Accepted for publication in *Spatial Statistics* published by Elsevier: https://doi.org/10.1016/j.spasta.2022.100615

Adaptive smoothing to identify spatial structure in global lake ecological processes using satellite remote sensing data

Mengyi Gong*[1,2], Ruth O'Donnell[2], Claire Miller[2], Marian Scott[2], Stefan Simis[3], Steve Groom[3], Andrew Tyler[4], Peter Hunter[4], Evangelos Spyrakos[4], Christopher Merchant[5], Stephen Maberly[6] and Laurence Carvalho[7]

[1] *Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom, LA1 4YF*
[2] *School of Mathematics and Statistics, University of Glasgow, University Place, Glasgow, United Kingdom, G12 8QQ*
[3] *Plymouth Marine Laboratory, Plymouth, United Kingdom, PL1 3JH*
[4] *School of Biological and Environmental Science, University of Stirling, Stirling, United Kingdom, KF9 4LA*
[5] *Department of Meteorology and National Centre for Earth Observation, University of Reading, Reading, United Kingdom, RG6 6AL*
[6] *UK Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, United Kingdom, LA1 4AP*
[7] *UK Centre for Ecology & Hydrology, Bush Estate, Penicuik, Midlothian, United Kingdom, EH26 0QB*

Abstract

Satellite remote sensing data are important to the study of environment problems at a global scale. The GloboLakes project aimed to use satellite remote sensing data to investigate the response of the major lakes on Earth to environmental conditions and change. The main challenge to statistical modelling is the identification of the spatial structure in global lake ecological processes from a large number of time series subject to incomplete data and varying uncertainty. This paper introduces a comprehensive modelling procedure, combining adaptive smoothing and functional data analysis, to estimate the smooth curves representing the trend and seasonal patterns in the time series and to cluster the curves over space. Two approaches, based on an irregular basis and an adaptive penalty matrix, are developed to account for the varying uncertainty induced by missing observations and specific constraints (e.g. substantive periods of measurement values of zero in winter). In particular, the adaptive penalty matrix applies a heavier penalty to smooth curve estimates where there is higher uncertainty to prevent over-fitting the noisy/biased data. The modelling procedure was applied to the lake surface water temperature (LSWT) time series from 732 largest lakes globally and the lake chlorophyll-a time series from 535 largest lakes globally. The procedure enabled the identification of nine global lake thermal regions based on the temporal dynamics of LSWT, and the extraction of eight global lake clusters based on the interannual variation in chlorophyll-a and ten clusters to differentiate the seasonal signals.

1 Introduction

Satellite remote sensing data are important to the understanding and management of the environmental problems facing the world today. The advancement of satellite remote sensing technology enables the regular monitoring of remote or even inaccessible places on Earth. The research project GloboLakes (http:

*Correspondence email: m.gong1@lancaster.ac.uk

//www.GloboLakes.ac.uk/) aimed to investigate the response of lakes to environmental change at a global scale through satellite remote sensing data. The project explored several key ecological variables associated with lake health, two of which are the lake surface water temperature (LSWT) time series, obtained from the Along-Track Spectrum Radiometer 2 and the Advanced Along-Track Spectrum Radiometer ((A)ATSRs), and the chlorophyll-a (Chl-a) time series, obtained from the MEdium Resolution Imaging Spectrometer (MERIS). Data from 932 largest lakes on Earth were retrieved. Using these data, scientists on the project were interested in investigating the temporal dynamics (e.g. seasonal and trend signals) of the lake ecological variables over the monitoring period and identifying the spatial patterns with respect to the temporal dynamics among the 932 lakes.

The LSWT data and the chlorophyll-a data are available as time series observed at the pixels covering the area of the lakes. The LSWT time series are available bi-monthly from summer 1995 to spring 2012, and the chlorophyll-a time series are available monthly from July 2003 to November 2011. In order to investigate the temporal dynamics and the spatial patterns in the LSWT and chlorophyll-a data from the 932 lakes, the spatially averaged time series (referred to as the "mean time series" hereafter) were created. Due to cloud cover, ice cover and satellite orbits, observations are not always available at an individual pixel over the observing period (MacCallum & Merchant, 2013). As a result, the number of pixel observations involved in the spatial average varies. In poorly observed months, such as the rainy seasons and the winter periods, only a small proportion of the lake may be observed, giving spatial averages with higher uncertainty. There may be a few continuous time points with no observation available at all, leaving a gap in the mean time series. This is common for the chlorophyll-a observations from lakes in cold climates, since the lakes freeze over the winter. Consequently, the mean time series are incomplete and are subject to varying uncertainty both within and between individual time series. This brings challenges to the statistical analysis that aims to investigate the temporal and spatial patterns in the data.

As a motivating example, two monthly mean chlorophyll-a time series from Lake Tanganyika in Congo and Lake Chany in Russia are presented in the top panels of Figure 1. The time series were log transformed to reduce the skewness. The bottom panels of Figure 1 show the time series of the proportion of pixels that have no observation out of all pixels covering the lake area. These time series, referred to as the "missing proportion time series", reflect the uncertainty of the spatial average to some extent. In general, the spatial average computed from incomplete observations tends to have higher uncertainty than the spatial average computed from complete observations in an satellite remote sensing image. The former may even be biased if the available observations are clustered in a particular area of the image which has consistently higher or lower measurement values. Here the mean chlorophyll-a time series of Lake Tanganyika (top left) is almost complete, but the proportion of missing observations varies over time (bottom left), resulting in changing uncertainty throughout the mean time series. The mean chlorophyll-a time series of Lake Chany (top right) displays distinctive seasonal gaps. The corresponding missing proportion time series (bottom right) reaches 1 (i.e. all pixels were unobserved) over the winter months.

Another modelling challenge is related to the specific constraints on the values of the observations. Consider the LSWT time series. By the law of physics, the LSWT is 0°C when ice forms during the winter period. Therefore, even if the pixels are unobserved and flagged as "ice cover", the uncertainty associated with the spatially averaged LSWT is low. Figure 2 shows the bi-monthly mean LSWT time series from Chamo Lake in Ethiopia (left) and Lac des Bois in northwest Canada (right), from 2000 to 2010. Whereas the uncertainty of the mean LSWT of the Chamo Lake can be affected by data availability, the confidence around the zero values over the winter period in the mean LSWT of Lac des Bois is high regardless of data availability due to its high latitude location. This again creates varying uncertainty within and between individual time series.

These features require a modelling approach that can account for the varying uncertainty in the time series to ensure the data from different lakes are comparable. At the same time, the approach needs to be interpretable so scientists can draw conclusion from the result, transferable so that it can be implemented on similar types

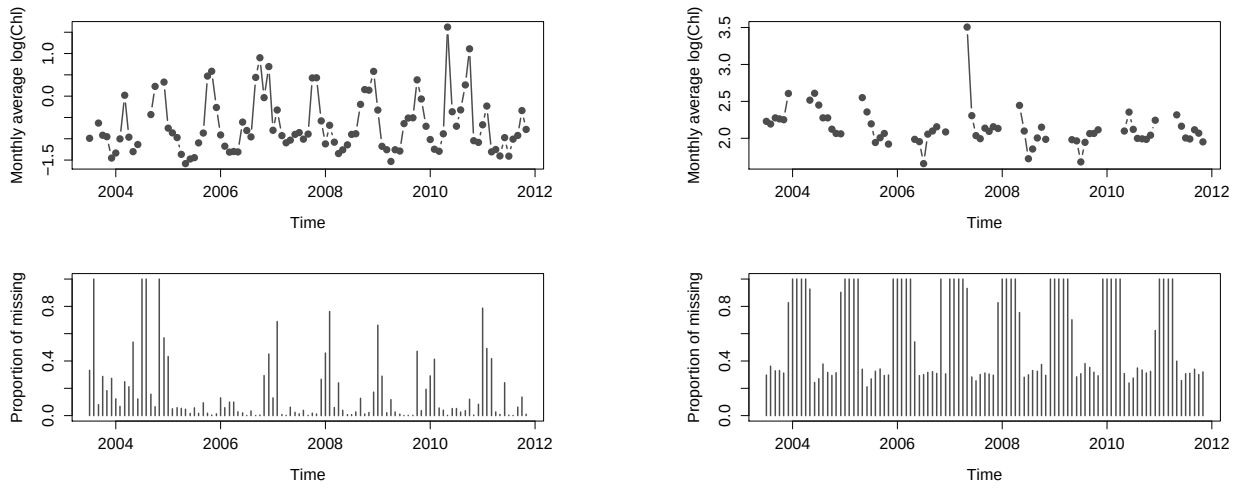

Figure 1: The monthly averaged log(Chl-a) time series of Lake Tanganyika (top left) and Lake Chany (top right). The corresponding missing proportion time series of Lake Tanganyika (bottom left) and Lake Chany (bottom right).

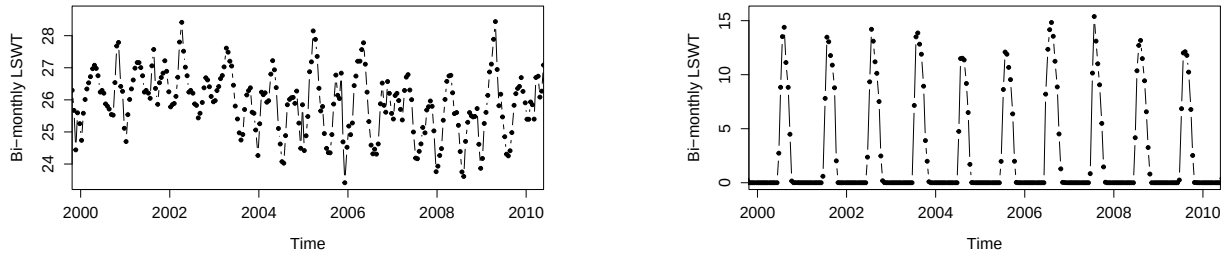

Figure 2: The bi-monthly averaged LSWT time series of Chamo Lake (left) and Lac des Bois (right).

of time series, and computationally efficient. Based on these considerations, this paper proposes a comprehensive modelling procedure, consisting of adaptive smoothing and functional data analysis, to tackle the challenges.

Functional data analysis (Ramsey & Silverman, 2005) is a natural choice to model time series data from a large number of objects. It reduces the data dimension through the functional representations of the time series and enables the simultaneous modelling of the data from all objects. The functional representations are typically created via smoothing the data. Using the functional representations helps to reduce the impact from the noisy observations and avoid the potential bias[1] in the data. To create the functional representations and account for the varying uncertainty associated with the mean time series of the LSWT and lake chlorophyll-a, this paper proposes two modifications of the standard P-spline smoothing (Eilers & Marx, 1996), namely the irregular basis and the adaptive penalty matrix. Together, the two methods address the problem of the changing uncertainty due to data availability and specific constraints. The resulting smoothed time series are considered as more appropriate functional representations of the underlying temporal patterns, suitable for clustering to explore spatial structure.

[1]The bias in the mean time series can originate from various sources, e.g. missing observations and satellite retrieval errors. The latter describe the deviations of the retrieved data from the truth. They are common in satellite remote sensing data retrievals (Povey & Grainger, 2015; MacCallum & Merchant, 2013). Although they may not be available in some data sets.

This rest of the paper is organized into three sections. Section 2 begins with a brief description of P-spline smoothing and the challenges in relation to lake ecological time series data. It then introduces the two modifications to the standard P-spline smoothing, with illustrations using the LSWT and chlorophyll-a time series. Section 3 presents the application of the proposed modelling procedure to the LSWT and chlorophyll-a time series to investigate the spatial patterns in the smoothed time series curves. Section 4 summarises the paper, discusses other plausible choices of the uncertainty measure with respect to the spatial average, and proposes some future extensions.

2 Adaptive smoothing

2.1 Standard P-spline smoothing and its problems

Initially, a standard P-spline approach was used to create the functional representations of the mean LSWT and chlorophyll-a time series. Denote the time series as Y_t, $t = 1, \cdots, n$, and its functional representation as $Y(t) = \Phi(t)\boldsymbol{\beta} + \epsilon(t)$, where $\Phi(t)$ is a 3rd order saturated B-spline basis evaluated at $t = 1, \cdots, n$. The standard P-spline smoothing applies the 2nd order difference penalty, $\lambda\|\boldsymbol{D}\boldsymbol{\beta}\|^2 = \lambda\boldsymbol{\beta}^\top\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{\beta}$, to control the smoothness of $Y(t)$ (Eilers & Marx, 1996). Specifically, let λ be the smoothing parameter and $\boldsymbol{D}$ be the 2nd difference matrix of the identity matrix, i.e. $\boldsymbol{D} = \triangle^2\mathbf{I}$. The penalized least squares criterion of the standard P-spline smoothing can be written as

$$\sum_t \|Y_t - \Phi(t)\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\top\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{\beta}\,. \tag{1}$$

In criterion (1), the penalty matrix $\mathcal{S} = \boldsymbol{D}^\top\boldsymbol{D}$ applies the same penalty to all the elements in the basis coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)^\top$, where K ($K = n + 3 + 1$) is the dimension of the B-spline basis, excluding the boundary values. In other words, β_k is penalized no more or no less than β_j, for $k, j \in \{3, \cdots, K-2\}$.

Although there is no technical problem in using the standard P-spline smoothing, it is not the most appropriate approach to model time series in the presence of changing uncertainty. On the one hand, applying the same penalty across all time points when some are subject to high uncertainty could result in a biased estimate. For example, the spatially averaged LSWT or chlorophyll-a at some time points may come from a small number of pixels in a particular area of the lake, where the measurement values are distinctively different from the rest of the lake. Consequently, the spatial averages at these time points would not be a reliable estimate of the condition of the lake. It is important to avoid over-fitting the data at these time points. On the other hand, the estimation of the basis coefficients corresponding to the time points where there are no data or where there are constraints on values can be problematic. To be precise, as the solution from the generalised least squares does not always respect the physical or biological constraints, it could produce values that are physically invalid for LSWT and chlorophyll-a.

Motivated by the problems above, this section proposes two adjustments to standard P-spline smoothing to better reflect the changing uncertainties and the features of the time series. The first adjustment is the use of an irregular basis. The idea is to remove the knots where there is no data or where there are specific constraints on the values of the data. This gives an irregular basis, which helps to mitigate the influence from the missing gaps and the constraints. The second adjustment is the use of an adaptive penalty matrix that accounts for the varying uncertainty throughout the time series. There are various ways of constructing an adaptive penalty, as described in Friedman (1991); Ruppert & Carroll (2000); Wood *et al.* (2002). This paper chooses to construct the adaptive penalty matrix based on the uncertainty associated with the spatial averaging. In particular, the adaptive penalty matrix places heavier penalties at the time points where there is higher uncertainty in the spatial averaging and lighter penalties where there is more confidence that the spatial averaging is reflective of the entire lake condition. The details of the two adjustments will be described in the next two sections.

Before moving on, it is worthwhile noting that, as the aim is to produce smoothed functional representations of a large number of time series in a computationally efficient manner, the methods presented in the following sections are appropriate heuristics to estimate the multiple time series curves in the data set. It is possible to improve the smoothing of a particular time series curve using a method tailored to it, but this is not the focus of this paper.

2.2 The irregular basis

In theory, there is no restriction on the number of the basis functions used in the P-spline fitting (Eilers *et al.*, 2015). A saturated basis, which has the same number of knots as the number of observations, is often a starting point. However, when there are gaps in the time series, or when there are specific constraints on the observations (e.g. the LSWT observations should be 0°C during the winter period), using a saturated basis can sometimes create artefacts.

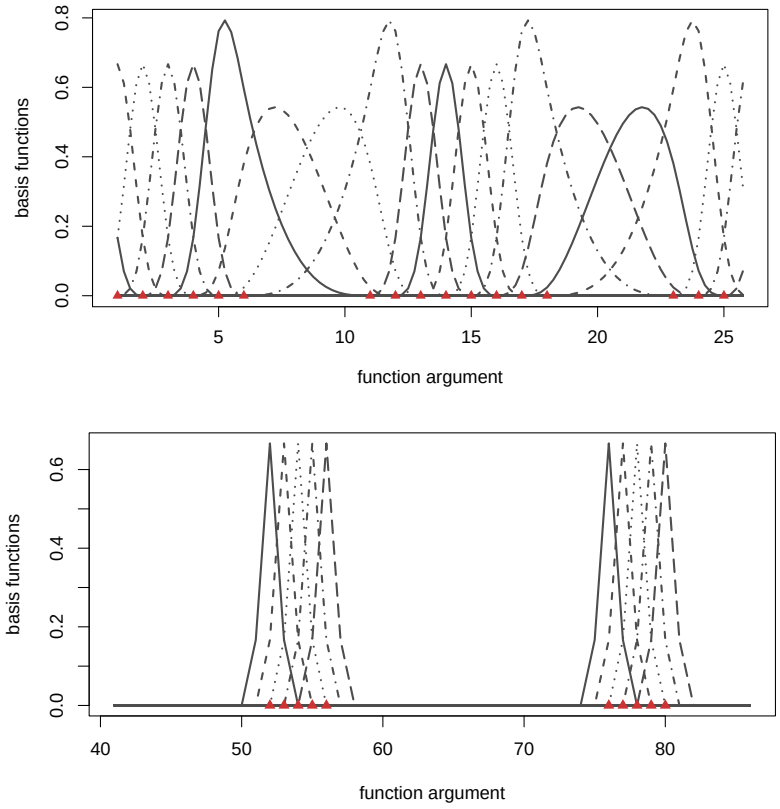

Figure 3: (top) An example of the first 15 basis functions from the B-spline basis used in the smoothing of Lake Chany log(Chl-a) time series. (bottom) An example of the irregular B-spline basis used in the smoothing of Lac des Bois LSWT time series. The red triangles indicate the locations of the knots.

A practical way to overcome this problem is to discard the knots in a saturated basis corresponding to the missing gaps in the time series and the time points where there are specific constraints. This gives an irregular basis that better represents the features of the time series. The top panel of Figure 3 gives an example of such an irregular basis. Here, the first 15 basis functions of the 3rd order B-spline basis created for the log(Chl-a)

time series of Lake Chany is plotted. The red triangles on the horizontal axis indicates the locations of the irregular spaced knots after removing the knots corresponding to the missing gaps. This construction can help to reduce the edge effects induced by the observations at the two ends of the missing gaps. The bottom panel of Figure 3 gives another example of the irregular 3rd order B-spline basis used in the smoothing of the LSWT time series of Lac des Bois. Here the knots corresponding to the winter periods were removed as the LSWT values were fixed at 0°C.

In addition, to avoid boundary effects, the knot sequence may be extended beyond the original boundaries and the boundary values Y_1 and Y_n are repeated multiple times. In this case, the splines at the two ends are the same as defined in Eilers *et al.* (2015). For a 3rd order B-spline basis, three extension points are added to each side of the original boundaries. By doing this, the period of interest (from time point 1 to n) would be free from any boundary effects.

2.3 The adaptive penalty matrix

The construction of the adaptive penalty matrix is based on the 2nd difference penalty matrix, $\boldsymbol{D} = \triangle^2\mathbf{I}$, in the standard P-spline smoothing. Whereas the standard 2nd difference penalty matrix applies the same penalty to all time points, with the adaptive penalty matrix, different time points would receive different penalties. One way to achieve this is to replace the $\boldsymbol{I}$ matrix with a diagonal matrix $\boldsymbol{A}$ whose diagonal entries reflect the relative high or low uncertainties of the data at different time points. Denote the adjustment matrix $\boldsymbol{A} = \text{diag}\{\alpha_1, \cdots, \alpha_K\}$, where α_k reflects the uncertainty of the datum associated with the k-th knot (which corresponds to time point k in a saturated basis). The adaptive penalty matrix, denoted as $\mathcal{S}_\alpha$, is then constructed by taking 2nd differences of matrix $\boldsymbol{A}$. Using similar notation as in section 2.1, $\mathcal{S}_\alpha$ can be written as

$$\mathcal{S}_\alpha = \boldsymbol{A}^\top \boldsymbol{D}^\top \boldsymbol{D}\boldsymbol{A} = \boldsymbol{D}_\alpha^\top \boldsymbol{D}_\alpha \,, \tag{2}$$

and the adaptive penalty term can be written as

$$\lambda \boldsymbol{\beta}^\top \mathcal{S}_\alpha \boldsymbol{\beta} = \lambda \boldsymbol{\beta}^\top \boldsymbol{D}_\alpha^\top \boldsymbol{D}_\alpha \boldsymbol{\beta} \,. \tag{3}$$

This is similar to the approach in Ruppert & Carroll (2000) for spatially adaptive splines, where a penalty of the form $\sum_k \alpha_k |\beta_k|^2$ was used, with β_k being the basis coefficient and α_k being the penalty function evaluated at knot k. Ruppert & Carroll (2000) referred to this as the "local penalty", as opposed to the "global penalty" which uses a constant α. Here a series of values α_k, $k = 1, \cdots, K$ is used to adjust for the uncertainties at different time points, which can also be considered as a local property. What remains is to determine the values of α_k, $k = 1, \cdots, K$, given the application, which will be discussed in detail in the following paragraphs. For convenience, the values of α_k are restricted to the range of [0, 1], as the absolute scale of α_k is indistinguishable from the smoothing parameter λ and only the relative scale difference among α_k, $k = 1, \cdots, K$, is important.

Adaptive smoothing has long been used to create smooth representations of the data when the homogeneity of smoothness cannot be assumed across the domain (Liu & Guo , 2010). Various ways of "adapting to the varying smoothness" have been proposed over the years. Some approached the problem by placing knots and basis functions adaptively (Friedman, 1991; Luo & Wahba, 1997), others by applying some versions of varying smoothing parameters, such as the spans of the kernel (Muller & Stadtmuller, 1987), the parameters in a penalty matrix (Ruppert & Carroll, 2000), the weights in a mixture of splines (Wood *et al.*, 2002), etc. In the latter approach, the adaptive smoothing parameters are sometimes constructed as a function of an independent variable. This could be the variable used in the smoothing or an external variable which can inform the variability of the smoothness of the data. The adaptive smoothing method proposed in this paper follows this approach and chooses to construct the adjustment matrix $\boldsymbol{A}$ based on an independent variable that can reflect the uncertainty of the data at each time point. This variable will be referred to as the "uncertainty variable" in the following content.

In practice, the adaptive penalty can be implemented through a two-stage method. The time series of the uncertainty variable is smoothed first using standard P-spline smoothing. This is referred to as the "uncertainty smooth". Then the data time series (e.g. the mean chlorophyll-a time series) is smoothed using an adaptive penalty matrix, constructed based on the basis coefficients of the uncertainty smooth. This is referred to as the "value smooth". Similar two-stage approaches using an additional smoothing step on auxiliary data to provide information on the smoothness of the time series of interest can be found in Denis *et al.* (2020). In their paper, the first smooth is used to determine the change point in the time series and the second smooth is applied taking into account the change point.

Here the process of creating the adaptive penalty matrix is illustrated using the mean chlorophyll-a time series. In this particular example, the uncertainty variable is chosen to be the proportion of data that are missing at each time point. Denote X_t as the missing proportion at time t, and $X(t)$ its functional form. A standard P-spline smoothing is first applied to the logit transformed missing proportion time series $X^*(t) = \log(\frac{X(t)}{1-X(t)})$, as

$$X^*(t) = \Phi(t)\boldsymbol{\alpha}^* + u(t)\,. \tag{4}$$

where $\Phi(t)$ is a saturated basis and $\boldsymbol{\alpha}^*$ the basis coefficient vector. A back transformation is then applied to get the smoothed missing proportions $\widehat{X}(t)$, and a "change of basis"[2] $\widehat{X}(t) = \Phi(t)\boldsymbol{\alpha}$ is applied to obtain the set of coefficients α_k, $k = 1, \cdots, K$ that will be used to construct the adjustment matrix $\boldsymbol{A}$. The main purpose of the back transformation and the change of basis is to obtain a set of coefficients α_k, $k = 1, \cdots, K$, that falls within the interval of [0, 1]. Although the elements in the coefficient vector $\boldsymbol{\alpha}^*$ reflect the relative highs and lows of the missing proportions, they often go beyond the [0, 1] interval. The elements in the coefficient vector $\hat{\boldsymbol{\alpha}}$, however, usually stay in the interval. Setting the sequence α_k, $k = 1, \cdots, K$, to be the diagonal entries of the adjustment matrix $\boldsymbol{A}$ gives the main building blocks of the adaptive penalty matrix.

In the rare cases where some elements in the estimated coefficient vector $\boldsymbol{\alpha}$ go beyond the interval, one simply replaces the values beyond 1 with 1 and the values below 0 with 0. Should the scale difference between the elements of the estimated coefficient vector $\boldsymbol{\alpha}$ becomes unrealistically large, further adjustment may be applied to avoid over or under penalising certain parts of the time series. An example of this will be given in section 2.5.

2.4 The selection of degrees of smoothness

The final step before one can implement the adaptive smoothing method is to determine the degrees of smoothness of the time series curve. This is usually controlled by the smoothing parameter λ in the penalty term (3). Depending on the application background, the degrees of smoothness may be determined through a combination of automatic selection methods, such as the generalized cross validation (GCV), and background knowledge.

For the smoothing of a time series from a particular lake, the generalized cross validation (Ramsey & Silverman, 2005) is applied. However, since the construction of the penalty matrix $\boldsymbol{D}_\alpha$ in the adaptive penalty (3) depends on the result of smoothing the uncertainty time series, which itself depends on another smoothing parameter, the computation of the GCV scores becomes less straightforward. To be precise, the penalised least squares problem for adaptive smoothing is actually

$$\sum_t \|Y_t - \Phi(t)\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\top\mathcal{S}_\alpha(\lambda_0)\boldsymbol{\beta}\,, \tag{5}$$

where λ_0 comes from the P-spline smoothing of the uncertainty time series X_t, with the minimization criterion,

$$\sum_t \|X_t - \Phi(t)\boldsymbol{\alpha}\|^2 + \lambda_0\boldsymbol{\alpha}^\top\mathcal{S}\boldsymbol{\alpha}\,.$$

[2]This is the smoothing where the curve $\Phi(t)\boldsymbol{\alpha}$ is required to interpolate $\widehat{X}(t)$, which is a smoothed version of $X(t)$. No smoothness penalty is used.

In addition, there may not be a simple expression for $\mathcal{S}_\alpha(\lambda_0)$ as a function of λ_0. For example, in the illustration in section 2.3, the uncertainty variable X_t was transformed before the P-spline smoothing was applied, which was then followed by a change of basis. In fact, a transformation may be a common practice to ensure that the entries of matrix $\boldsymbol{A}$ lie in the interval of [0, 1]. Therefore, there is usually no explicit expression for $\mathcal{S}_\alpha(\lambda_0)$ as a function of λ_0 and hence no explicit expression for the GCV score as a function of λ and λ_0.

One possible solution is to carry out a grid search to find the appropriate combination of λ and λ_0. However, this could be computationally intensive. As a practical approach, in terms of the LSWT and the chlorophyll-a time series, this paper proposes to combine a coarse grid of λ_0 and a fine grid of λ. That is, a search on a coarse grid of λ_0 values is carried out first. Then for each candidate value of λ_0, the GCV scores

$$GCV_{\lambda_0}(\lambda) = \frac{\| [\boldsymbol{I} - \boldsymbol{H}_{\lambda_0}(\lambda)] \boldsymbol{Y} \|^2}{T^{-1}\text{trace}\{\boldsymbol{I} - \boldsymbol{H}_{\lambda_0}(\lambda)\}^2} , \tag{6}$$

are computed, where $\boldsymbol{H}_{\lambda_0}(\lambda) = \boldsymbol{\Phi}[\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\,\mathcal{S}_\alpha(\lambda_0)]^{-1}\boldsymbol{\Phi}^\top$ is the "hat matrix", for a sequence of λ on a much finer grid. The combination of λ_0 and λ that produce the smallest GCV score would usually be considered as an appropriate choice. This approach breaks the selection problem into two steps. It makes sense for the application in this paper because there are different priorities in terms of the uncertainty smooth and the value smooth. To some extent, the purpose of uncertainty smooth is only to extract some general information on the varying uncertainty of the data. Therefore, the selection of λ_0 is less crucial than the selection of λ and it is preferable to have a relatively large value for λ_0. In situations where over-fitting is to a concern, double cross validation, which aims to obtain the closest match between the predicted value of a time point when its observation is excluded and included in the fitting (Wood, 2017), may be used.

For the smoothing of a large number of time series, this paper proposes to select the effective degrees of freedom (EDF) of the smoothed time series directly, instead of tuning the smoothing parameters λ_0 and λ. This is based on the following consideration. Whenever the irregular basis is used, the number of basis functions used in the smoothing would be different from one lake to another. As a result, using the same smoothing parameters λ_0 and λ across all the time series would result in different degrees of smoothness. This could be problematic when the smoothed time series are used in the functional data analysis, as they are not directly comparable. A more sensible option is to choose the effective degrees of freedom that is appropriate for the majority of the time series using some (automatic) selection procedure. Then apply the same EDF to all time series.

According to Cantoni & Hastie (2002), there is a strictly monotone relationship between the smoothing parameter λ and the effective degrees of freedom df_λ in the type of smoothing problem $\widehat{\boldsymbol{Y}} = (\boldsymbol{I} + \lambda\boldsymbol{Q})^{-1}\boldsymbol{Y}$. This relationship can be written as $\text{df}_\lambda = \sum_t \frac{1}{1+\lambda d_t}$, where d_t is the t-th diagonal element of matrix $\boldsymbol{Q}$. It can be shown that a similar result holds for the adaptive smoothing approach proposed in this paper. The 1 to 1 mapping can be written as

$$\text{df}_\lambda = \sum_t \frac{c_t}{1 + \lambda d_t} ,$$

where c_t is the t-th diagonal element of another matrix associated with the penalty matrix (see Appendix A for more details). Therefore, following Cantoni & Hastie (2002), the problem of selecting the smoothing parameter λ can be converted to the problem of selecting the effective degrees of freedom df_λ.

Based on the above result, this paper proposes to investigate the GCV score as a function of the effective degrees of freedom df_λ, rather than a function of the smoothing parameter λ. The problem remains to compute the GCV score for all the time series. This is different from the smoothing of one time series, as different smoothing parameters are required for different time series in order to reach the same EDF. The basis matrix also differs if the irregular basis is used. Considering these aspects, this paper proposed to compute the GCV scores of an augmented penalised least squares problem, where all time series are column stacked into a long vector, $\tilde{\boldsymbol{Y}} = (\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_N^\top)^\top$, all basis matrices are constructed into a block diagonal matrix $\tilde{\boldsymbol{\Phi}} = \text{diag}\{\boldsymbol{\Phi}_1^\top, \cdots, \boldsymbol{\Phi}_N^\top\}$

and all basis coefficient vectors are column stacked into a long vector $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1^\top, \cdots, \boldsymbol{\beta}_N^\top)^\top$. The associated penalised least squares problem can be written as

$$\|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\beta}}\|^2 + \tilde{\boldsymbol{\beta}}^\top \tilde{\mathcal{S}}_{df} \tilde{\boldsymbol{\beta}} \,, \tag{7}$$

where $\tilde{\mathcal{S}}_{df}$ is

$$\tilde{\mathcal{S}}_{df} = \begin{pmatrix} \lambda_1 \boldsymbol{D}_1^\top \boldsymbol{D}_1 & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \lambda_N \boldsymbol{D}_N^\top \boldsymbol{D}_N \end{pmatrix} ,$$

and $\lambda_1, \cdots, \lambda_N$ are chosen such that the effective degrees of freedom of all N smoothed time series are the same. Since all matrices involved in the calculation of this GCV score are block diagonal, it is relatively easy to compute its value using the sum of squared errors from the smoothing of individual time series.

Finally, the standard errors of the smoothed time series can be computed as $\text{Cov}[\widehat{\boldsymbol{Y}}] = \boldsymbol{H}_{\lambda_0}(\lambda)\boldsymbol{\Sigma}_Y \boldsymbol{H}_{\lambda_0}(\lambda)^\top$, where $\boldsymbol{\Sigma}_Y$ is the estimated covariance matrix of $\boldsymbol{Y}$. Hence, the standard errors of the smoothed time series can be obtained from the diagonal elements of the covariance matrix. Note that in the adaptive smoothing of a large number of time series, each time series may have its own smoothing parameters λ_0 and λ for the selected effective degrees of freedom as shown in problem (7).

2.5 The impact of adaptive smoothing

To illustrate the impact of adaptive smoothing, this section presents an example using the complete reconstruction of the lake surface water temperature of Lake Victoria in Africa. This data set is available from http://www.laketemp.net/home_ARCLake/index.php. It contains monthly satellite images of the LSWT of Lake Victoria from June 1995 to April 2012.

The area covering Lake Victoria consists of 2313 pixels, which is shown by the image in the left panel of Figure 4. Initially, the mean LSWT time series was computed by taking the spatial average of the complete reconstruction of the LSWT. It was then centred to have zero mean. To mimic the situation where averaging over a small number of clustered pixels results in bias in the mean LSWT time series, four time points were selected and the mean LSWT at these time points were replaced by the average of the observations in the northeast corner and the southeast corner of the lake. The two corners are depicted by the black boxes in the left panel of Figure 4. For this particular lake, the northeast corner tends to have higher LSWT and the southeast corner tends to have lower LSWT than the rest of the lake. The resulting mean LSWT time series with bias is shown in the top right panel of Figure 4. The proportion of missing observations at each time point was used to create the adaptive penalty matrix in this case. The missing proportion time series is shown in the middle right panel as the black vertical lines. The elements α_k, $k = 1, \cdots, K$ of the adaptive penalty matrix, which were obtained from the smoothing of the missing proportion time series, were shown as the black dots. The effective degrees of freedom used in this stage is 34, corresponding to placing one knot every six months. The bottom right panel shows the smoothed mean LSWT time series from the standard P-spline smoothing (blue curve) and that from the adaptive smoothing (red curve), both with effective degrees of freedom 81, which corresponds to placing one knot every 2.5 months. The mean LSWT from the complete reconstruction was also shown in the bottom right panel as the grey curve with dots.

It can be seen from the bottom right panel of Figure 4 that the blue curve from standard P-spline smoothing tracks the pattern throughout the time series in the same manner. It did not distinguish a datum with higher accuracy (i.e. a spatial average from a complete image) from a datum with higher uncertainty (i.e. a spatial average from a few clustered pixels). Therefore, if the datum contains biased information, the resulting smoothed time series is also likely to be biased. In this case, although the resulting smoothed time series reflects the patterns in the time series in the top right panel, it under/over estimated the true mean LSWT (the grey curve

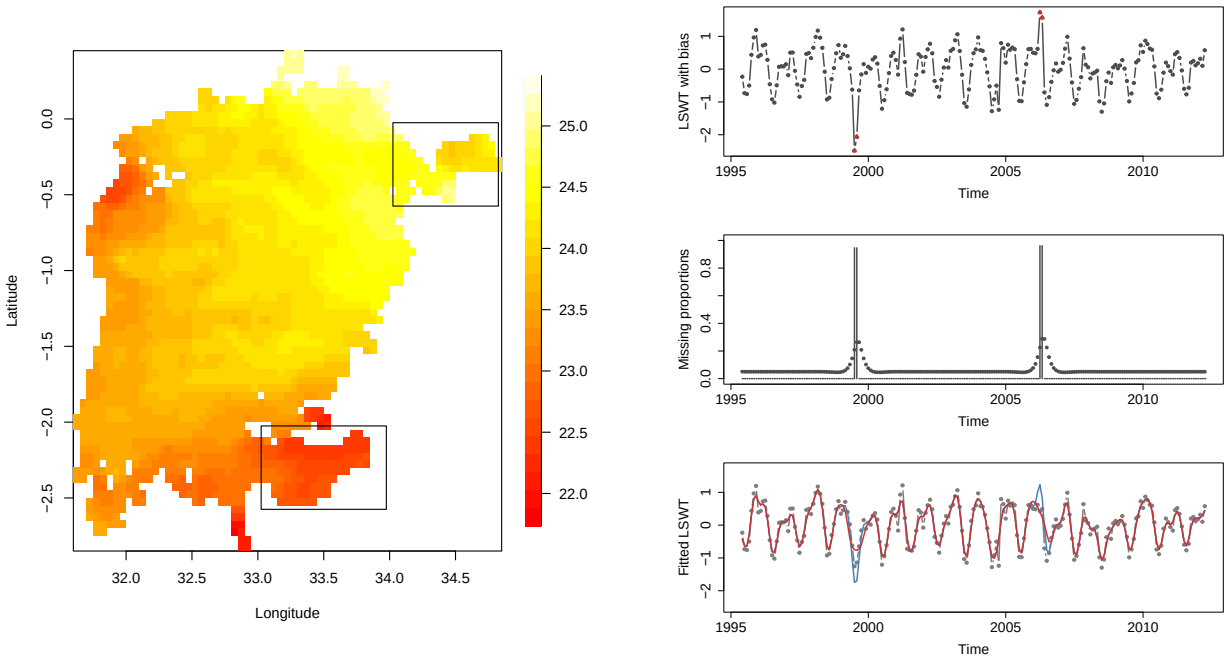

Figure 4: (Left) A map showing the lake surface water temperature data (in °C) from Lake Victoria in July 1999 and the two areas (black boxes) used to created the biased mean LSWT. (Right) The centred LSWT time series from Lake Victora with inserted biased data shown as red triangles (top), the missing proportion time series and the α_k, $k = 1, \cdots, K$ series (middle), and the smoothed LSWT time series from standard P-spline smoothing (blue curve) and adaptive smoothing (red curve) (bottom).

in the bottom right panel) around the time points where there are manually introduced bias. The red curve from adaptive smoothing avoided this problem by penalising the datum with higher uncertainty. The resulting smoothed value stays closer to the mean of the time series, which in this case reflect the patterns in the true mean LSWT more appropriately.

It may happen that the spatial average from the clustered pixels in the two corners coincides with the true mean. In such case, the adaptive smoothing would produce a seemingly worse result by not following the datum with higher uncertainty closely. However, the truth is often unknown in practice. Hence an estimation closer to the mean value of the time series and an estimation closer to the datum are equally likely to be biased, with the latter having a higher chance of bringing in extreme values to the result. This is a situation that people try to avoid in most applications, which is also where adaptive smoothing shows its advantage. Further examples showing the impact of adaptive smoothing using the two lake Chl-a time series in Figure 1 can be found in Appendix B.

The method for obtaining the adaptive penalty matrix may occasionally result in over adjustment due to the large scale difference between $\hat{\alpha}_k$, $k = 1, \cdots, K$. This may introduce artefacts to the smoothed time series. For example, some parts of the smoothed time series may be pulled towards the mean of the time series by the overly large penalties introduced by those large $\hat{\alpha}_k$ values. A motivating example can be found in Appendix B. Different methods may be used to modify the sequence $\hat{\alpha}_1, \cdots, \hat{\alpha}_K$ (denoted as $\{\hat{\alpha}_k\}$ for convenience). Ideally, the modified adjustment vector should have (i) all values in the range $(0, 1)$, (ii) the scale differences smaller than a specific threshold, (iii) elements with the same order (ties are allowed) as the original vector, i.e. for $\hat{\alpha}_k > \hat{\alpha}_j$, it is required that $f(\hat{\alpha}_k) \geq f(\hat{\alpha}_j)$ after transformation. For example, a square root transformation

with a lower cap may be used, i.e.

$$\alpha_k^* = \sqrt{\max\{\hat{\alpha}_k, \delta^2\}}\,. \tag{8}$$

This operation reduces the scale difference between the elements in sequence $\{\hat{\alpha}_k\}$. The lower cap δ^2 further ensures that the scale difference is less than $1/\delta$.

In summary, the adaptive smoothing method proposed in section 2 produces smoothed LSWT time series that are less likely to over-fit the data produced through averaging a small number of lake pixels, which are often associated with higher uncertainty. The smoothed time series curve displays less dramatic fluctuations than that from the standard P-spline smoothing which uses a standard 2nd difference penalty matrix. This can be an advantage in application as the outcomes are less likely to be affected by noise or biased information.

3 Investigating the spatial structure in lake ecological process

This section presents two applications of the proposed modelling procedure combining adaptive smoothing and functional data analysis to the mean LSWT time series and the mean chlorophyll-a time series from 932 lakes in the GloboLakes repository. The application to the mean LSWT time series aimed at identifying the spatial structure at a global scale in the seasonal dynamics of LSWT from the major lakes on Earth. The application to the mean chlorophyll-a time series sought to explore the potential spatial structure in the seasonal and trend signals of lake chlorophyll-a from the major lakes on Earth.

3.1 Preparing the data via adaptive smoothing

The GloboLakes repository contains the satellite remote sensing lake surface water temperature and lake chlorophyll-a data from 932 largest lakes on Earth. Based on data availability, 732 lakes were selected in the analysis of the LSWT and 535 lakes were selected in the analysis of the lake chlorophyll-a.

The bi-monthly lake surface water temperature time series consist of 398 time points, covering a period from summer 1995 to spring 2012. After the spatial averaging within lakes, most of the lakes have (near) complete time series and the data availability is usually high. Therefore, the use of an adaptive penalty matrix is not necessary. However, the LSWT observations can be affected by ice cover. When ice was detected, the water below the ice was set to 0°C. To reflect this property, the irregular basis created using the method in section 2.2 was adopted, which helped to ensure that the smoothed LSWT time series would respect the 0°C constraint. Figure 5 gives two examples of the smoothed LSWT time series. The smoothed time series (red curve) in the top left panel represents Chamo Lake, a tropical lake with good data availability. The top right panel represents Lac des Bois, which is a high latitude lake. An irregular basis was used here to account for the 0°C constraint over the winter periods. Part of this irregular basis was presented in the bottom panel of Figure 3.

The spatially averaged monthly lake chlorophyll-a time series consist of 101 time points from July 2003 to November 2011. To enable the analysis, the time series were log transformed and centred. The resulting time series will be referred to as the "centred log(Chl-a)" time series. Two types of smoothed signals were considered, the smoothed trend series and the smoothed seasonal series. Depending on data availability and modelling aims, either standard P-spline smoothing or adaptive smoothing was applied to the data. For the latter, the adaptive penalty matrix was constructed from the coefficients of the smoothing of the missing proportion time series as in (4). In some rare cases, further adjustment using the transformation (8) was made to the adaptive penalty matrix to avoid over-penalising the data.

The degrees of freedom of the smoothed time series was selected using the method in section 2.4. In particular, the EDF used in the uncertainty smooth was selected from a list of values, corresponding to placing a knot every 2, 3, 4, 6 months. Double cross validation as described in Wood (2017) was used to prevent over-fitting.

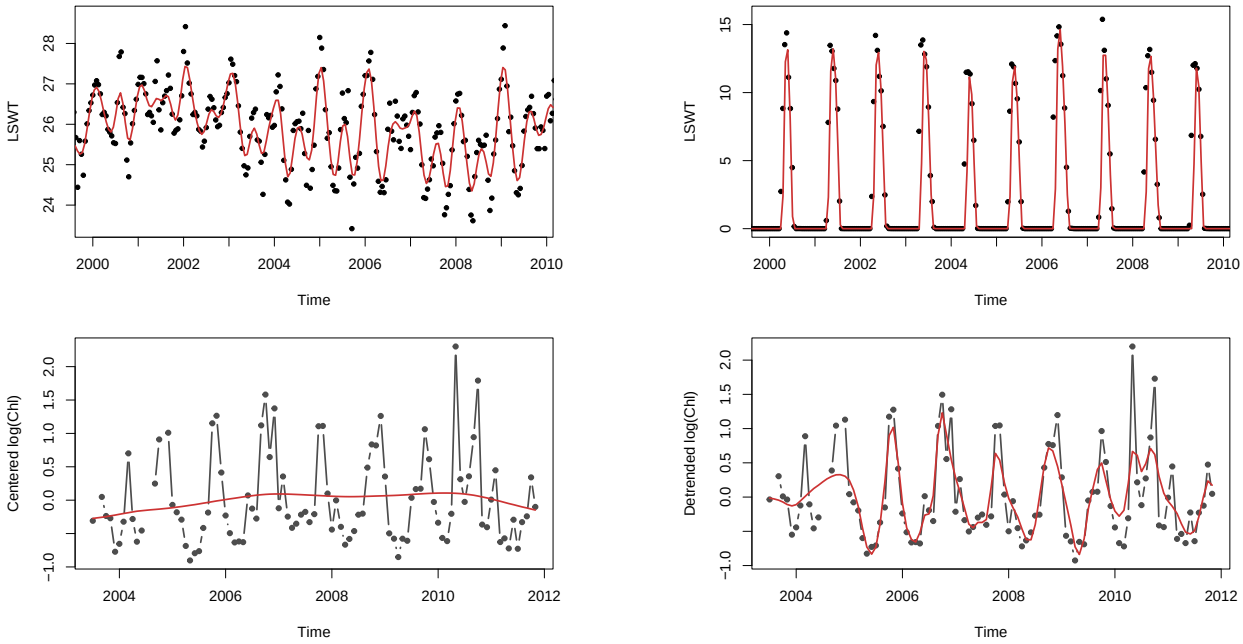

Figure 5: The observed (black dots) and the smoothed (red curve) mean LSWT time series of Chamo Lake (top left) and Lac des Bois (top right), from 2000 to 2010. The centred log(Chl-a) time series and the smoothed trend signal of Lake Tanganyika (bottom left). The de-trended log(Chl-a) time series and the smoothed seasonal signal of Lake Tanganyika (bottom right).

The cross validation scores indicated that a degrees of freedom of 16.83 (i.e. one knot every six months) was appropriate. The EDF used in creating the smoothed trend of the centred log(Chl-a) time series was set to 4 to capture the large scale temporal pattern over time. The EDF used in the smoothing of the time series after removing the smoothed trend was selected from another list of values, corresponding to placing a knot every 2, 2.25, ..., 4.75, 5, 5.5, 6 months. The plot of GCV scores against the EDF values has a bend at around 28 (see Appendix C), which is equivalent to placing one knot every 3.5 months. Considering the interpretation of the result in the real application, a final decision was made to use an effective degrees of freedom of 34, which is equivalent to placing one knot every three months (i.e. one knot per season). Using the selected effective degrees of freedom, the smoothed trend signals and the smoothed seasonal signals in the centred log(Chl-a) time series were extracted. Examples of the smoothed trend and seasonal curves from the centred log(Chl-a) time series of Lake Tanganyika are presented in the bottom left and bottom right panel of Figure 5.

3.2 Identifying the spatial structure via functional data analysis

After obtaining the smoothed representations of the LSWT time series and the smoothed trend and seasonal signals of the lake chlorophyll-a time series, functional data analysis techniques were applied to the smoothed time series respectively to explore the spatial structure in the temporal dynamics of the two ecological variables.

A novel investigation on global lake thermal region shift using the smoothed LSWT time series was conducted in Maberly *et al.* (2020). In particular, nine lake thermal regions were identified through functional principal component analysis (PCA) (Ramsay *et al.*, 2021), followed by quadratic discriminant analysis on the functional principal component scores. Figure 1 in Maberly *et al.* (2020) presents the cluster memberships on a map, along with the curves representing the cluster centers. An R-shiny app was developed to help visualise the thermal regions.

It can be found on GitHub through the link https://github.com/ruth-odonnell/LakeThermalRegions/.

In the analysis of the centred log(Chl-a) time series, functional PCA was first applied to the smoothed trend and seasonal time series respectively. Then cluster analysis was applied to the subsets of the identified functional principal components (PCs) which explain over 50% of the variation in the smoothed trend and seasonal time series. In particular, 11 functional PCs were used in the clustering of the trend signals and 22 functional PCs were used in the clustering of the seasonal signals. Various clustering methods were explored, including K-means clustering (Hartigan & Wong, 1979) and model based clustering using mixture models (Fraley & Raftery, 2002). The results from K-means clustering tend to be the most robust and they were presented in this section.

After removing four outlying curves, the number of clusters were selected considering both the gap statistics (Tibshirani *et al.*, 2001) and the ecological interpretation. The Rand index (Scrucca *et al.*, 2016) was computed and the functional ANOVA (Febrero-Bande & Fuente, 2012) was applied to the clustering result to examine the similarities and differences between the identified clusters. As suggested by the gap statistics, both four and eight clusters for the smoothed trend signals seem to be appropriate. Here the result from clustering the smoothed trend signals into eight clusters is presented in Figure 6. The top of Figure 6 shows eight different types of interannal variation identified from the smoothed trend signals. The bottom of Figure 6 shows the locations of the lakes in each cluster on a map. The result from clustering the smoothed seasonal signals into ten clusters is presented in Figure 7. There are some subtle spatial patterns in the Northern Hemisphere in terms of the seasonal signals where the members in cluster 4, 7 and 8 appear to share common features in their latitude. However, the majority of the clusters seem to consist of lakes from across the globe, which do not appear to follow particular geographic patterns constrasting to the clustering result of the LSWT time series. More details of the clustering are available in the R-Shiny app with the link https://github.com/GMY2018/ChlCluster. In this app, individual maps of each cluster are presented, which provides more information on the spatial patterns. Future work involves detailed interpretation of the clustering result for ecology and limnology.

4 Conclusion and discussion

4.1 Conclusion

Motivated by the real application problem of the GloboLakes project, this paper introduced a modelling procedure that combines adaptive smoothing and functional data analysis to identify the spatial structure in the temporal dynamics of lake ecological variables. To account for the varying uncertainty in the spatially averaged satellite remote sensing LSWT time series and lake chlorophyll-a time series, the paper proposed two adjustments to standard P-spline smoothing based on the idea of adaptive smoothing. Firstly, an irregular basis that matches the missing patterns in the time series and respects the specific constraints on the observations was used to mitigate the impact of these features. Secondly, an adaptive penalty matrix that assigns different penalties to different time points was created to adjust the smoothness of the time series and to prevent over-fitting based on the uncertainty associated with the proportion of data used in the estimation. The paper also presented a GCV-based method to select the degrees of freedom for a large number of smoothed time series so that they were comparable in the functional data analysis to investigate the spatial structure.

The proposed methods were applied to the mean LSWT time series from 732 lakes and the mean lake chlorophyll-a time series from 535 lakes globally. In the case of the mean LSWT time series, the irregular basis was used to handle the missing gaps and the measurement constraint on surface water temperature. In the case of the mean lake chlorophyll-a time series, the adaptive penalty matrix was constructed based on the missing proportion time series which reflect the uncertainty associated with the spatial average. This produced smoothed time series that are less prone to bias or noise. Functional principal component analysis, coupled with quadratic discriminant analysis and K-means clustering, were then applied to the smoothed LSWT time series and the smoothed lake chlorophyll-a trend and seasonal signals. The analysis of the smoothed LSWT time series identified nine

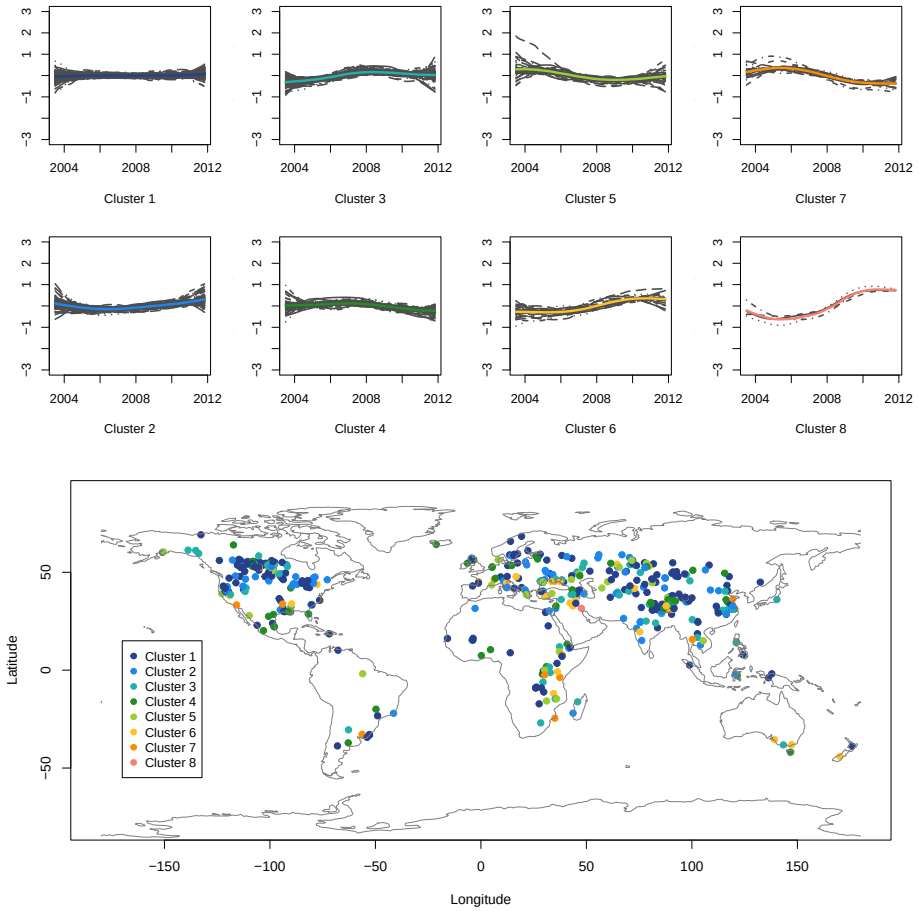

Figure 6: (Top) The smoothed trend signals extracted from the centred log(Chl-a) from eight clusters (grey curves) and the curves representing the cluster centers (colourful curves). (Bottom) The clustering result shown on a map, where each dot represents a lake and the colour represents the cluster membership.

clusters representing the global lake thermal regions. The clustering of the smoothed lake chlorophyll-a data extracted eight global lake clusters describing the interannual variation and ten clusters differentiating the seasonal signals. The two applications demonstrated the advantages of the proposed method in modelling a large number of lake ecological time series and the interesting information that can be extracted from the modelling result.

4.2 Potential choices of the uncertainty measure

Although some implementation details presented in the paper are problem specific, the method itself is flexible. It can be generalised to other problems with appropriate choices of basis, uncertainty measure and functional data analysis techniques. Here potential choices of the uncertainty measure of the spatial average in more general settings are discussed.

Satellite remote sensing data product may contain uncertainty measurements and/or error measurements (i.e. deviation from the truth) of the retrieved data at different resolutions (Povey & Grainger, 2015). These data can provide valuable information to the implementation of the adaptive smoothing method. Summary statistics, such as mean, standard deviations and quantiles, can be computed from the pixel level uncertainty/error

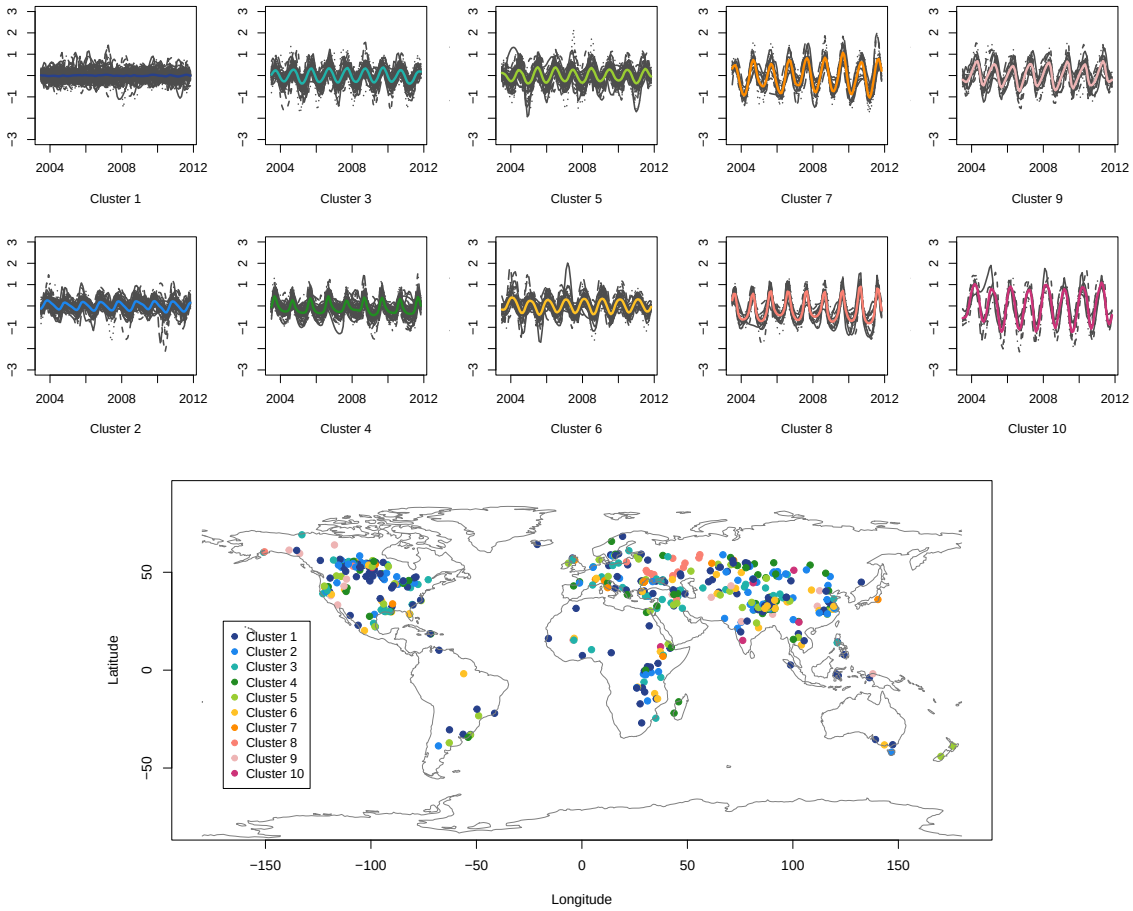

Figure 7: (Top) The smoothed seasonal signals extracted from the centred log(Chl-a) from ten clusters (grey curves) and the curves representing the cluster centers (colourful curves). (Bottom) The clustering result shown on a map, where each dot represents a lake and the colour represents the cluster membership.

measurements to quantify the uncertainty of the spatial average. The summary statistics can then be used to construct the adaptive penalty matrix. There have been extensive research on the uncertainties of the retrieved data within the satellite remote sensing community, along with discussions on the appropriate communication of the uncertainty (Povey & Grainger, 2015; von Clarmann *et al.*, 2020). It can be expected that such data will become increasingly accessible in the future.

Geostatistical methods may also be used to obtain the uncertainty measure. Aubry & Debouzie (2000) described a problem of quantifying the uncertainty of the spatial average using geostatistics tools. They defined the uncertainty measure as $\boldsymbol{E}[(Z_R - Z_R^*)^2]$, where Z_R is the spatial mean over region R and Z_R^* is the spatial average computed from the observations $z_1, \cdots, z_n$ in region R. They showed that this measure can be calculated using the estimated variogram model $\gamma(\cdot,\cdot)$ as $2\sum_{i=1}^n \lambda_i \overline{\gamma}(s_i, R) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \overline{\gamma}(s_i, s_j) - \overline{\gamma}(R, R)$, where λ_i, $i = 1, \cdots, n$, are the weights used in the spatial averaging, s_i represents the location of observation z_i, and $\overline{\gamma}(s_i, R) = \int_R \gamma(u - s_i)du/|R|$, $\overline{\gamma}(R, R) = \int_R \int_R \gamma(u - v)dudv/|R|^2$. The derivation of this expression can be traced back to Matheron's work in 1965. An estimation of the spatial averaging uncertainty may also be obtained through the mean-squared prediction error (MSPE) from kriging or block kriging as described in Chapter 3 of Cressie (1993). For example, one may obtain an prediction of the region R and its MSPE based

on data from sub-regions $R_1, \cdots, R_m$, using simple (block) kriging with the variogram model, $\gamma(\cdot, \cdot)$, estimated from the data. In particular, the MSPE can be calculated as $\sigma^2(R) = \boldsymbol{\lambda}^\top(R)\boldsymbol{\gamma}(R) - \gamma(R, R)$, where $\boldsymbol{\lambda}(R)$ is the vector of kriging coefficients, $\boldsymbol{\gamma}(R) = (\gamma(R, R_1), \cdots, \gamma(R, R_m), 1)^\top$, and $\gamma(R, R_i) = \int_R \int_{R_i} \gamma(u - s_i) ds_i du / |R_i||R|$ (Cressie, 1993).

Both measures require the inspection of the individual spatial images in order to estimate the variogram model and carry out the kriging interpolation, which can be computationally expensive. Hence, they are not suitable for the application in this paper, where computational efficiency was prioritised. The occasional high proportion of missing data during winter periods presents another challenge. Nevertheless, when the conditions are right, these measures will provide an appropriate uncertainty quantification for the spatially averaged time series.

4.3 Future extensions

Finally, potential extensions to the proposed methods are considered. As a result of the two-stage approach when using the adaptive penalty, two smoothing parameters are required and each of them need to be selected appropriately. Here the selection of smoothing parameters were guided by the application. The smoothing parameters used in the uncertainty smooth was chosen from a coarser grid to capture a general pattern, and the smoothing parameter of the value smooth was selected over a finer grid to find the optimal solution. For smoothing problems where no prior or background knowledge is available, the grid search on a tensor could be tedious and computationally expensive. Further investigation is needed to develop a method where the two smoothing parameters can be determined in a computationally efficient way.

The appropriate levels of smoothness of the time series investigated in this paper are similar. Hence it is relatively easy to describe them using smooth functions of the same degrees of freedom and apply functional data analysis. When it comes to time series with distinctively different levels of smoothness, the methods developed in this paper may not be appropriate. This is a more general problem that is worthy of investigation, but it goes beyond the scope of this paper.

Acknowledgement

We thank Iestyn Woolway, Mark Cutler, Ian Jones, Eirini Politi and Stephen Thackeray for the helpful discussions on the spatial structure in the temporal dynamics of the lake surface water temperature time series and the lake chlorophyll-a time series.

This work was supported by GloboLakes funded by the Natural Environment Research Council (grant number NE/J021717/1).

Data availability

The satellite-derived chlorophyll-a data used in this paper were produced within the NERC GloboLakes project and can be made available on request from calimnos-support@pml.ac.uk. A follow-on dataset is operationally available in the form of 10-day aggregated Trophic State Index and Turbidity, freely available from the operational Copernicus Land Monitoring Service: https://land.copernicus.eu/global/products/lwq. The lake surface water temperature data are publicly available from the ARC Lake project: http://www.laketemp.net/home_ARCLake/data_access.php.

References

Aubry, P., Debouzie, D., 2000. Geostatistical estimation variance for the spatial mean in two-dimensional systematic sampling. *Ecology*, 81:2, 543–553.

Cantoni, E., Hastie, T., 2002. Degrees-of-freedom test for smoothing splines. *Biometrika*, 89 (2), 251–263.

Craven, P., Whaba, G., 1979. Smoothing noisy data with spline functions Estimate the correct degree of smoothing. *Numerische Mathematic*, 31, 377–403.

Cressie, N. A. C., 1993. *Statistics for Spatial Data (revised edition)*. John Wiley & Sons, Incorporated.

Denis, D., Lebarbier, E., Lévy-Leduc, C., Martin, O., Sansonnet, L., 2020. A novel regularized approach for functional data clustering: an application to milking kinetics in dairy goats. *Journal of Royal Statistical Society, Series C (Applied Statistics)*, 69:3, 623–640.

Eilers, P. H. C., Marx, B. D., 1996. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:2, 89–121.

Eilers, P. H. C., Marx, B. D., Durban, M., 2015. 20 years of P-splines. *Statistics and Operations Research Transactions*, 39:2, 149–186.

Febrero-Bande, M., de la Fuente, M. O., 2012. Statistical Computing in Functional Data Analysis: The R Package fda.usc. *Journal of Statistical Software*, 51:2, 1–28, http://www.jstatsoft.org/v51/i04/

Fraley, C., Raftery, A., 2002. Model-based clustering, discriminant analysis, and density estimation *Journal of the American Statistical Association*, 97:458, 611–631.

Friedman, J. H., 1991. Multivariate adaptive regression splines. *The Annals of Statistics*, 19:1, 1–67.

Hartigan, J. A., Wong, M. A., 1979. Algorithm AS 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28:1, 100–108.

Maberly, S. C., O'Donnell, R. A., Woolway, R. I., Cutler, M. E. J., Gong, M., Jones, I. D., Marchant, C. J., Miller, C. A., Politi, E., Scott, E. M., Thackeray, S. J., Tyler, A. N., 2020. Global lakes thermal regions shift under climate change. *Nature Communications*, 11:1232. https://doi.org/10.1038/s41467-020-15108-z

Muller, H. G., Stadtmuller, U., 1987. Variable bandwidth kernel estimators of regression curves. *The Annals of Statistics*, 15:1, 182–201.

MacCallum, S., Merchant, C., (2013). ATSR reprocessing for climate lake surface water temperature: ARC-Lake: algorithm theoretical basis document. (Available from http://www.geos.ed.ac.uk/arclake/ARC-Lake-ATBD-v1.4.pdf)

Liu, Z., Guo, W., 2010. Data driven adaptive spline smoothing. *Statistica Sinica*, 20, 1143–1163.

Luo, Z., Whaba, G., (1997). T Hybrid Adaptive Splines. *Journal of the American Statistical Association*, 92:437, 107–116, https://doi.org/10.1080/01621459.1997.10473607

Povey, A. C., Grainger, R. G., 2015. Known and unknown unknowns: uncertainty estimation in satellite remote sensing. *Atmospheric Measurement Techniques*, 8, 4699–4718. https://doi.org/10.5194/amt-8-4699-2015

Febrero-Bande, M., de la Fuente, M. O., 2012. fda: Functional Data Analysis. R package version 5.5.1. https://CRAN.R-project.org/package=fda

Ramsay, J. O., Silverman, B. W., 2005. *Functional Data Analysis (second edition)*. Springer, New York.

Ruppert, D., Carroll, R. J., 2000. Spatially adpative penalties for spline fitting. *Australia & New Zealand Journal of Statistics*, 42:2, 205–223.

Scrucca L., Fop M., Murphy T. B. and Raftery A. E., 2016 mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8 (1), 289–317.

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63:2, 411–423.

von Clarmann, T., Degenstein, D. A., Livesey, N. J., Bender, S., Braverman, A., Butz, A., Compernolle, S., Damadeo, R., Dueck, S., Eriksson, P., Funke, B., Johnson, M. C., Kasai, Y., Keppens, A., Kleinert, A., Kramarova, N. A., Laeng, A., Langerock, B., Payne, V. H., Rozanov, A., Sato, T. O., Schneider, M., Sheese, P., Sofieva, V., Stiller, G. P., von Savigny, C., Zawada, D., 2020. Overview: Estimating and reporting uncertainties in remotely sensed atmospheric composition and temperature. *Atmospheric Measurement Techniques*, 13, 4393–4436. https://doi.org/10.5194/amt-13-4393-2020

Wood, S. A., Jiang, W., Tanner, M., 2002. Bayesian mixture of splines for spatially adaptive nonparametric regression. *Birmetrika*, 89:3, 513–528.

Wood, S. N., 2017. *Generalized Additive Models An introduction with R (second edition)*. Chapman & Hall, Boca Raton, FL.

A The relationship between λ and df_λ

The relationship between the smoothing parameter λ and the corresponding degrees of freedom df_λ in Cantoni & Hastie (2002), $\mathrm{df}_\lambda = \sum_t \frac{1}{1+\lambda d_t}$, is derived for smoothing splines and a particular smoothing problem with the hat (or projection) matrix being $(\boldsymbol{I}+\lambda\boldsymbol{Q})^{-1}$. Here a slightly different projection matrix $\boldsymbol{H} = \boldsymbol{\Phi}[\boldsymbol{V}(\boldsymbol{I}+\lambda\boldsymbol{Q})^{-1}\boldsymbol{V}^\top]\boldsymbol{\Phi}^\top$. In case an orthogonal basis $\boldsymbol{\Phi}$ is used, then the same relationship between λ and df_λ applies. For the rest of the situations, start by rewriting the matrix inverse in the hat matrix $\boldsymbol{H} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\mathcal{S}_\alpha)^{-1}\boldsymbol{\Phi}^\top$ using the simplification method in Chapter 5 of Ramsey & Silverman (2005). This starts with the eigenproblem $\mathcal{S}_\alpha\boldsymbol{V} = \boldsymbol{\Phi}^\top\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{V}\boldsymbol{Q}$, where $\boldsymbol{Q}$ is the eigenvalue matrix of $\mathcal{S}_\alpha$ in the metric defined by $\boldsymbol{\Phi}^\top\boldsymbol{W}\boldsymbol{\Phi}$ with weight matrix $\boldsymbol{W}$ ($\boldsymbol{W} = \boldsymbol{I}$ in this paper), and $\boldsymbol{V}$ is the corresponding eigenvector matrix satisfying $\boldsymbol{V}^\top\boldsymbol{\Phi}^\top\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{V} = \boldsymbol{I}$. It then follows that $\boldsymbol{V}^\top\boldsymbol{\Phi}^\top\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{V} = \boldsymbol{I}$ and $(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \lambda\mathcal{S}_\alpha)^{-1} = \boldsymbol{V}(\boldsymbol{I}+\lambda\boldsymbol{Q})^{-1}\boldsymbol{V}^\top$ as in Ramsey & Silverman (2005), where the inverse is much easier to compute as the eigenvalue matrix $\boldsymbol{Q}$ is diagonal. The trace of the hat matrix then follows as

$$\begin{aligned}
\mathrm{df}_\lambda &= \mathrm{trace}\left\{\boldsymbol{\Phi}\left[\boldsymbol{V}(\boldsymbol{I}+\lambda\boldsymbol{Q})^{-1}\boldsymbol{V}^\top\right]\boldsymbol{\Phi}^\top\right\} \\
&= \mathrm{trace}\left\{(\boldsymbol{I}+\lambda\boldsymbol{Q})^{-1}\boldsymbol{V}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{V}\right\} \\
&= \sum_{t=1}^{T}\frac{c_t}{1+\lambda d_t}\,,
\end{aligned}$$

where c_t, $t = 1, \cdots, n$, are diagonal elements of matrix $\boldsymbol{V}^\top\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\boldsymbol{V}$ and d_t, $t = 1, \cdots, n$, are diagonal elements of matrix $\boldsymbol{Q}$. This represents a strictly monotone relationship between λ and df_λ as both c_t and d_t come from matrices that are determined by the basis matrix and the penalty matrix, and do not involve λ.

B Further information on the impact of adaptive smoothing

Figure 1 in the introduction shows the spatially averaged log(Chl-a) time series of Lake Tanganyika and Lake Chany. To create the smoothed representations of the two time series, the irregular basis and the adaptive penalty matrix introduced in section 2 were used. Both time series have length equals to 101, covering a period from autumn 2003 to winter 2011. Uncertainty smooth was carried first. The effective degrees of freedom (EDF) was set to 17, corresponding to one knot every 6 months. The estimated basis coefficients from uncertainty smooth were then used to construct the adaptive penalty matrix. Finally, value smooth was applied to the log(Chl-a) time series respectively. The EDF was set to 34, corresponding to one knot every 3 months. The result is presented in Figure 8.

The log(Chl-a) time series of Lake Tanganyika is nearly complete. The missing data proportions are relatively low for most of the time, apart from the few months between 2004 and 2005. It is straightforward to see the impact from the higher penalties applied to the beginning of the time series from the bottom left panel. Whereas the blue curve (from standard P-spline smoothing) follows the peaks and troughs in the time series; the red curve (from adaptive smoothing) displays a more dampened signal in response to the higher uncertainty. The log(Chl-a) time series of Lake Chany has clear seasonal missing pattern, which is typical to high latitude lakes. Higher penalties were applied to the time points at the beginning of 2004, 2007 and 2008. Their impact was reflected by the difference between the blue curve and the red curve in the bottom right panel. Using adaptive smoothing prevented the smoothed time series from tracking the distinctively high value in the early spring of 2007, where the corresponding missing proportion is 0.932.

For an motivating example of the adjustment discussed in section 2.5, the log(Chl-a) time series of Lake Ngangze in China and its adaptive smoothing outcome is presented in the right panel of Figure 9. In this case, the largest of the $\hat{\alpha}_k$ is over 100 times larger than the smallest of $\hat{\alpha}_k$. As a result, some parts of the smoothed time series were dragged towards 0, as shown by the red curve in the bottom right panel of Figure 9. Therefore, a modification using the square root transformation was made. The result of this modification is shown as the blue

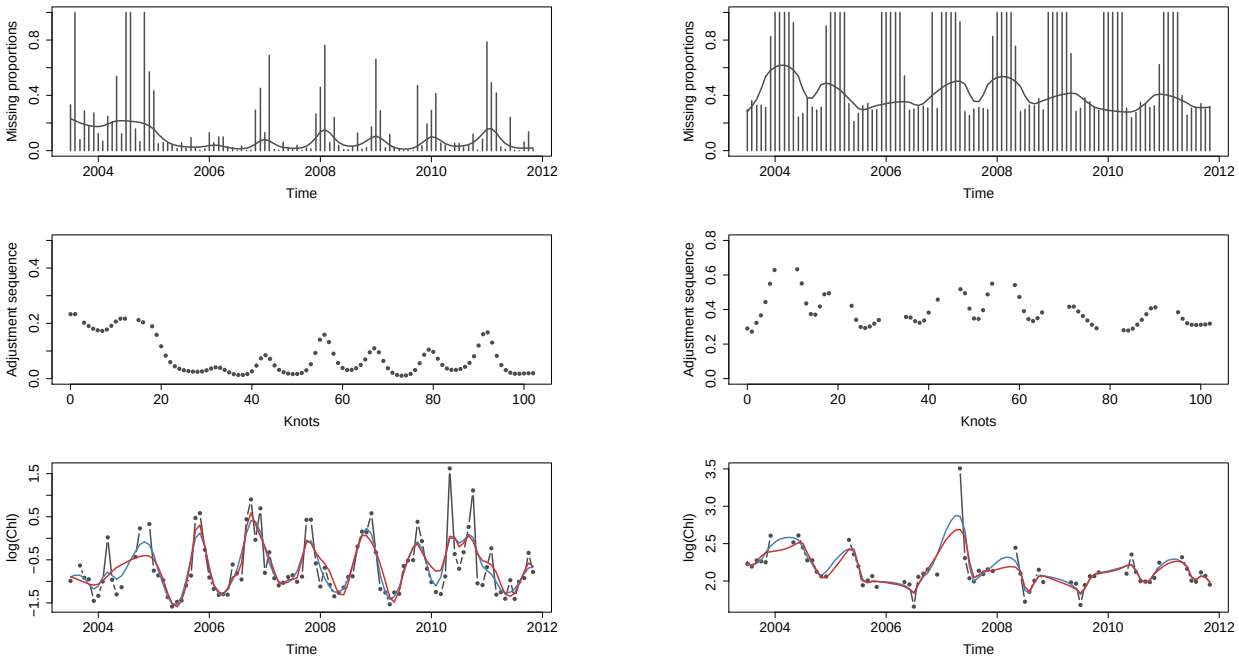

Figure 8: The smoothed missing proportion time series (top), the adjustment sequence used to create the adaptive penalty matrix (middle) and the smoothed log(Chl-a) time series from standard P-spline smoothing (blue curve) and adaptive smoothing (red curve) (bottom) of Lake Tanganyika (left) and Lake Chany (right).

dots in the middle right panel of Figure 9, representing the adjusted $\{\hat{\alpha}_k\}$, and the blue curve in the bottom right panel of Figure 9, representing the smoothed time series. The threshold used in this illustration is $\delta = 0.01$.

C GCV scores for selecting the degrees of freedom

The GCV scores for selecting the degrees of freedom for the lake chlorophyll-a time series was presented in Figure 10.

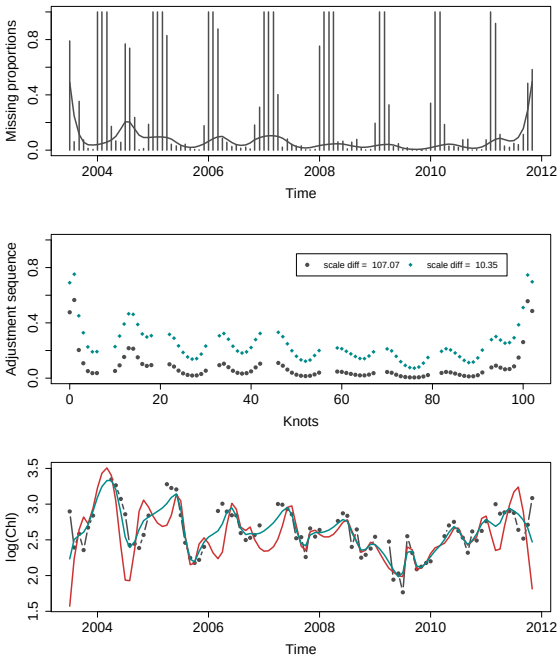

Figure 9: An example of the over adjustment using the log(Chl-a) time series of Lake Ngangze. The middle panel shows the basis coefficients $\hat{\alpha}_k$, $k = 1, \cdots, K$ (black dots), and the square root adjusted version (blue dots). The bottom panel shows the two smoothed curves using the usual adaptive penalty matrix (red curve) and the adjusted adaptive penalty matrix (blue curve).

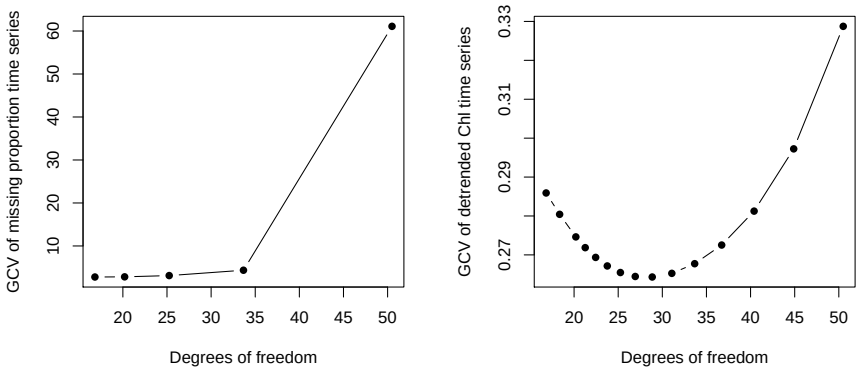

Figure 10: (Left) The double cross validation score of the 535 smoothed missing proportion time series against different degrees of freedom. (Right) The GCV scores of the 535 smoothed log(Chl-a) time series after removing the temporal trend against different degrees of freedom.