Globolakes; Functional clustering and coherence of Lake Water Quality

R. A. Haggarty (ruth.haggarty@glasgow.ac.uk)
C. A. Miller (claire.miller@glasgow.ac.uk)
E. M. Scott (marian.scott@glasgow.ac.uk)
School of Mathematics and Statistics, University of Glasgow,
15 University Gardens, Glasgow, G12 8QW

Abstract

Lakes are considered as sensitive indicators of environmental change which are impacted by both natural and anthropogenic drivers. The potential impact of climate change on freshwater resources is critical, and improved understanding of the observed changes is key to ensure better management of aquatic resources. While previous studies have often focussed on individual lakes, identifying synchronous temporal patterns observed across multiple lakes on a larger scale may indicate the existence of global common drivers and pressures. The GloboLakes project includes statistical analysis of remotely sensed data for lakes across the world in order to investigate how lake water quality responds to environmental change at a global scale. The aim of this paper is to investigate different clustering approaches for functional data applied to Lake Surface Water Temperature data. The clustering will be used to explore temporal coherence of multiple time series, with a view to establishing ecologically valid groups of lakes which are similar in terms of observed trends and seasonal patterns.

Key Words: functional, clustering, principal component analysis, water quality

1. Introduction

Freshwater ecosystems are vital components of the global biosphere. However, these environments are vulnerable to the forces of climate and human induced change. Consequently it is of great importance to monitor and assess the impact and extent of any such changes. In recent years, developments in Earth Observation systems, and the increased capability to retrieve in-water constituents, has resulted in the availability of exciting new data sets (Xie et al., 2008, Matthews, 2011). These expansive spatiotemporal data sets simultaneously enable global assessment of environmental changes and present new statistical challenges.

GloboLakes (www.globolakes.ac.uk) is a 5-year Natural Environment Research Council (NERC) consortium project involving 6 UK research groups whose goal is to investigate the state of 1000 lakes and their response to climatic and other environmental drivers of change at a global scale using a 20-year archive of satellite based observations. The synchrony between major fluctuations in a set of time series is often described as temporal coherence (Salisbury et al., 2011). Two of the key aims of the GloboLakes project are to identify patterns of temporal coherence for individual remotely sensed lake characteristics, and to explore the spatial extent of coherence for the lakes.

The aim of this paper is to compare different clustering approaches to investigate the temporal coherence of Lake Surface Water Temperature (LSWT) for a large number (675) of lakes distributed globally, and to obtain clusters of lakes based on common features across time.

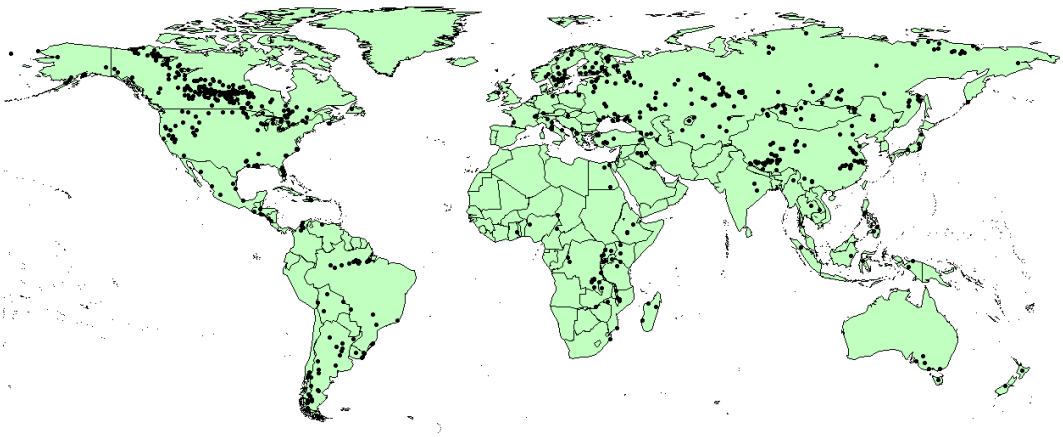

Figure 1: Global map with lake locations marked by a dot.

1.1 Data

The European Space Agency funded ARC-Lake project (MacCallum and Merchant, 2012) has employed the use of the Along Track Scanning Radiometers (ATSRs) instrument on-board the Envisat satellite in order to derive observations of LSWT for a large number of lakes across the globe. The data considered in this paper are a subset of the ARC-Lake version 3 data-set (see www.geos.ed.ac.uk/arclake/data for details). The data-set we have considered is comprised of bi-monthly lake average LSWTs for 675 lakes covering an 18-year period from 1995 to 2012, providing 405 observations for each of the lakes. All of the 675 lakes selected for this study are also being investigated as part of the GloboLakes project. Figure 1 shows a map with the location of each of the lakes considered in this paper.

The LSWT data products have been derived from spatially and temporally complete reconstructions of the ARC-Lake LSWT products. Spatially complete reconstructions have been obtained from observations using techniques based on interpolating empirical orthogonal functions. Full details of the interpolation of the spatial images are provided in MacCallum and Merchant (2010).

2. Methods

Our aim is to group the LSWT time series corresponding to the lakes into clusters, where two time series belong to the same cluster if they are coherent with each other. A functional data analysis approach has been taken. Regarding the data in this functional setting enables any long-term temporal and seasonal patterns in the data to be estimated and then compared across lakes. Several applications of functional data analysis to environmental data are available in the literature (Damon and Guillas, 2002 and Estvez-Prez and Vilar, 2013). More specifically, functional clustering techniques have been applied to environmental data in Pastres et al. (2011) and Haggarty et al. (2012), who consider the application of model based functional clustering to water quality data and Garcia-Escudero

and Gordaliza (2005) and Ignaccolo et al. (2008), who explore variations on k-means based clustering for air quality network data.

As the number and structure of any clusters in the data is unknown, several different functional clustering approaches are considered. This enables us to investigate the underlying structure in the data identified by each of the approaches and to examine the robustness of the results.

2.1 Functional Data Analysis

In functional data analysis, the observed time series are viewed as potentially noisy realisations of unobserved functions. To estimate each of these functions, the time series are described in terms of linear combinations of basis functions. For example, each LSWT time series can be expressed as;

$$y_i(t) = G_i(t) + \varepsilon_i(t) \tag{1}$$

where G_i is a smooth curve and ε_i is a normally distributed independent random error term ($i = 1, .., n$). The curve G_i is a spline function of degree d which can be expressed as a linear combination of B-splines, written in the following functional form for the spline $s_i\left(t; \beta_i\right)$

$$s_i\left(t; \beta_i\right) = \sum_{l=1}^{K+d-1} \beta_{i,l} B_l\left(t\right) \tag{2}$$

where $\beta_i = \left(\beta_{i,1}, ..., \beta_{i,K+d-1}\right)'$ is a vector of real-valued coefficients, $(B_1(t), ..., B_{K+d-1}(t))$ are the B-spline basis functions and K is the number of knots.

As detailed in Ignaccolo et al. (2008), the β_i vector is estimated by least squares and the G_i curve is approximated by $\hat{G}_i\left(t\right) = s_i\left(t; \hat{\beta}_i\right)$. If the polynomial of degree d, the number of knots K and the knot positions are the same for all the time series, then the B-spline basis functions are fixed and the spline coefficients β_i describe the same features for each of the time series.

Three approaches to obtain coherent clusters based on functional data are considered here; model based clustering of functional principal component scores, the k-means algorithm and a hierarchical clustering algorithm.

2.2 Functional Principal Components Analysis

One approach to find clusters based on functional data is to first decompose the variation in the data by applying a functional principal components analysis (FPCA) and subsequently, to cluster the corresponding principal component scores. The application of FPCA enables the dimension of the functional data to be substantially reduced and hence provides a very computationally efficient way of exploring any underlying structure in the data.

As in the standard setting, FPCA can be used to identify the dominant modes of variation in a data set. In the functional case, both the data and the estimated functional principal components (FPCs) are curves. The FPCs can be thought of as a set of orthogonal basis functions constructed so as to account for as much variation as possible.

In the non-functional data setting, for a centered data matrix X of dimension $n \times p$, where the n rows represent observations and the p columns represent variables, the application of PCA yields an orthogonal decomposition of X that is optimal for a given number of principal components. Singular value decomposition of X can be used to decompose the matrix and obtain principal component scores (Jolliffe, 2002). Furthermore, Varimax rotation of the principal components can also be applied to make them more interpretable (Ramsay and Silverman, 1997). To extend PCA to the functional setting each $1 \times p$ vectors representing an individual are replaced by sets of basis coefficients which define functions. Hence, matrices are replaced by compact linear operators and covariance matrices by covariance operators. Full details of functional principal components are provided in Ramsay and Silverman (1997).

After deciding on the appropriate number of functional principal components to retain, clustering approaches can be applied to the corresponding set of principal component scores. Model based clustering approaches can then be applied to the principal component scores in order to identify groups of observations which are coherent in terms of their temporal variation. The model based clustering method used in this paper is described in (Fraley and Raftery, 1998).

2.3 Functional K-means

The k-means algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Functional clustering based on the k-means algorithm has been introduced in Abraham et al. (2003) and applied in Ignaccolo et al. (2008). In the functional case the algorithm is applied to the spline coefficient vectors which define the curves representing the time series. Following this, the clustering result directly provides the clustering of the time series. For a given number of clusters, in order to reduce the influence of the starting values, the k-means algorithm is applied several times.

2.4 Functional Hierarchical Clustering

To apply hierarchical clustering to a set of points, a distance matrix, D, is first calculated which contains the distance between all possible pairs of points. The i, j^{th} entry of D is the distance between points i and j as determined by whichever metric has been chosen. Henderson (2006) states that the idea of measuring distances is easily transferable from pairs of points to pairs of curves and defines a method of computing a functional distance matrix as follows. The distance between two curves can be written as

$$d_{ij} = (\beta_i - \beta_j)^T W (\beta_i - \beta_j) \tag{3}$$

In the above expression, W is a symmetric matrix the elements of which are given by $w_{l,m} = \int B_l(t) B_m(t)' dt$, with $l, m = 1, ..., K + d - 1$. For each set of basis functions, W can be evaluated using numerical integration, if necessary, and the functional distance matrix D with entries d_{ij} can be computed. Standard algorithms for hierarchical clustering can then be applied to the elements of D.

2.5 Stopping criteria

Well developed methods exist to choose the optimal number of clusters. For k-means and hierarchical clustering the gap statistic (Tibshirani et al., 2001) approach has been con-

sidered here. The gap statistic uses the within cluster dispersion to determine the statistically optimal number of clusters and compares the average within cluster dispersion for the observed data, to the average within cluster dispersion for a null reference distribution which assumes there is no underlying clustering structure. The clustering based on FPC scores is model based and so model selection criteria such as BIC can be used to determine the best model. After calculating BIC for models with different numbers of mixture components corresponding to different numbers of clusters, the model which minimises BIC is selected.

While both approaches are data-driven and can be used to select the statistically optimal number of clusters for the different clustering methods applied here, they can be computationally expensive due to the repeated calculations required, and simulations in the case of the Gap statistic.

3. Results

To estimate the curves representing the average lake surface water temperature for each individual lake a cubic B-spline basis function was used. The raw time series were regular and reasonably smooth to start with, and so no additional smoothing penalty term was applied. When fitting the curves, a knot was placed at every 6 observations in order to correspond to intervals of duration 3 months, or one 'season'. Consequently, 68 basis functions were used to estimate each of the smooth curves representing each lake. This number of knots appeared to result in curves which provided a sensible fit to the data and neither resulted in a fit which was deemed to be too smooth, nor too locally variable.

Following estimation of the functional data, all three clustering approaches were applied. The gap statistic suggested that 9 clusters was statistically optimal for the hierarchical clustering of the curves while 13 clusters was identified as being best when K-means clustering was applied.

For the functional PCA approach, the first three principal components were found to account for 99% of variability in the curves, and hence scores corresponding to the first three principal components were retained. Figure 2 shows the mean-centered functional mean curve (black solid line) plus and minus the rotated functional principal components (green and blue lines). It is clear that the rotated fPCs have meaningful interpretations; component 1, accounts for the most variability in the curves (60.7%) and corresponds to the difference between the timing of the seasonal patterns of lakes located in the Northern and Southern hemispheres whilst components 2 and 3 account for a similar quantity of variability in the curves (18.5 and 20%) and correspond to deviations in the spring-summer months, and deviations in summer-autumn months respectively. Using mixture model based clustering via the mclust package in R (Fraley et al., 2012) and BIC to select the optimal number of clusters, 10 clusters were identified as being most appropriate.

The results are summarised in Figures 3, 4 and 5 which show maps with lake locations colored to represent the different clusters. The different sizes of each of the clusters obtained (in terms of number of lakes) using the three different approaches are shown in Table 1. Due to the different number of clusters and number of curves in each cluster the cluster numbers are nominal labels only.

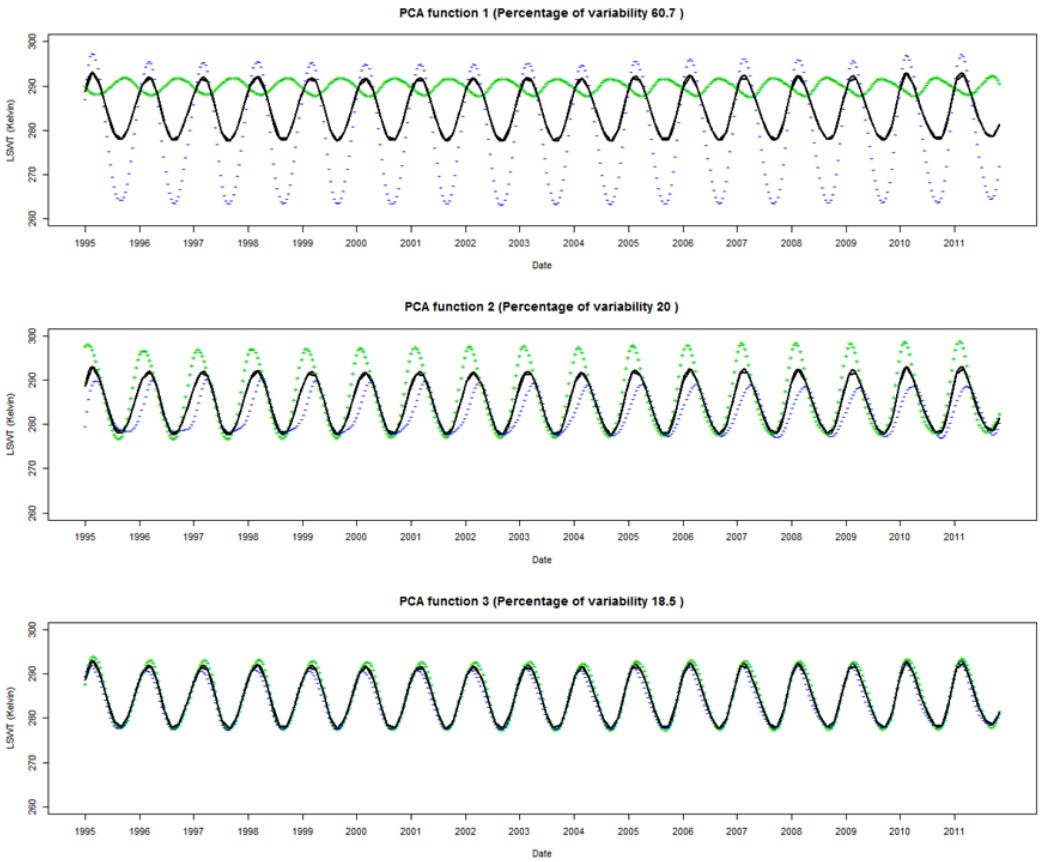

Figure 2: Functional mean (black solid line) plus and minus the rotated functional principal components (green/blue lines)

3.1 Discussion

Although there is no single number of clusters identified as being statistically optimal, there are undoubtedly broad similarities between the results of all of the different clustering methods investigated here. It is clear from the maps that in all approaches considered, as could be expected, the key distinction between the clusters is the latitude of the lakes. For all three approaches, the Northern hemisphere and Equatorial band clusters are defined primarily by relatively smaller scale differences in the phase and amplitude of the seasonal patterns than those clusters of primarily Southern hemisphere lakes.

While the number of clusters identified as statistically optimal is most different for the K-means and hierarchical approaches, these partitions share many similar features, such as the distribution of clusters in the Southern hemisphere. In fact, hierarchical clusters 6 and 8 are identical to K-means clusters 9 and 12. One of the main distinctions between the K-means and hierarchical clustering results is the difference in the number of clusters describing the variability in the Northern hemisphere. Hierarchical clustering suggests that fewer, larger clusters are suitable to describe the variability in the LSWT here. On the other hand, K-means clustering identifies smaller clusters that are separated by more subtle differences between the lakes. The FPCA clusters also share similarities with the other approaches, for example, all lakes in FPCA clusters 6 and 10 are contained within hierarchical

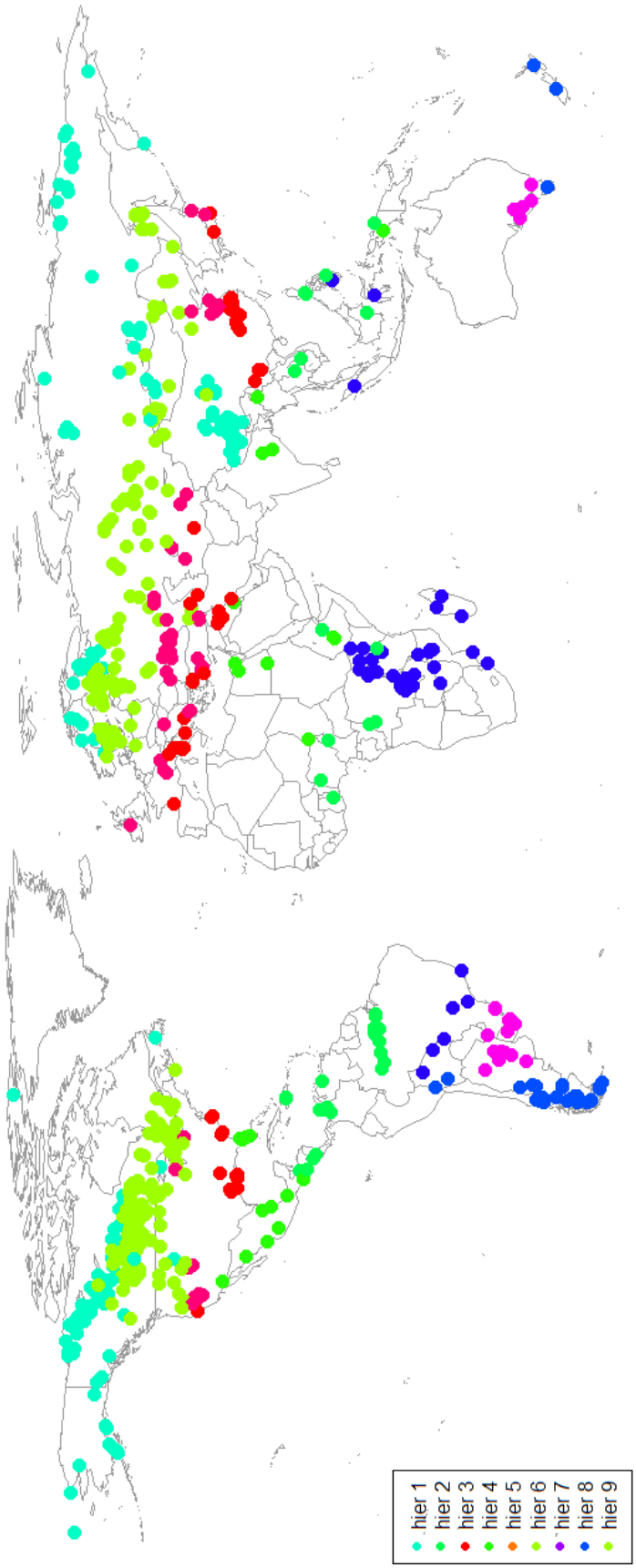

Figure 3: Map showing geographical distributions of clusters determined by hierarchical functional clustering

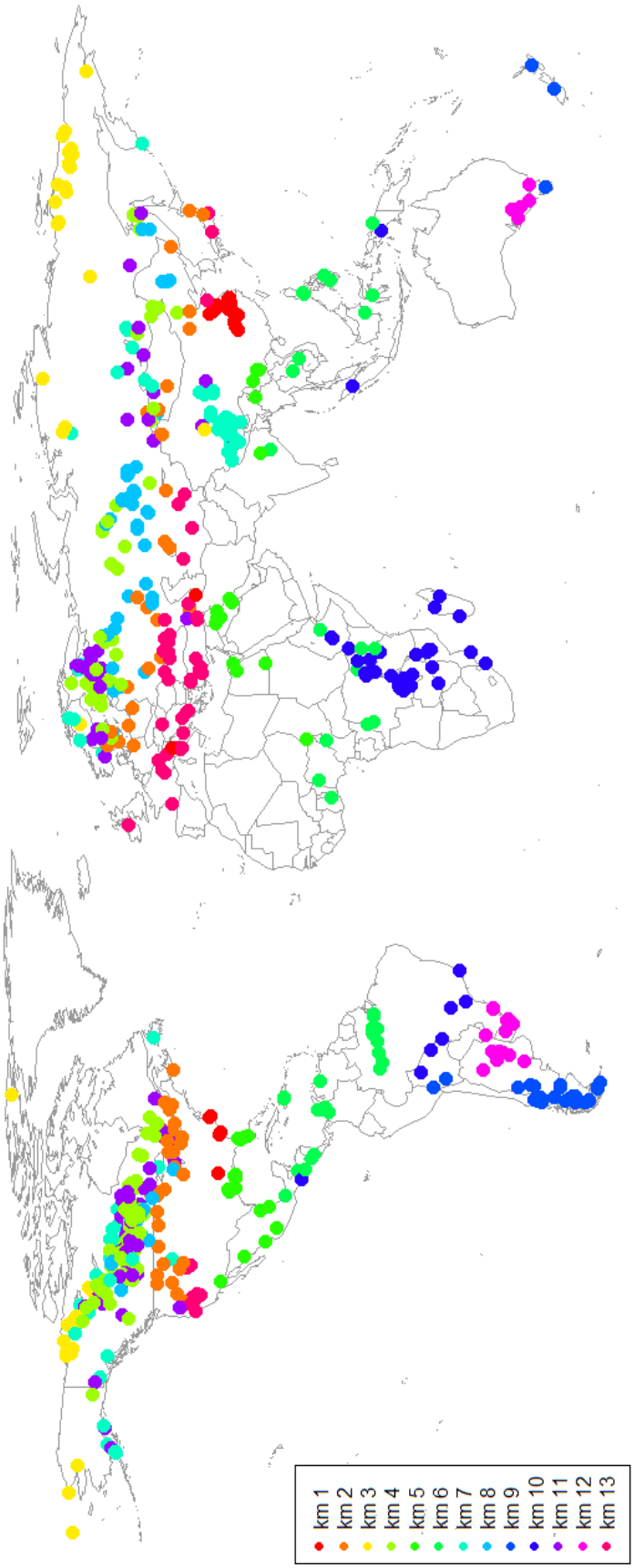

Figure 4: Map showing geographical distributions of clusters determined by K-means functional clustering

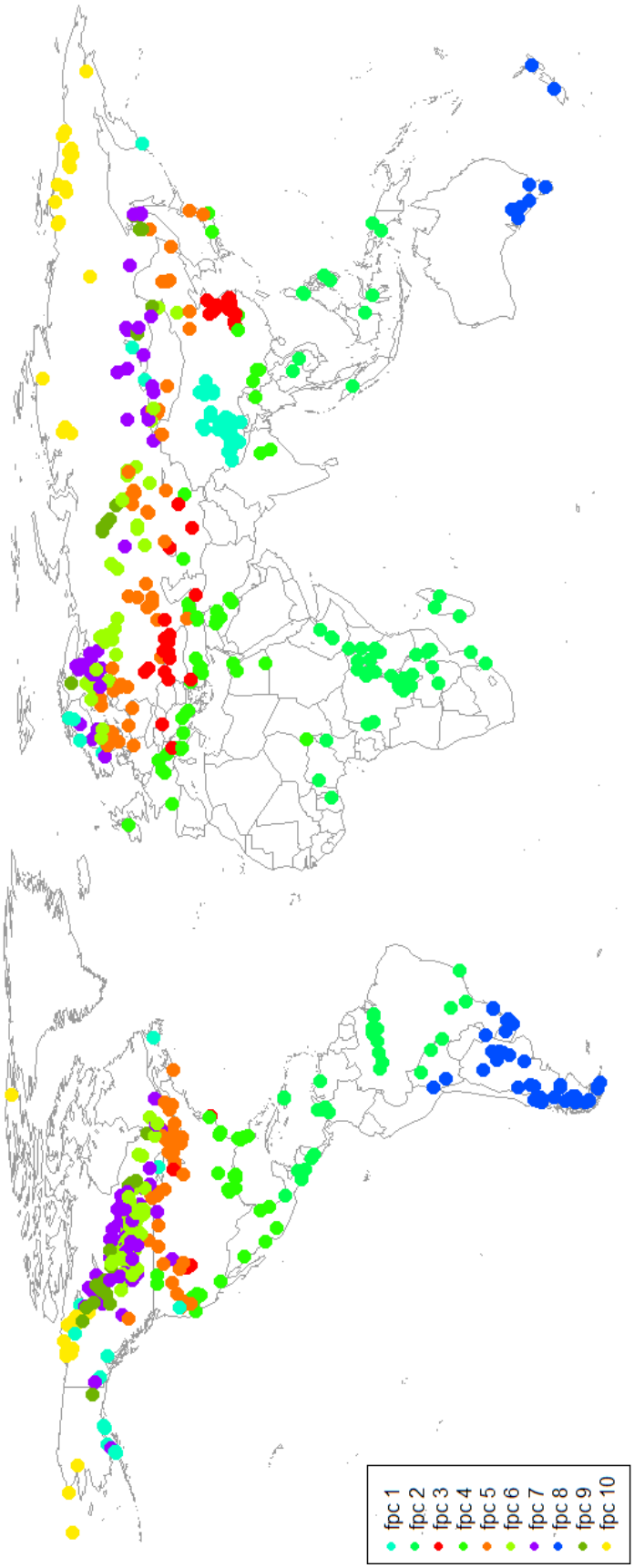

Figure 5: Map showing geographical distributions of clusters determined by functional principal component score based clustering

Table 1: Number of time series/curves in each cluster obtained using the three approaches

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13
Hierarchical	152	40	47	272	24	29	40	20	51	-	-	-	-
K-means	25	60	43	106	33	46	68	55	29	40	100	20	50
FPCA	56	85	41	73	92	103	102	49	32	42	-	-	-

clusters 4 and 1, respectively.

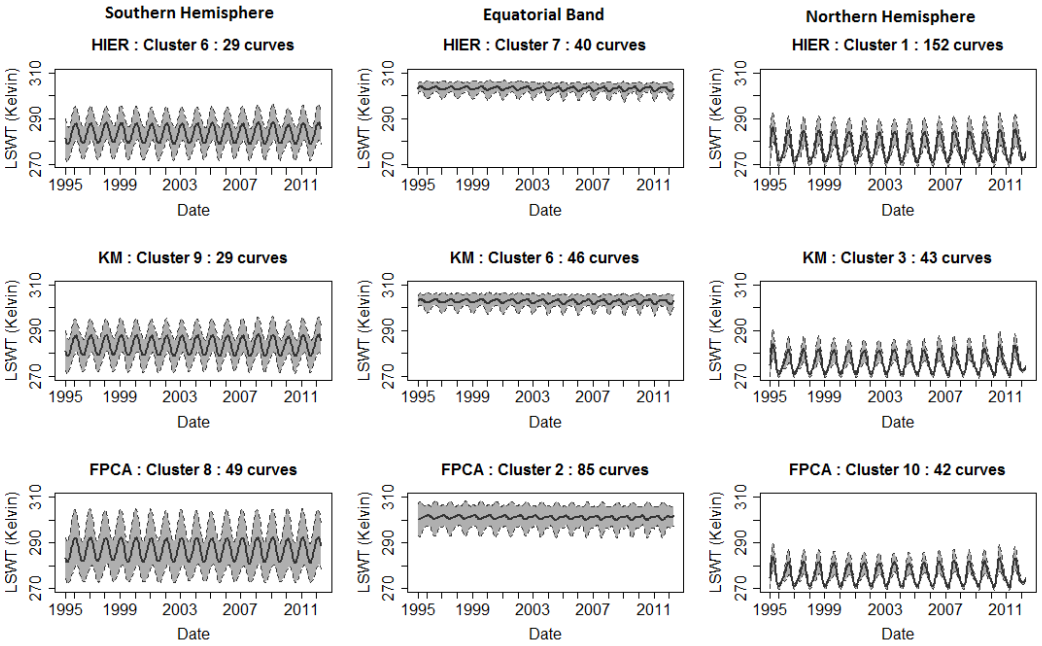

Figure 6: Comparison of a subset of clusters across the different approaches considered. Solid lines representing cluster means are shown for a Southern hemisphere cluster (left hand side), an Equatorial band cluster (middle) a Northern hemisphere cluster (right hand side). The shaded bands represent the uncertainty associated with the cluster means

Figure 6 shows a comparison between a subset of the clusters across the three methods. The left hand column of Figure 6 shows cluster means for a Southern hemisphere cluster, the middle column of plots shows the means for an Equatorial band cluster and the right hand plots a Northern hemisphere cluster. The shaded bands show the cluster mean plus and minus twice the functional standard deviation of the curves within that cluster. Despite the different cluster sizes, it can be seen that the broad patterns being identified are similar across the three approaches. However, there are some distinctions between the clusters obtained via each of the approaches. For example, both the k-means and hierarchical clustering methods identify one more cluster in the Southern hemisphere than the clustering based on principal component scores.

In order to attempt to quantify how similar the results of each method were the Adjusted Rand Index (Hubert and Arabie, 1985) was computed for all pairs of clustering approaches.

The Adjusted Rand Index (ARI) is an index which is based upon counting the pairs of points on which two clusterings agree or disagree and is corrected for the possibility that agreement between two sets of clusters may simply be due to chance. The maximum value of the ARI is one, which corresponds to perfect agreement between two partitions. Conversely, if the ARI is zero, the two partitions are mutually independent.

The greatest degree of agreement is between the partitions based on FPCA and K-means (ARI=0.49) with marginally poorer agreement between the clusters of lakes determined using the K-means and Hierarchical approaches (ARI=0.35) and the FPCA based and Hierarchical approaches (ARI=0.33). The moderate size of all of the ARI values computed may be surprising in light of the broad similarities between the clusters shown in Figures 3, 4 and 5. This indicates that, while informative to some degree, the ARI should be viewed with caution and should not be considered without also comparing the distribution of the partitions via either cluster mean curves or geographical maps.

4. Conclusions

It has been demonstrated with the ARC-lake data that all three of the approaches considered here are robust and computationally efficient for large numbers of time series of potentially noisy data. The use of functional data enables the data dimension to be reduced substantially. While the temperature time series are already reasonably smooth, creating functional data objects via smoothing can be very useful when highly noisy time series are required to be clustered. Initially, smoothing the time series before clustering enables the key features of interest in the data to be retained while, excluding local variability which, if included, may result in an overestimation of the the number of clusters required to describe the variability in the curves.

The clustering approaches enable clusters of curves to be identified which are coherent in terms of temporal dynamics. As there is no single correct answer when clustering data such as global lake temperature is necessary to consider different clustering approaches and compare results. It is very reassuring, that although not identical, the approaches considered, have produced results which are consistent with each other and which make sense when interpreted in an ecological context.

The analysis presented here gives us an overview of global coherence, however there may be distinct basins within lake which behave very differently to one another and so our future work will also focus on within lake coherence. The computational efficiency of FPCA could be an attractive feature for this.

Acknowledgements

Haggarty, Scott and Miller were funded for this work through the NERC GloboLakes project (NE/J022810/1). The authors gratefully acknowledge the ARC lake project for access to the data.

References

Abraham, C., P. A. Cornillion, E. Matzner-Lober, and N. Molinari (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 30*, 581–595.

Damon, J. and S. Guillas (2002). The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics 13*(7), 759–774.

Estvez-Prez, G. and J. Vilar (2013). Functional anova starting from discrete data: an application to air quality data. *Environmental and Ecological Statistics 20*(3), 495–517.

Fraley, C. and A. E. Raftery (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal 41*(8), 578–588.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (2012). mclust version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification. Technical Report 597, Department of Statistics, University of Washington.

Garcia-Escudero, L. A. and A. Gordaliza (2005). A proposal for robust curve clustering. *Journal of Classification 22*(2), 185–201.

Haggarty, R., C. Miller, E. Scott, F. Wyllie, and M. Smith (2012). Functional clustering of water quality data in scotland. *Environmetrics 23*(8), 685–695.

Henderson, B. (2006). Exploring between site difference in water quality trends: a functional data analysis approach. *Environmentrics 17*, 65–80.

Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Ignaccolo, R., S. Ghigo, and E. Giovenali (2008). Analysis of air quality monitoring networks by functional clustering. *Environmetrics 19*(7), 672–686.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer, New York.

MacCallum, S. N. and C. J. Merchant (2010). Arc-lake algorithm theoretical basis document. Technical report, School of GeoSciences, The University of Edinburgh.

MacCallum, S. N. and C. J. Merchant (2012). Surface water temperature observations of large lakes by optimal estimation. *Canadian Journal of Remote Sensing 38*(1), 25–45.

Matthews, M. W. (2011). A current review of empirical procedures of remote sensing in inland and near-coastal transitional waters. *International Journal of Remote Sensing 32*(21), 6855–6899.

Pastres, R., A. Pastore, and S. F. Tonellato (2011). Looking for similar patterns among monitoring stations. venice lagoon application. *Environmetrics 22*(6), 712–724.

Ramsay, J. and B. W. Silverman (1997). *Functional Data Analysis (Springer Series in Statistics)* (1st ed.). Springer.

Salisbury, J., D. Vandemark, J. Campbell, C. Hunt, D. Wisser, N. Reul, and B. Chapron (2011). Spatial and temporal coherence between amazon river discharge, salinity, and light absorption by colored organic carbon in western tropical atlantic surface waters. *Journal of Geophysical Research: Oceans 116*(C7).

Tibshirani, R., G. Walther, and T. Hastie (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), pp. 411–423.

Xie, Y., Z. Sha, and M. Yu (2008). Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology 1*(1), 9–23.