Predictive analytics and data driven approach to algae bloom prediction

Kavya Kompella[1](kavya10@ksu.edu); Collaborators: Dr. Lior Shamir[1], Dr. Trisha Moore[2], Dr. Aleksey Sheshukov[2], Dr. Daniel Flippo[2], Laura J. Krueger[2]

[1]Department of Computer Science, [2]Department of Biological and Agricultural Engineering, Kansas State University, Manhattan, KS

Background

Algae growth is a matter of major concern for water preservation. Toxic algae can contaminate water reservoirs, making them harmful for agriculture, household, and wildlife purposes. In Kansas, recurring blooms of toxin-forming cyanobacteria species raise health and ecological concerns. A possible solution to avoid or delay algae growth is to apply designated chemicals which has its own drawbacks. Additional research is needed to understand and predict blooms in freshwater reservoirs.

Research Objectives

- Develop a machine learning-based model using high frequency monitoring data to predict algal blooms before its growth escalates
- Identify most influential variables in bloom prediction
- Inform measures to prevent and mitigate blooms

Data Description

- In-lake water quality data: A multi-parameter sonde is used to record water quality variables in Marion Reservoir (Fig. 3), including phycocyanin (a pigment-based indicator of cyanobacteria), chlorophyll-a, dissolved oxygen, turbidity, specific conductivity, water temperature and pH.

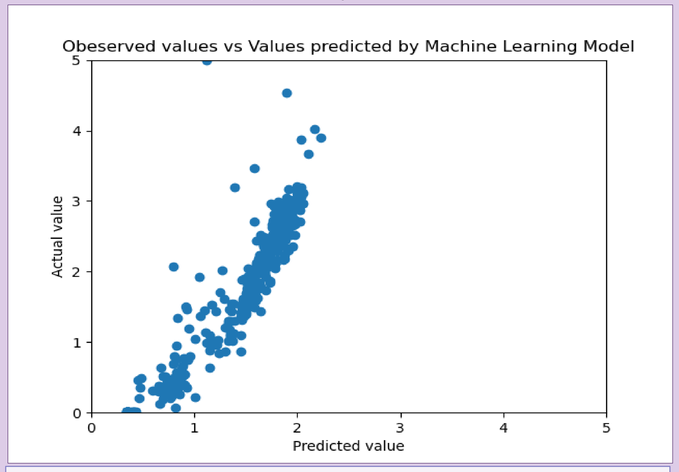
- Environmental data: Reservoir and meteorological data include precipitation, air temperature, wind direction and speed, relative humidity, solar radiation, pool elevation, reservoir storage, inflow, and outflows. These data are collected at the Marion Reservoir dam by the USACE (US Army Corps of Engineers).

Machine Learning Models

- **Ensemble Machine Learning Model**- This model is an ensemble of Random Forest, Decision Tree, KNeighbors, GaussianProcess, and SupportVector. This is the most efficient model so far for the data.
- **Long Short-term Memory (LSTM) Model**- A type of Recurrent Neural Network which is capable of learning long-term dependencies.
- As a part of the project, made future predictions of Phycocyanin and determined the features influencing the prediction.

Results

Preliminary model tends to underpredict at high phycocyanin levels, but captures overall trends



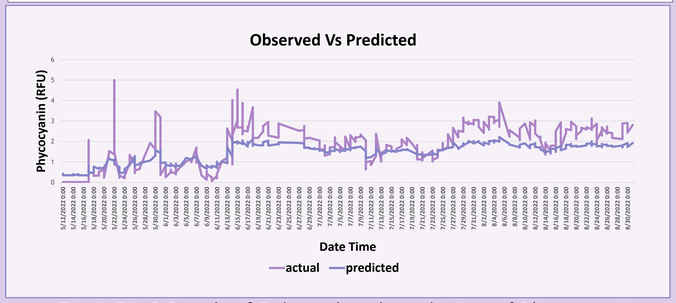


Fig.1, Fig.2 Result of 9 days ahead prediction of Phycocyanin (RFU) by the Ensemble Model

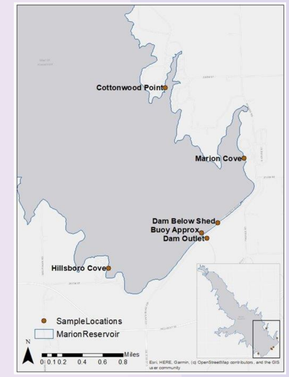


Fig. 3. Picture of cyanobacteria bloom in Marion reservoir study site.
Picture and map provided by Laura Krueger

Observations

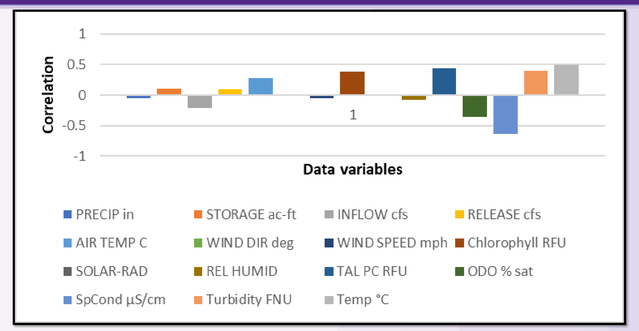


Fig.4. Pearson between variables of data and observed Phycocyanin (TAL PC)

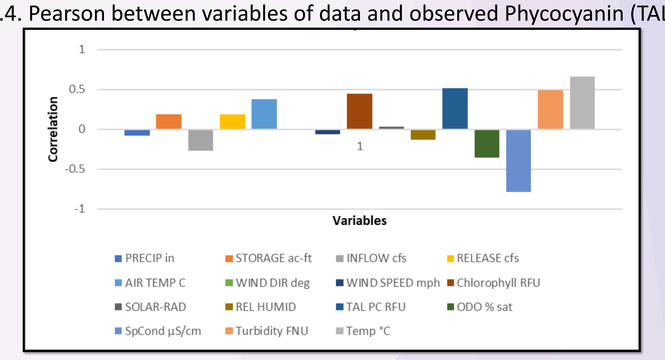


Fig.5. Pearson between variables of data and predicted Phycocyanin (TAL PC)

- The contribution of each tested variable are comparable in both observed and model predicted Phycocyanin
- Temperature, Turbidity, Specific Conductivity , Dissolved oxygen, Chlorophyll, Air Temperature, and Drainage Basin Inflow had good correlations with the predictions.

References

- https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Future Research

- Improve the efficiency of LSTM model
- To ensemble Machine Learning ensemble model with Long short-term memory model.