Model Comparison for Dauphin Island Station, Mobile Bay, AL

Antony, Aravind, Katy, Kevin

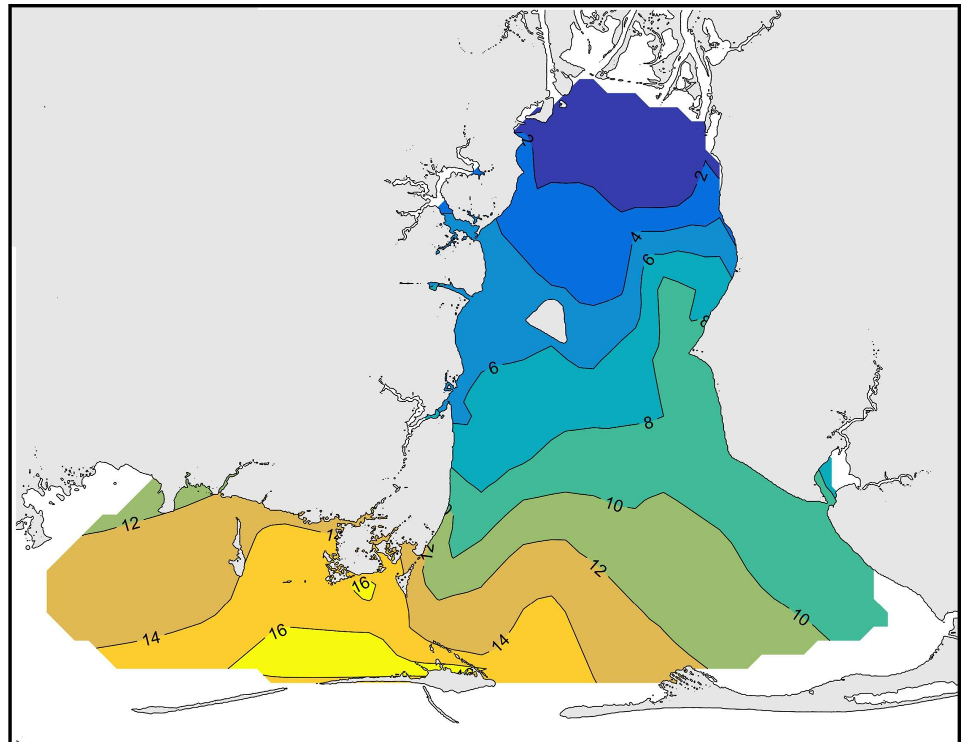
Project Goals

1. Identify independent and dependent variables with an emphasis on physical processes at Dauphin Island in Mobile Bay

2. Collect and clean necessary data using Python and Google Colab

3. Make a collaborative Github Repository for data and code sharing

4. Compare machine learning and deep learning models to see which method is more accurate for our selected dataset

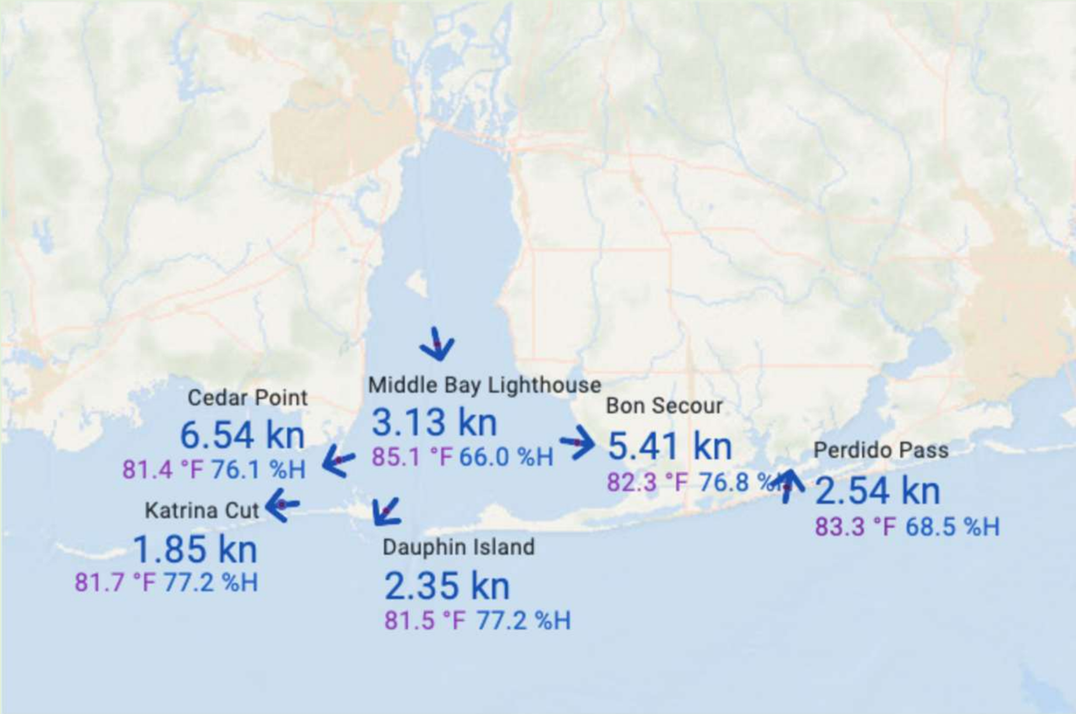


Mobile Bay Survey June 25-26, 2019
Surface Salinity Map (PSU)
University of South Alabama & Dauphin Island Sea Lab
Funded by USA Center for Environmental Resilience

Data background – ARCOS Data

- Dauphin Island, Mobile Bay, Alabama
- Why did we pick this station?
  - Minimal missing data over a long period of time (2003 – 2022)
  - Good location at exchange point with the estuary and Gulf of Mexico
- Observational station data – downloaded as .csv files
- ARCOS data – Alabama's Real Time Coastal Observing System at Dauphin Island Sea Lab
- Dependent Variables:

  Temperature, Salinity, Dissolved Oxygen



Data Background – ERA5 Reanalysis Data

- ERA5 -- Climate reanalysis that combine past observations with models using data assimilation technique to generate consistent time series of multiple climate variables across a grid.

1. Download as NetCDF format
2. Extract from the nearest location to Dauphin Station
3. Clean and converted to a central .csv file

- Independent Variables:
  - Atmospheric temperature, wind components (u+v) and speed, solar radiation (net short wave (ssr), latent heat (slhf), long wave radiation (str), sensible heat (sshf))
  - River discharge (USGS) – Tombigbee River, AL – we are using this river discharge data – key upstream river, main source of discharge into Mobile Bay

Data Background – River Discharge and Tides

Station DILA1 - 8735180- Dauphin Island Station

1. NOAA Tides and Current
2. Downloaded as .csv file
3. Cleaned and added to central .csv file

River discharge (USGS) – Tombigbee River, AL – we are using this river discharge data – key upstream river, main source of discharge into Mobile Bay

- Independent Variables:
  - Tidal elevation (WL) and river discharge (Q)

Methods

1. Data consolidation and cleaning in Google Colab - timestep uniform (monthly average), one primary .csv file for all data

   2010 – 2020 10 years total data-
   9 years for training and 2 years for testing
   Monthly average for each independent and dependent variable

2. Make a GitHub repository Hackathon2023 - share access with all four members as collaborators

   Upload all datasets and codes to this repository for collaboration

3. Develop models

   Deep learning model – Antony + Katy + Aravind
      13 total variables

   Machine learning model - Kevin + Antony
      Variable analysis specified by initial correlation analysis – highest correlation

4. Validate models - Comparison of predicted data versus observed data

   Compare model use for selected data

5. Determine next steps

Deep Learning Model Results

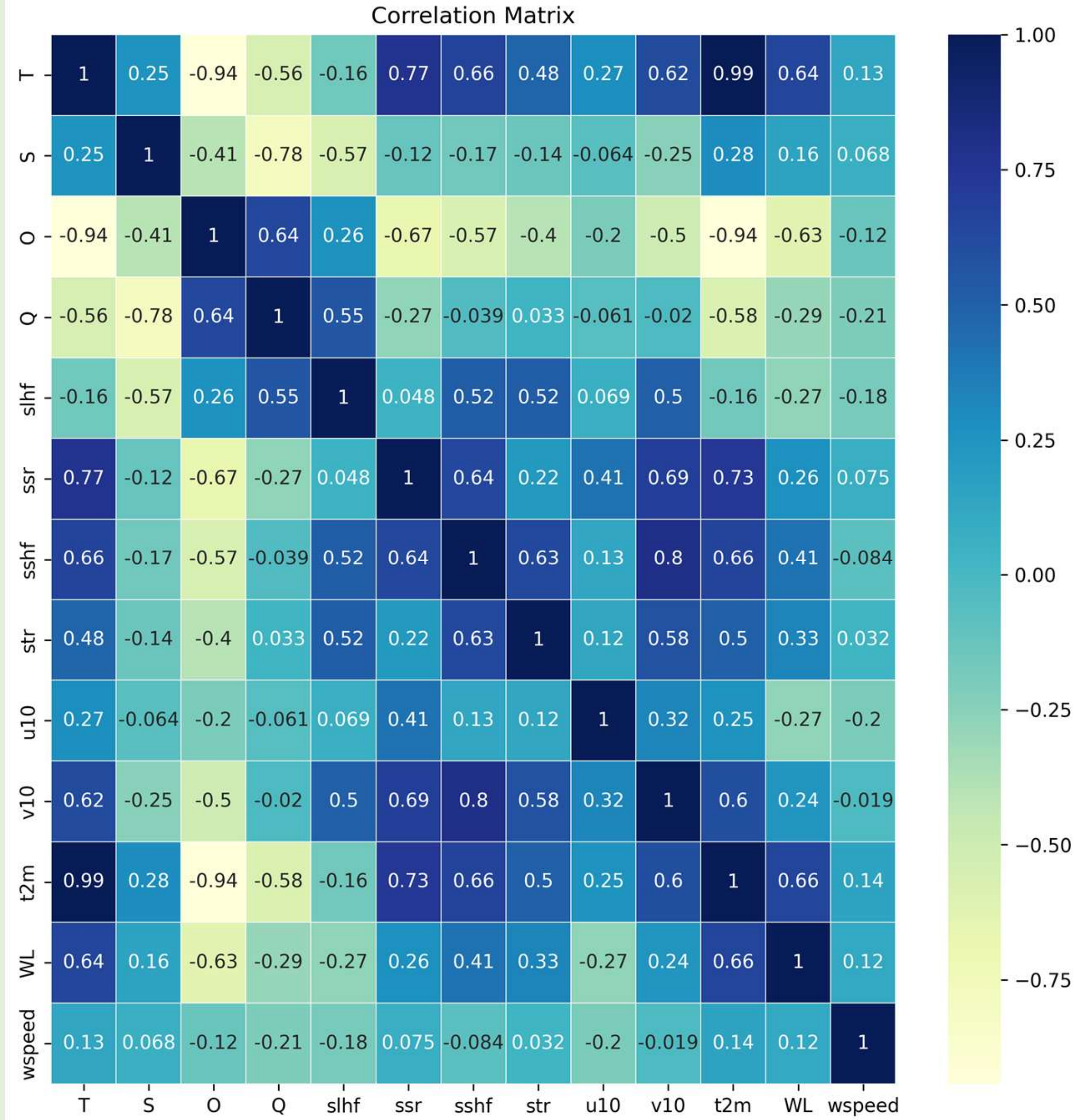
Deep Learning Model Development

- No separation of independent and dependent
- Train and test split – 9/2 years
  - Feature scaling
  - databatch = 4
  - 3- layer LSTM model (64, 8, 4)
  - 13 nodes for 13 features
  - 2000 epochs – long runtime – 5-10 mins, higher for more epochs
  - batch_size=18 (changed range outcome for predicted data v. large batch size)
- Made predictions for 2019-2020 & conduct RMSE comparison
- Comparison of predicted and observed for three main dependent variables T, S, DO

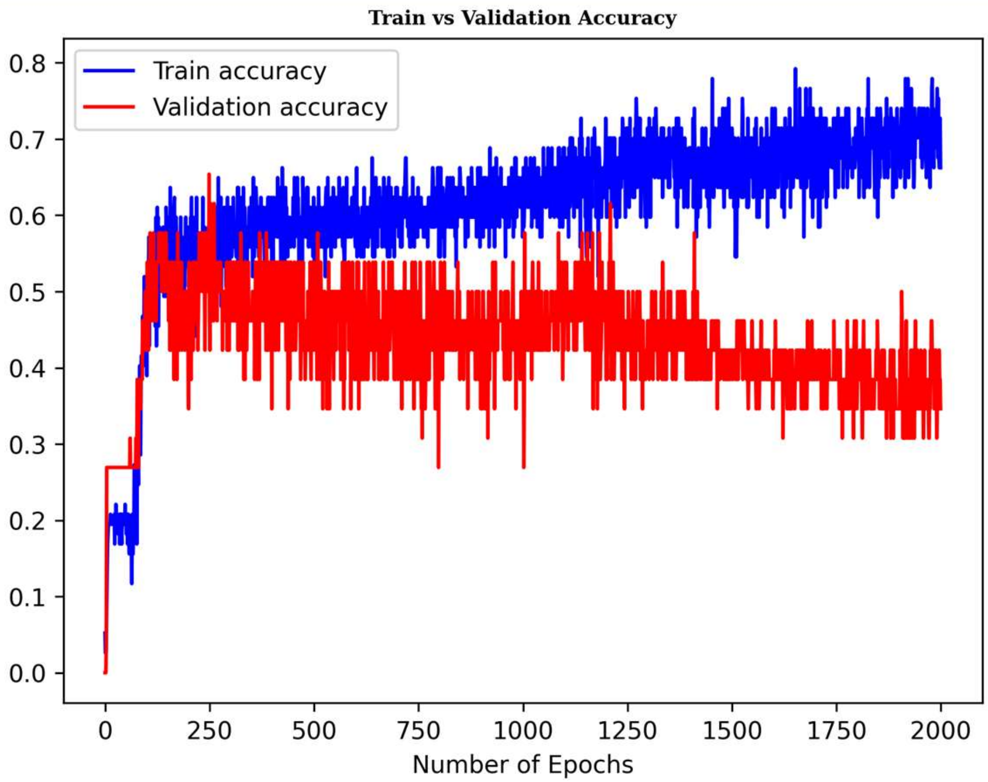
Correlation table to determine variables to incorporate in both models

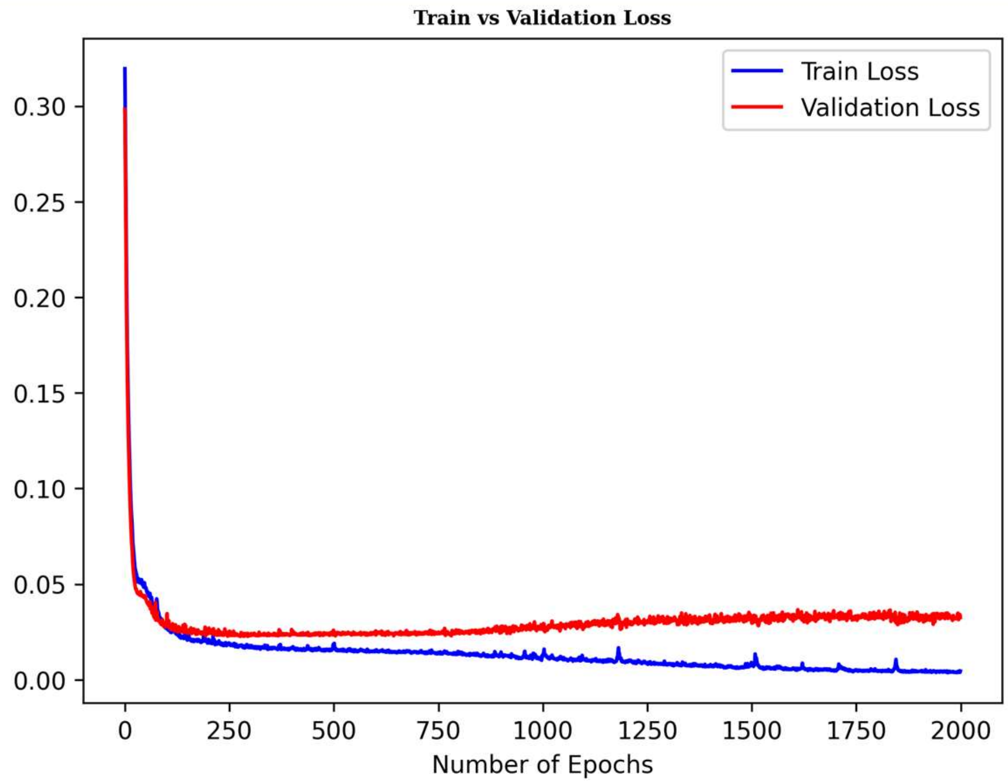
Machine Learning Model

1. Air temp and water temp - positive
2. River discharge and salinity – negative
3. Air temperature and DO – negative
4. Latent Heat Flux and Salinity - negative

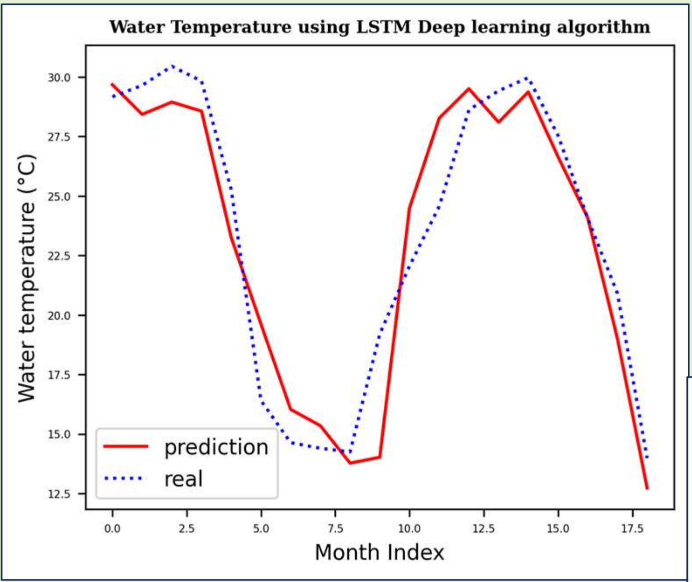


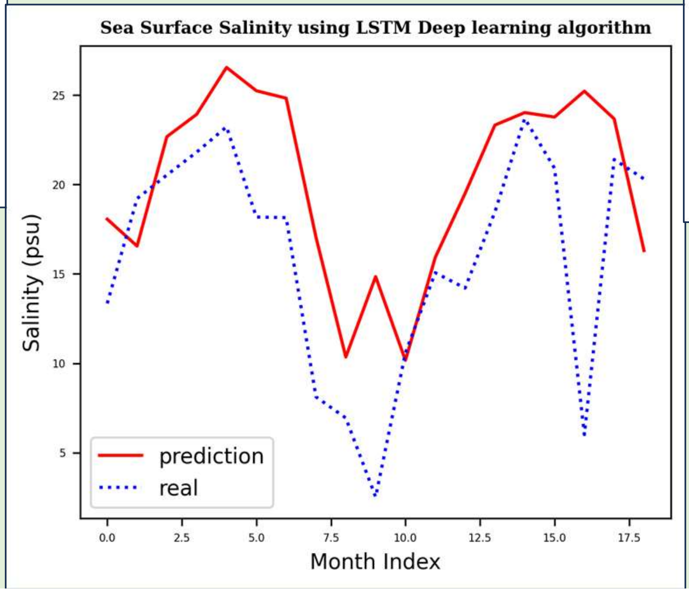
Model Performance
Train vs Validation Accuracy/Loss

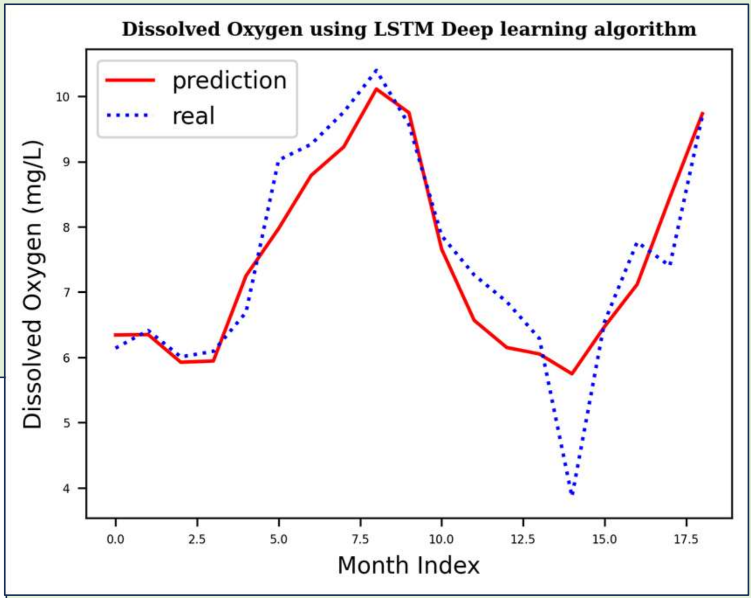




Predicted versus Observed Values







Machine Learning Model Results

Machine Learning Model Development

- Linear regression
  - Insert dummy column – linear regression did not work with one independant and one dependent , needed 2D array for independent variable inputs (one variable and one constant)

- Separate into independent and dependent – three pairs

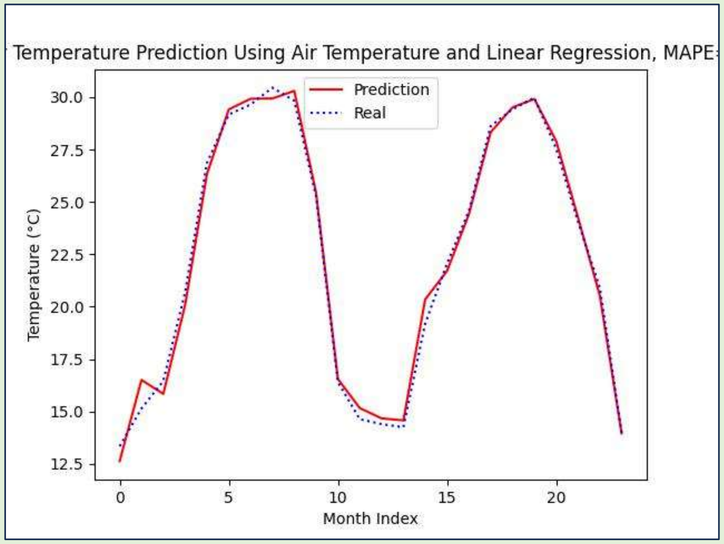
- Created training and testing data, tested model – 9/2 years

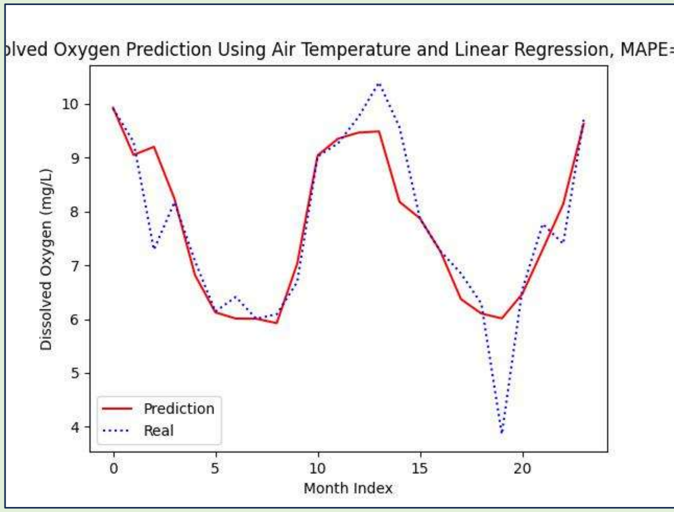
- Made predictions for 2019-2020

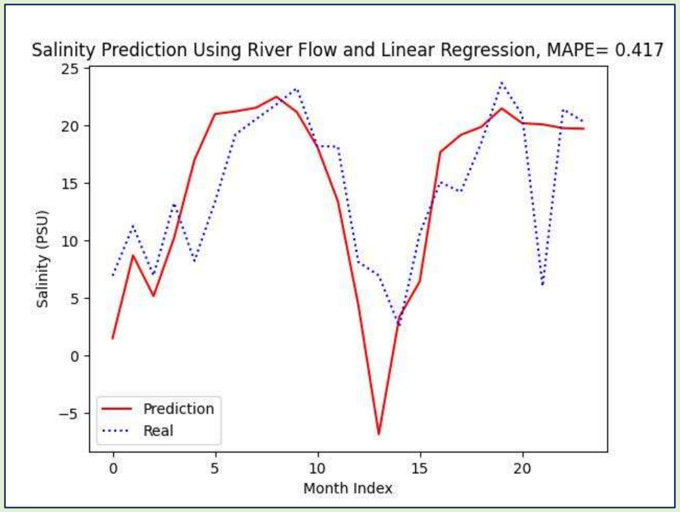
- Compared model prediction to to observational/testing data

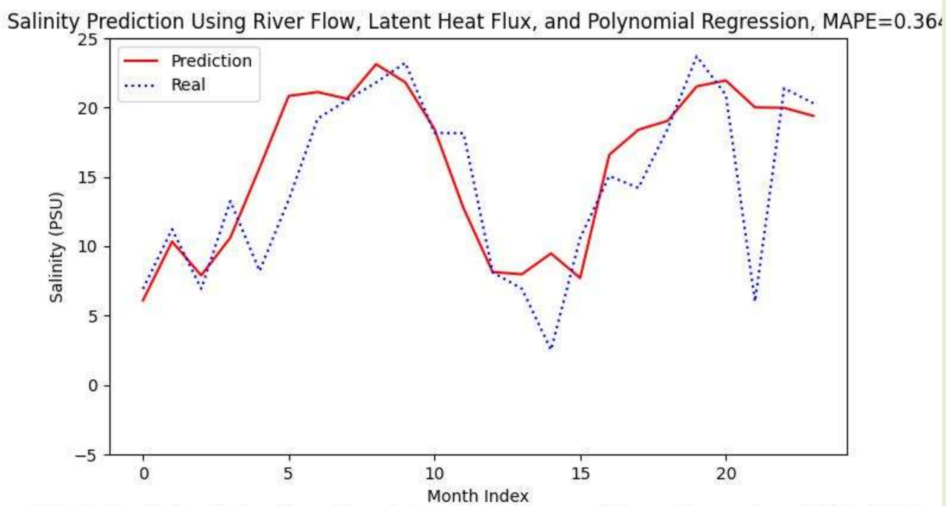
- Made MAPE calculations and graphs

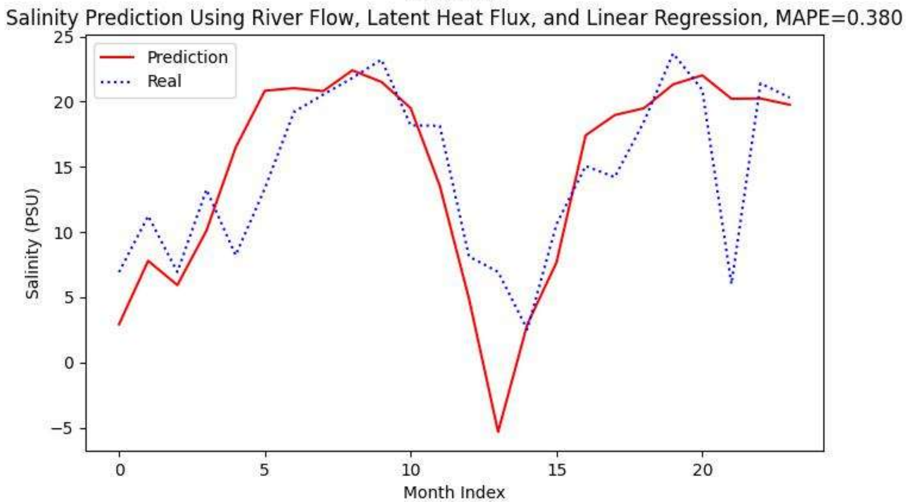
Machine Learning - Linear Regression Analyses







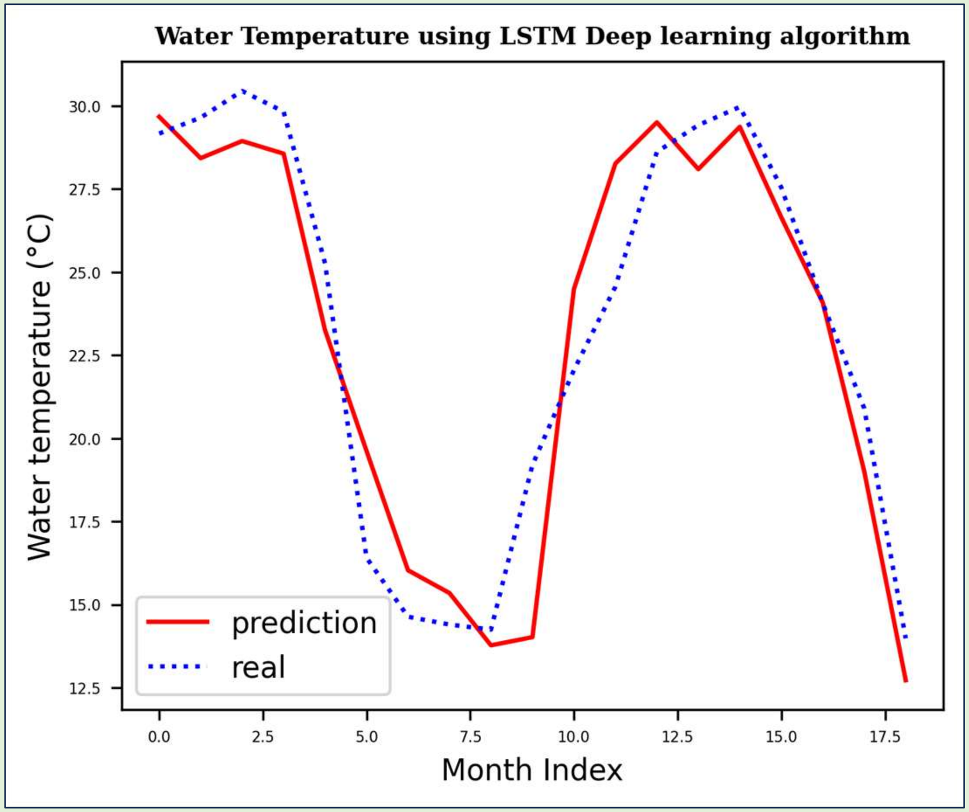


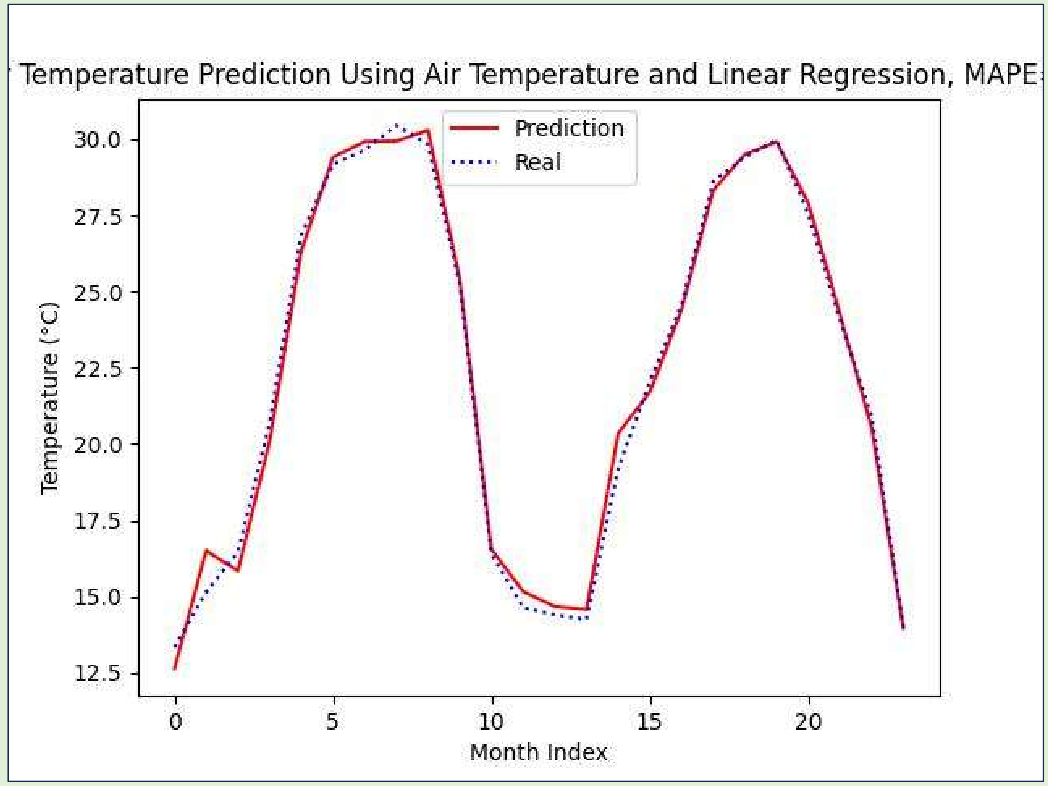


Linear versus Polynomial (2) Regression

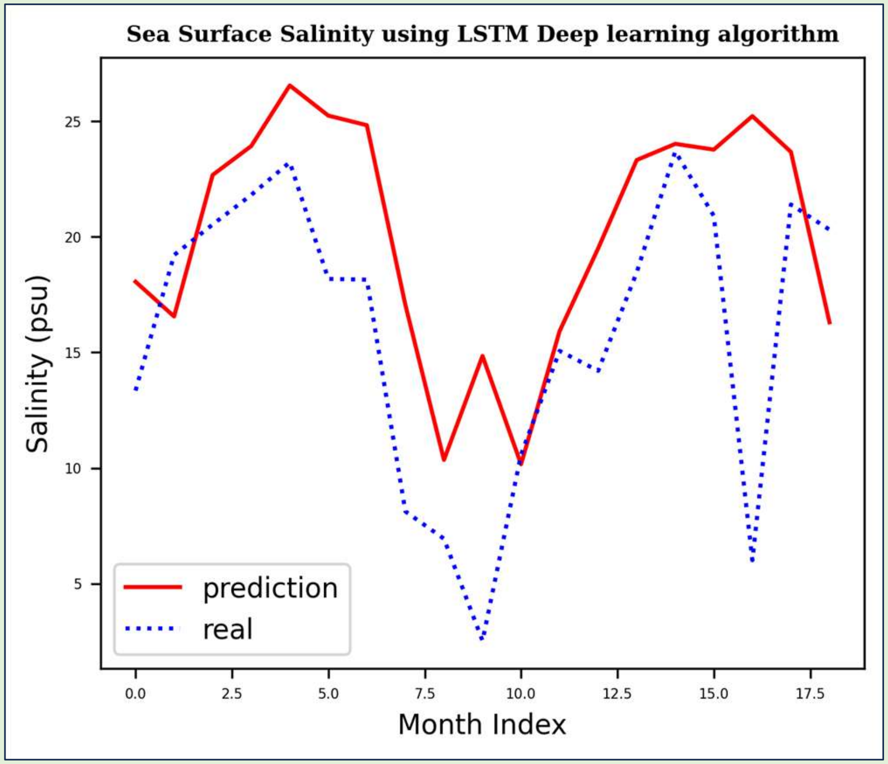
Conclusions

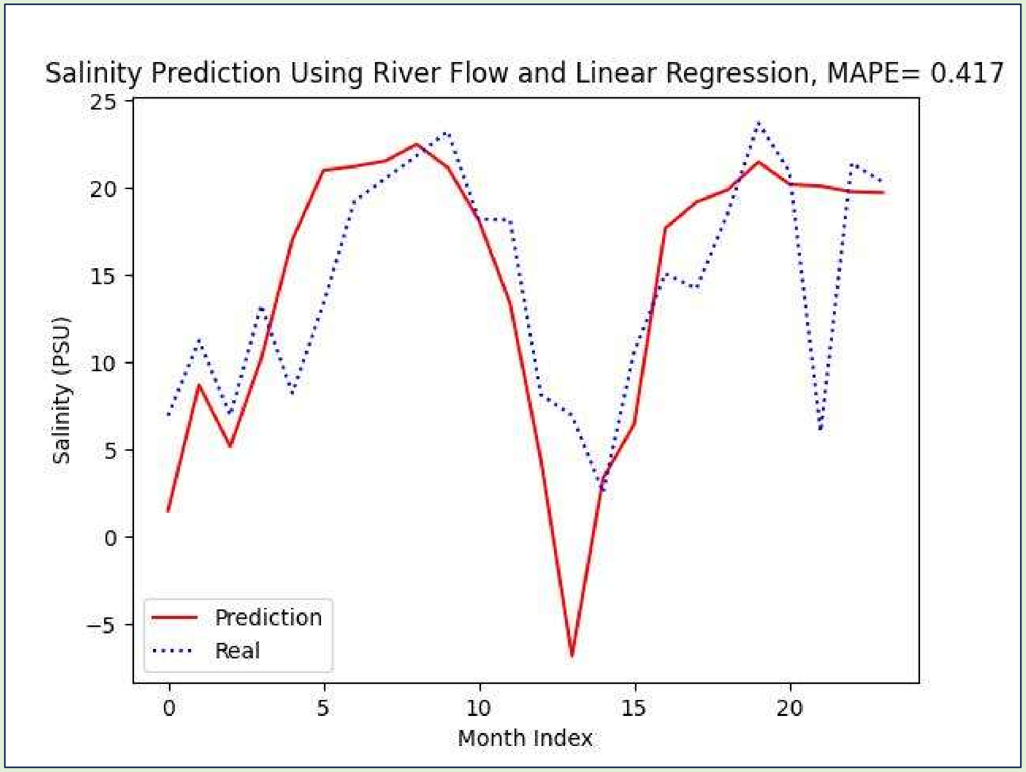
Machine Learning vs Machine Learning -- Temperature



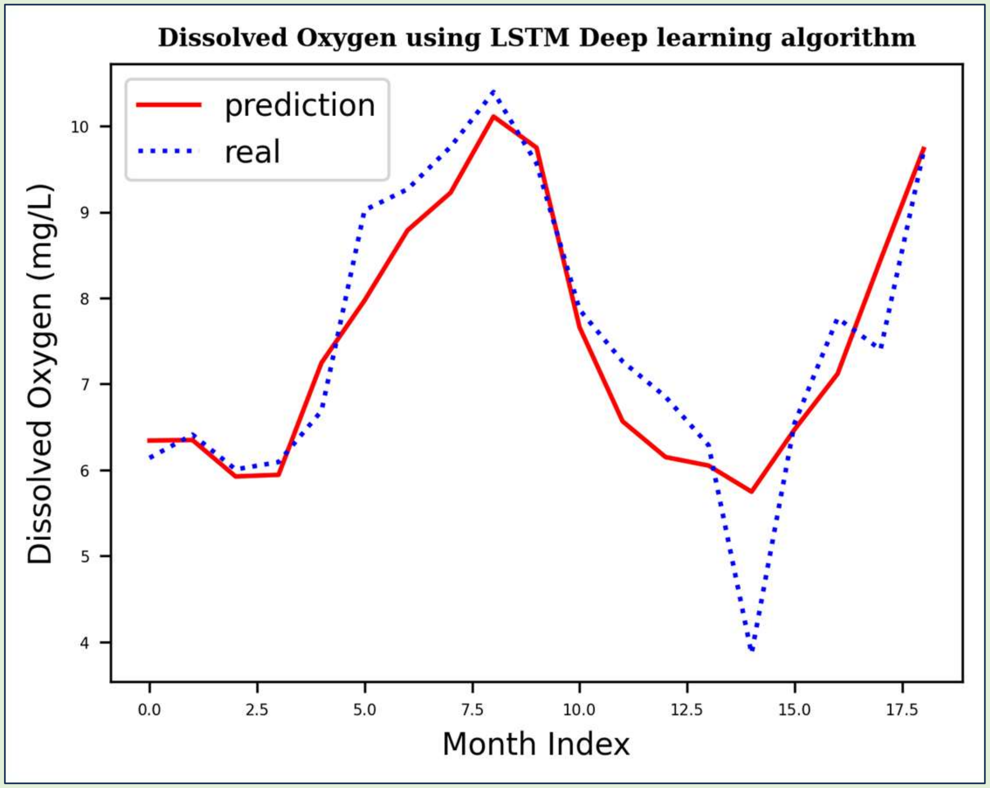


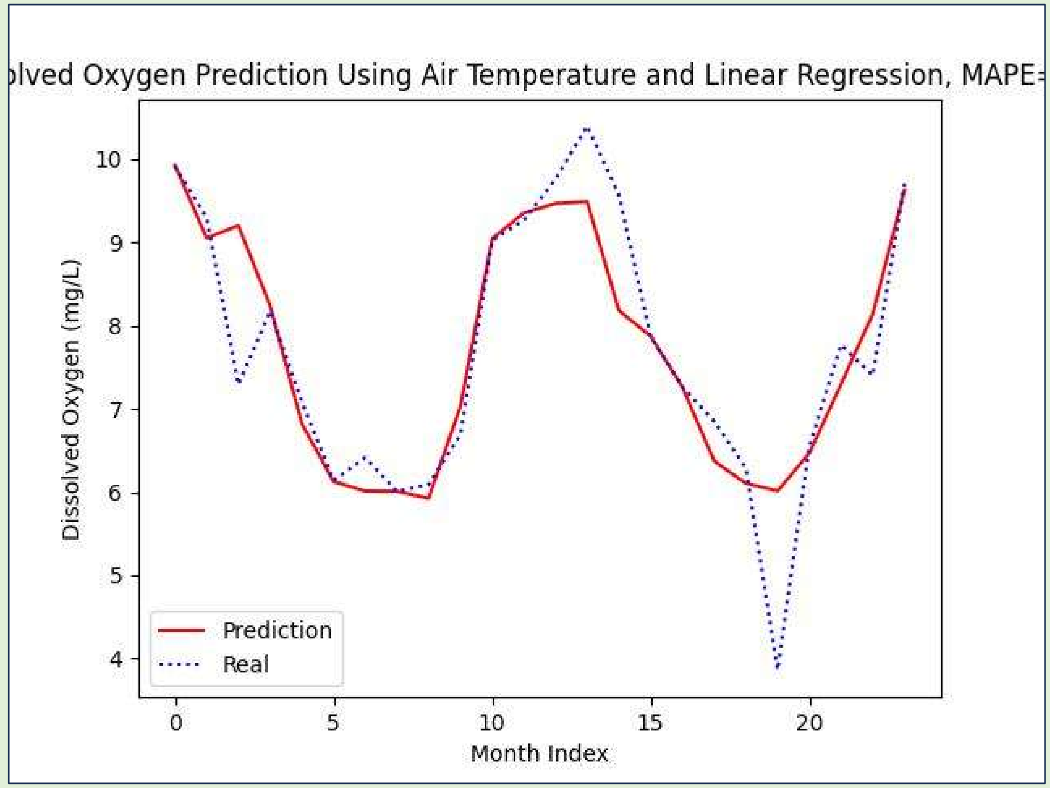
Machine Learning vs Machine Learning -- Temperature





Predicted versus Observed Values





Conclusions

- Troubleshooting – difference in output scale for DL and ML predicted values
- Overall, machine learning model was a better fit for our cleaned data than deep learning model
- Quantity of data – 12*24 -- monthly values were not enough

Next Steps – machine learning

- Testing more pairs with linear regression – only 4+ tested so far

  Multiple independent, one dependent
- Polynomial regression?
- Looking into other predictive models for machine learning

Next Steps – deep learning

- More develope deep learning model – daily averages instead of monthly averages
- Further manipulation of model type and settings
- Alternative

Next Steps – Data Analysis

- Daily averages instead of monthly averages for a better fit for our deep learning model
- Incorporate NDBC stations in the region for further development with ERA5 data
- Increase quantity of data
  - Incorporate other river discharges into our models
  - Comparison of more stations throughout Mobile Bay for further model validation

Questions?