Globolakes
WP5: Detecting spatial and temporal patterns

January 2014

GloboLakes

Global Observatory of Lake Responses to Environmental Change

CEH Centre for Ecology & Hydrology NATURAL ENVIRONMENT RESEARCH COUNCIL

DUNDEE

University of Reading

University of Glasgow

PML Plymouth Marine Laboratory

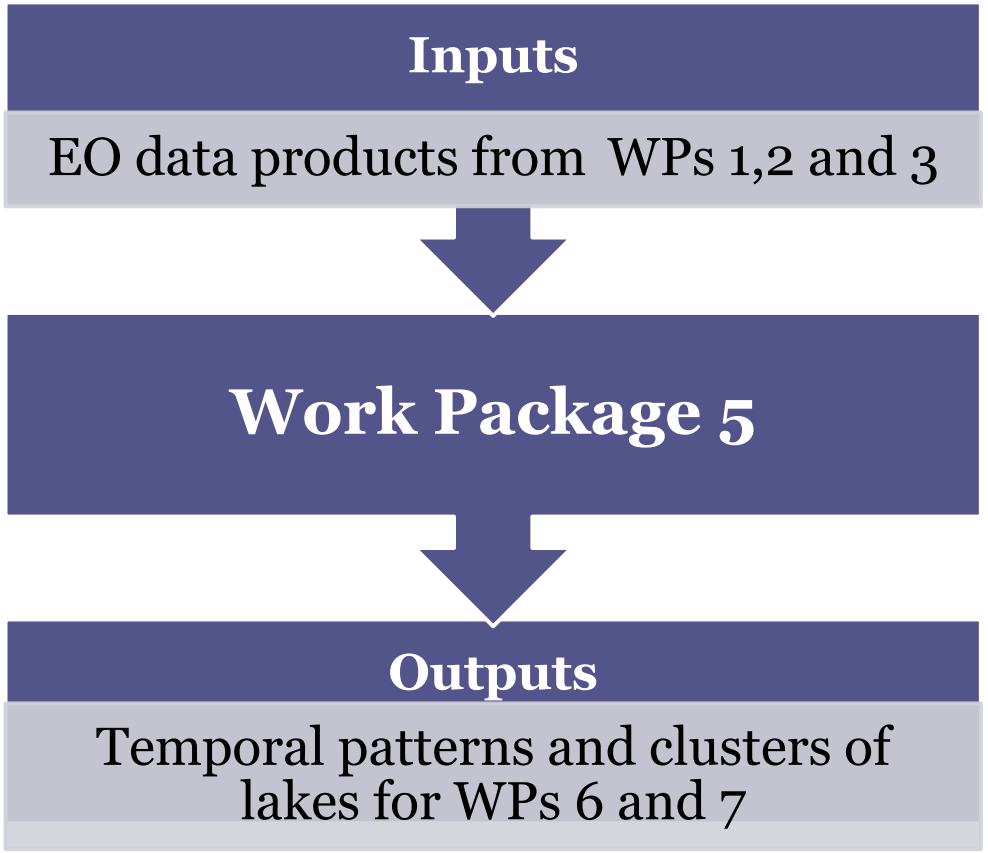
UNIVERSITY OF STIRLING

Aim:

To assess the extent of temporal coherence for individual remotely-sensed lake characteristics & to define the nature of any clusters of coherent lakes.

Contributors:

University of Glasgow
Centre for Ecology & Hydrology



Objectives

5.1 Assess the present state & evidence for long-term change in the 1000 lakes.

5.2 Identify patterns of temporal coherence for individual remotely sensed lake characteristics & the spatial extent of coherence.

5.3 Identify phenological patterns of change in remotely sensed lake characteristics.

Globolakes
WP5: Lake Coherence

Ruth Haggarty, Claire Miller, Marian Scott, Francesco Finazzi

January 2014

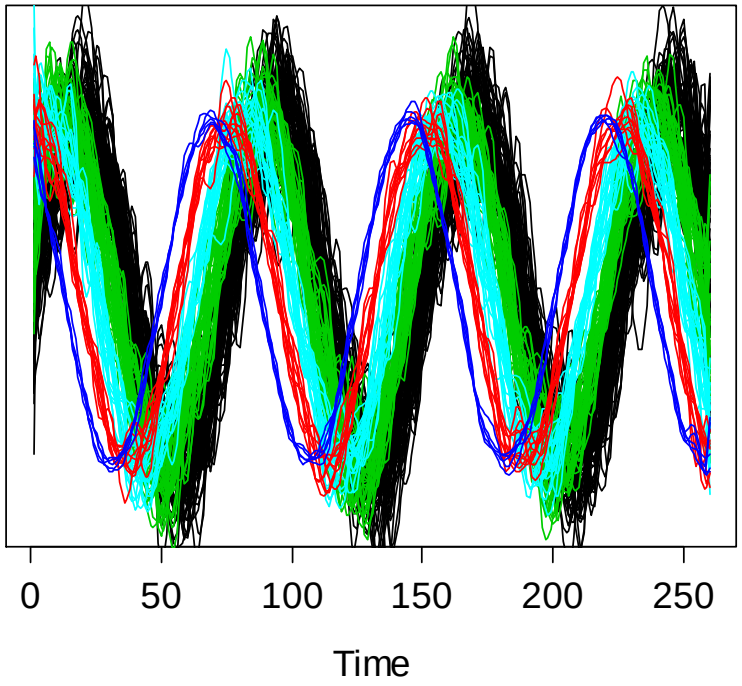
GloboLakes

Global Observatory of Lake Responses to Environmental Change

CEH Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL
DUNDEE
University of Reading
University of Glasgow
PML Plymouth Marine Laboratory
UNIVERSITY OF STIRLING

Coherence

- The synchrony between major fluctuations in a set of time series is often described as **temporal coherence**.
- Group the time series into a suitable number of clusters where two time series belong to the same cluster if they are coherent with each other.
- We want to focus on comparing a large number of time series and aim to obtain clusters based on common trends, seasonal patterns and other features across time.



Statistical Approaches

- We are investigating (and developing) two different statistical approaches:
  - State space model
  - Functional data analysis
- Both approaches can be applied to (potentially) thousands of time series.
- The main difference is that one approach is based on the raw data and the other uses smoothing.
- These techniques have been applied to a set of data from the ARC-lake project which will be presented later.

State Space Approach

Each individual lake time series (*y*) is represented in terms of an underlying (latent) time series *(z)*.

$$\mathbf{y}(t) = \mathbf{K}\mathbf{z}(t) + \mathbf{e}(t)$$
$$\mathbf{z}(t) = \mathbf{G}\mathbf{z}(t-1) + \mathbf{h}(t)$$

i.e. each individual lake time series is clustered into one and only one cluster, with error vectors **e** and **h**.

This is done on the basis of the temporal pattern in the time series and the model is fitted using the EM algorithm.

This is a modification to a class of models known as Dynamic Factor Analysis and is based on initial work by Finazzi (Fassò and Finazzi, 2011).

Functional Data Analysis

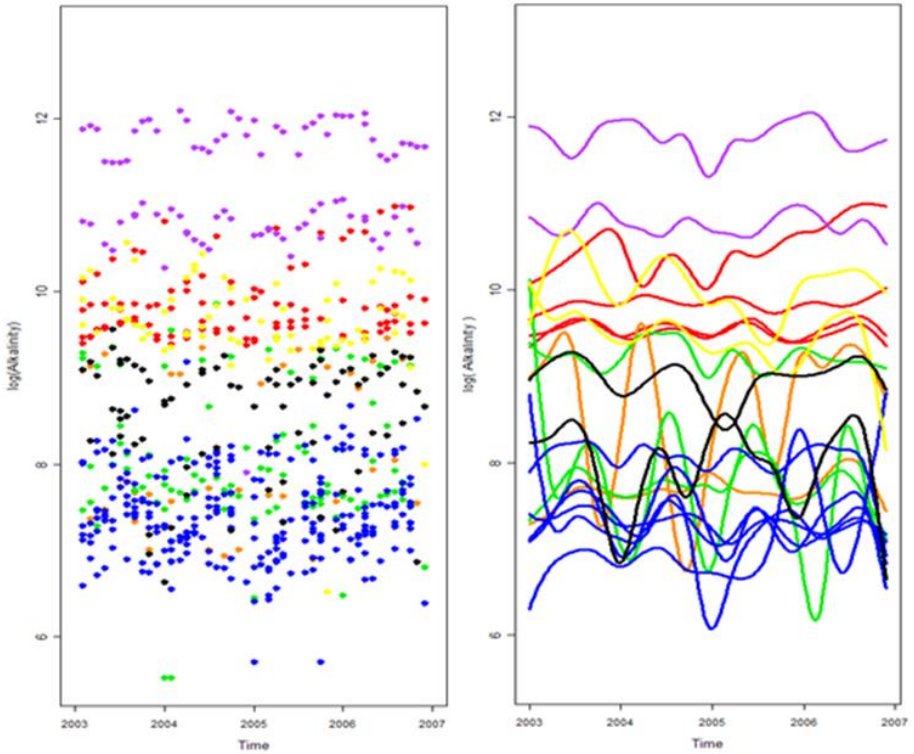
Each observed time series can be expressed as

$$y_i(t) = G_i(t) + e_i(t)$$

where G_i is a smooth curve and e_i is an independent random error term, $i=1, \ldots, N$ lakes.

The curve G_i is a spline function of degree d which can be expressed as a linear combination of B-splines,

$$\hat{G}_i(t) = \sum_{l=1}^{K+d-1} \beta_{i,l} B_l(t)$$



Functional Clustering

Hierarchical

- The distance between the pairs of curves $G_i(t)$ and $G_j(t)$, $i,j = 1,\ldots,N$ is first estimated as

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

where W is the symmetric matrix based on basis vectors.

- Standard algorithms for hierarchical clustering can then be applied to the elements of the functional distance matrix D with entries d_{ij}.

K-means

An iterative partitioning procedure where the number of groups is first specified, and then objects are moved from group to group, until the within-group sums of squares is minimised.

Determining the optimal number of clusters

State space

Two approaches have been investigated and proposed:

The number of clusters is increased (from 1) until the observed data log-likelihood stabilizes and/or an additional cluster is empty.

Functional Data Analysis

Two standard approaches have been applied:

L-curve

Gap statistic

Both involve minimising the within cluster dispersion until it stabilises to determine the number of clusters.

Clustering the ARC-lake data

- Clustering approaches are applied to the LSWT time series of the ARC-Lake data set (www.geos.ed.ac.uk/arclake) in order to cluster the lakes into homogeneous groups with respect to their temporal coherence.

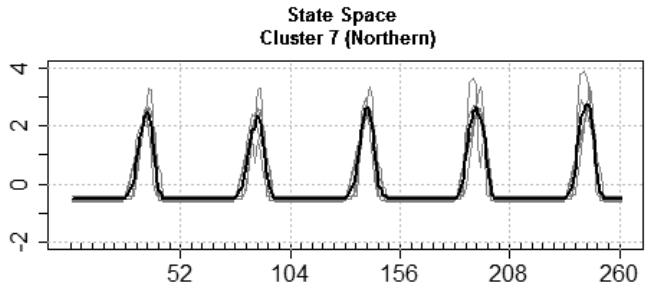
- 5 years of weekly mean values were used in the analysis (2006-2010) for 261 lakes.

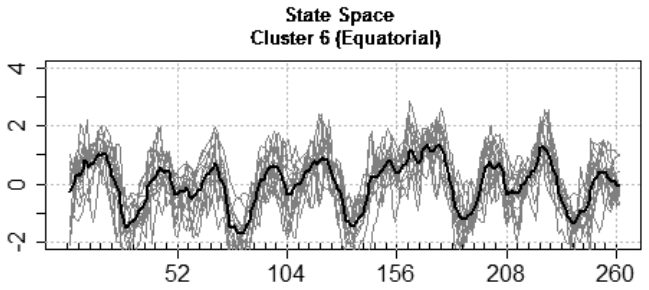
Clustering the ARC-lake data

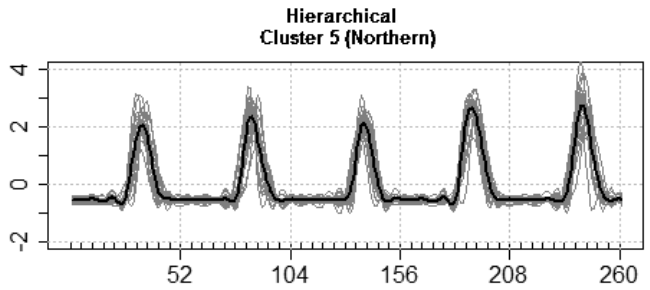
- **State space** and **hierarchical functional clustering** identified **11 clusters** as optimal,
- **k-means** identified **7 clusters** as being most appropriate.
- In general, the results for all three approaches were consistent however the state space model identified one cluster with a single time series.

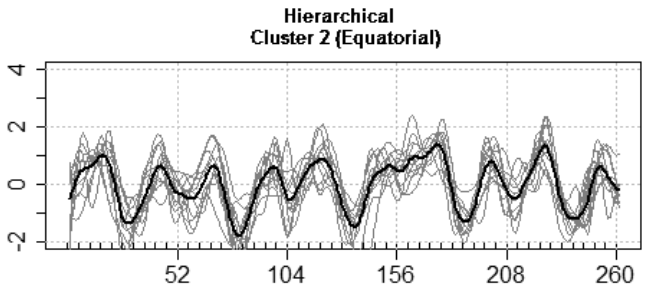
Comparing the clusters

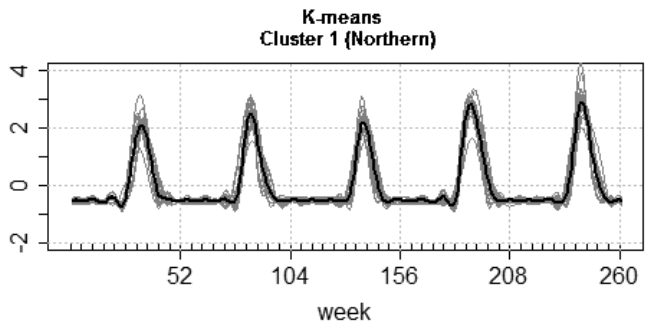
Each approach provides a different clustering result, however, the temporal patterns they identify are similar. Results for two clusters are shown.

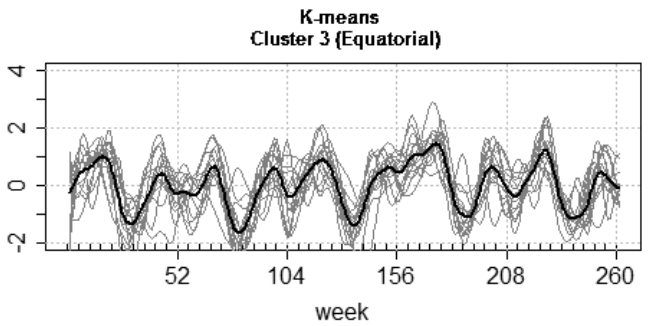






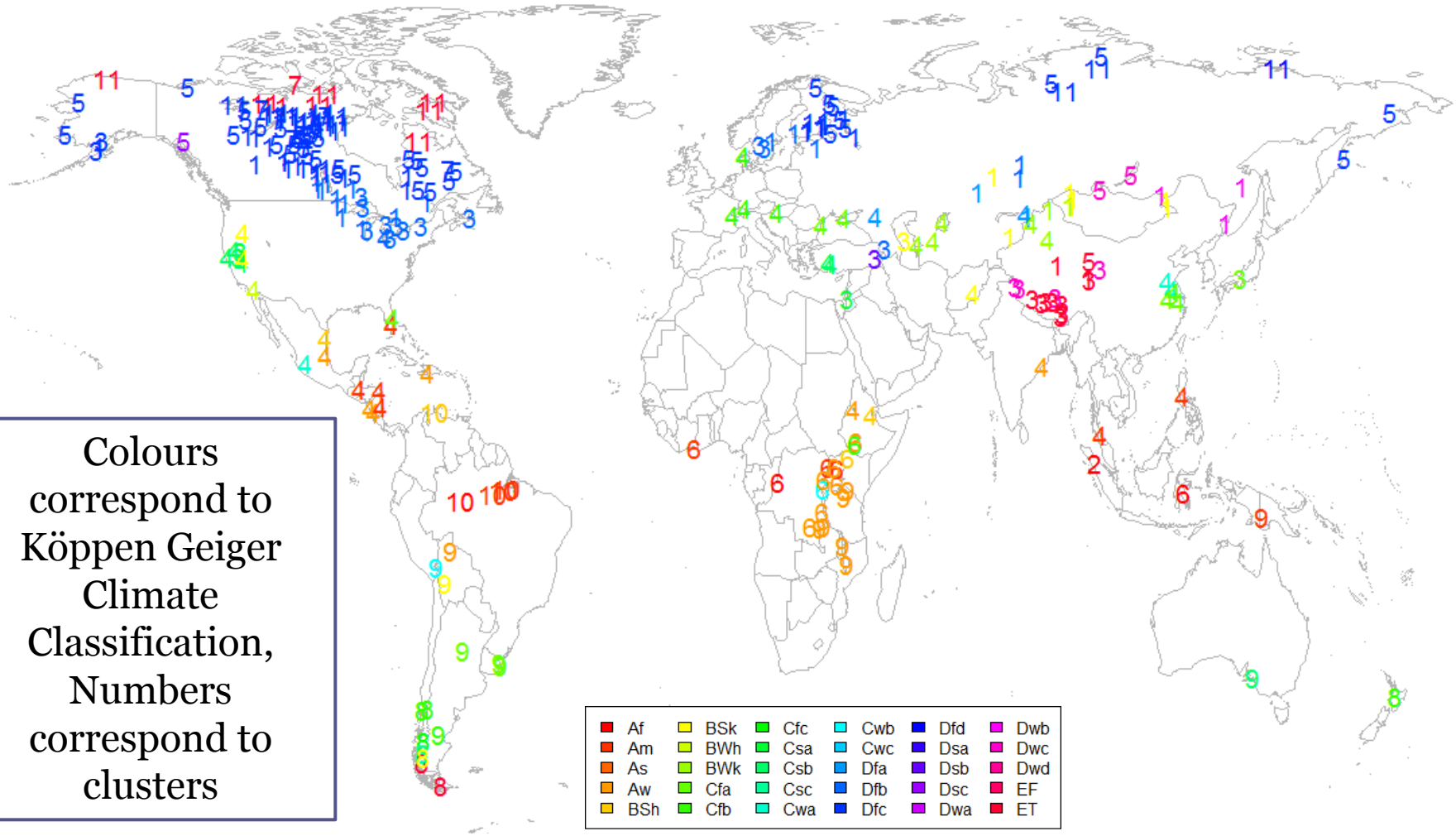






ARC-lake Clustering

State Space Model: 11 Clusters



Colours correspond to Köppen Geiger Climate Classification, Numbers correspond to clusters

Summary

- These approaches are suitable for large numbers of time series of potentially noisy data and enable clusters of curves to be identified which are coherent in terms of temporal dynamics.

- The approaches considered all, in general, produce results which are consistent with each other.

Summary

- The model-based approach does not require the observed time series to be smoothed and so the results obtained are not influenced by the degree of smoothing applied.

- However, smoothing can be useful when highly noisy time series are to be clustered, in which case the model-based approach might over-estimate the number of clusters.

References

- Abraham, C., Cornillon, P.A., Matzner-Lber, E., Molinari, N. Unsupervised curve clustering using b-splines. *Scandinavian Journal of Statistics* 30(3), 581–595 (2003).
- de Boor, C. *A Practical Guide to Splines*. No. 27 Applied Mathematical Sciences. Springer (2001).
- Fassò, A., Finazzi, F.: Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics* 22(6), 735–748 (2011).
- Finazzi, F., Haggarty, R., Miller, C., Scott, M., Fasso, A. A comparison of clustering approaches for the study of the temporal coherence of multiple time series, *Stochastic Environment Research and Risk Assessment* (Under Review, 2013) .
- Henderson, B. Exploring between site differences in water quality trends: a functional data analysis approach. *Environmetrics* 17(1), 65–80 (2006).
- MacCallum, S., Merchant, C.: Arc-lake v2.0, 1995-2011 [alidxxxx plrec9d ts366lm]. University of Edinburgh, School of GeoSciences / European
- Tibshirani, R., Walther, G., Hastie, T. 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 (2001).