State tagging for improved earth and environmental data quality assurance

Michael Tso (@Michael_ts0), Pete Henrys,
Sue Rennie, and John Watkins

8th May, 2020

European Geosciences Union (online)

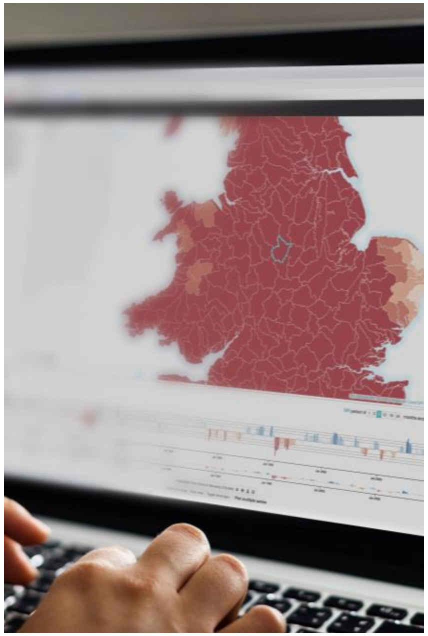
Manuscript just published in
Frontier in Environmental Science (2020)
DOI: 10.3389/fenvs.2020.00046

UK Centre for Ecology & Hydrology

@UK_CEH

Highlights

- A clustering-based state tagging framework is proposed to improve QA of environmental data
- Very efficient and applicable to virtually any type of point-based time series data
- Give greater confidence for users to use third-party data and encourage collaborative research
- Web applications available to explore the method



Environmental data in a big data age

- Long-term monitoring: (i.) form the foundation against which hypotheses can be formed and tested, (ii.) emerging trends determined and (iii.) future scenarios projected

- Environmental data explosion: more likely to use open/third party data to validate and compare observations, potentially from collaborative platforms in the cloud

- Data providers should not depend on users to verify the quality of datasets individually, but provide QA and QC information to assist this

- Can we provide a general tool to give users some idea about data quality?

Motivation

- Currently, static range check is the most common QC procedure for environmental data
- A generic and efficient machine learning tool to provide contextual information to produce out-of-range flags and understand variability of data
- The idea of "state" recognizes the acceptable or likely range of observed values depends on the state in which the system is in
- Goal: tag each observed value with an arbitrary calculated state number from contextual data and flag if out of the predicted state range (e.g. exceed mean +/- 2 std. dev.)

UKEOF ENVIRONMENTAL OBSERVATION FRAMEWORK

eLTER

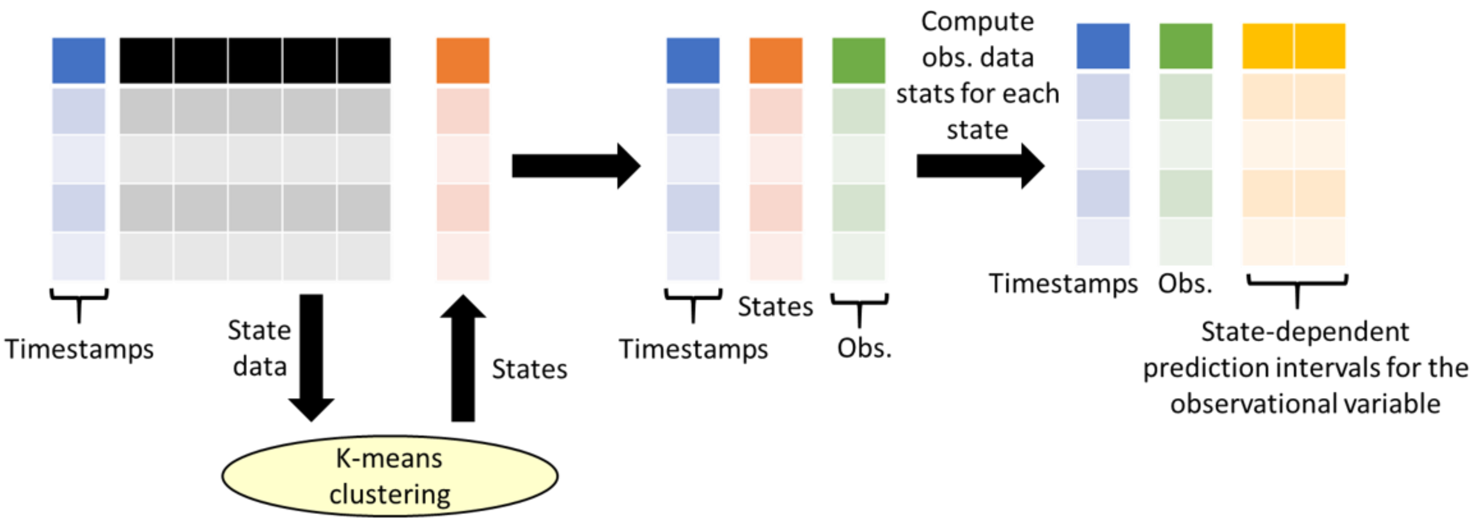
UK-SCAPE

UK Centre for Ecology & Hydrology

UK Environmental Change Network

State tagging: overview

Method overview



State tagging: the concept and design considerations

- Unsupervised and efficient: quick and flexible to implement to a large variety of datasets; labelled data may not be available

- A first-pass: Experts or users can interpret the state tagging results and conduct further analysis and quality checks using their subject-specific knowledge

- One state per data point: fuzzy methods are not suitable

- The definition of the identified states is purely statistical and is open to expert interpretation

Applications:
try these apps yourself now!

- Moth and butterfly data, UK Environmental Change Network (ECN), part of LTER-Europe

  https://statetag-ecnmoth.datalabs.ceh.ac.uk

- Lake chemistry data, UK Cumbrian Lakes Monitoring scheme

  https://statetag-lakes.datalabs.ceh.ac.uk

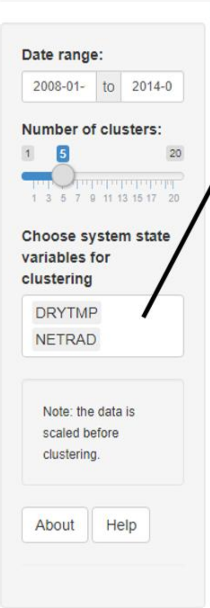
- A generic version: upload your own data (R Shiny source code included)

  https://statetag-generic.datalabs.ceh.ac.uk

1.

Change
if needed

2.



Date range:

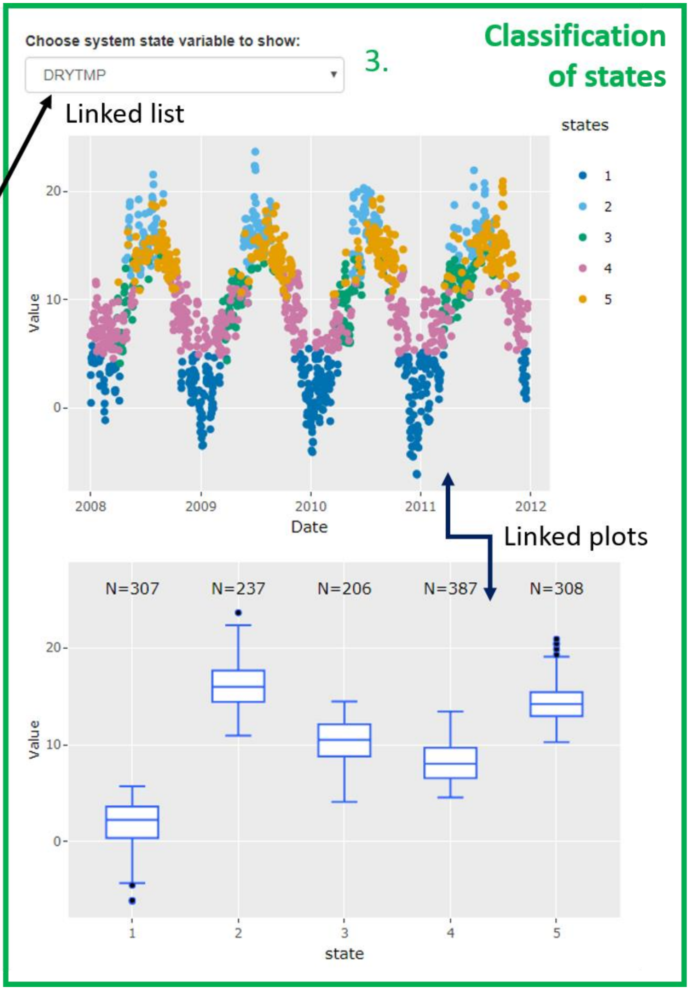
2008-01- to 2014-0

Number of clusters: 5

Choose system state variables for clustering

DRYTMP
NETRAD

Note: the data is scaled before clustering.



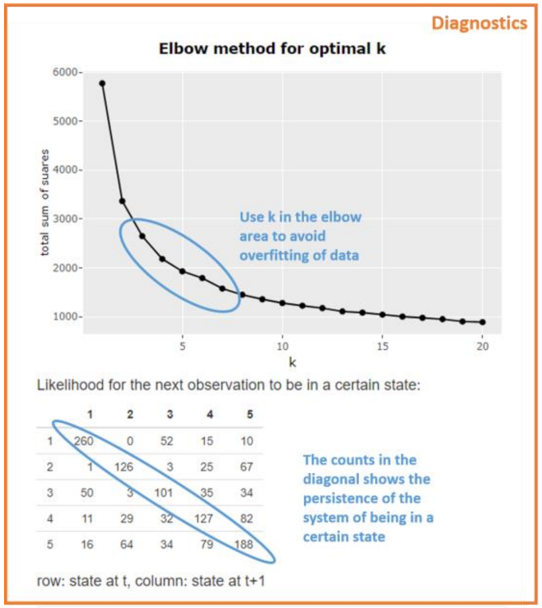
Classification of states

Choose system state variable to show: DRYTMP

3.

Linked list

Linked plots



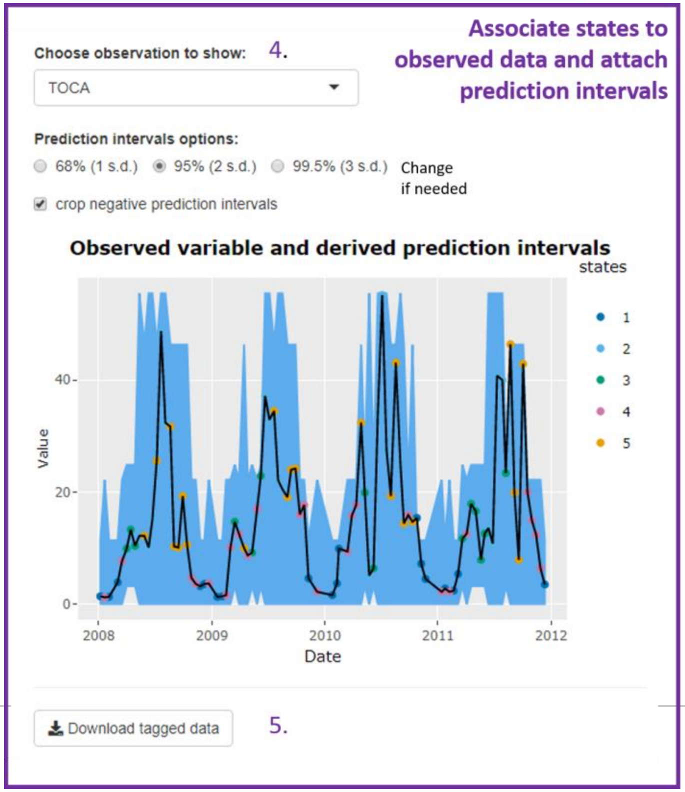
Diagnostics

Likelihood for the next observation to be in a certain state:

	1	2	3	4	5
1	260	0	52	15	10
2	1	126	3	25	67
3	50	3	101	35	34
4	11	29	32	127	82
5	16	64	34	79	188

The counts in the diagonal shows the persistence of the system of being in a certain state

row: state at t, column: state at t+1



Associate states to observed data and attach prediction intervals

Choose observation to show: TOCA

4.

Prediction intervals options:

○ 68% (1 s.d.) ● 95% (2 s.d.) ○ 99.5% (3 s.d.)

Change
if needed

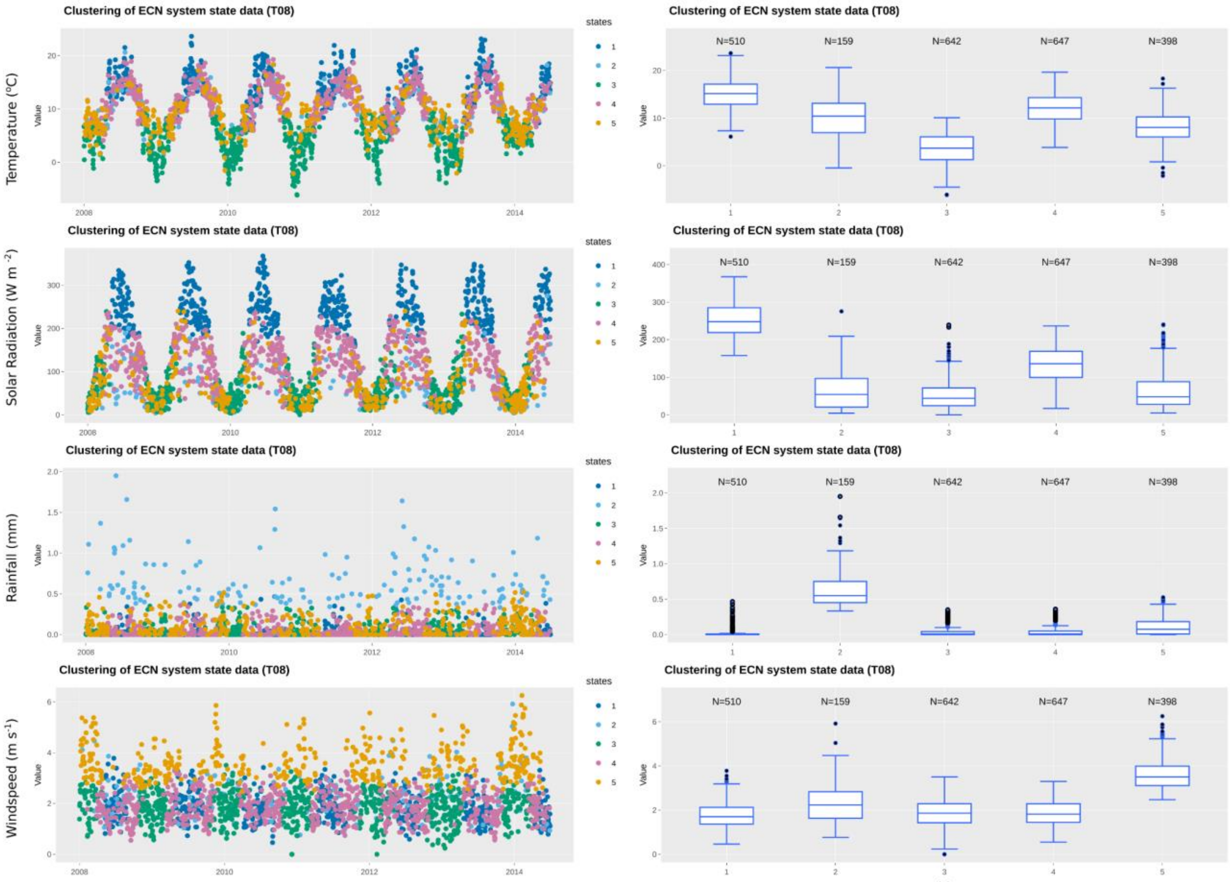
☑ crop negative prediction intervals

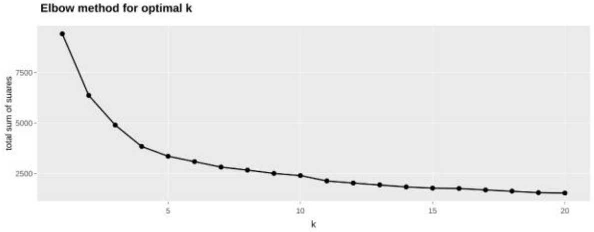
Observed variable and derived prediction intervals

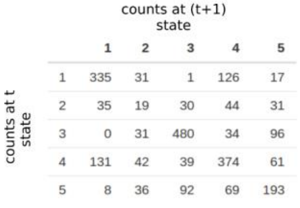
5.

UK Centre for
Ecology & Hydrology

ECN example: state definition







counts at t state	counts at (t+1) state 1	2	3	4	5
1	335	31	1	126	17
2	35	19	30	44	31
3	0	31	480	34	96
4	131	42	39	374	61
5	8	36	92	69	193

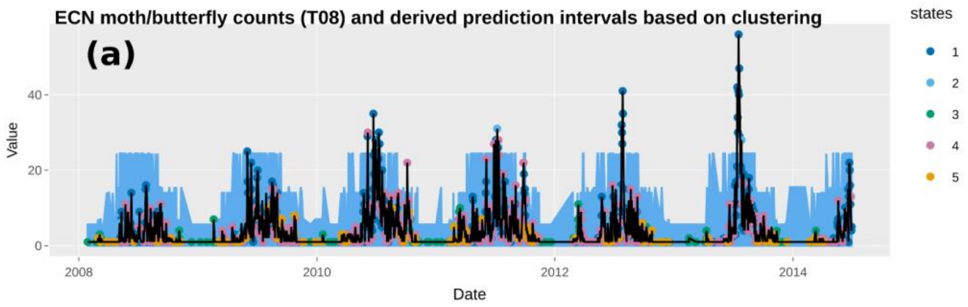
Example from the UK ECN site of Wytham (part of the LTER network).

Automatic weather station data are used for state tagging via K-means clustering.

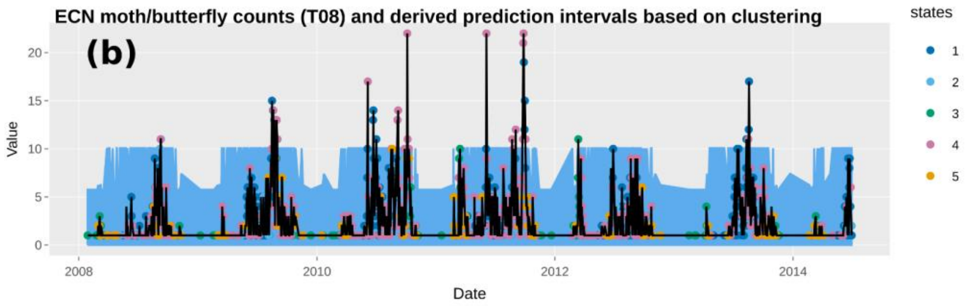
Observational variables are from daily moth traps and (seasonal) butterfly traps.

ECN example: 95% prediction intervals

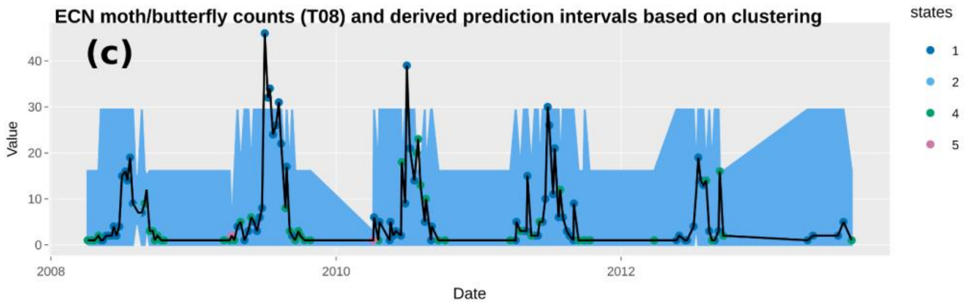
Moth counts (all):



Moth counts (noctuidae):

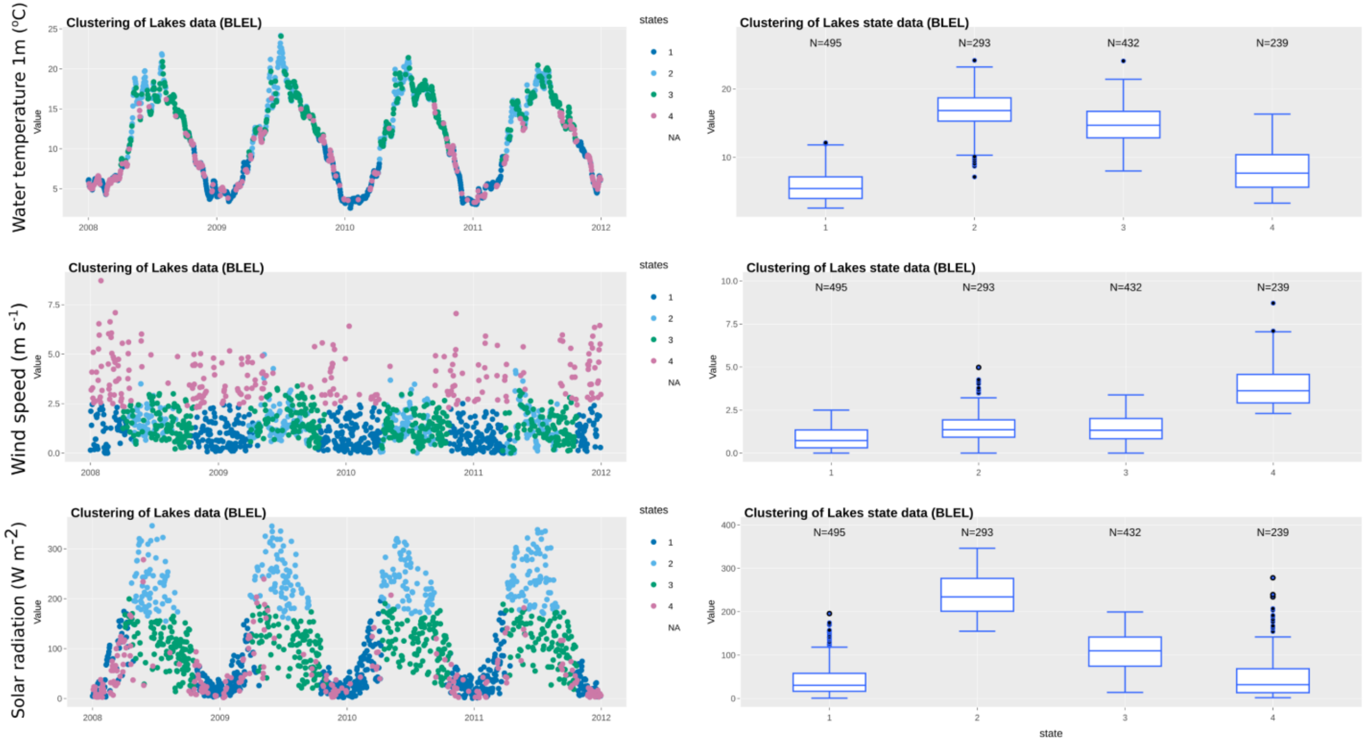


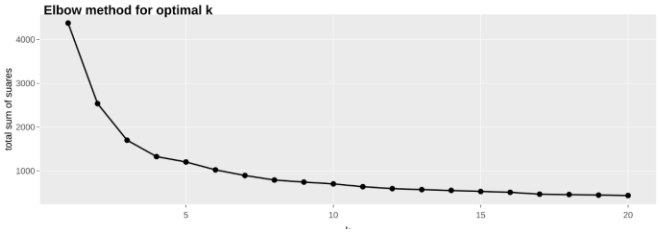
Butterfly counts (all):

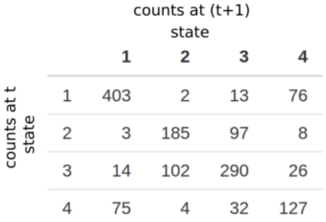


Note: the ECN data presented here has been range-checked.

Lakes example: state definition







		counts at (t+1) state			
counts at t state		1	2	3	4
	1	403	2	13	76
	2	3	185	97	8
	3	14	102	290	26
	4	75	4	32	127

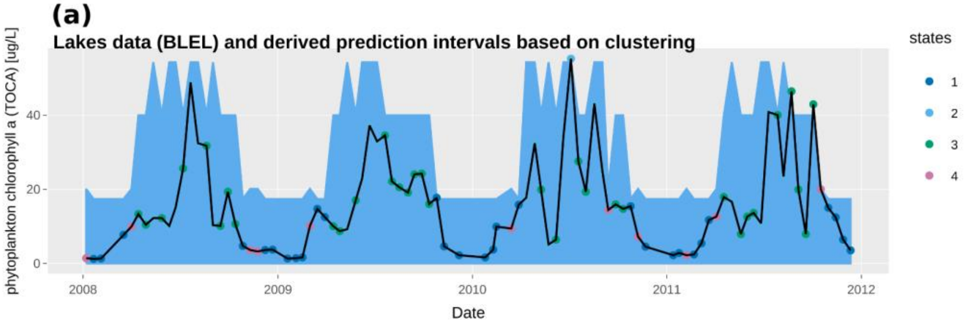
Example from the small English lake of Blelham Tarn.

Automatic buoy data are used for state tagging via K-means clustering.

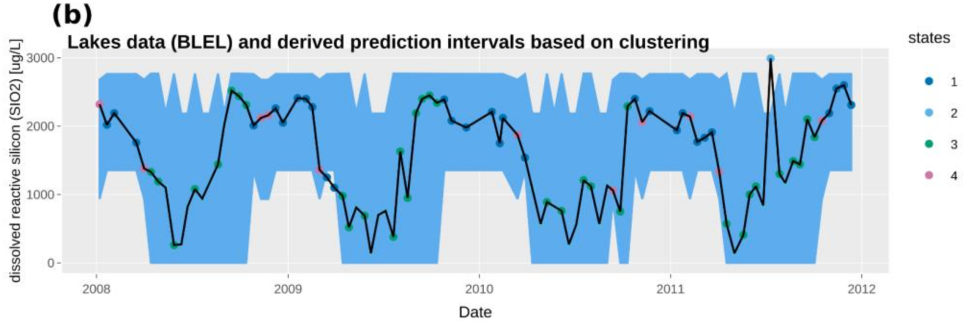
Observational variables are from manual sampling of lake biochemistry.

Lakes example: 95% prediction intervals

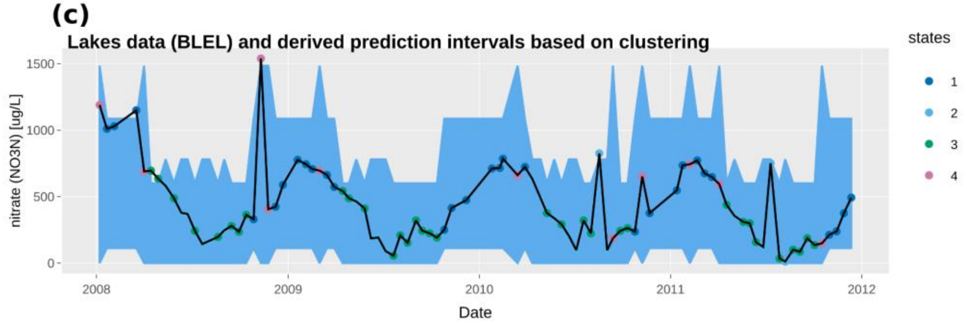
Total chlorophyll a:



Dissolved silicon:

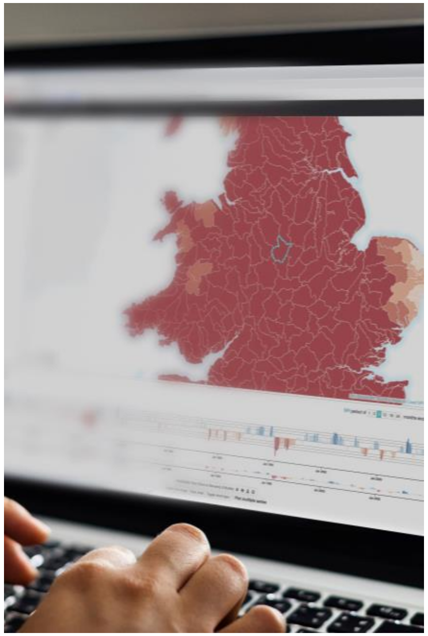


Nitrate:



Discussion and outlook

- Our method works for any time series of point data, which is very common in many earth and environmental applications
- It currently takes no consideration of time (i.e. the order of data is not important)
- Future work can extend its application to various types of spatial data
- It can potentially be used to identify whether there are systematic change in the system over time



Thank you

mtso@ceh.ac.uk

@Michael_ts0

Join the discussion using the EGU online forum, email, or Twitter.

Data availability

All data used are available through the following DOIs, hosted by the Environmental Information Data Centre (EIDC), a NERC Data Centre hosted by UKCEH.

App source code (generic version): https://doi.org/10.5285/1de712d3-081e-4b44-b880-b6a1ebf9fcd8 (Tso 2020)

ECN data

- Butterflies: https://doi.org/10.5285/5aeda581-b4f2-4e51-b1a6-890b6b3403a3 (Rennie et al., 2017a)
- Moths: https://doi.org/10.5285/a2a49f47-49b3-46da-a434-bb22e524c5d2 (Rennie et al., 2017b)

UK CEH Cumbrian Lakes monitoring scheme data(Blelham Tarn)

- Automatic buoy: https://doi.org/10.5285/38f382d6-e39e-4e6d-9951-1f5aa04a1a8c (Jones and 509Feuchtmayr, 2017)
- Long-term manual sampling data: https://doi.org/10.5285/393a5946-8a22-4350-80f3-a60d753beb00511 (Maberly et al., 2017)