Can regular, long-term monitoring of lakes improve short-term machine learning forecasts for algal blooms?

Daniel Atton Beckmann
supervised by Dr Ian Jones, Dr Peter Hunter, Dr Evangelos Spyrakos

UNIVERSITY of STIRLING

Scottish Government gov.scot
Hydro Nation Scholars Programme

The need for short-term algal bloom forecasts:

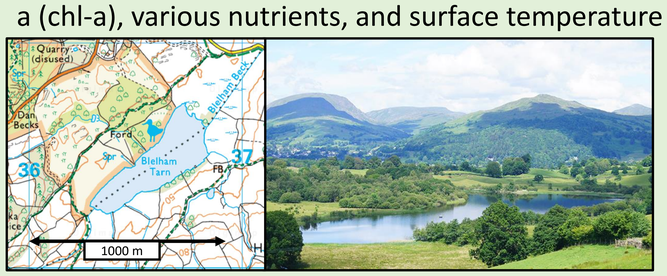
- Some algae (e.g. blue-green / cyanobacteria) can produce toxins harmful to human health and ecosystems
- Allow acting early to prevent drinking water supply issues
- Give warning to wild swimmers and dog walkers
- Further our understanding of aquatic ecosystems

The Research Question:

- Machine learning (ML) models have the potential to be used to forecast algal blooms in the short-term
- ML approaches typically require large training data-sets, which may mean that regular monitoring is required for many years to create reliable forecasts
- Here, we ask the questions:
  - **Does increasing the number of training years available always improve performance?**
  - **Do long-term changes in the behavior of algal blooms make older data less useful for training ML models?**

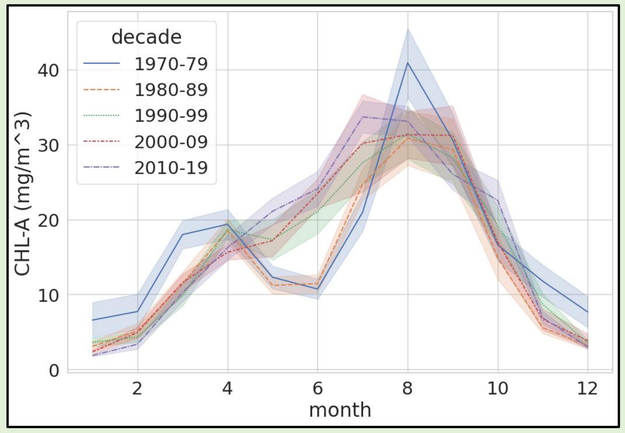
Study Site

- **Blelham Tarn,** English Lake District
- Eutrophic, suffers blue-green algal blooms in summer
- Sampled fortnightly by CEH (previously FBA)
- For this study, we use 47 years data, including chlorophyll-a (chl-a), various nutrients, and surface temperature



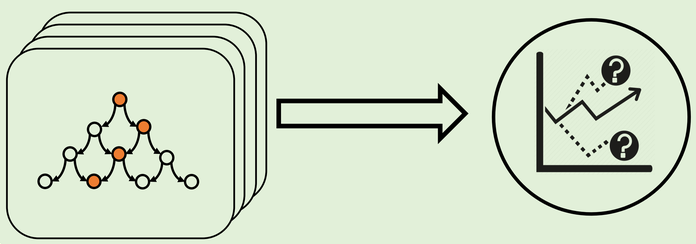
Long-term bloom history

- Decadal trends in the annual chl-a cycle suggest that in the 1970/80s there were clearly defined separate spring and summer blooms
- In more recent years, the spring and summer blooms are less separable,
- This change may suggest that older data is less useful for training ML models to predict recent blooms



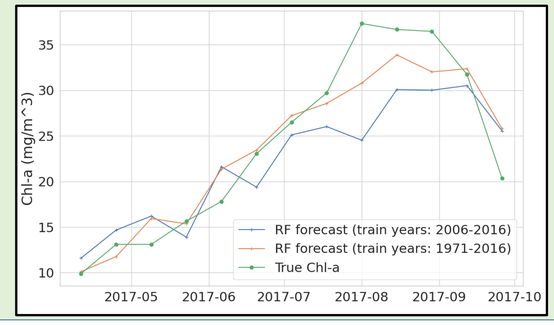
ML Model: *Random Forest (RF)*

- RF is a well-established ML approach which uses an ensemble of decision trees to make predictions
- **Input data:** all measurements from the last 4 weeks
- **Forecast:** Chl-a 2 weeks in the future



Forecasting Blooms in summer 2017

- A RF model trained with 45 years of data outperforms a RF model trained with 11 years data (as well as a persistence forecast benchmark model)
- The shape is good, but peaks in chl-a are consistently under-estimated



Does increasing the number of training years improve performance?

- Multiple RFs were trained on various sets of randomly selected training years
- **Increasing the number of training years generally improves performance**, but with significant variations between test years
- Further work required to investigate **some cases where adding additional training years reduces performance**