

Integration of long-term collocated ecological datasets: examples from the UK Environmental Change Network (ECN)

Chak-Hau Michael Tso, Aaron Lowther, Don Monteith, Linsay Flynn Banin, William Simm, Susannah Rennie, Michael Hollaway, Peter Henrys, Rebecca Killick, John Watkins, and Gordon S. Blair

Centre for Excellence for Environmental Data Science
UKCEH and Lancaster University

European Geoscience Union General Assembly 2021 (DOI:10.5194/egusphere-egu21-2293)

Session BG3.13 – Whole system approach for in-situ & long-term environmental system research on life supporting systems (WAILS)

Structuring, integration and visualisation of Research Infrastructures [Highlight]

Virtual labs to support WAILS research

- Integration of long-term datasets present not only methodological challenges, but also an infrastructure challenge.
- DataLabs:** A cloud infrastructure where an interdisciplinary team can seamlessly collaborate.
 - Everyone in the same space so they can work on the same datasets, codes, software environments and utilize elasticity of the cloud.
 - Easy to run notebooks and apps live– facilitate discussions and build trust, encourage co-development of methods.
- In each example, we will discuss the challenge, solutions provided via DataLabs, and research findings:
 - Rainfall chemistry (3rd party data)
 - A multi-sampling rate changepoint detection algorithm
 - A state tagging framework (data analytics)

The image shows a screenshot of the DataLabs web interface. The top section is the 'Landing page' which displays a list of notebooks under the 'Notebooks' tab. The notebooks listed are 'cptdiffreq', 'Demo', 'ecnmet', and another 'Ready' notebook. Each notebook has an 'Open' button and a status indicator.

Below the landing page is a screenshot of an 'Rstudio notebook'. The notebook is titled 'Exploratory data analysis of moth data across all ECN sites'. It contains R code for data manipulation and analysis, including loading packages like 'tidyverse', 'feather', 'dplyr', 'ggplot2', 'readr', and 'reticulate'. The code is organized into sections for data loading, cleaning, and analysis.

At the bottom of the Rstudio notebook is a 'JupyterLab notebook' section, which shows a list of files and a code editor with R code for loading packages and setting up the environment.

UK Environmental Change network (ECN)

An eLTER member network

- Long-term collocated data since 1992 to support the whole-system approach (WSA) to integrate data and models from various sub-systems
- Widespread use of WSA requires infrastructure, methodologies, and a culture shift in order to facilitate seamless collaboration and re-deployment of workflows.
- We present recent progress of using DataLabs to address some of these challenges.

DataLabs: a gentle introduction

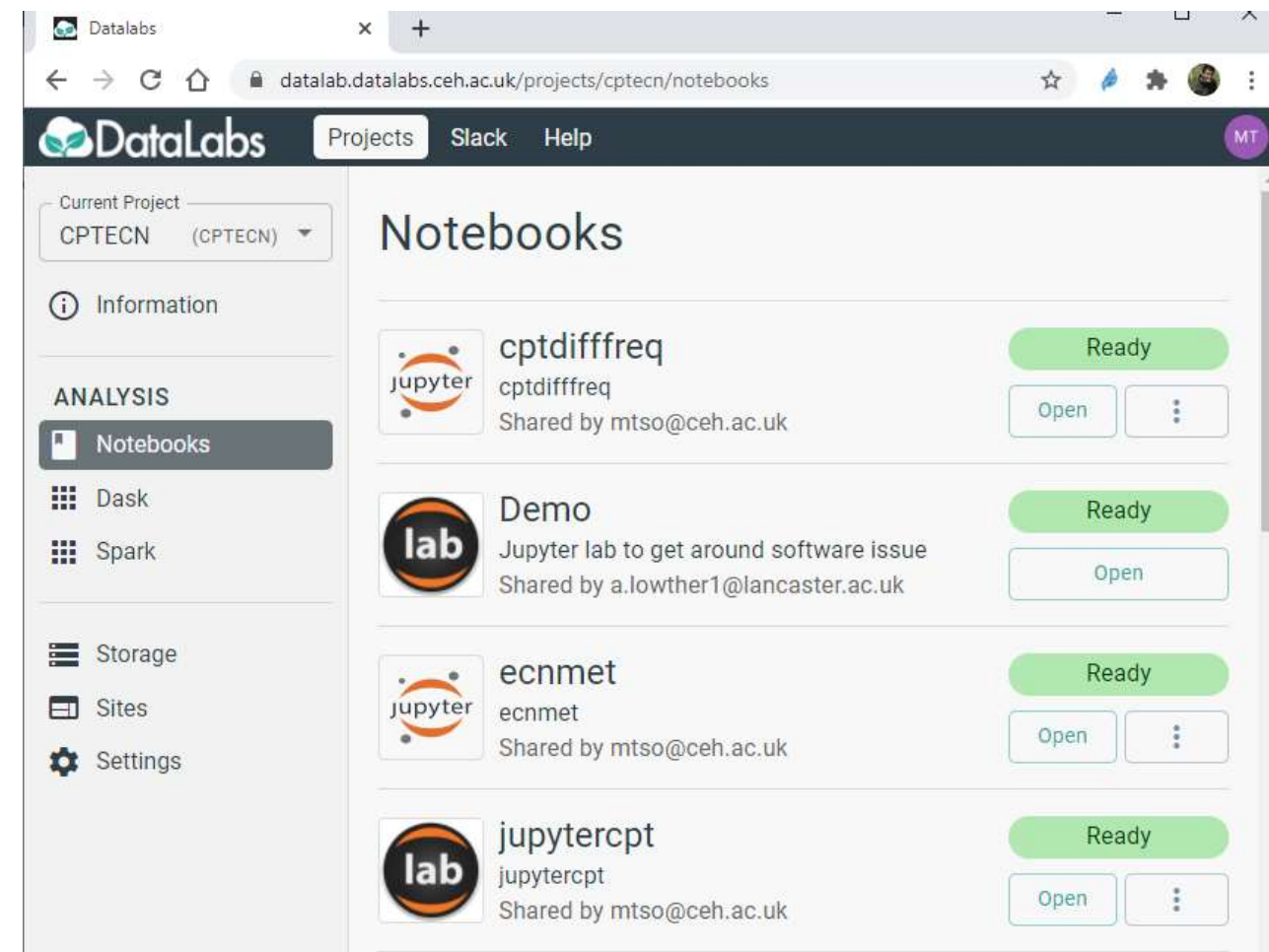


- Transdisciplinary **collaboration** space
- **Cloud-based** compute and storage
- Develop **common** resources and integrate with novel/existing analytical workflows
- Access **variety** of data, analytical methods and assessment tools
- Support **wide range** of stakeholders

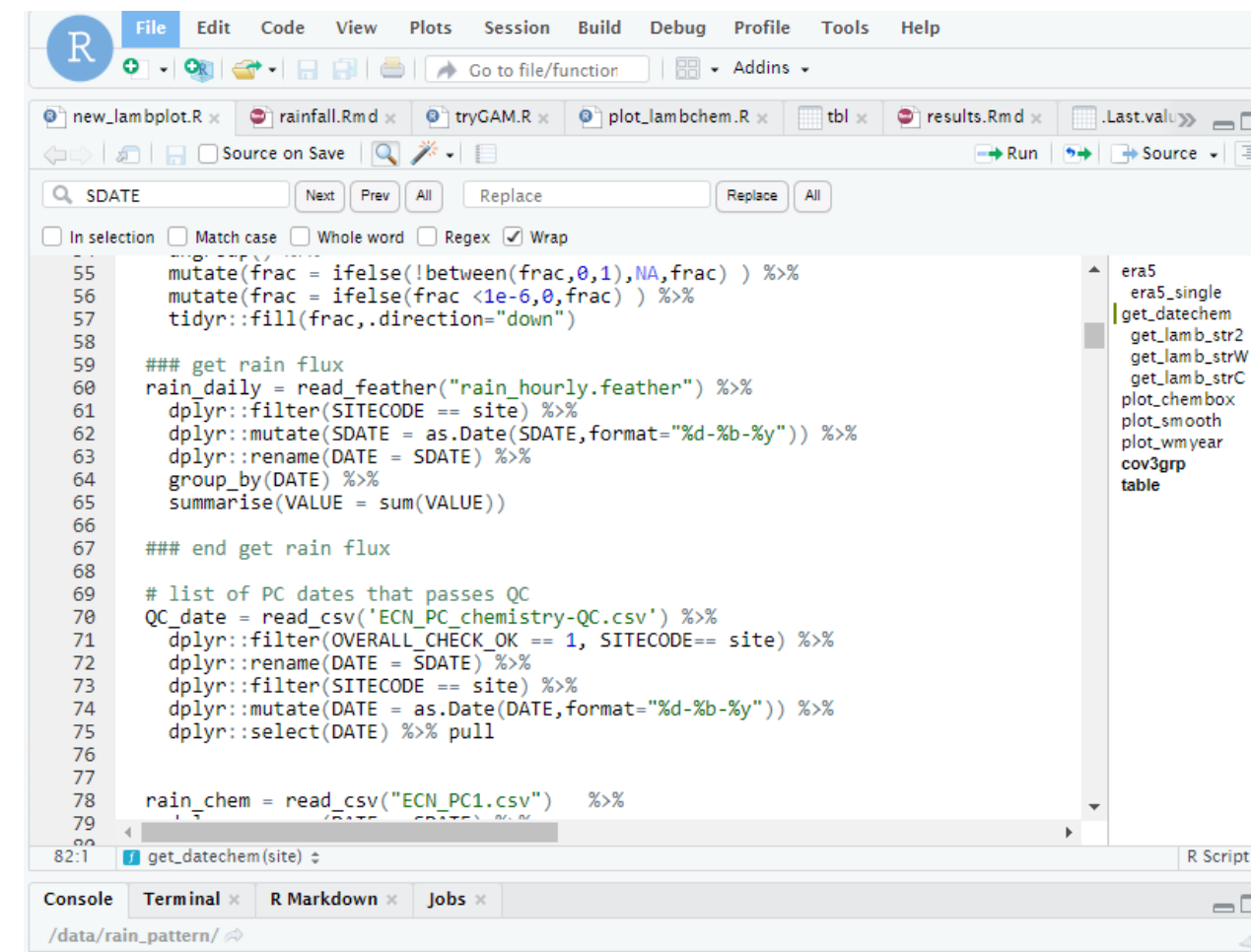
How can DataLabs supports WAILS research?

- Data, code, software package **in one cloud project space**
 - Controlled and shared **software environment**
 - You can see **what others have done** in this project and similar work.
- **Fostering collaboration** across disciplines and sub-disciplines
- **Seamless integration of workflows**
 - Easy to re-run the entire analysis by mix-and-matching **different components**
 - Allows **end-to-end analysis** of uncertainties
- Support **wide range** of stakeholders
 - Easy-to-use tools to **publish** notebooks and web applications (e.g. R Shiny)

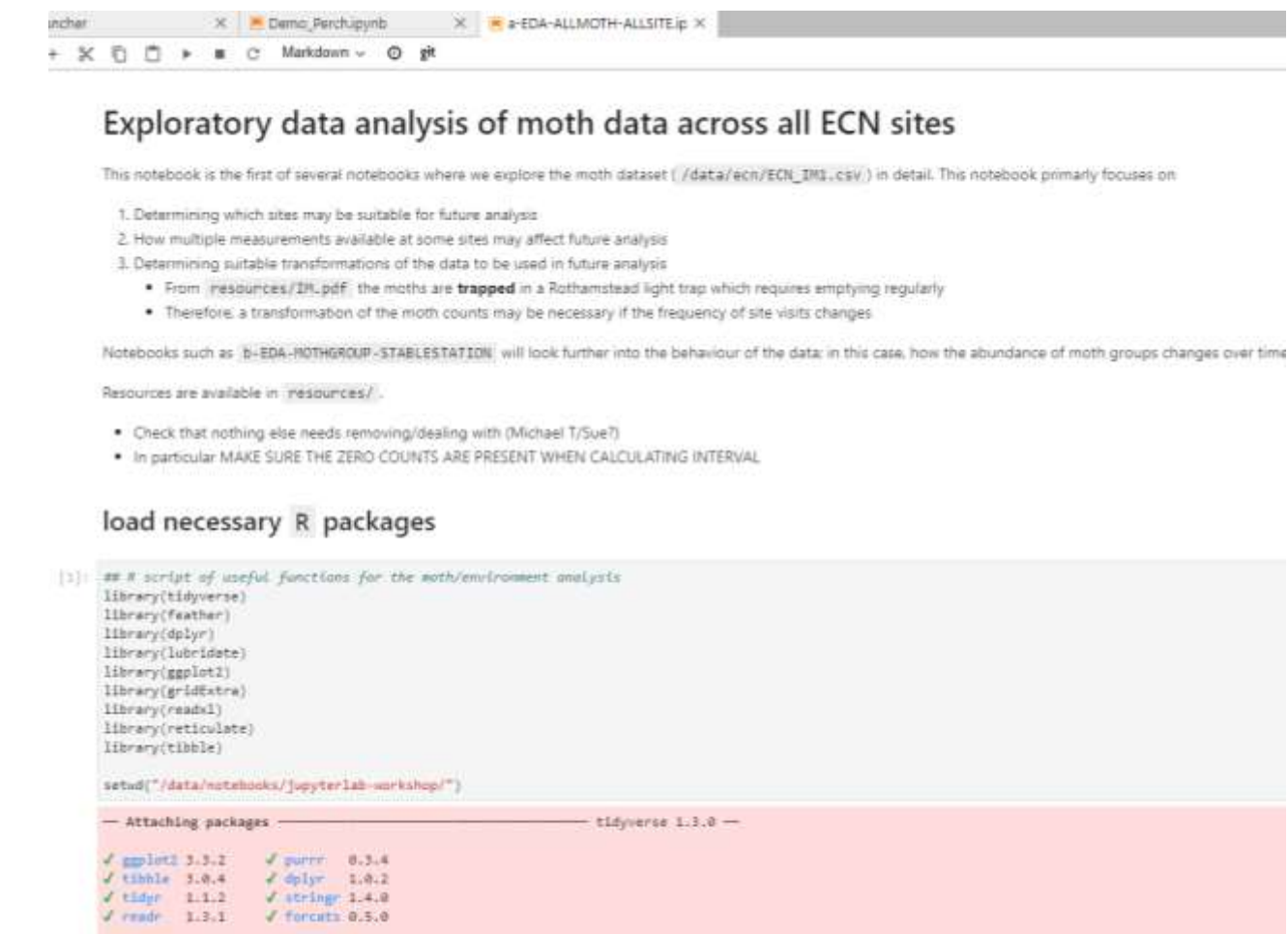
The ECN DataLabs project space



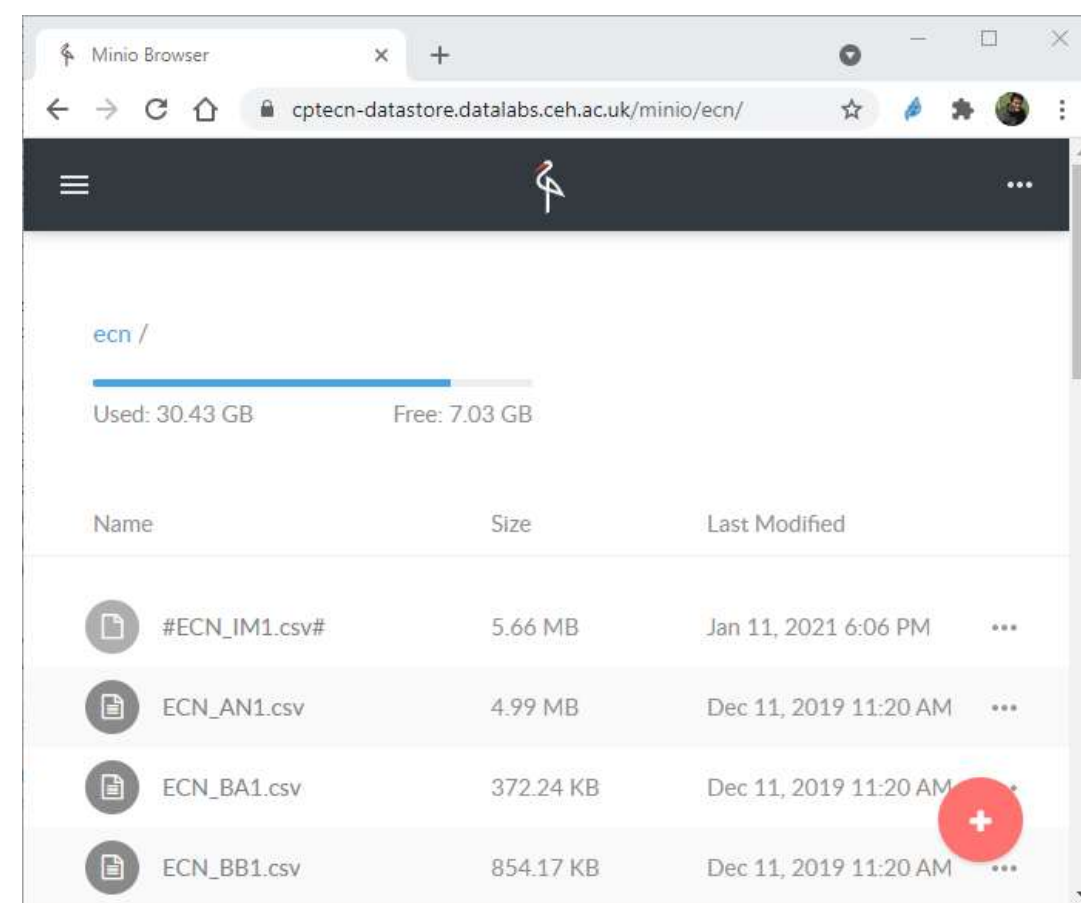
Landing page



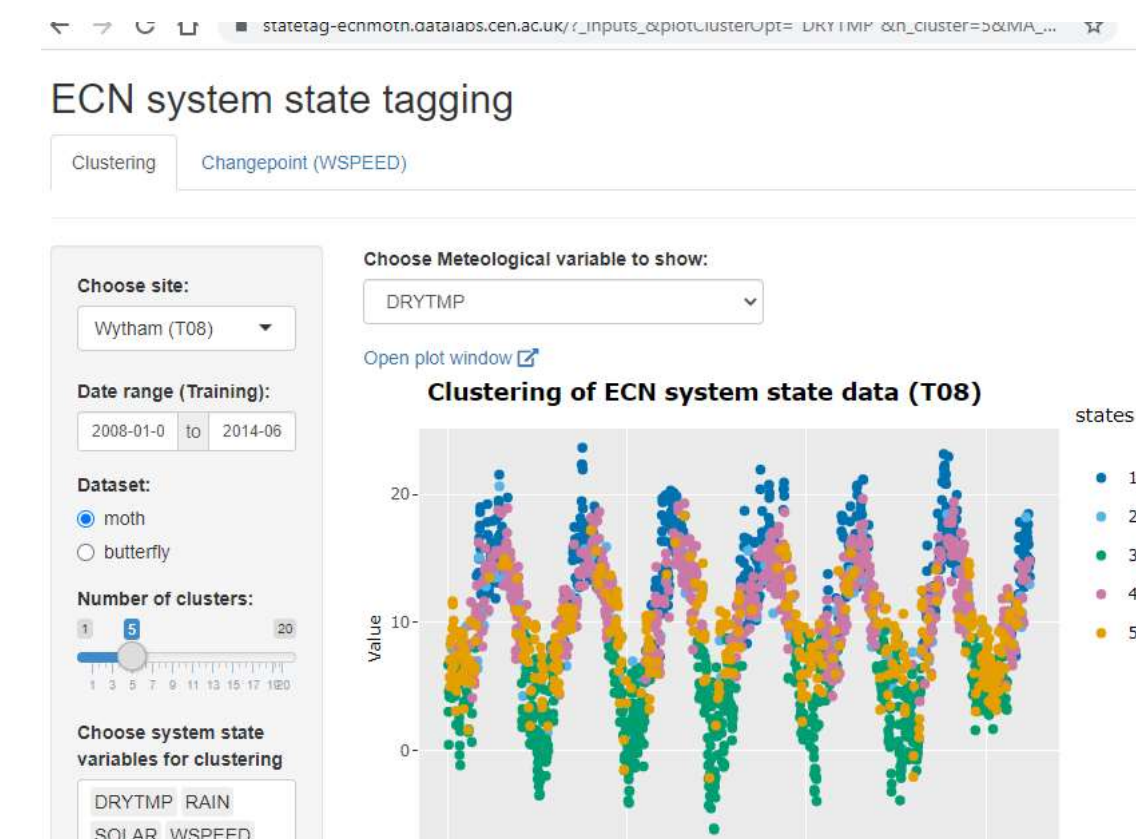
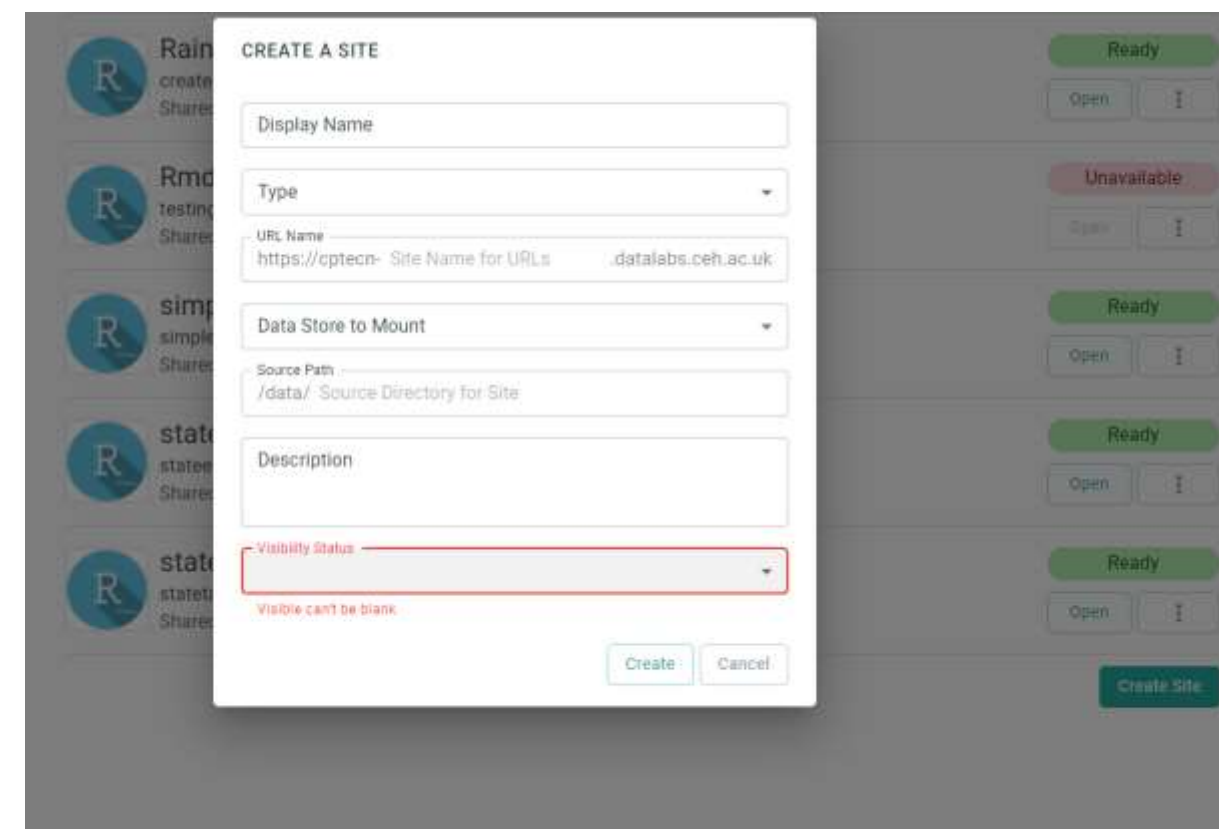
Rstudio notebook



JupyterLab notebook



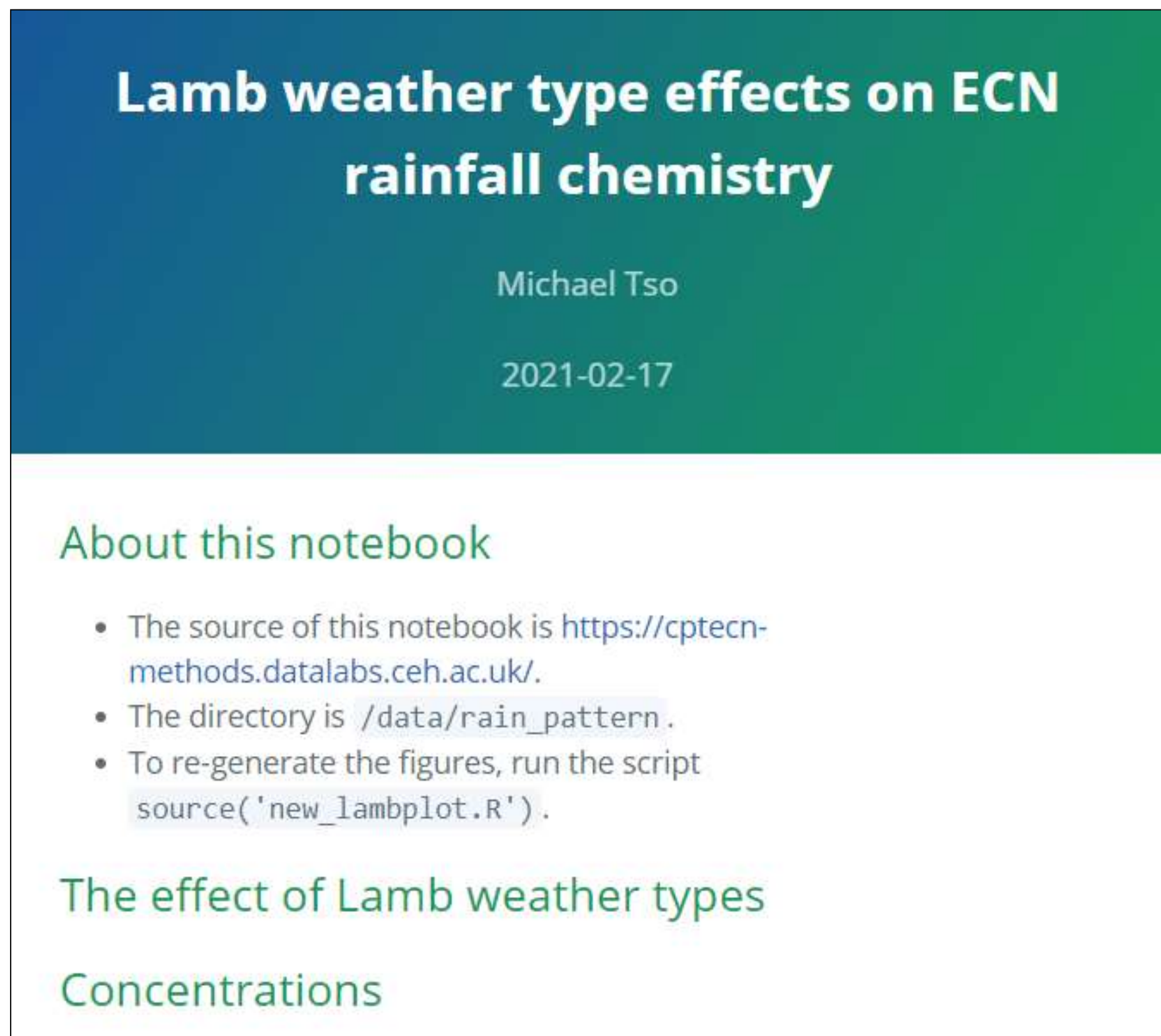
MinIO data store



Publish R Shiny apps with one click

Example 1: GB rainfall chemistry (1/3)

How DataLabs facilitate the use of third-party data



Lamb weather type effects on ECN rainfall chemistry

Michael Tso

2021-02-17

About this notebook

- The source of this notebook is <https://cptecnmethode.data-labs.ceh.ac.uk/>.
- The directory is `/data/rain_pattern`.
- To re-generate the figures, run the script `source('new_lambplot.R')`.

The effect of Lamb weather types

Concentrations

R markdown notebook documenting results generated on DataLabs

What are the effects of weather and rainfall type on GB rainfall chemistry?

Needs to bring in 3rd party data

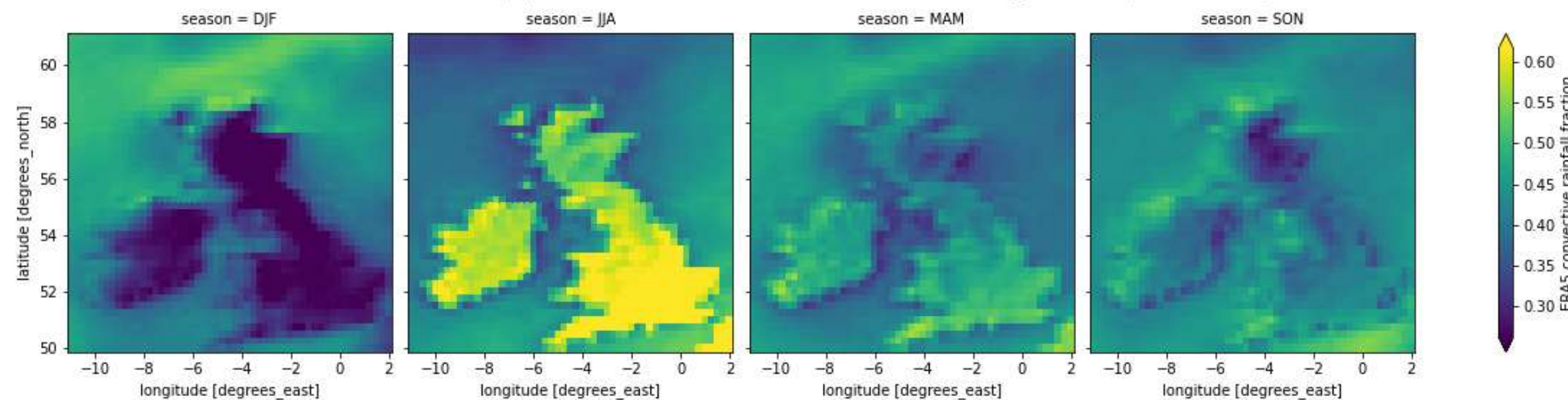
- Lamb weather types (Daily: cyclones, westerlies, etc.)
- Convective rainfall fraction (Gridded hourly, ERA5)
- Air mass trajectory data (NOAA)

Everything is stored and done in the cloud in DataLabs

Example 1: GB rainfall chemistry (2/3)

How DataLabs facilitate the use of third-party data

(a.) ERA5 Mean convective rainfall fraction by season (1979-2019)



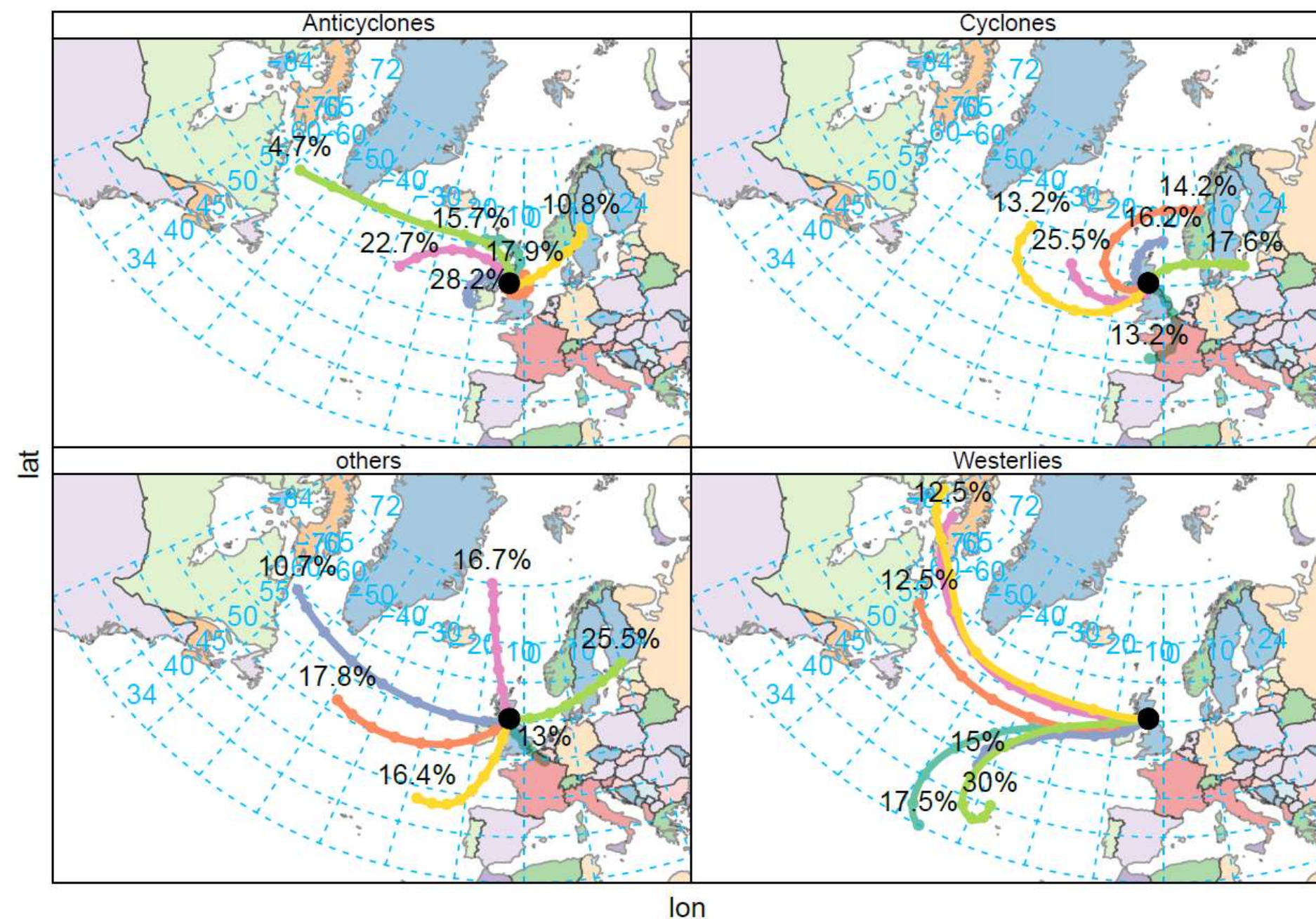
Solutions:

Most analysis occur in R Studio in DataLabs

R markdown reports

ERA5 data (.nc files, 25 years) downloaded using python API to DataLabs, where cloud storage and analysis occurs

Downloaded global gridded weather data to run back trajectory analysis on DataLabs → cloud resources, can run in the background using R Studio jobs even when browser is closed

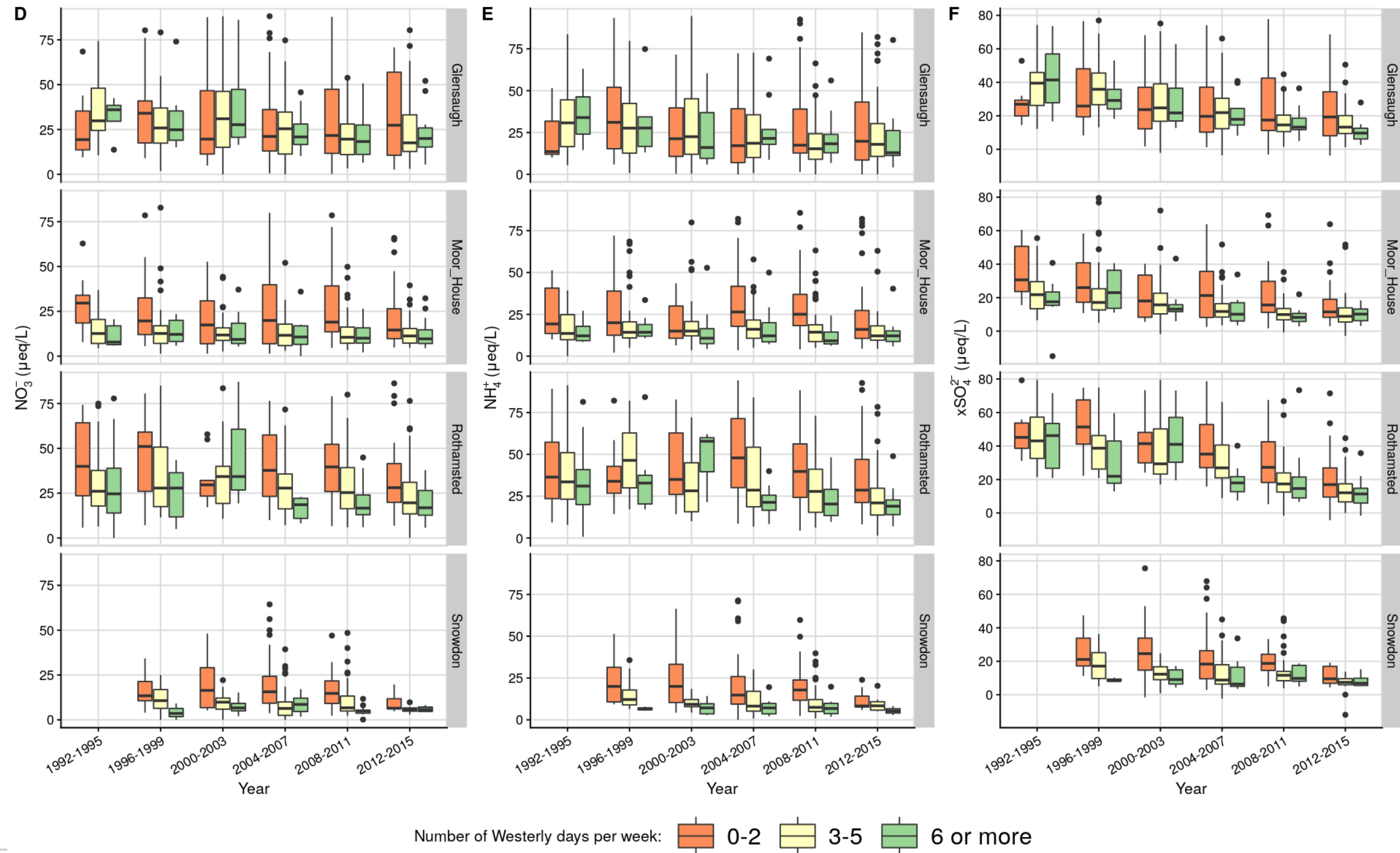


Example 1: GB rainfall chemistry (3/3)

How DataLabs facilitate the use of third-party data

Research Findings:

- Lamb westerly/cyclone days per week shows strong negative correlation with pollutant concentrations
- LWT effects on pollutant fluxes depends on the interplay between rainfall amount and concentration



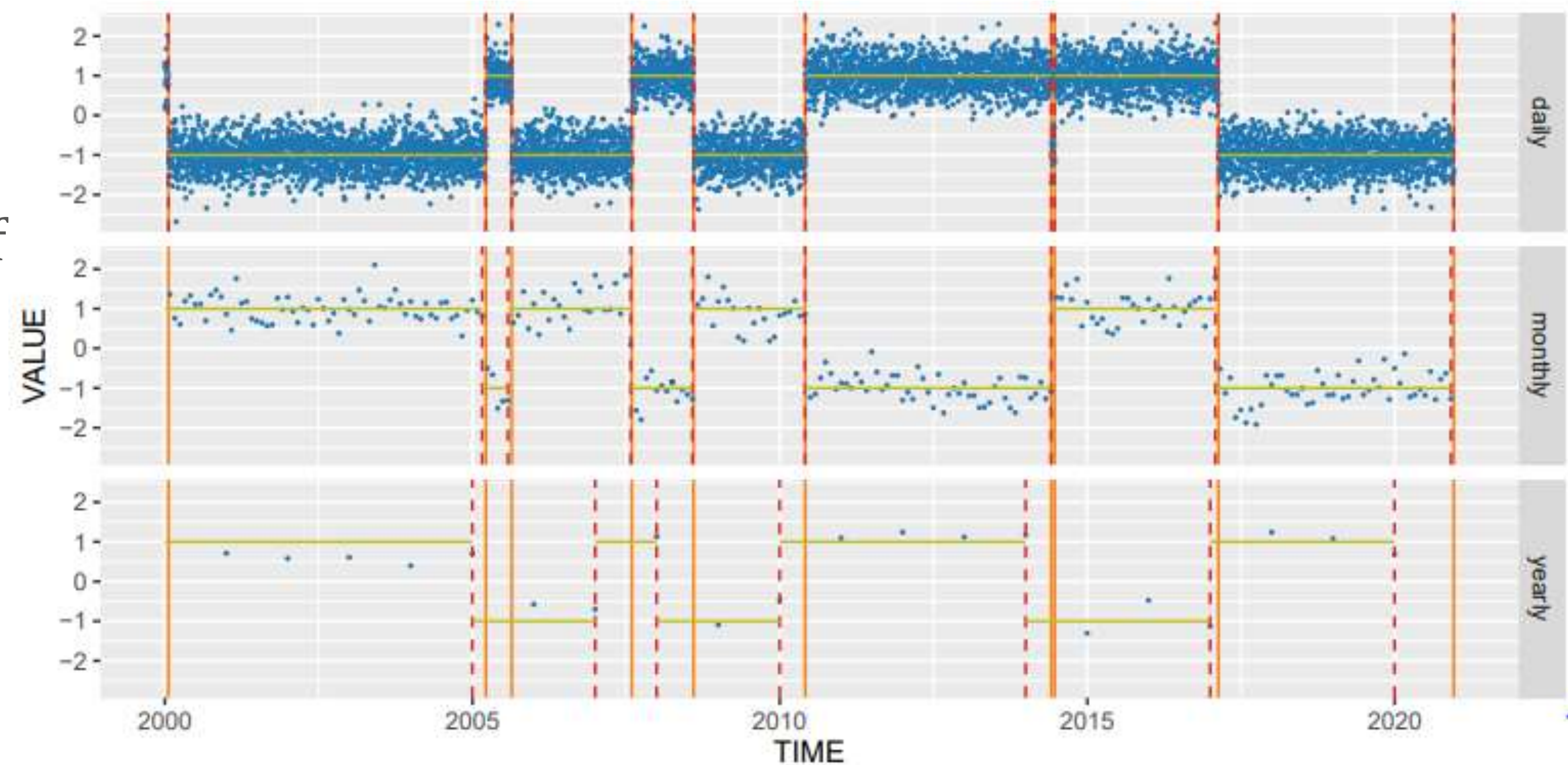
Example 2: A multi-sampling rate (MSR) changepoints detection method

- A changepoint is a location in a time series where we are about to observe a change in statistics (e.g. mean, variance, or both) in the segment before and after it. Statistics algorithms are used to detect changepoints objectively.
- Changepoints are important in environmental science because a step change in the system may have catastrophic effects (i.e. tipping points) or shows that the system is in a new 'state'.
- Challenge: Existing multivariate changepoint algorithms assumes the same frequency of all the time series and are properly aligned (as in engineered systems). Environmental data often has gaps and are measured in different frequencies.

Example 2: A multi-sampling rate (MSR) changepoints detection method

Solution:

- A MSR changepoint method is developed for multiple time series of different sampling frequency or different observation times.
- It takes the union of all sampling time as candidate changepoints and the detection of changepoints is informed by all datasets.
- The method is co-developed by statisticians and environmental scientists in DataLabs and is demonstrated with a series of Jupyter notebooks.
- Virtual workshop held with live analysis in DataLabs



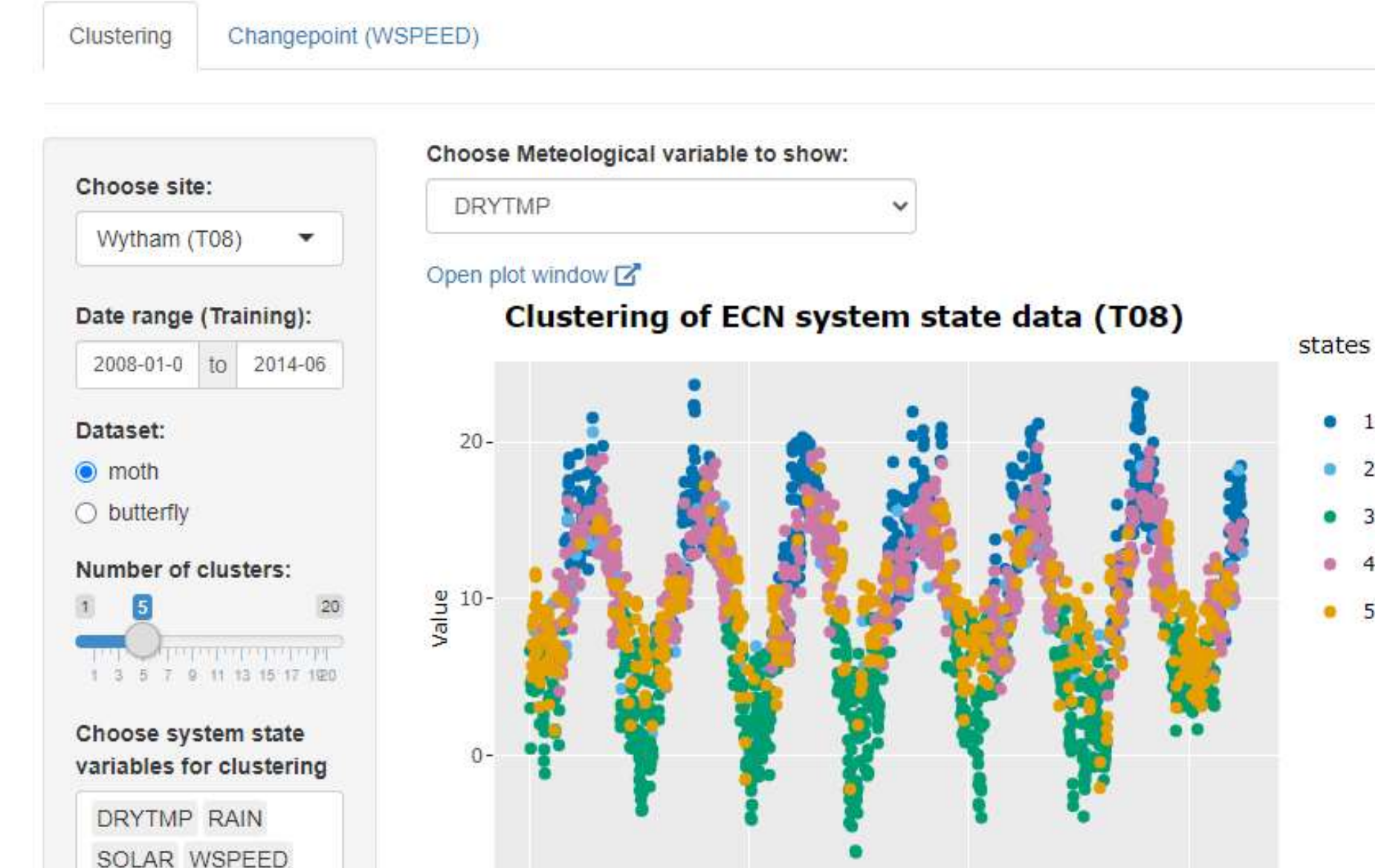
Example 3: Advancing environmental data analytics

A state tagging framework

How to include quality assurance of environmental data?

- Many different sub-disciplines
- Needs unsupervised method
- Quick check for anomalous data/data of interest

ECN system state tagging

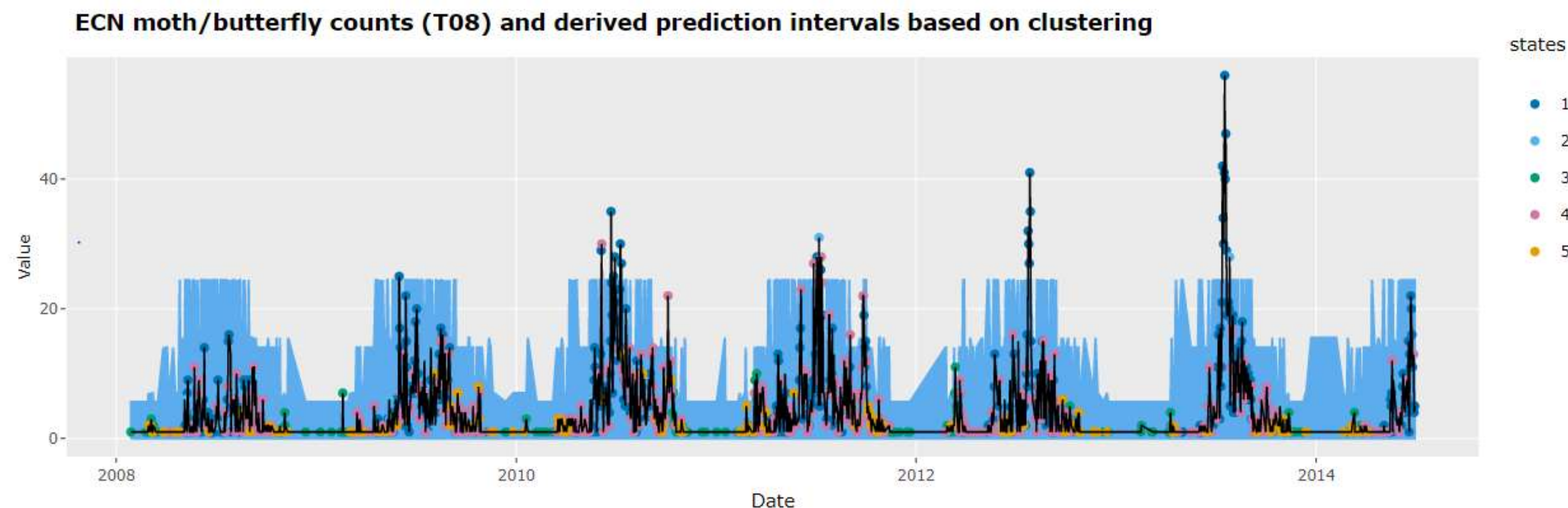


Example 3: Integrating long-term collocated datasets

How does DataLabs help?

Solution:

- Allow users to upload and select state variables to determine states
- States are based on clustering or other unsupervised methods
- Return state ID for each observation and prediction intervals for each state
- Flag data outside prediction intervals
- Applicable to any point data time series



Summary of use cases

How DataLabs contributes as part of the solution to their challenges

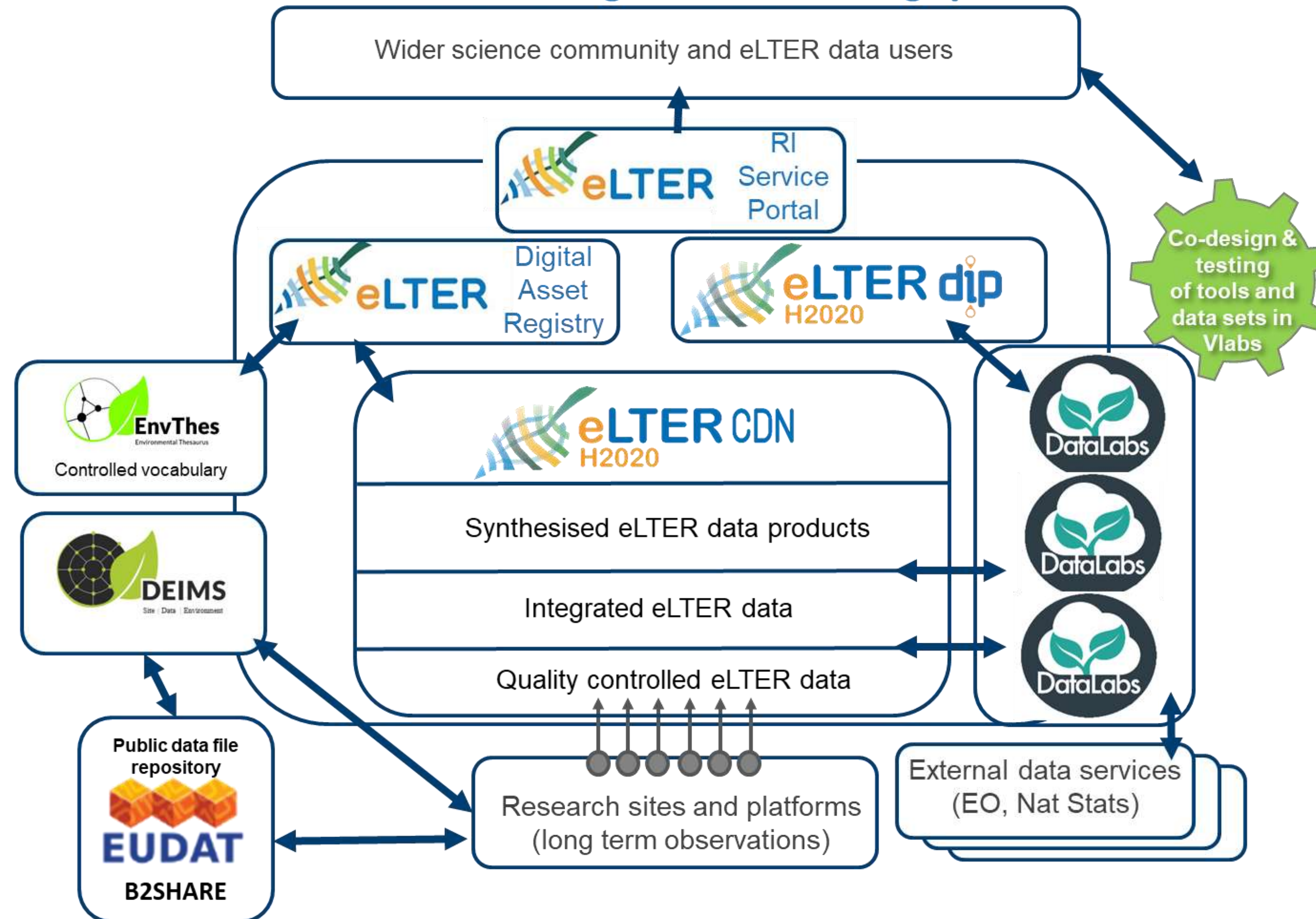
Case study:		Env. Science	Cultural	Computing/Software
GB rainfall chemistry	Challenge:	How does weather types affect rainfall chemistry?	Sharing research output	Third party data (e.g. ERA5 reanalysis)
	DL Solution:	Shiny apps, interactive notebooks and workflows	Interactive notebooks	Project data store
Multi-sampling rate changepoints	Challenge:	How to identify simultaneous change when data may have different frequencies	Develop a data science methods for env. science	Ability to use the methods as a cloud service on demand
	DL Solution:	Co-developed a new changepoint methods with data scientists	Workshops held using DataLabs	Server-less functions
Data analytics	Challenge:	Method that works for a wide variety of datasets	Workflows developed in silo for each datasets	Streaming or heterogeneous data
	DL Solution:	A suite of unsupervised methods	Consistent and modular design of DataLabs allow easy re-deployment of workflow with new data and methods	Host (or access via API) all data in DataLabs. Cloud service and methods for data discovery and integration

Final remarks

- 3 examples from a pilot study using **ECN** data in **DataLabs** for WAILS research
- They showcased DataLabs is a **key infrastructure** to advancing environmental data science, especially WAILS research
 - Cloud storage, computing, and collaboration
 - Easy implementation of whole-system workflows
 - All eLTER data in one place, plus any third-party data previously used
- The DataLabs approach **adds value** to ECN data

DataLabs and current eLTER programs

eLTER IT architecture: existing modules and gaps to be filled



Acknowledgments



CEEDS

CENTRE OF EXCELLENCE IN
ENVIRONMENTAL DATA SCIENCE

This work is generously supported by:

- UKRI Strategic Priority Fund— NERC Constructing a Digital Environment Pilot Study: Methodologically Enhanced Virtual Labs for Early Warning of Significant or Catastrophic Change in Ecosystems: Changepoints for a Changing Planet (NE/T006102/1)
- UK-SCAPE (<https://www.ceh.ac.uk/ukscape>)
- eLTER (<https://www.lter-europe.net/>)

Thank you

@michael_ts0

mtso@ceh.ac.uk



UK Centre for
Ecology & Hydrology