HABNet: Spatio-Temporal Based Machine Learning for Harmful Algal Bloom Detection

Dr P.R. Hill
The University of Bristol UK. e-mail: (paul.hill@bristol.ac.uk)

BRISTOL VISION INSTITUTE

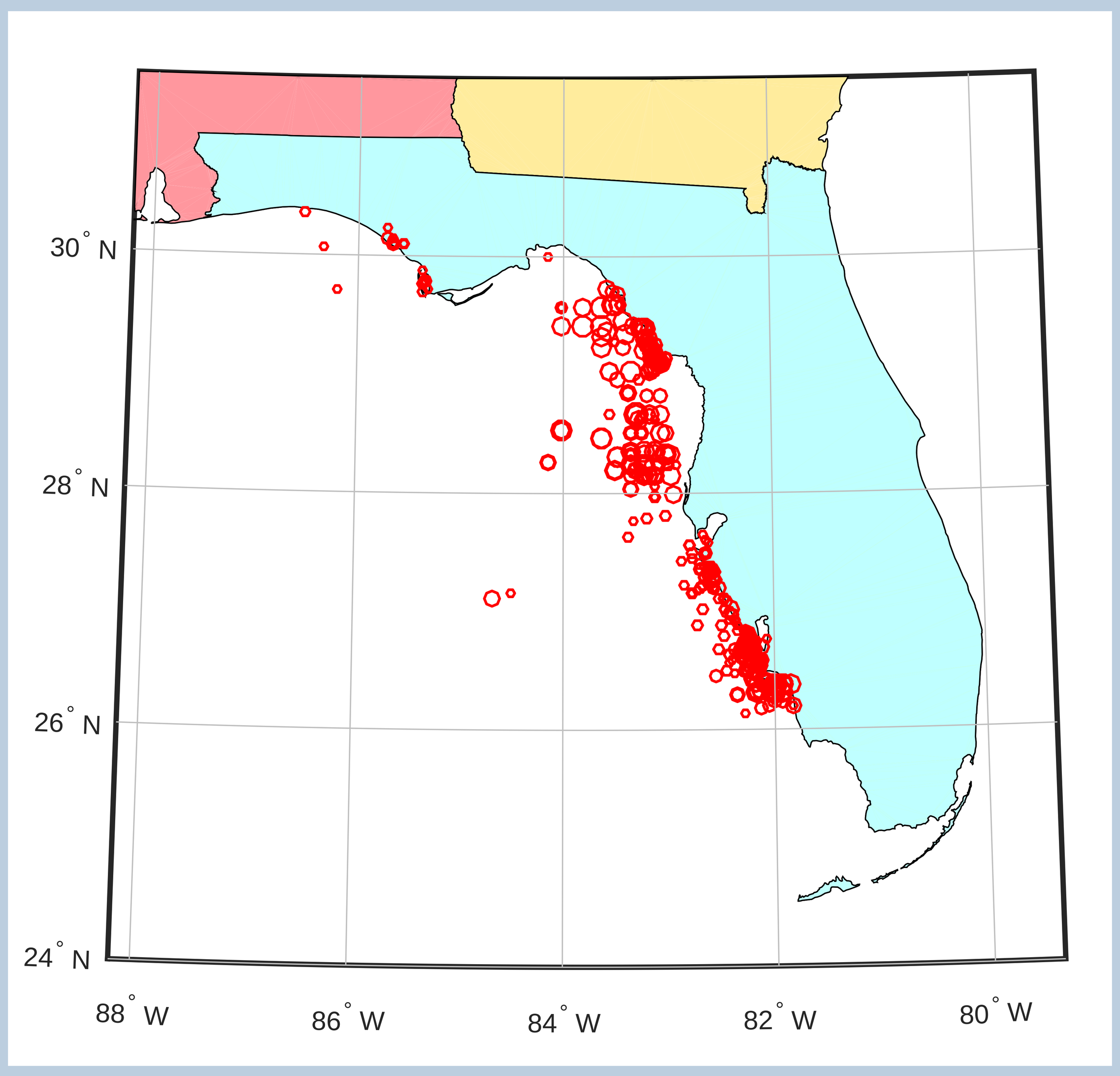
University of BRISTOL

Introduction / Contributions

- Application of machine learning techniques to develop a state of the art classifier and predictor of Harmful Algal Bloom events (HABs).
- HABs cause a large variety of human health and environmental issues together with associated economic impacts.
- HAB Detection system based on: a ground truth historical record of HAB events, a novel spatiotemporal datacube representation of each event (from MODIS-Aqua, MODIS-Terra and GEBCO bathymetry data) and a variety of machine learning architectures.
- ML tools include Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) components.
- This work has focused specifically on the case study of the detection of Karenia Brevis Algae HAB events within the coastal waters of Florida (over 5000 events from 2003 to 2018).

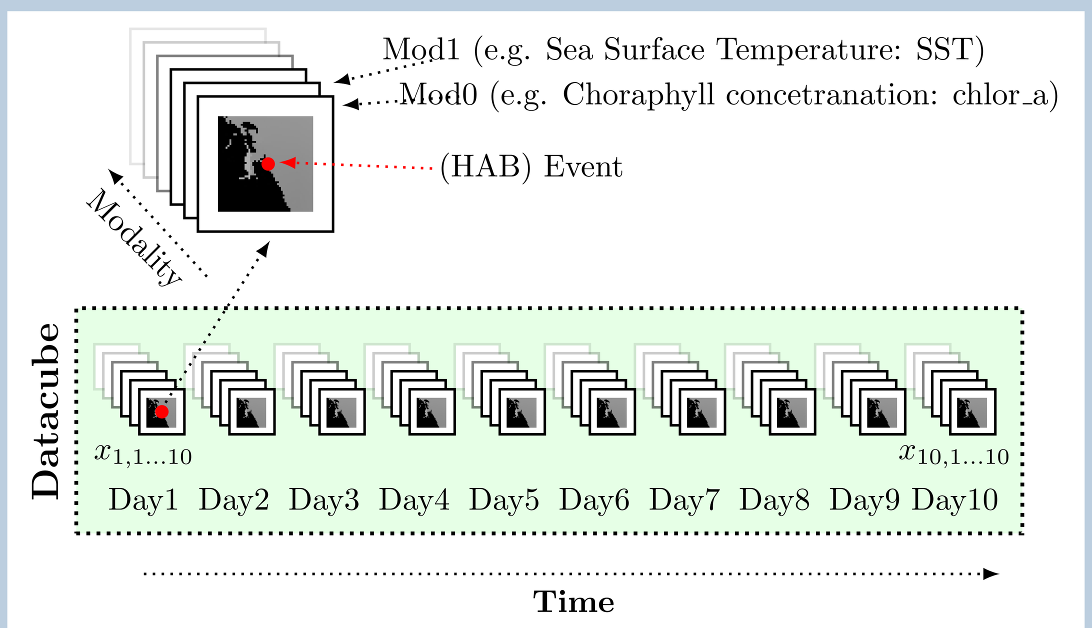
HAB Ground Truth Events

- The proposed HAB detection system uses a supervised machine learning method. Supervised machine learning requires a detailed ground truth dataset i.e. labelled positive and negative HAB events defined in time and location together with remote sensing data.
- The data is from the Florida Fish and Wildlife Conservation Commission (FWC) [1]. This dataset is extremely large (of both positive and negative HAB events) together with spanning the dates between 2001 and 2019.
- Only K. brevis algae events were extracted from this dataset in order provide a tractable solution (K. brevis is considered to be the most serious cause of HAB events).
- In order to further reduce the size of the dataset an HAB event was considered to have occurred when the event count algae abundance in cells/litre is in excess of 50,000. This is chosen as it was the threshold used in previous work by Lamp et al. [10]. The selection of K. brevis events and the 50,000 threshold led to the number of positive events being 2,768 (between 2003 and 2018). Negative events were selected from the entire dataset where the algae count in cells/litre were 0.



Spatial distribution of a selection of these positive events (circle size reflecting the count).

Datacube Structure



Structure of a datacube used in this paper

Algorithm 1: Creation of ML Datacube

```
Input :Groundtruth File
for ∀ HAB events in Groundtruth File do
    Extract HAB event Lat, Lon, Date Window
    for ∀ List of Modalities do
        Generate list of granules using NASA CMR
          search (within 10 days previously of HAB event date)
        for ∀ NetCDF Granules in Date Range do
            wget NetCDF Granule
            Extract modality data in spatial window
            Place cropped data in output Datacube
        end
    end
    Output Datacube
    :
end
```

Algorithm to generate the datacubes

Selected Modalities

MODIS-Aqua Products
- **chlor_a**: Estimated Chlorophyll
- **par**: MODIS Daily Mean Photosynthetically Available Radiation
- **SST** Estimated Sea Surface Temperature

MODIS-Aqua Remote sensing reflectance (Rrs) Bands (in nm)
- 412, 443, 488, 531, 555

MODIS-Terra Products
- **chlor_a**: Estimated Chlorophyll

Bathymetry
- **GEBCO from 500m grid [2]**

Machine Learning Structure

The index of the considered time sequence is denoted t where $\forall t \in \{1, 2, \ldots, T\}$ where in this case $T = 10$. The modality index is denoted m where $\forall m \in \{1, 2, \ldots, M\}$ where in this case $M = 10$. There are therefore 100 input images (10 modalities per each of the 10 time steps) per HAB event (each image denoted $x_{t,m}$). The concatenated outputs z_t of the CNNs are therefore created as follows.

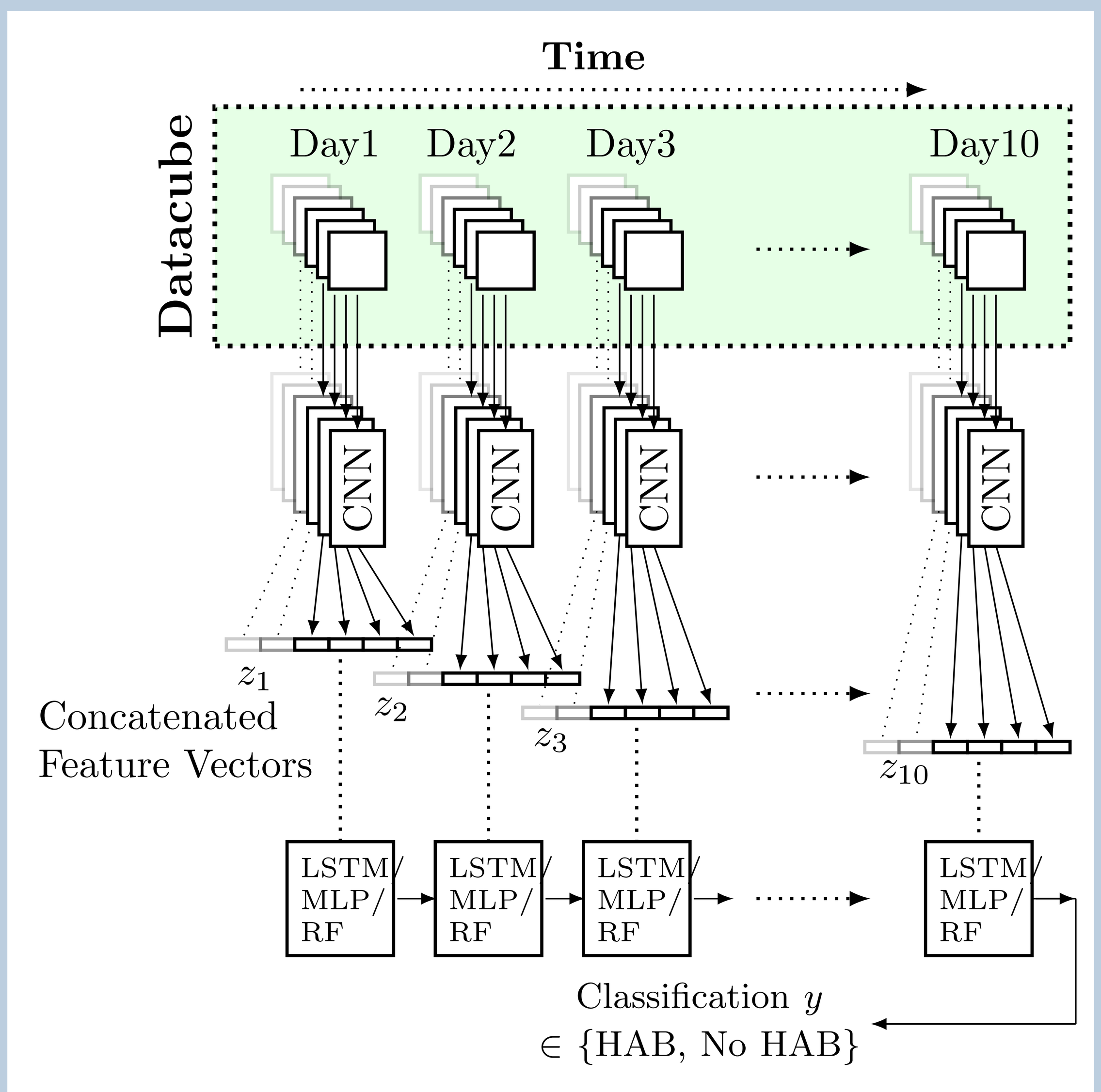
$$z_t = \{\phi(x_{t,1}), \phi(x_{t,2}), \ldots, \phi(x_{t,M})\}, \qquad (1)$$

where $\phi(\cdot)$ is the operation of the CNN that outputs the bottleneck 1D features before the softmax classification. The LSTM model (or equivalent) then takes as input all of the concatenated outputs z_t to generate the classification y where $y \in \{HAB, NoHAB\}$

$$y = \psi(\{z_1, z_2, \ldots, z_T\}) \qquad (2)$$

Where $\psi(\cdot)$ is the temporal classification operation (LSTM, MLP or Random Forest classifier) that outputs the HAB/No HAB classification.

Machine Learning Structure



Structure of Machine Learning system for datacube classification: CNN spatial characterisation followed by MLP, LSTM or Random Forest (RF) time series classification.

Results

Random Forest: RF Standard python (`sklearn`) implementation of RF with grid search of best parameters using validation set

MLP1: Two dense layers each with 512 nodes. Each layer combined with batch normalisation

MLP2: Two dense layers each with 512 nodes. Each layer combined with dropout (0.5)

LSTM1: One LSTM layer and one dense layer each with 512 nodes. Each layer combined with batch normalisation

LSTM2: One LSTM layer and one dense layer each with 128 nodes. Each layer combined with dropout (0.5)

LSTM3: One LSTM layer and one dense layer each with 512 nodes. Each layer combined with batch normalisation and dropout (0.5)

Temporal Classifier (using NASNET:Mobile) (CNN first stage)	Classification Accuracy Performance
RF	81.7 %
MLP1	85.25 %
MLP2	85.52 %
LSTM1	87.34 %
LSTM2	89.25 %
LSTM3	89.70 %

Conclusions

- Flexible detection machine learning architecture utilitsing CNN and LSTM components
- Up to 89.7% classification performance for LSTM and NasNet CNN architecture
- Applicable structure for other remote sensing detection systems

References

[1] "Florida fish and wildlife conservation commission monitoring database," http://myfwc.com/research/redtide/

[2] "General Bathymetric Chart of the Oceans (GEBCO)," https://www.gebco.net.