**Project Topic: Comparative Analysis of Models for Predicting Coastal Water Levels in California**

*Hongxin Wu*

**What are we trying to solve?**

The project aims to develop several predictive models and compare their performances to forecast the water level on California's coastline over the upcoming days, precise to every hour for the next 24 hours, utilizing historical and current climate data. Additionally, the project will provide less precise predictions for the coming years, such as 2050, 2100, and 2200. These short-term, hourly forecasts will help coastal communities, emergency responders, and local authorities better prepare for and manage potential risks associated with rapidly changing water levels, such as coastal flooding, erosion, and navigation hazards. The long-term predictions, although less precise, will offer valuable insights into the potential future impacts of sea-level rise on California's coastline. By comparing the performance of different models, the project seeks to identify the most accurate and reliable approach for predicting water levels along California's coastline, both in the short-term and long-term.

**Why is this important?**

Accurate short-term forecasting of water levels along California's coastline is crucial for ensuring public safety, mitigating potential damage to coastal infrastructure, and managing marine activities. Hourly predictions for the next 24 hours enable coastal communities, emergency responders, and local authorities to take timely action in response to rapidly changing water levels, such as issuing flood warnings, coordinating evacuation efforts, and closing threatened beaches or harbors. Additionally, these forecasts are valuable for marine navigation, as they help ship captains and port authorities make informed decisions to ensure safe passage and docking procedures. The long-term predictions, while less precise, are essential for understanding the potential future impacts of sea-level rise on California's coastline. These projections inform long-term planning decisions, such as coastal development, infrastructure investments, and adaptation strategies, which are necessary for building resilience against the effects of climate change. By comparing the performance of different predictive models, this project contributes to the development of more accurate and reliable tools for forecasting water levels, ultimately benefiting coastal communities and decision-makers in California.

**Where will the data come from?**

Data will be sourced from:

* Satellite observations for sea surface height (e.g., from NASA's satellite missions like Jason-3).
* <https://podaac.jpl.nasa.gov/dataset/JASON_3_L2_OST_OGDR_GPS>
* Historical sea level records from tide gauges (e.g., NOAA's National Water Level Observation Network).
* <https://tidesandcurrents.noaa.gov/sltrends/>
* <https://tidesandcurrents.noaa.gov/tide_predictions.html>
* <https://tidesandcurrents.noaa.gov/stations.html?type=Water+Levels>
* Climate models and projections (e.g., from the IPCC's Assessment Reports).
* <https://www.ipcc.ch/site/assets/uploads/2018/02/ar4-wg1-chapter8-1.pdf>
* <https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_Chapter09_FINAL.pdf>
* Ocean Surface Topography from Space
* <https://sealevel.jpl.nasa.gov/data/get-data/>
* <https://sealevel.jpl.nasa.gov/resources/?page=0&per_page=25&order=pub_date+desc&search=&condition_1=1%3Ais_in_resource_list&category=211>
* <https://www.epa.gov/sites/default/files/2021-04/documents/sea-level_td.pdf>

**How will it need to be refined?**

The data refinement process will include the following:

* Cleaning: Removing inaccurate or irrelevant data points.
* Normalization: Standardizing the scale of different datasets to enable meaningful comparisons.
* Interpolation: Filling in missing values in historical records to maintain consistency.
* Outlier Detection: Identifying and handling statistical outliers to improve model accuracy.

**What method will be used?**

The project will implement a multi-model approach to forecasting water levels along California's coastline:

1. **Time Series Analysis with ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal ARIMA):** These methods will be used to analyze linear trends and capture seasonal patterns in the water level data. ARIMA is suitable for datasets showing stationarity or where linear trends have been identified, while SARIMA extends the capabilities of ARIMA to handle seasonal fluctuations in the data.
2. **Machine Learning with Random Forest and Gradient Boosting Machines (GBM):** These ensemble learning methods will be employed to capture complex, non-linear relationships in the data and handle a wide range of predictors. Random Forest combines multiple decision trees to improve prediction accuracy and reduce overfitting, while GBM sequentially builds weak prediction models and combines them to create a strong, accurate model.
3. **Deep Learning with LSTM (Long Short-Term Memory) Networks:** LSTM networks, a type of recurrent neural network, will be utilized for their ability to process sequences of data and capture long-term dependencies. This method is especially useful for integrating various predictors of water levels, such as tidal patterns, wind speed, and atmospheric pressure, and learning from past sequences to make accurate predictions.

The choice of these diverse models allows for a comprehensive analysis, leveraging the strengths of each approach. ARIMA and SARIMA provide a solid foundation for handling time series data and seasonal trends, while Random Forest and GBM offer the flexibility to capture complex relationships and handle a variety of predictors. LSTM networks bring the power of deep learning to learn from sequential data and make accurate predictions based on long-term dependencies. By comparing and combining the results from these models, the project aims to develop a robust and reliable system for forecasting water levels along California's coastline.

**What metrics will be used to evaluate success?** The model's performance will be evaluated using the following metrics:

* **Root Mean Square Error (RMSE):** Measures the model's prediction accuracy by quantifying the square root of the average squared differences between predicted and actual values.
* **Mean Absolute Error (MAE):** Provides an understanding of the average magnitude of the prediction errors, which is easy to interpret.
* **R-squared (Coefficient of Determination):** Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, useful for assessing the goodness of fit for linear models like ARIMA.

Additionally, model validation will involve comparing the forecasts against observed water level changes not used in model training and against projections from authoritative sources like the IPCC and NOAA. This step ensures that the model's predictions are realistic and aligned with current scientific understanding.

**Contribution and Impact**

The project's success is primarily determined by its ability to provide accurate and reliable predictions of water levels along California's coastline, which can have significant implications for coastal communities, emergency responders, and local authorities. By developing a robust multi-model approach that leverages the strengths of ARIMA, SARIMA, Random Forest, GBM, and LSTM, this project contributes to the advancement of water level forecasting techniques and demonstrates the potential of combining traditional time series analysis with machine learning and deep learning methods.

The project's impact extends beyond its immediate application, as it serves as a valuable case study for researchers and practitioners working on similar challenges in other coastal regions. By documenting the methodology, comparing the performance of different models, and sharing insights gained throughout the project, this work contributes to the growing body of knowledge on coastal water level prediction and adaptation strategies. The project's findings can inform future research, guide the development of similar forecasting systems in other locations, and support decision-making processes related to coastal management and resilience planning.

Moreover, the project demonstrates a strong commitment to sustainable AI practices by embedding CodeCarbon to document the carbon cost of the project, from data processing to model training and evaluation. This aspect highlights the environmental impact of computational methods used in coastal water level research and contributes to a more comprehensive understanding of the trade-offs involved in deploying DS/ML/AI solutions. By transparently reporting the carbon footprint of the project, this work sets an example for responsible and sustainable research practices in the field.

Through its innovative approach, the potential for knowledge transfer, and dedication to sustainable AI practices, this project not only aims to provide actionable insights into coastal water level fluctuations but also to advance the methodologies and best practices in the field while promoting sustainability in scientific research.