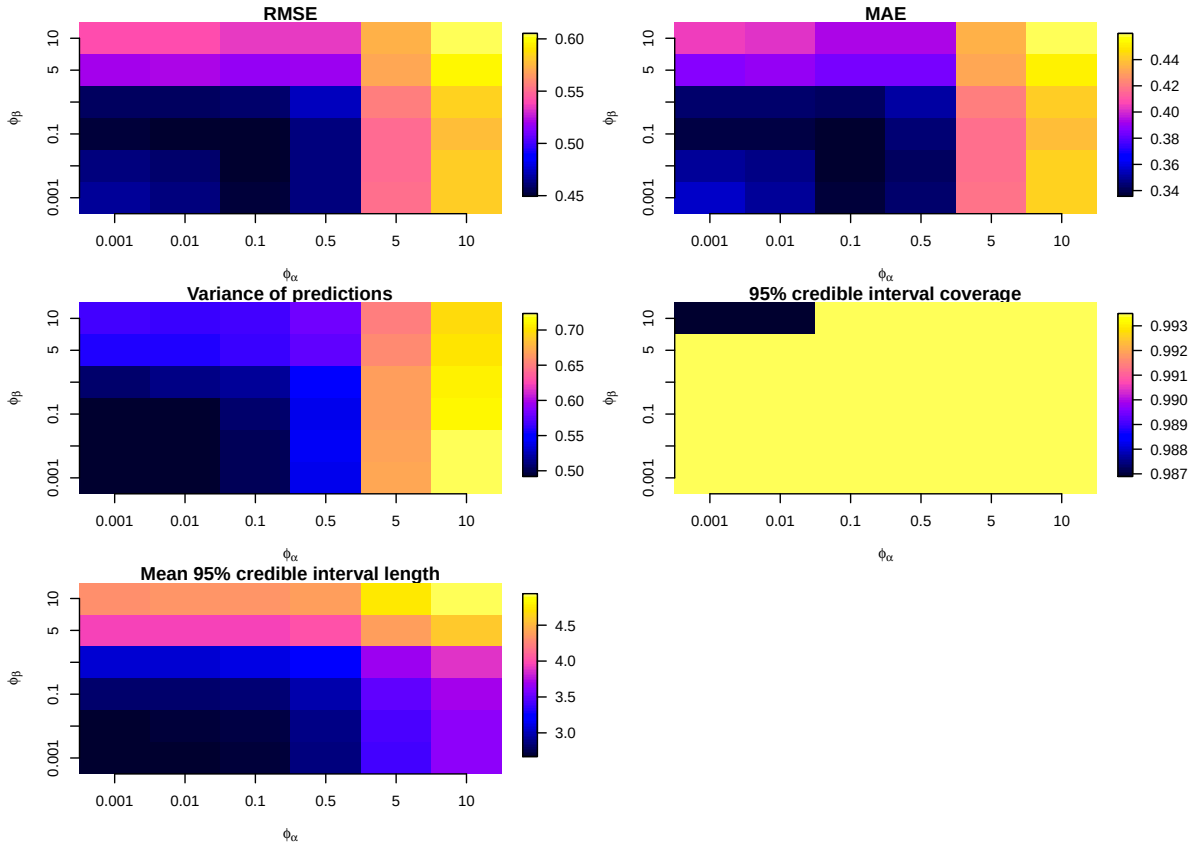

Figure 3.6: Plots of cross-validation summary statistics for model 3.1, for each combination of $\phi_\alpha = 0.001, 0.01, 0.1, 0.5, 5, 10$ and $\phi_\beta = 0.001, 0.01, 0.1, 0.5, 5, 10$.

of predictions reaches its minimum for $\phi_\alpha = 0.001$ and $\phi_\beta = 0.001$. The 95% credible interval coverage lies close to the nominal value for all values of ϕ_α and ϕ_β. From these results, values around $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$ are chosen as the most appropriate values for this model, although values up to say 0.1 and 0.1 could still be justified from these results. These small values of spatial decay parameters mean that the estimated correlations in intercept and slope coefficients decay slowly towards zero, with coefficients for locations far apart in the lake still showing correlation. No further investigation is carried out, to improve the estimates of ϕ_α and ϕ_β, since the small differences in the values of the summary statistics in the plots show that the model is not particularly sensitive to the values of these parameters, for this dataset. Figure 3.6 shows that the results for RMSE, MAE and variance of

predictions are similar for the model using Inv-Ga$(2, 1)$ prior distributions, compared to the results for the model using Inv-Ga$(0.001, 0.001)$ prior distributions. The 95% credible interval coverage is, however, slightly greater, while the mean 95% credible interval length is slightly longer, for the model using Inv-Ga$(2, 1)$ prior distributions.

These results do not provide evidence that the model is particularly sensitive to the choice of parameters for the inverse-gamma prior distributions for the variance parameters, or that the Inv-Ga$(0.001, 0.001)$ prior distributions have any adverse effects on the model predictions. For consistency, Inv-Ga$(0.001, 0.001)$ prior distributions are used for the variance parameters of each model in the remainder of this thesis, allowing a comparison between the models to be made. Further investigation of the effects of changing these prior distributions, for example to Inv-Ga$(2, 1)$ distributions, is left for future work.

Plots are produced of the original remotely-sensed data for these 997 locations and the corresponding calibrated values, with the *in situ* data overlaid on both plots (see Figure 3.7). The original remotely-sensed data display clear spatial patterns, with lower log(chlorophyll$_a$) values in the northeast basin of the lake, with other areas of lower values along the southern edge of the lake. The *in situ* data are available only along the length of the lake, so do not capture many of the spatial patterns in the lake. However, the *in situ* data do have higher values in the northeast lake basin. The corresponding predictions can be thought of as calibrated remotely-sensed data. These predictions are calibrated, so that predicted values near *in situ* data locations are much closer to these values than the original remotely-sensed data are. Without the aid of the white circles on the plot, it is difficult to identify the *in situ* data overlaid on the calibrated data, indicating the success of the modelling process in calibrating the remotely-sensed data. Plots of the predicted values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are produced (see Figure 3.8). These plots show that the values of the intercept and slope parameters do vary across the lake.

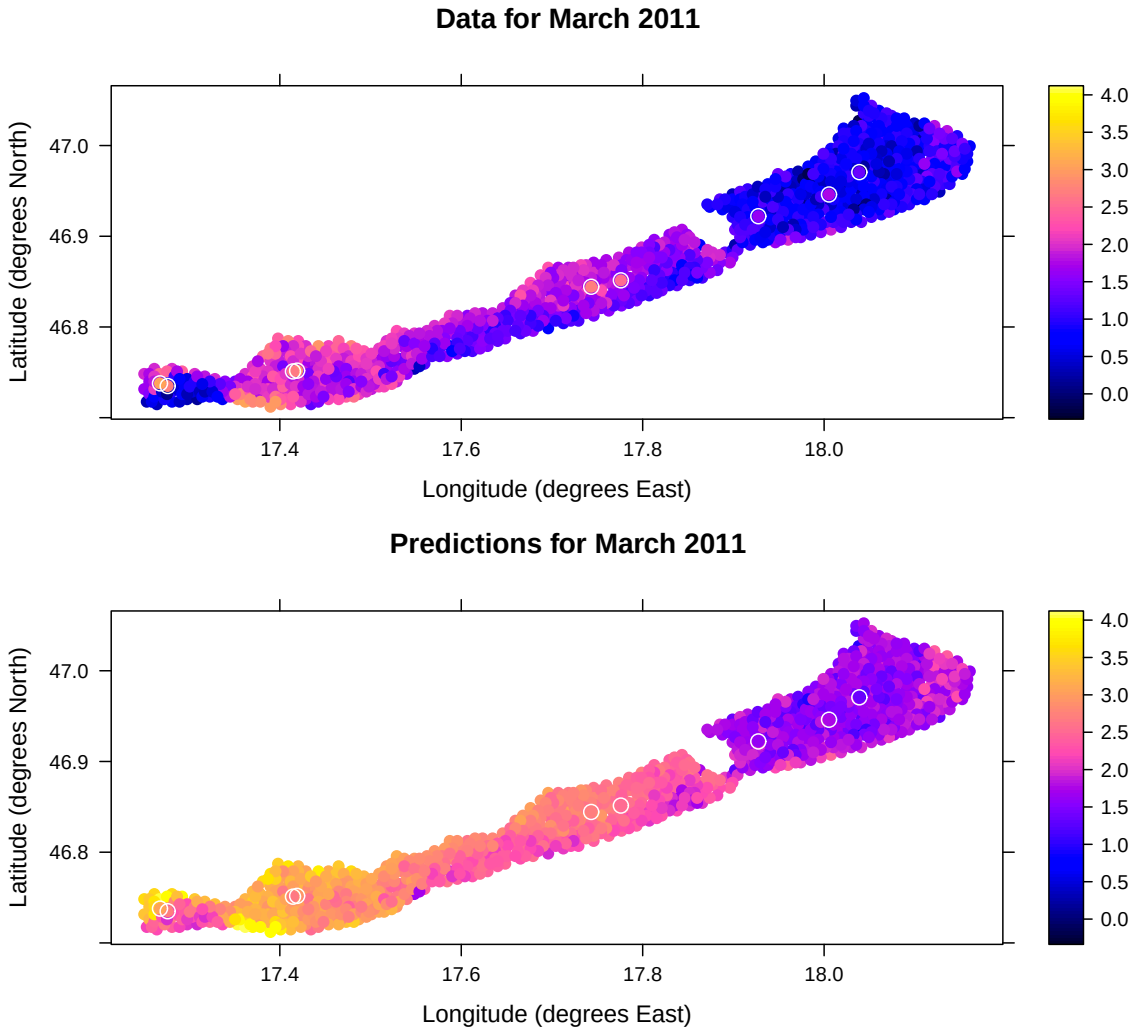

Figure 3.7: Data for March 2011 (top) and resulting predictions from model 3.1 (bottom). *In situ* data are overlaid on each plot, surrounded by white circles.

These parameters appear to be positively correlated, with similar patterns along the lake. The fact that the highest values of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are found in the southwesternmost basin of the lake is reflected by the fact that this part of the lake has the largest difference between the remotely-sensed data and the predictions from model 3.1, as shown in Figure 3.7. These plots show that the posterior predictions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are different from their prior mean values of $\mathbf{0}$ and $\mathbf{1}$, with ranges covering around 1.1 to 1.9 and 0.72 to 0.88, respectively. This means that the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are far from their prior means, for this particular month, with the relationship between the *in*

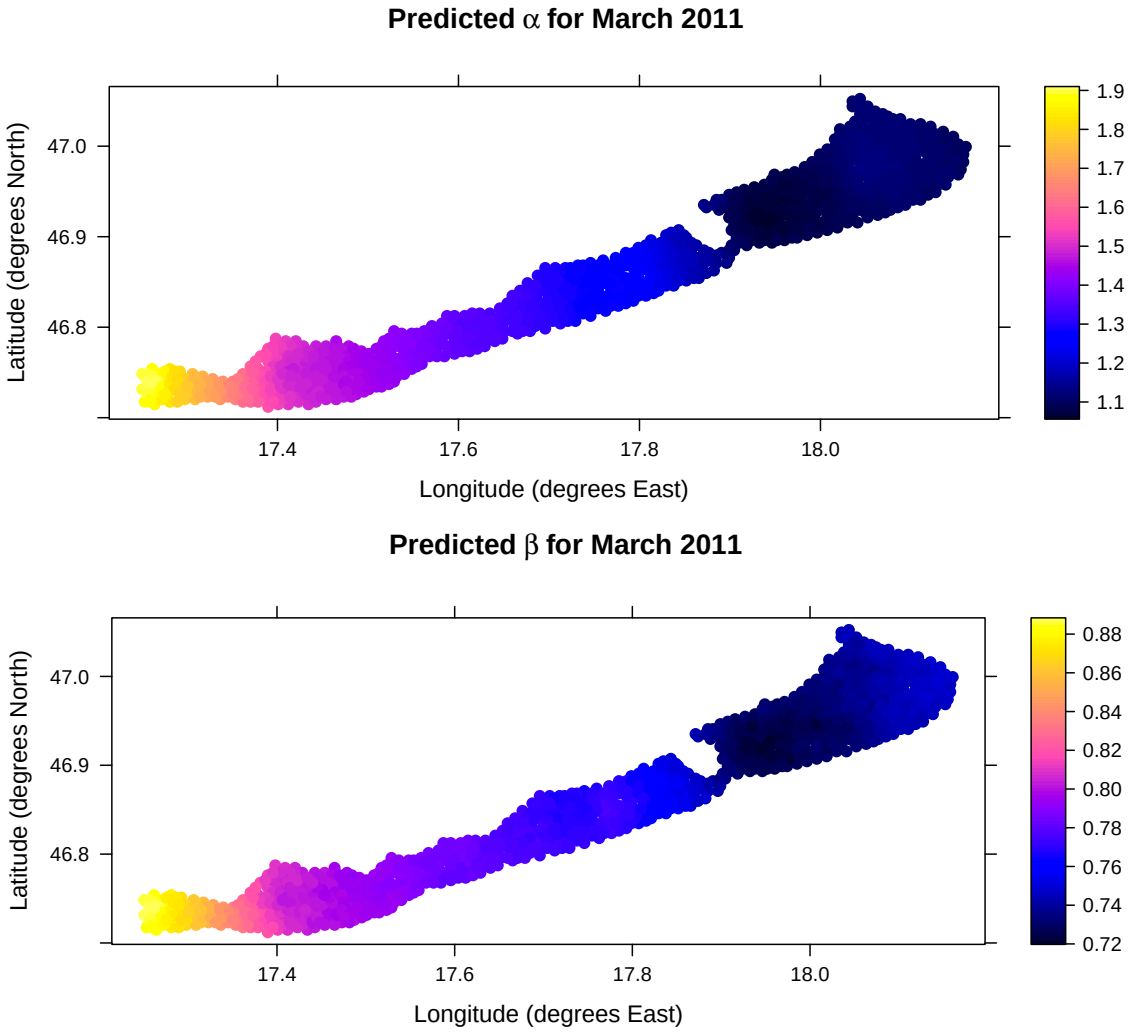

Figure 3.8: Predicted $\boldsymbol{\alpha}$ (top) and $\boldsymbol{\beta}$ (bottom) from model 3.1, for March 2011.

situ and remote sensing data having a positive intercept and a slope that is not parallel to the line of equality. This demonstrates that the calibration of the satellite data, through fusion with the *in situ* data, is required for this month, as the remotely-sensed data show a biased level of $\log(\text{chlorophyll}_a)$.

Finally, plots of the 95% credible interval bounds for predictions from model 3.1 for March 2011 are produced in Figure 3.9. The lower bounds lie between 0 and 2.5, while the upper bounds lie between 2 and 6.

Taken together, the information provided by the model for the example month of March 2011 is enough to conclude that the $\log(\text{chlorophyll}_a)$ levels for that month are highest on average in the southwest basins of the

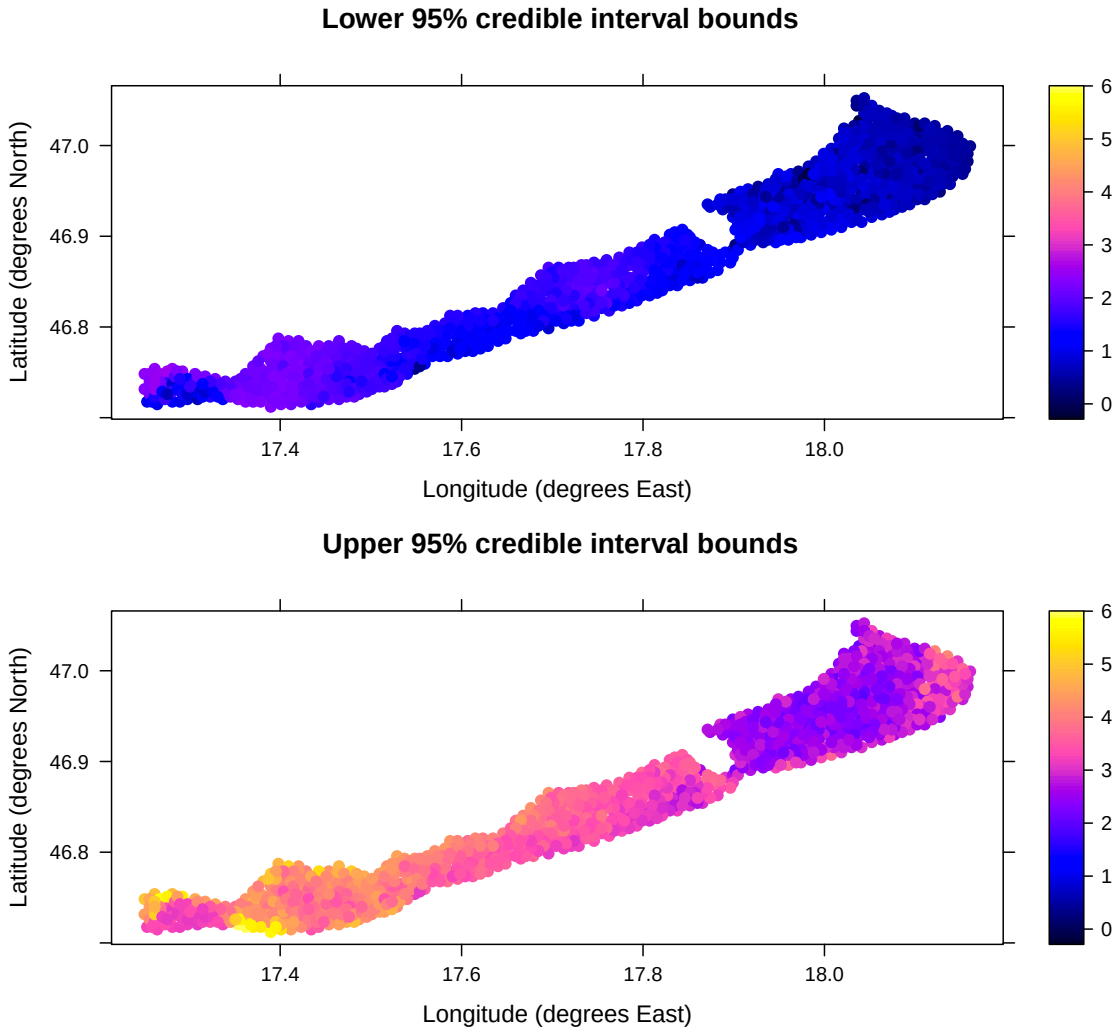

Figure 3.9: Upper (top) and lower (bottom) bounds for 95% credible interval for March 2011 predictions from model 3.1.

lake, near to the main inflow from the river Zala, with lower levels in the northeasternmost basin, near to the main outflow of the Sió canal. Median levels of log(chlorophyll$_a$) lie somewhere between 2 and 4 units, with 95% credible interval bounds from 0 to 6 units, for the southwesternmost part of the lake. Levels of log(chlorophyll$_a$) in the northeasternmost basin lie somewhere around 1 to 2 units, with 95% credible intervals from around -2 to 3 units. The model provides both estimates and quantified uncertainties for the investigator.

3.2.3 The Berrocal et al. (2010*b*) spatial downscaling model

Berrocal et al. (2010*b*) present a statistical downscaling model, building upon the spatially-varying coefficients model of Gelfand et al. (2003). Their application is a fusion of *in situ* air quality data $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ with modelled air quality data $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, with many *in situ* data available. Their model uses coregionalisation to model the correlation between slope and intercept parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$ and may be preferred in the case where correlation is expected between each pair of spatially-varying intercept and slope parameters. The model is:

$$
\begin{aligned}
\mathbf{y} &\sim \mathrm{N}_n\big(\gamma + \boldsymbol{\alpha} + (\delta + \boldsymbol{\beta}) \odot \mathbf{x}, \sigma_\varepsilon^2 \mathbf{I}_n\big), \\
\gamma &\sim \mathrm{N}(0, \sigma_\gamma^2), \\
\delta &\sim \mathrm{N}(0, \sigma_\delta^2), \\
\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} &= \begin{pmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} w_{0i} \\ w_{1i} \end{pmatrix}, \\
a_{11} &\sim \mathrm{logN}(0, \sigma_{11}^2), \\
a_{21} &\sim \mathrm{N}(0, \sigma_{21}^2), \\
a_{22} &\sim \mathrm{logN}(0, \sigma_{22}^2), \\
\mathbf{w}_0 &\sim \mathrm{N}_n\big(\mathbf{0}, \exp(-\phi_0 \mathbf{D})\big), \\
\mathbf{w}_1 &\sim \mathrm{N}_n\big(\mathbf{0}, \exp(-\phi_1 \mathbf{D})\big),
\end{aligned}
\tag{3.2}
$$

for $i = 1, \ldots, n$, where n is the number of *in situ* spatial locations in the model, a_{11} and a_{22} are given log-Normal prior distributions (since they act as spatial variances and should not go below zero), a_{21} acts as a kind of correlation parameter, $\mathbf{D}$ is the $n \times n$ matrix of distances between all *in situ* locations and ϕ_0 and ϕ_1 are spatial decay parameters, which are given uniform prior distributions (Berrocal et al. 2010*b*).

The connection between each pair of intercept and slope parameters α_i

and β_i is through the spatial decay parameter ϕ_0, since $\alpha_i = a_{11}w_{0i}$ and $\beta_i = a_{21}w_{0i} + b_{22}w_{1i}$. This means that it is assumed that the spatial pattern of the slope coefficients is related to the spatial pattern of the intercept parameters, with the strength of the relationship modelled using the parameter a_{21}.

The model is fitted to the log(chlorophyll$_a$) data for Lake Balaton and trace and density plots (see Figure B.2 on page 219) indicate that the convergence of the MCMC chains has been reached, while diagnostic plots (see Figure B.25 on page 241) provide no evidence against the model assumptions of homoscedasticity of residuals and mean-zero Normality of residuals. The parameters a_{11} and α_1 in Figure B.2 have some high values, which suggest that there may be a problem with the estimation of the parameter a_{11} for this dataset. It may be of interest to investigate different prior distributions for this parameter.

A leave-one-out cross-validation is carried out, to compare the performance of model 3.2 to that of model 3.1, for the Lake Balaton log(chlorophyll$_a$) data. Data corresponding to each of the 9 *in situ* locations are removed in turn and predicted using the model fitted to the remaining data. This process is carried out separately for data for each of the 17 months and root mean squared error (RMSE), mean absolute error (MAE), variance of predictions, mean 95% empirical interval coverage and mean 95% credible interval length are calculated using all data and predictions. These summary statistics are shown in Table 3.1. RMSE and MAE are lower for model 3.1 than for model

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.1	0.409	0.321	0.489	0.967	1.984
Model 3.2	0.460	0.348	0.526	0.941	2.290

Table 3.1: Table of summary statistics for leave-one-out cross-validation for models 3.1 (with $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$) and 3.2 (with $\phi_0 = 0.01$ and $\phi_1 = 0.1$).

3.2, so that 3.1 has more accurate predictions than model 3.2 for this dataset. The variance of predictions and the mean 95% credible interval length are

also lower for model 3.1, so that the predictions are more precise. The 95% empirical interval coverage is close to the nominal 95% for both models, so that there is no evidence against the suitability of the models for the data.

3.2.4 Simulation study

Since model 3.2 does not outperform model 3.1, a simulation study is carried out in order to investigate the circumstances in which one model outperforms the other, as measured by accuracy and precision of their predictions. The number of *in situ* locations is varied between 5 and 20, to give a range of values greater than and smaller than 9 *in situ* data locations for Lake Balaton. The values of the spatial decay parameters are varied between 0.001 and 10, to give a range of parameter values that cover the range assuming slow decline to zero (for 0.001) and fast decline to zero (for 10).

Firstly, a grid of simulated data values is created, with dimensions approximately equal to those of Lake Balaton. The data are simulated using the `R` package `RandomFields` (Schlather et al. 2015), using a Matérn covariance structure, with parameters estimated from the observed *in situ* data for Lake Balaton for one month. A grid of remotely-sensed data is simulated by fitting a linear model with the remotely-sensed observed data for Lake Balaton as the response and the *in situ* data as the explanatory variable and using this fitted relationship to predict a remotely-sensed data value for each simulated *in situ* data grid cell, with a small amount of random error added.

For spatial decay parameter values 0.001, 0.01, 0.1, 0.5, 5 and 10, $k + 20$ locations are randomly chosen 500 times, for $k = 5$, 9 and 20, where k is the number of simulated *in situ* data to which the model is fitted and 20 is the number of additional simulated *in situ* data locations at which predictions are made. The Berrocal et al. (2010*b*) model and model 3.1 are fitted to each set of k locations and predictions made for each corresponding set of 20 locations. Predictions are then compared to the simulated *in situ* data values and appropriate summary statistics calculated.

The two simulated datasets are plotted (see Figure 3.10), showing that

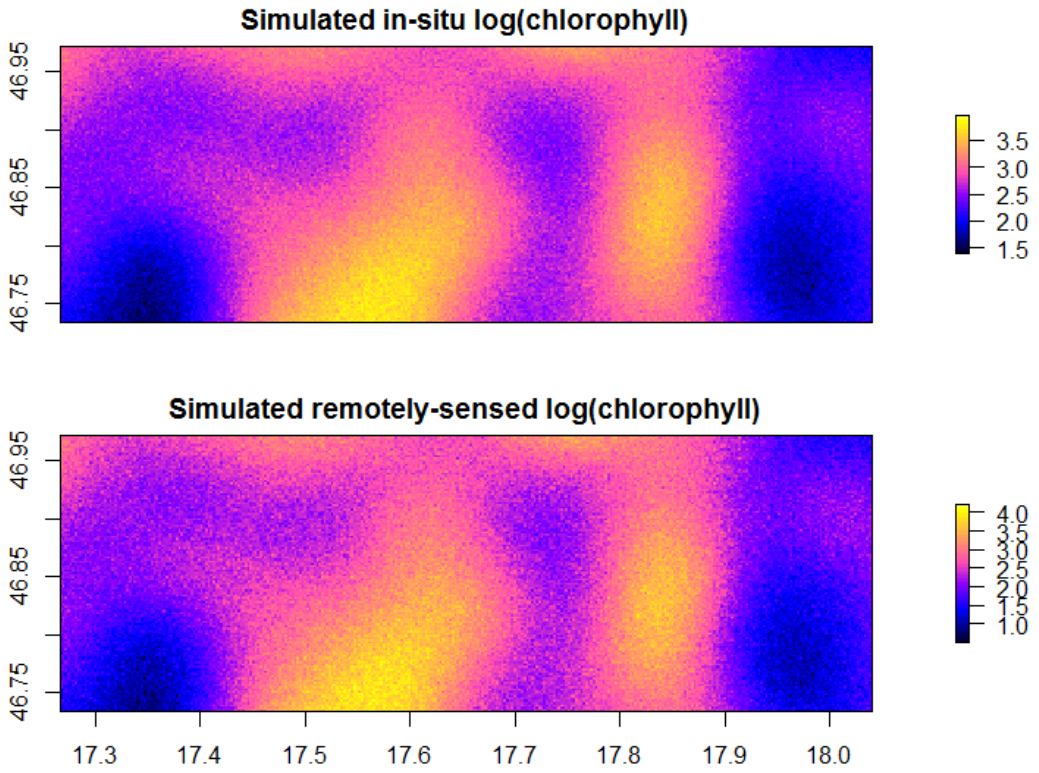

Figure 3.10: Plot of simulated *in situ* (top) and remotely-sensed (bottom) data.

both datasets follow similar spatial patterns, but that the remotely-sensed dataset generally has higher values than the *in situ* dataset.

Both models are fitted using `JAGS` (Plummer 2003) via `R`, for 2 chains each, for 500 iterations of initialisation (where the program generates initial values for each parameter), followed by a burn-in period of 10,000 iterations and the sampling period of 10,000 iterations, with every second iteration recorded (in order to save computer memory). For both models, the assumptions are checked (through plots of residuals versus fitted values and quantile-quantile plots, for a selection of model runs) and convergence is checked (through trace and density plots for a selection of parameters and model runs), giving no evidence against the validity of the model assumptions or the assumption of convergence of MCMC chains.

The resulting model performance summary statistics are plotted in Figure 3.11. RMSE, MAE, variance of predictions, 95% credible interval empirical

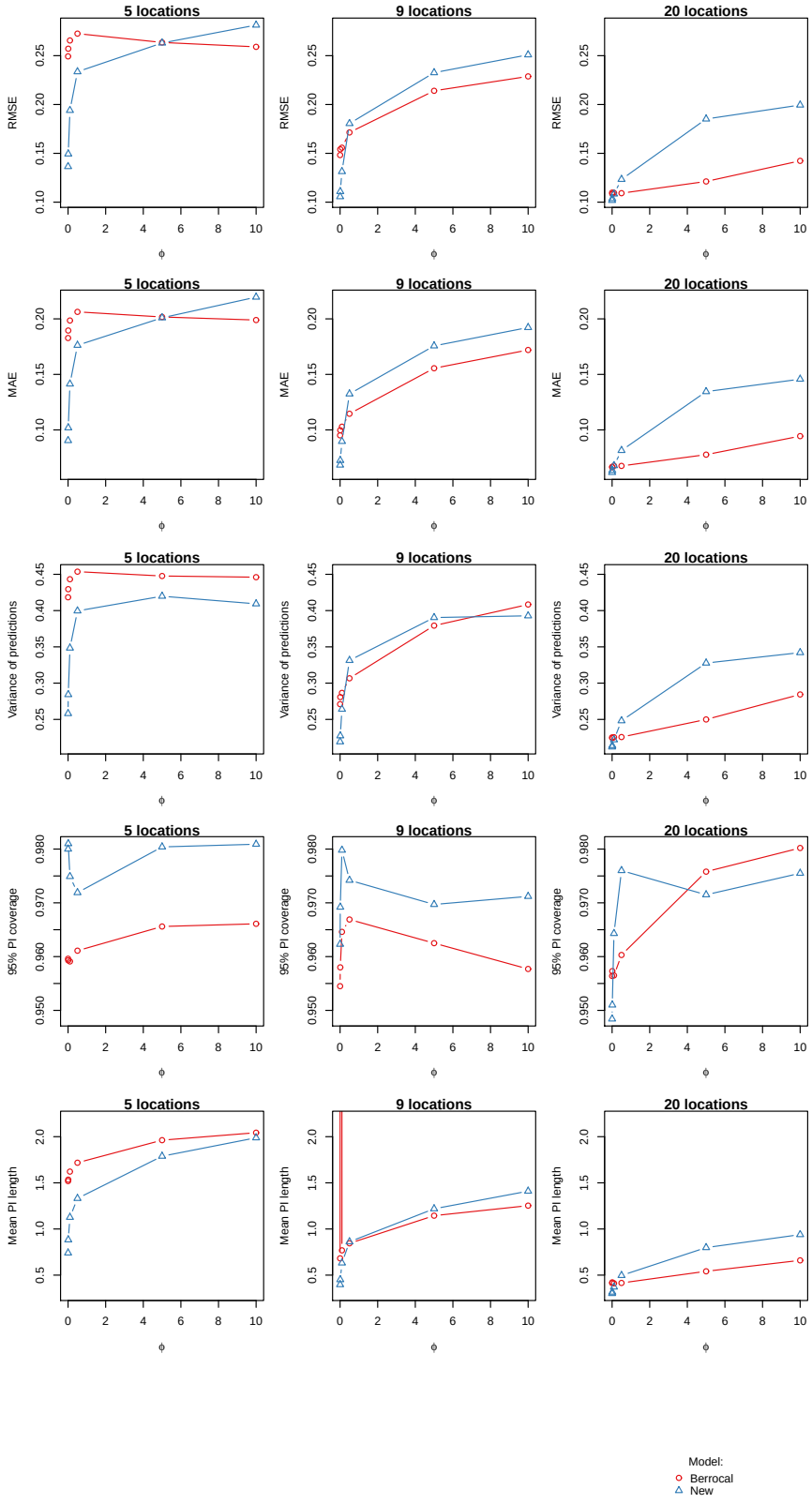

Figure 3.11: Plots of performance statistics from the simulation study for model 3.1 and the model of Berrocal et al. (2010*b*).

coverage and mean 95% credible interval length are all calculated and are assessed in turn. The RMSEs for both models generally increase with increasing ϕ, with model 3.1 having lower RMSE for smaller values of ϕ, for 5, 9 and 20 locations. The same pattern appears for MAE. Variance of predictions is lower for model 3.1 than for the Berrocal et al. (2010*b*) model, for small values of ϕ and for all values of ϕ for the models fitted to 5 sampling locations. Empirical 95% credible interval coverage is higher than the nominal 95% for almost all values of ϕ for all numbers of locations, for both models, indicating that both models are appropriate for the simulated data under study. Finally, mean 95% credible interval length generally increases with increasing ϕ and decreases with increasing numbers of sampling locations. For small values of ϕ, the interval length is smaller for model 3.1 than for the Berrocal et al. (2010*b*) model, but this reverses for higher values of ϕ, for 9 and 20 sampling locations.

These results suggest that model 3.1 should be preferred for this application, for small numbers of sampling locations, for small values of the spatial decay parameters ϕ. In the following chapters, the simpler model 3.1 will be extended and developed. Although this model does not explicitly model the intercept and slope parameters as being correlated, model 3.1 is both simpler and more computationally efficient than that of Berrocal et al. (2010*b*) and is shown to perform similarly to it and outperform it for small numbers of sampling locations and small values of spatial decay parameters ϕ.

3.3 Spatiotemporal statistical downscaling model development

In this section, spatiotemporal developments of the statistical downscaling model with spatially-varying coefficients (3.1) are discussed. The aim of this section is to allow the calibration of data for multiple months at once, motivated by the idea of improving parameter estimates and hence prediction

accuracy through the sharing of information over time.

3.3.1 Spatiotemporal development of model 3.1

Model 3.1 is re-written as:

$$\mathbf{y}_j \sim \mathrm{N}_n(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma^2_{\varepsilon j}\mathbf{I}_n), \tag{3.3}$$

where $\mathbf{y}_j$ is the vector of *in situ* data at time j ($j = 1, \ldots, t$) for locations 1 to n (where $\mathbf{y}_j = (y_{j,1}, \ldots, y_{j,n})^{\mathrm{T}}$), $\mathbf{x}_j$ is the vector of remote sensing data at time j for the grid cells containing locations 1 to n (where $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,n})^{\mathrm{T}}$) and $\sigma^2_{\varepsilon j}$ is the error variance for time j. The terms $\boldsymbol{\alpha}_j$ ($= (\alpha_{j,1}, \ldots \alpha_{j,n})^{\mathrm{T}}$) and $\boldsymbol{\beta}_j$ ($= (\beta_{j,1}, \ldots \beta_{j,n})^{\mathrm{T}}$) are modelled directly as spatially-varying terms, through their prior distributions:

$$\begin{aligned}\boldsymbol{\alpha}_j &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\alpha j}\exp(-\phi_{\alpha j}\mathbf{D})\big) \text{ and}\\ \boldsymbol{\beta}_j &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma^2_{\beta j}\exp(-\phi_{\beta j}\mathbf{D})\big),\end{aligned}$$

where $\mathbf{D}$ is the matrix of distances between *in situ* locations. Other prior distributions are:

$$\begin{aligned}(\sigma^2_{\alpha j})^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha),\\ (\sigma^2_{\beta j})^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta) \text{ and}\\ (\sigma^2_{\varepsilon j})^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon).\end{aligned}$$

This model results in identical computations to those resulting from model 3.1, but re-writing the model in this form makes clear the ways in which information can be shared over time, potentially improving the model performance and resulting in more accurate calibrated data from which to draw inferences. There is no temporal dependence structure within the model. The three variance parameters $\sigma^2_{\alpha j}$, $\sigma^2_{\beta j}$ and $\sigma^2_{\varepsilon j}$ are estimated separately in model 3.3, but can be estimated as σ^2_α, σ^2_β and σ^2_ε, with information pooled

across times. Should this be appropriate, each case of $\sigma^2_{\alpha j}$, $\sigma^2_{\beta j}$ and $\sigma^2_{\varepsilon j}$ in model 3.3 can be replaced with σ^2_α, σ^2_β and σ^2_ε, giving:

$$\begin{aligned}
\mathbf{y}_j &\sim \mathrm{N}_n(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma^2_\varepsilon \mathbf{I}_n), \\
\boldsymbol{\alpha}_j &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_\alpha \exp(-\phi_\alpha \mathbf{D})\big), \\
\boldsymbol{\beta}_j &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma^2_\beta \exp(-\phi_\beta \mathbf{D})\big), \\
(\sigma^2_\alpha)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\
(\sigma^2_\beta)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta) \text{ and} \\
(\sigma^2_\varepsilon)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon).
\end{aligned} \tag{3.3a}$$

The derivations of full conditional posterior distributions for models 3.3 and 3.3a are given in the appendix (see sections A.1 and A.2 on pages 196 and 201, respectively).

Models 3.3 and 3.3a are fitted to the log(chlorophyll$_a$) data for Lake Balaton and trace and density plots (see Figures B.3 and B.4 on pages 220 and 221) provide no evidence that the MCMC chains have not converged, while diagnostic plots (see Figures B.26 and B.27 on page 242) provide no evidence against the validity of the assumptions that residuals have zero mean and are homoscedastic and Normally distributed. Firstly, the values of $\sigma^2_{\alpha j}$, $\sigma^2_{\beta j}$ and $\sigma^2_{\varepsilon j}$ for each month j in model 3.3 are compared, to explore whether using a pooled estimate appears to be appropriate. The model (3.3) is fitted to the dataset of 17 months of log(chlorophyll$_a$) data for 9 locations in Lake Balaton, with $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$. Estimates for the three variance parameters are summarised in Table 3.2. This table shows that the estimates

	Minimum	1st quartile	Median	Mean	3rd quartile	Maximum
$\sigma^2_{\alpha j}$	1.352	2.710	3.708	3.610	4.805	5.401
$\sigma^2_{\beta j}$	0.063	0.474	0.762	0.888	1.140	2.538
$\sigma^2_{\epsilon j}$	0.013	0.041	0.056	0.081	0.119	0.234

Table 3.2: Summary table for estimates of variance parameters of model 3.3 for each month j.

of the variance parameters do vary between months, but a pooling estimate for each variance parameter can still be investigated. The two models 3.3 and 3.3a are therefore compared through a leave-one-out cross-validation, with data for one location of the 9 removed in turn and predicted using the remaining data for the 17 months, with $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$. The resulting summary statistics are presented in Table 3.3, showing that model

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.3	0.409	0.321	0.489	0.967	1.984
Model 3.3a	0.393	0.300	0.407	0.941	1.432

Table 3.3: Table of summary statistics of leave-one-out cross-validations for models 3.3 and 3.3a.

3.3a with the pooled variance estimates has improved prediction ability in comparison to model 3.3 (as assessed from RMSE and MAE), slightly lower prediction variance and improved precision of predictions (with narrower 95% credible intervals). Empirical 95% credible interval coverage is close to the nominal value for both models.

As written, model 3.3a cannot be fitted to months that have missing data, without using Bayesian methods for the imputation of missing data, since it is assumed that n is the same for each timepoint. A simple solution is to swap each occurrence of n in the model to n_j $(j = 1, \ldots, t)$, so that only the available data are used in the calculations for each month. This allows the calculation of the posterior distributions. The distance matrix $\mathbf{D}$ is replaced by $\mathbf{D}_j$, where $\mathbf{D}_j$ is the $n_j \times n_j$ matrix of distances between the n_j available *in situ* locations for month j. For clarity, the model is re-written below, as:

$$\mathbf{y}_j \sim \mathrm{N}_{n_j}(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma_\varepsilon^2 \mathbf{I}_{n_j}), \tag{3.4}$$

with spatially-varying coefficients

$$\boldsymbol{\alpha}_j \sim \mathrm{N}_{n_j}\big(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D}_j)\big) \text{ and}$$
$$\boldsymbol{\beta}_j \sim \mathrm{N}_{n_j}\big(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D}_j)\big),$$

and other prior distributions

$$(\sigma_\alpha^2)^{-1} \sim \mathrm{Ga}(a_\alpha, b_\alpha),$$
$$(\sigma_\beta^2)^{-1} \sim \mathrm{Ga}(a_\beta, b_\beta) \text{ and}$$
$$(\sigma_\varepsilon^2)^{-1} \sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon).$$

The main conclusion from this section is that model 3.3a and its variant model 3.4 are useful models for statistical downscaling of $\log(\text{chlorophyll}_a)$ data for Lake Balaton. Model 3.4 is also applicable in the presence of missing data.

In the remainder of this and in the following chapter, the aspects of model development that are discussed are the ability (or lack of ability) of the models to predict and calibrate data at new times, the potential of sharing information between variables, when multiple variables require simultaneous calibration, and the potential to calibrate data for multiple lakes at once. The next subsection focusses on the possibility of extending model 3.3a to include smoothing over time, as a preliminary investigation of whether prediction and calibration of data is possible for times outwith the set of times in the data, using the current statistical downscaling framework.

3.3.2 Spatiotemporal models including smoothing over time

This subsection consists of an investigation of extending model 3.3a to include smoothing over time, through both a model including autoregressive errors and a model including a temporal application of spatial covariance functions.

Spatiotemporal model with autoregressive errors

The first model extension is the model including autoregressive errors. The use of these errors induces correlation over time between model predictions, but assumes that the data are equally-spaced over time, which is not the case for the log(chlorophyll$_a$) data available for Lake Balaton. The model is:

$$\mathbf{y}_j \sim \mathrm{N}_n(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma_\varepsilon^2 \mathbf{I}_n), \tag{3.5}$$

for $j = 1, \ldots, t$, where $\mathbf{y}_j$ is the vector of *in situ* data at time j at locations 1 to n, $\mathbf{x}_j$ is the vector of remote sensing data at time j at the grid cells containing locations 1 to n and σ_ε^2 is the error variance. The terms $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are modelled directly as spatially-varying terms, through their prior distributions:

$$\begin{aligned}\boldsymbol{\alpha}_j &\sim \mathrm{N}_n\big(\boldsymbol{\mu}_j, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big) \text{ and} \\ \boldsymbol{\beta}_j &\sim \mathrm{N}_n\big(\boldsymbol{\nu}_j, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big),\end{aligned}$$

where $\boldsymbol{\mu}_j = (\mu_{j,1}, \ldots, \mu_{j,n})^{\mathrm{T}}$ and $\boldsymbol{\nu}_j = (\nu_{j,1}, \ldots, \nu_{j,n})^{\mathrm{T}}$ are mean vectors, with autoregressive hyperprior distributions:

$$\begin{aligned}\mu_{j,i} &\sim \mathrm{N}(\psi_\mu \mu_{j-1,i}, \theta_\mu^2) \text{ and} \\ \nu_{j,i} &\sim \mathrm{N}(\psi_\nu \nu_{j-1,i}, \theta_\nu^2),\end{aligned}$$

for $j = 1, \ldots, t$ and $i = 1, \ldots, n$. Other prior distributions are:

$$\begin{aligned}\psi_\mu &\sim \mathrm{Unif}(0, 1), \\ \psi_\nu &\sim \mathrm{Unif}(0, 1), \\ (\theta_\mu^2)^{-1} &\sim \mathrm{Ga}(a_\mu, b_\mu), \\ (\theta_\nu^2)^{-1} &\sim \mathrm{Ga}(a_\nu, b_\nu), \\ (\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha),\end{aligned}$$

$$(\sigma_\beta^2)^{-1} \sim \text{Ga}(a_\beta, b_\beta) \text{ and}$$
$$(\sigma_\varepsilon^2)^{-1} \sim \text{Ga}(a_\varepsilon, b_\varepsilon).$$

The autoregressive process coefficients ψ_μ and ψ_ν display poor convergence, so need to be chosen, instead of being fitted within the model. This poor convergence may be due to the small number of times for which data are available for Lake Balaton, of 17 months, making the estimation of these parameters difficult. Appropriate values of these parameters are chosen through a leave-one-out cross-validation, where for a selection of combinations of values for ψ_μ and ψ_ν, data corresponding to one of the 9 *in situ* data locations are removed in turn and predicted using the remaining data. The cross-validation is carried out for model 3.5 for ψ_μ and ψ_ν each set equal to values from the length-6 sequence $0.01, 0.2, 0.4, 0.6, 0.8, 0.99$, giving a total number of 36 combinations. The spatial decay parameters are set equal to the near-optimal values for this dataset, selected earlier, of $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$. Trace and density plots (see Figure B.5 on page 222) show good convergence for the parameters in the model, while diagnostic plots (see Figure B.28 on page 243) provide no evidence against the model assumptions that residuals have mean zero, are homoscedastic and Normally distributed. Root mean squared error (RMSE), mean absolute error (MAE), variance of predictions, 95% credible interval coverage and mean 95% credible interval length are calculated for each combination of parameters and are displayed in Figure 3.12. These plots show that the smallest RMSE and MAE are found for lower values of ψ_μ, with the lowest value of RMSE found for $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$ and the lowest value of MAE found for $\psi_\mu = 0.01$ and $\psi_\nu = 0.2$. Variance of predictions displays the opposite pattern, with values decreasing with increasing ψ_μ. This makes sense, since higher ψ_μ means more correlation over time and hence lower variability of predictions. Empirical 95% credible interval coverage is close to the nominal 95% for all combinations of parameters. Mean 95% credible interval length increases with increasing ψ_μ, with a minimum value reached at $\psi_\mu = 0.01$ and $\psi_\nu = 0.4$. The summary

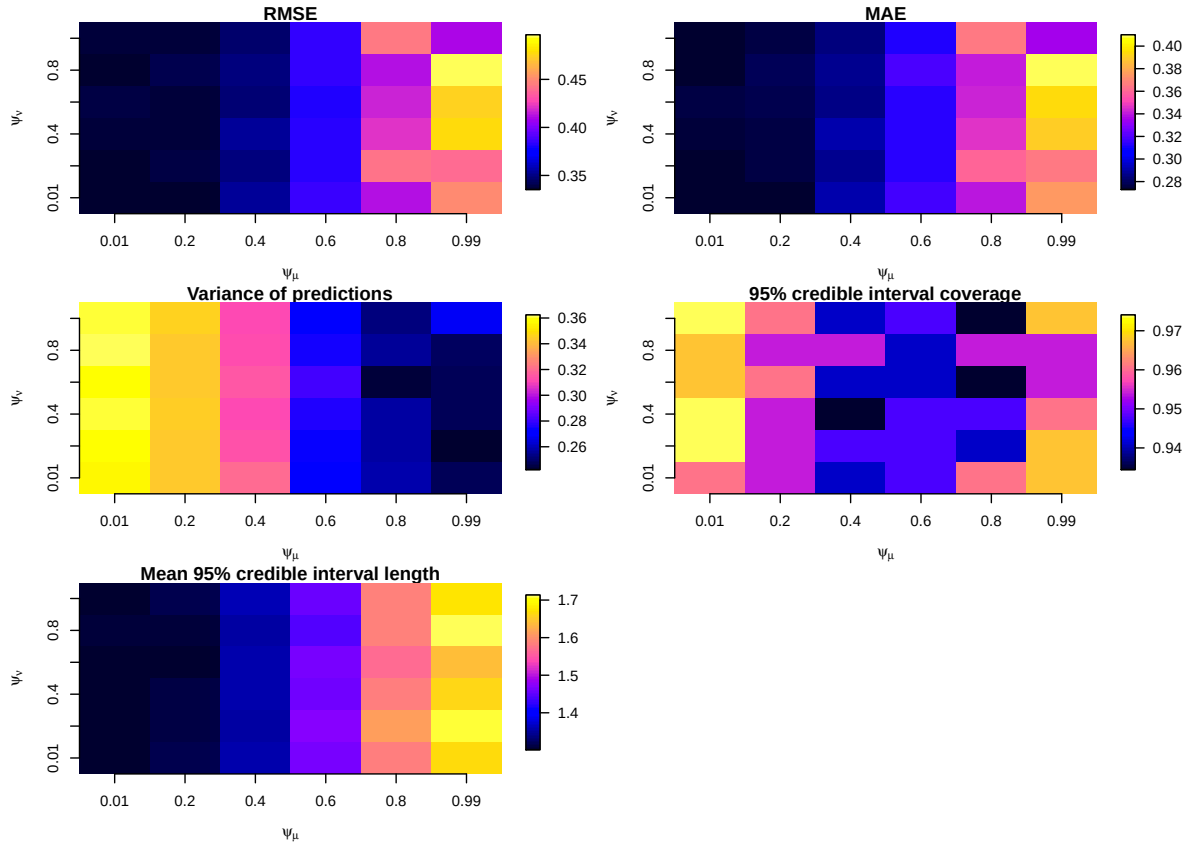

Figure 3.12: Plots of summary statistics for a leave-one-out cross-validation for model 3.5 for sequences of values of ψ_μ and ψ_ν.

statistics for model 3.5 with $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$ are given in Table 3.4. RMSE, MAE, variance of predictions and mean 95% credible interval length

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.5	0.336	0.275	0.343	0.954	1.316

Table 3.4: Table of summary statistics for leave-one-out cross-validation for model 3.5, with $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$.

are all slightly less for model 3.5, with $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$, than for model 3.3a, so there is some evidence here that including smoothing over time does improve the accuracy of the predictions from the model.

Spatiotemporal model using temporal covariance functions

The second model extension incorporates spatial covariance functions over time. Let this model be called (3.5a). Specifically, the exponential spatial covariance function is applied to a $t \times t$ matrix $\mathbf{T}$ of time periods between *in situ* data sampling times. The model equation is identical to that for model 3.5 on page 105, but the terms $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are modelled directly as spatially-varying terms, through their prior distributions:

$$\begin{aligned} \mathbf{y}_j &\sim \mathrm{N}_n(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma_\varepsilon^2 \mathbf{I}_n), \\ \boldsymbol{\alpha}_j &\sim \mathrm{N}_n\big(\boldsymbol{\mu}_j, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big), \\ \boldsymbol{\beta}_j &\sim \mathrm{N}_n\big(\boldsymbol{\nu}_j, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big), \end{aligned} \tag{3.5a}$$

with temporally-varying means $\boldsymbol{\mu}_j = (\mu_{j,1}, \ldots, \mu_{j,n})^{\mathrm{T}}$ and $\boldsymbol{\nu}_j = (\nu_{j,1}, \ldots, \nu_{j,n})^{\mathrm{T}}$, which are given hyperprior distributions:

$$\begin{aligned} \boldsymbol{\mu}_i &\sim \mathrm{N}_t\big(\mathbf{0}, \theta_\mu^2 \exp(-\psi_\mu \mathbf{T})\big) \text{ and} \\ \boldsymbol{\nu}_i &\sim \mathrm{N}_t\big(\mathbf{1}, \theta_\nu^2 \exp(-\psi_\nu \mathbf{T})\big), \end{aligned}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\mu}_i = (\mu_{1,i}, \ldots, \mu_{t,i})^{\mathrm{T}}$, $\boldsymbol{\nu}_i = (\nu_{1,i}, \ldots, \nu_{t,i})^{\mathrm{T}}$ and $\mathbf{T}$ is a $t \times t$ matrix of time periods between times $j = 1, \ldots, t$, calculated in the same way as for distances between spatial locations. The other prior distributions are:

$$\begin{aligned} (\theta_\mu^2)^{-1} &\sim \mathrm{Ga}(a_\mu, b_\mu), \\ (\theta_\nu^2)^{-1} &\sim \mathrm{Ga}(a_\nu, b_\nu), \\ (\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\ (\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta) \text{ and} \\ (\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon). \end{aligned}$$

The temporal decay parameters display poor convergence, so they are set equal to appropriate values, which are selected through a leave-one-out cross-

validation. The parameters ψ_μ and ψ_ν are each set equal to values from the length-6 sequence $0.001, 0.01, 0.1, 1, 5, 10$, giving a total of 36 combinations of values of ψ_μ and ψ_ν. A leave-one-out cross-validation is performed, where data for each of the 9 *in situ* locations for the Lake Balaton data are removed in turn and predicted from model 3.5a fitted to the remaining data, for each combination of ψ_μ and ψ_ν. The spatial decay parameters are set equal to the near-optimal values for this dataset, selected earlier, of $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$. Trace and density plots (see Figure B.6 on page 223) show evidence of convergence for the model parameters, while diagnostic plots (see Figure B.29 on page 243) provide no evidence against the model assumptions that residuals have zero mean, are homoscedastic and are Normally distributed. Plots of the resulting summary statistics are shown in Figure 3.13 for each combination of the parameters. There does not appear to be a

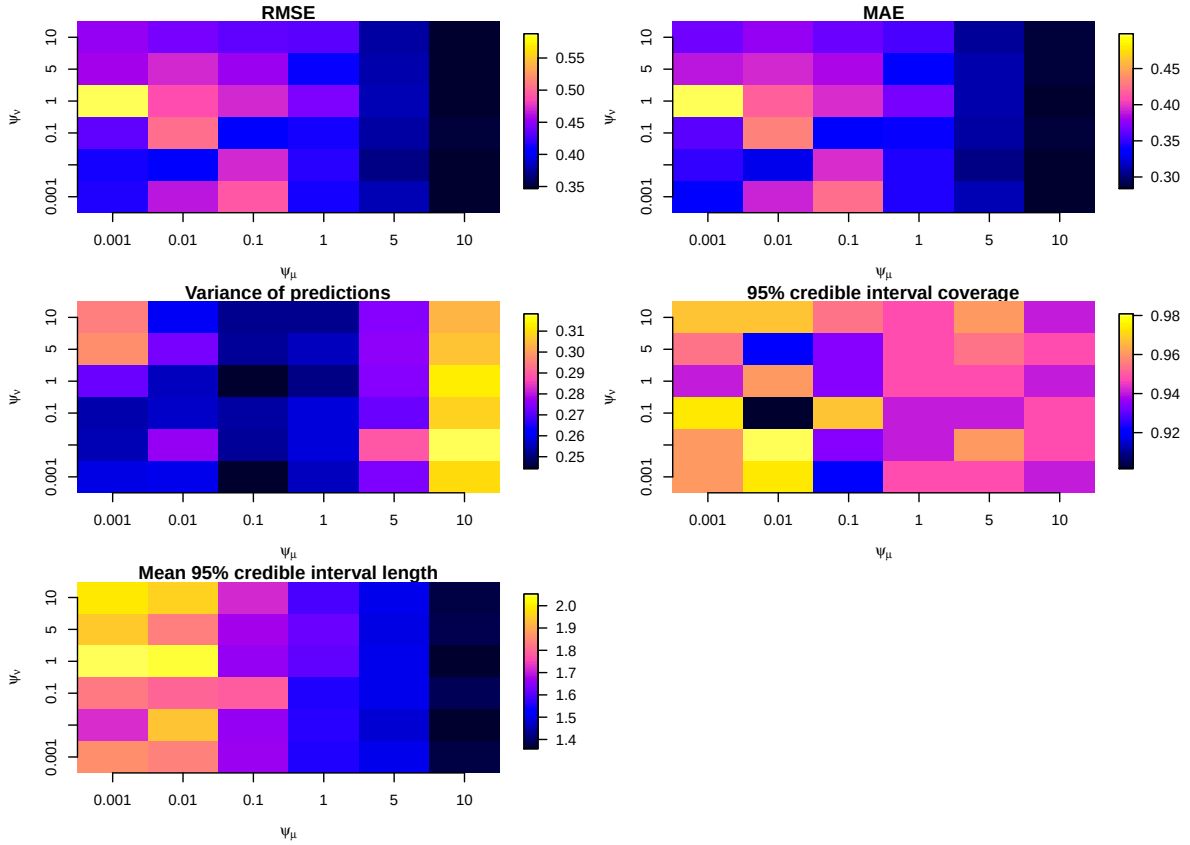

Figure 3.13: Plots of summary statistics for a leave-one-out cross-validation for model 3.5a for sequences of values of ψ_μ and ψ_ν.

clear relationship between any of the summary statistic values and ψ_ν, but RMSE, MAE and mean 95% credible interval length have their lowest values for $\psi_\mu > 0.1$. RMSE and MAE both reach their lowest values for $\psi_\mu = 10$ and $\psi_\nu = 0.001$. Variance of predictions reaches its lowest value for $\psi_\mu = 0.1$. $\psi_\mu > 1$ leads to generally higher variances of predictions. Empirical 95% credible interval coverage lies close to the nominal value for all fitted values of ψ_μ and ψ_ν. Taking $\psi_\mu = 10$ and $\psi_\nu = 0.001$, the values of summary statistics for the leave-one-out cross-validation are given in Table 3.5. This table

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.5a	0.348	0.285	0.31	0.941	1.382

Table 3.5: Table of summary statistics for leave-one-out cross-validation for model 3.5a, with $\psi_\mu = 20$ and $\psi_\nu = 15$.

shows that results from model 3.5a with $\psi_\mu = 10$ and $\psi_\nu = 0.001$ are similar to those from model 3.5 (i.e. the model using autoregressive parameters) with $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$. Both of these models result in lower values of RMSE and MAE than the spatiotemporal models without smoothing over time, models 3.3 and 3.3a. However, RMSE and MAE are the lowest for the model with autoregressive parameters, model 3.5.

Discussion of spatiotemporal models with smoothing over time

Two models have been developed in this subsection, to deal with smoothing over time in addition to smoothing over space. For model 3.5, which makes use of autoregressive parameters to smooth over time, the two parameters ψ_μ and ψ_ν control how strong the dependence is in the intercept and slope parameters over time. A leave-one-out cross-validation shows that the values of these parameters that lead to the most accurate predictions are $\psi_\mu = 0.2$ and $\psi_\nu = 0.01$, suggesting that correlation is low over time for both the intercept and slope parameters. For model 3.5a, which uses temporal covariance matrices to smooth over time, the parameters ψ_μ and ψ_ν are

temporal decay parameters, which control how fast correlation decays over time as the time period between observations increases, so that larger values indicate faster decay in correlation. Through a cross-validation, the values of ψ_μ and ψ_ν are estimated as 10 and 0.001, respectively, although any value of ψ_ν seems reasonable. The predictions are not sensitive to the choices of the values of these parameters, as shown by the small ranges of the scales of the plots on Figure 3.13. These results agree with those from model 3.5 that there is little evidence of correlation over time for the intercept parameter, but that there may be correlation over time for the slope parameter. Results from both of these models are similar, when their respective parameters ψ_μ and ψ_ν are set equal to the values that produced the most accurate predictions. With these parameter values, the models perform better than the spatiotemporal model 3.3a, which does not include smoothing over time. This provides evidence that smoothing over time may be helpful in improving predictions. However, these models do not take the patterns of *in situ* and remotely-sensed data fully into account, since they make use of correlation over time, ignoring the cyclical temporal patterns of $\log(\text{chlorophyll}_a)$, so may not produce very accurate predictions at new timepoints. These models assume that the *in situ* data and the remotely-sensed data are collected at the same time each month, ignoring the temporal change-of-support problem. This motivates further development, in order to address the different spatiotemporal support for the *in situ* and remotely-sensed data, which is discussed in Chapter 5.

3.4 Applications to the Lake Erie data

It is of interest whether the same conclusions regarding the optimal model for prediction are drawn when a different dataset is under investigation. There are *in situ* and remotely-sensed $\log(\text{chlorophyll}_a)$ data available for 20 months, for 20 locations in Lake Erie. As shown in Figure 1.2, the spa-

tial coverage of the 20 EPA locations in Lake Erie is extensive. However, the temporal coverage is poor in comparison to the data for Lake Balaton, with only 20 times for which data are available, spread over 10 years. A plot of the *in situ* data for a single location in Lake Erie and the remote sensing data for the corresponding grid cell and months is shown in Figure 3.14. The *in situ* and remote sensing data follow similar patterns over time,

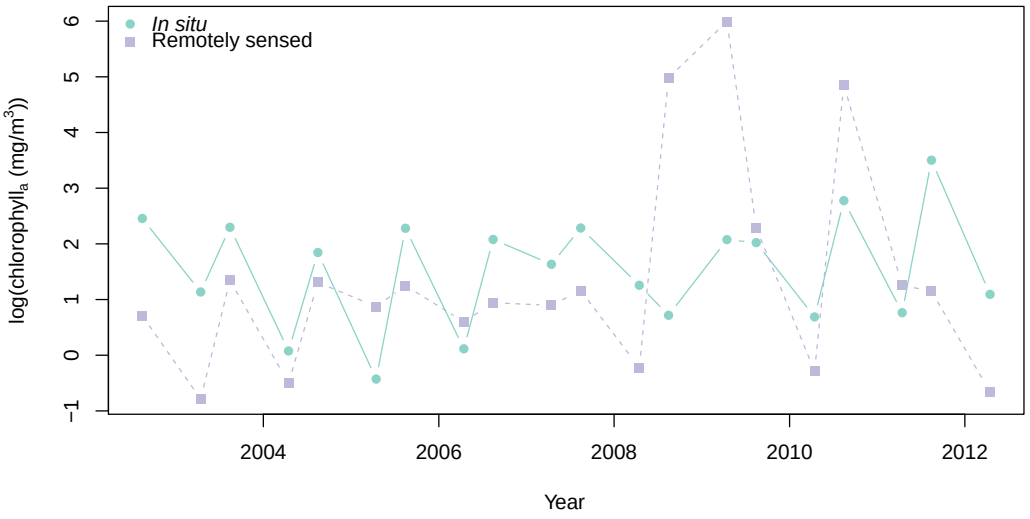

Figure 3.14: *In situ* and remotely sensed log(chlorophyll$_a$) data versus time for one location in Lake Erie.

but the poor temporal coverage of the *in situ* data makes this difficult to assess. There are two months of remotely sensed data that have particularly high values in comparison to the corresponding *in situ* data, in August 2008 and April 2009. Without associated information on the uncertainty of these values, which can be lost through the conversion process of the algorithm, it is difficult to know whether these data are reliable without a comparison to the *in situ* data. Although these data bear some resemblance to those for Lake Balaton, there are important differences. Notably, the ecological and hydrological processes within the lake could be different, since Lake Erie is a large lake, which is part of the Great Lakes ecosystem (Botts & Krushelnicki 1995), while Lake Balaton is a much shallower lake that is affected greatly by the inflow from a single source (Palmer et al. 2015). Additionally, the *in situ*

data for Lake Erie have a greater spatial extent than those for Lake Balaton, but are sampled much less frequently, with trips to all 20 locations only twice each year. Due to the improved spatial coverage, but diminished temporal coverage, the estimation of spatial patterns at each time could potentially be improved for the Lake Erie data compared to the Lake Balaton data, while patterns over time may be less well estimated.

Models are fitted to the Lake Erie data, with comparisons made through a leave-one-out cross-validation for each model, where data are removed for each of the 20 locations in turn and the model re-fitted to the remaining 19 locations. This allows a fair comparison between the models and gives an idea of how the models perform in general. As with the Lake Balaton data, the spatial decay parameters ϕ_α and ϕ_β are estimated as part of the leave-one-out cross-validation, since otherwise these parameters are difficult to estimate within the model. For each model, the assumptions and convergence are checked. Trace and density plots (see Figures B.7 and B.8 on pages 224 and 225, respectively) provide no evidence that the MCMC chains have not converged, while diagnostic plots (see Figures B.30 and B.31 on page 244) provide no information against the model assumptions that residuals are homoscedastic and mean-zero, Normally distributed.

For the spatial downscaling model 3.1 (see 78), the resulting summary statistics from the leave-one-out cross-validation are plotted against ϕ_α and ϕ_β, in Figure 3.15. These plots show that smaller values of ϕ_α and ϕ_β lead to more accurate and more precise predictions, with the minimum values of RMSE and MAE reached for $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$ and the minimum value of 95% credible interval length reached for $\phi_\alpha = 0.1$ and $\phi_\beta = 0.001$. The variance of the predictions is also smaller for smaller values of ϕ_β. The 95% credible interval coverage is close to the nominal 95% for all values of ϕ_α and ϕ_β, with slightly higher coverages for higher values of ϕ_α and ϕ_β. From this information, $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$ are chosen as near-optimal values for this dataset. These values are similar to those for the Lake

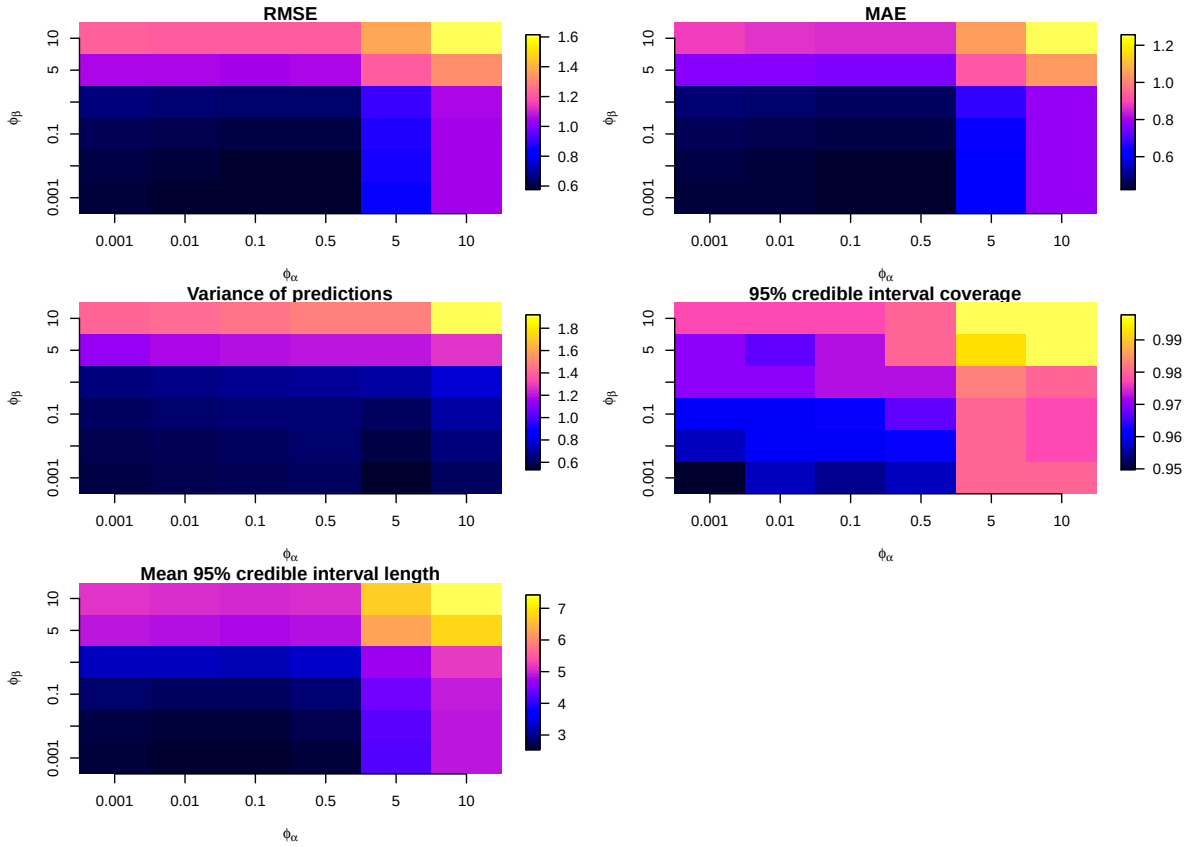

Figure 3.15: Plots of cross-validation summary statistics for model 3.1, for Lake Erie data, for each combination of $\phi_\alpha = 0.001, 0.01, 0.1, 0.5, 5, 10$ and $\phi_\beta = 0.001, 0.01, 0.1, 0.5, 5, 10$.

Balaton dataset, which are $\phi_\alpha = 0.01$ and $\phi_\beta = 0.001$. However, the larger estimated value of the spatial decay parameter ϕ_α for Lake Erie means that it is estimated that correlation between the intercept parameters decreases more quickly to zero as distance between spatial locations increases, for Lake Erie compared to Lake Balaton. As with the Lake Balaton data, a range of values of ϕ_α and ϕ_β could be selected and justified from these results, for model-fitting.

The same process is carried out for the spatiotemporal downscaling model 3.3a, with pooled variances over time (see page 101), resulting in the same near-optimal values of $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$ selected. Table 3.6 shows the resulting model summary statistics, with ϕ_α and ϕ_β set equal to 0.5 and 0.001, respectively. This table shows that the spatial and spatiotemporal

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.1	0.582	0.427	0.621	0.958	2.649
Model 3.3a	0.568	0.420	0.613	0.943	2.235

Table 3.6: Table of summary statistics for leave-one-out cross-validation for models 3.1 and 3.3a, with $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$, for Lake Erie data.

models lead to fairly similar levels of predictive accuracy, as assessed from RMSE and MAE. RMSE is, however, slightly lower for the spatiotemporal model 3.3a than for the spatial model 3.1. The variance of predictions, 95% credible interval coverage and 95% credible interval length are all very similar between both models. This suggests that sharing information over time helps in the estimation of model parameters for this dataset.

Plots of the data and predictions are shown in Figure 3.16 for model 3.3a fitted to the log(chlorophyll$_a$) data for Lake Erie. The *in situ* data for August 2007 are slightly lower than the remotely-sensed data in the northeast of the lake, but are similar to the remotely-sensed data in the centre and southwest of the lake. The predictions from model 3.3a are highest in the southwest of the lake, reflecting the pattern seen in the *in situ* and remotely-sensed data. This may be due to the main inflows into the lake, which enter in the southwest, bringing in more nutrients to this part of the lake. The northeast part of the lake is estimated to have lower levels of log(chlorophyll$_a$). This is the part of the lake closest to the main outflows, where nutrient-laden water leaves the lake. Predicted levels of $\tilde{\boldsymbol{\alpha}}$ are highest in the southwest of the lake, with areas of high levels also towards the lake centre and northeast. Predicted levels of $\tilde{\boldsymbol{\beta}}$ are highest in the southwest of the lake and lowest in the northeast of the lake, where they become negative over a small area. It is estimated that the remotely-sensed and *in situ* data have a positive relationship over much of the lake, but a weak, or inverse, relationship in the northeast of the lake.

Models 3.5 and 3.5a, which perform smoothing over time through the use of AR(1) coefficients and temporal covariance matrices, respectively (see the

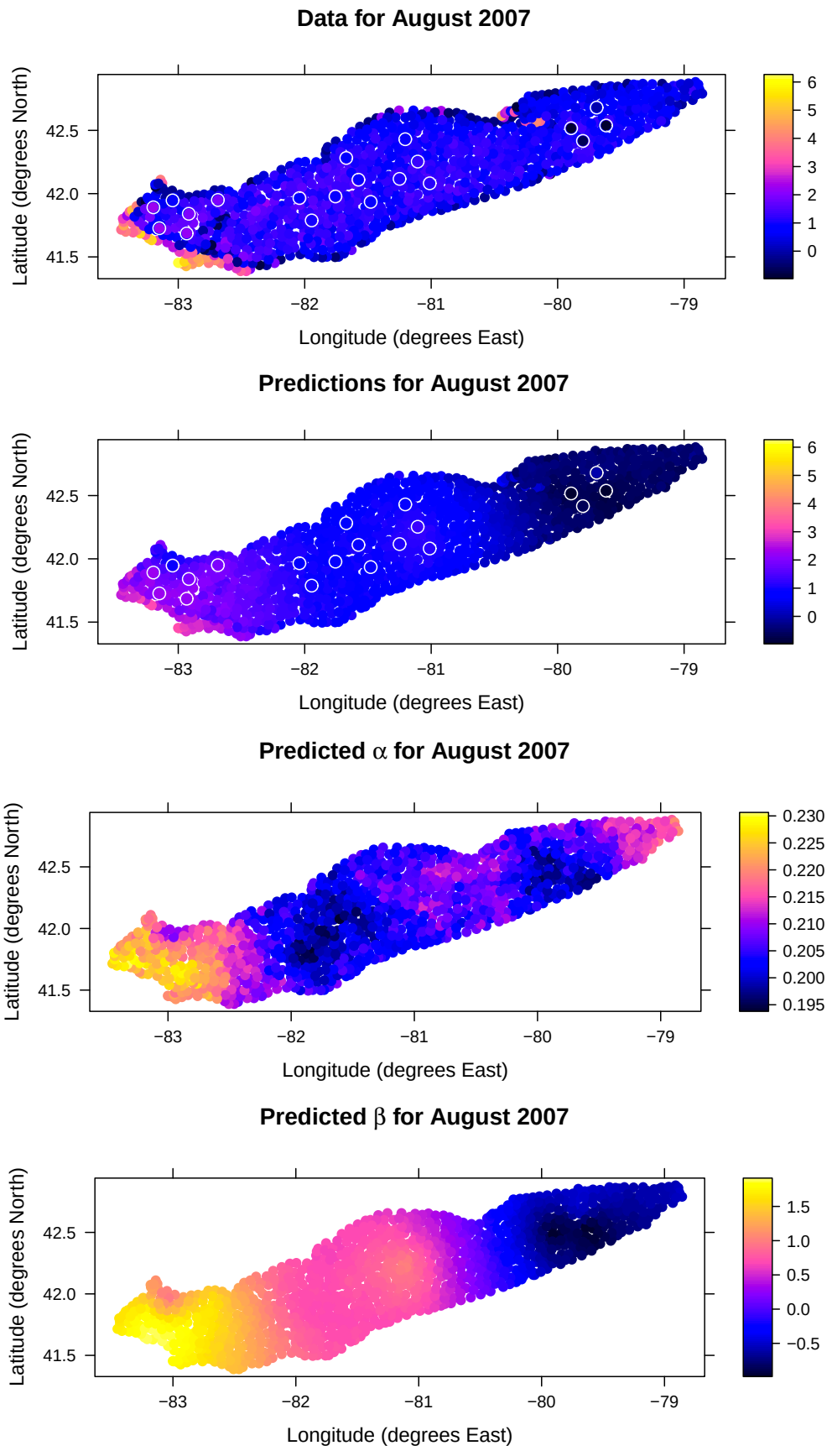

Figure 3.16: Remotely-sensed data and predicted $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ for model 3.3a fitted to the log(chlorophyll$_a$) data for Lake Erie. *In situ* data are overlaid on the top two plots, surrounded by white circles.

subsection starting on page 104), are also fitted to the Lake Erie dataset, to investigate how smooth the changes in relationship between the *in situ* and remotely-sensed data are over time and to investigate whether incorporating smoothing into the models leads to improved accuracy of predictions for these models. As with the Lake Balaton data, the values of the parameters ψ_μ and ψ_ν are chosen outwith the model-fitting process, as otherwise convergence does not occur for these parameters. As usual, the best values of these parameters are selected through a leave-one-out cross-validation, with the model refitted for several values of ψ_μ and ψ_ν. Diagnostic plots (see Figures B.32 and B.33 on page 245) suggest that the model assumptions of zero-mean Normality and homoscedasticity of residuals are valid for models 3.5 and 3.5a, while the trace and density plots provide no evidence against the assumption of convergence of MCMC chains for the model parameters (see Figures B.9 and B.10 on pages B.32 and B.33)

The values of the resulting summary statistics are displayed in Figure 3.17. For model 3.5, which performs smoothing over time using AR(1) parameters, the RMSE, MAE and mean 95% credible interval length increase with increasing ψ_μ, while the variance of predictions decreases. There is no obvious pattern of change with changing ψ_ν, for this model. Mean 95% credible interval empirical coverage lies slightly below the nominal 95% level for all investigated values of ψ_μ and ψ_ν, with values generally closer to 93% coverage. The minimum value of RMSE occurs at $\psi_\mu = 0.01$ and $\psi_\nu = 0.8$, while the minimum value of MAE occurs at $\psi_\mu = 0.01$ and $\psi_\nu = 0.2$, so that both of these sets of spatial decay parameter values are near-optimal for this model and dataset. For model 3.5a, which performs smoothing over time using temporal covariance matrices, the main pattern is that small values of ψ_ν lead to smaller RMSE, MAE and 95% credible intervals, while small values of ψ_μ lead to smaller variances of predictions, but wider 95% credible intervals. Mean 95% credible interval coverage is slightly below the nominal 95% level for all values of ψ_μ and ψ_ν, generally lying close to 0.925. The minimum

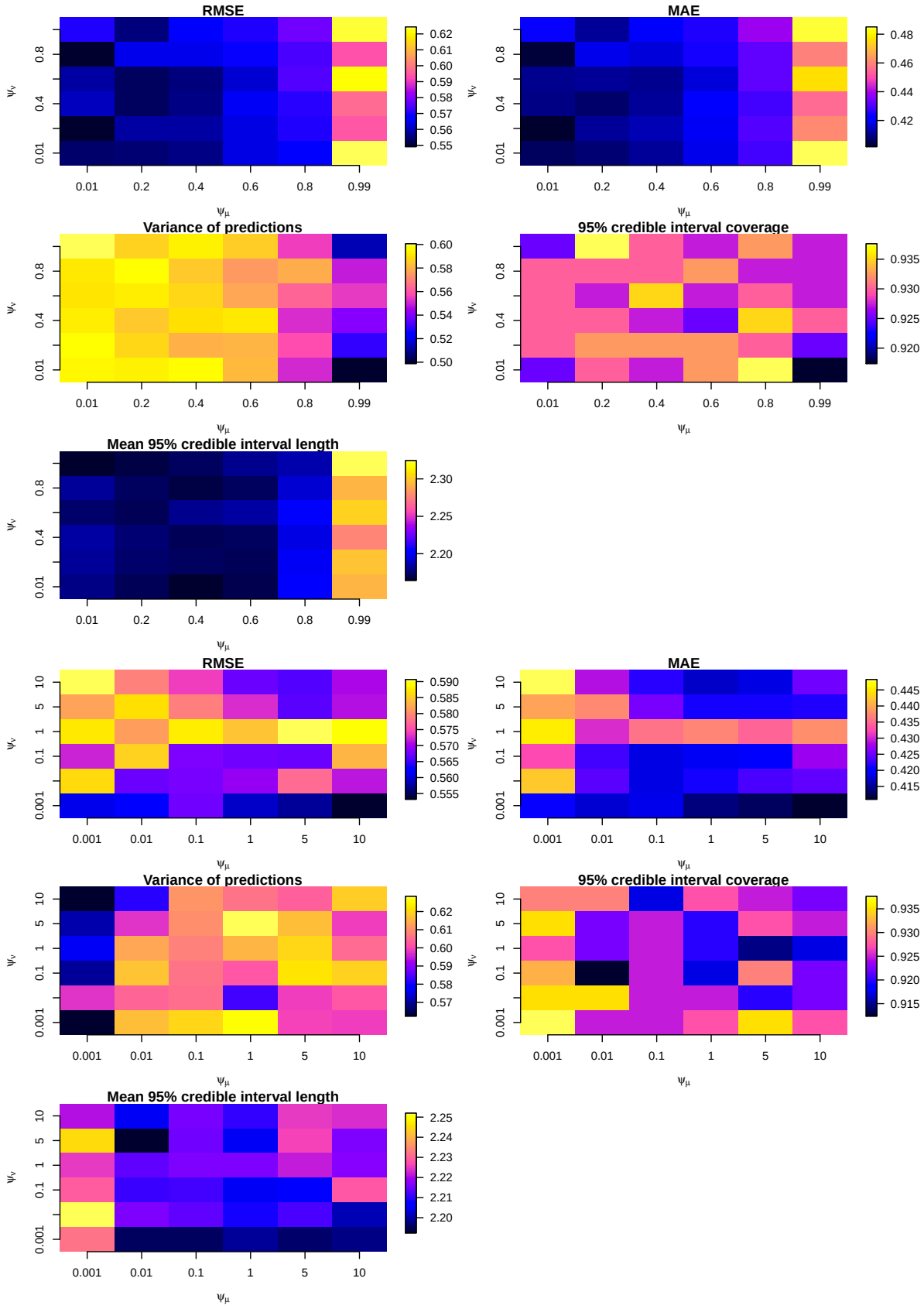

Figure 3.17: Plots of cross-validation summary statistics for models 3.5 (top) and 3.5a (bottom), for Lake Erie data, for each combination of $\phi_\mu = 0.01, 0.2, 0.4, 0.6, 0.8, 0.99$ and $\phi_\nu = 0.01, 0.2, 0.4, 0.6, 0.8, 0.99$ (top), and $\phi_\mu = 0.001, 0.01, 0.1, 1, 5, 10$ and $\phi_\nu = 0.001, 0.01, 0.1, 1, 5, 10$ (bottom).

values of RMSE and MAE are both reached at $\psi_\mu = 10$ and $\psi_\nu = 0.001$, so that these are near-optimal values of these parameters for this model and dataset. The ranges of the scales of these plots are all very small, showing that the model predictions for both models are not particularly sensitive to the parameter values being investigated. In particular, model 3.5 is not sensitive to the value of ψ_ν and model 3.5a is not sensitive to the value of ψ_μ.

Table 3.7 shows the resulting summary statistics for a leave-one-out cross-validation with parameter values chosen as the near-optimal estimates identified above, namely $\psi_\mu = 0.01$ and $\psi_\nu = 0.8$ for model 3.5 and $\psi_\mu = 10$ and $\psi_\nu = 0.001$ for model 3.5a. These summary statistics allow a comparison be-

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.5	0.550	0.403	0.592	0.903	2.183
Model 3.5a	0.553	0.411	0.599	0.928	2.198

Table 3.7: Table of summary statistics for leave-one-out cross-validation for models 3.5 and 3.5a, with $\phi_\alpha = 0.5$ and $\phi_\beta = 0.001$, and with ψ_μ and ψ_ν set equal to their chosen values (0.01 and 0.8 for model 3.5 and 10 and 0.001 for 3.5a), for the Lake Erie data.

tween these models and models 3.3 and 3.3a. Both model 3.5 and 3.5a have lower RMSE, MAE, variance of predictions and 95% credible interval length than models 3.3 and 3.3a, suggesting that incorporating correlation over time improves the accuracy and precision of estimates of $\log(\text{chlorophyll}_a)$ data for Lake Erie. However, the 95% credible interval coverage lying slightly below the nominal level suggests that the estimates are not as precise as claimed for either models 3.5 or 3.5a.

3.5 Conclusions and discussion

The aim of this chapter was to develop statistical downscaling models, motivated by the task of calibrating grid-scale remotely-sensed data for

log(chlorophyll$_a$), using point-scale *in situ* log(chlorophyll$_a$) data. Methodology was developed that enables fusion of these two data types, allowing predictions to be made at any location, along with associated measures of uncertainty. The utility of the models was demonstrated through the use of plots for an example month, allowing comparison of the remotely-sensed and *in situ* data and the resulting downscaled surface.

Spatiotemporal developments of the models allow both the fitting of models to data for various times at once and also the sharing of information across times, leading to improved estimates of model parameters and to improved accuracy of the resulting calibrated log(chlorophyll$_a$) surfaces.

Although the spatial model 3.1 and the spatiotemporal models 3.3, 3.3a and 3.4 are able to spatially calibrate data, allowing prediction to be made at any location within the area covered by the remotely-sensed data, they are unable to predict at times outwith those in the dataset. Additionally, these models assume that the support for the *in situ* data and the remotely-sensed data is the same, ignoring the fact that the *in situ* data have a point-time scale, while the remotely-sensed data are monthly-averaged. These issues are addressed in Chapter 5.

Chapter 4

Bivariate and multiple lakes downscaling

The previous chapter introduced developments and applications of statistical downscaling models, for a single lake, e.g. Lake Balaton or Lake Erie, and for a single variable, e.g. log(chlorophyll$_a$). It was demonstrated that sharing information, specifically sharing over time, helped to improve the accuracy of model predictions. Taking this into account, this chapter investigates whether sharing information between related variables and between neighbouring lakes helps to improve the accuracy of predictions.

This chapter presents developments in bivariate statistical downscaling, where two variables are downscaled at once in order to share information between them. A second development, multiple lakes downscaling, is also presented. This is where data are calibrated over multiple lakes at once, enabling the sharing of information between lakes, while allowing for differences in data patterns between lakes. The background and motivation are presented, followed by model development and an application to data. Conclusions and discussion of the methodology are then presented.

4.1 Bivariate statistical downscaling

Bivariate statistical downscaling is a method for data fusion of two related variables simultaneously, with the aim of producing improved calibration in comparison to downscaling each variable separately. The aim of the methodology is to accomplish the fusion through sharing information on the relationship between the *in situ* and remotely-sensed data between the two variables being modelled. This methodology makes the assumption that there is a similar relationship between the *in situ* and remotely-sensed data for the two variables selected, so these variables must be chosen carefully, based on both expert ecological knowledge and empirical evidence, such as exploratory plots of the data. Berrocal et al. (2010*a*) develop a bivariate version of the univariate statistical downscaling model that was introduced in Berrocal et al. (2010*b*), where they model the *in situ* data for each of two related variables on the remotely-sensed data for the variable of interest and the related variable, with the spatially-varying parameters modelled through coregionalisation. The approach taken in this section will instead build within the framework of models developed in the previous chapter. Additionally, the aim here is to calibrate the remotely-sensed data for each variable using as much information as possible from the *in situ* data for both variables, which contrasts with the approach in the paper, where remotely-sensed data for multiple variables are calibrated using *in situ* data for multiple variables. The approach described here introduces correlation between variables, through correlated errors.

4.1.1 Motivational exploratory analysis

In order to motivate the development of bivariate downscaling methodology, plots of data for Lake Balaton are produced. The variables available for both the *in situ* and remotely-sensed data are log(chlorophyll$_a$), log(total suspended matter) and water temperature. Scatterplots of data for these

variables (see Figures 4.1 and 4.2) show that the strongest relationship by

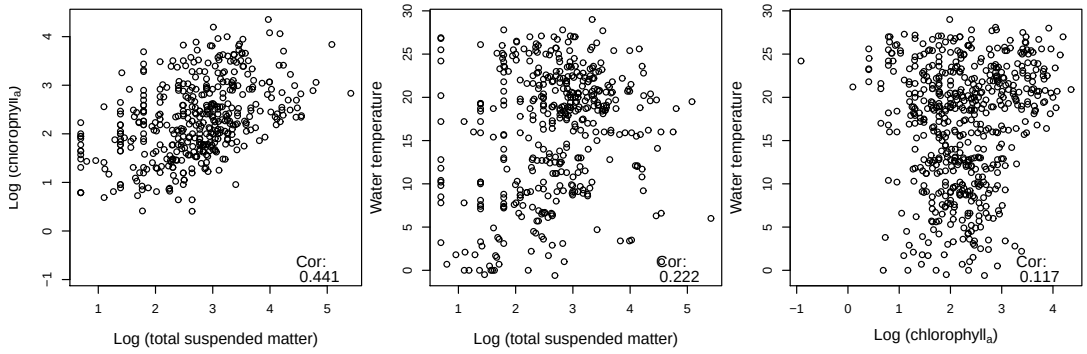

Figure 4.1: Plots showing relationships between *in situ* log(chlorophyll$_a$) (mg/m^3), log(total suspended matter) (g/m^3) and lake surface water temperature (°C), with correlations shown in lower-right corners.

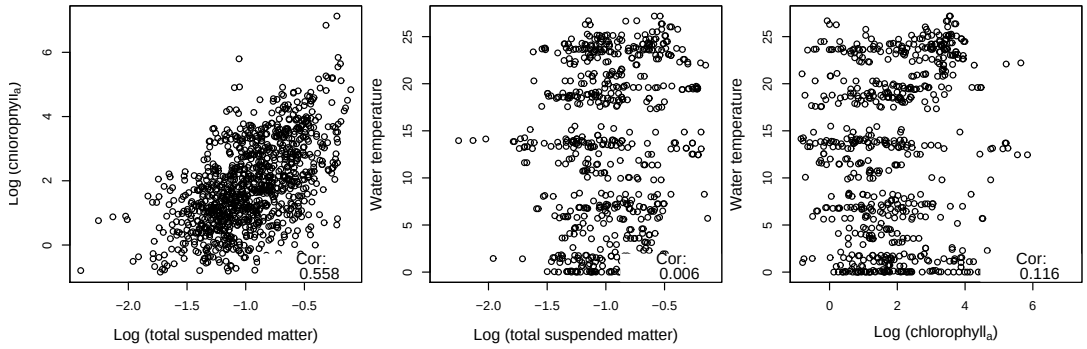

Figure 4.2: Plots showing relationships between remotely-sensed log(chlorophyll$_a$) (mg/m^3), log(total suspended matter) (g/m^3) and lake surface water temperature (°C), with correlations shown in lower-right corners.

far is between log(chlorophyll$_a$) and log(total suspended matter), for both data types, with correlations of 0.441 and 0.558 for the *in situ* and remotely-sensed data, respectively. This may be explained by the fact that changes in these two variables may be caused by similar environmental or catchment drivers. Water temperature has a fairly smooth change over time, without the same local effects that log(chlorophyll$_a$) and log(total suspended matter) respond to, explaining the weaker relationship between water temperature and the other variables, with weak correlations displayed on the plots. This section will therefore focus on modelling log(chlorophyll$_a$) and log(total sus-

pended matter) together.

Model 3.1 is firstly fitted separately to the data for log(chlorophyll$_a$) and log(total suspended matter), for Lake Balaton. The dataset for Lake Balaton only contains 5 locations for which both log(chlorophyll$_a$) and log(total suspended matter) *in situ* data are available for the same set of times. These data are available for 18 months at all 5 of these locations, for each of the two variables. Therefore, this section of analysis works with an 18×5 dataset for each variable, which has fewer locations than the 17×9 dataset that was analysed in the previous chapter, but still contains sufficient information on the spatial patterns in the data in order to fit statistical downscaling models. The residuals from model 3.1 fitted to data for the two variables can be investigated. Figure 4.3 shows that there is not a strong relationship between

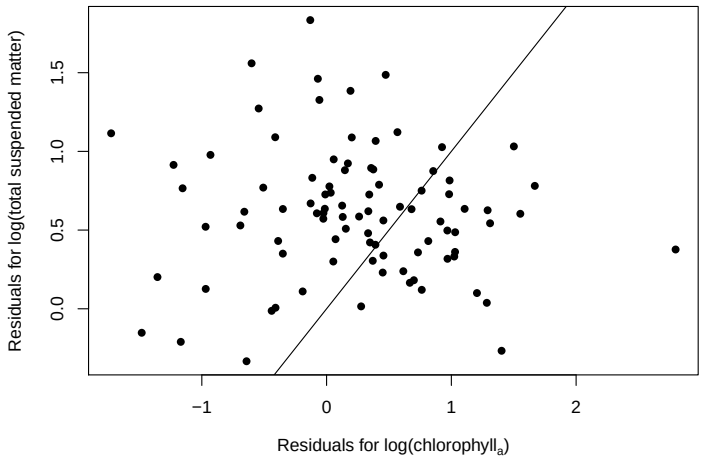

Figure 4.3: Plot of residuals for model 3.1 fitted to the 18×5 log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton.

the residuals for the two variables, with correlation estimated to be -0.06, i.e. very close to zero.

Given the small number of available data, it is of interest to investigate whether sharing information between variables is useful. Since sharing information over time improved estimates of σ_α^2 and σ_β^2 in model 4.1a, it makes sense to investigate whether these parameters can be estimated more accurately by sharing information between variables in a bivariate modelling

framework.

4.1.2 Spatial bivariate downscaling model

This subsection introduces the bivariate downscaling model, in its spatial form. The following subsection describes an application to the data for Lake Balaton.

A spatial bivariate statistical downscaling model is:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} \alpha_{1i} + \beta_{1i} x_{1i} \\ \alpha_{2i} + \beta_{2i} x_{2i} \end{pmatrix}, \boldsymbol{\Sigma}_\varepsilon \right), \tag{4.1}$$

where:

$$\boldsymbol{\Sigma}_\varepsilon = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

is the variance-covariance matrix for errors, modelling the correlation between errors for both variables. Prior distributions for this model are:

$$\begin{aligned} \boldsymbol{\alpha}_1 &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_{\alpha_1}^2 \exp(-\phi_{\alpha_1}\mathbf{D})\big), \\ \boldsymbol{\alpha}_2 &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_{\alpha_2}^2 \exp(-\phi_{\alpha_2}\mathbf{D})\big), \\ \boldsymbol{\beta}_1 &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_{\beta_1}^2 \exp(-\phi_{\beta_1}\mathbf{D})\big), \\ \boldsymbol{\beta}_2 &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_{\beta_2}^2 \exp(-\phi_{\beta_2}\mathbf{D})\big), \\ (\sigma_{\alpha_1}^2)^{-1} &\sim \mathrm{Ga}(a_{\alpha_1}, b_{\alpha_1}), \\ (\sigma_{\alpha_2}^2)^{-1} &\sim \mathrm{Ga}(a_{\alpha_2}, b_{\alpha_2}), \\ (\sigma_{\beta_1}^2)^{-1} &\sim \mathrm{Ga}(a_{\beta_1}, b_{\beta_1}), \\ (\sigma_{\beta_2}^2)^{-1} &\sim \mathrm{Ga}(a_{\beta_2}, b_{\beta_2}) \text{ and} \\ \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} &\sim \text{Inv-W}(\mathbf{A}_\varepsilon, b_\varepsilon), \end{aligned}$$

where $\boldsymbol{\alpha}_1 = (\alpha_{1,1}, \ldots, \alpha_{1,n})^{\mathrm{T}}$, $\boldsymbol{\alpha}_2 = (\alpha_{2,1}, \ldots, \alpha_{2,n})^{\mathrm{T}}$, $\boldsymbol{\beta}_1 = (\beta_{1,1}, \ldots, \beta_{1,n})^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (\beta_{2,1}, \ldots, \beta_{2,n})^{\mathrm{T}}$, and where a_{α_1}, b_{α_1}, a_{α_2}, b_{α_2}, a_{β_1}, b_{β_1}, a_{β_2}, b_{β_2},

$\mathbf{A}_\varepsilon$ and b_ε are parameters that must be chosen *a priori*. For non-informative prior distributions, each a and b are set to small values, such as 0.001, while $\mathbf{A}_\varepsilon$ is set equal to the identity matrix of dimension 2, and b_ε is set equal to 3 (Gelman et al. 2014).

A bivariate or multivariate relationship can also be modelled using the linear model of coregionalisation (LMC), as carried out by Berrocal et al. (2010*a*) and in the spBayes R package (Finley et al. 2007, 2013). The LMC models the correlation between variables through correlated intercept and slope parameters, rather than through correlated errors.

An adjustment to model 4.1 is to have common prior distributions for the spatial variances for both variables, giving an alternative set of prior distributions:

$$
\begin{aligned}
\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} &\sim \mathrm{N}_2\left(\begin{pmatrix} \alpha_{1i} + \beta_{1i}x_{1i} \\ \alpha_{2i} + \beta_{2i}x_{2i} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), \\
\boldsymbol{\alpha}_1 &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big), \\
\boldsymbol{\alpha}_2 &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big), \\
\boldsymbol{\beta}_1 &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big), \\
\boldsymbol{\beta}_2 &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big), \\
(\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\
(\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta), \\
\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} &\sim \text{Inv-W}(\mathbf{A}_\varepsilon, b_\varepsilon).
\end{aligned}
\tag{4.1a}
$$

These models fit regressions on the *in situ* data, with the corresponding remotely-sensed data as the only explanatory variable. This means that the regressions for the two variables are only related through correlated errors. The full conditional distributions for the parameters of model 4.1a are derived in the appendix (see section A.3 on page 204).

In order to assess the performance of models 4.1 and 4.1a and compare to the performance of the univariate model (i.e. the spatial statistical down-

scaling model 3.1, presented on page 78), these models are fitted to data for Lake Balaton, for the 18 months of log(chlorophyll$_a$) and log(total suspended matter) data for which data for 5 *in situ* locations were available. The model performances are compared through a leave-one-out cross-validation, where data corresponding to each of the 5 *in situ* locations are removed in turn and predicted from each model fitted to data for the remaining 4 locations. The convergence of MCMC chains in each model is checked using trace and density plots (see Figures B.11 and B.12 on pages 228 and 229). The model assumptions are checked using diagnostic plots (see Figures B.34 and B.35 on page 246). Summary statistics are then calculated to compare the predictions and observed *in situ* data. These resulting summary statistics are displayed in Table 4.1. These results show that model 4.1 (i.e. the bivariate

	Model	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Chl	3.1	0.809	0.632	1.031	0.967	6.162
	4.1	1.500	1.416	0.169	1.000	10.503
	4.1a	0.635	0.532	0.189	1.000	7.735
TSM	3.1	0.733	0.629	0.181	1.000	16.115
	4.1	1.574	1.483	0.196	1.000	12.367
	4.1a	0.644	0.527	0.206	1.000	8.214

Table 4.1: Table of summary statistics of leave-one-out cross-validations for models 3.1 (i.e. the univariate spatial model), 4.1 (i.e. the bivariate spatial model) and 4.1a (i.e. the bivariate spatial model with pooled spatial variances across variables), fitted to Lake Balaton log(chlorophyll$_a$) and log(total suspended matter) data.

spatial model without pooled spatial variance parameters across variables) performs poorly compared to the univariate model 3.1, when fitted to the 18 by 5 dataset for Lake Balaton. This is perhaps due to the small number of available data for each month causing difficulty in estimating all of the parameters well. On the other hand, model 4.1a, which is a bivariate spatial model with correlated errors and pooled spatial variance parameters across variables, performs much better, with much lower RMSE and MAE in comparison to the univariate model and the bivariate model without pooling

across variables (i.e. model 4.1). This is true for both the log(chlorophyll$_a$) data and the log(total suspended matter) data. The variance of predictions and mean 95% credible interval coverage are both fairly similar for all three models under consideration. The main conclusions from this analysis are that the bivariate spatial model 4.1 does not produce improved levels of prediction compared to the univariate spatial model 3.1 and in fact performs badly for the dataset of interest. Model 4.1a, which is a bivariate spatial model that pools estimates for spatial variance parameters across parameters, performs much better than the other models investigated and so can be preferred for this dataset.

Predictions are made from model 4.1a, using data for March 2011 as an example, for 997 locations as determined by a Delaunay triangulation (see Figure 3.1 and the related description on page 84). These predictions are shown, along with the original remotely-sensed and *in situ* data, on Figures 4.4 and 4.5. These plots show that there is some difference between the *in situ* and surrounding remotely-sensed log(chlorophyll$_a$) data for all 5 locations in the lake, but that this difference is no longer evident in the resulting calibrated surface of predictions from model 4.1a. For the log(total suspended matter) data, there are much larger differences between each *in situ* data point and its surrounding remotely-sensed data. However, the resulting calibrated surface from model 4.1a lies much closer to the observed values of *in situ* log(total suspended matter) data. These plots show that calibration has been achieved for both log(chlorophyll$_a$) and log(total suspended matter), with the resulting predictions for each variable over the lake retaining the spatial patterns from the remotely-sensed data for that variable, but pulled towards the values of the accurate *in situ* data.

The estimated values of ρ for model 4.1a, when fitted to each of the 18 months of data for Lake Balaton separately, lie between 0.006 and 0.06, with a median value around 0.028. This means that the model is fitting very little correlation between the errors for the two variables, which agrees with the

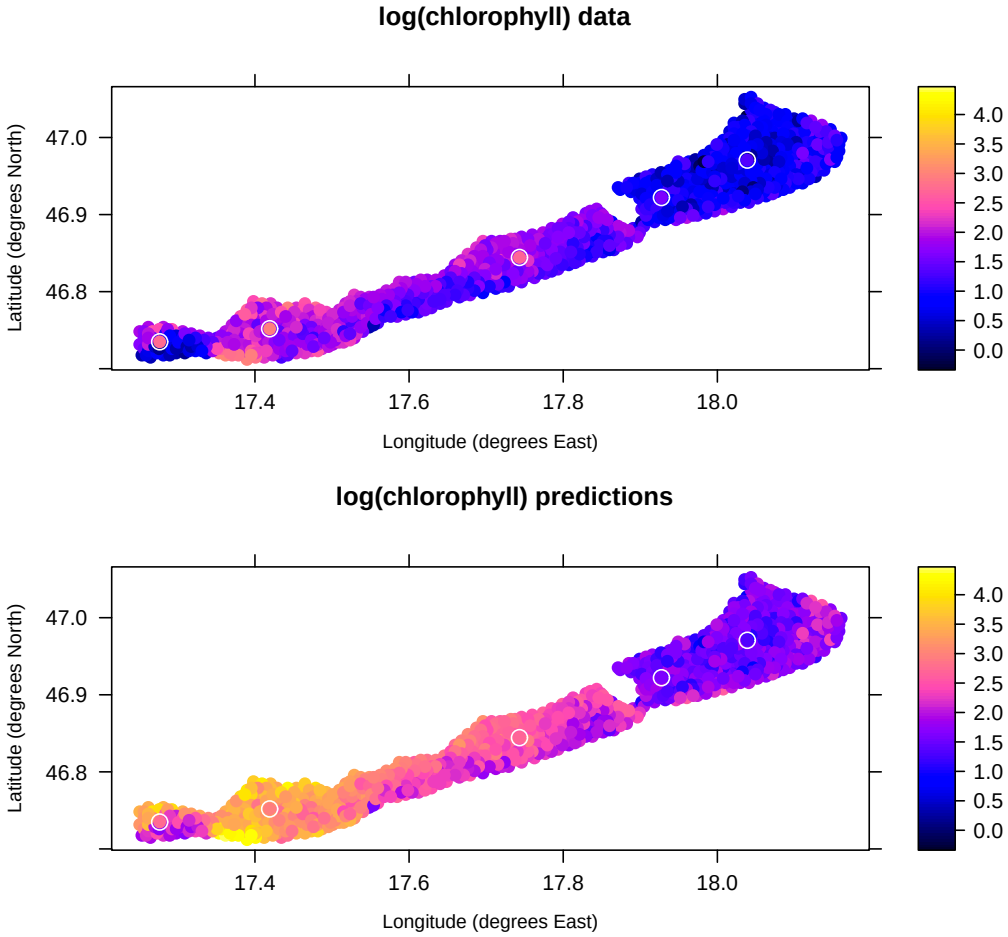

Figure 4.4: Plots of remotely-sensed log(chlorophyll$_a$) data for March 2011 and resulting predictions from model 4.1a, with *in situ* data overlaid and circled in white.

results from the exploratory fitting of the univariate spatial model 3.1, which suggested that there was little correlation between the model residuals for log(chlorophyll$_a$) and log(total suspended matter), when fitted separately to data for each variable.

4.1.3 Spatiotemporal bivariate downscaling model

In order to share information over time, a spatiotemporal extension to the bivariate model is developed, based upon model 4.1. The model has the additional subscript j, which represents time ($j = 1, \ldots, t$). The spatial variance parameters, $\sigma^2_{\alpha_1}$, $\sigma^2_{\alpha_2}$, $\sigma^2_{\beta_1}$ and $\sigma^2_{\beta_2}$, and the error matrix, $\mathbf{\Sigma}_\varepsilon$, are

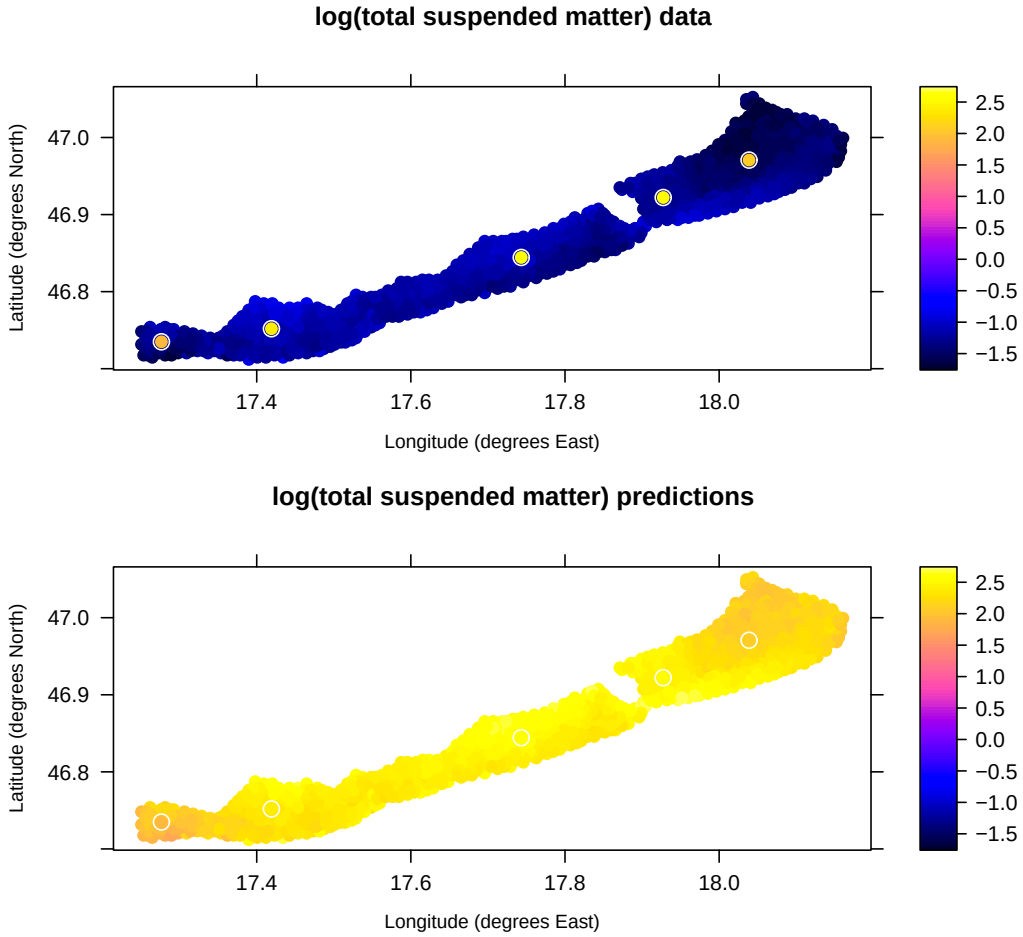

Figure 4.5: Plots of remotely-sensed log(total suspended matter) data for March 2011 and resulting predictions from model 4.1a, with *in situ* data overlaid and circled in white.

estimated using data for all timepoints. The model is:

$$\begin{pmatrix} y_{1ji} \\ y_{2ji} \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} \alpha_{1ji} + \beta_{1ji} x_{1ji} \\ \alpha_{2ji} + \beta_{2ji} x_{2ji} \end{pmatrix}, \boldsymbol{\Sigma}_\varepsilon \right), \tag{4.2}$$

where y_{1ji} is the *in situ* data for log(chlorophyll$_a$) at time j ($j = 1, \ldots, t$) for *in situ* location i ($i = 1, \ldots, n$), y_{2ji}, x_{1ji} and x_{2ji} are the corresponding values of data for *in situ* log(total suspended matter), remotely-sensed log(chlorophyll$_a$) and remotely-sensed log(total suspended matter), respec-

tively, and:

$$\boldsymbol{\Sigma}_{\varepsilon} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

is the variance-covariance matrix for errors, modelling the correlation between errors for both variables. Prior distributions for this model are:

$$\begin{aligned}
\boldsymbol{\alpha}_{1j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\alpha_1}\exp(-\phi_{\alpha_1}\mathbf{D})\big), \\
\boldsymbol{\alpha}_{2j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\alpha_2}\exp(-\phi_{\alpha_2}\mathbf{D})\big), \\
\boldsymbol{\beta}_{1j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\beta_1}\exp(-\phi_{\beta_1}\mathbf{D})\big), \\
\boldsymbol{\beta}_{2j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\beta_2}\exp(-\phi_{\beta_2}\mathbf{D})\big), \\
(\sigma^2_{\alpha_1})^{-1} &\sim \mathrm{Ga}(a_{\alpha_1}, b_{\alpha_1}), \\
(\sigma^2_{\alpha_2})^{-1} &\sim \mathrm{Ga}(a_{\alpha_2}, b_{\alpha_2}), \\
(\sigma^2_{\beta_1})^{-1} &\sim \mathrm{Ga}(a_{\beta_1}, b_{\beta_1}), \\
(\sigma^2_{\beta_2})^{-1} &\sim \mathrm{Ga}(a_{\beta_2}, b_{\beta_2}) \text{ and} \\
\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} &\sim \text{Inv-W}(\mathbf{A}_{\varepsilon}, b_{\varepsilon}).
\end{aligned}$$

A revised model (4.2a), based upon model 4.1a, is created, with $\sigma^2_{\alpha_1}$ and $\sigma^2_{\alpha_2}$ replaced by the common parameter σ^2_{α}, and $\sigma^2_{\beta_1}$ and $\sigma^2_{\beta_2}$ replaced by σ^2_{β}. Similarly, ϕ_{α_1} and ϕ_{α_2} are replaced by ϕ_{α}, while ϕ_{β_1} and ϕ_{β_2} are replaced by ϕ_{β}, resulting in further sharing of information between variables:

$$\begin{aligned}
\begin{pmatrix} y_{1ji} \\ y_{2ji} \end{pmatrix} &\sim \mathrm{N}_2\left(\begin{pmatrix} \alpha_{1ji} + \beta_{1ji}x_{1ji} \\ \alpha_{2ji} + \beta_{2ji}x_{2ji} \end{pmatrix}, \boldsymbol{\Sigma}_{\varepsilon}\right), \\
\boldsymbol{\alpha}_{1j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\alpha}\exp(-\phi_{\alpha}\mathbf{D})\big), \\
\boldsymbol{\alpha}_{2j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\alpha}\exp(-\phi_{\alpha}\mathbf{D})\big), \\
\boldsymbol{\beta}_{1j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\beta}\exp(-\phi_{\beta}\mathbf{D})\big), \\
\boldsymbol{\beta}_{2j} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\beta}\exp(-\phi_{\beta}\mathbf{D})\big), \\
(\sigma^2_{\alpha_1})^{-1} &\sim \mathrm{Ga}(a_{\alpha_1}, b_{\alpha_1}),
\end{aligned} \tag{4.2a}$$

$$(\sigma^2_{\alpha_2})^{-1} \sim \text{Ga}(a_{\alpha_2}, b_{\alpha_2}),$$
$$(\sigma^2_{\beta_1})^{-1} \sim \text{Ga}(a_{\beta_1}, b_{\beta_1}),$$
$$(\sigma^2_{\beta_2})^{-1} \sim \text{Ga}(a_{\beta_2}, b_{\beta_2}) \text{ and}$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \sim \text{Inv-W}(\mathbf{A}_{\varepsilon}, b_{\varepsilon}).$$

4.1.4 Application of spatiotemporal bivariate downscaling model to data for Lake Balaton

Models 4.2 and 4.2a are fitted to the Lake Balaton data for the 18 months for which *in situ* log(chlorophyll$_a$) data and log(total suspended matter) data are both available for 5 locations. In order to compare the performance of the bivariate models 4.2 and 4.2a with that of univariate model 3.3a (i.e. the spatiotemporal statistical downscaling model with pooled spatial variances over time, described on page 101), a leave-one-out cross-validation is carried out, where data for each of the 5 *in situ* locations can be removed in turn and predicted. The convergence of MCMC chains was checked from trace and density plots for each model (see Figures B.13 and B.12 on pages 230 and 229), while diagnostic plots (see Figures B.36 and B.37 on pages 247 and 247) provide no evidence against the validity of the model assumptions that residuals have mean zero, are homoscedastic and are Normally distributed. The summary statistics calculated from the resulting predictions are given in Table 4.2. This table shows that there are very little differences between

	RMSE	MAE	Variance of predictions	95% credible interval coverage	Mean 95% credible interval length
Model 3.3a	0.500	0.338	2.668	0.965	2.116
Model 4.2	0.501	0.340	2.669	0.965	2.428
Model 4.2a	0.474	0.322	2.689	0.959	2.286

Table 4.2: Table of summary statistics for leave-one-out cross-validation for models 3.3a (i.e. the univariate spatiotemporal model described on page 101), 4.2 and 4.2a. (i.e. the bivariate spatiotemporal models without and with pooled estimates of spatial variance parameters between variables, respectively.)

the univariate model 3.3a and the bivariate models 4.2 and 4.2a in terms

of their accuracy and precision at predicting. This can be explained, if the estimated correlation between the errors for the two variables is examined. Since $\mathbf{\Sigma}_\varepsilon = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$ (see, for example, model 4.2), then $\rho = \frac{\Sigma_{\varepsilon,1,2}}{\sqrt{\Sigma_{\varepsilon,1,1}\Sigma_{\varepsilon,2,2}}}$, which has a posterior median estimate of 0.245 for model 4.2 and 0.215 for model 4.2a. This shows that there is very little estimated correlation between the residuals for the two variables, so it appears that not much is gained from fitting the bivariate model to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. It is of note, however, that the predictions for model 4.2a are more accurate than those for model 4.2, as measured by RMSE and MAE, showing that sharing information between the variables through common parameters σ_α^2 and σ_β^2 can improve the accuracy of the predictions.

4.2 Multiple-lakes statistical downscaling

The concept of multiple-lakes statistical downscaling is motivated by the data available for the Great Lakes of North America. There are both *in situ* and remotely-sensed data available for Lakes Superior, Michigan, Huron and Erie. The *in situ* data are available from the Great Lakes Monitoring website (greatlakesmonitoring.org) and the remotely-sensed data are available from the Diversity II project (www.diversity2.info). Since these lakes are located nearby in space and are interconnected, they share some features of their ecology. On the other hand, each lake has a different hydrological status, due both to the shapes of the lakes and increasing nutrient levels as water moves downstream from Lake Superior. This can be accounted for, by including lake-specific parameters within any statistical downscaling model.

In this section, multiple-lakes statistical downscaling models will be developed, discussed and applied to the Great Lakes. These models will be compared and contrasted, to understand whether these models are really beneficial and, if so, in which circumstances.

4.2.1 Model development

The spatial-only single lake model was described in the previous chapter. This model could be applied to data for each lake individually, or alternatively to data for all lakes at once, effectively treating all Great Lakes as part of the same larger lake. Both of these possibilities are investigated in the application subsection, where these models are applied to the data.

Full multiple-lakes statistical downscaling models

The multiple lakes downscaling model takes into account the overall patterns in the intercept and slope coefficients of the relationship between the *in situ* and remotely-sensed log(chlorophyll$_a$) data, while also allowing for lake-specific adjustments and spatially-varying adjustments. A first development of a multiple-lakes model is:

$$y_{j(i)} \sim \mathrm{N}(\alpha + \beta_i + \gamma_{j(i)} + (\delta + \eta_i + \kappa_{j(i)}) \odot x_{j(i)}, \sigma^2_\epsilon), \tag{4.3}$$

for $i = 1, \ldots, l$ and $j = 1, \ldots, n_i$, where l is the number of lakes and n_i is the number of *in situ* sampling locations within lake i, where $y_{j(i)}$ is the value of the *in situ* data point at location j within lake i, $x_{j(i)}$ is the corresponding remotely-sensed data value, α and δ are the overall intercept and slope parameter, respectively, β_i and η_i are the lake-specific intercept and slope parameters, modelled as adjustments to the overall parameters, and $\gamma_{j(i)}$ and $\kappa_{j(i)}$ are the spatially-varying parameters, which vary smoothly over space, within each lake. Prior distributions are as follows:

$$\begin{aligned}
(\sigma^2_\varepsilon)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \\
\alpha &\sim \mathrm{N}(0, \sigma^2_\alpha), \\
\beta_i &\sim \mathrm{N}(0, \sigma^2_\beta) \text{ for } i = 1, \ldots, l-1, \\
\beta_l &= -\sum_{i=1}^{n-1} \beta_i,
\end{aligned}$$

$$\begin{aligned}
\delta &\sim \mathrm{N}(1, \sigma_\delta^2), \\
\eta_i &\sim \mathrm{N}(0, \sigma_\eta^2) \text{ for } i = 1, \ldots, l-1, \text{ and} \\
\eta_l &= -\sum_{i=1}^{n-1} \eta_i,
\end{aligned}$$

with the sum-to-zero constraints for $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ ensuring that $\sum_{i=1}^{l} \beta_i = 0$ and $\sum_{i=1}^{l} \eta_i = 0$, so that these parameters behave as lake-specific adjustments to the overall intercept and slope parameters. The spatially-varying parameters are given different prior distributions, depending upon whether these parameters are assumed to vary smoothly within each lake separately or over all lakes. Assuming a separate spatially-varying intercept and slope for each lake, their prior distributions are:

$$\begin{aligned}
\boldsymbol{\gamma}_i &\sim \mathrm{N}_{n_i}\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D}_i)\big) \text{ and} \\
\boldsymbol{\kappa}_i &\sim \mathrm{N}_{n_i}\big(\mathbf{0}, \sigma_\kappa^2 \exp(-\phi_\kappa \mathbf{D}_i)\big),
\end{aligned} \tag{4.3a}$$

for $i = 1, \ldots, l$, where $\boldsymbol{\gamma}_i = (\gamma_{1(i)}, \ldots, \gamma_{n_i(i)})^{\mathrm{T}}$, $\boldsymbol{\kappa}_i = (\kappa_{1(i)}, \ldots, \kappa_{n_i(i)})^{\mathrm{T}}$ and $\mathbf{D}_i$ is the $n_i \times n_i$ matrix of distances between locations within lake i, i.e.

$$\mathbf{D}_i = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n_i} \\ \vdots & & \vdots \\ d_{n_i,1} & \cdots & d_{n_i,n_i} \end{pmatrix},$$

where $d_{i,j}$ is the distance between locations i and j. Alternatively, assuming that the spatially-varying intercept and slope parameters vary smoothly across all lakes, the prior distributions of the spatially-varying parameters are:

$$\begin{aligned}
\boldsymbol{\gamma} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D})\big) \text{ and} \\
\boldsymbol{\kappa} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\kappa^2 \exp(-\phi_\kappa \mathbf{D})\big),
\end{aligned} \tag{4.3b}$$

where $n = \sum_{i=1}^{l} n_i$ is the total number of *in situ* locations within all lakes, $\mathbf{D}$ is the matrix of distances between all n locations, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^{\mathrm{T}} =$

$(\gamma_{1(1)}, \ldots, \gamma_{n_1(1)}, \gamma_{1(2)}, \ldots\ldots, \gamma_{n_l(l)})^{\mathrm{T}}$ and $\boldsymbol{\kappa} = (\kappa_{1(1)}, \ldots, \kappa_{n_1(1)}, \kappa_{1(2)}, \ldots\ldots,$ $\kappa_{n_l(l)})^{\mathrm{T}} = (\kappa_1, \ldots, \kappa_n)^{\mathrm{T}}$. The remaining hyperprior distributions are specified as:

$$
\begin{aligned}
(\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\
(\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta), \\
(\sigma_\gamma^2)^{-1} &\sim \mathrm{Ga}(a_\gamma, b_\gamma), \\
(\sigma_\delta^2)^{-1} &\sim \mathrm{Ga}(a_\delta, b_\delta), \\
(\sigma_\eta^2)^{-1} &\sim \mathrm{Ga}(a_\eta, b_\eta) \text{ and} \\
(\sigma_\kappa^2)^{-1} &\sim \mathrm{Ga}(a_\kappa, b_\kappa).
\end{aligned}
$$

Each of the a and b parameters are chosen *a priori*, with small values, such as 0.001 and 0.001, leading to noninformative prior distributions (Lunn et al. 2013). As mentioned in Chapter 3, it should be noted that Gelman et al. (2014) state that posterior distributions may in fact be sensitive to the choice of values for a and b, while Sahu et al. (2006) and Sahu et al. (2010) suggest choosing $a = 2$ and $b = 1$ instead, giving a distribution with mean 1 and infinite variance. This chapter continues to use the values $a = 0.001$ and $b = 0.001$, to allow comparison between the models fitted in earlier chapters and since there was no evidence that model 3.1 was particularly sensitive to the choice of a and b.

Treating model 4.3 as two separate models, 4.3a and 4.3b, named after the respective equations for the prior distributions of their spatially-varying parameters given above, model 4.3b has the advantage over model 4.3a of potentially being able to estimate the spatial variance parameters more accurately, using the larger number of data available over all lakes in comparison to using data for each lake separately. However, a potential problem with model 4.3b is that lake-specific differences in the relationships between the *in situ* and remotely-sensed data are only accounted for by the lake-specific parameters.

Convergence issues Initial model runs confirmed the need for the sum-to-zero constraint for the lake-specific parameters, as otherwise these could not easily be separated from the estimates of the overall intercept and slope parameters. Poor convergence was also discovered in the estimates of the spatially-varying slope parameters and several variance parameters. Further work could investigate whether different prior distributions for the variance parameters may be more appropriate and possibly lead to improved convergence. Several authors note the problem of the lack of identifiability of spatially-varying slopes in statistical downscaling models and conclude that the spatially-varying slope parameter can be replaced with a constant slope, while retaining the spatially-varying intercept parameter (Fuentes & Raftery 2005, Berrocal et al. 2010*b*, 2012, 2014, Rundel et al. 2015). Replacing a spatially-varying slope with a constant slope is justified by Fuentes & Raftery (2005) on the basis that their preliminary investigation found that the bias of their numerical model output was mostly additive, with little multiplicative bias. Motivated by this, Berrocal et al. (2010*b*) fit several models with a spatially-varying multiplicative bias and several with a constant multiplicative bias, and find that models with constant multiplicative bias perform slightly better for their data.

Resulting reduced models

After removing the spatially-varying slope parameter, $\boldsymbol{\kappa}$, from model 4.3, the resulting reduced model is:

$$
\begin{aligned}
y_{j(i)} &\sim \mathrm{N}\big(\alpha + \beta_i + \gamma_{j(i)} + (\delta + \eta_i) \odot x_{j(i)}, \sigma_\epsilon^2\big) \text{ for } j = 1, \ldots, n \text{ and } i = 1, \ldots, l, \\
\alpha &\sim \mathrm{N}(0, \sigma_\alpha^2), \\
\delta &\sim \mathrm{N}(0, \sigma_\delta^2), \\
\beta_i &\sim \mathrm{N}(0, \sigma_\beta^2) \text{ for } i = 1, \ldots, l-1,
\end{aligned}
$$

(continued on next page)

(4.4)

$$
\begin{aligned}
\beta_l &= -\sum_{i=1}^{l-1} \beta_i, \\
\eta_i &\sim \mathrm{N}(0, \sigma_\eta^2) \text{ for } i = 1, \ldots, l-1, \\
\eta_i &= -\sum_{i=1}^{l-1} \eta_i, \\
\boldsymbol{\gamma}_i &\sim \mathrm{N}_{n_i}\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D}_i)\big) \text{ for } i = 1, \ldots, l \text{ (for model 4.4}a\text{)}, \\
\boldsymbol{\gamma} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D})\big) \text{ (for model 4.4}b\text{)}, \\
(\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \\
(\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\
(\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta), \\
(\sigma_\gamma^2)^{-1} &\sim \mathrm{Ga}(a_\gamma, b_\gamma), \\
(\sigma_\delta^2)^{-1} &\sim \mathrm{Ga}(a_\delta, b_\delta), \\
(\sigma_\eta^2)^{-1} &\sim \mathrm{Ga}(a_\eta, b_\eta),
\end{aligned}
$$

where model 4.4a fits a smooth surface for the spatially-varying intercept parameter over each lake separately and model 4.4b fits a smooth surface for the spatially-varying intercept parameter over all lakes at once.

These models are simplified by removing the lake-specific parameters, to give:

$$
\begin{aligned}
y_{j(i)} &\sim \mathrm{N}(\alpha + \gamma_{j(i)} + \delta \odot x_{j(i)}, \sigma_\epsilon^2) \text{ for } j = 1, \ldots, n \text{ and } i = 1, \ldots, l, \\
\alpha &\sim \mathrm{N}(0, \sigma_\alpha^2), \\
\delta &\sim \mathrm{N}(0, \sigma_\delta^2), \\
\boldsymbol{\gamma}_i &\sim \mathrm{N}_{n_i}\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D}_i)\big) \text{ for } i = 1, \ldots, l \text{ (for model 4.5}a\text{)}, \\
\boldsymbol{\gamma} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D})\big) \text{ (for model 4.5}b\text{)}, \\
(\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \\
(\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\
(\sigma_\gamma^2)^{-1} &\sim \mathrm{Ga}(a_\gamma, b_\gamma), \\
(\sigma_\delta^2)^{-1} &\sim \mathrm{Ga}(a_\delta, b_\delta),
\end{aligned}
\tag{4.5}
$$

where model 4.5a fits a smooth surface for the spatially-varying intercept parameter over each lake separately and model 4.5b fits a smooth surface for the spatially-varying intercept parameter over all lakes at once.

Comparing results from this model to results from models 4.4a or 4.4b allows the assessment of whether lake-specific parameters are needed in this modelling framework, for the datasets investigated.

Single-lakes model for comparison

A single-lakes version of model 4.4 is:

$$\begin{aligned}
y_{j(i)} &\sim \mathrm{N}(\beta_i + \gamma_{j(i)} + \eta_i x_{j(i)}, \sigma^2_{\varepsilon,i}) \text{ for } j = 1, \ldots, n_i \text{ and } i = 1, \ldots, l, \\
\beta_i &\sim \mathrm{N}(0, \sigma^2_{\beta,i}), \\
\eta_i &\sim \mathrm{N}(0, \sigma^2_{\eta,i}), \\
\boldsymbol{\gamma}_i &\sim \mathrm{N}_{n_i}\big(\mathbf{0}, \sigma^2_{\gamma,i}\exp(-\phi_\gamma \mathbf{D}_i)\big), \\
(\sigma^2_{\varepsilon,i})^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \\
(\sigma^2_{\beta,i})^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta), \\
(\sigma^2_{\gamma,i})^{-1} &\sim \mathrm{Ga}(a_\gamma, b_\gamma), \\
(\sigma^2_{\eta,i})^{-1} &\sim \mathrm{Ga}(a_\eta, b_\eta).
\end{aligned} \tag{4.6}$$

This is a single-lakes model written in the same framework as the multiple-lakes models. It is fitted to each lake i separately, for $i = 1, \ldots, l$. This allows a comparison, to assess whether single-lakes or multiple-lakes models produce the best calibration of the remotely-sensed data with the *in situ* data. A version of this model, to be fitted to data for all lakes at once, is:

$$\begin{aligned}
y_j &\sim \mathrm{N}(\beta + \gamma_j + \eta x_j, \sigma^2_\varepsilon) \text{ for } j = 1, \ldots, n, \\
\beta &\sim \mathrm{N}(0, \sigma^2_\beta), \\
\eta &\sim \mathrm{N}(0, \sigma^2_\eta), \\
\boldsymbol{\gamma} &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_\gamma \exp(-\phi_\gamma \mathbf{D})\big),
\end{aligned} \tag{4.7}$$

$$(\sigma_\varepsilon^2)^{-1} \sim \text{Ga}(a_\varepsilon, b_\varepsilon),$$
$$(\sigma_\beta^2)^{-1} \sim \text{Ga}(a_\beta, b_\beta),$$
$$(\sigma_\gamma^2)^{-1} \sim \text{Ga}(a_\gamma, b_\gamma),$$
$$(\sigma_\eta^2)^{-1} \sim \text{Ga}(a_\eta, b_\eta).$$

Model 4.7 treats all of the data as if they are from the same lake.

Spatiotemporal multiple-lakes statistical downscaling

The spatial statistical downscaling models can be developed into spatiotemporal models, allowing the models to be fitted to data for various times at once and also allowing the sharing of information across time, in order to improve the estimation of the parameters within each model. Spatiotemporal versions of models 4.4a, 4.4b, 4.5a, 4.5b, 4.6 and 4.7 are:

$$y_{h,j(i)} \sim \text{N}(\alpha + \beta_i + \gamma_{h,j(i)} + (\delta + \eta_i) \odot x_{h,j(i)}, \sigma_\epsilon^2), \tag{4.4-ST}$$

$$y_{h,j(i)} \sim \text{N}\big(\alpha + \gamma_{h,j(i)} + \delta \odot x_{h,j(i)}, \sigma_\epsilon^2\big), \tag{4.5-ST}$$

$$y_{h,j(i)} \sim \text{N}(\beta_i + \gamma_{h,j(i)} + \eta_i x_{h,j(i)}, \sigma_{\varepsilon,i}^2) \tag{4.6-ST}$$

and

$$y_{h,j} \sim \text{N}(\beta + \gamma_{h,j} + \eta x_{h,j}, \sigma_\varepsilon^2), \tag{4.7-ST}$$

where all prior distributions are the same as those for the spatial-only multiple-lakes downscaling models, with the exceptions of:

$$\boldsymbol{\gamma}_{h,i} \sim \text{N}_{n_i}\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D}_i)\big) \tag{a}$$

and

$$\boldsymbol{\gamma}_h \sim \text{N}_n\big(\mathbf{0}, \sigma_\gamma^2 \exp(-\phi_\gamma \mathbf{D})\big), \tag{b}$$

to give models 4.4a-ST, 4.5a-ST and 4.6-ST (each using prior distribution (a) for $\boldsymbol{\gamma}_{h,i}$) and 4.4b-ST, 4.5b-ST and 4.7-ST (each using prior distribution (b) for $\boldsymbol{\gamma}_h$).

4.2.2 Model fitting and results

In order to gain an understanding of how well the models described in the previous subsections fit to the data for the Great Lakes, a k-fold cross-validation is carried out. There are *in situ* $\log(\text{chlorophyll}_a)$ data available for 19, 11, 14 and 20 locations for lakes Superior, Michigan, Huron and Erie, respectively, collected by the US Environmental Protection Agency (EPA) and made available through the Great Lakes Monitoring website (greatlakesmonitoring.org). These data are available for 20 months (April and August each year, between August 2002 and April 2012).

A k-fold cross-validation is where data are split into k subsets, before fitting the model with each of the k data subsets removed in turn. Predictions are then made at these removed locations and various model performance statistics calculated, based upon the predictions and their relationship with the observed *in situ* data. Here, k is chosen as 16, so that data for 4 of the 64 *in situ* locations are left out each time. The data are sorted randomly into these 16 data subsets, using the `sample` function in `R`. This high value of k and the resulting small number of locations to remove each time removes the possibility that all data for a lake could be put into a single subset and removed entirely for a model run.

The convergence of parameters is checked for each model, using trace and density plots (see Figures B.15, B.16, B.17, B.18 and B.19 on pages 232 to 236), while diagnostic plots (see Figures B.38, B.39, B.40, B.41 and B.42 on pages 248 to 250) provide no evidence against the validity of the assumptions that residuals are mean-zero Normally distributed and homoscedastic.

The RMSE, MAE, variance of predictions, empirical 95% credible interval (CI) coverage and mean interval length are calculated for each model, to

compare the model performances at predicting the observed *in situ* data.

The results of this k-fold cross-validation are given in Tables 4.3 and 4.4. Models 4.4a-ST, 4.4b-ST (which are both spatiotemporal multiple-lakes models without spatially-varying slope parameters) and 4.5a-ST (the spatiotemporal multiple-lakes model without lake-specific parameters) are fitted to the data. In addition, model 4.6-ST (the spatiotemporal single-lakes model) is fitted to the data for each lake separately and to data for all lakes (as model 4.7-ST), treating the four Great Lakes as a single large lake.

Model	RMSE	MAE	Variance of predictions	95% CI coverage probability	Mean CI length
Model 4.4a-ST	0.424	0.301	0.845	0.949	1.582
Model 4.4b-ST	0.392	0.278	0.858	0.940	1.675
Model 4.5a-ST	0.439	0.309	0.855	0.947	1.594
Model 4.6-ST, separate lakes	0.425	0.302	0.849	0.950	1.582
Model 4.7-ST, all lakes	0.395	0.280	0.844	0.941	1.754

Table 4.3: Performance statistics for several models for four Great Lakes, for $\log(\text{chlorophyll}_a)$.

Model	Lake	RMSE	MAE	Variance of predictions	95% CI coverage probability	Mean CI length
Model 4.4a-ST	Superior	0.290	0.220	0.068	0.941	1.150
	Michigan	0.261	0.211	0.068	0.981	1.153
	Huron	0.277	0.213	0.032	0.954	1.075
	Erie	0.658	0.504	0.609	0.930	2.598
Model 4.4b-ST	Superior	0.289	0.218	0.066	0.991	1.838
	Michigan	0.261	0.210	0.069	0.976	1.143
	Huron	0.278	0.213	0.033	0.954	1.070
	Erie	0.626	0.480	0.445	0.938	2.593
Model 4.5a-ST	Superior	0.289	0.219	0.078	0.941	1.175
	Michigan	0.257	0.206	0.073	0.995	1.733
	Huron	0.264	0.196	0.046	0.989	1.640
	Erie	0.564	0.423	0.475	0.833	1.514
Model 4.6-ST, separate lakes	Superior	0.290	0.220	0.068	0.941	1.148
	Michigan	0.263	0.212	0.069	0.981	1.146
	Huron	0.278	0.214	0.033	0.950	1.068
	Erie	0.627	0.480	0.448	0.940	2.592
Model 4.7-ST, all lakes	Superior	0.292	0.221	0.087	0.991	2.001
	Michigan	0.264	0.209	0.102	0.995	1.818
	Huron	0.280	0.207	0.077	0.986	1.690
	Erie	0.562	0.419	0.545	0.840	1.529

Table 4.4: Individual performance statistics for several models for four Great Lakes, for $\log(\text{chlorophyll}_a)$.

Table 4.3 shows that models 4.4a-ST and 4.6-ST, fitted to each lake separately, both perform fairly similarly. This is also true, when the performance of the models is compared for each lake, separately (see Table 4.4). This indicates that having the overall parameters α and δ in model 4.4a-ST, with β_i and η_i treated as lake-specific adjustments, may be unnecessary, since treating β_i and η_i as parameters in their own right, as in model 4.6-ST, leads to similar results.

Model 4.4b-ST performs very similarly to model 4.7-ST, fitted to data for all lakes at once. This suggests that the lake-specific parameters of model 4.4b-ST may not be required, for this dataset. This appears to be the case for all four lakes investigated.

The models that fit a spatial surface over all four lakes perform better at predicting *in situ* data than those that fit a separate spatial surface to each lake. From these results, model 4.7-ST, fitted to all lakes at once, performs as well as the model developed specifically for multiple-lakes downscaling (4.4b-ST), providing no evidence that a multiple-lakes downscaling model is needed. This means that model 4.7-ST can be fitted to the data for several related lakes simultaneously, providing a continuous spatial surface of predictions over all four lakes.

An example plot of predictions for August 2003 is shown in Figure 4.6. In order to do this, model 4.7-ST is fitted to data for 15 months for which data are available for all 64 *in situ* sampling locations across four Great Lakes. A Delaunay triangulation is carried out, in order to select around 1000 prediction locations within the lakes. This ensures that a large enough number of prediction locations, with optimal coverage, is used to gain a good understanding of spatial patterns in the lakes, while ensuring that the model is not too computationally expensive. This plot helps to explain why model 4.7-ST is preferred over the multiple lakes models. Since the spatial patterns are fairly similar over the four lakes, it makes sense to fit a smooth spatial surface covering all lakes at once. The bottom plot shows that the *in situ*

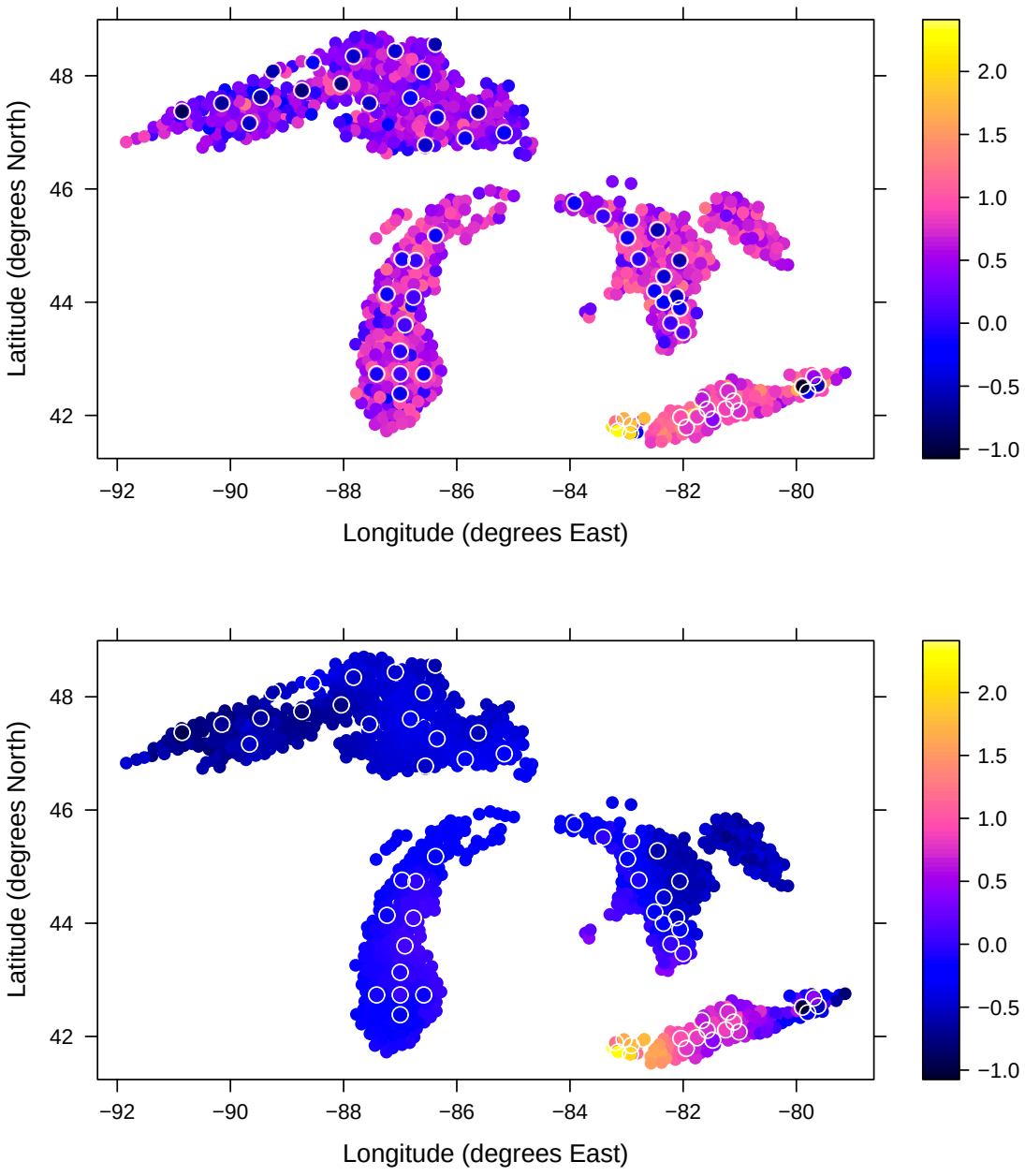

Figure 4.6: Remotely-sensed log(chlorophyll$_a$) (top) and downscaled surface (bottom) from model 4.7-ST for August 2003, for 1005 locations in four Great Lakes. *In situ* data are overlaid, surrounded by white circles.

data are similar to their surrounding remotely-sensed data, demonstrating that calibration has been achieved here. Even with this common spatial surface over all lakes, the differences in spatial patterns between the lakes

are accounted for by the model, with Lake Michigan having generally higher values of log(chlorophyll$_a$) than Lake Superior, for example. Additionally, the much higher levels of log(chlorophyll$_a$) in Lake Erie are accounted for by the model, without the necessity for lake-specific parameters.

4.2.3 Conclusions and discussion

It was found that a specific multiple-lakes model is not necessary in the context of statistical downscaling of data for four Great Lakes. It is possible to instead fit a model over all four lakes, ignoring the effect of the lake boundaries. This adequately calibrates the remotely-sensed data for all four lakes, using the *in situ* data. The model predicts reasonable levels of log(chlorophyll$_a$) over all four lakes, with lower levels in Lake Superior and higher levels (with more variation) in Lake Erie. These predictions reflect the patterns in the *in situ* data and also reflect the known ecological patterns in the lake system, with water flowing from Superior through to Lake Erie, picking up more nutrients as it flows.

It was found that fitting a model with a smooth surface over all four lakes at once leads to improved predictions in comparison to fitting a model with separate spatial surfaces over all four lakes. This makes sense, since fitting a single spatial surface over all four lakes involves the sharing of information, which leads to improved estimates of the parameters and hence improved accuracy of predictions.

For this application, it is fortunate that there are *in situ* data available for the same months for each of 64 locations in four lakes. This is due to the fact that all four lakes are sampled by the US Environmental Protection Agency, using the same equipment. In other cases, there may not be *in situ* data available at similar times, so that the model could not be fitted to several lakes at once. In order to downscale data of this type, the model would need to account for the different sampling times of the data. A modelling framework that accomplishes this, using functional data methodology,

is developed in the following chapter.

The methodology is able to downscale data for several nearby lakes at once. Downscaling data for several distant lakes is another issue. In the case of lakes that are far apart, it may make more sense to simply fit separate downscaling models to each lake. This would require further study.

4.3 Overall conclusions and discussion

In this chapter, developments of the spatiotemporal statistical downscaling model were introduced. These involved the sharing of information between simultaneously-downscaled variables and also between neighbouring lakes.

The application of bivariate downscaling is motivated by the availability of two related variables in the Lake Balaton dataset, namely log(total suspended matter) and log(chlorophyll$_a$), which have positive correlation. The model that was developed fits correlation between the errors for the two variables. It was found that simultaneously downscaling these two variables, using this bivariate downscaling model, led to a slightly improved accuracy of calibration. Again, sharing information enabled an improved accuracy of predictions.

Multiple lakes downscaling is motivated by the data for the Great Lakes of North America. The data for Lakes Superior, Michigan, Huron and Erie were investigated here. All four lakes are connected as part of the same lake system. However, each lake differs in characteristics such as size and shape and position in the lake system, so differences between the lakes were considered in the analysis. The second section in this chapter investigated multiple-lakes downscaling models, that accounted for differences between lakes, while also sharing information between them on the relationship between *in situ* and remotely-sensed data. Various models were developed and it was discovered that downscaling data for multiple lakes simultaneously

could in fact be accomplished by treating all data as being from a single lake and fitting a single smooth surface over all lakes at once. It was found that this method resulted in more accurate predictions than fitting separate models to each lake individually.

Chapter 5

Nonparametric statistical downscaling

This chapter presents the development and application of a method for the fusion of data of differing spatial and temporal support, incorporating aspects of both datasets. Nonparametric statistical downscaling is a novel technique to accomplish this data fusion, incorporating aspects of both statistical downscaling and functional data analysis. This chapter describes the motivation behind the method, the development of the nonparametric statistical downscaling model, applications to lake water quality datasets for Lakes Balaton and Erie and, finally, a brief discussion of the utility of the method and the conclusions reached.

5.1 Background and motivation

So far, the question of data fusion between data of differing spatial support has been addressed through statistical downscaling, specifically in the context of fusing *in situ* and remotely-sensed log(chlorophyll$_a$) data for various lakes. However, these methods require that both *in situ* and remotely-sensed data are available on the same temporal scale, an assumption that is often not met in practice. While both the *in situ* and remotely-sensed

data are available at a fixed set of sampling locations (point locations and grid cells, respectively), the *in situ* data are sampled irregularly over time, whereas the remotely-sensed data are available for monthly-averages. Methodology in earlier chapters simply averages *in situ* data over their months, to put them on the same temporal support as their corresponding remotely-sensed data. However, this still leaves gaps in the *in situ* data, with some months having no *in situ* data sampled at any location, so that the statistical downscaling model is unable to calibrate the remotely-sensed data for these months.

Functional data analysis is an approach that treats the data as observations of smooth functions, rather than simply points, assuming an underlying continuous process. There has been extensive research into fitting curves to temporal data, in the context of functional data analysis, and so the use of this approach appears to offer a solution to the temporal support problem. This motivates the development of a model for nonparametric statistical downscaling, which is described and explained later in this chapter, after some preliminary applications to the data.

5.2 Preliminary application to the data

Before developing the nonparametric statistical downscaling model, some preliminary applications of functional data methodology are carried out, using the $\log(\text{chlorophyll}_a)$ data for Lake Balaton. To recap briefly, there are *in situ* data available for 9 locations within the lake, sampled frequently between 2002 and 2012, but with some gaps, especially during the year of 2007 for the data collected by the BLI (locations 1 to 5 of 9). The remotely-sensed data are available for 115 monthly averages for 7616 grid cells across the lake, so there are no issues of missing data to deal with for the remotely-sensed data.

In this section, cubic B-splines are chosen as the basis type. Basis di-

mension differs between locations, in this preliminary demonstration, but equally-spaced breakpoints will be used.

5.2.1 Preliminary application of frequentist model

The frequentist model is:

$$f(y_i) = \sum_{k=1}^{m} \phi_k(x_i) c_k,$$

for $i = 1, \ldots, n$, where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is a vector of data recorded at times $t_1, \ldots, t_n$, ϕ_k is the kth basis function and c_k is the coefficient corresponding to the kth basis function. The model is fitted by least squares (see equations 1.15 to 1.18 on page 32) to the *in situ* data for locations 1 and 9, separately, in order to demonstrate the differing fit to sparse and more abundant data.

For location 9, data are available between June 2002 and December 2011, with 93 data points available in total. The pattern of log(chlorophyll$_a$) in Lake Balaton has two peaks per year (Palmer et al. 2015), which should require around 4 basis functions per year. Taking into account the additional knots required at each endpoint, the number of basis functions for this preliminary application is calculated as 41. The 41 cubic B-spline basis functions are plotted (see Figure 5.1, top). Using the `fda` (Ramsay et al. 2014) package in `R` (R Core Team 2017), a smooth function is fitted to the *in situ* data for location 9 and, separately, to the remotely-sensed data for location 9, which is made up of 115 monthly averages between June 2002 and December 2011. Model fits to the data are plotted (see Figure 5.1, centre and bottom). These plots show that the two-peaks-per-year pattern in the *in situ* log(chlorophyll$_a$) data is adequately captured by the fitted smooth function, which follows the data closely. The remotely-sensed data for the grid cell corresponding to location 9 has a less regular pattern than the *in situ* data, so that the smooth function does not follow the data so closely, but it does follow most of the patterns in the data.

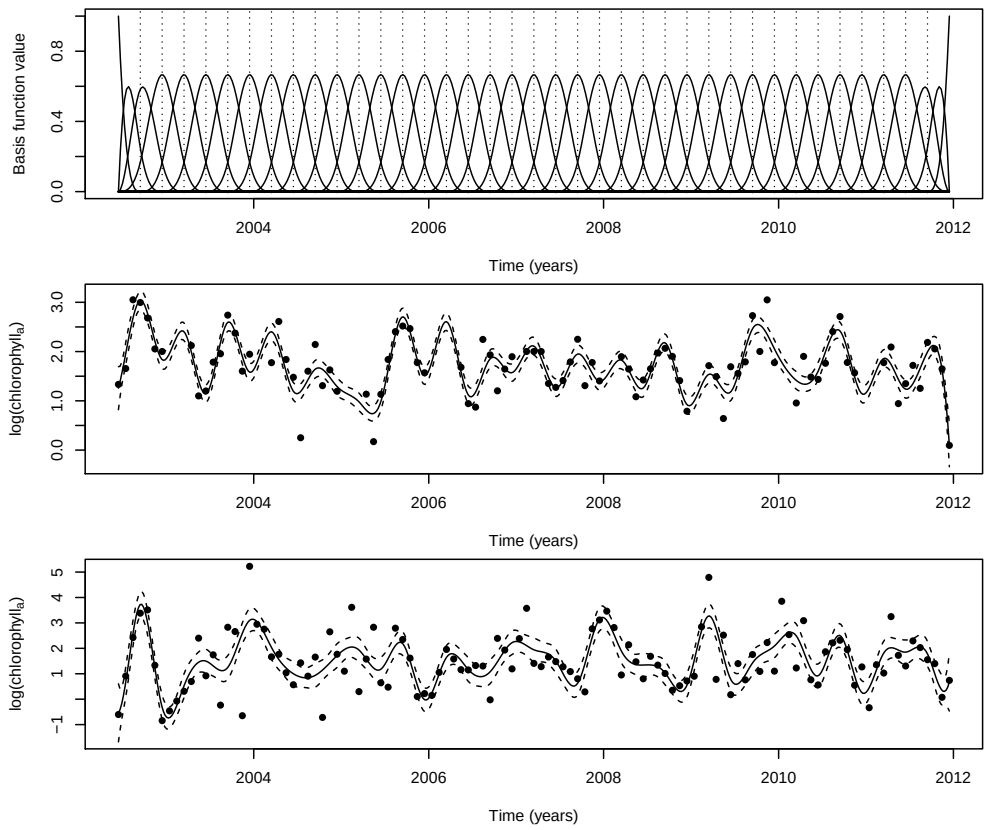

Figure 5.1: Basis functions (top), smooth function fitted to *in situ* data (centre) and smooth function fitted to remotely-sensed data (bottom), for Lake Balaton location 9. 95% confidence intervals for fitted function values are shown as dashed lines and data are solid points.

For location 1, the *in situ* data are only available between May 2006 and December 2011, with no data during 2007, giving a total of 45 available data points. In order to adequately fit the smooth function to the pattern of two peaks in log(chlorophyll$_a$) per year, the basis dimension should be close to 25. Fitting a smooth function using this basis dimension is, however, not possible, since the matrix $(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})$ is singular and so cannot be inverted, due to having too few data points. A smaller dimensional basis is instead used, allowing for the smooth function to be fitted to the data, but this has the expense of being only able to model longer-term patterns in the data. Figure 5.2 shows the basis functions and the fitted smooth curve using cubic B-splines of dimension 15. Here, the smooth curve can only fit a pattern of one peak of log(chlorophyll$_a$) per year. However, the basis dimension is

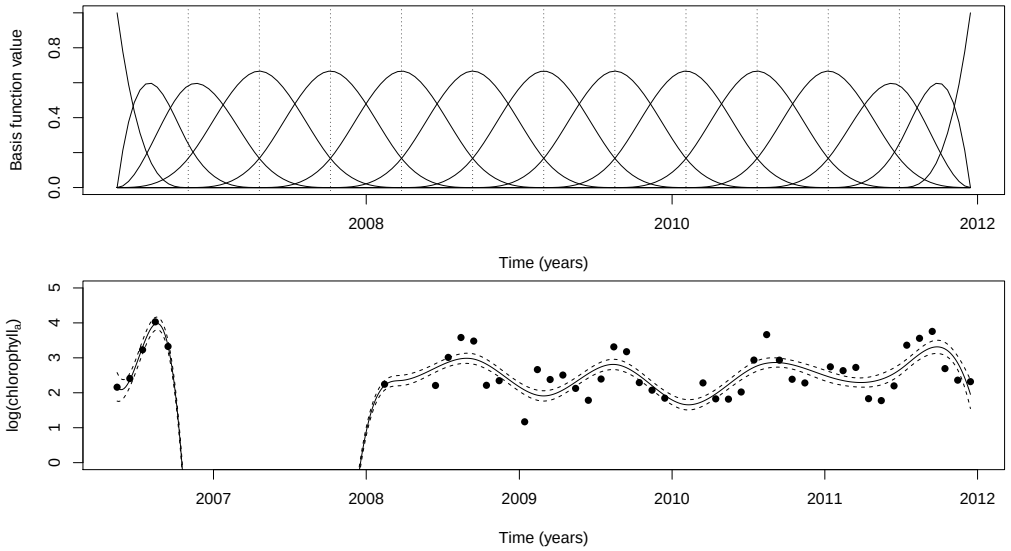

Figure 5.2: Basis functions (top) and smooth function fitted to *in situ* data (bottom), for Lake Balaton location 1. 95% confidence intervals for fitted function values are shown as dashed lines and data are solid points.

still too large, since the estimated curve and confidence intervals during 2007 lie far from the values of observed data for any other year, reaching values close to -50. Any B-spline basis fitted to these data must not have multiple breakpoints during the gap in the data in 2007, so a much smaller basis dimension is required. A dimension 5, cubic B-spline basis avoids extreme curve estimates, but does not show useful patterns in the data, meaning that it is of limited use here.

5.2.2 Preliminary application of Bayesian model

In order to resolve the difficulties found when fitting the frequentist model to data with gaps over time, the Bayesian formulation of the model is investigated. As stated in equations 1.23 and 1.24 on page 33, this model is:

$$\begin{aligned}
\mathbf{y}|\mathbf{c}, \sigma_\varepsilon^2 &\sim \mathrm{N}_n(\mathbf{\Phi c}, \sigma_\varepsilon^2 \mathbf{I}_n), \\
(\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a, b), \\
\mathbf{c} &\sim \mathrm{N}_m(\boldsymbol{\mu}, \mathbf{\Sigma}),
\end{aligned}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ is the n-length vector of data, $\mathbf{\Phi}$ is the $n \times m$ matrix of basis functions evaluated at the times of data collection $t_1, \ldots, t_n$, $\mathbf{c} = (c_1, \ldots, c_m)^{\mathrm{T}}$ is the vector of basis coefficients corresponding to each basis function, σ_ε^2 is the variance of the errors, $\mathbf{I}_n$ is the $n \times n$ identity matrix and a, b, $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are parameters that are chosen *a priori*.

The prior distributions on the basis coefficients and error variance provide additional information that comes in useful around gaps in the data. The aim of fitting a model with equal basis dimension for all locations motivates the investigation of the Bayesian model. Before fitting the model, the coefficients for the prior distributions in equation 1.24 are chosen as $\boldsymbol{\mu} = \mathbf{0}$, $\mathbf{\Sigma} = 100 \times \mathbf{I}_m$, $a = 0.001$ and $b = 0.001$. These are chosen to provide minimal prior information, to avoid influencing estimates for times when data were available, but to provide enough information to enable model fitting without extreme curve estimates at times when no data were available. $\mathbf{0}$ is a reasonable prior mean value to choose, when no further information is available, since it is not known whether the true mean value is positive or negative (Denison et al. 2002). $\mathbf{I}_m$ is sensible, since no further information on the dependence structure is available *a priori* (Denison et al. 2002). $a = 0.001$ and $b = 0.001$ are common choices for coefficients of the Gamma distribution, since small values of these coefficients only slightly affect the values of the posterior estimates for parameter σ_ε^2. The choice of 100, in the mean of the prior covariance matrix $100 \times \mathbf{I}_m$, is data-dependent, chosen so as not to be too small (such as 1, which leads to a strong prior distribution and pulls posterior basis coefficient estimates close to the prior mean), while also not too large (such as 10^6, which leads to an almost uninformative prior distribution that does not provide enough information in the presence of gaps in the data). The observed variance in the basis coefficients is observed to vary only up to values around 10, when fitted in the frequentist framework. A sensitivity analysis was carried out, where the model was re-fitted with 100 varied from 20 to 1000, to ensure that this made little difference to the model

estimates.

Again, examples of the model fit to *in situ* data for locations 1 and 9 are given here. For both locations, the same basis is used, namely the cubic B-spline basis of dimension 41. Models are run using `C++` code using the `R` packages `Rcpp` (Eddelbuettel & François 2011, Eddelbuettel 2013) and `RcppArmadillo` (Eddelbuettel & Sanderson 2014). The model is run for 10,000 iterations, for each location, and checked for convergence using trace and density plots. The fact that the model predictions (and credible intervals) over time display smooth patterns is also an indication that the convergence of parameters to their stationary distributions has been reached, since otherwise a jagged pattern would be observed.

The fitted smooth curve is plotted for the *in situ* data for location 9 (see Figure 5.3), along with the 95% credible intervals. The fitted smooth curve

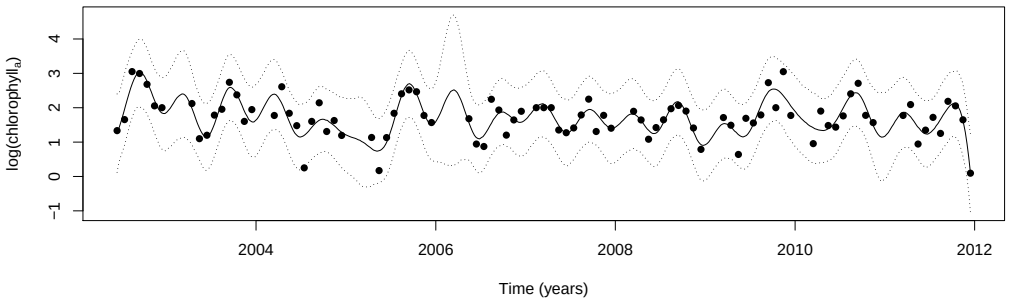

Figure 5.3: Smooth function fitted to the Lake Balaton location 9 *in situ* data using the Bayesian model (solid line), along with the corresponding 95% credible intervals (dotted lines). The data are shown as points.

is identical to that for the frequentist model. The fitted credible intervals are fairly wide and include almost all of the observed *in situ* data, suggesting that the empirical coverage probabilities are close to their nominal 95%.

The benefit to fitting the Bayesian model is illustrated for the *in situ* data for location 1. A plot of the resulting fitted smooth curve is shown in Figure 5.4. The fitted smooth curve follows the two-peaks-per-year pattern, where data are available. During the gap in available data during 2007, the model predicts that the data reach as low as -2, but credible intervals are appropriately wide for this time period, covering the range of values

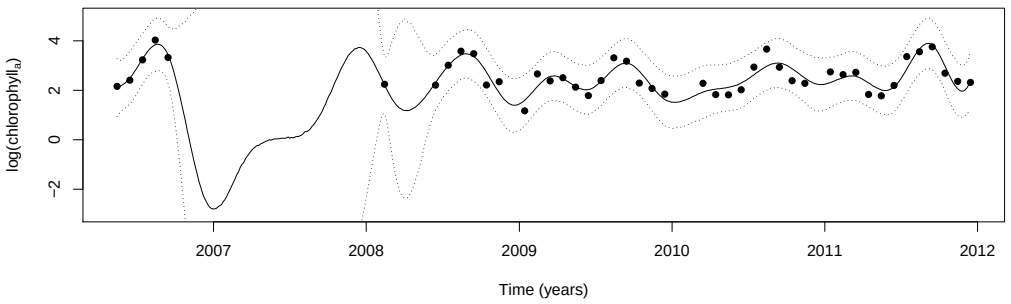

Figure 5.4: Smooth function fitted to the Lake Balaton location 1 *in situ* data using the Bayesian model (solid line), along with the corresponding 95% credible intervals (dotted lines). The data are shown as points.

observed in the available data. These intervals lie between -14 and 14 at their extremes. Although these wide intervals are unavoidable during the periods without data, the Bayesian model is able to fit a smooth function allowing for the fitting of up to two peaks per year, while appropriately expressing the lack of certainty in function estimates during periods without data.

5.3 Developing a model for nonparametric statistical downscaling

Nonparametric statistical downscaling combines the method of expressing the shape of the data through functions defined by their basis coefficients, with data fusion of the *in situ* and remotely-sensed data. Since the *in situ* data for a location and the remotely-sensed data for the corresponding grid cell are measures of the same variable, it is assumed that their basis coefficients are positively related, as long as the same basis is used for both datasets. This idea is explored in the next subsection. Should this be a reasonable assumption to make, the fully Bayesian nonparametric statistical downscaling model makes use of these basis coefficients for different locations, in order to fuse data of different spatiotemporal support.

This section focusses on model development, firstly using a simple linear

regression to test the idea behind nonparametric downscaling, before moving on to the fully Bayesian functional downscaling model.

5.3.1 Examining the correspondence of *in situ* and remotely-sensed basis coefficients

Firstly, the relationship between basis coefficients for the *in situ* and remotely-sensed data for one location is explored, taking the example of Lake Balaton location 9. A scatterplot of these coefficients (see Figure 5.5) shows that there is a positive, probably linear relationship between the two

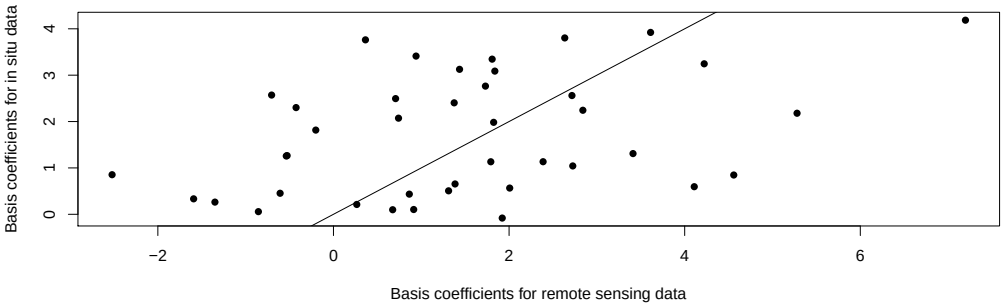

Figure 5.5: Scatterplot of *in situ* and remote sensing data basis coefficients, from fitting curves using a cubic B-spline basis of dimension 41.

sets of basis coefficients. A linear model is fitted, of the form:

$$c_j = \alpha + \beta d_j + \varepsilon_j, \tag{5.1}$$

where c_j $(j = 1, \ldots, m)$ are the basis coefficients for the *in situ* data, d_j are the basis coefficients for the remotely-sensed data and $\varepsilon_j \sim \mathrm{N}(0, \sigma_\varepsilon^2)$ are the random errors. Predictions are:

$$\hat{c}_j = \hat{\alpha} + \hat{\beta} d_j, \tag{5.2}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimated intercept and slope, respectively, from model 5.1. The back-transformation to give the predicted *in situ* data value at time

t, i.e. $\hat{y}(t)$, is:

$$\hat{y}(t) = \sum_{j=1}^{m} \hat{c}_j \phi_j(t), \tag{5.3}$$

where t is some time at which to predict and $\phi_j(t)$ is the jth basis function evaluated at this time. These models are fitted in the frequentist framework, through least squares. The resulting predicted curve from this method is plotted, for location 9, in Figure 5.6. The predicted smooth curve from this

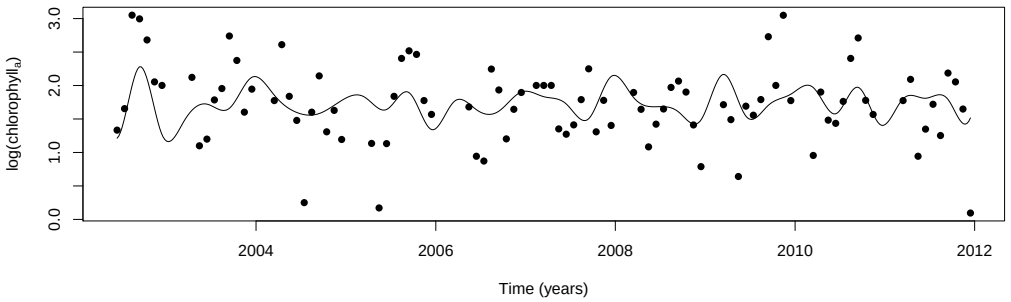

Figure 5.6: Predicted smooth curve using basis coefficients estimated using a linear model, for Lake Balaton location 9 data.

model resembles the observed *in situ* data, so that this preliminary analysis suggests that the idea of calibrating data through fitting the relationship between basis coefficients for *in situ* and remote sensing data is appropriate and warrants further development.

5.3.2 Combining a linear model and functional data analysis methodology

A first step to developing a fully Bayesian nonparametric statistical downscaling model is to investigate the combination of functional data analysis methodology and a linear model. Since the previous section shows that the basis coefficients for *in situ* and remote sensing data for a single location show a positive, possibly linear relationship, it makes sense to pursue this challenge.

Given *in situ* log(chlorophyll$_a$) data $\mathbf{y}$ at a single location, remotely-sensed log(chlorophyll$_a$) data $\mathbf{x}$ for the grid cell containing the location of $\mathbf{y}$, a matrix of basis functions $\mathbf{\Phi}$ evaluated at the *in situ* sampling times and a matrix of basis functions $\mathbf{\Psi}$ evaluated at the remote sensing sampling times, separate Bayesian models are given for the estimation of the basis coefficients for the *in situ* data (Model 5.4), for the estimation of the basis coefficients for the remotely-sensed data (Model 5.5) and for the linear regression of the remotely-sensed basis coefficients on the *in situ* basis coefficients (Model 5.6). The model for the *in situ* basis coefficients is:

$$\mathbf{y}|\mathbf{c}, \sigma_y^2 \sim \mathrm{N}_q(\mathbf{\Phi c}, \sigma_y^2 \mathbf{I}_q), \tag{5.4}$$

where q is the number of *in situ* data $\mathbf{y}$, and where prior distributions are:

$$\begin{aligned} \mathbf{c} &\sim \mathrm{N}(\boldsymbol{\mu}_y, \mathbf{\Sigma}_y) \text{ and} \\ (\sigma_y^2)^{-1} &\sim \mathrm{Ga}(a_y, b_y). \end{aligned}$$

The model for the remotely-sensed basis coefficients is:

$$\mathbf{x}|\mathbf{d}, \sigma_x^2 \sim \mathrm{N}_p(\mathbf{\Psi d}, \sigma_x^2 \mathbf{I}_p), \tag{5.5}$$

where p is the number of remotely-sensed data $\mathbf{x}$, and where prior distributions are:

$$\begin{aligned} \mathbf{d} &\sim \mathrm{N}(\boldsymbol{\mu}_x, \mathbf{\Sigma}_x) \text{ and} \\ (\sigma_x^2)^{-1} &\sim \mathrm{Ga}(a_x, b_x). \end{aligned}$$

The linear model regressing the remotely-sensed basis coefficients on the *in situ* basis coefficients is:

$$\mathbf{c} \sim \mathrm{N}(\alpha \mathbf{1} + \beta \mathbf{d}, \sigma_\varepsilon^2 \mathbf{I}_m), \tag{5.6}$$

with prior distributions:

$$\begin{aligned} \alpha &\sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2), \\ \beta &\sim \mathrm{N}(\mu_\beta, \sigma_\beta^2) \text{ and} \\ (\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon). \end{aligned}$$

The next stage in model development is to combine these three models together. This model is:

$$\mathbf{y}|\mathbf{c}, \sigma_y^2 \sim \mathrm{N}_q(\mathbf{\Phi c}, \sigma_y^2 \mathbf{I}_q), \tag{5.7}$$

with prior distributions:

$$\begin{aligned} (\sigma_y^2)^{-1} &\sim \mathrm{Ga}(a_y, b_y) \text{ and} \\ \mathbf{c}|\alpha, \beta, \mathbf{d}, \sigma_\varepsilon^2 &\sim \mathrm{N}(\alpha \mathbf{1} + \beta \mathbf{d}, \sigma_\varepsilon^2 \mathbf{I}_m) \end{aligned}$$

and hyperprior distributions:

$$\begin{aligned} (\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \\ \alpha &\sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2) \text{ and} \\ \beta &\sim \mathrm{N}(\mu_\beta, \sigma_\beta^2). \end{aligned}$$

To estimate the basis coefficients for the remotely-sensed data within the model, the following three hyperpriors are added:

$$\begin{aligned} \mathbf{x}|\mathbf{d}, \sigma_x^2 &\sim \mathrm{N}_p(\mathbf{\Psi d}, \sigma_x^2 \mathbf{I}_p), \\ (\sigma_x^2)^{-1} &\sim \mathrm{Ga}(a_x, b_x) \text{ and} \\ \mathbf{d} &\sim \mathrm{N}_m(\boldsymbol{\mu}_d, \mathbf{\Sigma}_d). \end{aligned}$$

Posterior estimates of all parameters in this model are obtained through Gibbs sampling, with the posterior distribution of predictions given as $(\tilde{\mathbf{y}}|\cdot) \sim \mathrm{N}_{\tilde{q}}(\tilde{\mathbf{\Phi}}\mathbf{c}, \sigma_y^2 \mathbf{I}_{\tilde{q}})$, where $\tilde{q}$ is the number of times at which predictions are to be

made and $\tilde{\mathbf{\Phi}}$ is the matrix of basis functions evaluated at the prediction times.

5.3.3 Nonparametric statistical downscaling: a fully Bayesian model for data fusion

The next step in this procedure is to incorporate a statistical downscaling model, rather than a simple linear model. The fully Bayesian nonparametric statistical downscaling model is written as:

$$\begin{aligned}
\mathbf{y}_i|\mathbf{c}_i, \sigma_y^2 &\sim \mathrm{N}_{q_i}(\mathbf{\Phi}_i\mathbf{c}_i, \sigma_y^2\mathbf{I}_{q_i}),\\
(\sigma_y^2)^{-1} &\sim \mathrm{Ga}(a_y, b_y),\\
c_{ij}|\alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 &\sim \mathrm{N}(\alpha_{ij} + \beta_{ij}d_{ij}, \sigma_c^2),\\
\boldsymbol{\alpha}_j|\sigma_\alpha^2 &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\alpha^2\mathbf{H}_{22}(\phi_\alpha)\big),\\
\boldsymbol{\beta}_j|\sigma_\beta^2 &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_\beta^2\mathbf{H}_{22}(\phi_\beta)\big),\\
(\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha),\\
(\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta),\\
(\sigma_c^2)^{-1} &\sim \mathrm{Ga}(a_c, b_c),\\
\mathbf{x}_i|\mathbf{d}_i, \sigma_x^2 &\sim \mathrm{N}_{p_i}(\mathbf{\Psi}_i\mathbf{d}_i, \sigma_x^2\mathbf{I}_{p_i}),\\
(\sigma_x^2)^{-1} &\sim \mathrm{Ga}(a_x, b_x),\\
\mathbf{d}_i &\sim \mathrm{N}_m(\boldsymbol{\mu}_d, \mathbf{\Sigma}_d),
\end{aligned} \tag{5.8}$$

where:

- y_{ij} is the value of *in situ* data at time j at location i ($i = 1, \ldots, n$ and $j = 1, \ldots, q_i$).
- x_{ij} is the value of remotely-sensed data at time j at location i ($i = 1, \ldots, n$ and $j = 1, \ldots, p_i$).
- q_i is the number of *in situ* data collected at location i.
- p_i is the number of remotely-sensed data collected at location i.

- n is the number of *in situ* data locations i.
- m is the number of basis functions in each $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_i$.
- $\mathbf{\Phi}_i$ are the basis functions evaluated at times of data collection for y_i.
- $\mathbf{\Psi}_i$ are the basis functions evaluated at times of data collection for x_i.
- $\mathbf{H}_{22}(\phi_\alpha) = \exp(-\phi_\alpha \times \mathbf{D}_{22})$, where ϕ_α is selected *a priori* and $\mathbf{D}_{22}$ is the matrix of distances between *in situ* locations i.
- $\mathbf{H}_{22}(\phi_\beta) = \exp(-\phi_\beta \times \mathbf{D}_{22})$, where ϕ_β is selected *a priori*, with $\mathbf{D}_{22}$ as above.
- $a_y, b_y, a_\alpha, b_\alpha, a_\beta, b_\beta, a_c, b_c, a_x, b_x, \boldsymbol{\mu}_d$ and $\mathbf{\Sigma}_d$ are values to be chosen *a priori*. 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, results in non-informative prior distributions. Sensible values for $\boldsymbol{\mu}_d$ and $\mathbf{\Sigma}_d$ are $\mathbf{0}$ and $\mathbf{I}_m$, reflecting lack of knowledge of the signs of the coefficients $\mathbf{d}_i$ and of their dependence structure.

A directed acyclic graph (DAG) can be created for this model (see Figure 5.7) and the full conditional posterior distributions of the model parameters are given in the appendix (see section A.4 on page 211).

5.4 Model fitting

In this section, the nonparametric statistical downscaling model (5.8) is applied to the log(chlorophyll$_a$) data for Lakes Balaton and Erie, to demonstrate and evaluate its effectiveness for data fusion.

5.4.1 Application to data for Lake Balaton

The effectiveness of the nonparametric statistical downscaling model (5.8) is demonstrated through fitting to the log(chlorophyll$_a$) data for Lake Balaton. This process highlights the advantages, disadvantages and issues as-

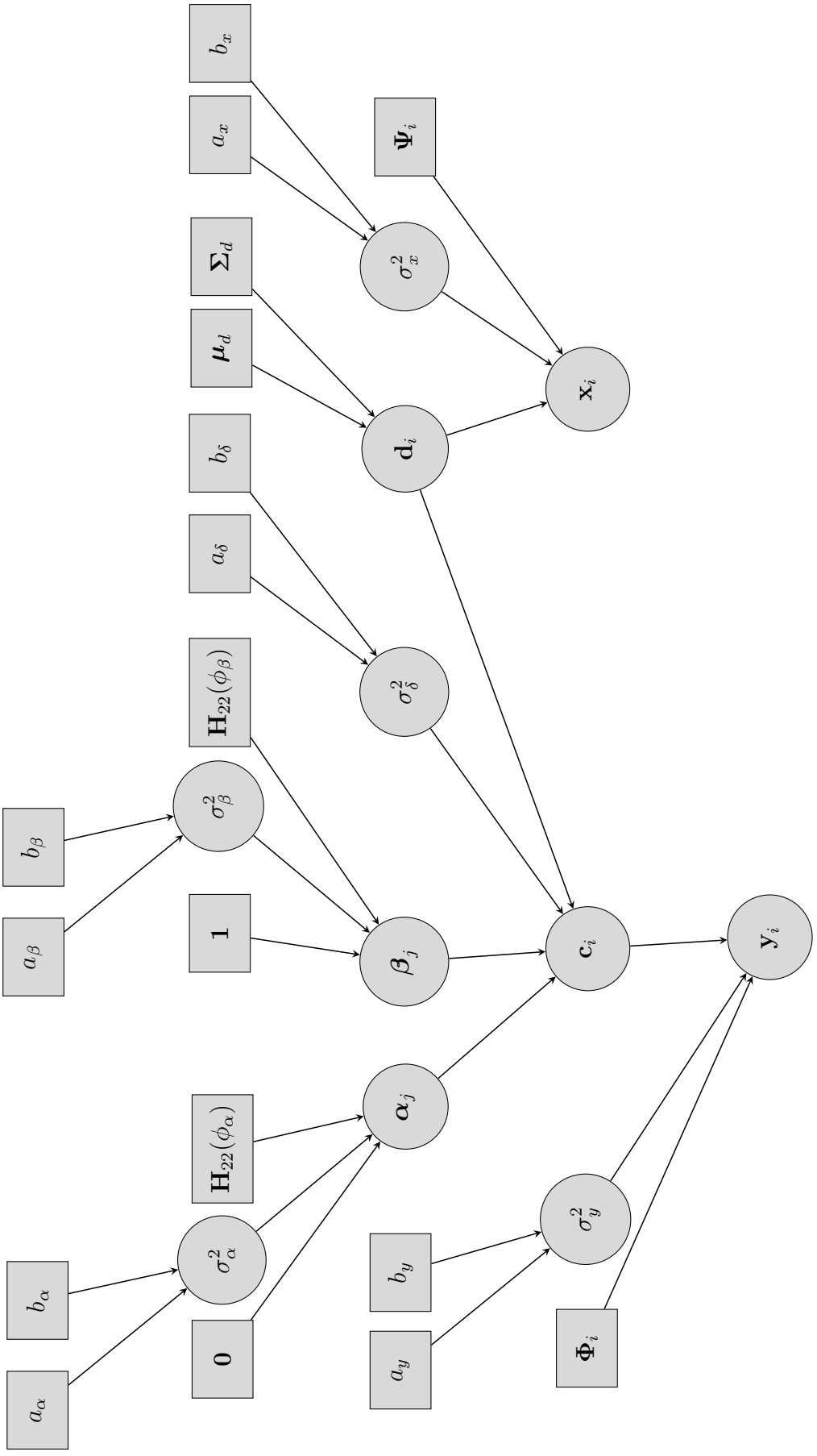

Figure 5.7: Directed acyclic graph (DAG) for the nonparametric statistical downscaling model (5.8). Circles are variables and rectangles are constants. Arrows represent direct dependencies between variables and constants.

sociated with fitting the model to log(chlorophyll$_a$) data. The process is illustrated using a B-spline basis and a Fourier basis. For these models, the spatial decay parameters ϕ_α and ϕ_β are each set equal to 0.1, as earlier results show that these are reasonable values for this dataset and that the model predictions are not sensitive to small changes in these values.

B-spline basis dimension for Lake Balaton

Given that a B-spline basis has been chosen, the basis dimension is estimated as:

$$\text{Basis dimension} = \left((t_{\max} - t_{\min}) \times \frac{2 \times r}{365} \right) + 3, \tag{5.9}$$

where r is the expected number of peaks in the data per year, $t_{\max}$ is the maximum sampling date and $t_{\min}$ is the minimum sampling date, with the difference $(t_{\max} - t_{\min})$ measured in days. The rationale behind this choice of formula is that the basis dimension must be large enough that the main patterns in the data are captured. The formula is based upon the assumption that there must be at least two basis functions per peak, in order to capture the pattern well enough. The additional three basis functions are required, since there are two additional breakpoints at each endpoint of the range of the basis and only one is otherwise already accounted for by the formula. This formula gives an estimate of 41 basis functions for the Lake Balaton dataset, based upon the assumption that two peaks in log(chlorophyll$_a$) are expected.

An empirical estimate of the optimal basis dimension is gained through GCV, DIC and a leave-one-out cross-validation, allowing the comparison of model performance for different numbers of basis functions. GCV and DIC both suggest that basis dimensions around 30 to 60 are reasonable. The model is fitted with basis dimension varying from 30 to 60 and a leave-one-out cross-validation is carried out for each dimension (with data for each lo-

cation in turn removed and predicted from the model fitted to the remaining data). For each basis dimension, the model is fitted to the log(chlorophyll$_a$) data for Lake Balaton and trace and density plots (see Figures B.20 and B.21 on pages 237 and 238 for an example using basis dimension 49) provide no evidence against the assumption that the MCMC chains have converged, while diagnostic plots (see Figure B.43 on page 250 for an example using basis dimension 49) provide no evidence against the model assumptions of homoscedasticity and mean-zero Normality of residuals. Root mean squared error (RMSE), mean absolute error (MAE), variance of predictions, mean 95% credible interval coverage and mean 95% credible interval width are calculated for each basis dimension (see Figure 5.8). RMSE and MAE should

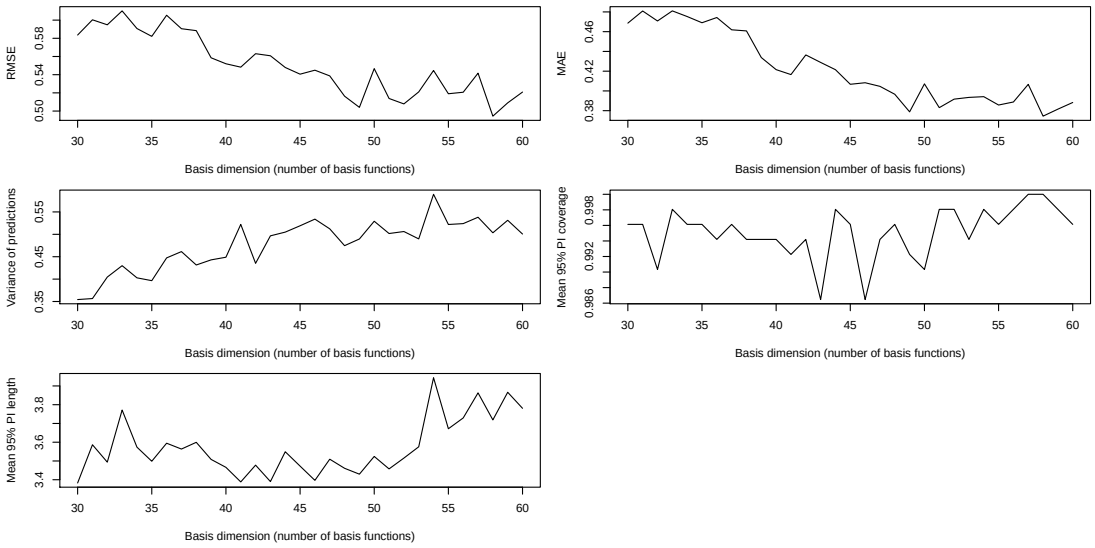

Figure 5.8: Plots of summary statistics versus basis dimension for a leave-one-out cross-validation for Lake Balaton, for B-spline basis.

give a good indication of how well the model predicts, when fitted using different basis dimensions. Both decrease with increasing basis dimension, until around 45, after which their values vary, but do not decrease further on average. The first time RMSE and MAE reach low points is at basis dimension 49, indicating that the model predicts fairly well using this basis dimension, and also indicating that the theoretical optimal value of 41 is slightly too small for this dataset. Increasing the dimension beyond 49 only serves to

increase computation time and mean credible interval length, without leading to an improved model predictive ability, so that 49 should be selected as the optimal basis dimension for the B-splines basis for the Lake Balaton $\log(\text{chlorophyll}_a)$ data.

Fourier basis dimension for Lake Balaton

If a Fourier basis is chosen, the formula to estimate the optimal basis dimension is:

$$\text{Basis dimension} = 2 \times \text{expected number of peaks in } \log(\text{chlorophyll}_a) \text{ per year} + 1, \tag{5.10}$$

which leads to an estimate of 5 as the optimal basis dimension, for the Lake Balaton $\log(\text{chlorophyll}_a)$ data. Similarly to the previous formula for estimating the basis dimension for the B-spline basis (equation 5.9), this formula assumes that two basis functions per peak, per year are required, with an additional knot required at an endpoint of the range of the basis. GCV and DIC also suggest that a small basis dimension is appropriate here. As with the B-spline basis, a leave-one-out cross-validation is a good way to estimate the optimal basis dimension empirically. This is carried out, similarly to that for the B-spline basis, with assumptions checked in the same way and found to be reasonable. Plots of resulting summary statistics (see Figure 5.9) show that RMSE and MAE decrease as basis dimension increases from 3 to 5, but then mostly level off. This agrees with the theory that at least 5 basis dimensions are required to model the two-peaks-per-year $\log(\text{chlorophyll}_a)$ pattern well, but the minimum value is reached at dimension 9. This may be due to the fact that 9 basis functions are able to model the shape of the two peaks better than 5 basis functions would be. The mean credible interval length reaches a minimum at 7, but is still low at 9. From this, 9 is a reasonable choice for the optimal basis dimension, for this dataset.

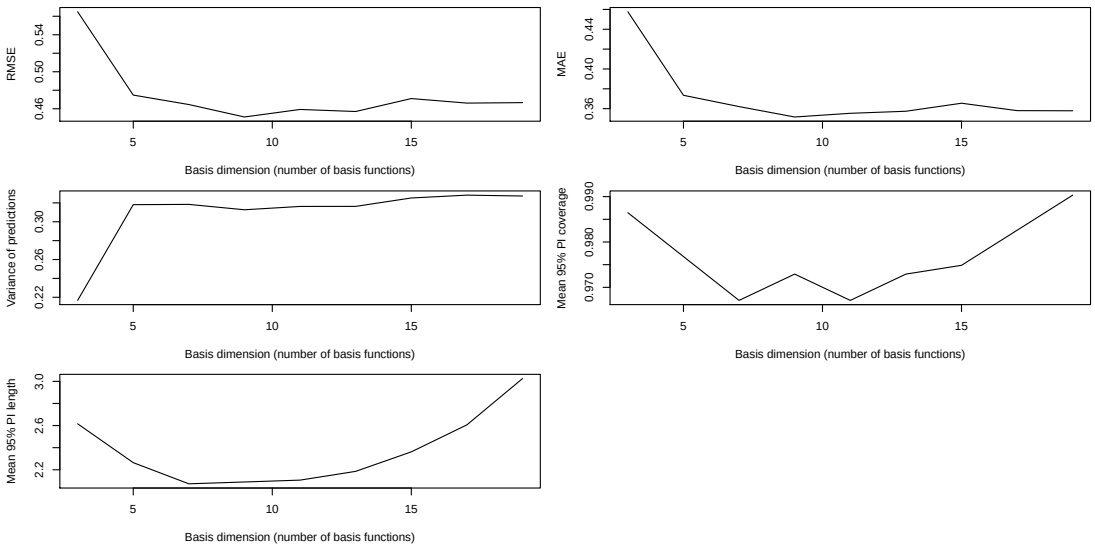

Figure 5.9: Plots of various summary statistics versus basis dimension for a leave-one-out cross-validation.

Comparison of the nonparametric and previously-fitted statistical downscaling models and choice of basis type

While the goal is to develop a model that fuses data of different spatiotemporal support, it is important that this model performs at least as well as the simpler models fitted previously, as otherwise its usefulness is limited. Once again, this is assessed empirically through a leave-one-out cross-validation. This allows a comparison between the two basis types and between the nonparametric downscaling model and its simpler counterpart (specifically model 3.4, the spatiotemporal statistical downscaling model with variance pooled over time), for which corresponding summary statistics have been calculated previously. These statistics are displayed in Table 5.1, for each model, for each of the two basis types, for their calculated optimal basis dimensions. The RMSE and MAE values demonstrate that nonparametric statistical downscaling can outperform traditional statistical downscaling, as long as an appropriate basis dimension is selected. It was mentioned earlier that either the B-spline or the Fourier basis are suitable, depending on whether periodicity is a reasonable assumption for the data. From these results, the Fourier basis leads to more accurate predictions than the B-spline

Model	RMSE	MAE	Variance of predictions	95% CI coverage	Mean 95% CI length
3.4	0.554	0.388	0.643	0.907	1.594
5.8 (B-spline, 49)	0.504	0.379	0.490	0.992	3.430
5.8 (Fourier, 9)	0.451	0.352	0.313	0.973	2.090

Table 5.1: Summary statistics for cross validation assessing nonparametric downscaling model performance, for the Lake Balaton data.

basis, although the difference is not so large that a B-spline basis is considered inappropriate. In fact, a B-spline basis can still be preferred, if the advice from ecological specialists is that an aperiodic relationship is expected over time. Although RMSE and MAE are important in comparing the models, other measures of performance are also important. The variances of predictions are lower for the nonparametric models and specifically for the Fourier basis, as expected due to the predictions varying more smoothly over time. The mean 95% credible interval (CI) coverage is close to 95% for all three models. Finally, the mean 95% credible interval length is greater for both nonparametric statistical downscaling models, so that estimates are less precise for the nonparametric models, which appears to be the only downside to these models, at least when fitting to this dataset. This section of analysis has demonstrated the ability of the nonparametric downscaling model to perform as well as previously fitted models. The next section focusses on understanding the patterns in predictions from the nonparametric model over space and over time.

Illustration of calibrations from the nonparametric statistical downscaling model

The previous section showed that the nonparametric statistical downscaling model is able to predict with ability equal to (and in fact slightly greater than that for) model 3.4. This section shows that the patterns in these predictions are sensible and relate to the patterns observed in the *in situ* data. All of the predictions from the nonparametric downscaling model are from

the model fitted with basis dimension 49 (for the B-spline basis) or 9 (for the Fourier basis). The patterns in the predictions over time are shown for the two bases, predicted at Lake Balaton location 1, based on data for locations 2 to 9 (see Figure 5.10). On each plot, the estimated smooth function for

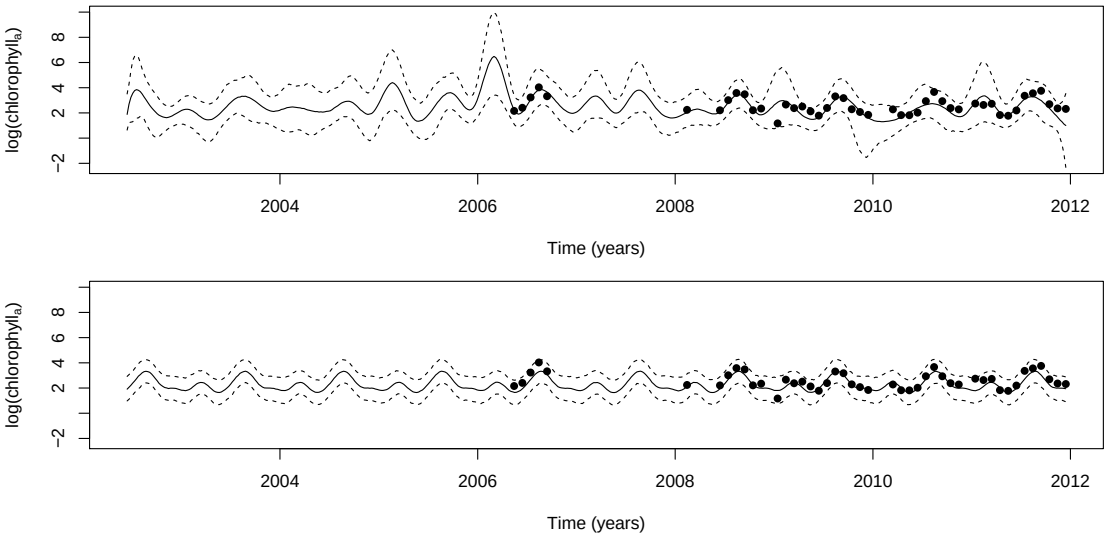

Figure 5.10: Predictions for Lake Balaton location 1 log(chlorophyll$_a$) data from a nonparametric downscaling model fitted to data for locations 2 to 9, using a B-spline basis of dimension 49 (top) and a Fourier basis of dimension 9 (bottom). Points are data, solid lines are predictions and dashed lines are 95% credible intervals.

location 1 is shown as a solid line, with 95% credible intervals in dotted lines. These predicted curves lie close to, and follow the main patterns in, the observed data for location 1, for both basis types. In addition to demonstrating the apparent good performance of the model at predicting patterns in the data, these plots also illustrate a possible reason for the improved performance of the Fourier basis-fitted model in comparison to the B-spline-fitted model. This is that, although the Fourier basis assumes that the pattern in log(chlorophyll$_a$) repeats every year, the B-spline basis has the problem of being more variable, probably due to the B-spline basis having to make do with few data over certain periods of time, while the Fourier basis uses data across years and hence smooths out excess variability.

The ability of the nonparametric statistical downscaling model to ad-

equately calibrate data over space is also of importance and can be illustrated through plots of predictions for March to May 2003, along with predictions from a traditional statistical downscaling model (3.4) and the original remotely-sensed data (see Figure 5.11). The top plots show the original data for these months, which are 3 of 115 for which remotely-sensed data are available. Predictions are made for the middle of each month. On plots for months that have available *in situ* data, these data are overlaid and surrounded by white circles. There are no *in situ* data for March 2003. For April, the available *in situ* data lie fairly close in value to their surrounding remotely-sensed data, while for May 2003, the available *in situ* data are higher than their surrounding remotely-sensed data, especially in the southwest of the lake, indicating the need for the calibration of the remotely-sensed data. This difference in the relationship between the *in situ* data and the remotely-sensed data between these months can be explained by differences in the quality of satellite readings, for example due to differences in cloud cover, or possibly to a change in the angle of the satellite, resulting in a change in the amount of the Earth's atmosphere that the light must pass through and therefore poorer calibration for one month. Predictions from the traditional downscaling model are good for April and May, in that they lie close to the *in situ* data values, but retain spatial characteristics from the remotely-sensed data, but no predictions can be made for March at all. The nonparametric model, in both its B-spline and Fourier forms, also fits predictions that lie close to the observed *in situ* data, with those from the B-spline-version model retaining more spatial structure from the remotely-sensed data. This may be a reason that a B-spline-based model would be preferred, although the Fourier-based model may be preferred if the average patterns over the years are of most interest. Finally, both versions of the nonparametric statistical downscaling model are able to predict at the month with only remotely-sensed data available, taking information across both space and time to calibrate the data, whereas the traditional model

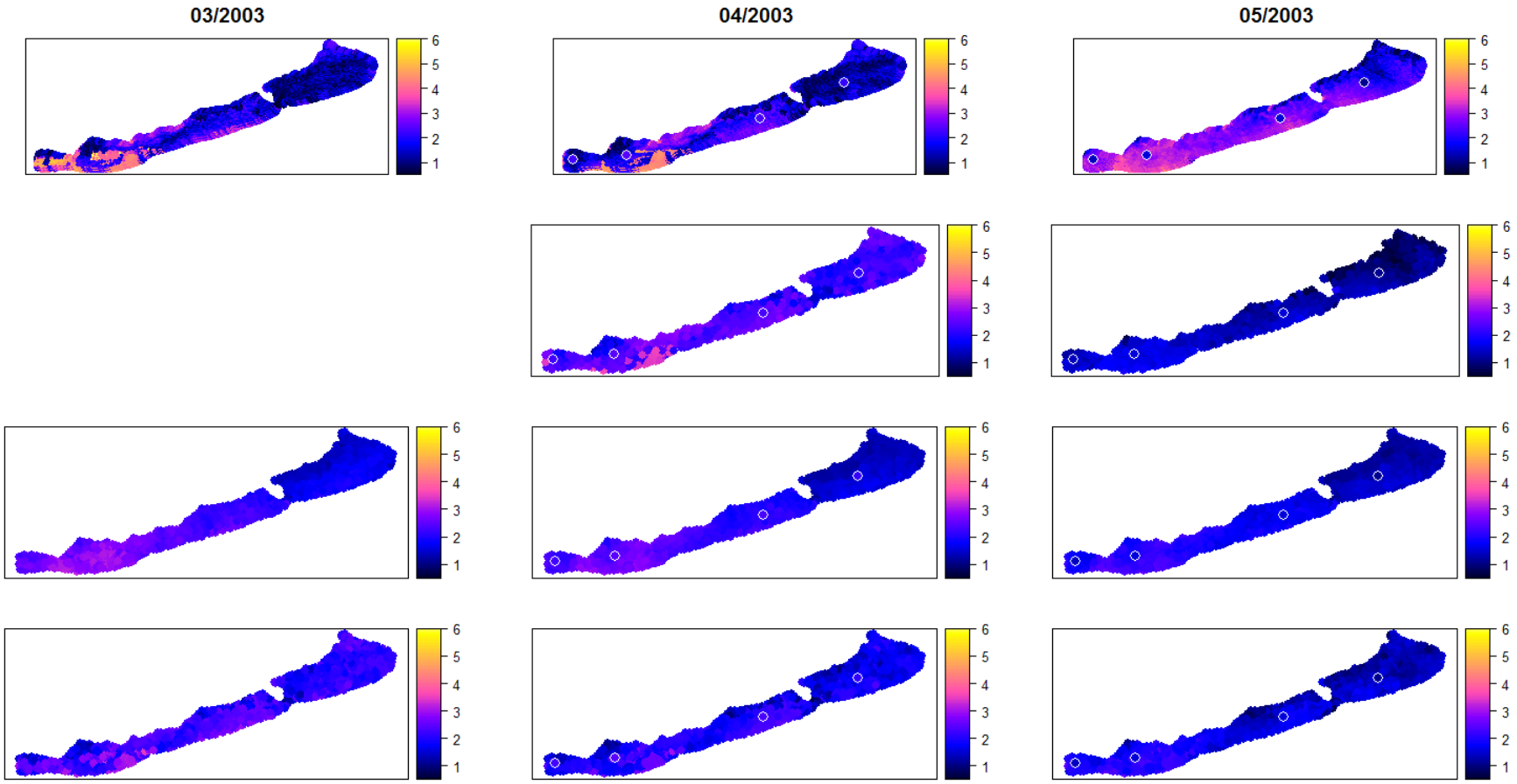

Figure 5.11: Plots of remotely-sensed data (top row) and predictions from statistical downscaling model 3.4 (second-top row) and nonparametric statistical downscaling model 5.8 with Fourier (second-bottom row) and B-spline (bottom row) bases, for March 2003 to May 2003, for Lake Balaton. *In situ* monthly mean data are overlaid, where available, surrounded by white circles. Prediction cannot be carried out for model 3.4 for March 2003, since no *in situ* data are available for that month.

cannot do this.

Plots of the standard errors for the traditional and nonparametric models (see Figure 5.12) show that standard errors are larger for the traditional model than for either version of the nonparametric model, for this particular dataset. This is not necessarily true for other datasets. For all models, the standard errors are much smaller than the variation over space within each month, which means that the model gives a useful understanding of patterns in $\log(\text{chlorophyll}_a)$ over space.

5.4.2 Application to data for Lake Erie

This subsection details the investigation of the data for Lake Erie, through the use of the nonparametric downscaling model.

The data for Lake Erie suffer from two quality problems. The first issue is that the *in situ* data are collected infrequently over time, with only twice-yearly sampling for the data collected by EPA and large gaps over winter of the data collected by LEC. The second issue is that the remotely-sensed data change in variability after the end of 2007, with temporal patterns much less clear after this time (see Figure 5.13, showing example patterns for two locations). It is unclear from the plot whether the *in situ* and remote sensing data follow the same patterns over time, since so few EPA data are available and since the available EPA *in situ* data may miss out one or more peaks in the data, apparently only covering a trough that does not appear so clearly in the remote sensing data. Despite the data issues, it is still of interest to apply the developed methodology to allow inference about the patterns in $\log(\text{chlorophyll}_a)$ over space and over time. The analysis proceeds, with the data from November 2008 onwards removed, giving 65 months of available data from June 2002 until October 2007.

Similarly to the Lake Balaton data, a leave-one-out cross-validation is carried out for both the B-spline and Fourier bases, to determine the empirically optimal basis dimension. Using formulae 5.9 and 5.10, the theoretically

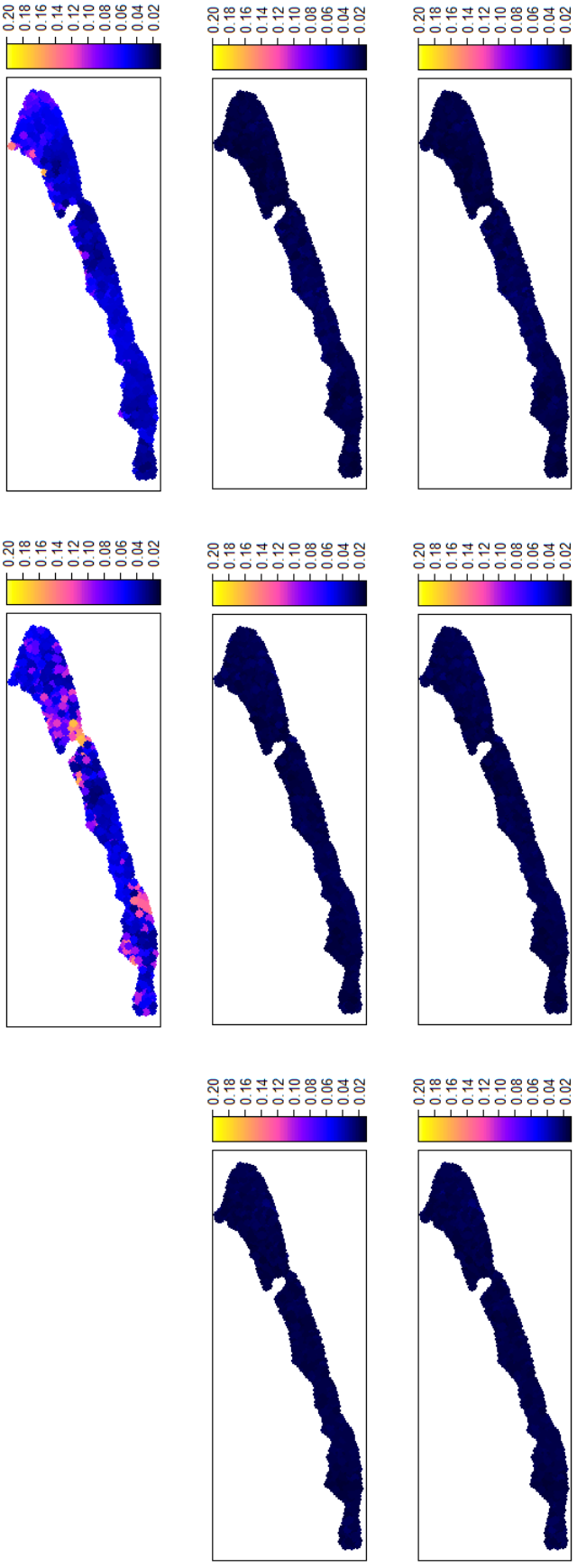

Figure 5.12: Plots of standard errors for predictions from statistical downscaling model 3.4 (top row) and nonparametric statistical downscaling model 5.8 with Fourier (middle row) and B-spline (bottom row) bases, for March 2003 to May 2003, for Lake Balaton.

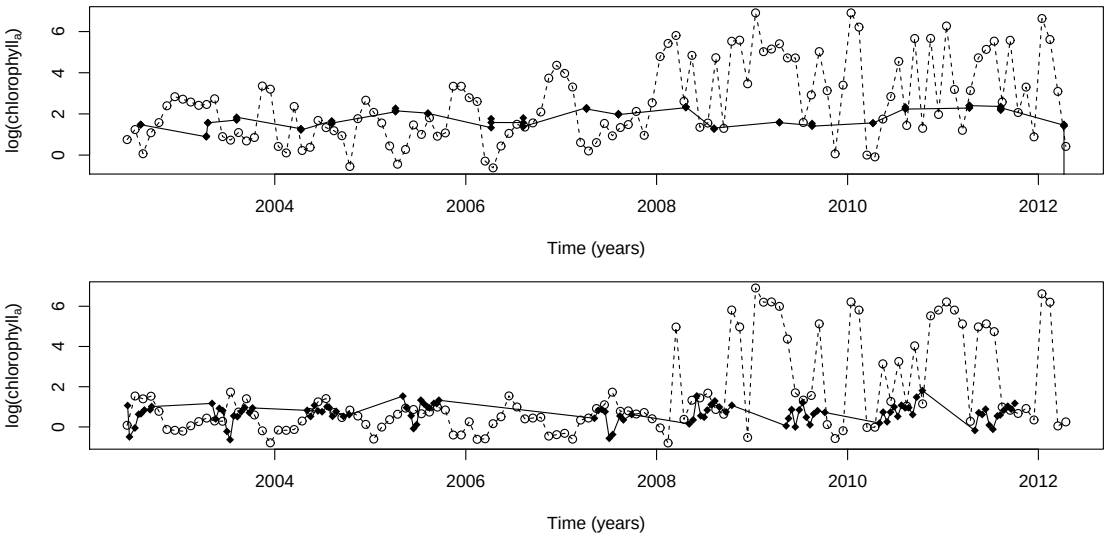

Figure 5.13: Plots of log(chlorophyll$_a$) over time for two locations in Lake Erie (top: EPA, location 1; bottom: LEC, location 21). Hollow circles are remotely-sensed data, while solid diagonal squares are *in situ* data.

optimal basis dimension is estimated as 14 or 24 for B-splines, depending on whether the number of peaks per year is 1 or 2, and 3 or 5 for the Fourier basis, again depending on the expected number of peaks per year. For each basis dimension, the convergence of the MCMC chains is checked using trace and density plots (see Figures B.22 and B.23 on pages 239 and 240 for an example using basis dimension 14), while diagnostic plots (see Figure B.44 on page 251 for an example using basis dimension 14) provide no evidence against the assumptions that the residuals have mean zero, are Normally distributed and are homoscedastic. Plots of summary statistics from the leave-one-out cross-validation are given in Figure 5.14, where the B-spline basis is fitted for a sequence between 5 and 29, increasing by 3 each time, and where the Fourier basis is fitted for the sequence of odd numbers between 3 and 19. The top plots show that RMSE reaches a minimum at 11, while MAE reaches a minimum at 14. Since 14 is the dimension required for at least one peak per year, it is selected over 11, giving both a theoretically and empirically reasonable choice. The mean prediction interval length increases with increasing basis dimension, but has very little difference be-

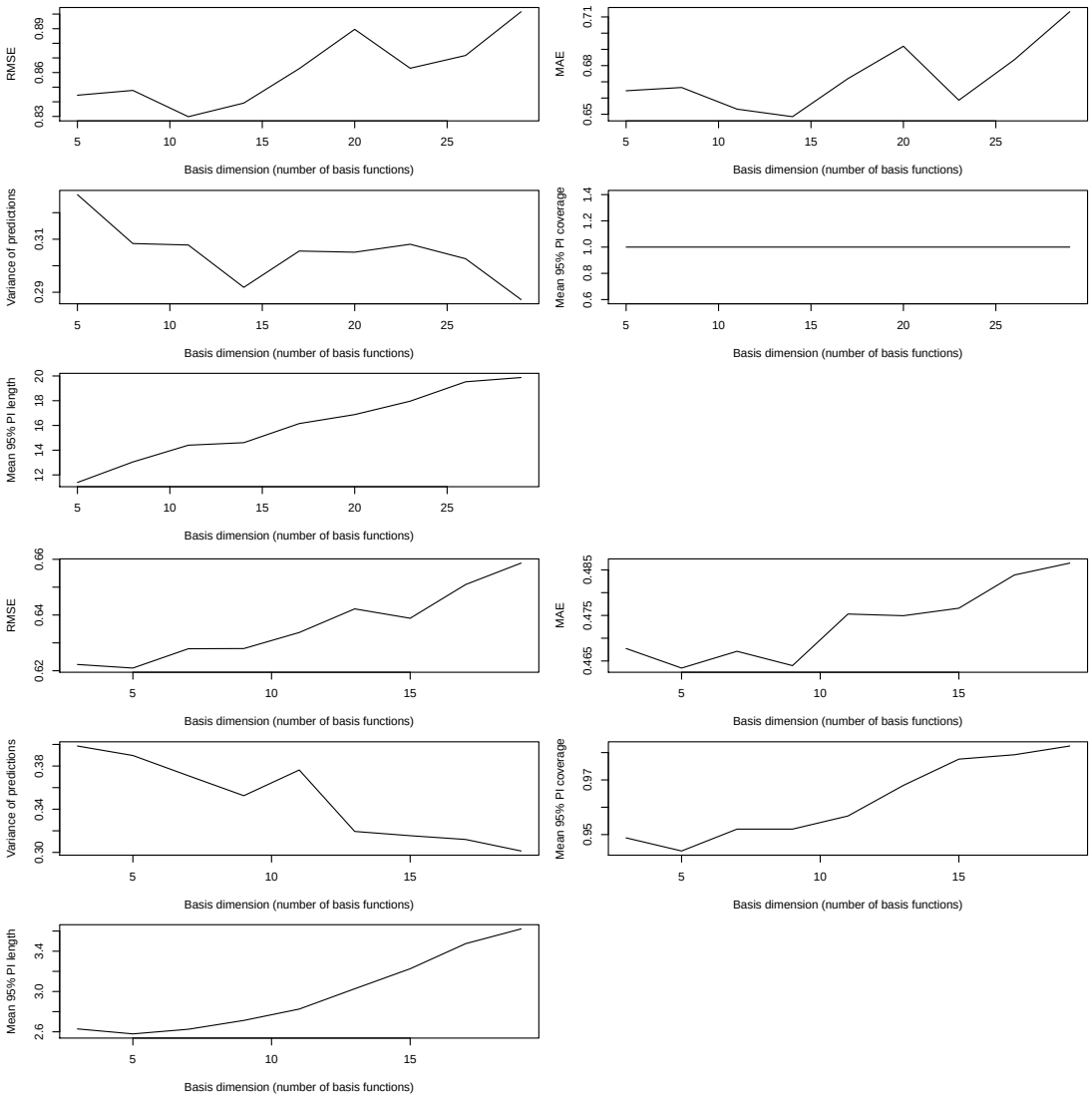

Figure 5.14: Summary statistics for Erie data from cross-validation using B-spline (top 5) and Fourier (bottom 5) basis.

tween dimensions 11 and 14. The bottom plots show that RMSE and MAE generally increase with increasing basis dimension, for the Fourier basis, with a minimum reached at dimension 5. Mean 95% credible interval length also increases with increasing basis dimension, with a minimum reached at dimension 5. Mean 95% credible interval coverage is slightly less than the nominal 95% for Fourier basis dimension 5, but not so much that alarm should be raised.

Choosing the basis dimensions to be 14 for B-splines and 5 for the Fourier

basis, the results from fitting the nonparametric model are compared to those from fitting model 3.4 (which is the spatiotemporal statistical downscaling model with pooling of estimates for the spatial variance parameters over time) (see Table 5.2). These results contrast with those for Lake Balaton,

Model	RMSE	MAE	Variance of predictions	95% CI coverage	Mean 95% CI length
3.4	0.624	0.452	0.512	0.818	2.097
5.8 (B-spline, 14)	0.840	0.649	0.292	1	14.607
5.8 (Fourier, 5)	0.621	0.463	0.390	0.944	2.579

Table 5.2: Summary statistics for cross validation comparing traditional and nonparametric downscaling models, for the Lake Erie data.

since the nonparametric downscaling model does not outperform the traditional model. In fact, RMSE and MAE are higher for the nonparametric downscaling model using the B-spline basis of optimal dimension 14 than for the traditional downscaling model. The mean 95% credible interval (CI) length is also extremely high for the model using the B-spline basis, compared to the other models. This may be due to the *in situ* data sparseness over time, meaning that understanding the smooth patterns in the *in situ* data is difficult. The model fitted using the B-spline basis leads to unusual patterns in the fitted smooth curve, with corresponding wide credible intervals. The nonparametric model using the Fourier basis, however, is found to perform very similarly to the traditional model, with a higher 95% credible interval coverage, which lies very close to the nominal 95% coverage. The improved performance of the model using the Fourier basis, compared to the model using the B-spline basis, may be due to the ability of the model to use data over all years, rather than relying on there being enough data in each year to understand the smooth pattern. These results show that the nonparametric statistical downscaling model is able to perform as well as the traditional model, even in the absence of densely-sampled *in situ* data with good temporal coverage. However, the requirement to check which basis type is most suitable is highlighted.

Plots of predictions over time (see Figure 5.15) show how overly smooth

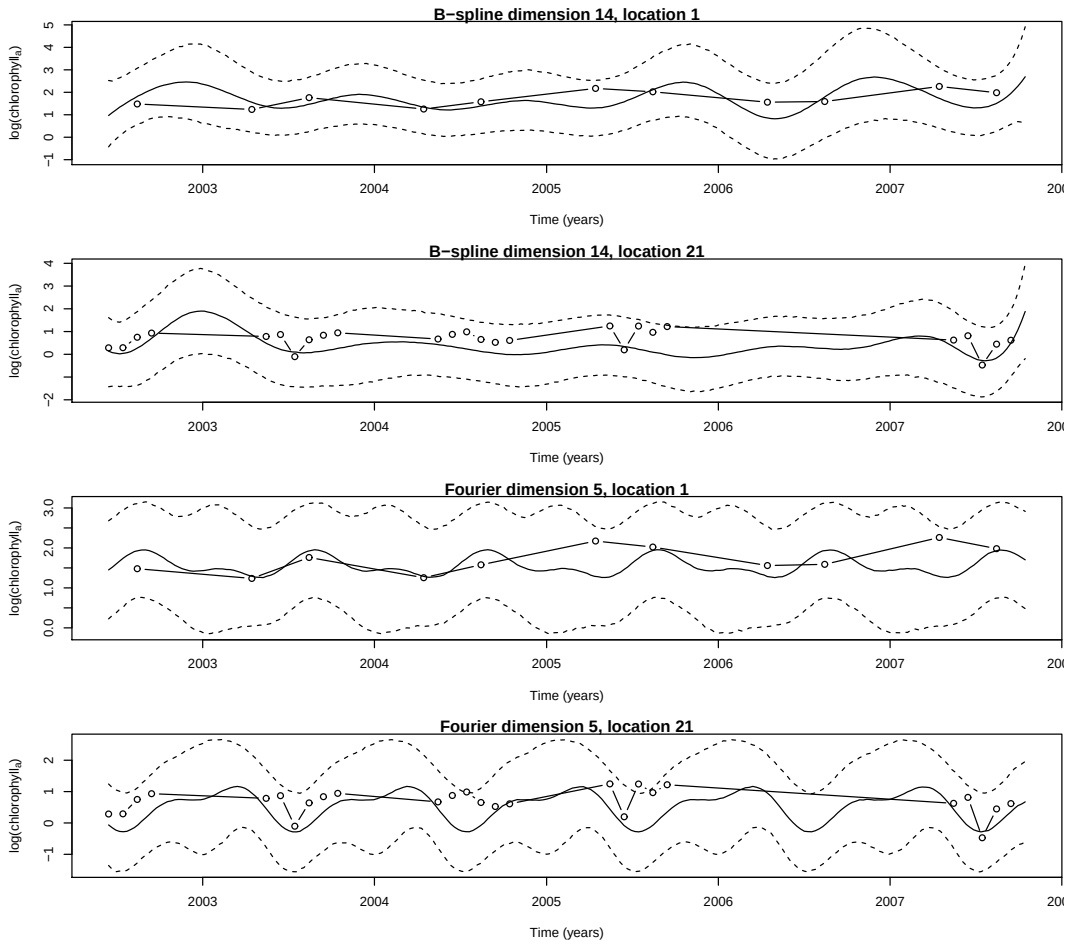

Figure 5.15: Plots of predictions from model 5.8 at Erie locations 1 and 21, for B-spline dimension 14 and Fourier dimension 5 bases. Points are data, solid lines are predictions and dashed lines are 95% credible intervals.

the predictions appear for the B-spline basis, as these do not follow the observed *in situ* data well where they are available. The predictions from the model with the Fourier basis do, however, follow the patterns in the observed *in situ* data for the LEC locations, as shown for the example in the bottom plot. The ability of the model using the Fourier basis to predict at the EPA locations is difficult to assess, due to the lack of *in situ* data over time for these locations.

Spatial predictions can also be produced for the nonparametric statistical downscaling model applied to Lake Erie (see Figure 5.16). These predictions

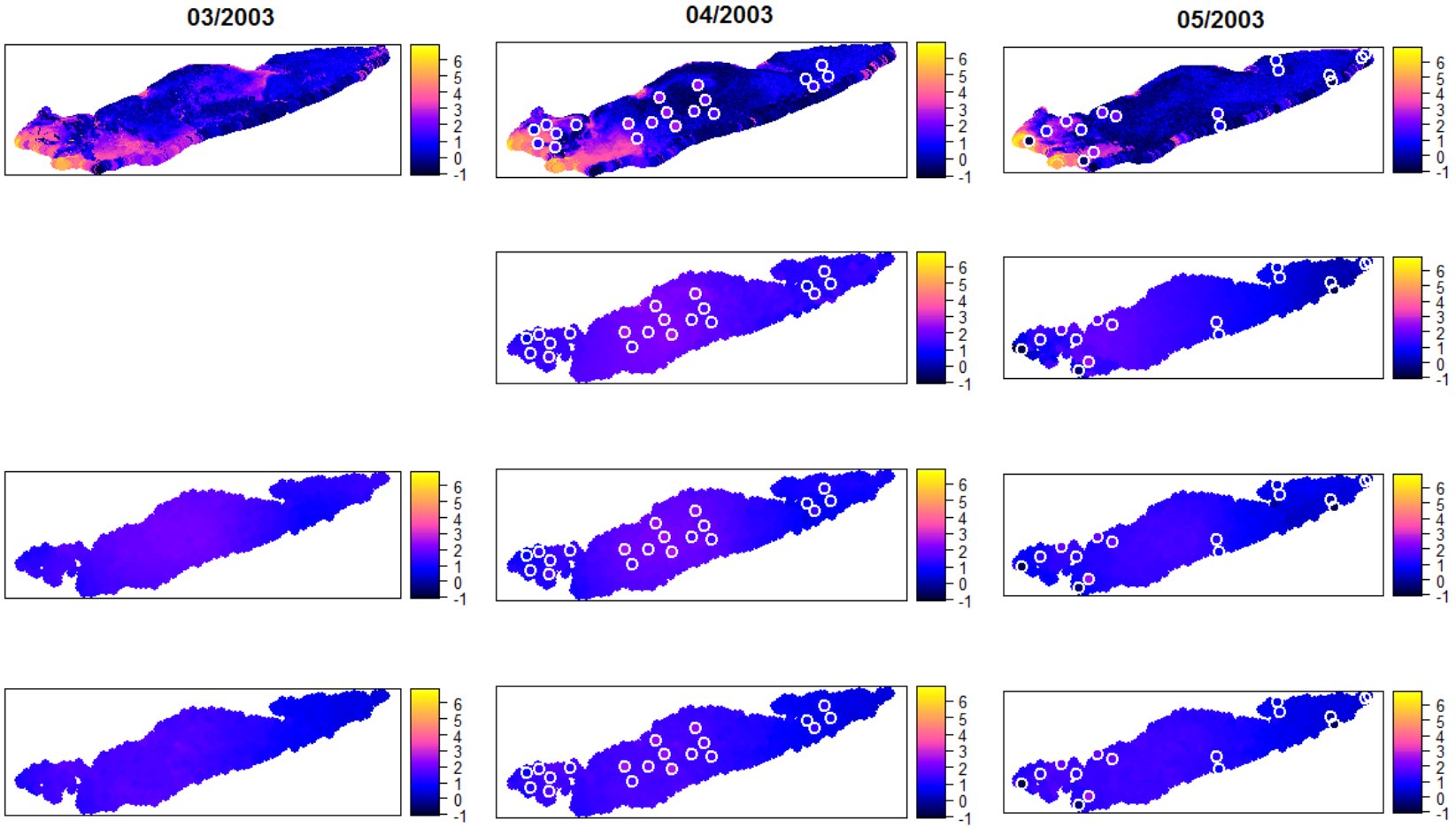

Figure 5.16: Plots of remotely-sensed data (top row) and predictions from statistical downscaling model 3.4 (second-top row) and nonparametric statistical downscaling model 5.8 with Fourier (second-bottom row) and B-spline (bottom row) bases, for March 2003 to May 2003, for Lake Erie. *In situ* monthly mean data are overlaid, where available, surrounded by white circles. Prediction cannot be carried out for the traditional downscaling model for March 2003, since no *in situ* data are available for that month.

are made at 1000 locations in the Lake, determined using a Delaunay triangulation in order to give the optimal coverage of the area, for the month-centres of March to May 2003. Also shown are predictions from the traditional model 3.4 and the original remotely-sensed data. *In situ* data are overlaid, surrounded by white circles. For April 2003, predictions for the nonparametric model fitted using both the Fourier and B-spline bases, and predictions from the traditional model, are similar, with lower values in the centre of the lake, compared to the east and west. None of the models take much information from the remotely-sensed data, since it does not relate very strongly to the *in situ* data for this lake. The same is true for May 2003. For March 2003, only remotely-sensed data are available, so predictions are only available from the nonparametric model, with similar patterns to those in the closest months with *in situ* data, such as April 2003. This demonstrates the strength of the nonparametric model, which is able to predict at any time, whether *in situ* data are available at that time or not. Standard errors are also plotted (see Figure 5.17). For this dataset, the standard errors are higher for the nonparametric downscaling model than for the traditional model, in contrast to the case for the Lake Balaton data. This might be due to the temporal sparseness of the *in situ* data, meaning that the predictions from a model that requires good temporal information are less certain than those from a model that does not.

5.5 Conclusions and further work

This section summarises the conclusions reached from the analysis and details further work that is required. In this chapter, a method for downscaling data of different spatiotemporal support was developed, incorporating methodology from the fields of both statistical downscaling and functional data analysis. The model calibrates remotely-sensed data, available on a grid-cell and monthly-average scale, using *in situ* data on a point-location

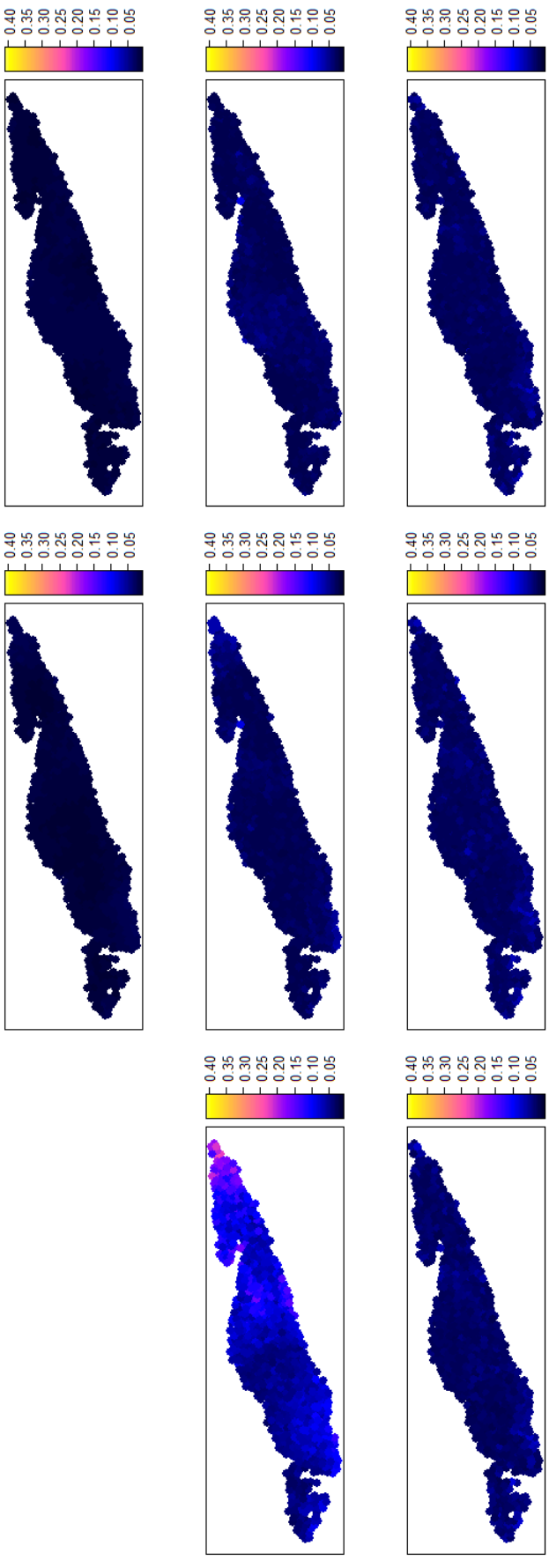

Figure 5.17: Plots of standard errors for predictions from statistical downscaling model 3.4 (top row) and nonparametric statistical downscaling model 5.8 with Fourier (middle row) and B-spline (bottom row) bases, for March 2003 to May 2003, for Lake Erie.

and point-time scale, through treating the observed data as observations of unknown smooth functions, fitted through use of basis functions. The two data types are related through their basis coefficients, which are modelled as spatially-varying, thus bringing in the spatial component of the model. The model is therefore able to predict at any point location and point time within the region and period of available data, whether *in situ* data are available for that point location or time, or not. The statistical novelty of the model is in the incorporation of functional data methodology within the statistical downscaling framework.

Two main choices to be made in fitting the model were discussed, namely the choices of basis type and basis dimension. The common basis types of the B-spline and Fourier basis were discussed, with the Fourier basis more suitable for periodic data and the B-spline basis able to model data more flexibly. The second important choice to make involved the selection of the basis dimension, i.e. the number of basis functions to use to model the smooth function. The basis dimension could be estimated, based on assumed or observed temporal patterns in the data, although it was found to be more appropriate to obtain a better estimate of the optimal basis dimension through cross-validation, in order to gain a good understanding of how well the model performs for various different basis dimensions. Another issue that was understood, through the analysis, was that either the B-spline or Fourier basis could be most appropriate for the data, if periodicity was a reasonable assumption for the data. In the data examined, the values of $\log(\text{chlorophyll}_a)$ were expected to have slightly differing patterns over the years, but the use of a Fourier basis still led to an improved model prediction accuracy, possibly due to the Fourier basis taking into account data over multiple years, compared to the reliance of the B-spline basis on possibly sparse data within each year separately.

The model was fitted to data for two lakes, one of which (Lake Balaton) had much better quality data available than the other (Lake Erie). For each

lake respectively, using the same dataset for all models, it was found that the nonparametric model outperformed the traditional model for the Lake Balaton data, but not for the Lake Erie data. For the Lake Balaton data, the nonparametric model provided improved estimates, even when these estimates were only required at times of *in situ* data collection. Along with the ability to predict at any timepoint, the nonparametric model has the benefit of being computationally efficient, with the spatial downscaling part of the model using parameters of dimension determined by the basis dimension, rather than by the possibly higher dimensional data. The Lake Erie data, however, present the potential drawback to the model, which is that the *in situ* data really need to give at least some idea of the temporal patterns in the data, which is not the case for Lake Erie. This case can be presented as a more typical case for lakes around the world, for which *in situ* data sampling is expensive and so carried out only infrequently. Even with the temporally sparse *in situ* data, however, the nonparametric downscaling model using the Fourier basis performed as well as the traditional model at predicting the values of the *in situ* data. This analysis highlighted the need to determine the most appropriate basis type from an application to the data, rather than making an assumption *a priori*, since the Fourier basis was found to greatly outperform the B-spline basis for the Lake Erie dataset.

For Lake Balaton, it can be concluded that the values of $\log(\text{chlorophyll}_a)$ are generally higher in the southwest of the lake (around 3.5 units for May 2003), while the values in the northeast of the lake are generally the lowest (closer to 1 unit for May 2003). The models estimated that there was a two-peaks per year pattern in $\log(\text{chlorophyll}_a)$ data for lake Balaton, with the highest values in summer and a secondary smaller peak in spring. The inferences made using the nonparametric statistical downscaling model agree with the knowledge in the literature on patterns of $\log(\text{chlorophyll}_a)$ data over both space and time for Lake Balaton.

For Lake Erie, there was a weaker relationship between the *in situ* and

remotely-sensed data. The model was unable to derive much information on temporal patterns from the *in situ* data, which could have led to the resulting surfaces of spatial predictions not reflecting many of the spatial patterns in the remotely-sensed data. It was clear, however, that the lake centre often had the highest levels of $\log(\text{chlorophyll}_a)$.

For both lakes, standard errors were small in comparison to the changes in $\log(\text{chlorophyll}_a)$ levels estimated over space and over time, so that a precise understanding of the patterns in the data is gained from fitting the model.

The nonparametric statistical downscaling model allows the production of plots of calibrated $\log(\text{chlorophyll}_a)$ data over both space and time, as shown throughout this chapter. This allows water quality investigators to understand how patterns change in the lake over space and time. The model addresses the spatiotemporal support issue identified in previous chapters and has been shown to perform as well as previously described traditional models. This statistically novel technique should be of practical use to water quality researchers in order to better understand lake health variation over space and time.

Chapter 6

Conclusions

This thesis has developed methodology for the data fusion of *in situ* and remotely-sensed data, accounting for the challenges that these data of different spatiotemporal support present. Motivated by the GloboLakes project and the data provided for $\log(\text{chlorophyll}_a)$ in Lake Balaton and the Great Lakes of North America, the need for data fusion of *in situ* and remotely-sensed data was identified, so that the accuracy from the *in situ* data could be combined with spatial and temporal information from the remotely-sensed data.

Statistical downscaling models, which fuse data of different spatial support, were developed and applied in the novel application area of data for $\log(\text{chlorophyll}_a)$, producing fully calibrated spatial surfaces, with associated comprehensive uncertainty estimates. In order to improve the accuracy of model predictions, bivariate and multiple-lakes statistical downscaling models were developed, so that information on the relationship between the *in situ* and remotely-sensed data could be shared between variables and between lakes. For all of these statistical downscaling methods, the data must be available on the same temporal support. Nonparametric statistical downscaling was developed, to allow the data fusion of data with different spatial and temporal support, filling a gap in the literature relating to temporal support in statistical downscaling.

The following sections provide more detail on the methodology, applications and conclusions from each chapter of this thesis.

6.1 Chapter 1: Introduction and background

Chapter 1 introduced the background to the research, the GloboLakes project and the data that are available for analysis. The importance of the variables chlorophyll$_a$, total suspended matter and lake surface water temperature were described and the main research aims and objectives were discussed.

The literature review detailed the methodology from the data fusion literature that are relevant to this research. For chlorophyll$_a$, methods such as linear modelling, a pixel-by-pixel algorithm, a genetic programming model and wavelet multiresolution analysis were among those used, while for air quality data, Bayesian melding, fixed rank kriging and statistical downscaling were amongst those used. Statistical downscaling was identified as an appropriate method for analysis of the log(chlorophyll$_a$) data, based upon the similarity to the change-of-support problem to which it was applied in the air quality data to that of the *in situ* and remotely-sensed log(chlorophyll$_a$) data.

The relevant methodology from geostatistics, nonparametric smoothing and Bayesian modelling was then presented, to give the basis upon which methods in the later chapters depend.

6.2 Chapter 2: Initial spatial and temporal analysis of data

Patterns in the data for Lake Balaton were identified through the use of exploratory plots. These showed that log(chlorophyll$_a$) and log(total suspended matter) had a positive relationship and that there were cyclical pat-

terns in each of log(chlorophyll$_a$), log(total suspended matter) and temperature over time. For log(chlorophyll$_a$), there was a clear two-peaks per year pattern, with a small peak in spring and a larger peak in summer, while for temperature, the pattern was a strong one-peak-per-year pattern, with the highest temperatures in summer. The cyclical pattern for log(total suspended matter) was less strong, but there were generally higher values in summer than in winter.

Mixed effects models were fitted to the log(chlorophyll$_a$) and log(total suspended matter) data. Log(chlorophyll$_a$) was modelled using a pattern of two peaks per year, with latitude as a fixed effect, providing evidence of a strong pattern over time and of variation over space, accounting for the effects of log(total suspended matter) and temperature. Similarly, log(total suspended matter) was modelled using a pattern of two peaks per year, but had no significant effect of longitude or latitude. There was, however, a large estimated contribution of the random effect of location, providing evidence of patterns over both space and time for this variable.

Kriging was used to interpolate the remotely-sensed temperature data for Lake Balaton spatially, in order to get an improved understanding of the spatial patterns in temperature for each month. Universal kriging, which allows for a trend in the mean level across space, was carried out separately for each month in the dataset and predictions were made over a dense grid covering the lake. The resulting plots of the fitted surface showed that different patterns of lake surface water temperature were observed during different months, suggesting that the changes in temperature over space were small, on average, in comparison with the changes over time.

Principal component analysis (PCA) was carried out on the remotely-sensed temperature data, to identify the common patterns over space and time, which could enable reduction of the spatial or temporal dimensions of the data. Two modes of PCA were applied, namely S-mode and T-mode, which apply the PCA to the matrix of times versus locations and its trans-

pose, respectively. S-mode aims to isolate groups of locations that covary similarly, while T-mode aims to isolate the groups of times with similar spatial patterns. The results of the S-mode PCA suggested that almost all locations covary similarly, so that the first component explained almost all of the variance in the data. The results of the T-mode PCA suggested that there were no groups of times with similar spatial patterns, since a large number of components were required, in order to explain a high proportion of variance in the data.

Generalised additive models (GAMs) were used to investigate the effect of smoothing data on improving the relationship between the *in situ* and remotely-sensed data. The remotely-sensed data were regressed on smooth functions of latitude, longitude, year and day of the year, with independent, Normally distributed random errors. For each time, predictions were made from the model at the *in situ* data locations. Predictions were also made, by taking the value of the remotely-sensed data at the cell nearest to each *in situ* location. These two sets of predictions were compared through the root mean squared error, showing that the predictions from the GAM were more accurate. This suggests that smoothing should be considered, when modelling the relationship between the *in situ* and remotely-sensed data.

6.3 Chapter 3: Statistical downscaling

The objective of Chapter 3 was to develop and apply statistical downscaling models for the fusion of the *in situ* and remotely-sensed log(chlorophyll$_a$) data. Firstly, a spatial statistical downscaling model was developed, which regressed the *in situ* data on the remote sensing data for the corresponding grid cells, with smoothly spatially-varying intercept and slope coefficients, within a hierarchical Bayesian modelling framework. Predictions were made over the lake and a plot of the resulting fused data showed that the remotely-sensed data had been calibrated with the *in situ* data, so that the resulting

surface of predictions was a fully calibrated surface, reflecting the spatial patterns in the data.

The spatial statistical downscaling model must be fitted to data for each month separately, but a spatiotemporal statistical downscaling model was developed that was fitted to the data for all times in the dataset at once. This model was adapted to share information over time, by using the same spatial variance parameters for all times in the data, which led to improved estimates of these parameters and hence improved predictions overall.

Models that fitted correlation over time, using an autoregressive process and temporal covariance matrices, were fitted to the data. These led to a slight improvement in the accuracy of predictions made from these models, compared to those from the previous spatiotemporal model, suggesting that accounting for smoothing over time may be appropriate.

These models were fitted to the data for both Lake Balaton, which has data for 17 months corresponding to 9 *in situ* locations, and Lake Erie, which has data available for 20 months corresponding to 10 *in situ* locations, but temporally-sparse, spread out over 10 years.

For the novel application to $\log(\text{chlorophyll}_a)$ data, the spatiotemporal statistical downscaling models were able to fuse *in situ* and remotely-sensed data successfully, resulting in spatial surfaces of fused data that were calibrated with the *in situ* data, with associated uncertainty measures.

6.4 Chapter 4: Bivariate and multiple lakes downscaling

Chapter 4 fitted bivariate and multiple-lakes downscaling models to data for Lake Balaton and the Great Lakes, respectively. Bivariate and multiple-lakes downscaling models were both motivated by the assumption that sharing information can improve the accuracy of predictions. Bivariate downscaling shares information between two variables, while multiple-lakes down-

scaling shares information between neighbouring lakes.

6.4.1 Bivariate statistical downscaling

Bivariate statistical downscaling was fitted as an extension to the spatial statistical downscaling model, with two variables modelled simultaneously. The correlation structure between the two variables was built in through a correlated error structure. A second model additionally shared spatial variance parameters between variables. It was found that the second model resulted in improved accuracy of predictions, in comparison to those from fitting a separate model to each variable separately, while the first model with correlated errors showed no improvement.

6.4.2 Multiple lakes statistical downscaling

Since groups of nearby lakes are likely to share similar patterns of ecological variables, information can be shared between lakes, in order to improve the estimation of the relationships between the *in situ* and remotely-sensed data for each lake.

The multiple-lakes downscaling model was developed in the same framework as the previous statistical downscaling models, so that the *in situ* data are regressed on the remotely-sensed data. The model fits an overall intercept parameter, with a lake-specific intercept parameter and a spatially-varying intercept parameter. Likewise, the model fits an overall slope parameter, a lake-specific slope parameter and a spatially-varying slope parameter. The lake-specific parameters are constrained to have mean zero and account for lake-specific changes in the intercept and slope parameters, which may occur due to differences in the hydrological or ecological structures between different lakes. The spatially-varying parameters were either assumed to vary smoothly over each lake separately, or to vary smoothly over all lakes.

The multiple-lakes model was fitted to the data for the Great Lakes and

it was found that there were convergence problems for some parameters. Simplified versions of the model were fitted, with some of the parameters removed, which solved the convergence problems. The results from fitting the models were compared to those from fitting a separate statistical downscaling model to each lake, and to those from fitting a statistical downscaling model to data for all lakes at once, without any lake-specific parameters. It was found that fitting a statistical downscaling model over all lakes at once resulted in predictions that were equally accurate, compared to those from the best of the multiple-lakes models that specifically accounted for lake-specific effects, so that the lake-specific models were deemed to be unnecessary for the Great Lakes data.

It would be of interest to apply the multiple-lakes downscaling models to data for different lakes around the world, to investigate whether there are circumstances where lake-specific parameters are required. This is, however, beyond the scope of the current research, and is suggested here as a topic for future work.

6.5 Chapter 5: Nonparametric statistical downscaling

Chapter 5 presented nonparametric statistical downscaling, which is the main methodological development of this thesis. Using methodology from both functional data analysis and statistical downscaling, this chapter described the motivation and development of the nonparametric statistical downscaling model, which fuses data of both different spatial and temporal support, addressing the temporal change-of-support problem that the existing statistical downscaling models do not account for.

The gap in the literature of accounting for changes in temporal support was identified. Since the *in situ* data for the Lake Balaton and the Great Lakes datasets were sampled at point times, while the corresponding

remotely-sensed data were available for monthly averages, the change-of-support had to be accounted for. This was accomplished through incorporating functional data analysis within the statistical downscaling framework. Specifically, this involved treating the *in situ* and remotely-sensed data at each location as observation of smooth curves over time. These curves were estimated using basis functions and the basis coefficients of the *in situ* and remotely-sensed data were then related through a spatially-varying coefficients regression. The curve estimation and spatially-varying coefficients regression were fitted within a Bayesian hierarchical model, so that errors were carried through the model to give a realistic, comprehensive uncertainty estimate for each model prediction.

There are four main benefits to this model, for the Lake Balaton and Lake Erie data. The most important is that it accommodates the spatiotemporal change of support between the *in situ* and remotely-sensed data, resulting in data fusion of the two datasets, with no data aggregation required. This leads to the second benefit, which is that predictions can be made at any time and any location within the lake, whether *in situ* data were available for that time and location or not. This is a benefit over the statistical downscaling models from previous chapters, which were only able to predict over space within a month for which some *in situ* data were available and not for months for which no *in situ* data were sampled. The third benefit is dataset specific and is that the nonparametric statistical downscaling model resulted in slightly more accurate predictions than those from spatiotemporal statistical downscaling models from earlier chapters. This third benefit is a good addition for these datasets, but even without it, this model may be preferred, since it addresses the problem of changing spatiotemporal support that other models have ignored. The fourth benefit is that the method is computationally efficient, since the spatially-varying coefficient regression part of the model is fitted using the basis coefficients, which are likely to be of a lower dimension that the data themselves.

6.6 Discussion, limitations and future work

This thesis aimed to develop methodology for data fusion of *in situ* and remotely-sensed data, accounting for the spatiotemporal change of support between the data types. After preliminary modelling identified the spatial and temporal patterns in the data, statistical downscaling was identified as an appropriate method for the fusion of $\log(\text{chlorophyll}_a)$ data, which is the type of data that this thesis has focussed on. Fitted in the Bayesian hierarchical model framework, a model with spatially-varying coefficients was fitted to data for each month in the dataset, resulting in a fully calibrated spatial surface of fused data for each month, with associated measures of uncertainty. The model accounted for the spatial change of support between the point-location *in situ* data and the grid-cell-scale remotely-sensed data, but ignored the temporal change of support, so that the *in situ* data had to be aggregated over each month. The model was extended to a spatiotemporal version, sharing information over time and improving the accuracy of the predictions made. Versions of the model that included smoothing over time were also developed. These models slightly improved the accuracy of the predictions, suggesting that smoothing over time was appropriate for gaining a better understanding of the data. Bivariate models and multiple lakes models were developed, based upon the assumptions that sharing information between variables and between neighbouring lakes would improve the accuracy of predictions from the models. It was found that sharing information between two variables, which were modelled simultaneously, did improve the performance of the model. However, it was found that multiple-lakes downscaling models with lake-specific parameters were unnecessary, since a simpler statistical downscaling model performed as well, without accounting for lake-specific effects. Finally, the temporal change of support problem was tackled, through the development of the nonparametric statistical downscaling model, which is a novel statistical development that incorporates the

estimation of the *in situ* and remotely-sensed data at each location as smooth functions over time. The nonparametric statistical downscaling model was found to accomplish the main objective of the research, through the fusion of data of different spatiotemporal support, to produce a fused dataset that incorporates the accuracy from the *in situ* data and the spatial and temporal information from the remotely-sensed data.

The main limitation of the earlier work on statistical downscaling in this thesis and of the methods in the literature, such as those of Berrocal et al. (2010*b*) and Berrocal et al. (2010*a*), is that the temporal change of support is ignored. It is important for the application of lake water quality data that this is addressed, since the *in situ* data for variables such as chlorophyll$_a$ and total suspended matter are often only available at irregular sampling times, while their corresponding remotely-sensed data are often available for monthly averages. This limitation and the resulting gap in the literature were addressed by the novel development of the nonparametric statistical downscaling model, allowing data to be input without any aggregation and predictions to be made for any time.

All of the statistical downscaling models investigated in this thesis make use of Inv-Ga$(0.001, 0.001)$ as prior distributions for their variance parameters. This distribution has been used extensively in the literature for this purpose, but has also been criticised by some authors. Since the main aim of this work is to present novel methodology, the investigation of whether alternative prior distributions would be more suitable is left for future work. An initial comparison of the first statistical downscaling model did not suggest that the results were particularly sensitive to a change in the prior distributions for the variance parameters, giving no cause for concern.

A limitation of all models developed in this thesis is the computational complexity, which increases with the number of times and locations in the data, and the number of times and locations at which to predict. Since the models are fitted in the Bayesian framework, through Gibbs sampling, this

computational complexity can become quite large. Steps were taken to reduce the computational complexity of the models, including coding them in `Rcpp` and identifying the parts of the full conditional posterior distributions that could be simplified. Additionally, the nonparametric statistical downscaling model carries out the spatially-varying coefficient regression part of the model on the basis coefficients, rather than the possibly higher dimensional data, keeping the computational complexity low. A Delaunay triangulation was carried out for each lake, to determine the locations at which to predict to give an optimal coverage of the lake surface. However, the maximum number of locations at which to predict was chosen to be around 1000 for each lake, since otherwise the memory used in the model-fitting process became too high. It would be useful to be able to predict at a larger number of locations, in order to gain an improved spatial understanding of the data for very large lakes. Suggestions for future work, to address this limitation, include investigating alternative methods for improving the mixing of the MCMC chains, so that convergence is reached after a smaller number of iterations, and investigating how the thinning of the MCMC chains can be used to improve the computational efficiency of the models. Alternatively, the possibility of implementing the models in the INLA framework could be investigated, since INLA avoids MCMC altogether (Rue et al. 2009, Blangiardo & Cameletti 2015).

For the nonparametric statistical downscaling model in particular, a suggestion for future work is an investigation of how the estimation of the basis dimension can be incorporated within the model, as this must currently be chosen outwith the model. This would be a challenge, since the same basis dimension must be used for the *in situ* and remotely-sensed data corresponding to each *in situ* location. A benefit to this would be that the number of decisions to be made by the modeller would be reduced, making the model more attractive to a non-statistician. Additionally, this would ensure that the prediction errors took into account the uncertainty associated with estimat-

ing the optimal basis dimension, since otherwise the dimension is assumed to be fixed and known.

A further suggestion for future work is the application of the models to different datasets. The models from Chapters 3 and 4 were investigated using datasets with few *in situ* sampling locations and few months for which data were available for all locations. Although the small number of *in situ* sampling locations is not unusual for lake water data, it would be of interest to apply the models to datasets with more frequent sampling over time at each location, to investigate how the performances of the models compared to their performances when applied to the Lake Balaton and Lake Erie data. A final important point to note about the models is that, although they were developed for the application of fusing $\log(\text{chlorophyll}_a)$ data, there is no reason that the models cannot be applied to data from different spheres of research. For example, the nonparametric statistical downscaling model can be applied to fuse point-location data with grid-cell data, each with either a point-time or averaged-time scale, for any variable.

Appendix A

Derivation of full conditional posterior distributions

This appendix presents the derivations of the full conditional posterior distributions for models 3.3, 3.4, 4.1 and 5.8. These full conditional distributions are required for Gibbs sampling and are the posterior distributions of each model parameter, given the values of every other parameter in the model. They are used in Gibbs sampling to iteratively update the values of each parameter, given the value of each other parameter at the previous iteration (Gelman et al. 2014).

In the following sections, the notation $f(x|\cdot)$ indicates the posterior distribution of the variable x, given the values of all of the other variables in the model.

A.1 Spatiotemporal statistical downscaling model 3.3

The spatiotemporal statistical downscaling model 3.3 is written as follows:

$$\begin{aligned} y_{ji} &\sim \mathrm{N}(\alpha_{ji} + \beta_{ji}x_{ji}, \sigma^2_{\varepsilon_j}), \\ \boldsymbol{\alpha}_j &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma^2_{\alpha_j}\exp(-\phi_\alpha \mathbf{D})\big), \\ \boldsymbol{\beta}_j &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma^2_{\beta_j}\exp(-\phi_\beta \mathbf{D})\big), \\ (\sigma^2_{\alpha_j})^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\ (\sigma^2_{\beta_j})^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta), \\ (\sigma^2_{\varepsilon_j})^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \end{aligned} \tag{A.1}$$

where y_{ji} is the value of *in situ* data for $\log(\text{chlorophyll}_a)$ at time j ($j = 1, \ldots, t$) and location i ($i = 1, \ldots, n$), x_{ji} is the value of remote sensing data at time j for the grid cell containing *in situ* location i, $\mathbf{D}$ is the $n \times n$ matrix of distances between *in situ* locations, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jn})^{\mathrm{T}}$ is the vector of intercept parameters for time j, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jn})^{\mathrm{T}}$ is the vector of slope parameters for time j, $\mathbf{0}$ is an n-length vector of zeros and $\mathbf{1}$ is an n-length vector of ones.

The probability density functions for the data distribution and for the distributions of each parameter in the model can be written as follows:

$$\begin{aligned} f(y_{ji}) &= \frac{1}{\sqrt{2\pi\sigma^2_{\varepsilon_j}}} \exp\left(-\frac{1}{2\sigma^2_{\varepsilon_j}}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right), \\ f(\boldsymbol{\alpha}_j) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2_{\alpha_j}\exp(-\phi_\alpha \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_j - \mathbf{0})^{\mathrm{T}}\big(\sigma^2_{\alpha_j}\exp(-\phi_\alpha \mathbf{D})\big)^{-1}(\boldsymbol{\alpha}_j - \mathbf{0})\right), \\ f(\boldsymbol{\beta}_j) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2_{\beta_j}\exp(-\phi_\beta \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_j - \mathbf{1})^{\mathrm{T}}\big(\sigma^2_{\beta_j}\exp(-\phi_\beta \mathbf{D})\big)^{-1}(\boldsymbol{\beta}_j - \mathbf{1})\right), \end{aligned}$$

$$\begin{aligned} f\big((\sigma^2_{\alpha_j})^{-1}\big) &= b_\alpha^{a_\alpha}(\sigma^2_{\alpha_j})^{-1(a_\alpha - 1)}\exp\big(-b_\alpha(\sigma^2_{\alpha_j})^{-1}\big)/\Gamma(a_\alpha), \\ f\big((\sigma^2_{\beta})^{-1}\big) &= b_\beta^{a_\beta}(\sigma^2_{\beta_j})^{-1(a_\beta - 1)}\exp\big(-b_\beta(\sigma^2_{\beta_j})^{-1}\big)/\Gamma(a_\beta), \\ f\big((\sigma^2_{\varepsilon})^{-1}\big) &= b_\varepsilon^{a_\varepsilon}(\sigma^2_{\varepsilon_j})^{-1(a_\varepsilon - 1)}\exp\big(-b_\varepsilon(\sigma^2_{\varepsilon_j})^{-1}\big)/\Gamma(a_\varepsilon), \end{aligned}$$

for $j = 1, \ldots, t$ and $i = 1, \ldots, n$. For calculation of the full conditional posterior distributions of $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$, it is useful to know the data distribution in vector form:

$$\begin{aligned} f(\mathbf{y}_j) &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma^2_{\varepsilon_j})^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2_{\varepsilon_j}} \sum_{i=1}^{n} \big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma^2_{\varepsilon_j}\mathrm{I}_n|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Big(\mathbf{y}_j - (\boldsymbol{\alpha}_j + \mathbf{X}_j\boldsymbol{\beta}_j)\Big)^{\mathrm{T}}(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}\Big(\mathbf{y}_j - (\boldsymbol{\alpha}_j + \mathbf{X}_j\boldsymbol{\beta}_j)\Big)\right), \end{aligned}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and:

$$\mathbf{X}_j = \begin{pmatrix} x_{j1} & 0 & \cdots & 0 \\ 0 & x_{j2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x_{jn} \end{pmatrix}$$

is an $n \times n$ diagonal matrix with the vector of remotely-sensed data at time j, $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})^{\mathrm{T}}$, as its diagonal.

The full conditional posterior distributions are calculated as follows:

$$\begin{aligned} f((\sigma^2_{\varepsilon_j})^{-1}|\cdot) &\propto (\sigma^2_{\varepsilon_j})^{-1(a_\varepsilon - 1)}\exp\big(-b_\varepsilon(\sigma^2_{\varepsilon_j})^{-1}\big) \\ &\quad \times (\sigma^2_{\varepsilon_j})^{-1(\frac{n}{2})}\exp\left(-(\sigma^2_{\varepsilon_j})^{-1}\left(\frac{1}{2}\sum_{i=1}^{n}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right)\right) \\ &\quad \propto \mathrm{Ga}\left(a_\varepsilon + \frac{n}{2}, b_\varepsilon + \frac{1}{2}\sum_{i=1}^{n}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right). \\ f((\sigma^2_{\alpha_j})^{-1}|\cdot) &\propto (\sigma^2_{\alpha_j})^{-1(a_\alpha - 1)}\exp\big(-b_\alpha(\sigma^2_{\alpha_j})^{-1}\big) \\ &\quad \times (\sigma^2_{\alpha_j})^{-1(\frac{n}{2})}\exp\left(-\frac{1}{2}\boldsymbol{\alpha}_j^{\mathrm{T}}\big(\sigma^2_{\alpha_j}\exp(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}_j\right) \\ &\quad \propto \mathrm{Ga}\left(a_\alpha + \frac{n}{2}, b_\alpha + \frac{1}{2}\boldsymbol{\alpha}_j^{\mathrm{T}}\exp^{-1}(-\phi_\alpha\mathbf{D})\boldsymbol{\alpha}_j\right). \\ f((\sigma^2_{\beta_j})^{-1}|\cdot) &\propto \mathrm{Ga}\left(a_\beta + \frac{n}{2}, b_\beta + \frac{1}{2}(\boldsymbol{\beta}_j - \mathbf{1})^{\mathrm{T}}\exp^{-1}(-\phi_\beta\mathbf{D})(\boldsymbol{\beta}_j - \mathbf{1})\right). \end{aligned}$$

$$
\begin{aligned}
f(\boldsymbol{\alpha}_j|\cdot) &\propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}_j^{\mathrm{T}}\big(\sigma^2_{\alpha_j}\exp(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}_j\right)\\
&\quad\times\exp\left(-\frac{1}{2}\Big(\boldsymbol{\alpha}_j-(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j)\Big)^{\mathrm{T}}(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}\Big(\boldsymbol{\alpha}_j-(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j)\Big)\right)\\
&=\exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\alpha}_j^{\mathrm{T}}\big((\sigma^2_{\alpha_j}\exp(-\phi_\alpha\mathbf{D}))^{-1}+(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}\big)\boldsymbol{\alpha}_j\\
&\quad-\boldsymbol{\alpha}_j^{\mathrm{T}}\big((\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j)\big)+\ldots\Big)\Bigg)\\
&\propto \mathrm{N}(\boldsymbol{\Sigma}_{\alpha_j}\mathbf{A}_{\alpha_j},\boldsymbol{\Sigma}_{\alpha_j}),\ \text{where}\\
&\boldsymbol{\Sigma}_{\alpha_j}=\Big(\big(\sigma^2_{\alpha_j}\exp(-\phi_\alpha\mathbf{D})\big)^{-1}+(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}\Big)^{-1}\ \text{and}\ \mathbf{A}_{\alpha_j}=(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j).\\
f(\boldsymbol{\beta}_j|\cdot) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_j-\mathbf{1})^{\mathrm{T}}\big(\sigma^2_{\beta_j}\exp(-\phi_\beta\mathbf{D})\big)^{-1}(\boldsymbol{\beta}_j-\mathbf{1})\right)\\
&\quad\times\exp\left(-\frac{1}{2}\Big(\mathbf{X}_j\boldsymbol{\beta}_j-(\mathbf{y}_j-\boldsymbol{\alpha}_j)\Big)^{\mathrm{T}}(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}\Big(\mathbf{X}_j\boldsymbol{\beta}_j-(\mathbf{y}_j-\boldsymbol{\alpha}_j)\Big)\right)\\
&=\exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\beta}_j^{\mathrm{T}}\big((\sigma^2_{\beta_j}\exp(-\phi_\beta\mathbf{D}))^{-1}+\mathbf{X}_j^{\mathrm{T}}(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}\mathbf{X}_j\big)\boldsymbol{\beta}_j\\
&\quad-\boldsymbol{\beta}_j^{\mathrm{T}}\big((\sigma^2_{\beta_j}\exp(-\phi_\beta\mathbf{D}))^{-1}\mathbf{1}+\mathbf{X}_j^{\mathrm{T}}(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}(\mathbf{y}_j-\boldsymbol{\alpha}_j)\big)+\ldots\Big)\Bigg)\\
&\propto \mathrm{N}(\boldsymbol{\Sigma}_{\beta_j}\mathbf{A}_{\beta_j},\boldsymbol{\Sigma}_{\beta_j}),\ \text{where}\\
&\boldsymbol{\Sigma}_{\beta_j}=\Big(\big(\sigma^2_{\beta_j}\exp(-\phi_\beta\mathbf{D})\big)^{-1}+\mathbf{X}_j^{\mathrm{T}}(\sigma^2_{\varepsilon}\mathbf{I}_n)^{-1}\mathbf{X}_j\Big)^{-1}\\
&\text{and}\ \mathbf{A}_{\beta_j}=\big(\sigma^2_{\beta_j}\exp(-\phi_\beta\mathbf{D})\big)^{-1}\mathbf{1}+\mathbf{X}_j^{\mathrm{T}}(\sigma^2_{\varepsilon_j}\mathbf{I}_n)^{-1}(\mathbf{y}_j-\boldsymbol{\alpha}_j).
\end{aligned}
$$

Predictions at new locations i ($i = 1, \ldots, p$, where p is the number of locations at which predictions are to be made), are made by sampling from the posterior predictive distribution:

$$\tilde{y}_{ji} \sim \mathrm{N}(\tilde{\alpha}_{ji} + \tilde{\beta}_{ji}\tilde{x}_{ji}, \sigma^2_{\varepsilon_j}), \tag{A.2}$$

for $j = 1, \ldots, t$ and $i = 1, \ldots, p$, where $\tilde{x}_{ji}$ is the value of remotely-sensed data for the grid cell containing prediction location i, at time j, $\sigma^2_{\varepsilon_j}$ is the error variance, estimated from fitting the model to the data, and the distributions of $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_j$ are conditional on the values of $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ obtained from fitting the model to the data. The joint distribution of predicted $\tilde{\boldsymbol{\alpha}}_j$ and observed

$\boldsymbol{\alpha}_j$ is:

$$\begin{pmatrix} \tilde{\boldsymbol{\alpha}}_j \\ \boldsymbol{\alpha}_j \end{pmatrix} \sim \mathrm{N}_{p+n}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right),$$

so that the conditional distribution of $\tilde{\boldsymbol{\alpha}}_j$, given $\boldsymbol{\alpha}_j$, is:

$$\tilde{\boldsymbol{\alpha}}_j | \boldsymbol{\alpha}_j \sim \mathrm{N}_p\big(\boldsymbol{\mu_1} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\alpha}_j - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\big),$$

where $\boldsymbol{\mu}_1 = \mathbf{0}$, $\boldsymbol{\mu}_2 = \mathbf{0}$, $\boldsymbol{\Sigma}_{11} = \sigma^2_{\alpha_j} \exp(-\phi_\alpha \mathbf{D}_{11})$, $\boldsymbol{\Sigma}_{12} = \sigma^2_{\alpha_j} \exp(-\phi_\alpha \mathbf{D}_{12})$, $\boldsymbol{\Sigma}_{21} = \sigma^2_{\alpha_j} \exp(-\phi_\alpha \mathbf{D}_{21})$ and $\boldsymbol{\Sigma}_{22} = \sigma^2_{\alpha_j} \exp(-\phi_\alpha \mathbf{D}_{22})$. The matrix of distances between prediction and *in situ* data locations is partitioned as follows:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix},$$

where $\mathbf{D}_{11}$ is the matrix of distances between prediction locations, $\mathbf{D}_{22}$ is the matrix of distances between *in situ* data locations, and $\mathbf{D}_{12}$ and $\mathbf{D}_{21}$ are the matrices of distances between locations of *in situ* data and predictions. The conditional distribution of $\tilde{\boldsymbol{\alpha}}_j | \boldsymbol{\alpha}_j$ therefore simplifies to:

$$\begin{aligned} \tilde{\boldsymbol{\alpha}}_j | \boldsymbol{\alpha}_j \sim \mathrm{N}_p\Big(& \mathbf{0} + \exp(-\phi_\alpha \mathbf{D}_{12}) \exp(-\phi_\alpha \mathbf{D}_{22})^{-1} (\boldsymbol{\alpha}_j - \mathbf{0}), \\ & \sigma^2_{\alpha_j} \big(\exp(-\phi_\alpha \mathbf{D}_{11}) - \exp(-\phi_\alpha \mathbf{D}_{12}) \exp(-\phi_\alpha \mathbf{D}_{22})^{-1} \exp(-\phi_\alpha \mathbf{D}_{21}) \big) \Big). \end{aligned}$$

Given that ϕ_α and ϕ_β are chosen before fitting the model, the only parts of the mean and covariance matrix of the distribution of $\tilde{\boldsymbol{\alpha}}_j | \boldsymbol{\alpha}_j$ that must be re-calculated at each iteration of the Gibbs sampler are $\boldsymbol{\alpha}_j$ and $\sigma^2_{\alpha_j}$.

Let $\boldsymbol{\theta} \sim \mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a length-d vector, with a multivariate Normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. A random sample from $\boldsymbol{\theta}$ can be drawn as follows (Gelman et al. 2014):

- Let $\mathbf{A}$ be a lower triangular matrix, such that $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$.

- Let $\mathbf{z}$ be a d-length vector of independent standard Normal draws.
- Then a draw from $\mathbf{x}$ is $\boldsymbol{\mu} + \mathbf{A}\mathbf{z}$.

In the case of sampling from $\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j$, this means that the Cholesky decomposition of $\sigma^2_{\alpha_j}\big(\exp(-\phi_\alpha \mathbf{D}_{11}) - \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}\exp(-\phi_\alpha \mathbf{D}_{21})\big)$ must be calculated at each iteration of the Gibbs sampler, which is computationally expensive for a large number p of prediction locations. Fortunately, if $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$, then $(\sqrt{b}\mathbf{A})(\sqrt{b}\mathbf{A})^{\mathrm{T}} = b\boldsymbol{\Sigma}$ is the Cholesky decomposition of $b\boldsymbol{\Sigma}$. Taking note of this fact, it can be seen that the Cholesky decomposition of the covariance matrix $\big(\exp(-\phi_\alpha \mathbf{D}_{11}) - \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}\exp(-\phi_\alpha \mathbf{D}_{21})\big)$ of the distribution of $\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j$ only needs to be calculated once. Updates at each iteration then simply involve the multiplication of this value with the updated value of $\sqrt{\sigma^2_{\alpha_j}}$, which is much more computationally efficient.

The conditional distribution of the slope coefficients at prediction locations, given the values of the slope coefficients at *in situ* data locations, is:

$$\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j \sim \mathrm{N}_p\Big(\mathbf{1} + \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}(\boldsymbol{\beta}_j - \mathbf{1}),$$
$$\sigma^2_{\beta_j}\big(\exp(-\phi_\beta \mathbf{D}_{11}) - \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}\exp(-\phi_\beta \mathbf{D}_{21})\big)\Big).$$

As with the conditional distribution of the intercept coefficients for prediction locations, it can be noted that the covariance matrix of this distribution is a scalar (which updates at each iteration of the Gibbs sampler) multiplied by a matrix (which does not update at each iteration of the Gibbs sampler), so that there is no need to perform a Cholesky decomposition on this matrix at each iteration of the sampler.

This model was fitted, using a Gibbs sampler, in `C++`, with predictions obtained at new locations using the above computationally efficient procedure. The results of fitting the model to data for $\log(\text{chlorophyll}_a)$ are presented in Chapter 3.

A.2 Spatiotemporal statistical downscaling model 3.3a

The spatiotemporal statistical downscaling model 3.3a is written as:

$$\begin{aligned} y_{ji} &\sim \mathrm{N}(\alpha_{ji} + \beta_{ji} x_{ji}, \sigma_\varepsilon^2), \\ \boldsymbol{\alpha}_j &\sim \mathrm{N}_n(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})), \\ \boldsymbol{\beta}_j &\sim \mathrm{N}_n(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})), \\ (\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha), \\ (\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta), \\ (\sigma_\varepsilon^2)^{-1} &\sim \mathrm{Ga}(a_\varepsilon, b_\varepsilon), \end{aligned} \tag{A.3}$$

where y_{ji} is the value of *in situ* data for $\log(\text{chlorophyll}_a)$ at time j ($j = 1, \ldots, t$) and location i ($i = 1, \ldots, n$), x_{ji} is the value of remote sensing data at time j for the grid cell containing *in situ* location i, $\mathbf{D}$ is the $n \times n$ matrix of distances between *in situ* locations, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jn})^{\mathrm{T}}$ is the vector of intercept parameters for time j, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jn})^{\mathrm{T}}$ is the vector of slope parameters for time j, $\mathbf{0}$ is an n-length vector of zeros and $\mathbf{1}$ is an n-length vector of ones.

The probability density functions for the data distribution and for the distribution of each parameter in the model can be written as follows:

$$\begin{aligned} f(y_{ji}) &= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2} \big(y_{ji} - (\alpha_{ji} + \beta_{ji} x_{ji})\big)^2 \right), \\ f(\boldsymbol{\alpha}_j) &= \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{\alpha}_j - \mathbf{0})^{\mathrm{T}} \big(\sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big)^{-1} (\boldsymbol{\alpha}_j - \mathbf{0}) \right), \\ f(\boldsymbol{\beta}_j) &= \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{\beta}_j - \mathbf{1})^{\mathrm{T}} \big(\sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big)^{-1} (\boldsymbol{\beta}_j - \mathbf{1}) \right), \\ f\big((\sigma_\alpha^2)^{-1}\big) &= b_\alpha^{a_\alpha} (\sigma_\alpha^2)^{-1(a_\alpha - 1)} \exp\big(- b_\alpha (\sigma_\alpha^2)^{-1} \big) / \Gamma(a_\alpha), \\ f\big((\sigma_\beta^2)^{-1}\big) &= b_\beta^{a_\beta} (\sigma_\beta^2)^{-1(a_\beta - 1)} \exp\big(- b_\beta (\sigma_\beta^2)^{-1} \big) / \Gamma(a_\beta), \\ f\big((\sigma_\varepsilon^2)^{-1}\big) &= b_\varepsilon^{a_\varepsilon} (\sigma_\varepsilon^2)^{-1(a_\varepsilon - 1)} \exp\big(- b_\varepsilon (\sigma_\varepsilon^2)^{-1} \big) / \Gamma(a_\varepsilon), \end{aligned}$$

for $j = 1, \ldots, t$ and $i = 1, \ldots, n$. For some derivations, it is useful to work with matrices or vectors, so the following probability density functions are also provided here:

$$f(\boldsymbol{\alpha}) = \frac{1}{(2\pi)^{\frac{tn}{2}}|\sigma_\alpha^2 \text{exp}(-\phi_\alpha \mathbf{D})|^{\frac{t}{2}}} \exp\left(-\frac{1}{2}\sum_{j=1}^{t}(\boldsymbol{\alpha}_j - \mathbf{0})^{\mathrm{T}}\big(\sigma_\alpha^2 \text{exp}(-\phi_\alpha \mathbf{D})\big)^{-1}(\boldsymbol{\alpha}_j - \mathbf{0})\right),$$

$$f(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{tn}{2}}|\sigma_\beta^2 \text{exp}(-\phi_\beta \mathbf{D})|^{\frac{t}{2}}} \exp\left(-\frac{1}{2}\sum_{j=1}^{t}(\boldsymbol{\beta}_j - \mathbf{1})^{\mathrm{T}}\big(\sigma_\beta^2 \text{exp}(-\phi_\beta \mathbf{D})\big)^{-1}(\boldsymbol{\beta}_j - \mathbf{1})\right),$$

$$\begin{aligned} f(\mathbf{y}_j) &= \frac{1}{(2\pi)^{\frac{n}{2}}(\sigma_\varepsilon^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right) \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma_\varepsilon^2 \mathrm{I}_n|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\big(\mathbf{y}_j - (\boldsymbol{\alpha}_j + \mathbf{X}_j\boldsymbol{\beta}_j)\big)^{\mathrm{T}}(\sigma_\varepsilon^2 \mathbf{I}_n)^{-1})\big(\mathbf{y}_j - (\boldsymbol{\alpha}_j + \mathbf{X}_j\boldsymbol{\beta}_j)\big)\right), \\ f(\mathbf{y}) &= \frac{1}{(2\pi)^{\frac{tn}{2}}(\sigma_\varepsilon^2)^{\frac{tn}{2}}} \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\sum_{j=1}^{t}\sum_{i=1}^{n}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right). \end{aligned}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix and:

$$\mathbf{X}_j = \begin{pmatrix} x_{j1} & 0 & \cdots & 0 \\ 0 & x_{j2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & x_{jn} \end{pmatrix}$$

is an $n \times n$ diagonal matrix with the vector of remotely-sensed data at time j, $\mathbf{x}_j = (x_{j1}, \ldots, x_{jn})^{\mathrm{T}}$, as its diagonal.

The full conditional posterior distributions are calculated as follows:

$$\begin{aligned} f((\sigma_\varepsilon^2)^{-1}|\cdot) &\propto (\sigma_\varepsilon^2)^{-1(a_\varepsilon - 1)}\text{exp}(-b_\varepsilon(\sigma_\varepsilon^2)^{-1}) \\ &\quad \times (\sigma_\varepsilon^2)^{-1(\frac{nt}{2})}\text{exp}\left(-(\sigma_\varepsilon^2)^{-1}\left(\frac{1}{2}\sum_{j=1}^{t}\sum_{i=1}^{n}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right)\right) \\ &\quad \propto \text{Ga}\left(a_\varepsilon + \frac{nt}{2}, b_\varepsilon + \frac{1}{2}\sum_{j=1}^{t}\sum_{i=1}^{n}\big(y_{ji} - (\alpha_{ji} + \beta_{ji}x_{ji})\big)^2\right) \end{aligned}$$

$$
\begin{aligned}
f((\sigma_\alpha^2)^{-1}|\cdot) &\propto (\sigma_\alpha^2)^{-1(a_\alpha-1)}\exp(-b_\alpha(\sigma_\alpha^2)^{-1}) \\
&\quad\times (\sigma_\alpha^2)^{-1(\frac{nt}{2})}\exp\left(-\frac{1}{2}\sum_{j=1}^{t}\boldsymbol{\alpha}_j^{\mathrm{T}}\big(\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}_j\right) \\
&\quad\propto \mathrm{Ga}\left(a_\alpha+\frac{nt}{2}, b_\alpha+\frac{1}{2}\sum_{j=1}^{t}\boldsymbol{\alpha}_j^{\mathrm{T}}\exp^{-1}(-\phi_\alpha\mathbf{D})\boldsymbol{\alpha}_j\right) \\
f((\sigma_\beta^2)^{-1}|\cdot) &\propto \mathrm{Ga}\left(a_\beta+\frac{nt}{2}, b_\beta+\frac{1}{2}\sum_{j=1}^{t}(\boldsymbol{\beta}_j-\mathbf{1})^{\mathrm{T}}\exp^{-1}(-\phi_\beta\mathbf{D})(\boldsymbol{\beta}_j-\mathbf{1})\right) \\
f(\boldsymbol{\alpha}_j|\cdot) &\propto \exp\left(-\frac{1}{2}\boldsymbol{\alpha}_j^{\mathrm{T}}\big(\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}_j\right) \\
&\quad\times \exp\left(-\frac{1}{2}\Big(\boldsymbol{\alpha}_j-(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j)\Big)^{\mathrm{T}}(\sigma_\varepsilon^2\mathbf{I}_n)\Big(\boldsymbol{\alpha}_j-(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j)\Big)\right) \\
&\quad= \exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\alpha}_j^{\mathrm{T}}\big((\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D}))^{-1}+(\sigma_\varepsilon^2\mathbf{I}_n)^{-1}\big)\boldsymbol{\alpha}_j \\
&\qquad -\boldsymbol{\alpha}_j^{\mathrm{T}}\big((\sigma_\varepsilon^2\mathbf{I}_n)^{-1}(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j)\big)+\ldots\Big)\Bigg) \\
&\quad\propto \mathrm{N}(\boldsymbol{\Sigma}_{\alpha_j}\mathbf{A}_{\alpha_j}, \boldsymbol{\Sigma}_{\alpha_j}), \text{ where} \\
&\quad\boldsymbol{\Sigma}_{\alpha_j} = \Big(\big(\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D})\big)^{-1}+(\sigma_\varepsilon^2\mathbf{I}_n)^{-1}\Big)^{-1} \text{ and } \mathbf{A}_{\alpha_j} = (\sigma_\varepsilon^2\mathbf{I}_n)^{-1}(\mathbf{y}_j-\mathbf{X}_j\boldsymbol{\beta}_j) \\
f(\boldsymbol{\beta}_j|\cdot) &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_j-\mathbf{1})^{\mathrm{T}}\big(\sigma_\beta^2\exp(-\phi_\beta\mathbf{D})\big)^{-1}(\boldsymbol{\beta}_j-\mathbf{1})\right) \\
&\quad\times \exp\left(-\frac{1}{2}\Big(\mathbf{X}_j\boldsymbol{\beta}_j-(\mathbf{y}_j-\boldsymbol{\alpha}_j)\Big)^{\mathrm{T}}(\sigma_\varepsilon^2\mathbf{I}_n)^{-1}\Big(\mathbf{X}_j\boldsymbol{\beta}_j-(\mathbf{y}_j-\boldsymbol{\alpha}_j)\Big)\right) \\
&\quad= \exp\Bigg(-\frac{1}{2}\Big(\boldsymbol{\beta}_j^{\mathrm{T}}\big((\sigma_\beta^2\exp(-\phi_\beta\mathbf{D}))^{-1}+\mathbf{X}_j^{\mathrm{T}}(\sigma_\varepsilon^2\mathbf{I}_n)^{-1}\mathbf{X}_j\big)\boldsymbol{\beta}_j \\
&\quad -\boldsymbol{\beta}_j^{\mathrm{T}}\big((\sigma_\beta^2\exp(-\phi_\beta\mathbf{D}))^{-1}\mathbf{1}+\mathbf{X}_j^{\mathrm{T}}(\sigma_\varepsilon^2\mathbf{I}_n)^{-}(\mathbf{y}_j-\boldsymbol{\alpha}_j)\big)+\ldots\Big)\Bigg) \\
&\quad\propto \mathrm{N}(\boldsymbol{\Sigma}_{\beta_j}\mathbf{A}_{\beta_j}, \boldsymbol{\Sigma}_{\beta_j}), \text{ where} \\
&\quad\boldsymbol{\Sigma}_{\beta_j} = \Big(\big(\sigma_\beta^2\exp(-\phi_\beta\mathbf{D})\big)^{-1}+\mathbf{X}_j^{\mathrm{T}}(\sigma_\varepsilon^2\mathbf{I}_n)^{-1}\mathbf{X}_j\Big)^{-1} \\
&\quad\text{and } \mathbf{A}_{\beta_j} = \big(\sigma_\beta^2\exp(-\phi_\beta\mathbf{D})\big)^{-1}\mathbf{1}+\mathbf{X}_j^{\mathrm{T}}(\sigma_\varepsilon^2\mathbf{I}_n)^{-1}(\mathbf{y}_j-\boldsymbol{\alpha}_j).
\end{aligned}
$$

Predictions are made from the posterior predictive distribution:

$$\tilde{y}_{ji} \sim \mathrm{N}(\tilde{\alpha}_{ji}+\tilde{\beta}_{ji}\tilde{x}_{ji}, \sigma_{\varepsilon_j}^2), \tag{A.4}$$

where the conditional distributions of the intercepts and slopes at prediction

locations, given their values at the locations of the *in situ* data, are:

$$\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j \sim \mathrm{N}_p\Big(\mathbf{0} + \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}(\boldsymbol{\alpha}_j - \mathbf{0}),$$
$$\sigma^2_\alpha\big(\exp(-\phi_\alpha \mathbf{D}_{11}) - \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}\exp(-\phi_\alpha \mathbf{D}_{21})\big)\Big) \text{ and}$$
$$\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j \sim \mathrm{N}_p\Big(\mathbf{1} + \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}(\boldsymbol{\beta}_j - \mathbf{1}),$$
$$\sigma^2_\beta\big(\exp(-\phi_\beta \mathbf{D}_{11}) - \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}\exp(-\phi_\beta \mathbf{D}_{21})\big)\Big)$$

The model was fitted and predictions were made, using a Gibbs sampler written in `C++`, using a computationally efficient method for obtaining samples from the distributions of $\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j$ and $\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j$, which makes use of the fact that their covariance matrices are made up of a scalar (which is updated at each iteration) and a matrix (which does not need to be updated at each iteration). See the previous section for further discussion of this. This computationally efficient algorithm is only possible, if ϕ_α and ϕ_β are estimated outwith the model. This model was fitted to data for $\log(\text{chlorophyll}_a)$, in Chapter 3.

A.3 Bivariate spatial model 4.1

The bivariate spatial statistical downscaling model 4.1 can be written as follows. The likelihood is:

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} \alpha_{1i} + \beta_{1i} x_{1i} \\ \alpha_{2i} + \beta_{2i} x_{2i} \end{pmatrix}, \boldsymbol{\Xi} \right), \tag{A.5}$$

where y_{1i} and y_{2i} are the *in situ* data for variables 1 and 2, respectively, at location i ($i = 1, \ldots, n$), and x_{1i} and x_{2i} are the remote sensing data for variables 1 and 2, for the grid cell corresponding to *in situ* location i.

The prior distributions are:

$$\begin{aligned}
\boldsymbol{\alpha}_1 &\sim \mathrm{N}_n(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})), \\
\boldsymbol{\alpha}_2 &\sim \mathrm{N}_n(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})), \\
\boldsymbol{\beta}_1 &\sim \mathrm{N}_n(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})), \\
\boldsymbol{\beta}_2 &\sim \mathrm{N}_n(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})), \\
\boldsymbol{\Xi} &\sim \text{Inv-W}(\mathbf{R}^{-1}, k),
\end{aligned}$$

where $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1n})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \ldots, \alpha_{2n})^{\mathrm{T}}$ are the vectors of intercept parameters for variables 1 and 2, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1n})^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2n})^{\mathrm{T}}$ are the vectors of corresponding slope parameters, $\mathbf{0}$ is an n-length vector of zeros, $\mathbf{1}$ is an n-length vector of ones, $\mathbf{D}$ is an $n \times n$ matrix of distances between *in situ* locations, and the parameters ϕ_α, ϕ_β and k, along with the 2×2 matrix of parameters $\mathbf{R}$, must be chosen outwith the modelling process.

For some calculations, it is more useful to work with the distributions for vectors and matrices of parameters:

$$\begin{aligned}
\boldsymbol{\alpha} &\sim \mathrm{N}_{2\times n}(\mathbf{0}, \mathbf{I}_2, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})) \\
\boldsymbol{\beta} &\sim \mathrm{N}_{2\times n}(\mathbf{0}, \mathbf{I}_2, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})) \\
\mathbf{y}_i &\sim \mathrm{N}_2(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i \odot \mathbf{x}_i, \boldsymbol{\Xi}) \\
\mathbf{y} &\sim \mathrm{N}_{2\times n}(\boldsymbol{\alpha} + \boldsymbol{\beta} \odot \mathbf{x}, \boldsymbol{\Xi}, \mathbf{I}_n), \\
\mathrm{vec}(\boldsymbol{\alpha}) &\sim \mathrm{N}_{2n}(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D}) \otimes \mathbf{I}_2), \\
\mathrm{vec}(\boldsymbol{\beta}) &\sim \mathrm{N}_{2n}(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D}) \otimes \mathbf{I}_2), \\
\mathrm{vec}(\mathbf{y}) &\sim \mathrm{N}_{2n}(\mathrm{vec}(\boldsymbol{\alpha}) + \mathrm{vec}(\boldsymbol{\beta}) \odot \mathrm{vec}(\mathbf{x}), \mathbf{I}_n \otimes \boldsymbol{\Xi}),
\end{aligned}$$

where $\mathbf{I}_2$ is the 2×2 identity matrix, $\mathbf{I}_n$ is the $n \times n$ identity matrix and the other parameters are $\boldsymbol{\alpha}_i = (\alpha_{1i}, \alpha_{21})^{\mathrm{T}}$, $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{21})^{\mathrm{T}}$, $\mathbf{y}_i = (y_{1i}, y_{21})^{\mathrm{T}}$,

$\mathbf{x}_i = (x_{1i}, x_{21})^{\mathrm{T}}$,

$$\boldsymbol{\alpha} = \Big((\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)^{\mathrm{T}}\Big) = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2n} \end{pmatrix},$$

$$\boldsymbol{\beta} = \Big((\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^{\mathrm{T}}\Big) = \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2n} \end{pmatrix},$$

$$\mathbf{y} = \Big((\mathbf{y}_1, \mathbf{y}_2)^{\mathrm{T}}\Big) = \begin{pmatrix} y_{11} & y_{12} & \dots & y_{1n} \\ y_{21} & y_{22} & \dots & y_{2n} \end{pmatrix},$$

$$\mathbf{x} = \Big((\mathbf{x}_1, \mathbf{x}_2)^{\mathrm{T}}\Big) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \end{pmatrix},$$

$$\mathrm{vec}(\boldsymbol{\alpha}) = (\alpha_{11}, \alpha_{21}, \alpha_{12}, \alpha_{22}, \dots, \alpha_{1n}, \alpha_{2n}),$$

$$\mathrm{vec}(\boldsymbol{\beta}) = (\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, \dots, \beta_{1n}, \beta_{2n}),$$

$$\mathrm{vec}(\mathbf{y}) = (y_{11}, y_{21}, y_{12}, y_{22}, \dots, y_{1n}, y_{2n}),$$

$$\mathrm{vec}(\mathbf{x}) = (x_{11}, x_{21}, x_{12}, x_{22}, \dots, x_{1n}, x_{2n}),$$

where $\odot$ represents the Hadamard, or Schur, product operation (i.e. element-by-element multiplication of two matrices) and $\otimes$ represents the Kronecker product operation, where e.g. $\mathbf{A} \otimes \mathbf{B} = \left(\begin{smallmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{smallmatrix}\right)$, for $\mathbf{A} = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right)$.

The probability density functions for the parameters in this model are written as follows:

$$f(\boldsymbol{\alpha}_1) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{\alpha}_1^{\mathrm{T}} \big(\sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big)^{-1} \boldsymbol{\alpha}_1\right)$$

$$f(\boldsymbol{\alpha}_2) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \boldsymbol{\alpha}_2^{\mathrm{T}} \big(\sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D})\big)^{-1} \boldsymbol{\alpha}_2\right)$$

$$f(\boldsymbol{\beta}_1) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{\beta}_1 - \mathbf{1})^{\mathrm{T}} \big(\sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big)^{-1} (\boldsymbol{\beta}_1 - \mathbf{1})\right)$$

$$f(\boldsymbol{\beta}_2) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\boldsymbol{\beta}_2 - \mathbf{1})^{\mathrm{T}} \big(\sigma_\beta^2 \exp(-\phi_\beta \mathbf{D})\big)^{-1} (\boldsymbol{\beta}_2 - \mathbf{1})\right)$$

$$
\begin{aligned}
f(\boldsymbol{\alpha}) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})|^{\frac{2}{2}}|\mathbf{I}_2|^{\frac{n}{2}}}\ \text{exp}\bigg(-\frac{1}{2}\Big(\boldsymbol{\alpha}_1^{\text{T}}\big(\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}_1 \\
&\quad + \boldsymbol{\alpha}_2^{\text{T}}\big(\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}_2\Big)\bigg) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})|^{\frac{2}{2}}|\mathbf{I}_2|^{\frac{n}{2}}}\ \text{exp}\left(-\frac{1}{2}\text{tr}\Big(\big(\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})\big)^{-1}\boldsymbol{\alpha}^{\text{T}}\mathbf{I}_2^{-1}\boldsymbol{\alpha}\Big)\right) \\
f(\boldsymbol{\beta}) &= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})|^{\frac{2}{2}}|\mathbf{I}_2|^{\frac{n}{2}}}\ \text{exp}\bigg(-\frac{1}{2}\Big((\boldsymbol{\beta}_1-\mathbf{1})^{\text{T}}\big(\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})\big)^{-1}(\boldsymbol{\beta}_1-\mathbf{1}) \\
&\quad + (\boldsymbol{\beta}_2-\mathbf{1})^{\text{T}}\big(\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})\big)^{-1}(\boldsymbol{\beta}_2-\mathbf{1})\Big)\bigg) \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}|\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})|^{\frac{2}{2}}|\mathbf{I}_2|^{\frac{n}{2}}} \\
&\quad \times \text{exp}\left(-\frac{1}{2}\text{tr}\Big(\big(\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})\big)^{-1}(\boldsymbol{\beta}-\mathbf{1})^{\text{T}}\mathbf{I}_2^{-1}(\boldsymbol{\beta}-\mathbf{1})\Big)\right) \\
f(\boldsymbol{\Xi}) &= \left(2^{\frac{2k}{2}}\pi^{\frac{2(2-1)}{4}}\Pi_{i=1}^2\Gamma\left(\frac{k+1-i}{2}\right)\right)^{-1}|\mathbf{R}|^{-\frac{k}{2}}|\boldsymbol{\Xi}|^{-\frac{k+2+1}{2}}\text{exp}\left(-\frac{1}{2}\text{tr}(\mathbf{R}\boldsymbol{\Xi}^{-1})\right) \\
f(\mathbf{y}_i) &= \frac{1}{(2\pi)^{\frac{2}{2}}|\boldsymbol{\Xi}|^{\frac{1}{2}}}\ \text{exp}\left(-\frac{1}{2}\Big(\mathbf{y}_i-(\boldsymbol{\alpha}_i+\boldsymbol{\beta}_i\odot\mathbf{x}_i)\Big)^{\text{T}}\boldsymbol{\Xi}^{-1}\Big(\mathbf{y}_i-(\boldsymbol{\alpha}_i+\boldsymbol{\beta}_i\odot\mathbf{x}_i)\Big)\right) \\
f(\mathbf{y}) &= \frac{1}{(2\pi)^{\frac{2n}{2}}|\boldsymbol{\Xi}|^{\frac{n}{2}}|\mathbf{I}_n|^{\frac{2}{2}}}\ \text{exp}\left(-\frac{1}{2}\text{tr}\Big(\mathbf{I}_n^{-1}\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)^{\text{T}}\boldsymbol{\Xi}^{-1}\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)\Big)\right) \\
f(\text{vec}(\boldsymbol{\alpha})) &= \frac{\text{exp}\Big(-\frac{1}{2}\big(\text{vec}(\boldsymbol{\alpha})-\mathbf{0}\big)^{\text{T}}\big(\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}\big(\text{vec}(\boldsymbol{\alpha})-\mathbf{0}\big)\Big)}{(2\pi)^{\frac{2n}{2}}|\sigma_\alpha^2\text{exp}(-\phi_\alpha\mathbf{D})\otimes\mathbf{I}_2|^{\frac{1}{2}}} \\
f(\text{vec}(\boldsymbol{\beta})) &= \frac{\text{exp}\Big(-\frac{1}{2}\big(\text{vec}(\boldsymbol{\beta})-\mathbf{1}\big)^{\text{T}}\big(\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}\big(\text{vec}(\boldsymbol{\beta})-\mathbf{1}\big)\Big)}{(2\pi)^{\frac{2n}{2}}|\sigma_\beta^2\text{exp}(-\phi_\beta\mathbf{D})\otimes\mathbf{I}_2|^{\frac{1}{2}}} \\
f(\text{vec}(\mathbf{y})) &= \frac{1}{(2\pi)^{\frac{2n}{2}}|\mathbf{I}_n\otimes\boldsymbol{\Xi}|^{\frac{1}{2}}}\ \text{exp}\left(-\frac{1}{2}\mathbf{K}^{\text{T}}(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\mathbf{K}\right),
\end{aligned}
$$

where $\mathbf{K} = \text{vec}(\mathbf{y}) - \big(\text{vec}(\boldsymbol{\alpha}) + \text{vec}(\boldsymbol{\beta})\odot\text{vec}(\mathbf{X})\big)$.

The full conditional posterior distributions for the bivariate spatial statistical downscaling model are as follows:

$$
\begin{aligned}
f\big((\sigma_\alpha^2)^{-1}|\cdot\big) &\propto (\sigma_\alpha^2)^{-1(a_\alpha-1)}\text{exp}(-(\sigma_\alpha^2)^{-1}b_\alpha) \\
&\quad \times (\sigma_\alpha^2)^{-1(\frac{2n}{2})}\text{exp}\left(-(\sigma_\alpha^2)^{-1}\left(\frac{1}{2}\Big(\boldsymbol{\alpha}_1^{\text{T}}\text{exp}^{-1}(-\phi_\alpha\mathbf{D})\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2^{\text{T}}\text{exp}^{-1}(-\phi_\alpha\mathbf{D})\boldsymbol{\alpha}_2\Big)\right)\right) \\
&\quad \propto \text{Ga}\left(a_\alpha+\frac{2n}{2}, b_\alpha+\frac{1}{2}\Big(\boldsymbol{\alpha}_1^{\text{T}}\text{exp}^{-1}(-\phi_\alpha\mathbf{D})\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2^{\text{T}}\text{exp}^{-1}(-\phi_\alpha\mathbf{D})\boldsymbol{\alpha}_2\Big)\right)
\end{aligned}
$$

$$
\begin{aligned}
f\big((\sigma_\beta^2)^{-1}|\cdot\big) &\propto (\sigma_\beta^2)^{-1(a_\beta-1)}\exp(-(\sigma_\beta^2)^{-1}b_\beta) \\
&\quad\times (\sigma_\beta^2)^{-1(\frac{2n}{2})}\exp\Bigg(-(\sigma_\beta^2)^{-1}\Bigg(\frac{1}{2}\Big((\boldsymbol{\beta}_1-\mathbf{1})^{\mathrm{T}}\exp^{-1}(-\phi_\beta\mathbf{D})(\boldsymbol{\beta}_1-\mathbf{1}) \\
&\quad+(\boldsymbol{\beta}_2-\mathbf{1})^{\mathrm{T}}\exp^{-1}(-\phi_\beta\mathbf{D})(\boldsymbol{\beta}_2-\mathbf{1})\Big)\Bigg)\Bigg) \\
&\quad\propto \mathrm{Ga}\Bigg(a_\beta+\frac{2n}{2}, b_\beta+\frac{1}{2}\Big((\boldsymbol{\beta}_1-\mathbf{1})^{\mathrm{T}}\exp^{-1}(-\phi_\beta\mathbf{D})(\boldsymbol{\beta}_1-\mathbf{1}) \\
&\quad+(\boldsymbol{\beta}_2-\mathbf{1})^{\mathrm{T}}\exp^{-1}(-\phi_\beta\mathbf{D})(\boldsymbol{\beta}_2-\mathbf{1})\Big)\Bigg) \\
f(\boldsymbol{\Xi}|\cdot) &\propto |\mathbf{R}|^{-\frac{k}{2}}|\boldsymbol{\Xi}|^{-\frac{k+2+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{R}\boldsymbol{\Xi}^{-1})\right) \\
&\quad\times |\boldsymbol{\Xi}|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\Big(\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)^{\mathrm{T}}\boldsymbol{\Xi}^{-1}\Big)\right) \\
&\quad\propto |\boldsymbol{\Xi}|^{-\frac{k+2+n+1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}\Big(\Big(\mathbf{R}+\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)^{\mathrm{T}}\Big)\boldsymbol{\Xi}^{-1}\Big)\right) \\
&\quad\propto \text{Inv-W}\Big(\mathbf{R}+\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)\big(\mathbf{y}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{x})\big)^{\mathrm{T}}, k+n\Big) \\
f\big(\mathrm{vec}(\boldsymbol{\alpha})|\cdot\big) &\propto \exp\left(-\frac{1}{2}\big(\mathrm{vec}(\boldsymbol{\alpha})-\mathbf{0}\big)^{\mathrm{T}}\big(\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}\big(\mathrm{vec}(\boldsymbol{\alpha})-\mathbf{0}\big)\right) \\
&\quad\times\exp\Bigg(-\frac{1}{2}\Big(\mathrm{vec}(\boldsymbol{\alpha})-\big(\mathrm{vec}(\mathbf{y})-(\mathrm{vec}(\boldsymbol{\beta})\odot\mathrm{vec}(\mathbf{x}))\big)\Big)^{\mathrm{T}}(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1} \\
&\quad\times\Big(\mathrm{vec}(\boldsymbol{\alpha})-\big(\mathrm{vec}(\mathbf{y})-(\mathrm{vec}(\boldsymbol{\beta})\odot\mathrm{vec}(\mathbf{x}))\big)\Big)\Bigg) \\
&\quad=\exp\Bigg(-\frac{1}{2}\Bigg(\big(\mathrm{vec}(\boldsymbol{\alpha})\big)^{\mathrm{T}}\Big(\big(\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}+(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\Big)\big(\mathrm{vec}(\boldsymbol{\alpha})\big) \\
&\quad-\big(\mathrm{vec}(\boldsymbol{\alpha})\big)^{\mathrm{T}}\Big((\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\big(\mathrm{vec}(\mathbf{y})-(\mathrm{vec}(\boldsymbol{\beta})\odot\mathrm{vec}(\mathbf{x}))\big)\Big)+\dots\Bigg)\Bigg) \\
&\quad\propto \mathrm{N}(\boldsymbol{\Sigma}_\alpha\mathbf{A}_\alpha,\boldsymbol{\Sigma}_\alpha),\ \text{where} \\
&\quad\boldsymbol{\Sigma}_\alpha=\Big(\big(\sigma_\alpha^2\exp(-\phi_\alpha\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}+(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\Big)^{-1}\ \text{and} \\
&\quad\mathbf{A}_\alpha=(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\Big(\mathrm{vec}(\mathbf{y})-\big(\mathrm{vec}(\boldsymbol{\beta})\odot\mathrm{vec}(\mathbf{x})\big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
f\big(\mathrm{vec}(\boldsymbol{\beta})|\cdot\big) &\propto \exp\left(-\frac{1}{2}\big(\mathrm{vec}(\boldsymbol{\beta})-\mathbf{1}\big)^{\mathrm{T}}\big(\sigma^2_\beta\exp(-\phi_\beta\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}\big(\mathrm{vec}(\boldsymbol{\beta})-\mathbf{1}\big)\right)\\
&\times\exp\left(-\frac{1}{2}\Big(\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)\mathrm{vec}(\boldsymbol{\beta})-\big(\mathrm{vec}(\mathbf{y})-\mathrm{vec}(\boldsymbol{\alpha})\big)\Big)^{\mathrm{T}}(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\right.\\
&\left.\times\Big(\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)\mathrm{vec}(\boldsymbol{\beta})-\big(\mathrm{vec}(\mathbf{y})-\mathrm{vec}(\boldsymbol{\alpha})\big)\Big)\right)\\
&\propto \mathrm{N}(\boldsymbol{\Sigma}_\beta\mathbf{A}_\beta,\boldsymbol{\Sigma}_\beta),\ \text{where}\\
\boldsymbol{\Sigma}_\beta &= \big(\sigma^2_\beta\exp(-\phi_\beta\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}+\Big(\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)\Big)^{\mathrm{T}}(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\Big(\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)\Big)\ \text{and}\\
\mathbf{A}_\beta &= \big(\sigma^2_\beta\exp(-\phi_\beta\mathbf{D})\otimes\mathbf{I}_2\big)^{-1}\mathbf{1}+\Big(\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)\Big)^{\mathrm{T}}(\mathbf{I}_n\otimes\boldsymbol{\Xi})^{-1}\big(\mathrm{vec}(\mathbf{y})-\mathrm{vec}(\boldsymbol{\alpha})\big),
\end{aligned}
$$

where $\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)$ is the diagonal matrix with $\mathrm{vec}(\mathbf{x})$ as its diagonal, i.e.:

$$
\mathrm{mat}\big(\mathrm{vec}(\mathbf{x})\big)=\begin{pmatrix}
x_{11} & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & x_{21} & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & x_{12} & 0 & \cdots & 0 & 0\\
0 & 0 & 0 & x_{22} & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots\\
0 & 0 & 0 & 0 & \cdots & x_{1n} & 0\\
0 & 0 & 0 & 0 & \cdots & 0 & x_{2n}
\end{pmatrix}.
$$

Predictions $\tilde{y}_{1i}$ and $\tilde{y}_{2i}$ for variables 1 and 2 at new location i ($i=1,\ldots,p$, where p is the number of locations at which to predict) are drawn from the posterior predictive distribution:

$$
\begin{pmatrix}\tilde{y}_{1i}\\ \tilde{y}_{2i}\end{pmatrix}\sim \mathrm{N}_2\left(\begin{pmatrix}\tilde{\alpha}_{1i}+\tilde{\beta}_{1i}\tilde{x}_{1i}\\ \tilde{\alpha}_{2i}+\tilde{\beta}_{2i}\tilde{x}_{2i}\end{pmatrix},\boldsymbol{\Xi}\right), \tag{A.6}
$$

where $\tilde{x}_{1i}$ and $\tilde{x}_{2i}$ are the values of remote sensing data for the grid cell that contains the *in situ* location i. The spatially varying parameters at prediction

locations are drawn from conditional multivariate Normal distributions:

$$\tilde{\boldsymbol{\alpha}}_1|\boldsymbol{\alpha}_1 \sim \mathrm{N}_p\Big(\mathbf{0} + \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}(\boldsymbol{\alpha}_1 - \mathbf{0}),$$
$$\sigma_\alpha^2\big(\exp(-\phi_\alpha \mathbf{D}_{11}) - \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}\exp(-\phi_\alpha \mathbf{D}_{21})\big)\Big),$$
$$\tilde{\boldsymbol{\alpha}}_2|\boldsymbol{\alpha}_2 \sim \mathrm{N}_p\Big(\mathbf{0} + \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}(\boldsymbol{\alpha}_2 - \mathbf{0}),$$
$$\sigma_\alpha^2\big(\exp(-\phi_\alpha \mathbf{D}_{11}) - \exp(-\phi_\alpha \mathbf{D}_{12})\exp(-\phi_\alpha \mathbf{D}_{22})^{-1}\exp(-\phi_\alpha \mathbf{D}_{21})\big)\Big),$$
$$\tilde{\boldsymbol{\beta}}_1|\boldsymbol{\beta}_1 \sim \mathrm{N}_p\Big(\mathbf{1} + \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}(\boldsymbol{\beta}_1 - \mathbf{1}),$$
$$\sigma_\beta^2\big(\exp(-\phi_\beta \mathbf{D}_{11}) - \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}\exp(-\phi_\beta \mathbf{D}_{21})\big)\Big) \text{ and}$$
$$\tilde{\boldsymbol{\beta}}_2|\boldsymbol{\beta}_2 \sim \mathrm{N}_p\Big(\mathbf{1} + \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}(\boldsymbol{\beta}_2 - \mathbf{1}),$$
$$\sigma_\beta^2\big(\exp(-\phi_\beta \mathbf{D}_{11}) - \exp(-\phi_\beta \mathbf{D}_{12})\exp(-\phi_\beta \mathbf{D}_{22})^{-1}\exp(-\phi_\beta \mathbf{D}_{21})\big)\Big).$$

This model was fitted using a Gibbs sampling algorithm, which was written in `C++`. Since the covariance matrices of the conditional multivariate Normal distributions of $\tilde{\boldsymbol{\alpha}}_1|\boldsymbol{\alpha}_1$, $\tilde{\boldsymbol{\alpha}}_2|\boldsymbol{\alpha}_2$, $\tilde{\boldsymbol{\beta}}_1|\boldsymbol{\beta}_1$ and $\tilde{\boldsymbol{\beta}}_2|\boldsymbol{\beta}_2$ are each made up of a scalar (which is updated at each iteration of the algorithm) and a matrix (which does not need to be updated at each iteration of the algorithm), the computationally efficient method of calculating the Cholesky decomposition of the matrix only once (discussed in the first section of this appendix) can be used, allowing fast and computationally efficient prediction of two variables at many locations. This reduction in computational complexity can only be achieved when ϕ_α and ϕ_β are estimated outwith the model.

A.4 Nonparametric downscaling model 5.8

The nonparametric statistical downscaling model (5.8) is written as:

$$
\begin{aligned}
\mathbf{y}_i|\mathbf{c}_i, \sigma_y^2 &\sim \mathrm{N}_{q_i}(\mathbf{\Phi}_i\mathbf{c}_i, \sigma_y^2\mathbf{I}_{q_i}),\\
(\sigma_y^2)^{-1} &\sim \mathrm{Ga}(a_y, b_y),\\
c_{ij}|\alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 &\sim \mathrm{N}(\alpha_{ij} + \beta_{ij}d_{ij}, \sigma_c^2),\\
\boldsymbol{\alpha}_j|\sigma_\alpha^2 &\sim \mathrm{N}_n\big(\mathbf{0}, \sigma_\alpha^2\mathbf{H}_{22}(\phi_\alpha)\big),\\
\boldsymbol{\beta}_j|\sigma_\beta^2 &\sim \mathrm{N}_n\big(\mathbf{1}, \sigma_\beta^2\mathbf{H}_{22}(\phi_\beta)\big),\\
(\sigma_\alpha^2)^{-1} &\sim \mathrm{Ga}(a_\alpha, b_\alpha),\\
(\sigma_\beta^2)^{-1} &\sim \mathrm{Ga}(a_\beta, b_\beta),\\
(\sigma_c^2)^{-1} &\sim \mathrm{Ga}(a_c, b_c),\\
\mathbf{x}_i|\mathbf{d}_i, \sigma_x^2 &\sim \mathrm{N}_{p_i}(\mathbf{\Psi}_i\mathbf{d}_i, \sigma_x^2\mathbf{I}_{p_i}),\\
(\sigma_x^2)^{-1} &\sim \mathrm{Ga}(a_x, b_x),\\
\mathbf{d}_i &\sim \mathrm{N}_m(\boldsymbol{\mu}_d, \mathbf{\Sigma}_d),
\end{aligned}
\tag{A.7}
$$

where:

- q_i is the number of *in situ* data collected at location i.
- p_i is the number of remotely-sensed data collected at location i.
- n is the number of *in situ* data locations i.
- m is the number of basis functions in each $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_i$.
- y_{ij} is the value of *in situ* data at time j at location i ($i = 1, \ldots, n$ and $j = 1, \ldots, q_i$), with $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq_i})^{\mathrm{T}}$.
- x_{ij} is the value of remotely-sensed data at time j at location i ($i = 1, \ldots, n$ and $j = 1, \ldots, p_i$), with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_i})^{\mathrm{T}}$.
- $\mathbf{\Phi}_i$ is the matrix of basis functions evaluated at times of data collection for the *in situ* data y_i at location i.

- $\mathbf{\Psi}_i$ is the matrix of basis functions evaluated at times of data collection for the remotely-sensed data x_i at location i.

- $\mathbf{H}_{22}(\phi_\alpha) = \exp(-\phi_\alpha \mathbf{D}_{22})$, where ϕ_α is selected *a priori* and $\mathbf{D}_{22}$ is the matrix of distances between *in situ* locations i.

- $\mathbf{H}_{22}(\phi_\beta) = \exp(-\phi_\beta \mathbf{D}_{22})$, where ϕ_β is selected *a priori*, with $\mathbf{D}_{22}$ as above.

- $a_y, b_y, a_\alpha, b_\alpha, a_\beta, b_\beta, a_c, b_c, a_x, b_x, \boldsymbol{\mu}_d$ and $\mathbf{\Sigma}_d$ are values to be chosen *a priori*. 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, results in non-informative prior distributions. Sensible values for $\boldsymbol{\mu}_d$ and $\mathbf{\Sigma}_d$ are the length-m vector of zeros ,$\mathbf{0}$, and the $m \times m$ identity matrix, $\mathbf{I}_m$, respectively, reflecting a lack of knowledge of the signs of the coefficients $\mathbf{d}_i$ and of their dependence structure.

The probability density functions for the likelihood and for the prior distributions are:

Likelihood:

$$
\begin{aligned}
f(\mathbf{y}) &= \prod_{i=1}^{n} f(\mathbf{y}_i) \\
&= \frac{1}{(2\pi)^{\sum_{i=1}^{n} q_i/2} \prod_{i=1}^{n} (|\sigma_y^2 \mathbf{I}_{q_i}|^{1/2})} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{\Phi}_i \mathbf{c}_i)^{\mathrm{T}} (\sigma_y^2 \mathbf{I}_{q_i})^{-1} (\mathbf{y}_i - \mathbf{\Phi}_i \mathbf{c}_i) \right).
\end{aligned}
$$

Prior distributions:

$$
\begin{aligned}
f(c_{ij}) &= \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{1}{2\sigma_c^2} \big(c_{ij} - (\alpha_{ij} + \beta_{ij} d_{ij})\big)^2 \right), \text{ so that} \\
f(\mathbf{c}_i) &= \prod_{j=1}^{m} f(\delta_{ij}) = \frac{1}{(2\pi)^{m/2} |\sigma_\delta^2 \mathbf{I}_m|^{1/2}} \\
&\quad \times \exp\left(-\frac{1}{2} \Big(\mathbf{c}_i - (\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i \odot \mathbf{d}_i) \Big)^{\mathrm{T}} (\sigma_\delta^2 \mathbf{I}_m)^{-1} \Big(\mathbf{c}_i - (\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i \odot \mathbf{d}_i) \Big) \right), \\
f(\mathbf{c}_j) &= \prod_{i=1}^{n} f(\delta_{ij}) = \frac{1}{(2\pi)^{n/2} |\sigma_\delta^2 \mathbf{I}_n|^{1/2}} \\
&\quad \times \exp\left(-\frac{1}{2} \Big(\mathbf{c}_j - (\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{d}_j) \Big)^{\mathrm{T}} (\sigma_\delta^2 \mathbf{I}_n)^{-1} \Big(\mathbf{c}_j - (\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{d}_j) \Big) \right) \text{ and}
\end{aligned}
$$

$$f(\mathbf{c}) = \prod_{i=1}^{n} f(\mathbf{c}_i) = \frac{1}{(2\pi)^{mn/2}|\sigma_\delta^2\mathbf{I}_n|^{m/2}|\mathbf{I}_m|^{n/2}}$$
$$\times \exp\left(-\frac{1}{2}\mathrm{tr}\Big((\sigma_\delta^2)^{-1}\big(\mathbf{c}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{d})\big)^{\mathrm{T}}(\mathbf{I}_m)^{-1}\big(\mathbf{c}-(\boldsymbol{\alpha}+\boldsymbol{\beta}\odot\mathbf{d})\big)\Big)\right).$$

$$f(\boldsymbol{\alpha}_j) = \frac{1}{(2\pi)^{n/2}|\sigma_\alpha^2\mathbf{H}_0(\phi_\alpha)|^{1/2}}\ \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_j-\mathbf{0})^{\mathrm{T}}\big(\sigma_\alpha^2\mathbf{H}_0(\phi_\alpha)\big)^{-1}(\boldsymbol{\alpha}_j-\mathbf{0})\right),\ \text{so that}$$

$$f(\boldsymbol{\alpha}) = \prod_{j=1}^{m} f(\boldsymbol{\alpha}_j) = \frac{1}{(2\pi)^{mn/2}|\sigma_\alpha^2\mathbf{H}_0(\phi_\alpha)|^{m/2}|\mathbf{I}_m|^{n/2}}$$
$$\times \exp\left(-\frac{1}{2}\mathrm{tr}\Big(\big(\sigma_\alpha^2\mathbf{H}_0(\phi_\alpha)\big)^{-1}(\boldsymbol{\alpha}_j-\mathbf{0})^{\mathrm{T}}(\mathbf{I}_m)^{-1}(\boldsymbol{\alpha}_j-\mathbf{0})\Big)\right).$$

$$f(\boldsymbol{\beta}_j) = \frac{1}{(2\pi)^{n/2}|\sigma_\beta^2\mathbf{H}_0(\phi_\beta)|^{1/2}}\ \exp\left(-\frac{1}{2}(\boldsymbol{\beta}_j-\mathbf{1})^{\mathrm{T}}(\sigma_\beta^2\mathbf{H}_0(\phi_\beta))^{-1}(\boldsymbol{\beta}_j-\mathbf{1})\right),\ \text{so that}$$

$$f(\boldsymbol{\beta}) = \prod_{j=1}^{m} f(\boldsymbol{\beta}_j) = \frac{1}{(2\pi)^{mn/2}|\sigma_\beta^2\mathbf{H}_0(\phi_\beta)|^{m/2}|\mathbf{I}_m|^{n/2}}$$
$$\times \exp\left(-\frac{1}{2}\mathrm{tr}\Big(\big(\sigma_\beta^2\mathbf{H}_0(\phi_\beta)\big)^{-1}(\boldsymbol{\beta}_j-\mathbf{1})^{\mathrm{T}}(\mathbf{I}_m)^{-1}(\boldsymbol{\beta}_j-\mathbf{1})\Big)\right).$$

$$f\big((\sigma_\alpha^2)^{-1}\big) = b_\alpha^{a_\alpha}(\sigma_\alpha^2)^{-1(a_\alpha-1)}\ \exp\big(-b_\alpha(\sigma_\alpha^2)^{-1}\big)/\Gamma(a_\alpha)$$

$$f\big((\sigma_\beta^2)^{-1}\big) = b_\beta^{a_\beta}(\sigma_\beta^2)^{-1(a_\beta-1)}\ \exp\big(-b_\beta(\sigma_\beta^2)^{-1}\big)/\Gamma(a_\beta)$$

$$f\big((\sigma_y^2)^{-1}\big) = b_y^{a_y}(\sigma_y^2)^{-1(a_y-1)}\ \exp\big(-b_y(\sigma_y^2)^{-1}\big)/\Gamma(a_y)$$

$$f\big((\sigma_\delta^2)^{-1}\big) = b_\delta^{a_\delta}(\sigma_\delta^2)^{-1(a_\delta-1)}\ \exp\big(-b_\delta(\sigma_\delta^2)^{-1}\big)/\Gamma(a_\delta)$$

$$f(\mathbf{x}_i) = \frac{1}{(2\pi)^{p_i/2}|\sigma_x^2\mathbf{I}_{p_i}|^{1/2}}\ \exp\left(-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\Psi}_i\mathbf{d}_i)^{\mathrm{T}}(\sigma_x^2\mathbf{I}_{p_i})^{-1}(\mathbf{x}_i-\boldsymbol{\Psi}_i\mathbf{d}_i)\right),\ \text{so that}$$

$$f(\mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}_i) = \frac{1}{(2\pi)^{\sum_{i=1}^n p_i/2}(\sigma_x^2)^{\sum_{i=1}^n p_i/2}\prod_{i=1}^n(|\mathbf{I}_{p_i}|^{1/2})}$$
$$\times \exp\left(-\frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\Psi}_i\mathbf{d}_i)^{\mathrm{T}}(\mathbf{I}_{p_i})^{-1}(\mathbf{x}_i-\boldsymbol{\Psi}_i\mathbf{d}_i)\right).$$

$$f\big((\sigma_x^2)^{-1}\big) = b_x^{a_x}(\sigma_x^2)^{-1(a_x-1)}\ \exp\big(-b_x(\sigma_x^2)^{-1}\big)/\Gamma(a_x)$$

$$f(\mathbf{d}_i) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_\gamma|^{1/2}}\ \exp\left(-\frac{1}{2}(\mathbf{d}_i-\boldsymbol{\mu}_\gamma)^{\mathrm{T}}\boldsymbol{\Sigma}_\gamma^{-1}(\mathbf{d}_i-\boldsymbol{\mu}_\gamma)\right)$$

The full conditional posterior distributions for model 5.8 are:

$$f\big((\sigma_\alpha^2)^{-1}|\cdot\big) \propto \text{Ga}\left(a_\alpha + \frac{mn}{2}, b_\alpha + \frac{1}{2}\text{tr}\Big(\mathbf{H}_0^{-1}(\phi_\alpha)\boldsymbol{\alpha}^{\text{T}}\boldsymbol{\alpha}\Big)\right)$$

$$f\big((\sigma_\beta^2)^{-1}|\cdot\big) \propto \text{Ga}\left(a_\beta + \frac{mn}{2}, b_\beta + \frac{1}{2}\text{tr}\Big(\mathbf{H}_0^{-1}(\phi_\beta)(\boldsymbol{\beta}-\mathbf{1})^{\text{T}}(\boldsymbol{\beta}-\mathbf{1})\Big)\right)$$

$$f\big((\sigma_y^2)^{-1}|\cdot\big) \propto \text{Ga}\left(a_y + \sum_{i=1}^{n}\frac{q_i}{2}, b_y + \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\mathbf{c}_i)^{\text{T}}(\boldsymbol{y}_i - \boldsymbol{\Phi}_i\mathbf{c}_i)\right)$$

$$f\big((\sigma_\delta^2)^{-1}|\cdot\big) \propto \text{Ga}\left(a_\delta + \frac{mn}{2}, b_\delta + \frac{1}{2}\text{tr}\Big(\mathbf{I}_n\big(\mathbf{c} - (\boldsymbol{\alpha} + \boldsymbol{\beta}\odot\mathbf{d})\big)^{\text{T}}\mathbf{I}_m\big(\mathbf{c} - (\boldsymbol{\alpha} + \boldsymbol{\beta}\odot\mathbf{d})\big)\Big)\right)$$

$$f(\boldsymbol{\alpha}_j|\cdot) \propto \text{N}(\boldsymbol{\Sigma}_{\alpha_j}\mathbf{A}_{\alpha_j}, \boldsymbol{\Sigma}_{\alpha_j}), \text{ where}$$

$$\boldsymbol{\Sigma}_{\alpha_j} = \Big(\big(\sigma_\alpha^2\mathbf{H}_0(\phi_\alpha)\big)^{-1} + (\sigma_\delta^2\mathbf{I}_n)^{-1}\Big)^{-1} \text{ and}$$

$$\mathbf{A}_{\alpha_j} = (\sigma_\delta^2\mathbf{I}_n)^{-1}(\mathbf{c}_j - \boldsymbol{\beta}_j\odot\mathbf{d}_j)$$

$$f(\boldsymbol{\beta}_j|\cdot) \propto \text{N}(\boldsymbol{\Sigma}_{\beta_j}\mathbf{A}_{\beta_j}, \boldsymbol{\Sigma}_{\beta_j}), \text{ where}$$

$$\boldsymbol{\Sigma}_{\beta_j} = \Big(\big(\sigma_\beta^2\mathbf{H}_0(\phi_\beta)\big)^{-1} + \mathbf{G}_j^{\text{T}}(\sigma_\delta^2\mathbf{I}_n)^{-1}\mathbf{G}_j\Big)^{-1} \text{ and}$$

$$\mathbf{A}_{\beta_j} = \big(\sigma_\beta^2\mathbf{H}_0(\phi_\beta)\big)^{-1}\mathbf{1} + \mathbf{G}_j^{\text{T}}(\sigma_\delta^2\mathbf{I}_n)^{-1}(\mathbf{c}_j - \boldsymbol{\alpha}_j)$$

$$f(\mathbf{c}_i|\cdot) \propto \text{N}(\boldsymbol{\Sigma}_{c_i}\mathbf{A}_{c_i}, \boldsymbol{\Sigma}_{c_i}), \text{ where}$$

$$\boldsymbol{\Sigma}_{c_i} = \Big(\boldsymbol{\Phi}_i^{\text{T}}(\sigma_y^2\mathbf{I}_{q_i})^{-1}\boldsymbol{\Phi}_i + (\sigma_\delta^2\mathbf{I}_m)^{-1}\Big)^{-1} \text{ and}$$

$$\mathbf{A}_{c_i} = \boldsymbol{\Phi}_i^{\text{T}}(\sigma_y^2\mathbf{I}_{q_i})^{-1}\mathbf{y}_i + (\sigma_\delta^2\mathbf{I}_m)^{-1}(\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i\odot\mathbf{d}_i)$$

$$f\big((\sigma_x^2)^{-1}|\cdot\big) = \text{Ga}\left(a_x + \sum_{i=1}^{n}\frac{p_i}{2}, b_x + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\Psi}_i\mathbf{d}_i)^{\text{T}}(\mathbf{I}_{p_i})^{-1}(\mathbf{x}_i - \boldsymbol{\Psi}_i\mathbf{d}_i)\right)$$

$$f(\mathbf{d}_i|\cdot) = \text{N}(\boldsymbol{\Sigma}_{d_i}\mathbf{A}_{d_i}, \boldsymbol{\Sigma}_{d_i}), \text{ where}$$

$$\boldsymbol{\Sigma}_{d_i} = \Big(\boldsymbol{\Sigma}_d^{-1} + \boldsymbol{\Psi}_i^{\text{T}}(\sigma_x^2\mathbf{I}_{p_i})^{-1}\boldsymbol{\Psi}_i + \mathbf{F}_i^{\text{T}}(\sigma_\delta^2\mathbf{I}_m)^{-1}\mathbf{F}_i\Big)^{-1} \text{ and}$$

$$\mathbf{A}_{d_i} = \boldsymbol{\Sigma}_d^{-1}\boldsymbol{\mu}_d + \boldsymbol{\Psi}_i^{\text{T}}(\sigma_x^2\mathbf{I}_{p_i})^{-1}\mathbf{x}_i + \mathbf{F}_i^{\text{T}}(\sigma_\delta^2\mathbf{I}_m)^{-1}(\mathbf{c}_i - \boldsymbol{\alpha}_i)$$

In the above equations, $\mathbf{G}_j$ and $\mathbf{F}_i$ represent diagonal matrices, with $\mathbf{d}_j$ and $\boldsymbol{\beta}_i$ as their diagonals, respectively. $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where n is the number of *in situ* sampling locations and m is the number of basis functions fitted for each location. p_i is the number of remotely-sensed data available for location i, while q_i is the number of *in situ* data available for

location i.

Predictions can be made at new times j ($j = 1, \ldots, \tilde{q}_i$, where $\tilde{q}_i$ is the number of times to predict at for location i) and at new locations i ($i = 1, \ldots, \tilde{n}$, where $\tilde{n}$ is the number of locations at which to predict), by drawing from the posterior predictive distribution:

$$\begin{aligned}
\tilde{\mathbf{y}}_i|\tilde{\mathbf{c}}_i, \sigma_y^2 &\sim \mathrm{N}_{\tilde{q}_i}(\tilde{\mathbf{\Phi}}_i\tilde{\mathbf{c}}_i, \sigma_y^2\mathbf{I}_{\tilde{q}_i}), \\
\tilde{c}_{ij}|\tilde{\alpha}_{ij}, \tilde{\beta}_{ij}, \tilde{d}_{ij}, \sigma_c^2 &\sim \mathrm{N}(\tilde{\alpha}_{ij} + \tilde{\beta}_{ij}\tilde{d}_{ij}, \sigma_c^2), \\
\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j &\sim \mathrm{N}\Big(\mathbf{0} + \exp(-\phi_\alpha\mathbf{D}_{12})\exp(-\phi_\alpha\mathbf{D}_{22})^{-1}(\boldsymbol{\alpha}_j - \mathbf{0}), \\
&\qquad \sigma_\alpha^2\big(\exp(-\phi_\alpha\mathbf{D}_{11}) - \exp(-\phi_\alpha\mathbf{D}_{12})\exp(-\phi_\alpha\mathbf{D}_{22})^{-1}\exp(-\phi_\alpha\mathbf{D}_{21})\big)\Big), \\
\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j &\sim \mathrm{N}\Big(\mathbf{1} + \exp(-\phi_\beta\mathbf{D}_{12})\exp(-\phi_\beta\mathbf{D}_{22})^{-1}(\boldsymbol{\beta}_j - \mathbf{1}), \\
&\qquad \sigma_\beta^2\big(\exp(-\phi_\beta\mathbf{D}_{11}) - \exp(-\phi_\beta\mathbf{D}_{12})\exp(-\phi_\beta\mathbf{D}_{22})^{-1}\exp(-\phi_\beta\mathbf{D}_{21})\big)\Big), \\
\tilde{\mathbf{d}}_i &\sim \mathrm{N}(\tilde{\mathbf{\Sigma}}_{d_i}\tilde{\mathbf{A}}_{d_i}, \tilde{\mathbf{\Sigma}}_{d_i}),
\end{aligned} \tag{A.8}$$

where:

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{d_i} &= \left(\mathbf{\Sigma}_d^{-1} + \tilde{\mathbf{\Psi}}_i^{\mathrm{T}}(\sigma_x^2\mathbf{I}_{\tilde{p}_i})^{-1}\tilde{\mathbf{\Psi}}_i\right)^{-1} \text{ and} \\
\tilde{\mathbf{A}}_{d_i} &= \mathbf{\Sigma}_d^{-1}\boldsymbol{\mu}_d + \tilde{\mathbf{\Psi}}_i^{\mathrm{T}}(\sigma_x^2\mathbf{I}_{\tilde{p}_i})^{-1}\tilde{\mathbf{x}}_i,
\end{aligned}$$

where $\tilde{\mathbf{\Phi}}$ is the matrix of basis coefficients evaluated at times of prediction for the *in situ* data at location i, $\tilde{\mathbf{\Psi}}$ is the matrix of basis coefficients evaluated at times of data collection for the remotely-sensed data for the grid cell containing location i, $\tilde{q}_i$ is the number of times at which to predict, for location i, $\tilde{p}_i$ is the number of remotely-sensed data collected at location i and $\tilde{\mathbf{x}}_i$ is the vector of remotely-sensed data for the grid cell containing the location i at which prediction is to be carried out.

The model is fitted using a Gibbs sampler, which is implemented in `C++`. Obtaining draws from $\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j$ and $\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j$ is potentially computationally expensive, if predictions are to be made at a large number of locations, since

the method of obtaining draws from the multivariate Normal distributions would involve taking the Cholesky decomposition of large matrices at each iteration of the Gibbs sampler. Noting, however, that the covariance matrices of the distributions of both $\tilde{\boldsymbol{\alpha}}_j|\boldsymbol{\alpha}_j$ and $\tilde{\boldsymbol{\beta}}_j|\boldsymbol{\beta}_j$ are each made up of a scalar that must be updated each time (i.e. σ_α^2 and σ_β^2, respectively) multiplied by a matrix that needs only to be calculated once, the computations can be reduced in complexity and sped up. The algorithm makes use of the fact that the Cholesky decomposition of $b\boldsymbol{\Sigma}$ is $(\sqrt{b}\mathbf{A})(\sqrt{b}\mathbf{A})^{\mathrm{T}} = b\boldsymbol{\Sigma}$, where $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}$, so that the computation at each iteration of the sampler multiplies a matrix by a scalar, rather than the more complex Cholesky decomposition. It can be seen that, if ϕ_α and ϕ_β were to be estimated within the model, rather than being chosen before fitting the model, then this reduction in computational complexity would not be possible and the Cholesky decomposition would be required at each iteration of the Gibbs sampler.

Appendix B

Diagnostic plots for statistical downscaling models

This appendix presents plots for diagnosing whether the assumptions of the statistical downscaling models have been met, along with trace and density plots for checking whether MCMC chains have converged. For each model, there are too many parameters to include trace and density plots for them all here. Plots are instead presented for a small number of parameters, to give an idea of how appropriate the assumption of convergence is in general.

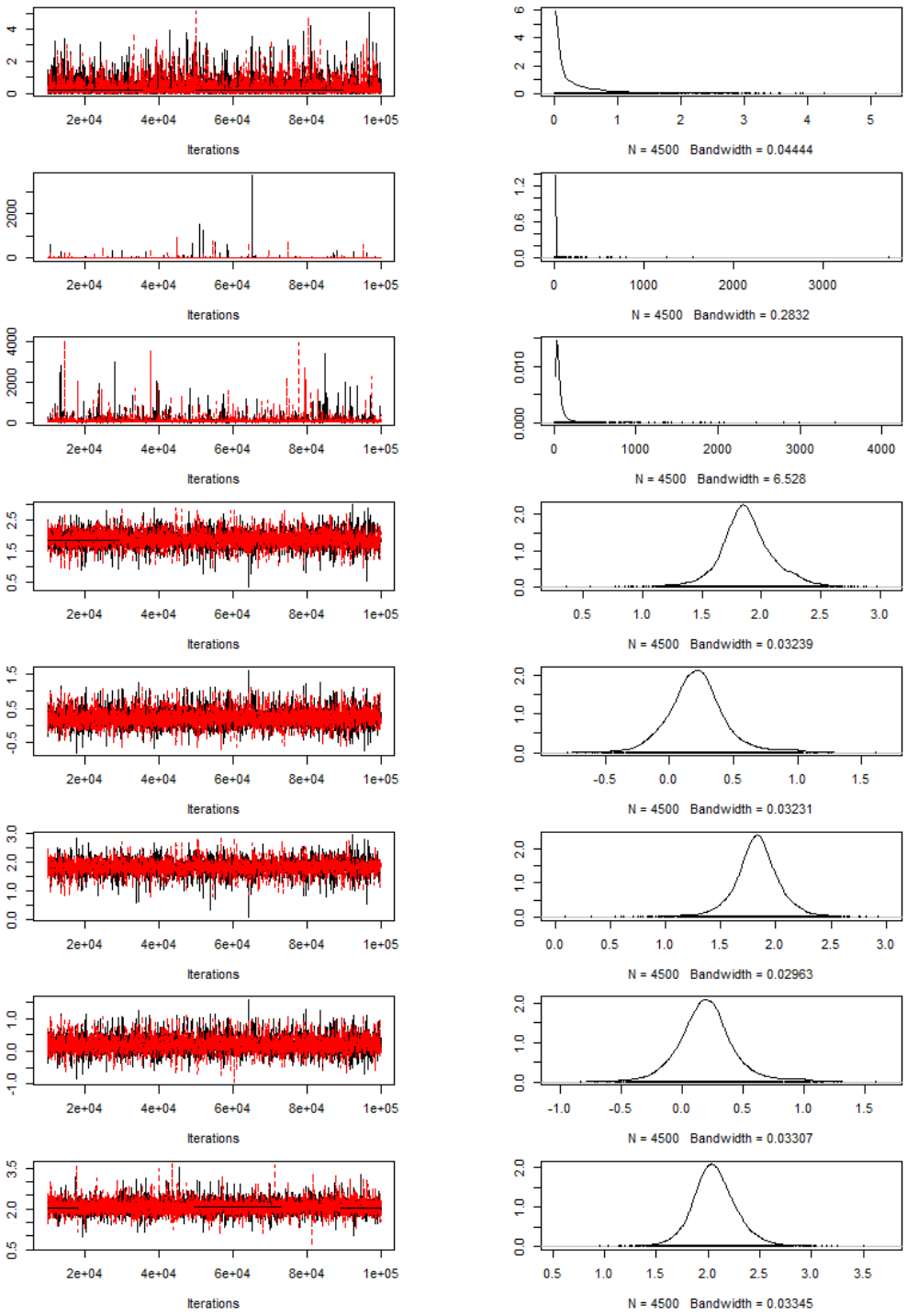

Figure B.1: Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\epsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$, of model 3.1, fitted to the log(chlorophyll$_a$) data for Lake Balaton for October 2008.

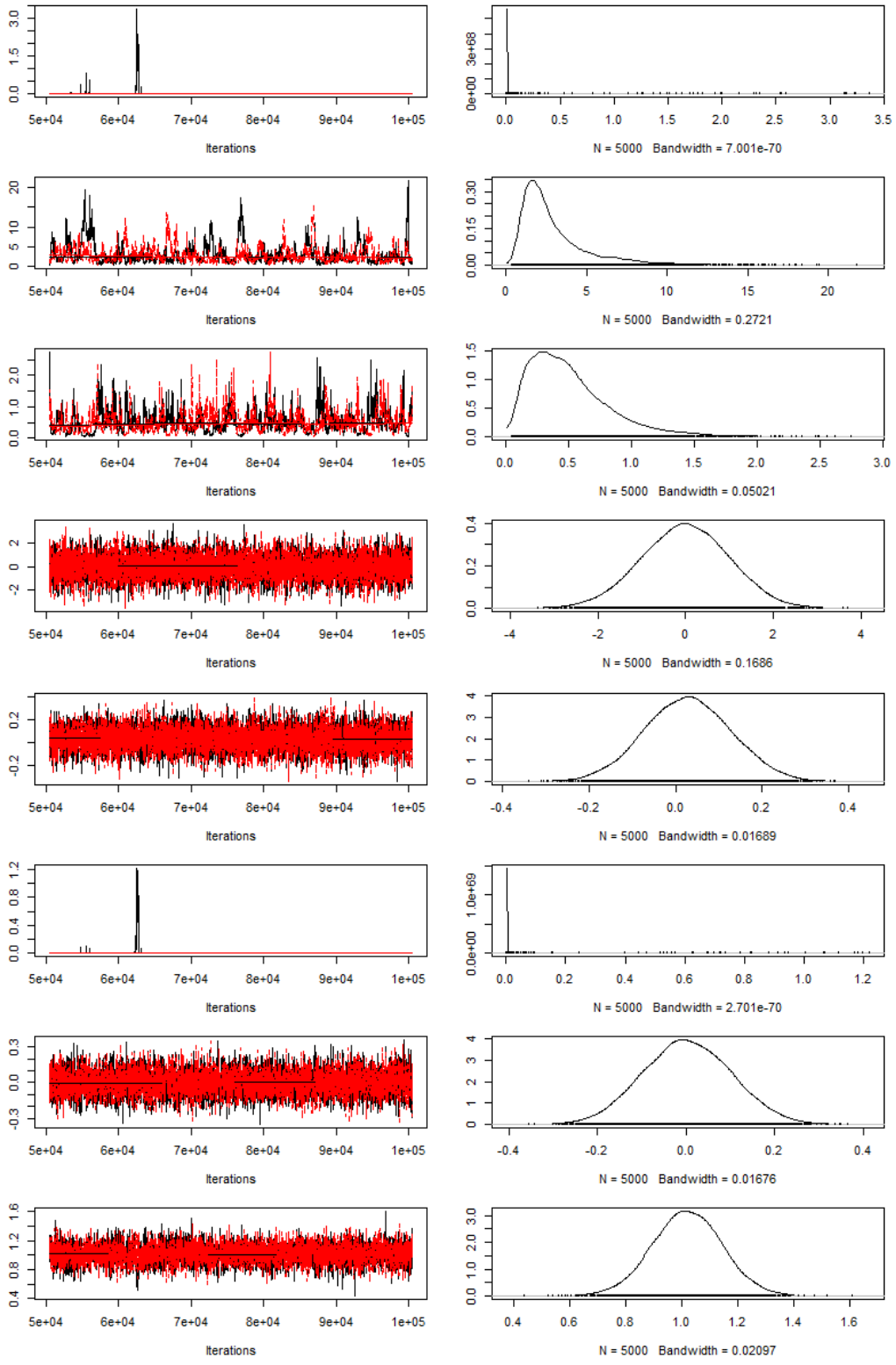

Figure B.2: Trace and density plots for the parameters a_{11}, a_{21}, $w_{0,1}$, $w_{1,1}$, γ, α_1, δ and β_1 of model 3.2, fitted to the log(chlorophyll$_a$) data for Lake Balaton for October 2008.

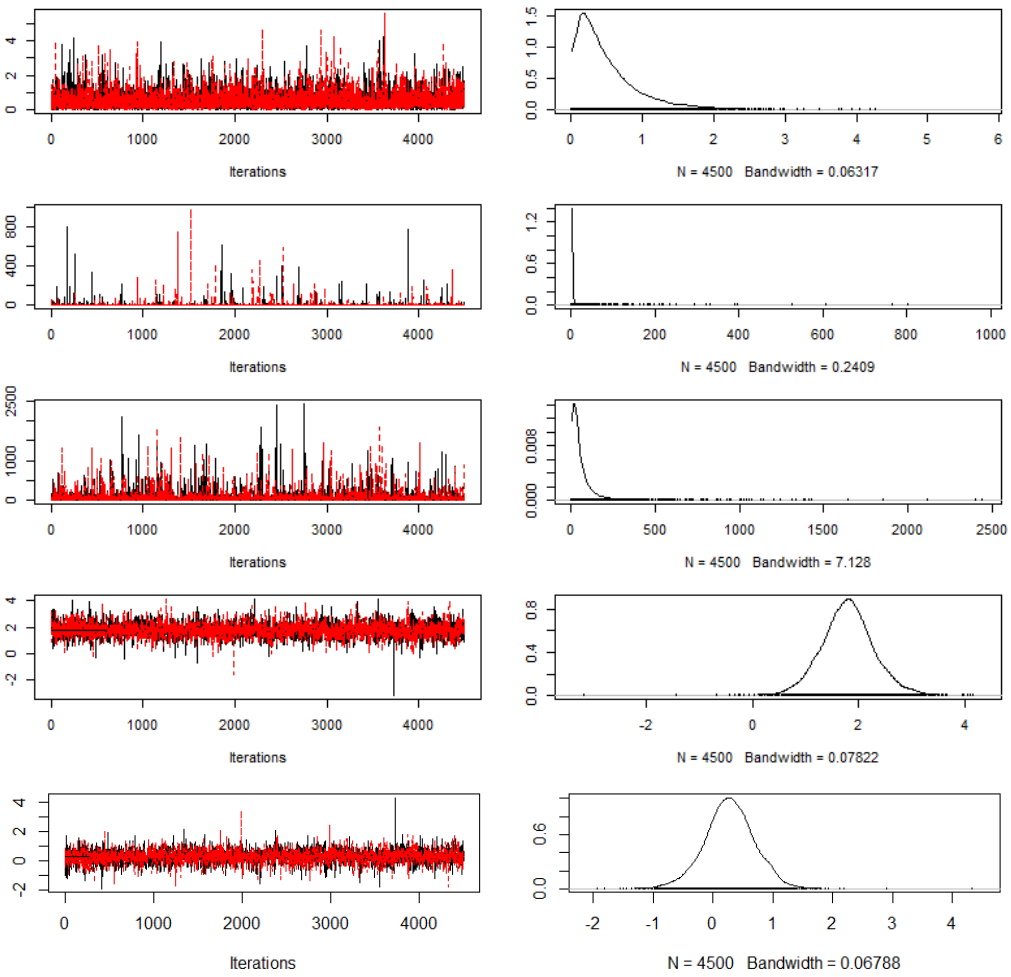

Figure B.3: Trace and density plots for the parameters $(\sigma^2_{\alpha 1})^{-1}$, $(\sigma^2_{\beta 1})^{-1}$, $(\sigma^2_{\varepsilon 1})^{-1}$, α_{11} and β_{11} of model 3.3, fitted to the log(chlorophyll$_a$) data for Lake Balaton.

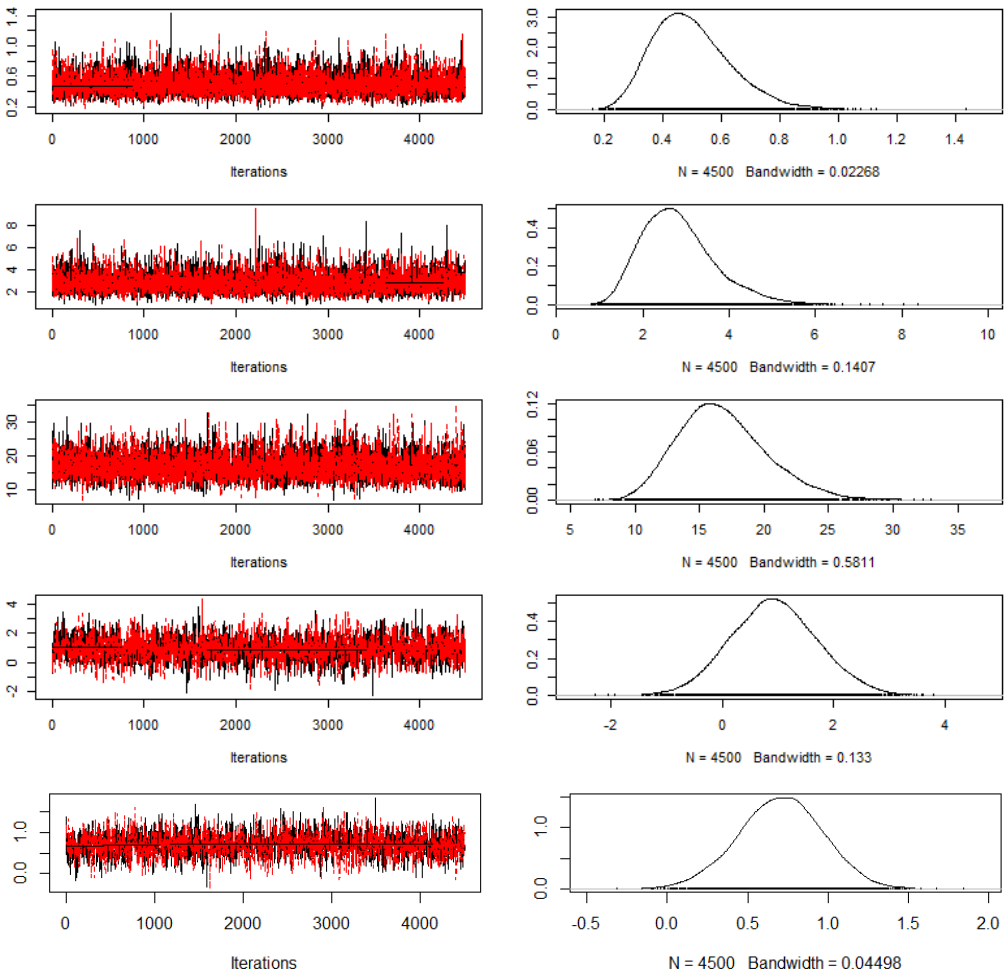

Figure B.4: Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\varepsilon^2)^{-1}$, α_{11} and β_{11} of model 3.3a, fitted to the log(chlorophyll$_a$) data for Lake Balaton.

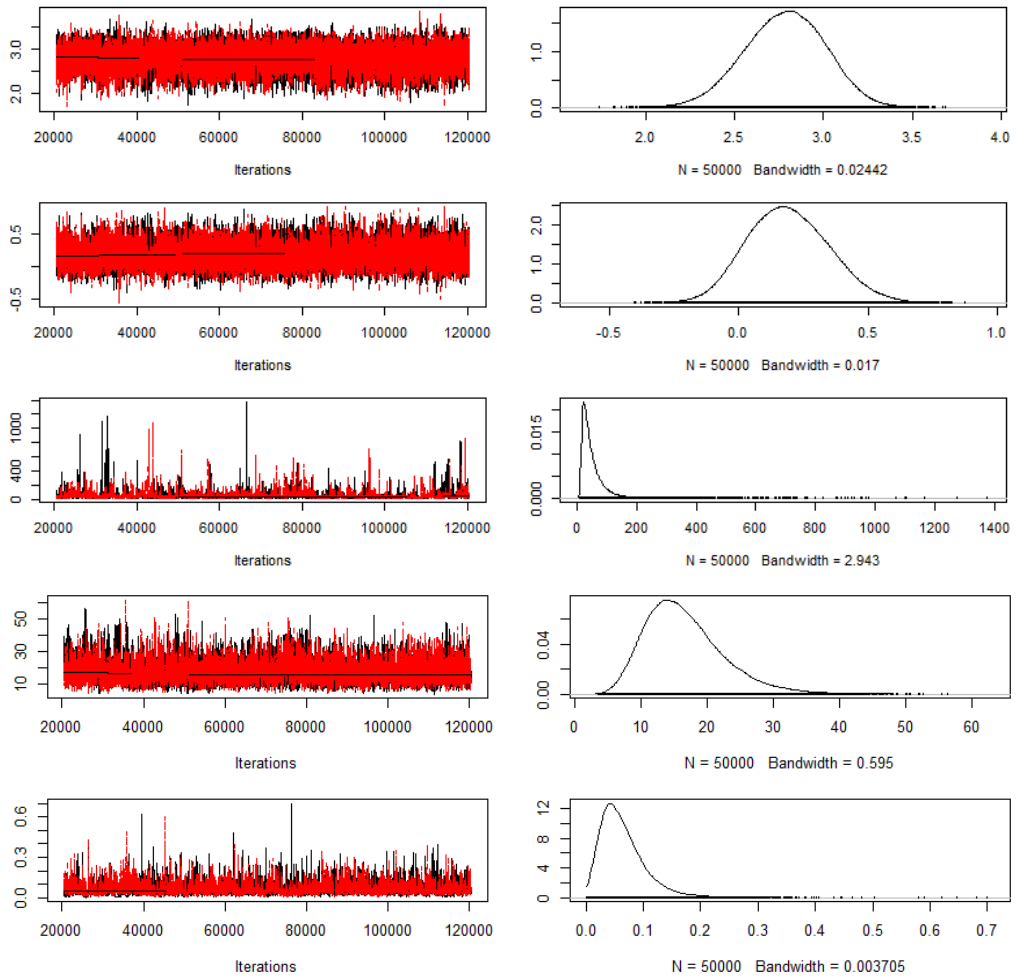

Figure B.5: Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_\varepsilon)^{-1}$, $(\sigma^2_\alpha)^{-1}$ and $(\sigma^2_\beta)^{-1}$ of model 3.5, fitted to the log(chlorophyll$_a$) data for Lake Balaton.

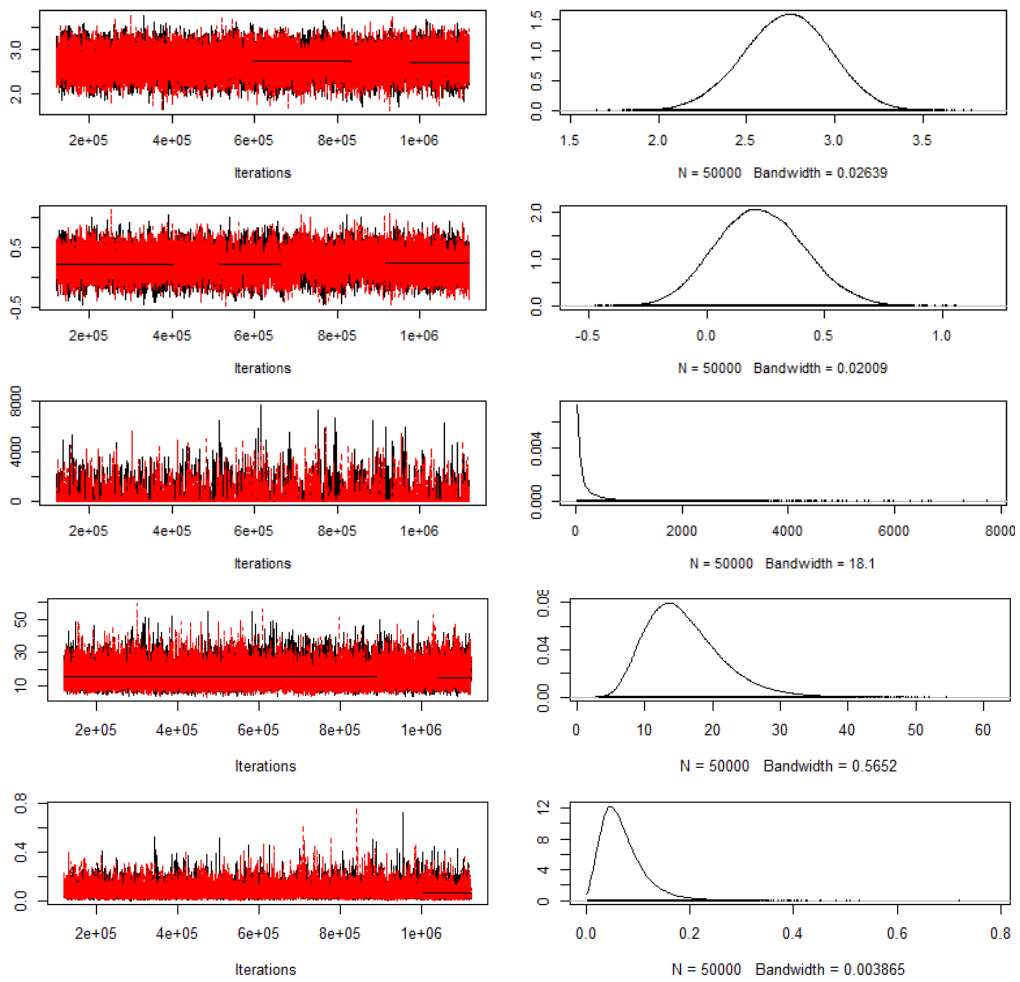

Figure B.6: Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_\varepsilon)^{-1}$, $(\sigma^2_\alpha)^{-1}$ and $(\sigma^2_\beta)^{-1}$ of model 3.5a, fitted to the log(chlorophyll$_a$) data for Lake Balaton.

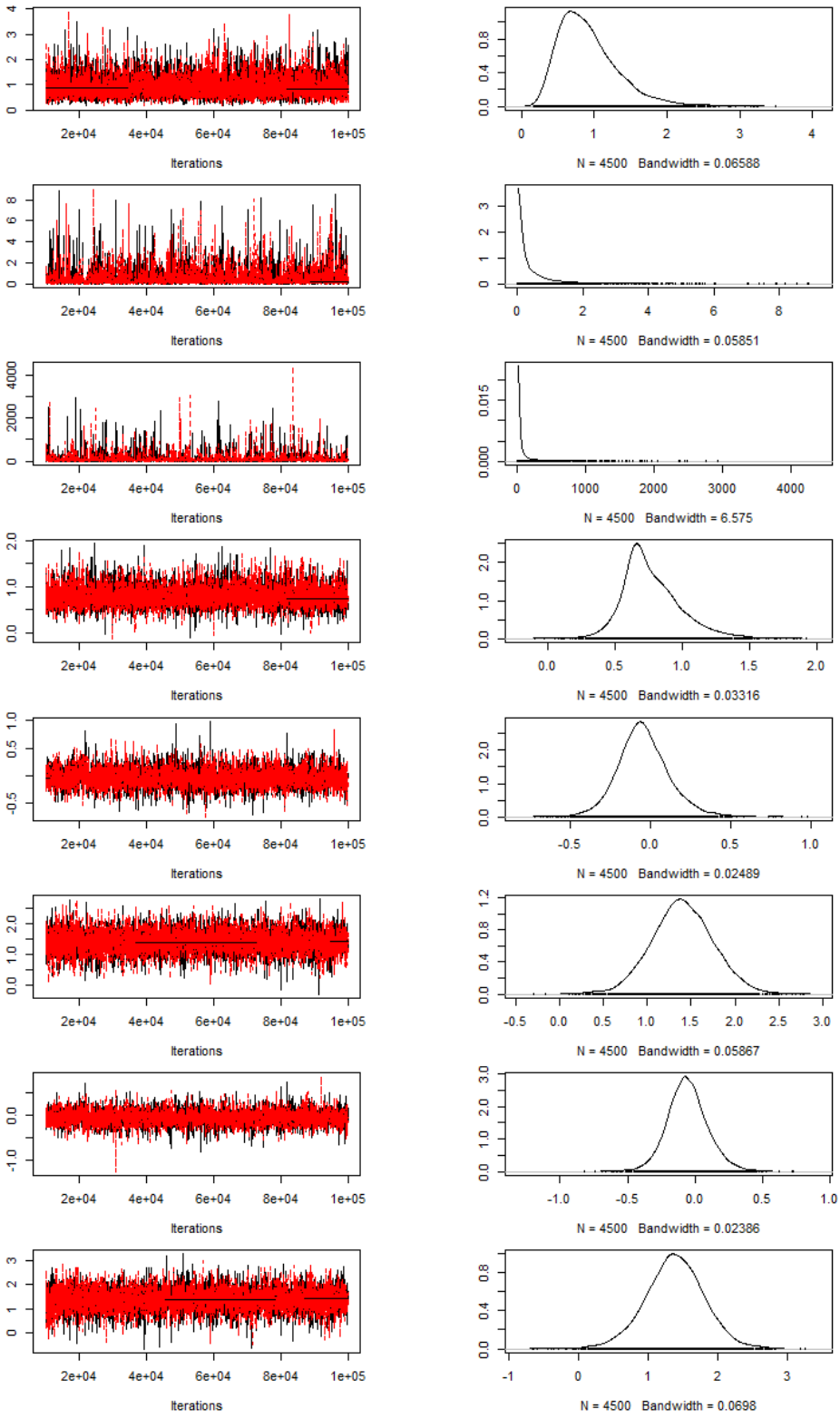

Figure B.7: Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\varepsilon^2)^{-1}$, α_1, β_1, $\tilde{\alpha}_1$, $\tilde{\beta}_1$ and $\tilde{y}_1$ of model 3.1, fitted to the log(chlorophyll$_a$) data for Lake Erie.

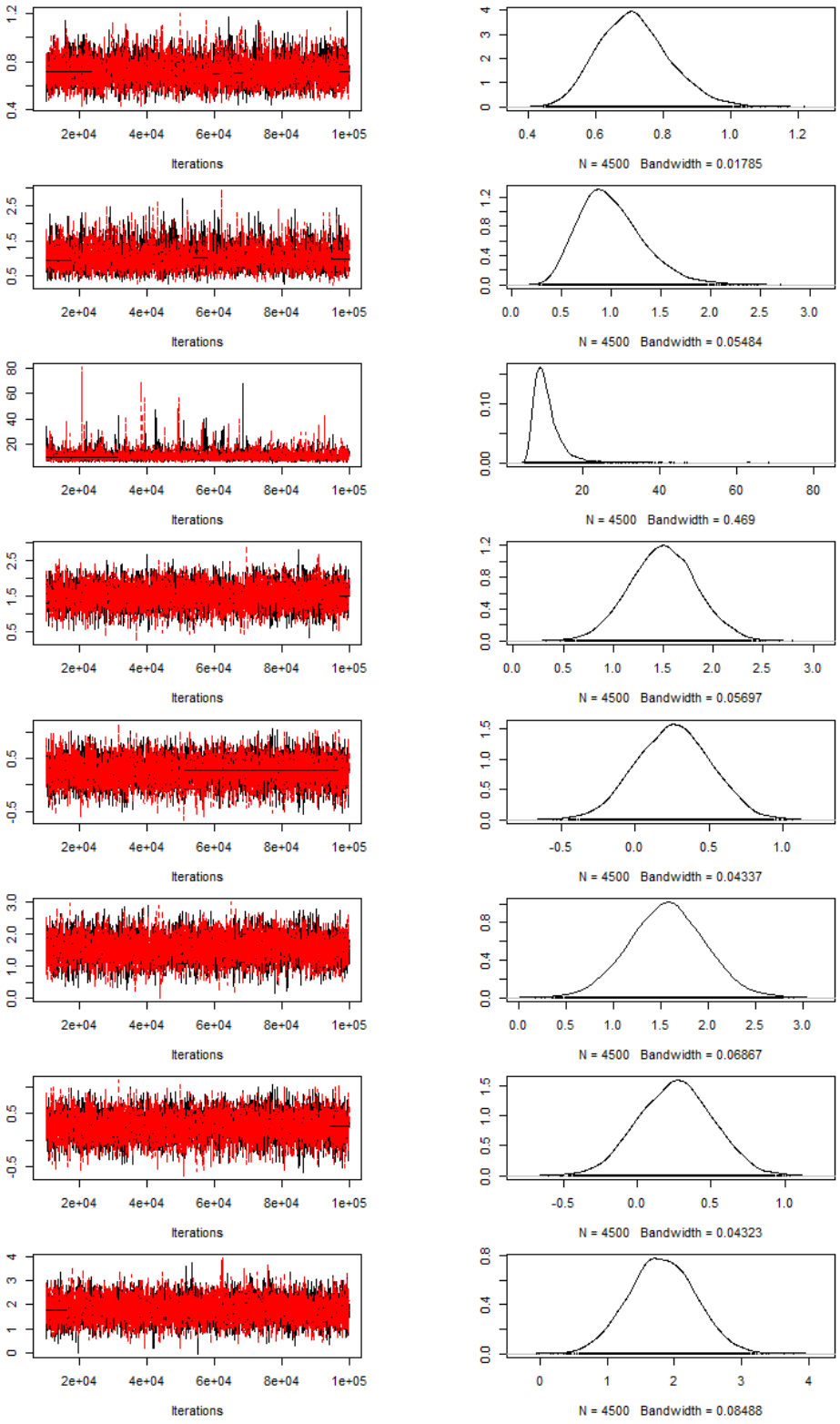

Figure B.8: Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_\varepsilon^2)^{-1}$, α_{11}, β_{11}, $\tilde{\alpha}_{11}$, $\tilde{\beta}_{11}$ and $\tilde{y}_{11}$ of model 3.3a, fitted to the $\log(\text{chlorophyll}_a)$ data for Lake Erie.

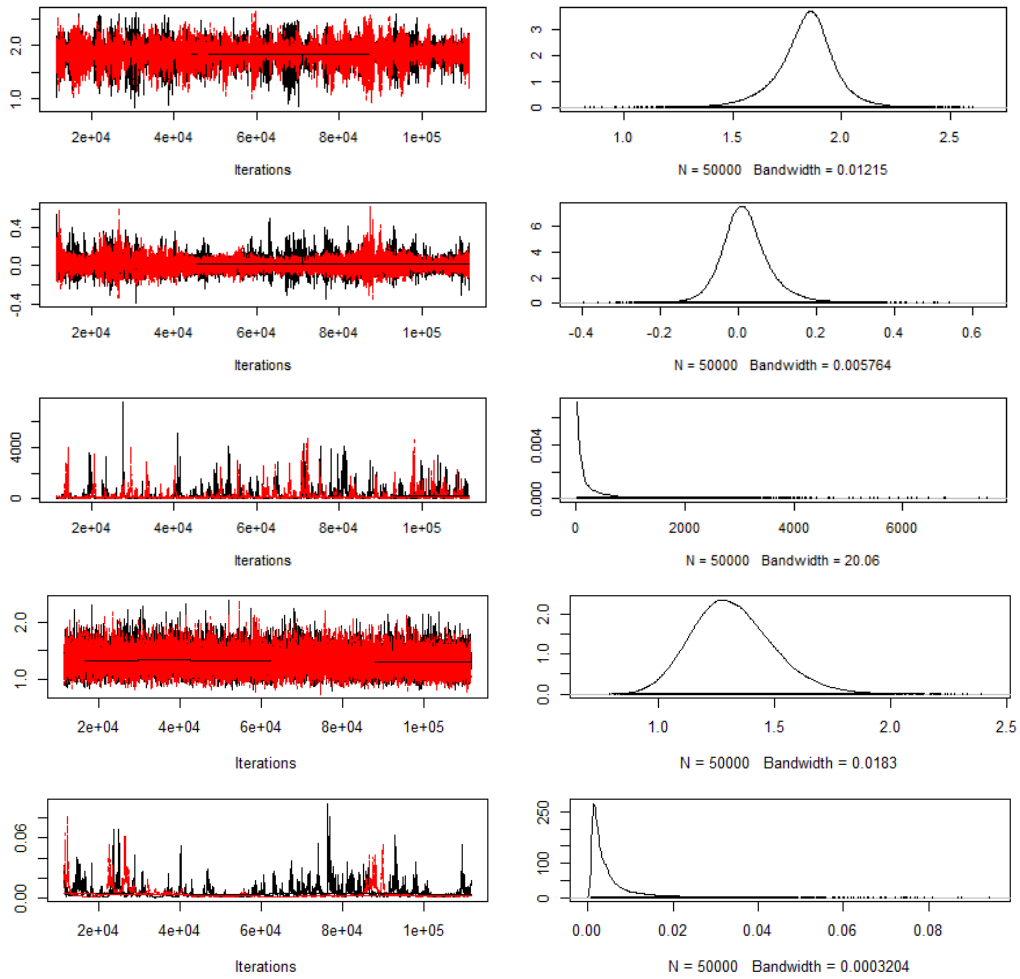

Figure B.9: Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma^2_\varepsilon)^{-1}$, $(\sigma^2_\alpha)^{-1}$ and $(\sigma^2_\beta)^{-1}$ of model 3.5, fitted to the log(chlorophyll$_a$) data for Lake Erie.

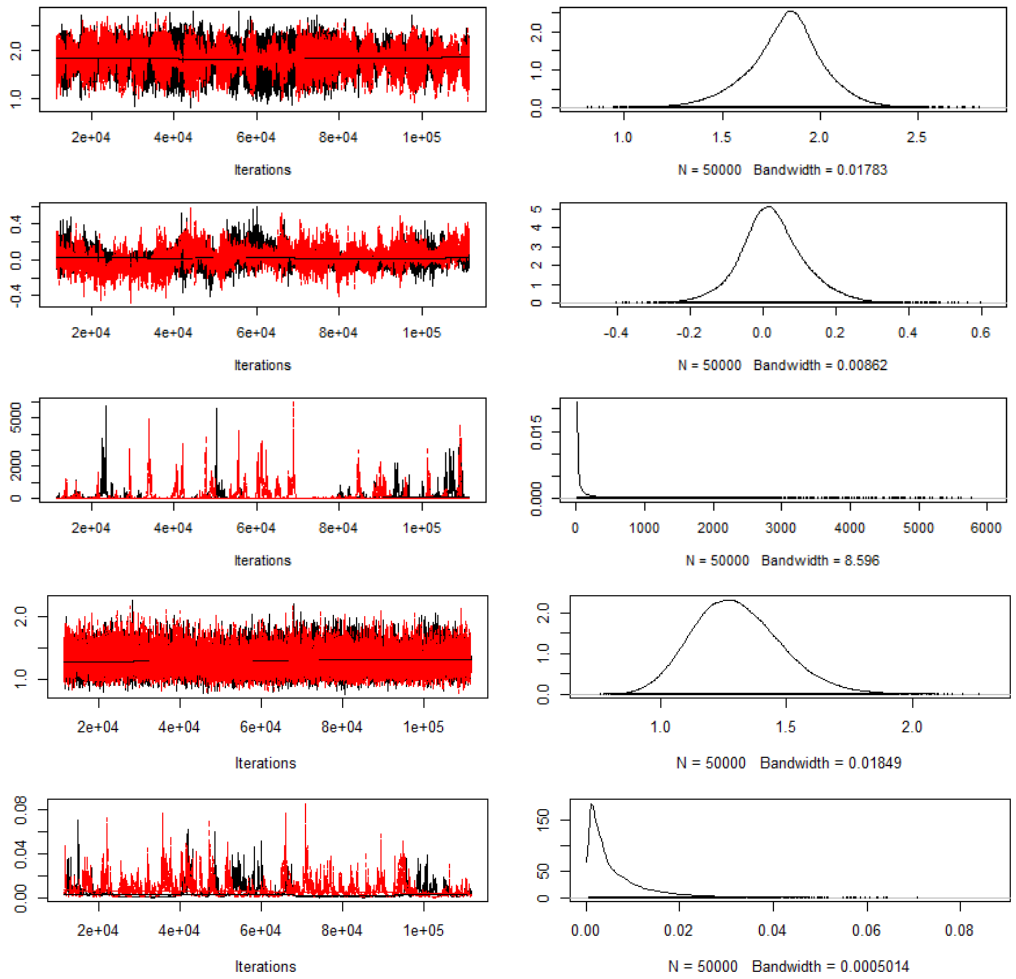

Figure B.10: Trace and density plots for the parameters α_{11}, β_{11}, $(\sigma_\varepsilon^2)^{-1}$, $(\sigma_\alpha^2)^{-1}$ and $(\sigma_\beta^2)^{-1}$ of model 3.5a, fitted to the log(chlorophyll$_a$) data for Lake Erie.

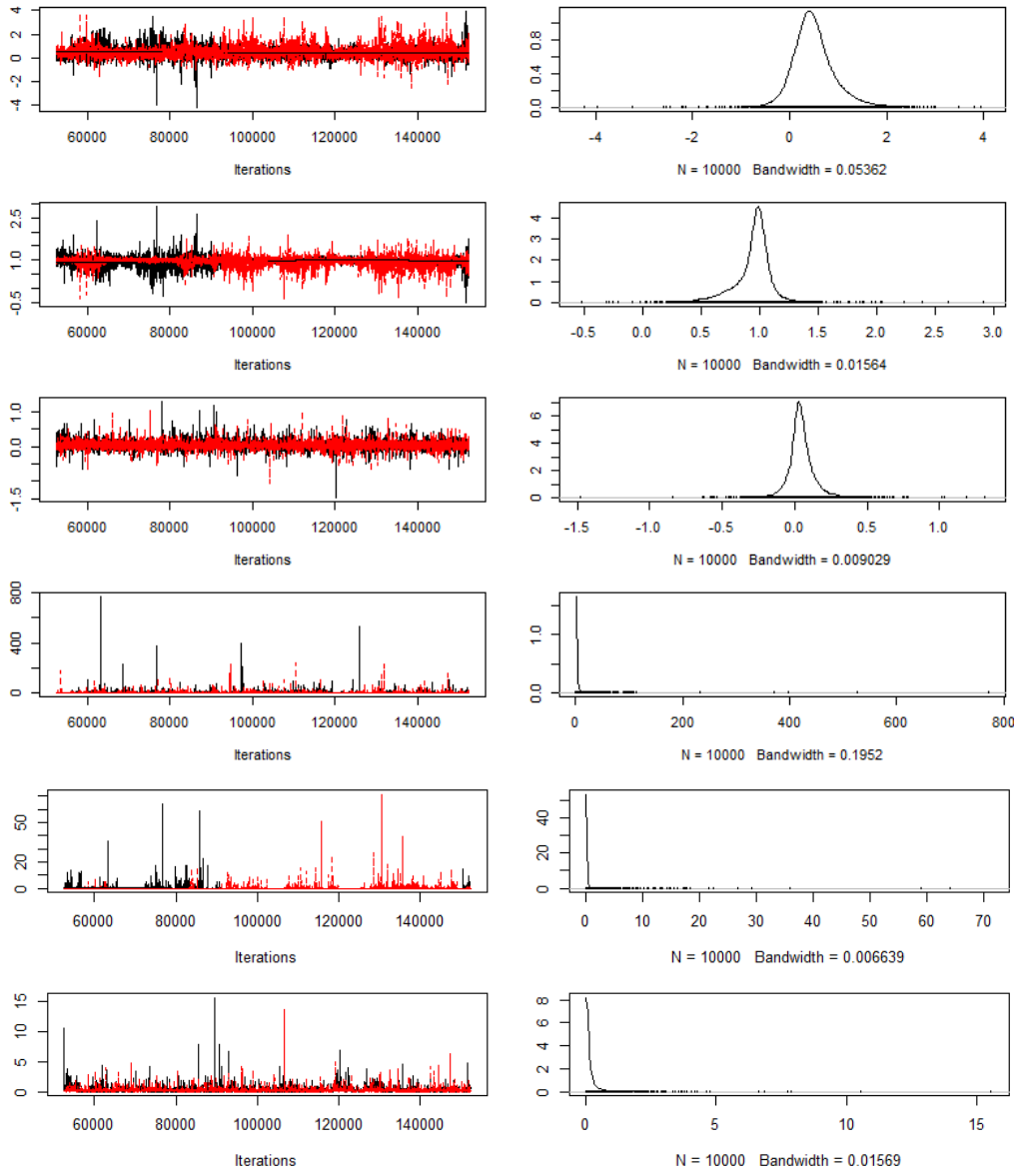

Figure B.11: Trace and density plots for the parameters α_{111}, β_{111}, ρ_1, $\sigma_{\alpha_{11}}$, $\sigma_{\beta_{11}}$ and $\sigma^2_{\varepsilon_1}$ of model 4.1, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton.

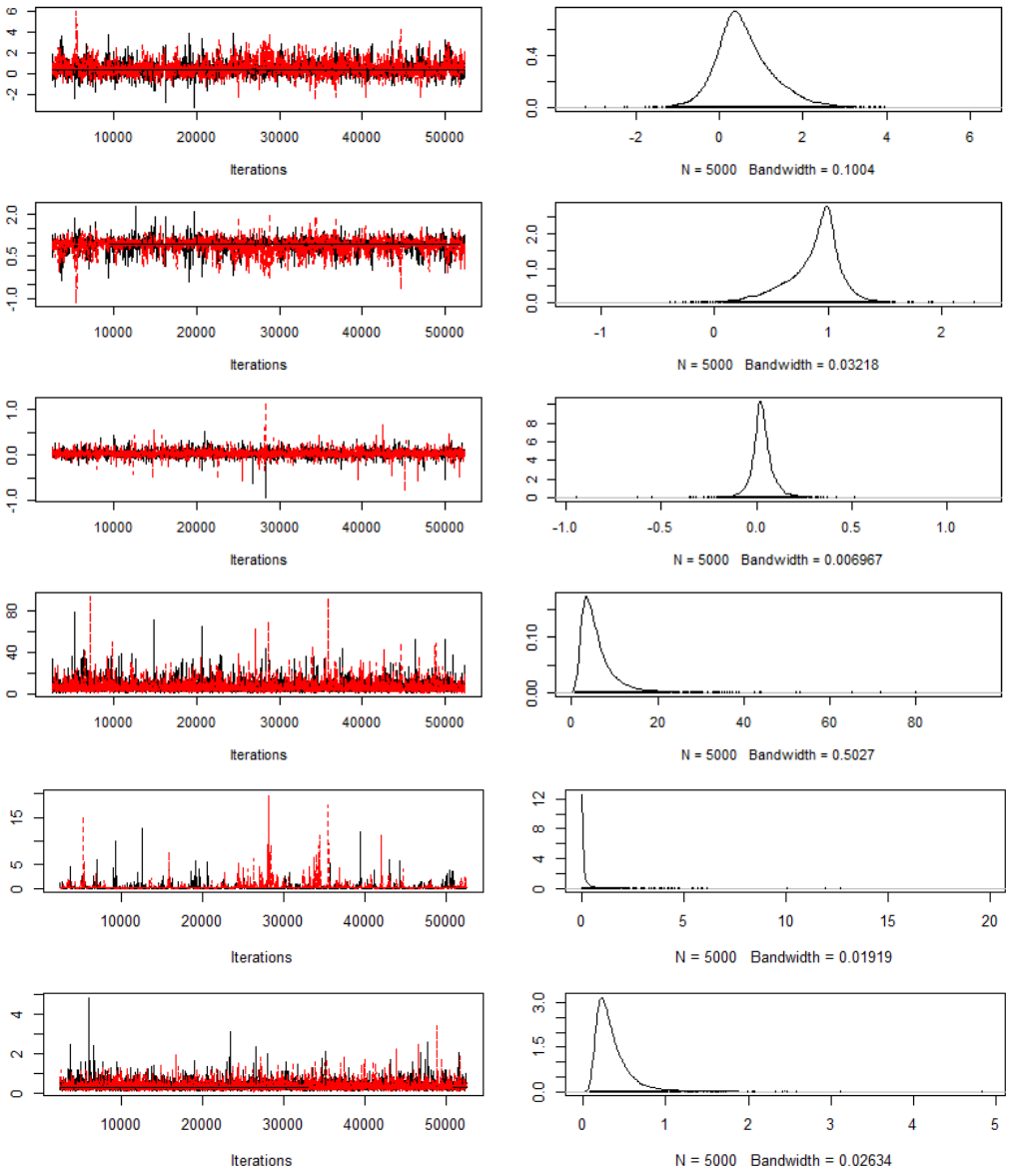

Figure B.12: Trace and density plots for the parameters α_{111}, β_{111}, ρ_1, $\sigma_{\alpha_{11}}$, $\sigma_{\beta_{11}}$ and $\sigma^2_{\varepsilon_1}$ of model 4.1a, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton.

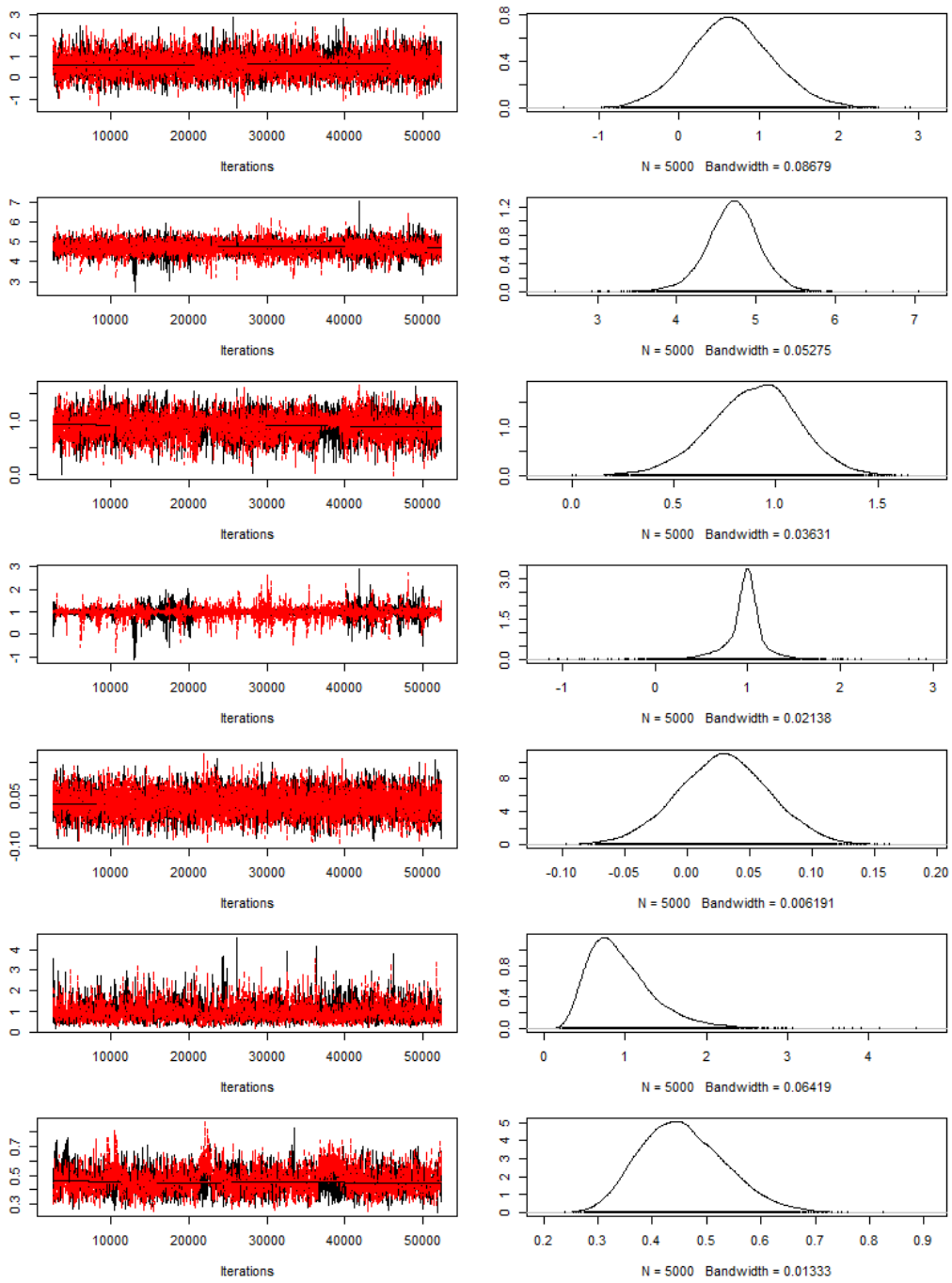

Figure B.13: Trace and density plots for the parameters $\alpha_{1,1,1}$, $\alpha_{1,1,2}$, $\beta_{1,1,1}$, $\beta_{1,1,2}$, ρ, $\sigma^{-1}_{\alpha^2_1}$ and $(\sigma^2_{\varepsilon,1})^{-1}$ of model 4.2, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton.

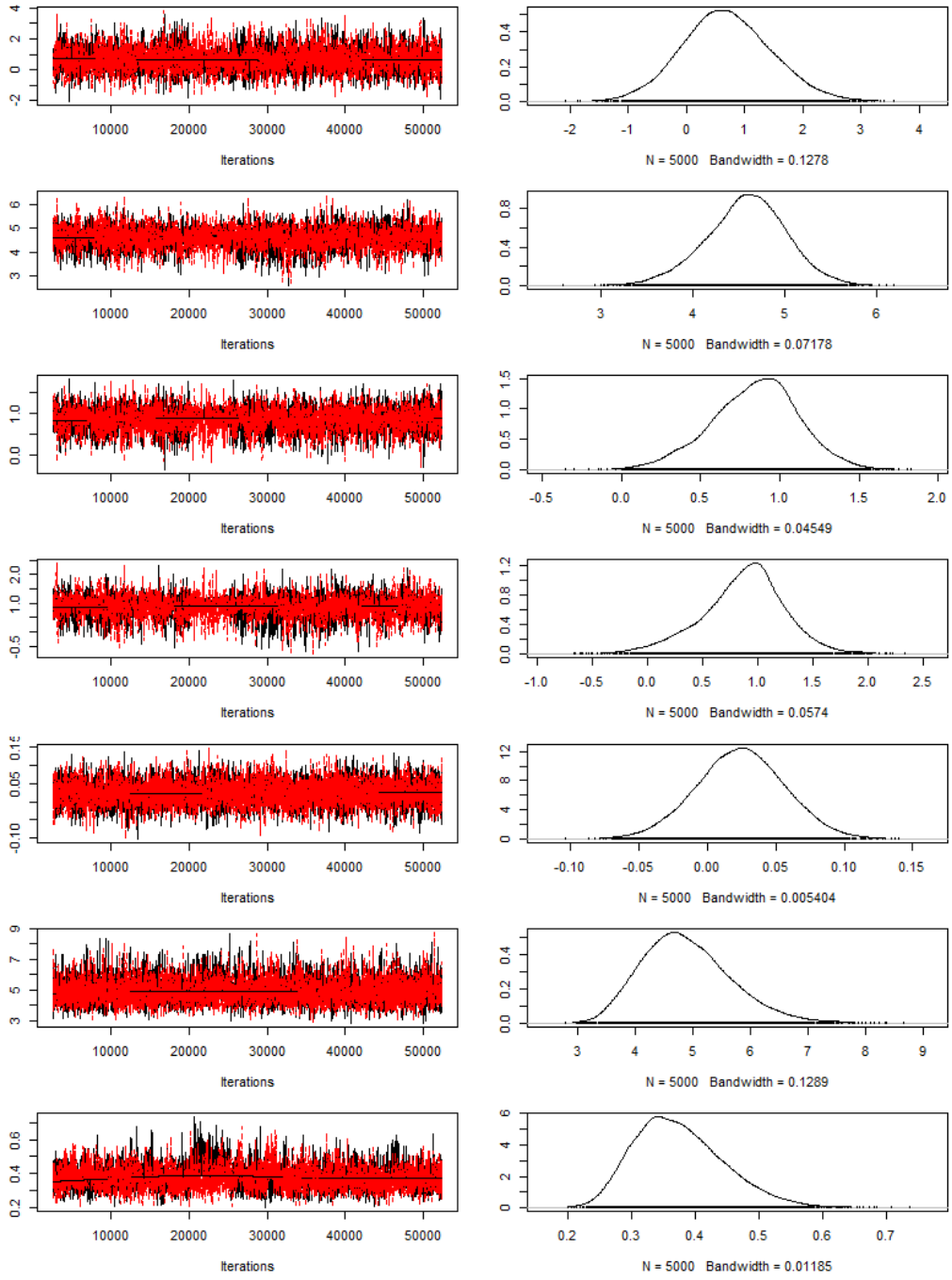

Figure B.14: Trace and density plots for the parameters $\alpha_{1,1,1}$, $\alpha_{1,1,2}$, $\beta_{1,1,1}$, $\beta_{1,1,2}$, ρ, $\sigma^{-1}_{\alpha^2_1}$ and $(\sigma^2_{\varepsilon,1})^{-1}$ of model 4.2a, fitted to the log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton.

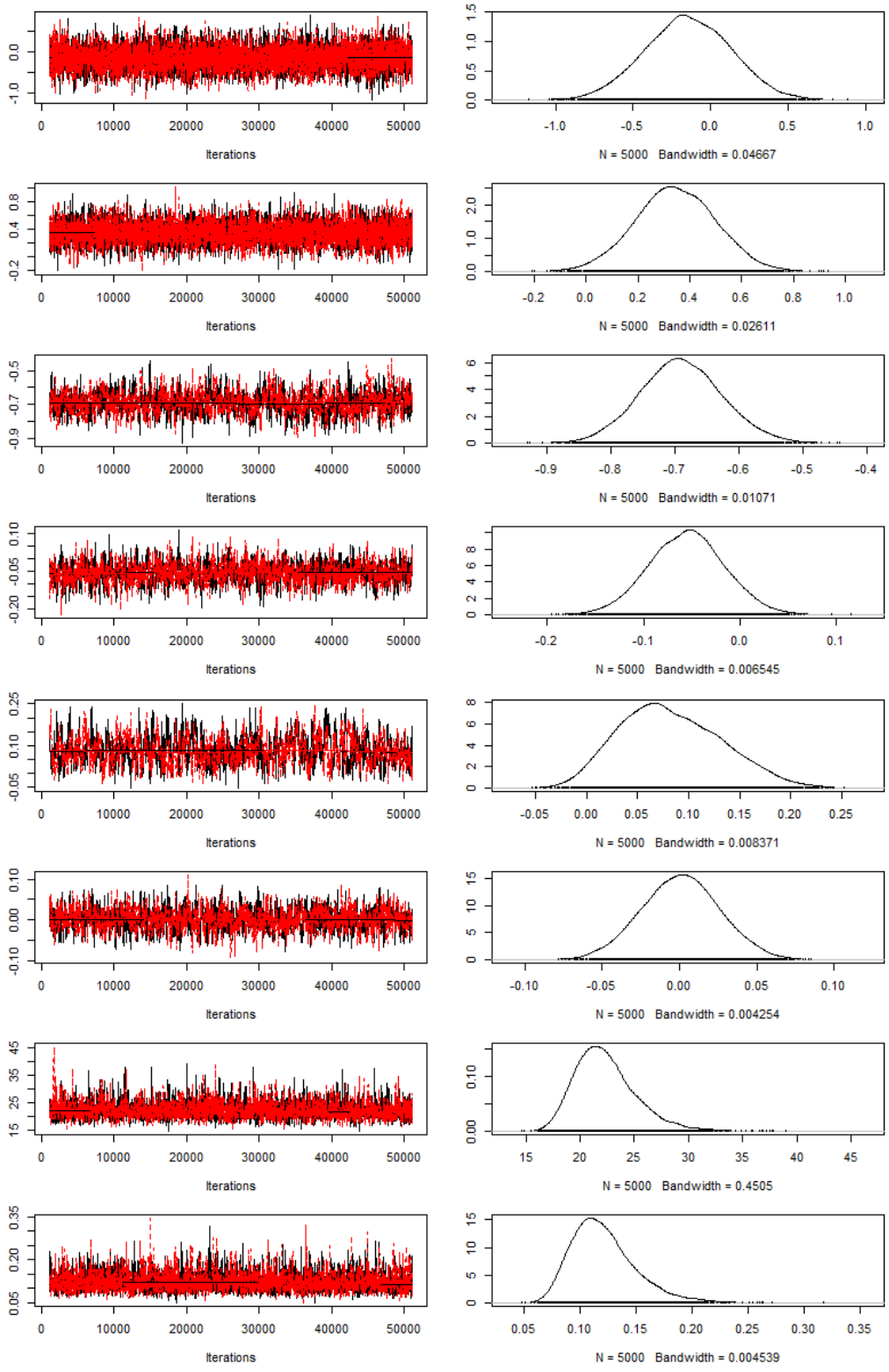

Figure B.15: Trace and density plots for the parameters $y_{1,1(1)}$, $\gamma_{1,1(1)}$, β_1, α, η_1, δ, $(\sigma_\varepsilon^2)^{-1}$ and $(\sigma_{\gamma_1}^2)^{-1}$ of model 4.4a-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes.

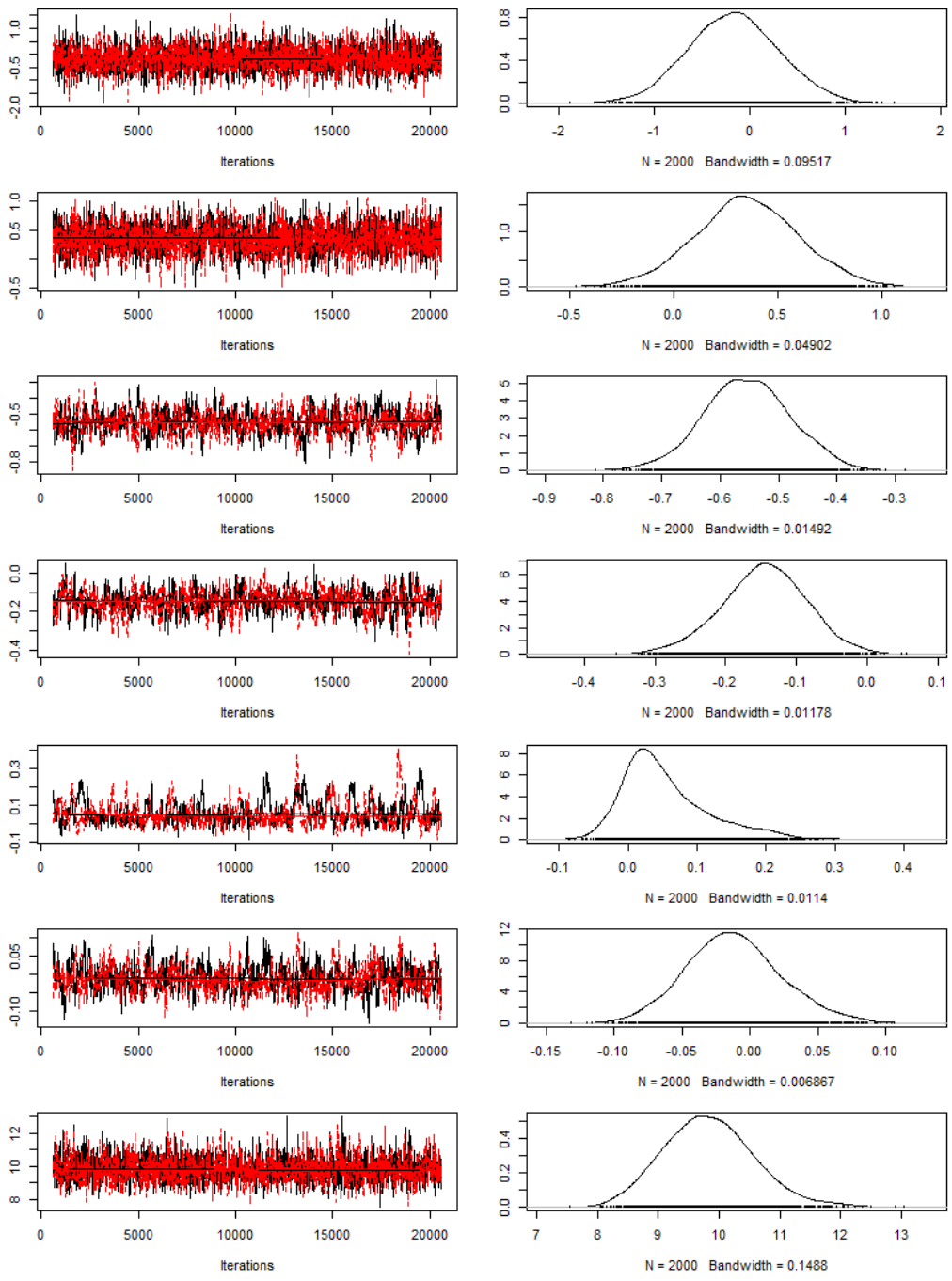

Figure B.16: Trace and density plots for the parameters $\tilde{y}_{1,1}$, $\gamma_{1,1}$, $\beta_{1,1}$, α, η_1, δ and $(\sigma_\varepsilon^2)^{-1}$ of model 4.4b-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes.

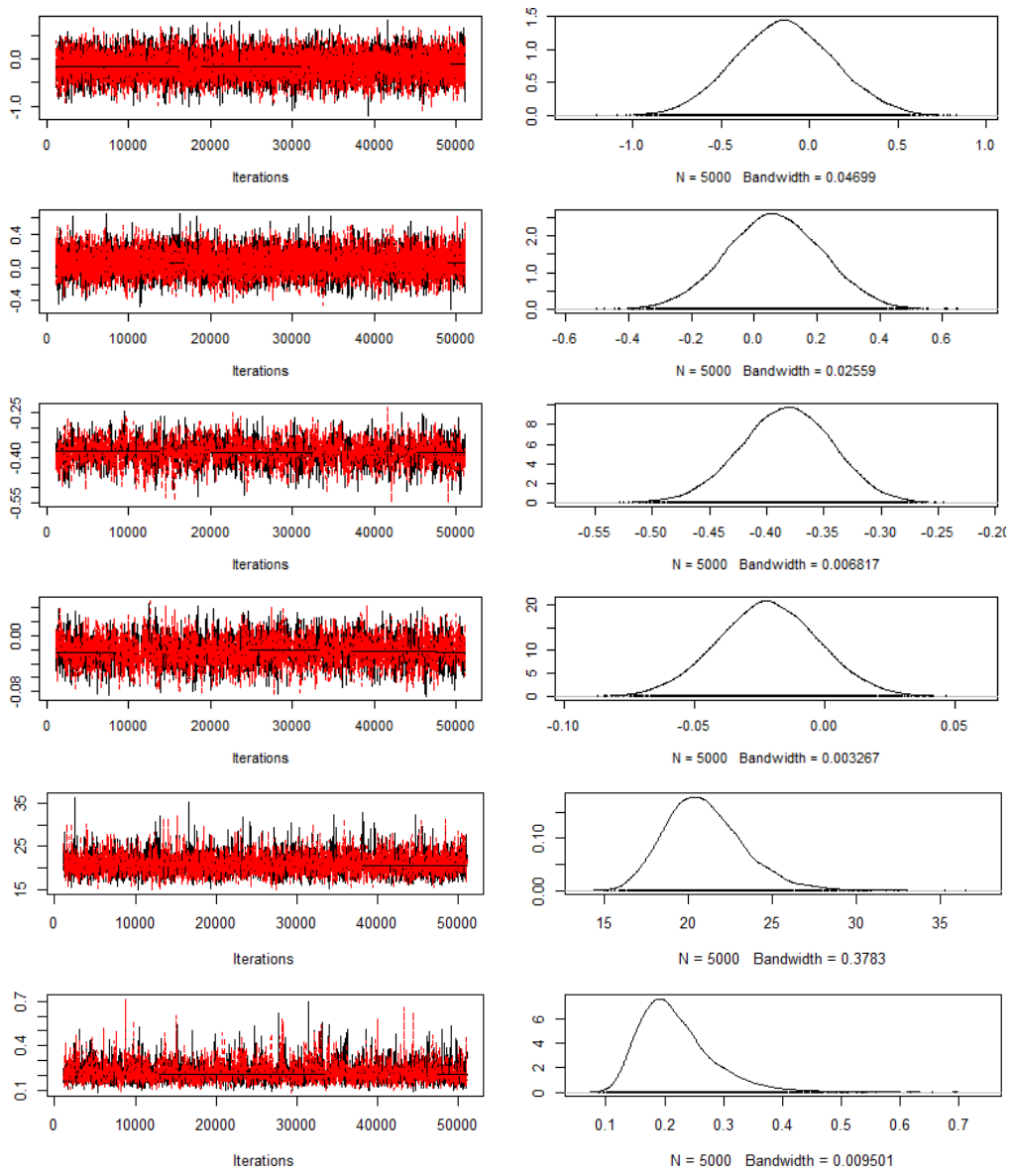

Figure B.17: Trace and density plots for the parameters $y_{1,1(1)}$, $\gamma_{1,1(1)}$, α, δ, $(\sigma_\varepsilon^2)^{-1}$ and $(\sigma_{\gamma_1}^2)^{-1}$ of model 4.5a-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes.

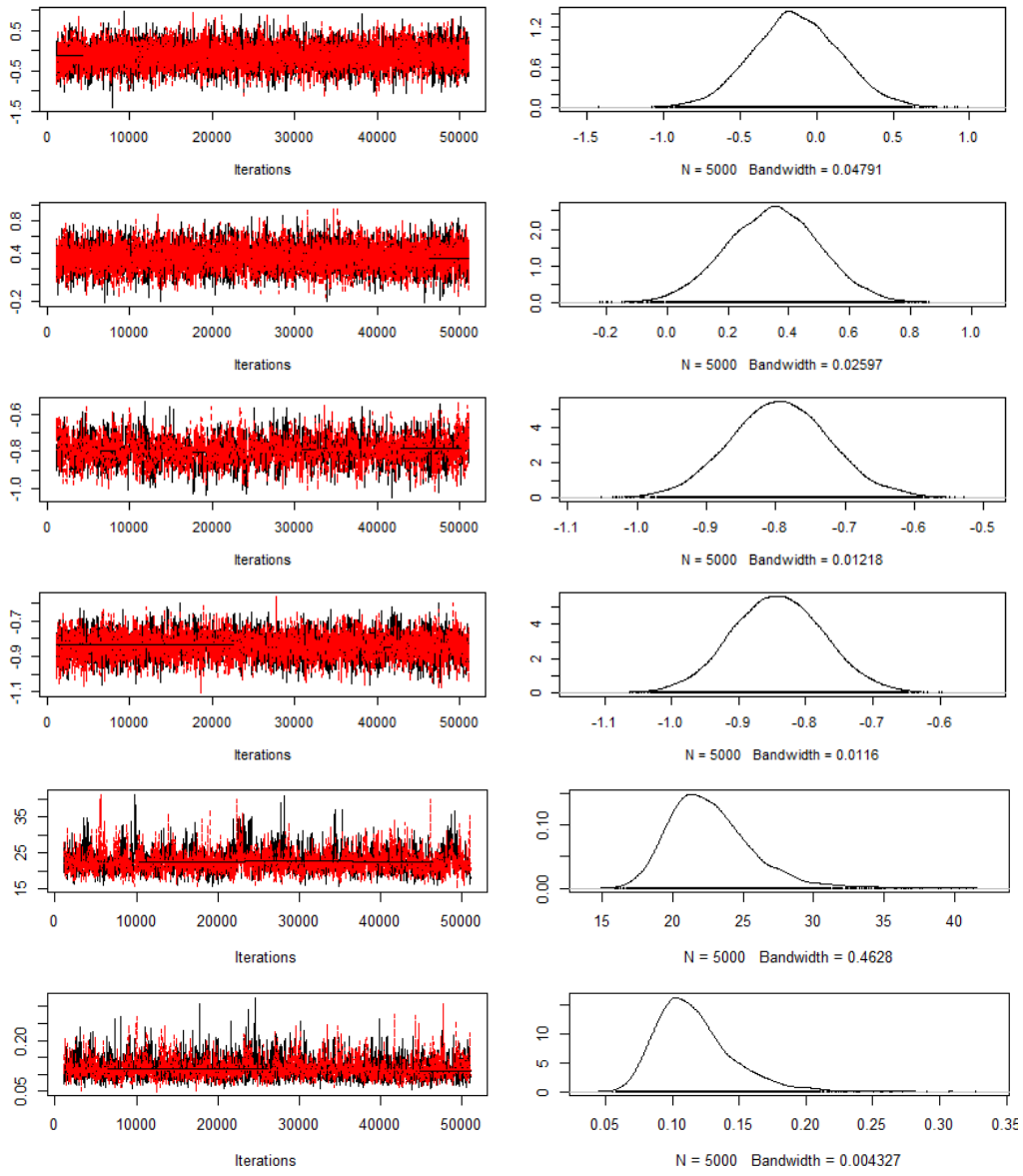

Figure B.18: Trace and density plots for the parameters $y_{1,1(1)}$, $\gamma_{1,1(1)}$, β_1, η_1, $(\sigma_\varepsilon^2)^{-1}$ and $(\sigma_{\gamma_1}^2)^{-1}$ of model 4.6-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes.

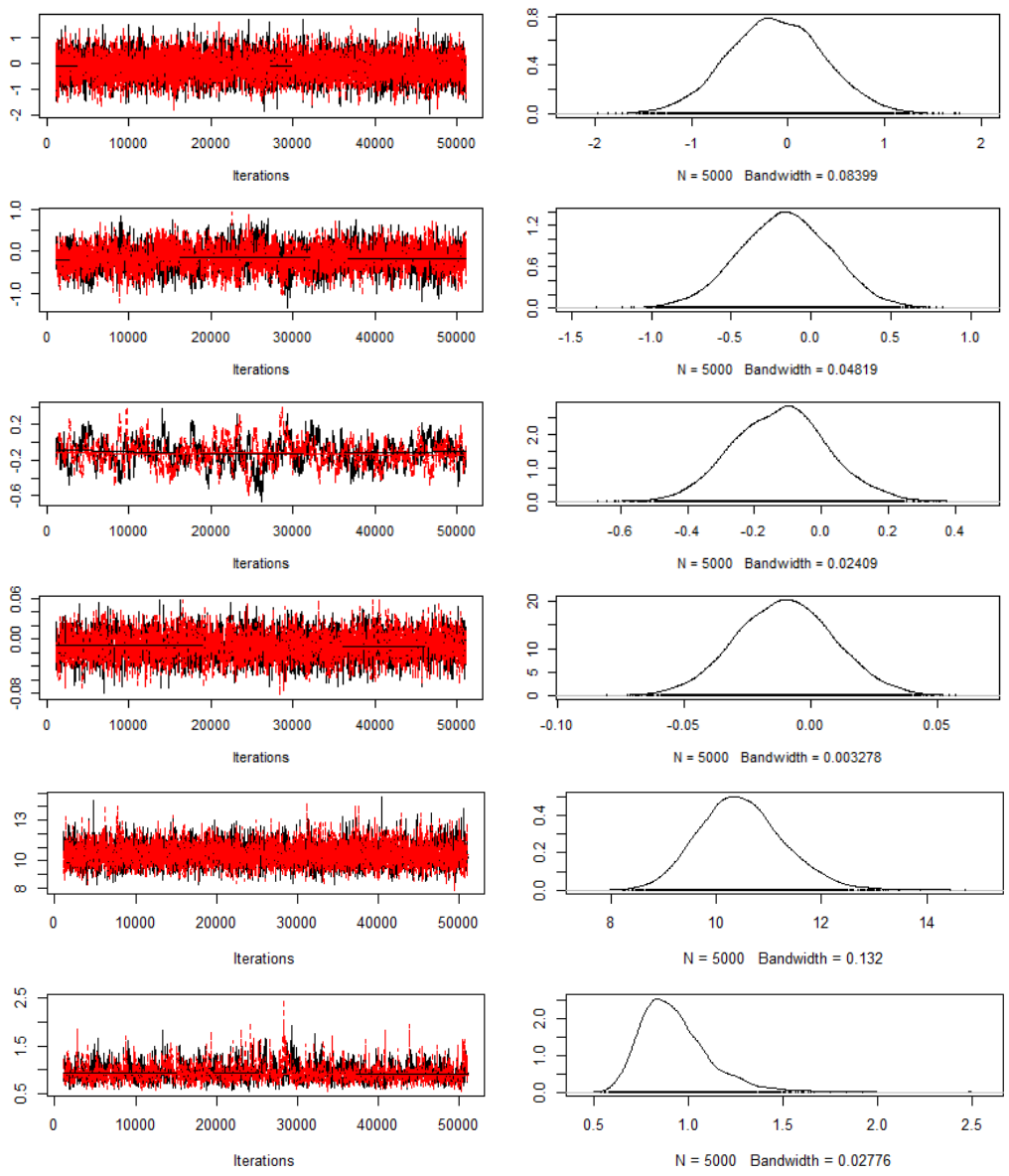

Figure B.19: Trace and density plots for the parameters $\tilde{y}_{1,1}$, $\gamma_{1,1}$, β, η, $(\sigma_\varepsilon^2)^{-1}$ and σ_γ^2 of model 4.7-ST, fitted to the log(chlorophyll$_a$) data for the Great Lakes.

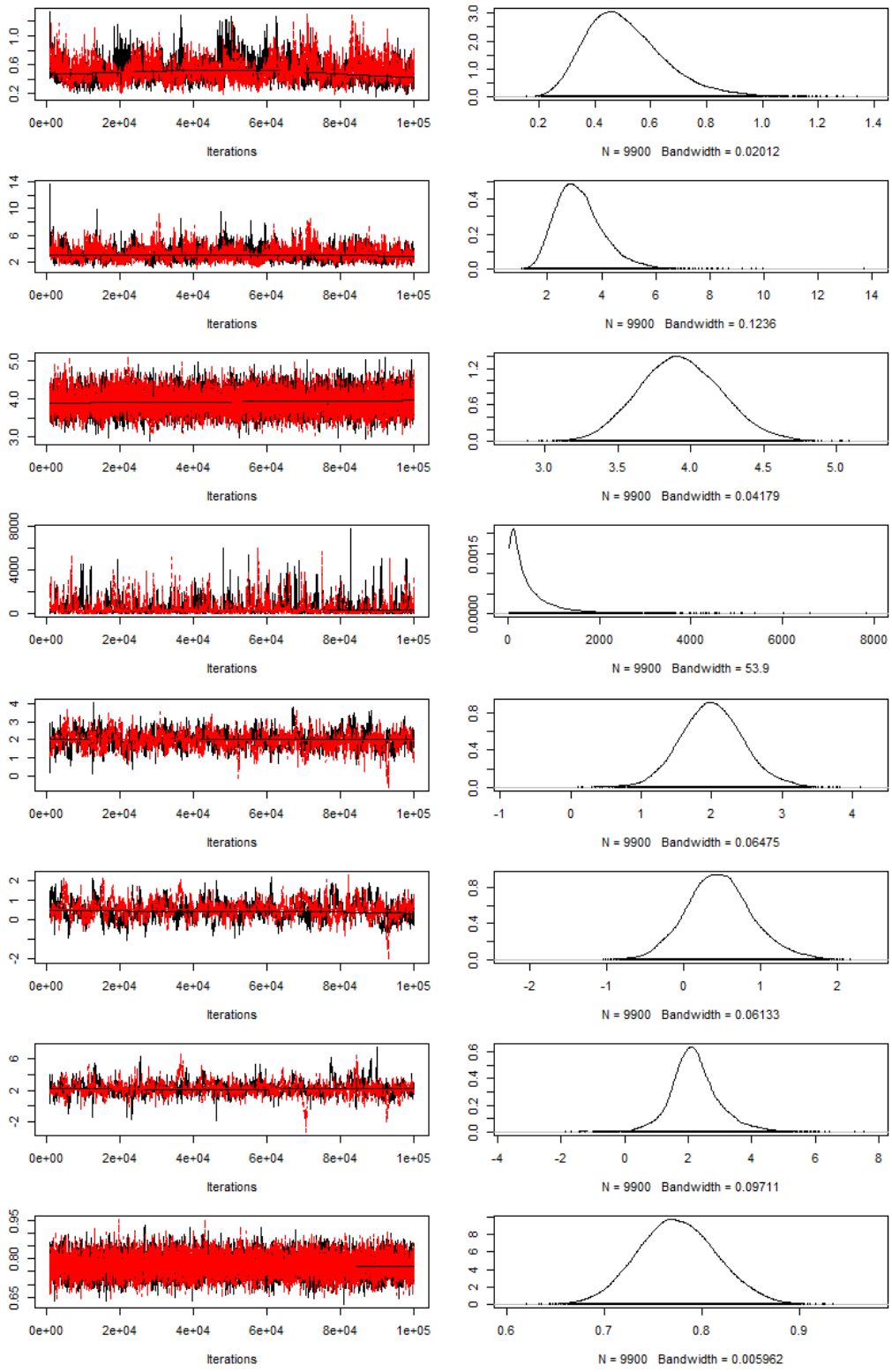

Figure B.20: Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_y^2)^{-1}$, $(\sigma_c^2)^{-1}$, $\alpha_{1,1}$, $\beta_{1,1}$, $c_{1,1}$ and $(\sigma_x^2)^{-1}$ of model 5.8, fitted to the $\log(\text{chlorophyll}_a)$ data for Lake Balaton.

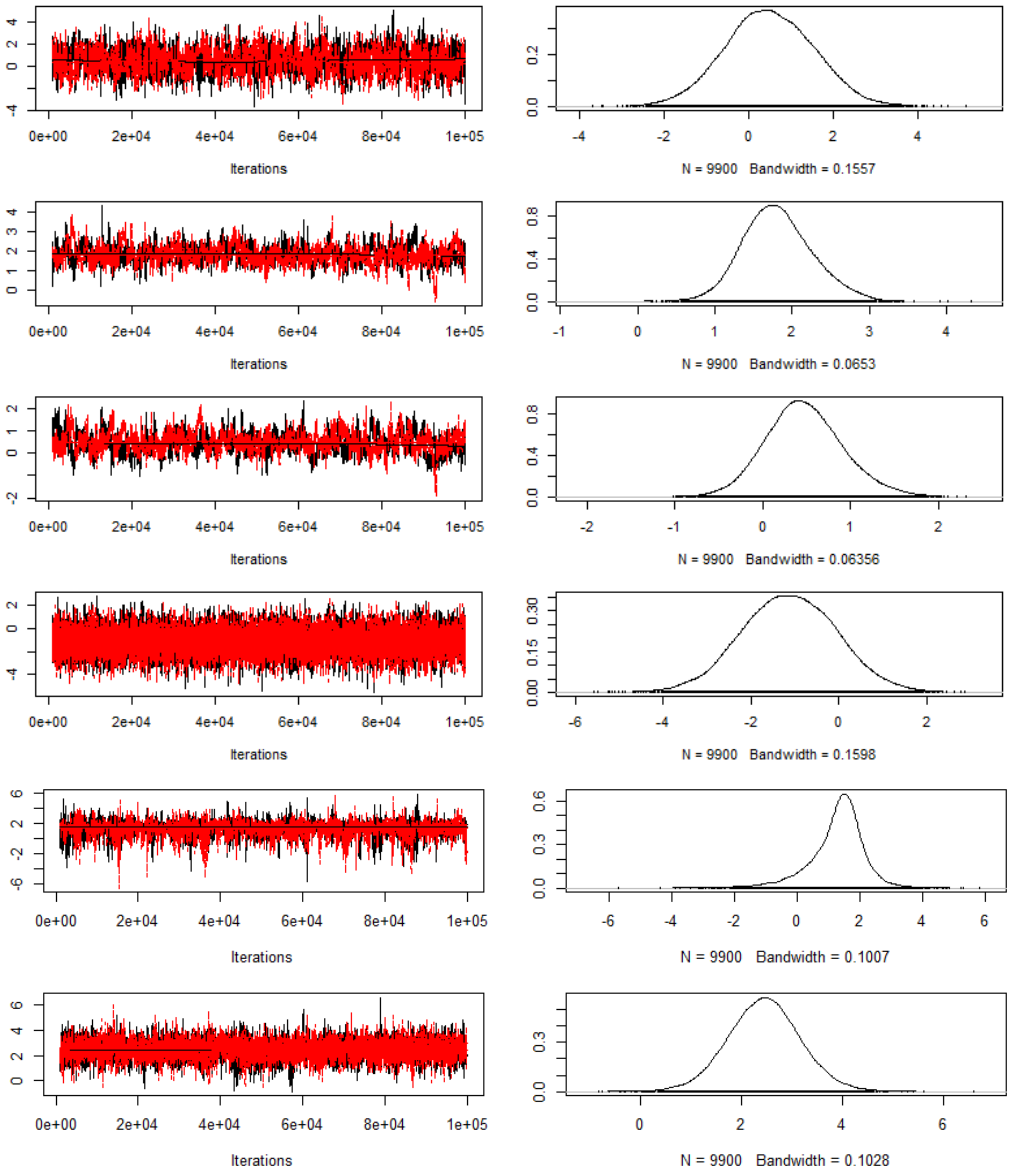

Figure B.21: Trace and density plots for the parameters $d_{1,1}$, $\tilde{\alpha}_{1,1}$, $\tilde{\beta_{1,1}}$, $\tilde{d}_{1,1}$, $\tilde{c}_{1,1}$ and $\tilde{y}_{1,1}$ of model 5.8, fitted to the log(chlorophyll$_a$) data for Lake Balaton.

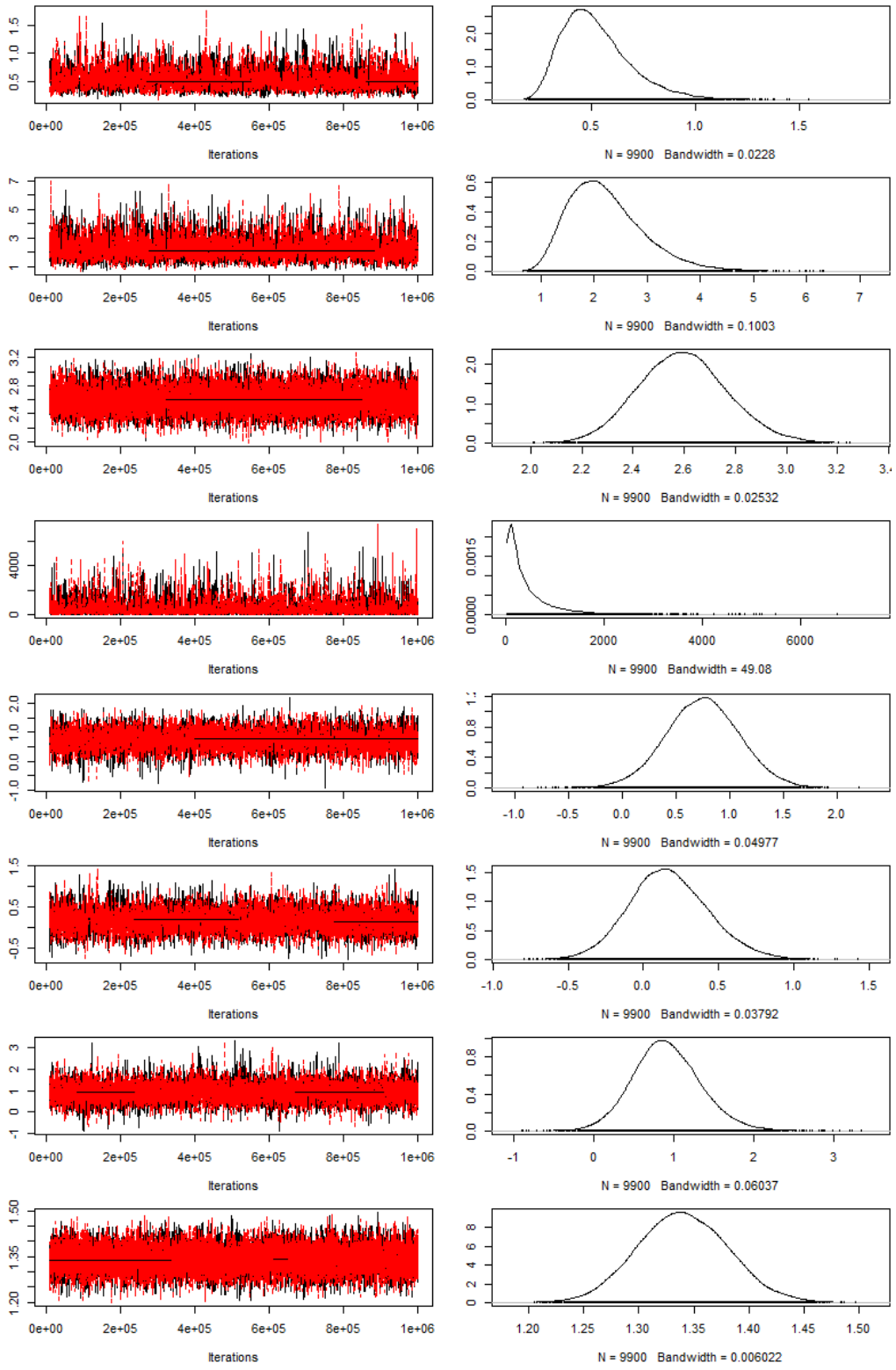

Figure B.22: Trace and density plots for the parameters $(\sigma_\alpha^2)^{-1}$, $(\sigma_\beta^2)^{-1}$, $(\sigma_y^2)^{-1}$, $(\sigma_c^2)^{-1}$, $\alpha_{1,1}$, $\beta_{1,1}$, $c_{1,1}$ and $(\sigma_x^2)^{-1}$ of model 5.8, fitted to the $\log(\text{chlorophyll}_a)$ data for Lake Erie.

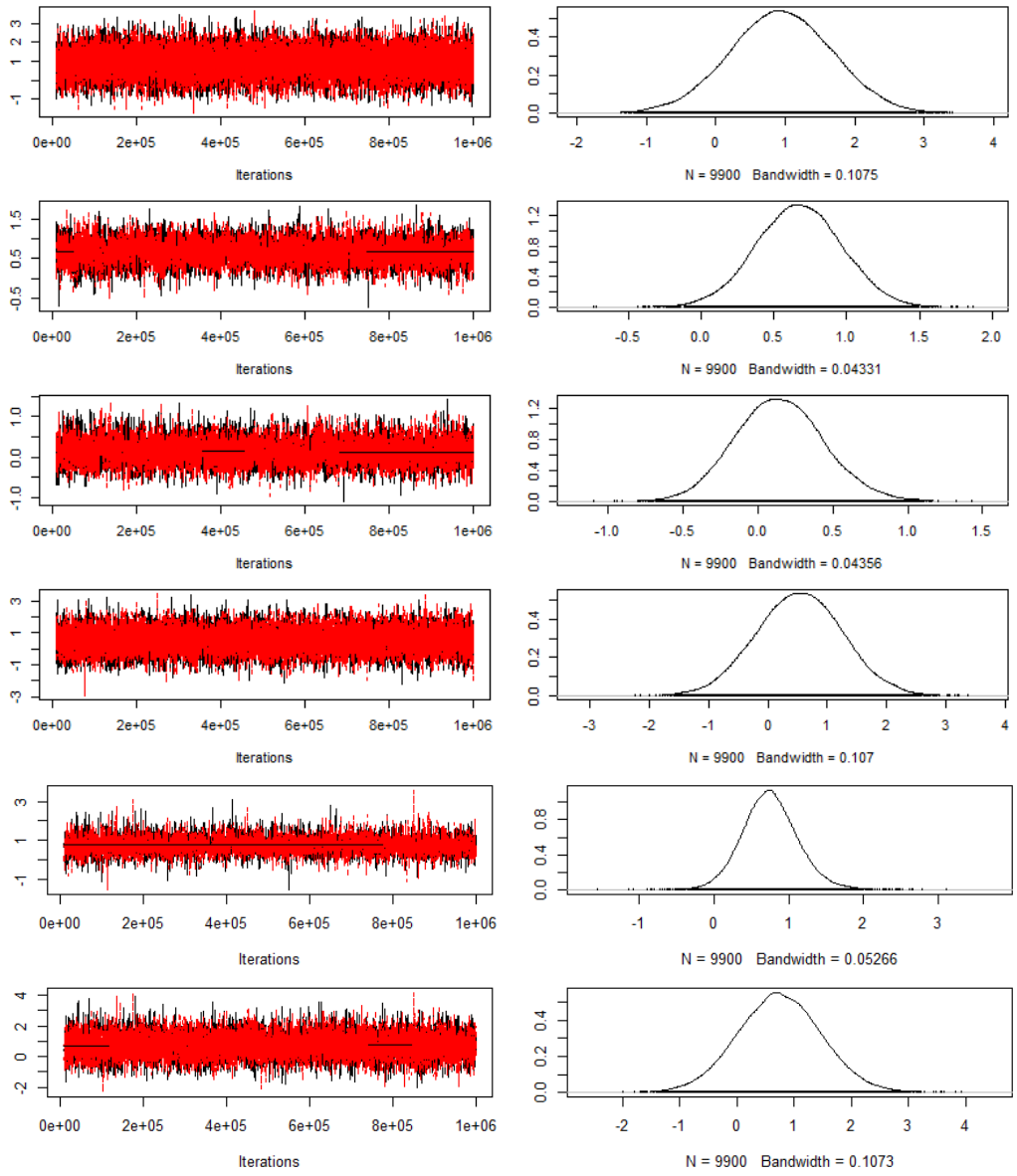

Figure B.23: Trace and density plots for the parameters $d_{1,1}$, $\tilde{\alpha}_{1,1}$, $\tilde{\beta_{1,1}}$, $\tilde{d}_{1,1}$, $\tilde{c}_{1,1}$ and $\tilde{y}_{1,1}$ of model 5.8, fitted to the log(chlorophyll$_a$) data for Lake Erie.

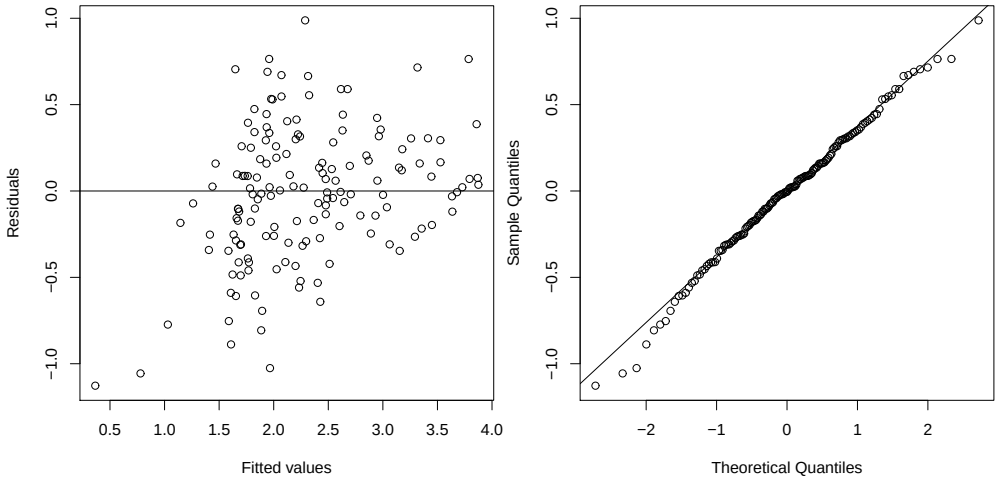

Figure B.24: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.1, fitted to log(chlorophyll$_a$) data, for Lake Balaton.

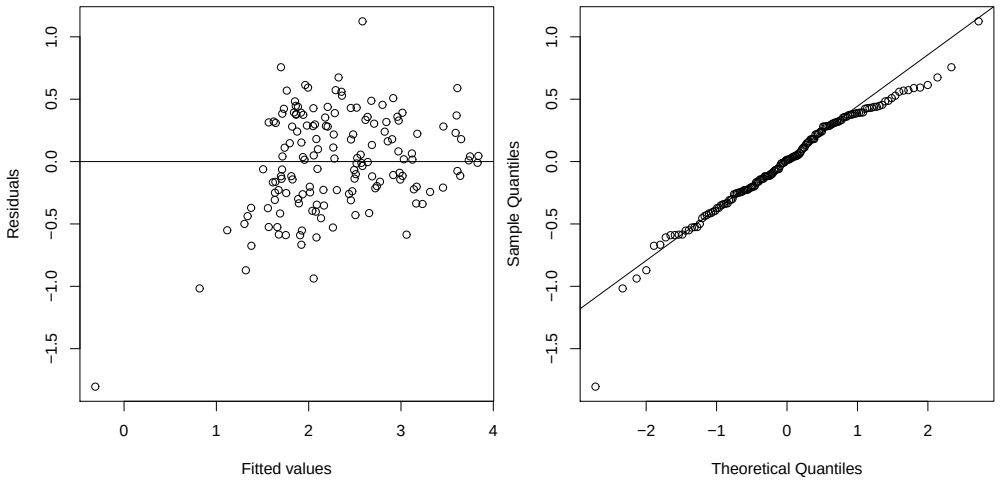

Figure B.25: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.2, fitted to log(chlorophyll$_a$) data for Lake Balaton.

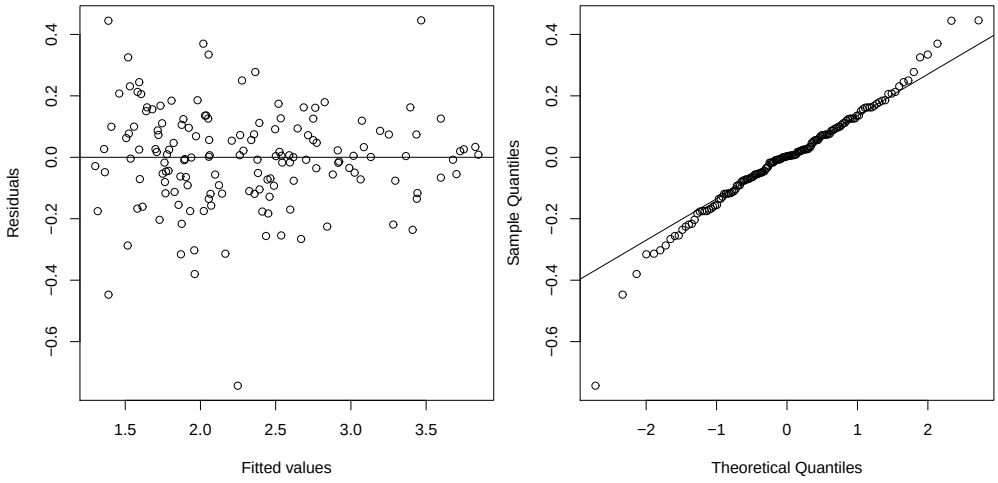

Figure B.26: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.3, fitted to log(chlorophyll$_a$) data for Lake Balaton.

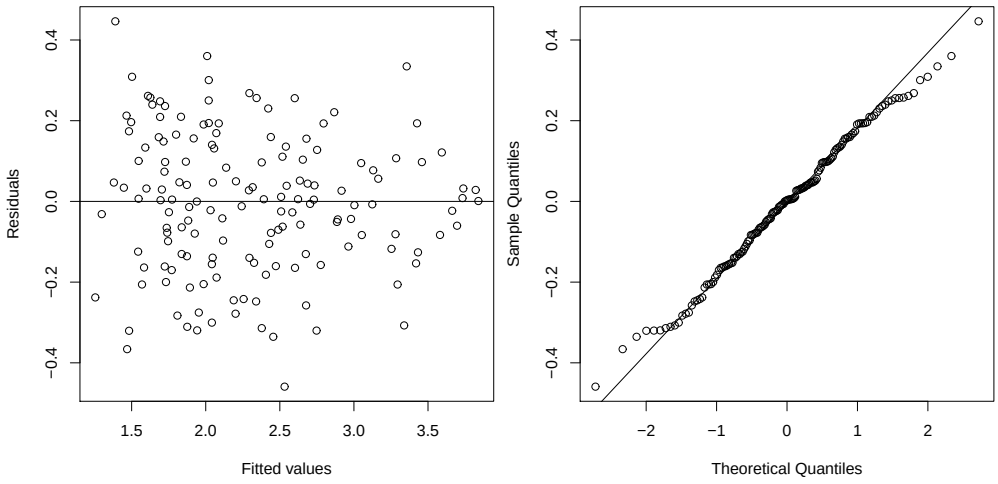

Figure B.27: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.3a, fitted to log(chlorophyll$_a$) data for Lake Balaton.

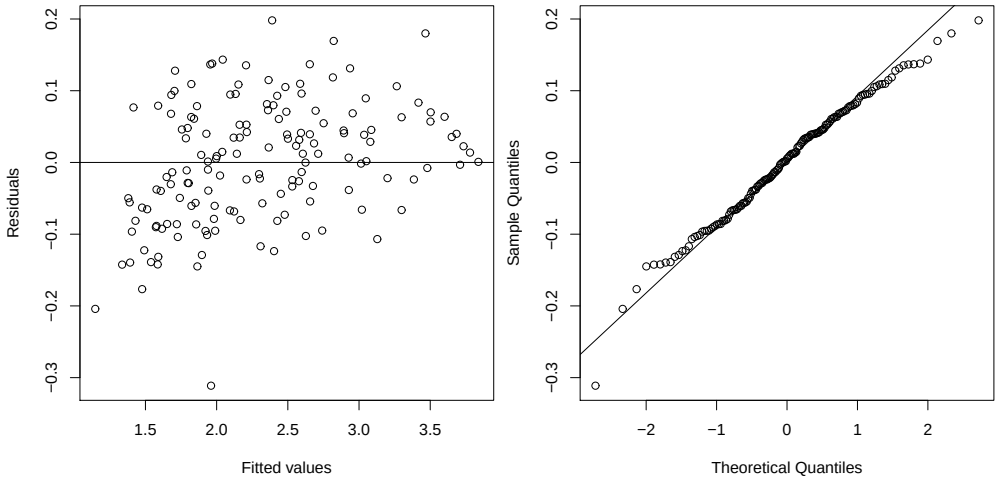

Figure B.28: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.5, fitted to log(chlorophyll$_a$) data for Lake Balaton.

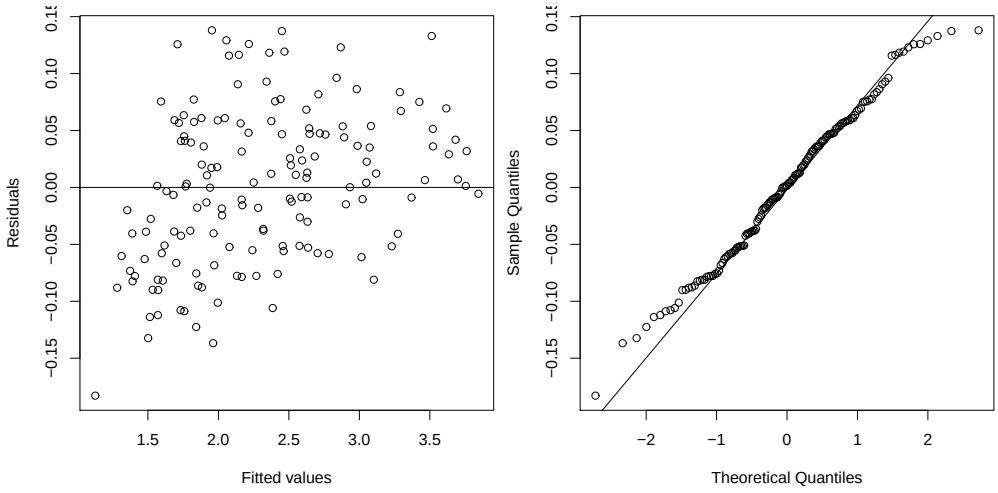

Figure B.29: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.5a, fitted to log(chlorophyll$_a$) data for Lake Balaton.

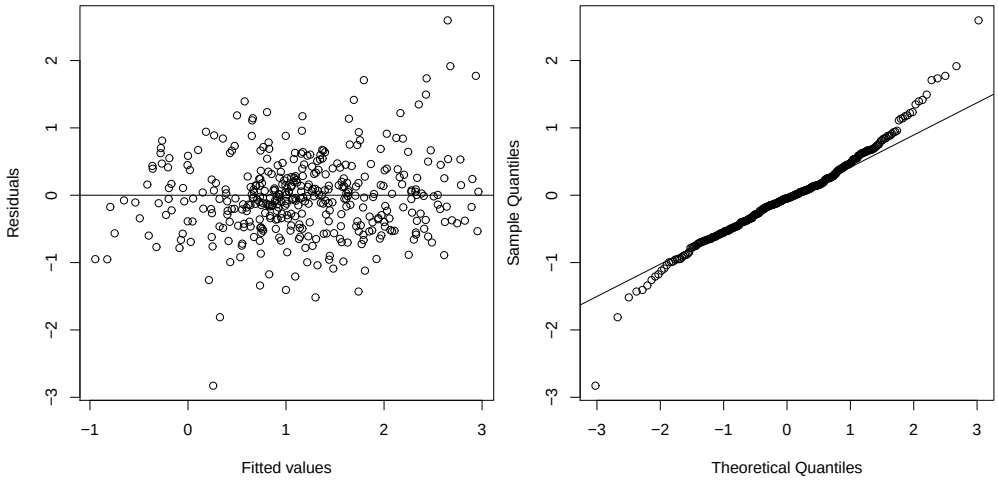

Figure B.30: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.1, fitted to log(chlorophyll$_a$) data for Lake Erie.

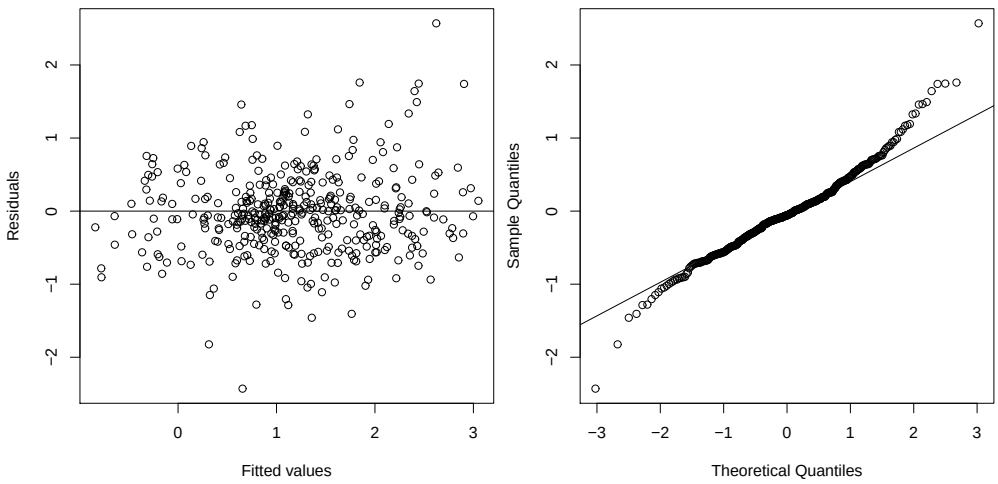

Figure B.31: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.3a, fitted to log(chlorophyll$_a$) data for Lake Erie.

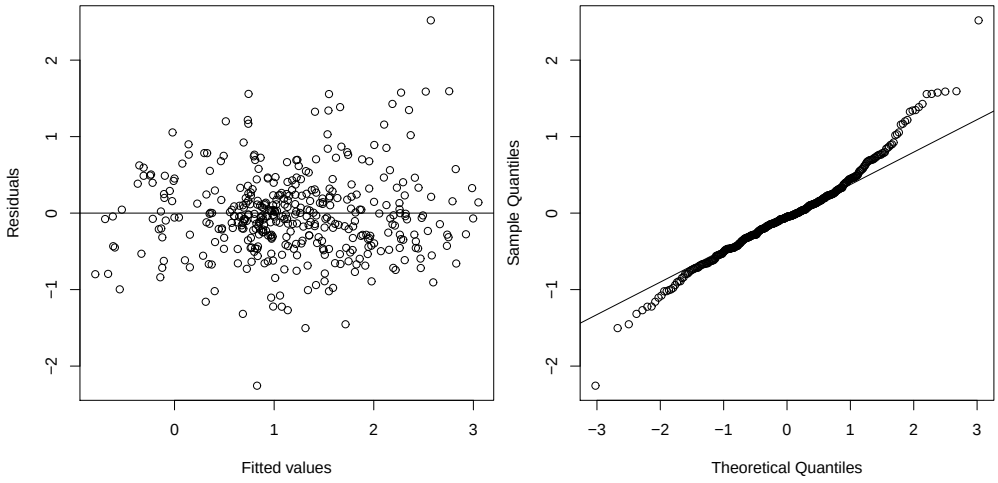

Figure B.32: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.5, fitted to log(chlorophyll$_a$) data for Lake Erie.

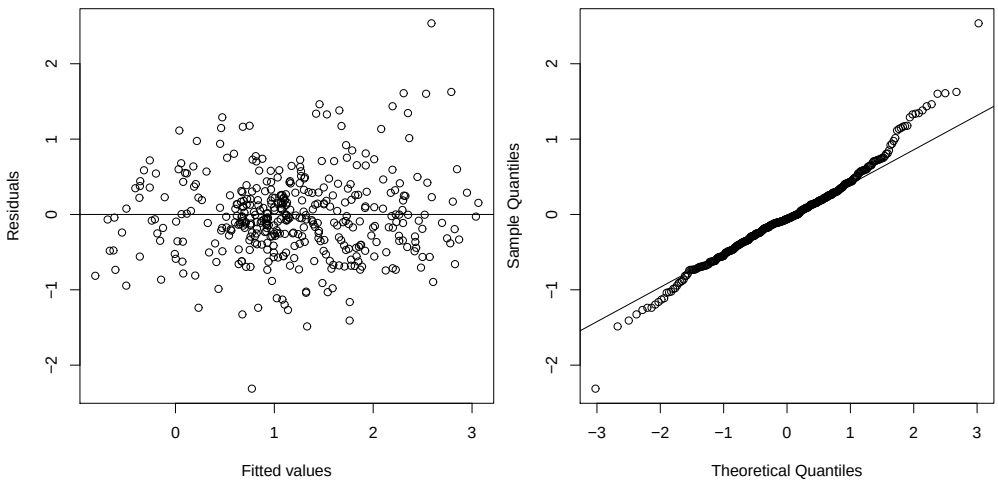

Figure B.33: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 3.5a, fitted to log(chlorophyll$_a$) data for Lake Erie.

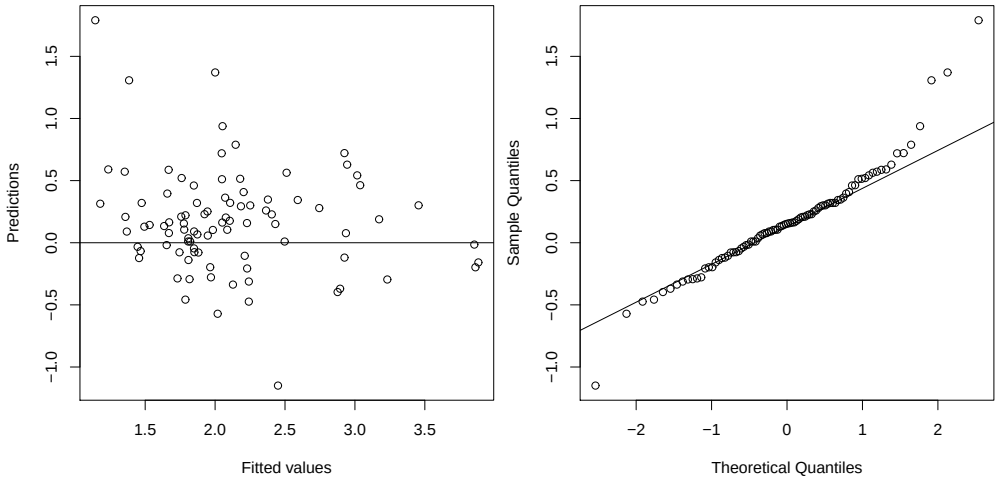

Figure B.34: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.1, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. Plots for log(chlorophyll$_a$) data shown.

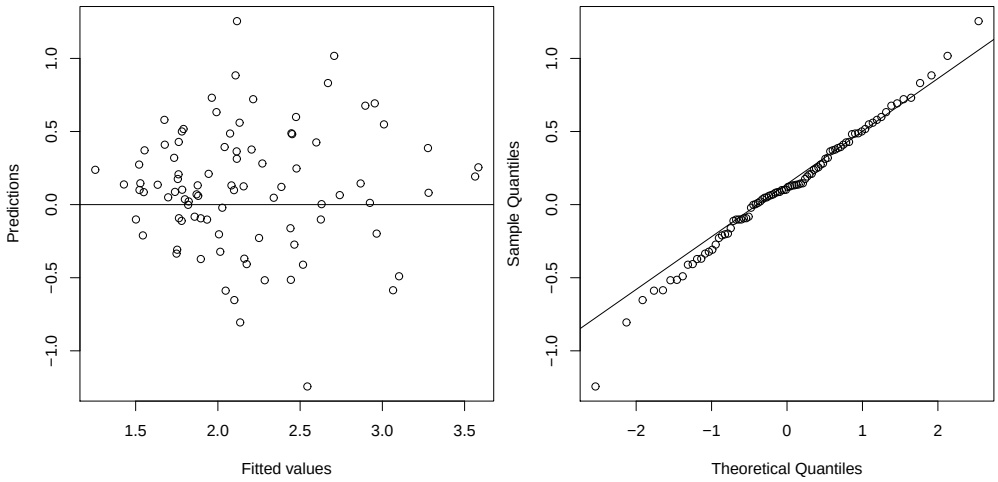

Figure B.35: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.1a, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. Plots for log(chlorophyll$_a$) data shown.

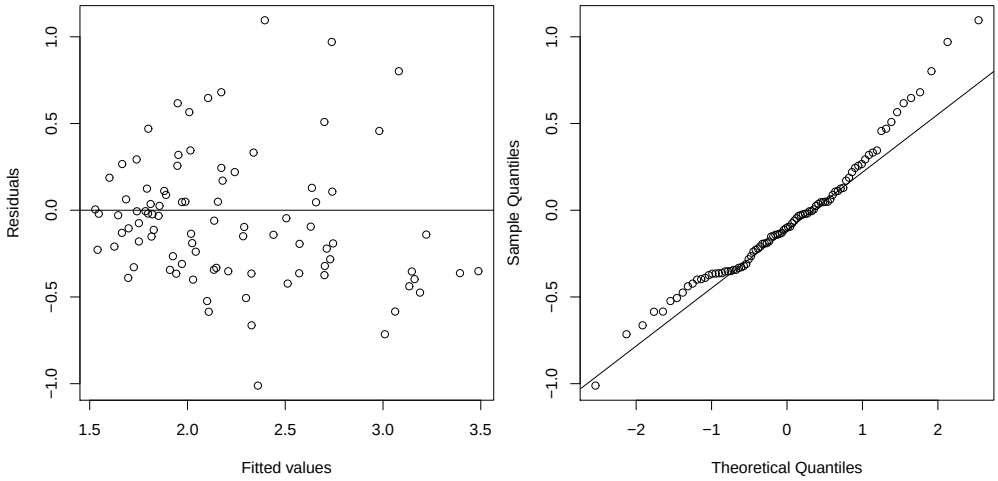

Figure B.36: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.2, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. Plots for log(chlorophyll$_a$) data shown.

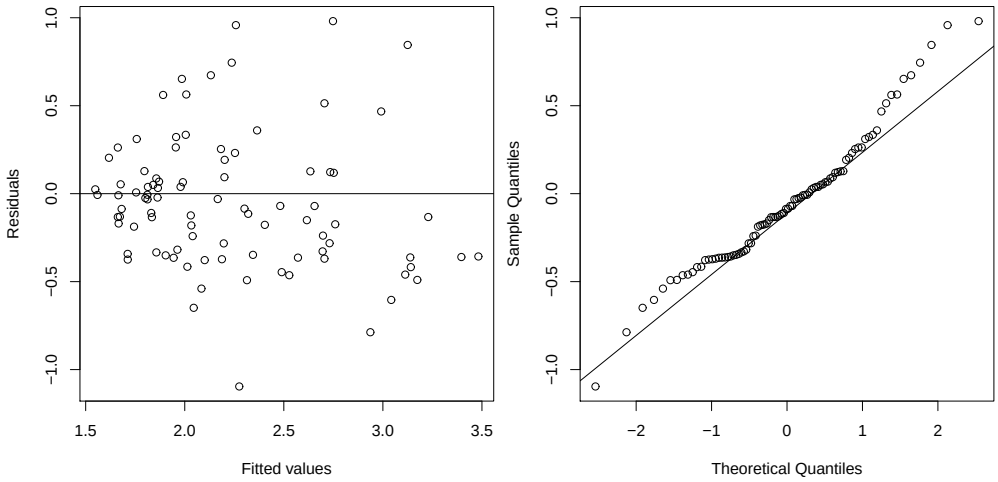

Figure B.37: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.2a, fitted to log(chlorophyll$_a$) and log(total suspended matter) data for Lake Balaton. Plots for log(chlorophyll$_a$) data shown.

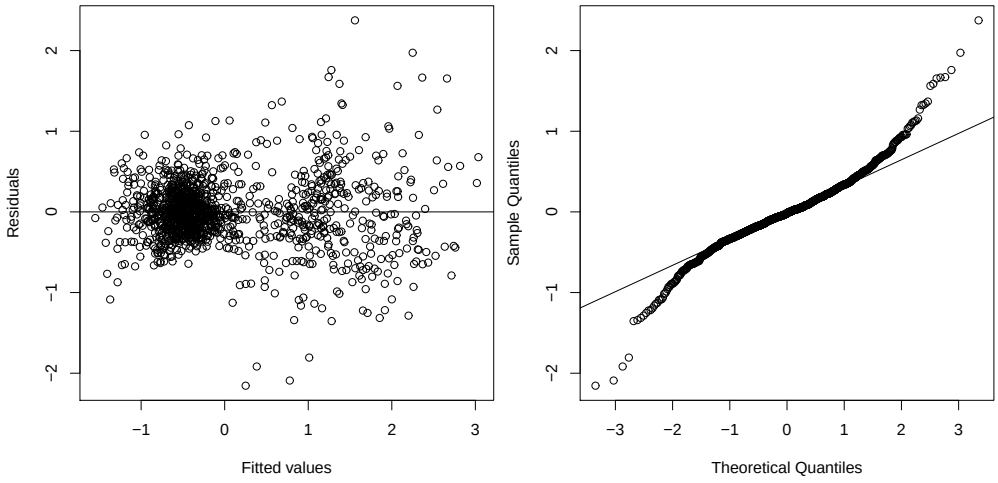

Figure B.38: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.4a-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes.

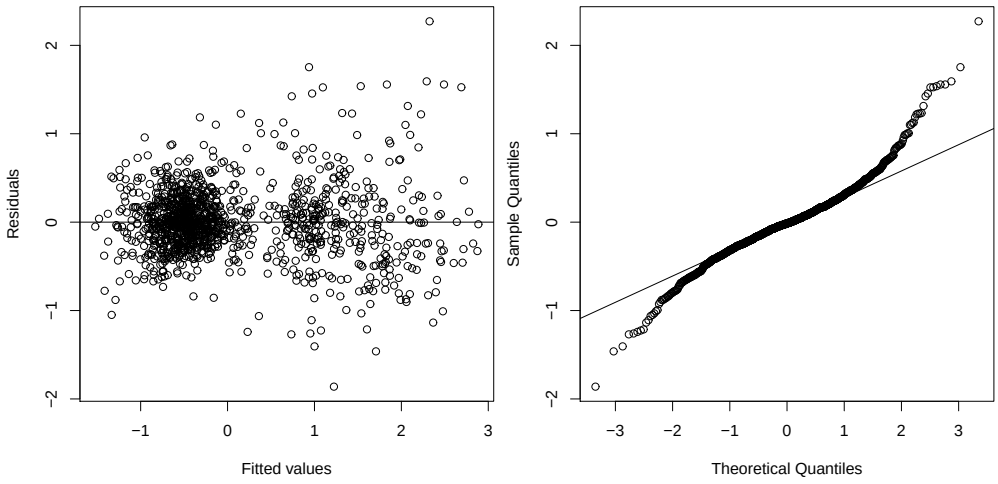

Figure B.39: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.4b-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes.

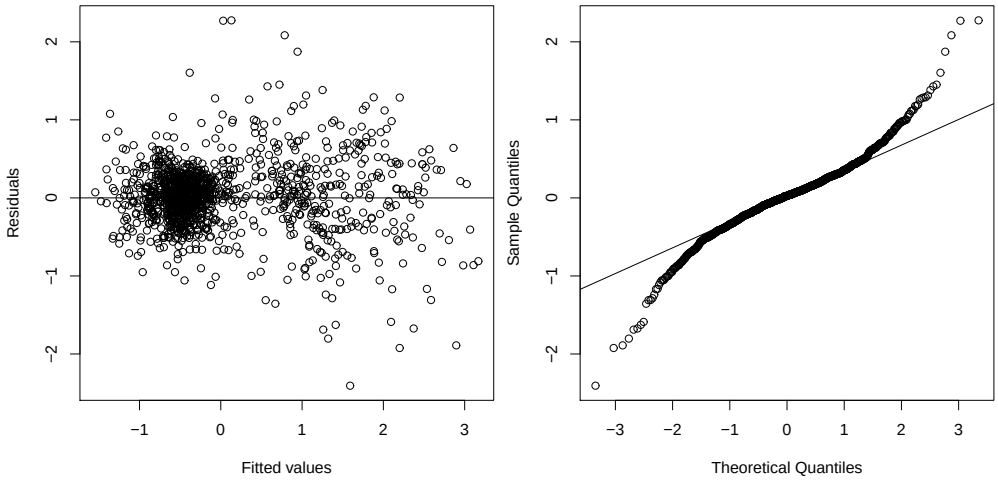

Figure B.40: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.5a-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes.

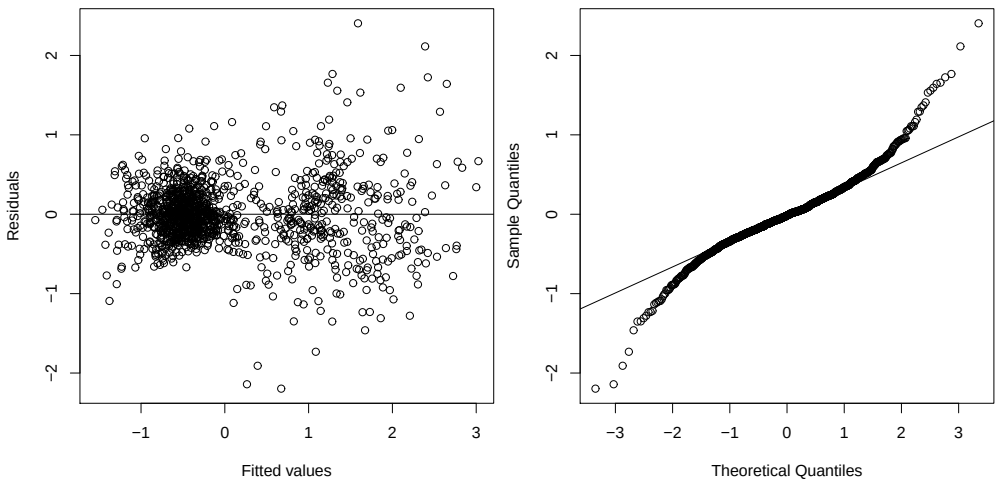

Figure B.41: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.6-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes.

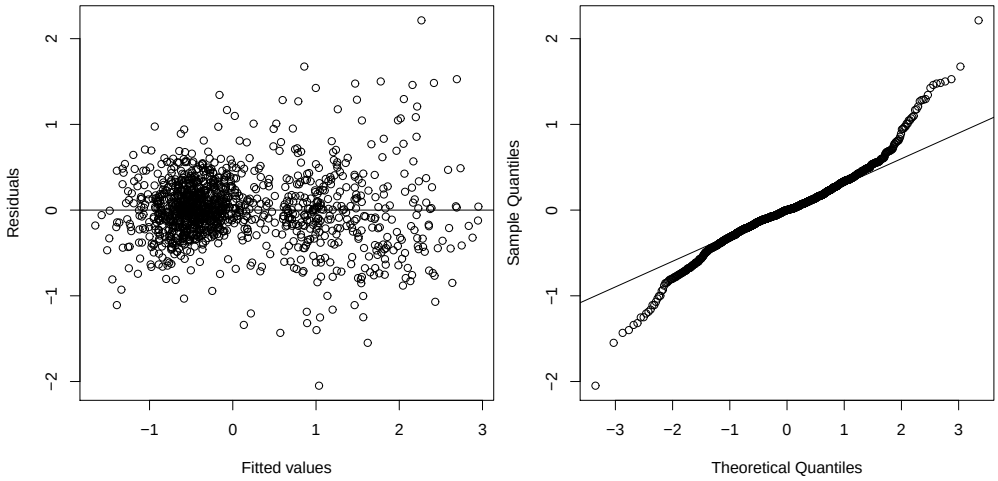

Figure B.42: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 4.7-ST, fitted to log(chlorophyll$_a$) data for the Great Lakes.

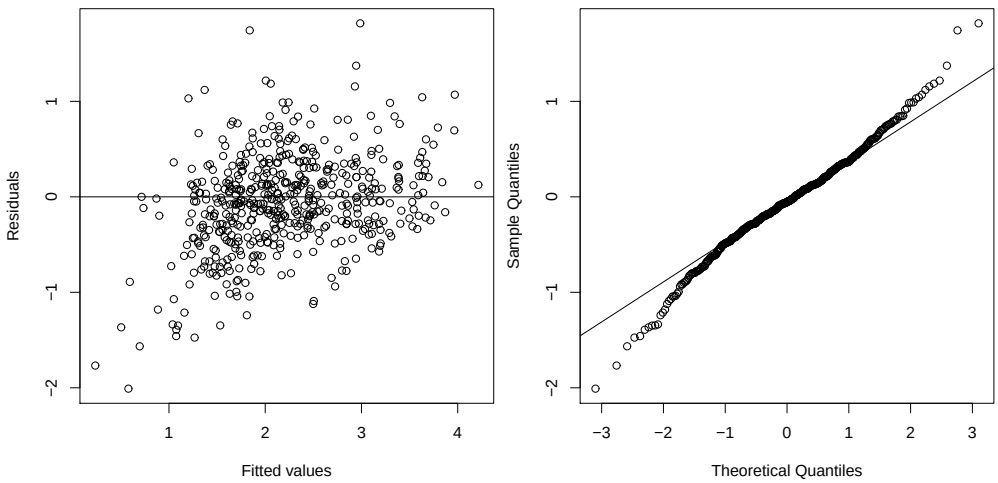

Figure B.43: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 5.8, fitted to log(chlorophyll$_a$) data for Lake Balaton.

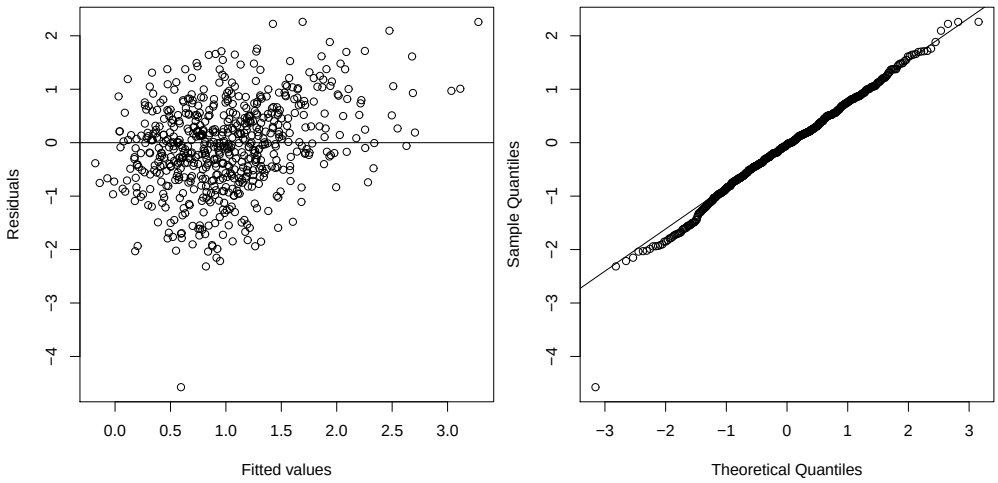

Figure B.44: Residuals versus fitted values (left) and theoretical versus sample quantiles of the distribution of the residuals (Q-Q plot, right) of model 5.8, fitted to log(chlorophyll$_a$) data for Lake Erie.

Bibliography

Abraham, C. & Khadraoui, K. (2015), 'Bayesian regression with B-splines under combinations of shape constraints and smoothness properties', *Statistica Neerlandica* **69**(2), 150–170.

Berrocal, V., Gelfand, A. & Holland, D. (2014), 'Assessing exceedance of ozone standards: a space-time downscaler for fourth highest ozone concentrations', *Environmetrics* **25**(4), 279–291.

Berrocal, V. J., Gelfand, A. E. & Holland, D. M. (2010*a*), 'A bivariate space-time downscaler under space and time misalignment', *The annals of applied statistics* **4**(4), 1942–1975.

Berrocal, V. J., Gelfand, A. E. & Holland, D. M. (2010*b*), 'A spatio-temporal downscaler for output from numerical models', *Journal of agricultural, biological, and environmental statistics* **15**(2), 176–197.

Berrocal, V. J., Gelfand, A. E. & Holland, D. M. (2012), 'Space-time data fusion under error in computer model output: An application to modeling air quality', *Biometrics* **68**(3), 837–848.

Bivand, R. S., Pebesma, E. & Gómez-Rubio, V. (2013), *Applied Spatial Data Analysis with R*, second edn, Springer.

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

Blangiardo, M. & Cameletti, M. (2015), *Spatial and Spatio-temporal Bayesian Models with R - INLA*, John Wiley and Sons, Inc.

BLI (n.d.), 'Welcome to BLI — Institute of Ecology and Botany'.
URL: *http://www.bli.okologia.mta.hu/en*

Botts, L. & Krushelnicki, B. (1995), *The Great Lakes: An Environmental Atlas and Resource Book*, third edn, U.S. Environmental Protection Agency; Government of Canada.
URL: *http://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=P1004ICU.txt*

Bröcker, J. (2012), 'Evaluating raw ensembles with the continuous ranked probability score', *Quarterly Journal of the Royal Meteorological Society* **138**(667), 1611–1617.

Brockmann, C., Peters, M., Poser, K. & Krämer, U. (2004), *DUE Coast-Colour Product User Guide Deliverable DEL-21*, second edn, Brockmann Consult.
URL: *http://www.coastcolour.org/publications/Coastcolour-PUG-v2.2.pdf*

Brockmann Consult, Geoville Information Systems, Brockmann Geomatics Sweden & CIBIO (2015), *Products User Handbook — Inland Waters*, 2.1 edn, Brockmann Consult.

Bruno, F. & Paci, L. (2014), Spatiotemporal model for short-term predictions of air pollution data, *in* 'The Contribution of Young Researchers to Bayesian Statistics', Springer, pp. 91–94.

Büttner, G., Korandi, M., Gyömörei, A., Köte, Z. & Szabó, G. (1987), 'Satellite remote sensing of inland waters: Lake Balaton and reservoir Kisköre', *Acta Astronautica* **15**(6), 305–311.

Clark, J. S. & Gelfand, A. E., eds (2006), *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*, Oxford University Press.

Clarke, B., Fokoué, E. & Zhang, H. H. (2009), *Principles and Theory for Data Mining and Machine Learning*, Springer.

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **31**, 377–403.

Cressie, N. (1985), 'Fitting variogram models by weighted least squares', *Journal of the International Association for Mathematical Geology* **17**(5), 563–586.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, revised edn, John Wiley & Sons, Ltd.

Cressie, N. & Johannesson, G. (2008), 'Fixed rank kriging for very large spatial data sets', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 209–226.

Cressie, N. & Wikle, C. K. (2011), *Statistics for spatio-temporal data*, John Wiley & Sons, Ltd.

de Boor, C. (1978), *A Practical Guide to Splines*, Vol. 27 of *Applied Mathematical Sciences*, Springer-Verlag.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S. & McLoone, S. (2013), 'Principal component analysis on spatial data: An overview', *Annals of the Association of American Geographers* **103**(1), 106–128.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Diggle, P. J. & Ribeiro, Jr., P. J. (2007), *Model-based Geostatistics*, Springer.

Doña, C., Chang, N.-B., Caselles, V., Sánchez, J. M., Camacho, A., Delegido, J. & Vannah, B. W. (2015), 'Integrated satellite data fusion and mining for monitoring lake water quality status of the Albufera de Valencia in Spain', *Journal of Environmental Management* **151**, 416–426.

Duan, H., Ma, R., Simis, S. G. & Zhang, Y. (2012), 'Validation of MERIS case-2 water products in Lake Taihu, China', *GIScience & remote sensing* **49**(6), 873–894.

Eddelbuettel, D. (2013), *Seamless R and C++ Integration with Rcpp*, Springer.

Eddelbuettel, D. & François, R. (2011), 'Rcpp: Seamless R and C++ Integration', *Journal of Statistical Software* **40**(8).

Eddelbuettel, D. & Sanderson, C. (2014), 'RcppArmadillo: Accelerating R with high-performance C++ linear algebra', *Computational Statistics and Data Analysis* **71**, 1054–1063.
URL: *http://dx.doi.org/10.1016/j.csda.2013.02.005*

ESA (n.d. *a*), 'MERIS — Earth Online — ESA'.
URL: *https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/meris*

ESA (n.d. *b*), 'Envisat — Earth Online — ESA'.
URL: *https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat*

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Series in Statistics, Springer.

Finley, A. O., Banerjee, S. & Carlin, B. P. (2007), 'spbayes: an r package for univariate and multivariate hierarchical point-referenced spatial models', *Journal of Statistical Software* **19**(4), 1.
URL: *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074178/*

Finley, A. O., Banerjee, S. & Gelfand, A. E. (2013), 'spbayes for large univariate and multivariate point-referenced spatio-temporal data models', *arXiv preprint arXiv:1310.8192* .

Fuentes, M. & Raftery, A. E. (2005), 'Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models', *Biometrics* **61**, 36–45.

Gelfand, A. E., Kim, H.-J., Sirmans, C. & Banerjee, S. (2003), 'Spatial modeling with spatially varying coefficient processes', *Journal of the American Statistical Association* **98**(462), 387–396.
URL: *http://www.jstor.org/stable/30045248*

Gelman, A., Carlin, B. P., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2014), *Bayesian Data Analysis*, third edn, CRC Press/ Chapman and Hall/ Taylor and Francis.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on pattern analysis and machine intelligence* **PAMI-6**(6), 721–741.

GloboLakes (2016), 'Globolakes Home'.
URL: *http://www.globolakes.ac.uk/*

Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Hall, D. L. & Llinas, J. (1997), 'An introduction to multisensor data fusion', *Proceedings of the IEEE* **85**(1), 6–23.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer, chapter 5, p. 120.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Jolliffe, I. T. (2002), *Principal Component Analysis*, 2nd edn, Springer.

Kasprzak, P., Padisak, J., Koschel, R., Krienitz, L. & Gervais, F. (2008), 'Chlorophyll a concentration across a trophic gradient of lakes: An estimator of phytoplankton biomass?', *Limnologica — Ecology and Management of Inland Waters* **38**(3), 327–338.

Khaleghi, B., Khamis, A., Karray, F. O. & Razavi, S. N. (2013), 'Multisensor data fusion: a review of the state-of-the-art', *Information Fusion* **14**, 28–44.

Kneubühler, M., Frank, T., Kellenberger, T. W., Pasche, N. & Schmid, M. (2007), Mapping chlorophyll-a in Lake Kivu with remote sensing methods, *in* 'Proceedings of Envisat Symposium'.

Kruschke, J. K. (2014), *Doing Bayesian Data Analysis: A Tutorial with R, JAGS and STAN*, second edn, Academic Press.

Kwiatkowska, E. J. & Fargion, G. S. (2002), Merger of ocean color information from multiple satellite missions under the NASA SIMBIOS Project Office, *in* 'Information Fusion, 2002. Proceedings of the Fifth International Conference on', Vol. 1, IEEE, pp. 291–298.

Lark, R. (2000), 'Estimating variograms of soil properties by the method-of-moments and maximum likelihood', *European Journal of Soil Science* **51**(4), 717–728.

Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2013), *The BUGS Book: A Practical Introduction to Bayesian Analysis*, Chapman and Hall/CRC.

MacCallum, S. N. & Merchant, C. J. (2012), 'Surface water temperature observations of large lakes by optimal estimation', *Canadian Journal of Remote Sensing* **38**(1), 25–45.

MacCallum, S. N. & Merchant, C. J. (2013), 'ARC-Lake v2.0, 1991-2011 [Dataset]'.

Maraun, D., Wetterhall, F., Ireson, A., Chandler, R., Kendon, E., Widmann, M., Brienen, S., Rust, H., Sauter, T., Themeßl, M. et al. (2010), 'Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user', *Reviews of Geophysics* **48**(3).

Matthews, M. W., Bernard, S. & Robertson, L. (2012), 'An algorithm for detecting trophic status (chlorophyll–a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters', *Remote Sensing of Environment* **124**, 637–652.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**(6), 1087–1092.

Nguyen, H., Cressie, N. & Braverman, A. (2012), 'Spatial statistical data fusion for remote sensing applications', *Journal of the American Statistical Association* **107**(499), 1004–1018.

Ormerod, S., Dobson, M., Hildrew, A. & Townsend, C. (2010), 'Multiple stressors in freshwater ecosystems', *Freshwater Biology* **55**(s1), 1–4.

Paci, L., Gelfand, A. E. & Holland, D. M. (2013), 'Spatio-temporal modeling for real-time ozone forecasting', *Spatial Statistics* **4**, 79–93.

Padisák, J. & Reynolds, C. S. (1998), 'Selection of phytoplankton associations in Lake Balaton, Hungary, in response to eutrophication and restoration measures, with special reference to the cyanoprokaryotes', *Hydrobiologia* **384**(1-3), 41–53.

Palmer, S., Odermatt, D., Hunter, P., Brockmann, C., Présing, M., Balzter, H. & Tóth, V. (2015), 'Satellite remote sensing of phytoplankton phenology

in Lake Balaton using 10 years of MERIS observations', *Remote Sensing of Environment* .

Parkinson, C. L. (1997), *Earth From Above: Using Color-Coded Satellite Images to Examine the Global Environment*, University Science Books, Sausalito, California.

Paul, J. F., Kasprzyk, R. & Lick, W. (1982), 'Turbidity in the western basin of Lake Erie', *Journal of Geophysical Research: Oceans* **87**(C8), 5779–5784.

Piegorsch, W. W. & Bailer, A. J. (2005), *Analyzing Environmental Data*, John Wiley & Sons, Ltd.

Pierce, D. (2017), *ncdf4: Interface to Unidata netCDF (Version 4 or Earlier) Format Data Files*. R package version 1.16.
URL: *https://CRAN.R-project.org/package=ncdf4*

Plummer, M. (2003), JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, *in* K. Hornik, F. Leisch & A. Zeileis, eds, 'Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)', Vienna.
URL: *http://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf*

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006), 'CODA: Convergence diagnosis and output analysis for MCMC', *R News* **6**(1), 7–11.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Pya, N. & Wood, S. N. (2016), 'A note on basis dimension selection in generalized additive modelling', *arXiv preprint arXiv:1602.06696* .

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ramsay, J. O., Hooker, G. & Graves, S. (2009), *Functional Data Analysis with R and MATLAB*, Springer.

Ramsay, J. O. & Silverman, B. W. (2006), *Functional Data Analysis*, Springer Series in Statistics, second edn, Springer.

Ramsay, J. O., Wickham, H., Graves, S. & Hooker, G. (2014), *fda: Functional Data Analysis*. R package version 2.4.3.
URL: *http://CRAN.R-project.org/package=fda*

Ribeiro, Jr., P. J. & Diggle, P. J. (2001), 'geoR: a package for geostatistical analysis', *R-NEWS* **1**(2), 14–18. ISSN 1609-3631.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Richman, M. B. (1986), 'Rotation of principal components', *Journal of climatology* **6**(3), 293–335.

Rue, H., Martino, S. & Chopin, N. (2009), 'Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2), 319–392.

Rundel, C. W., Schliep, E. M., Gelfand, A. E. & Holland, D. M. (2015), 'A data fusion approach for spatial analysis of speciated $PM_{2.5}$ across time', *Environmetrics* **26**, 515–526.

Ruppert, D. (2002), 'Selecting the number of knots for penalized splines', *Journal of Computational and Graphical Statistics* **11**(4), 735–757.

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**(1), 61–86.

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**(1), 77–103.

Sakuno, Y. (2013), Chlorophyll data fusion in Tachibana Bay using COMS GOCI and MODIS data by the LCI method, *in* 'Geoscience and Remote Sensing Symposium (IGARSS), 2013 IEEE International', IEEE, pp. 1594–1597.

Sari, W. J., Wigenac, A. H. & Djuraidah, A. (2017), 'Quantile regression with functional principal component in statistical downscaling to predict extreme rainfall', *International Journal of Ecological Economics and Statistics* **38**(1), 1–9.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., Martini, J., Ballani, F., Moreva, O., Berreth, C., Menck, P., Gross, S., Ober, U., Burmeister, K., Manitz, J., Ribeiro, P., Singleton, R., Pfaff, B. & R Core Team (2015), *RandomFields: Simulation and Analysis of Random Fields*. R package version 3.1.1.
URL: *https://cran.r-project.org/web/packages/RandomFields/index.html*

Schmidli, J., Goodess, C., Frei, C., Haylock, M., Hundecha, Y., Ribalaygua, J. & Schmith, T. (2007), 'Statistical and dynamical downscaling of precipitation: An evaluation and comparison of scenarios for the European Alps', *Journal of Geophysical Research: Atmospheres* **112**(D4).

Schmitt, M. & Zhu, X. X. (2016), 'Data fusion and remote sensing: an ever-growing relationship', *IEEE Geoscience and Remote Sensing Magazine* **4**(4), 6–23.

Sellinger, C. E., Stow, C. A., Lamon, E. C. & Qian, S. S. (2008), 'Recent water level declines in the Lake Michigan-Huron System', *Environmental Science & Technology* **42**(2), 367–373.

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 of

Lecture Notes in Computer Science, Springer-Verlag, pp. 203–222. From the First ACM Workshop on Applied Computational Geometry.

Shewchuk, J. R. (1997), Delaunay Refinement Mesh Generation, PhD thesis, Carnegie Mellon University, Pittsburgh.

Simis, S. G., Peters, S. W. & Gons, H. J. (2005), 'Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water', *Limnology and Oceanography* **50**(1), 237–245.

Tátrai, I., Mátyás, K., Korponai, J., Paulovits, G. & Pomogyi, P. (2000), 'The role of the Kis-Balaton Water Protection System in the control of water quality of Lake Balaton', *Ecological Engineering* **16**(1), 73–78.

Teta, R., Romano, V., Della Sala, G., Picchio, S., De Sterlich, C., Mangoni, A., Di Tullio, G., Costantino, V. & 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**(2).

Waller, L. & Carlin, B. (2010), *Handbook of Spatial Statistics*, CRC Press/ Chapman and Hall/ Taylor and Francis, chapter 14, pp. 217–244.

Wan, N. & Hu, J. (2013), 'Optimal layout of borehole location based on Delaunay refinement', *21st International Conference on Geoinformatics* .

White, F. E. (1991), Data fusion lexicon, Technical report, Data Fusion Panel, Joint Directors of Laboratories, Technical Panel for C3.

Wikle, C. K. & Berliner, L. M. (2005), 'Combining information across spatial scales', *Technometrics* **47**(1).

Wilby, R. L. & Wigley, T. (1997), 'Downscaling general circulation model output: a review of methods and limitations', *Progress in Physical Geography* **21**(4), 530–548.

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**, 124–127.

Williamson, C. E., Saros, J. E., Vincent, W. F. & Smold, J. P. (2009), 'Lakes and reservoirs as sentinels, integrators, and regulators of climate change', *Limnology and Oceanography* **54**(6part2), 2273–2282.

Wood, S. N. (2003), 'Thin plate regression splines', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1), 95–114.

Wood, S. N. (2006), *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC Press.