Graph-based river network analysis for rapid discovery and analysis of linked hydrological data

Matt Fry
UK Centre for Ecology and Hydrology
mfry@ceh.ac.uk
Jan Rozecky, Epimorphics Limited
jan.rosecky@epimorphics.com

EGU2020 online
Tuesday 5th May 2020

UK Centre for Ecology & Hydrology

Integration of data across the hydroscape

Humans and machines need to understand the **context** to derive genuine **understanding** from data

Environment Agency has data from river samples across >20k sites:

- > 12M measurements of **nitrogen**
- > 2.6M measurements of **phosphate**
- > 2.5M measurements of **BOD**

Can a machine understand this data without information on:

Images courtesy Pixabay

But how do we incorporate this information?

To find and use data related to rivers, the connectivity of sites to rivers is essential information:

- Find monitoring data upstream of a location
- Find nearest flow site for a water quality site
- Which of these sites has a lake upstream?
- How far is the nearest upstream sewage effluent discharge?

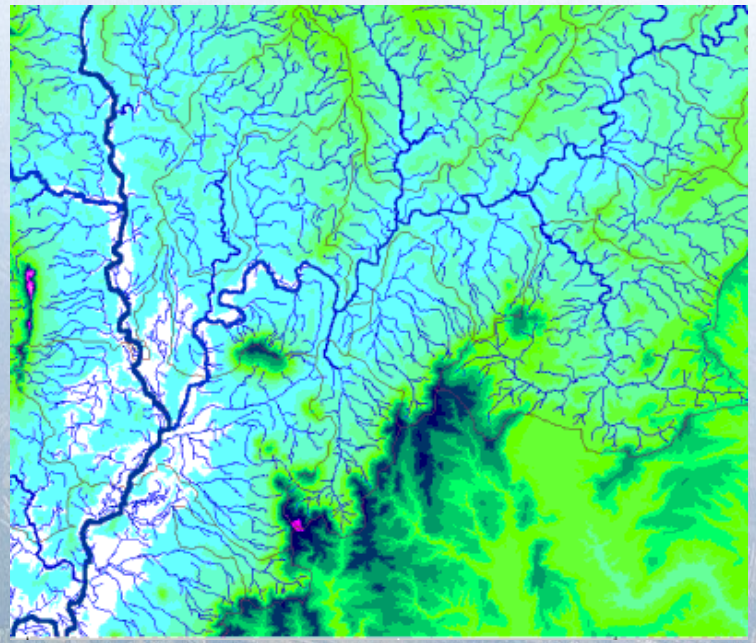
Data discovery may need to undertake these queries in real-time

Machine-driven analyses may need to perform these queries across many sites

Integration of data across the hydroscape

Building a digital representation of the freshwater environment

Using digital rivers to link:



- Monitoring sites and sensors
- Man made drivers of water quality:
  - Sewage treatment works
  - Crop maps
- Natural influences:
  - Soils
  - Lakes
  - Riparian tree cover

Could extend to include roads, livestock, abstractions, hedges.....

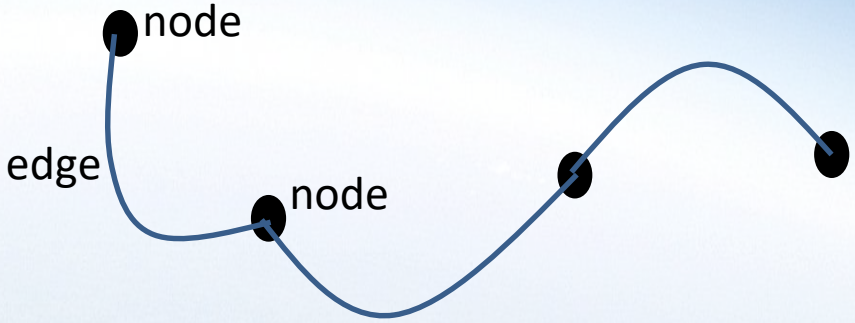
Integration of data across the hydroscape

- Digital rivers with rapidly accessible connectivity information
- Rivers attributed with key geo-spatial information
- Monitoring sites linked to digital rivers
- Usable summary information to identify sites of interest

- Connectivity data could be pre-processed, but for how many combinations of sites / features?
- GIS networking tools can provide this functionality but there may be limitations in speed and integrating in analytical code-bases
- networkX is a python package for analysis of network datasets

networkX

- undirected graph



- directed graph

- multi-graph

(can include multiple edges between the same node to describe bifurcations)

Representing a river network as a graph

- edges are river stretches
- nodes are points at intersections or other points of change along a stretch

attributes:

- length
- name
- % length through woodland, urban area, clay soil, etc.

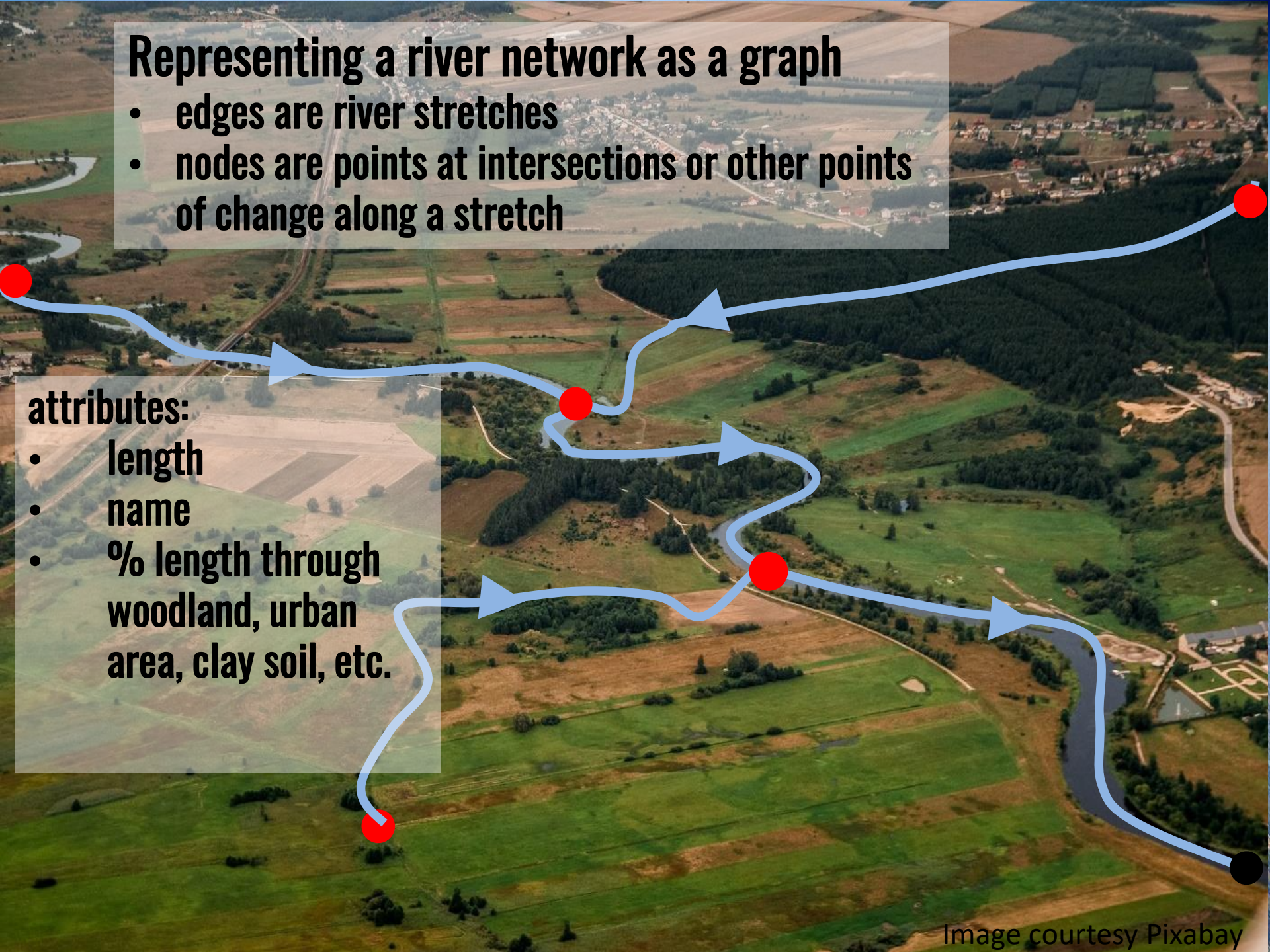
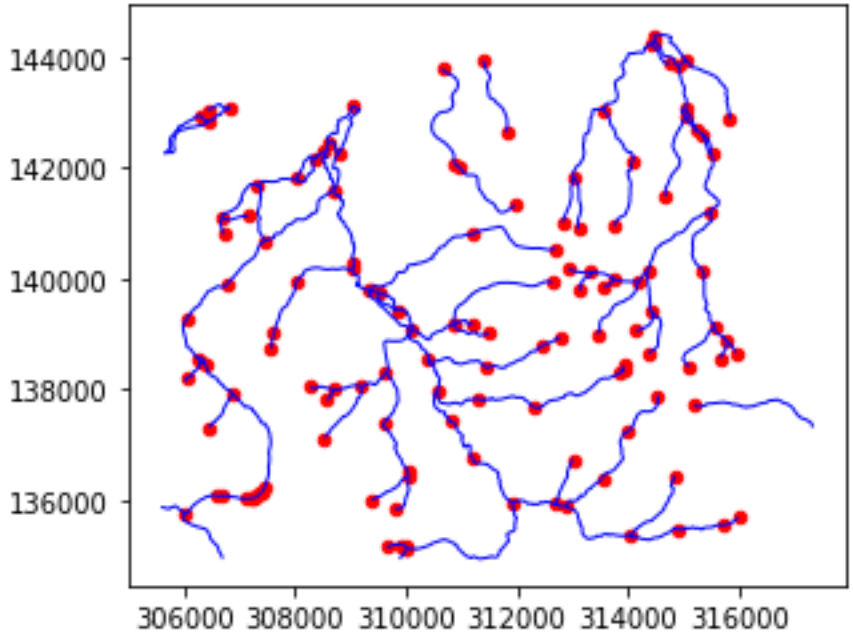


Image courtesy Pixabay

UK river network:

- from 1:50k maps
- 433k nodes, 436k edges,
- total length: 290k km
- networked
- attributed with river name and hydrometric area

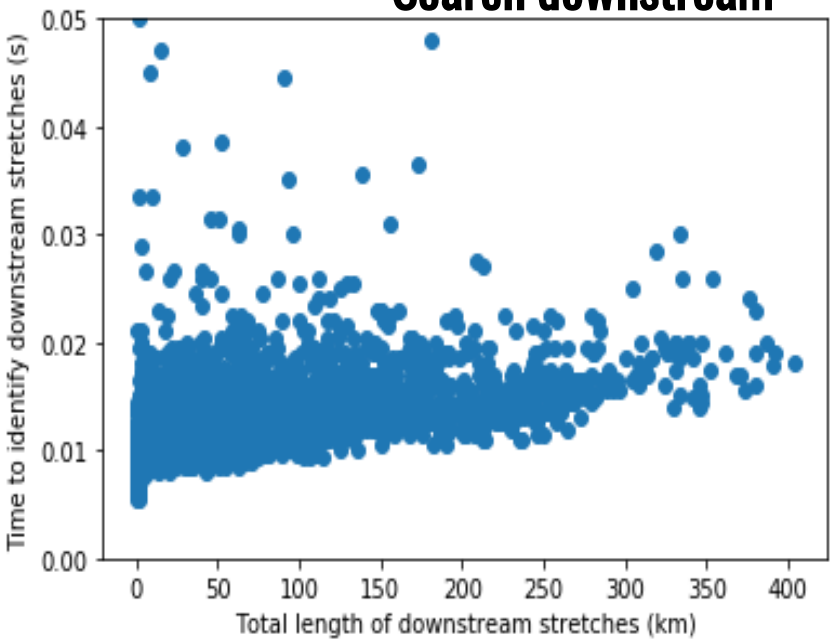


Searching river network

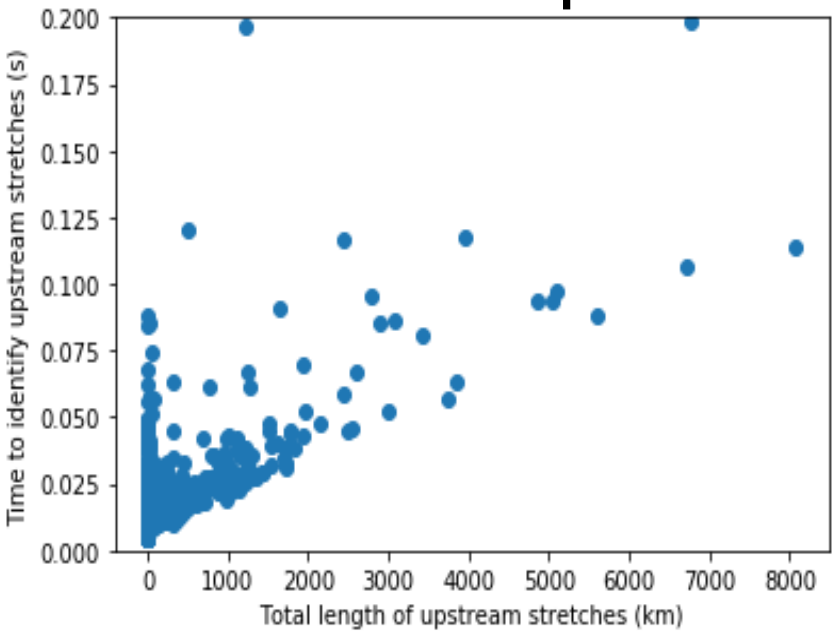
- Identify all up/downstream stretches from a point
- Calculate total length of stretches
- <20ms for 98/94% of down/upstream searches

Speeds viable to support real-time discovery / analysis tools

Search downstream



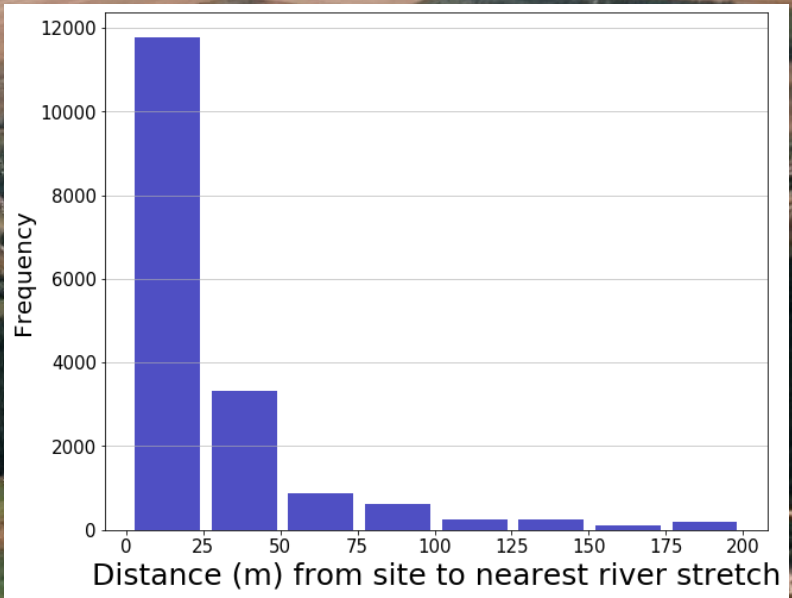
Search upstream



Linking sites and rivers

Water quality data for England

- ~18k water quality sample sites
- Spatial locations not precise wrt rivers
- Matching based on distance
- Many errors, with distance providing uncertainty
- Similarity between river name and site name also provides uncertainty information



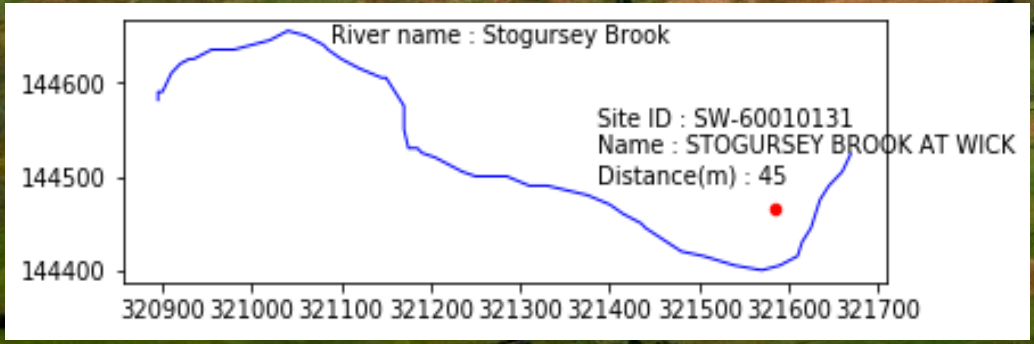
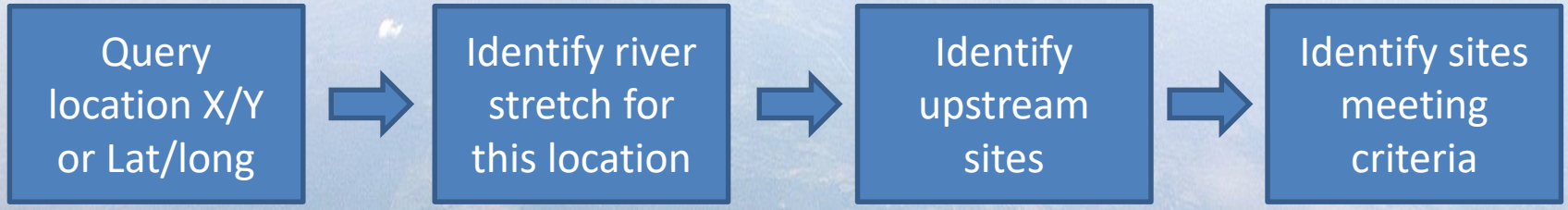


Image courtesy Pixabay

Data discovery

Use case: Query site on river stretch, identify all upstream water quality monitoring sites that have more than 5 years data on both phosphorous and nitrate concentrations.



Data discovery for scientific analysis of monitoring data benefits from summary information on availability of data:

- Determinands measured at the site
- Data available for these determinands
- Frequency of measurements
- Measurement procedure / processing (e.g. sampled with lab analyses vs in situ sonde)

Potentially, Data by date (e.g. year), Completeness, Summary statistics?

Data discovery: Working towards expressing monitoring time series data using common standards

- JSON-LD format based on concepts drawn from SSN/SOSA and INSPIRE EMF.
- Enables exposure and linking of vocabulary terms for observed properties, sensors, etc.

Describing:

- Monitoring networks
- Feature of interest / Ultimate feature of interest
  (e.g. location on river / the river itself)
- Statistical measure
- Sensor and procedure
- Complex model of properties (e.g. phosphorous, dissolved, concentration)

Review of sensor metadata stanadrds: http://nora.nerc.ac.uk/id/eprint/526628/

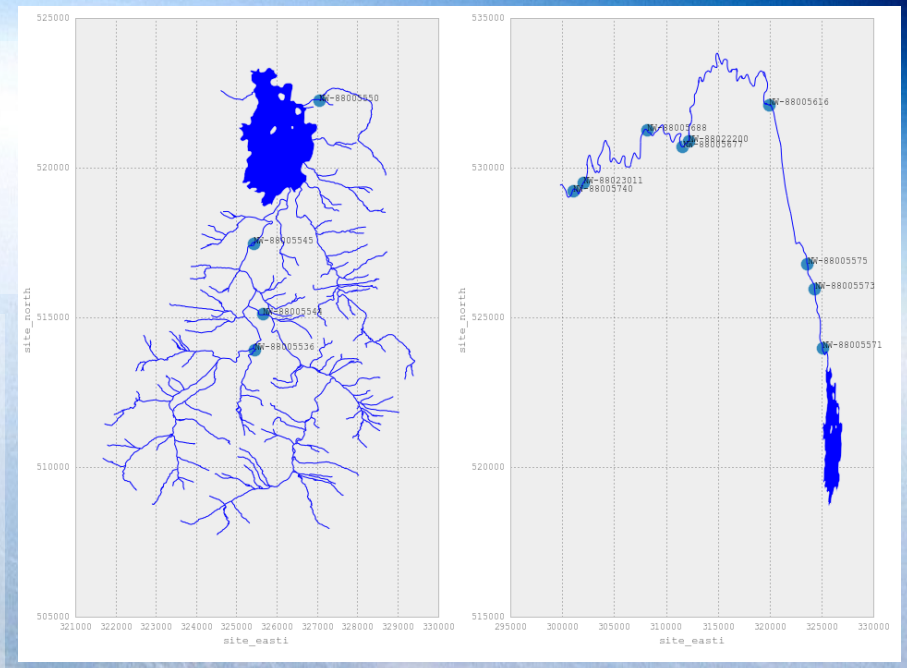
Integration of data across the hydroscape

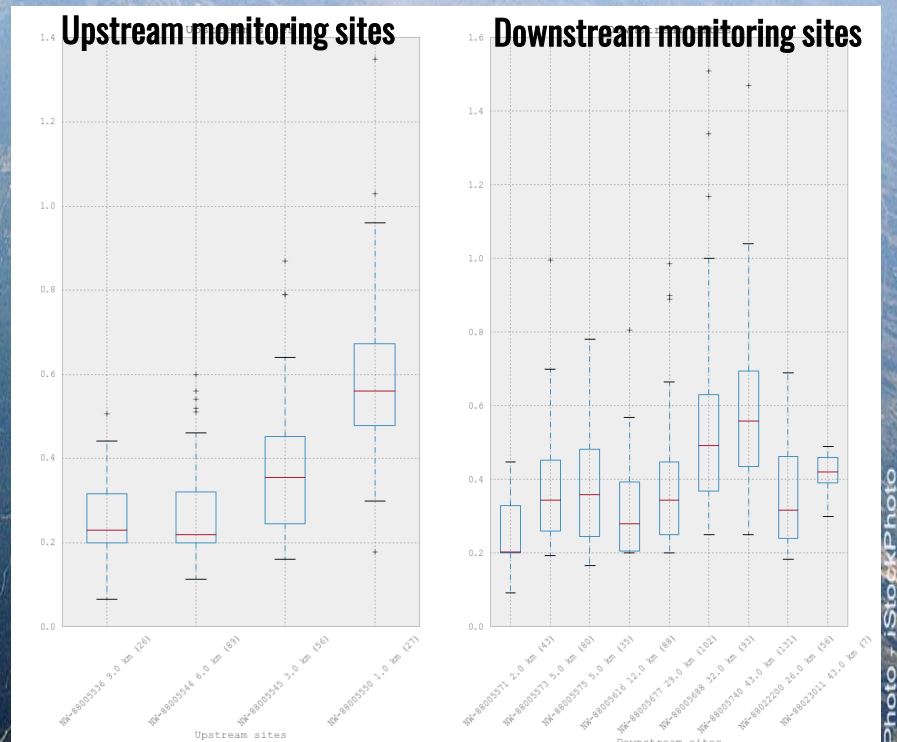
Example of analysis:

Automating the analysis of the impact of lakes on nitrogen and phosphorous

Identify all monitoring sites up and downstream of >1000 lakes in UK Lakes Database

The integration of key national scale datasets makes research (human and machine driven) and the development of evidence simpler and more effective

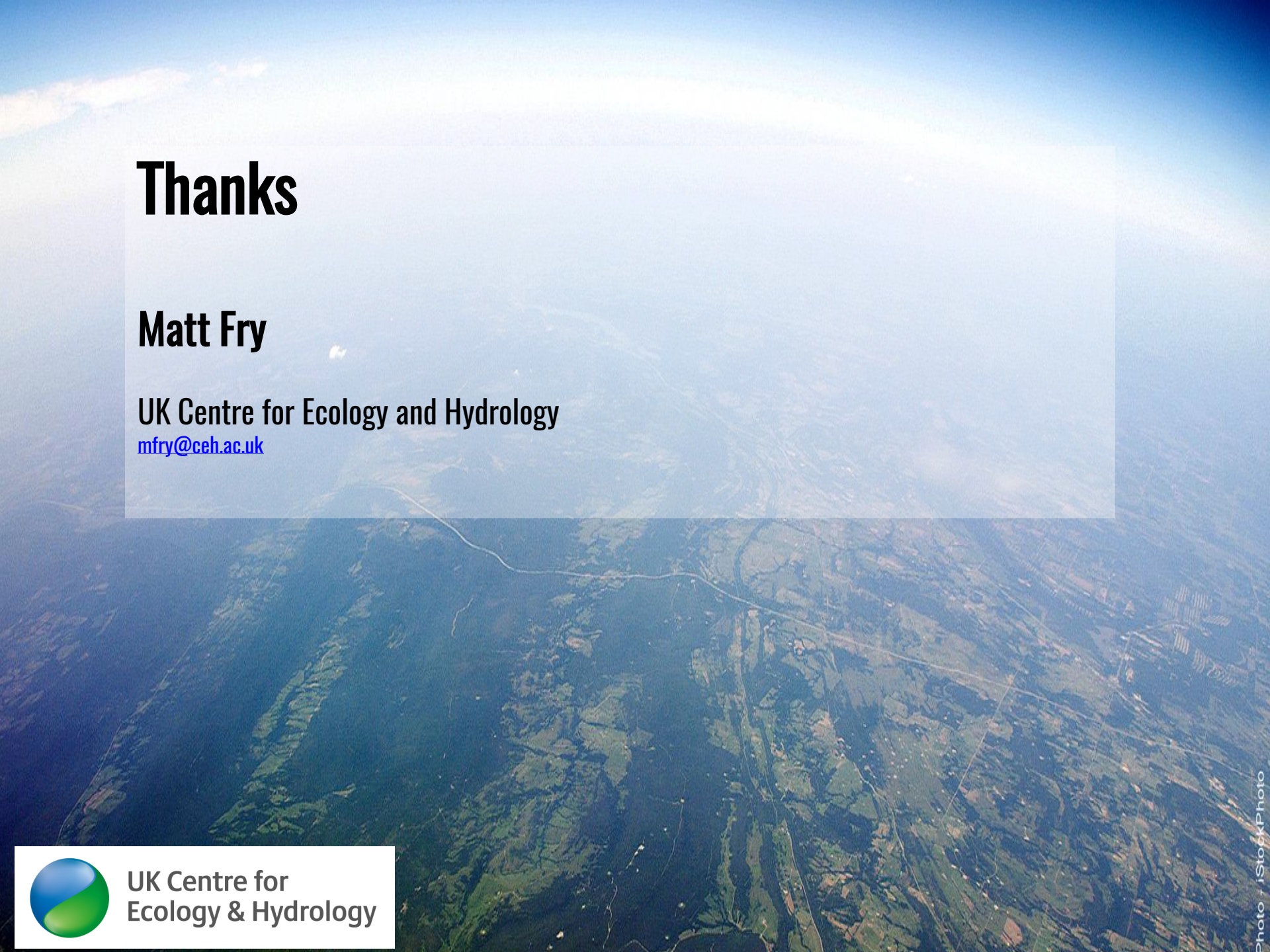




Thanks

Matt Fry

UK Centre for Ecology and Hydrology
mfry@ceh.ac.uk



UK Centre for Ecology & Hydrology