Comparative Analysis of Models for Predicting Coastal Water Levels in California

Hongxin Wu*
Graduate School of Arts and Sciences
Georgetown University
Washington, DC 20057
hw487@georgetown.edu

Abstract

Accurate prediction of coastal water levels is crucial for ensuring public safety, mitigating potential damage to infrastructure, and managing marine activities in California. This project aims to develop and compare the performance of several predictive models, including time series analysis (ARIMA and SARIMA), machine learning (Random Forest and Gradient Boosting Machines), and deep learning (LSTM Networks), to forecast water levels along California's coastline. Utilizing historical and current climate data from NOAA's National Water Level Observation Network and other meteorological sources like satellite observations, tide gauge records, and climate models, the models provide hourly predictions for the next 24 hours. The results demonstrate the effectiveness of the multi-model approach in capturing both short-term and long-term water level variations, with Random Forest exhibiting the highest accuracy for hourly predictions. The project also emphasizes sustainable AI practices by documenting the carbon cost using CodeCarbon, setting an example for responsible research in the field. The findings contribute to the advancement of coastal water level prediction techniques and provide valuable insights for coastal communities and decision-makers in California.

1 Introduction

1.1 Background

Coastal communities worldwide are increasingly vulnerable to the impacts of climate change, with rising sea levels posing significant risks to human settlements, infrastructure, and ecosystems. In California, where a substantial portion of the population resides along the coast, accurate predictions of coastal water levels are essential for effective risk management and adaptation planning. Short-term forecasts enable emergency responders and local authorities to take timely action in response to rapidly changing water levels, such as issuing flood warnings and coordinating evacuation efforts. Long-term projections, on the other hand, inform strategic decisions related to coastal development, infrastructure investments, and resilience-building measures.

1.2 Problem Statement and Motivation

Existing coastal water level prediction methods often rely on a single modeling approach, which may not capture the complex interactions between various factors influencing water levels, such as tidal patterns, wind speed, and atmospheric pressure. Time series analysis methods, such as ARIMA and SARIMA, effectively handle linear trends and seasonal patterns but may struggle with non-linear relationships. Machine learning techniques, including Random Forest and Gradient Boosting

*https://github.com/IslaAshina/dsan5550-final-project.git

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Machines, can capture complex interactions but may not fully exploit the sequential nature of water level data. Deep learning methods, such as LSTM Networks, have shown promise in processing sequential data and learning long-term dependencies but have not been extensively explored in the context of coastal water level prediction in California.

Additionally, the lack of comparative studies evaluating different models' performance limits our understanding of the most suitable approach for specific coastal regions and time scales. Comparing multiple models and combining their strengths through ensemble methods could potentially improve the accuracy and reliability of coastal water level predictions. Furthermore, considering sustainable AI practices, such as documenting the carbon cost of the project, is crucial for promoting responsible research and mitigating the environmental impact of computational methods used in this field.

1.3 Research Objectives and Contributions

This project aims to address these limitations by:

- Developing a multi-model approach that combines time series analysis (ARIMA and SARIMA), machine learning (Random Forest and Gradient Boosting Machines), and deep learning (LSTM Networks) techniques to predict coastal water levels in California.
- Comparing the performance of these models in forecasting hourly water levels for the next 24 hours.
- Exploring the potential of ensemble methods to leverage the strengths of different models and improve prediction accuracy.
- Documenting the carbon cost of the project using CodeCarbon to promote sustainable AI practices in coastal water level research.

The project's contributions include:

- Advancing the understanding of the strengths and weaknesses of different modeling approaches for coastal water level prediction in California.
- Providing a framework for combining multiple models to improve the accuracy and reliability of water level forecasts.
- Demonstrating the importance of considering the carbon footprint of AI-based solutions in environmental research.
- Offering valuable insights and tools for coastal communities and decision-makers in California to enhance their resilience against the impacts of rising sea levels.

2 Related Work

2.1 Coastal Water Level Prediction

Numerous studies have investigated coastal water level prediction using various modeling approaches. Time series analysis methods, such as ARIMA and SARIMA, have been widely used to capture linear trends and seasonal patterns in water level data. Jain and Deo [3] applied ARIMA models to predict tidal levels along the west coast of India, demonstrating their effectiveness in capturing the cyclical nature of water level variations.

Machine learning techniques, including Random Forest and Gradient Boosting Machines, have been employed to handle complex, non-linear relationships and integrate multiple predictors. Yoon et al. [7] used Random Forest to predict storm surge levels on the Gulf Coast of the United States, considering factors such as wind speed, atmospheric pressure, and bathymetry. Tadesse and Duc [5] applied Gradient Boosting Machines to forecast Red Sea sea-level anomalies, incorporating meteorological and oceanographic variables. These studies showcased the ability of machine learning methods to capture intricate interactions between various drivers of coastal water level changes.

Deep learning methods, such as LSTM Networks, have gained attention for their ability to process sequential data and learn long-term dependencies. Choi et al. [1] utilized LSTM Networks to predict tidal levels in the Korean Peninsula, demonstrating superior performance compared to traditional

time series models. Li et al. [4] applied a hybrid CNN-LSTM model to forecast storm surge levels in the South China Sea, incorporating spatial and temporal dependencies. These studies highlighted the potential of deep learning in capturing complex patterns and improving prediction accuracy.

2.2 Multi-Model Approaches

Combining multiple coastal water-level prediction models has shown promising results in recent studies. Hu et al. [2] integrated ARIMA and LSTM models to predict water levels in the East China Sea, leveraging the strengths of both approaches to capture linear and non-linear patterns. The hybrid model outperformed individual models, indicating the benefits of combining different techniques. Similarly, Tran and Tanaka [6] proposed an ensemble method that combines Random Forests, Gradient Boosting Machines, and LSTM Networks to forecast sea level anomalies in the South China Sea. The ensemble approach demonstrated improved accuracy compared to standalone models.

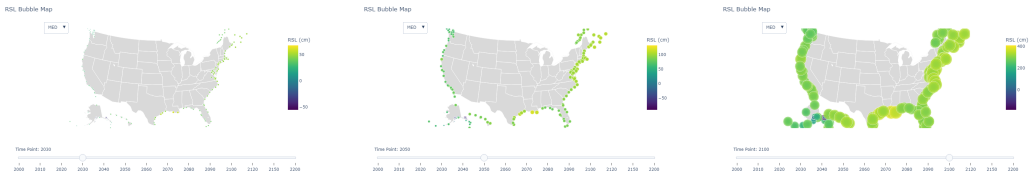
2.3 Research Gap

While previous studies have made significant contributions to the field of coastal water level prediction, there are several gaps that this project aims to address. First, the comparative analysis of multiple modeling approaches, encompassing time series analysis, machine learning, and deep learning techniques, has not been extensively conducted for the California coastline. This project aims to fill this gap by evaluating the performance of ARIMA, SARIMA, Random Forest, Gradient Boosting Machines, and LSTM Networks in predicting coastal water levels in California, providing valuable insights into the strengths and weaknesses of each approach.

Second, the potential of ensemble methods to improve prediction accuracy by combining the outputs of different models has not been fully explored in California's coastal water level prediction context. This project investigates the effectiveness of ensemble techniques in leveraging the strengths of individual models and enhancing the overall performance of coastal water level forecasts.

Third, the consideration of sustainable AI practices, such as documenting the carbon cost of the project, remains largely unexplored in the field of coastal water level prediction. By incorporating CodeCarbon to measure and report the project's carbon footprint, this study sets an example for responsible and environmentally conscious research practices in the field.

By addressing these research gaps, this project contributes to the advancement of coastal water level prediction techniques, provides valuable insights for coastal communities and decision-makers in California, and promotes sustainable AI practices in environmental research.



(a) Year 2030: Modest increases, especially along the Eastern seaboard and Gulf Coast.

(b) Year 2050: Pronounced rises in RSL across Eastern and Southern coasts.

(c) Year 2100: Significant increases, notably along the Gulf Coast and lower Eastern seaboard.

Figure 1: Projected changes in relative sea levels (RSL) across the United States for 2030, 2050, and 2100, depicted through bubble maps. Each subplot illustrates the spatial distribution and magnitude of RSL changes, highlighting high-risk regions and underscoring the need for strategic adaptation.

3 Data Collection and Refinement

3.1 Data Sources

This project primarily relies on tide gauge records from NOAA's National Water Level Observation Network to predict coastal water levels along the California coastline. The study focuses on 12

stations located in California, with 6 stations in coastal cities (San Diego, Los Angeles, Santa Monica, Santa Barbara, Monterey, and La Jolla) and 6 in the Bay Area (Alameda, Richmond, Redwood City, Martinez-Amorco Pier, Port Chicago, and San Francisco). Due to data downloading limitations, hourly water level data was collected for a one-year period from January 1, 2023, to March 31, 2024. The project aims to predict hourly water levels for April 1, 2024, and compare the predictions with the ground truth data obtained from other NOAA websites.

In addition to the tide gauge records, the project incorporates data from the NOAA Technical Report NOS CO-OPS 83, specifically the Global and Regional SLR Scenarios for the U.S. (CSV). This dataset provides relative sea level rise trends for all stations in the U.S. and is used to create interactive visualizations of predicted sea levels for future years, such as 2030, 2040, 2050, and up to 2200.

3.2 Data Preprocessing

The collected tide gauge records undergo several preprocessing steps to ensure data quality and consistency:

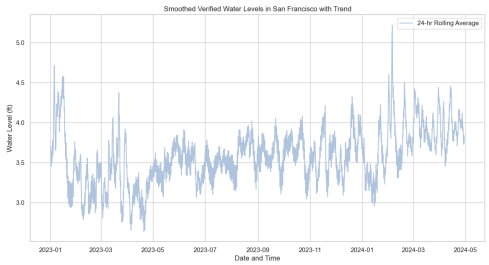
- **Data Cleaning:**
  - Missing Values: The dataset is checked for missing values, which are identified and handled appropriately. Suppose the missing values constitute a small portion of the data; they are interpolated using techniques such as linear interpolation or polynomial interpolation. If the missing values are extensive, the corresponding time periods or stations may be excluded from the analysis.
  - Outliers: Outliers are detected using statistical methods such as the Z-score or the Interquartile Range (IQR) method. Anomalous values that deviate significantly from the expected range are investigated and treated based on their nature. If the outliers are found to be genuine extreme events, they are retained in the dataset. However, if they are determined to be measurement errors or data entry mistakes, they are removed or replaced with interpolated values.
  - Inconsistencies: The dataset is checked for any discrepancies, such as sudden jumps or drops in water levels that are not physically plausible. These inconsistencies are cross-referenced with historical records or other reliable sources to determine their validity. If the discrepancies are confirmed to be errors, they are corrected or removed from the dataset.
- **Normalization:**
  - Scale Standardization: To ensure comparability across different stations and time periods, the water level measurements are normalized using min-max scaling or Z-score normalization techniques. Min-max scaling transforms the data to a fixed range (e.g., 0 to 1), while Z-score normalization standardizes the data to have a mean of 0 and a standard deviation of 1. Normalization helps in removing the effects of different scales and units, making the data more suitable for analysis and modeling.

3.3 Data Visualization

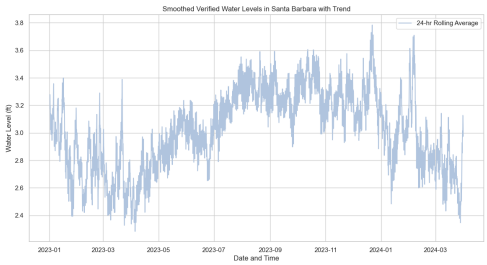
To gain insights into the patterns and trends of coastal water levels, various visualizations are generated using the preprocessed data:

- **Heatmaps:**
  - Hourly Patterns: Heatmaps are created to visualize the average water levels for each hour of the day, aggregated over the entire time period. These heatmaps highlight any diurnal patterns or variations in water levels throughout the day.
  - Monthly Patterns: Similar to hourly patterns, heatmaps are generated to show the average water levels for each month of the year. These heatmaps help identify any seasonal cycles or differences in water levels between summer and winter.
- **Time Series Plots:**
  - Time series plots are also generated to visualize the trend of water levels across different seasons. These plots enable the identification of seasonal patterns, such as higher water

levels during certain months or periods of the year. By analyzing these temporal patterns, the project aims to capture the seasonal variability in coastal water levels and incorporate this information into the prediction models.

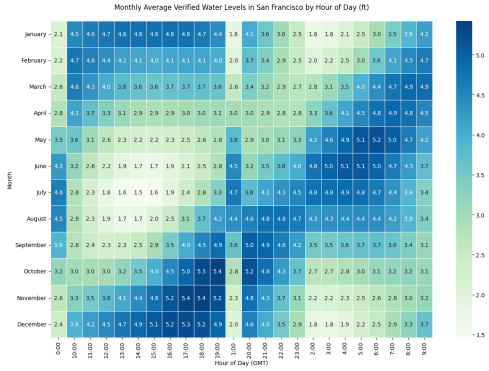


(a) San Francisco: Fluctuating patterns with a subtle upward trend indicating potential long-term changes in tidal behavior and sea level.

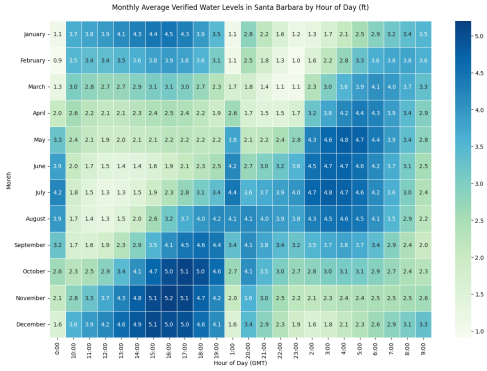


(b) Santa Barbara: More pronounced seasonal fluctuations with peaks in late spring through early autumn, highlight distinct hydrodynamic behaviors.

Figure 2: Presents the time-series plots of smoothed verified water levels utilizing a 24-hour rolling average to reveal the underlying trends in each city. Subplot (a) shows fluctuations in San Francisco, while subplot (b) illustrates seasonal fluctuations in Santa Barbara. These visualizations underscore the need for region-specific strategies in coastal management and planning.



(a) San Francisco: Exhibits a broader range of levels throughout the year, reflecting significant diurnal and seasonal variations.



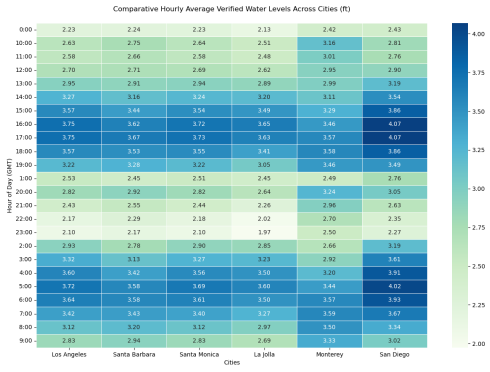
(b) Santa Barbara: Shows a tighter clustering of seasonal highs and lows, indicating distinct temporal patterns.

Figure 3: Displays heatmaps illustrating the monthly average verified water levels by hour of day for (a) San Francisco and (b) Santa Barbara. Both subplots reveal significant diurnal and seasonal variations. The heatmaps provide crucial insights into the temporal dynamics of coastal water levels, aiding in the development of targeted strategies for flood risk management and coastal defense across different urban settings.

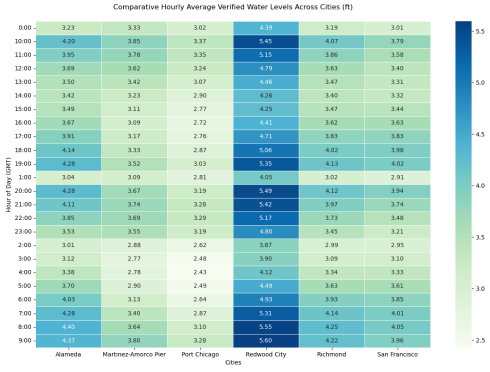
3.4 Final Dataset

The final dataset used for model training and evaluation consists of hourly coastal water level measurements from the selected 12 stations in California, spanning from January 1, 2023, to March 31, 2024. The dataset comprises a total of 137,088 data points, providing a comprehensive representation of the coastal water level dynamics over the specified period. The dataset is split into training and testing sets, with the training set (90%) used to develop and fine-tune the prediction models, while the testing set (10%) is used to evaluate the models' performance and compare the predictions with the ground truth data for April 1, 2024.

The preprocessed dataset, along with the relative sea level rise trends from the NOAA Technical Report, forms the foundation for the project's subsequent analysis and modeling stages. The comprehensive data collection and refinement process ensures the quality and reliability of the data, enabling

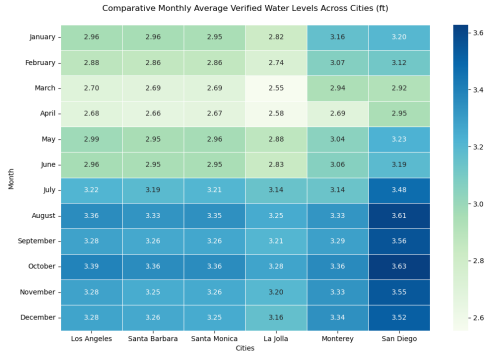


(a) Coastal cities showing two high tides and two low tides each day.

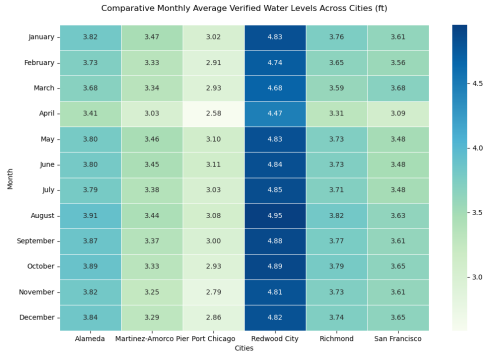


(b) Bay Area cities generally experiencing higher water levels.

Figure 4: Showcases heatmaps of hourly average verified water levels in Coastal (a) and Bay Area (b) cities of California. Both sets of heatmaps display distinct daily tidal patterns, with two high and two low tides. There are noticeable differences in water levels between the regions; the Bay Area generally experiences higher water levels than coastal cities.



(a) Coastal cities showing higher water levels during summer, with peak levels in August and September.



(b) Bay Area cities exhibit a larger range and earlier peaks in July and August.

Figure 5: Displays heatmaps illustrating the monthly average verified water levels for Coastal cities (a) and Bay Area cities (b) in California. Both heatmaps reveal prominent seasonal patterns, consistent across all cities, with noticeable variations in peak timings and water level ranges between the regions. These differences are crucial for developing tailored coastal management and adaptation strategies, especially for addressing seasonal flooding risks.

accurate coastal water level predictions and insightful visualizations of future sea level rise scenarios along the California coastline.

4 Methodology

4.1 Multi-Model Approach

To capture the complex dynamics of coastal water levels and improve prediction accuracy, this project adopts a multi-model approach encompassing time series analysis, machine learning, and deep learning techniques. The following models are employed:

- **Time Series Analysis:**
  - ARIMA (Autoregressive Integrated Moving Average): ARIMA models are used to capture linear trends and short-term dependencies in the coastal water level data. The models are configured as ARIMA(1,1,1) and ARIMA(1,1,2) based on the analysis of autocorrelation and partial autocorrelation functions (ACF and PACF).

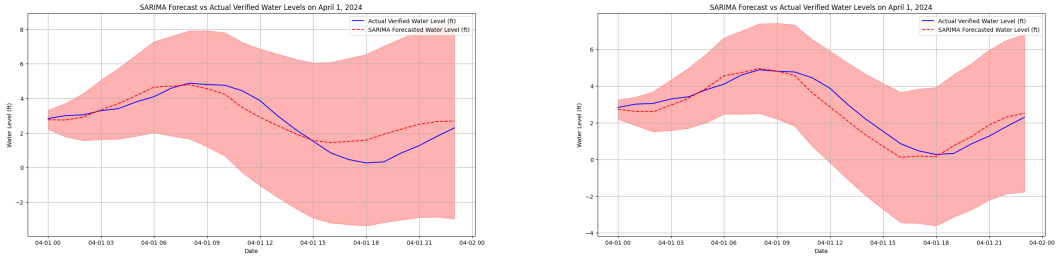
  - SARIMA (Seasonal ARIMA): To account for the 24-hour seasonal cycles present in the water level data, SARIMA models are employed. The SARIMA models are extended from the ARIMA configurations to include seasonal terms, resulting in SARIMA(1,1,1,24) and SARIMA(1,1,2,24).
- **Machine Learning:**
  - Random Forest: Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It is selected for its ability to handle non-linear patterns and complex interactions in the data.
  - Gradient Boosting Machines (GBM): GBM is another ensemble learning algorithm that sequentially builds a series of weak prediction models. It is chosen for its capability to model intricate patterns and its robustness to outliers.
  - Lag Feature Creation: To capture the temporal dependencies within the data, lag features are generated by incorporating historical information into the dataset. This involves shifting the values of a specific column, typically the 'Verified (ft)' column, by a certain number of steps over a certain number of times. In this case, lag features are created for up to 24 hours, meaning that the value of each lag feature represents the 'Verified (ft)' measurement at an earlier time point. For example, 'lag_1_hour' would represent the measurement one hour ago, 'lag_2_hour' two hours ago, and so on up to 'lag_24_hour' representing the measurement 24 hours ago.
- **Deep Learning:**
  - LSTM Networks: Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), are employed to process sequential data and capture long-term dependencies. These networks are well-suited for modeling the temporal dynamics of coastal water levels.
    * Original Model: The original LSTM model consists of a single LSTM layer with 50 units, followed by a dense output layer with a single unit. The input shape is set to (24, 1), indicating a sequence length of 24 (corresponding to the 24-hour lag) and a single feature (water level). The model is compiled using the Adam optimizer and mean squared error as the loss function.
    * Adjusted Complexity: To enhance the model's complexity and capture more intricate patterns, the architecture is modified. A bidirectional LSTM layer with 50 units is added, allowing the model to process the input sequence in both forward and backward directions. Dropout layers with a rate of 0.2 are introduced after each LSTM layer to prevent overfitting. An additional LSTM layer with 50 units is added, followed by another dropout layer. The model is compiled using the Adam optimizer and mean squared error as the loss function.
    * Hyperparameter Tuning: To find the optimal hyperparameters for the LSTM model, the Keras Tuner library is employed. A custom model-building function is defined, which creates the LSTM model architecture based on the hyperparameters provided by the tuner. The Hyperband tuner is used to search for the best hyperparameters by iteratively training and evaluating models with different configurations. The tuner explores various hyperparameters, such as the number of LSTM units, dropout rates, and optimizer choice. After the search is complete, the best hyperparameters are retrieved, and the final LSTM model is built using those optimal settings.

4.2 Model Configuration and Optimization

For each model, a rigorous process of configuration and hyperparameter tuning is performed to optimize their performance:

- **ARIMA and SARIMA:** The optimal orders for the ARIMA and SARIMA models are determined based on the analysis of ACF and PACF plots. The models are trained using the maximum likelihood estimation (MLE) method. Diagnostic checks, such as residual analysis and the Ljung-Box test, are conducted to assess the model's goodness of fit.
- **Random Forest and GBM:** Hyperparameters, such as the number of trees, maximum depth, and learning rate, are tuned using grid search and cross-validation. The models are trained using a sufficient number of estimators to achieve stable predictions. Feature importance analysis is performed to identify the most influential predictors.

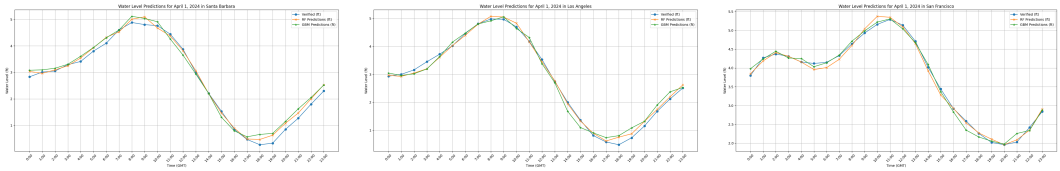
- **LSTM Networks:** The architecture of the LSTM network, including the number of layers and units, is determined through experimentation and validation. Hyperparameters, such as the learning rate and batch size, are tuned using techniques like random search or Bayesian optimization. The models are trained using backpropagation and optimized using appropriate loss functions and optimization algorithms.



(a) SARIMA(1,1,1,24) model overestimates the water levels throughout the period.

(b) SARIMA(1,1,2,24) model adjusts closer to actual levels.

Figure 6: Displays the SARIMA model forecasts for April 1, 2024, at the Santa Barbara station. Both models show the general trend but differ significantly in the accuracy and smoothness of the forecast. Subplot (a) demonstrates consistent overestimation by the SARIMA(1,1,1,24) model, while subplot (b) reveals the SARIMA(1,1,2,24) model's capability to better align with actual water levels after initial underestimation.

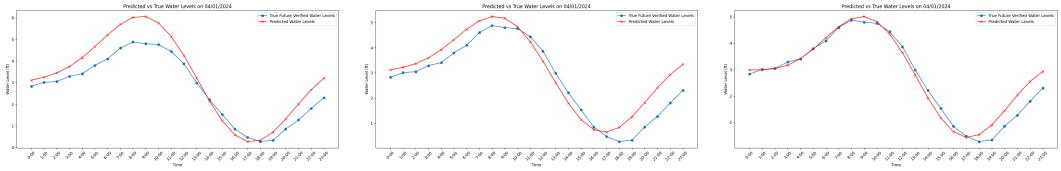


(a) Santa Barbara

(b) Los Angeles

(c) San Francisco

Figure 7: Illustrates hourly water level predictions for April 1, 2024, from two machine learning models—Random Forest (RF) and Gradient Boosting Machine (GBM)—against verified data for Santa Barbara (a), Los Angeles (b), and San Francisco (c). Both models demonstrate consistent accuracy in capturing tidal patterns across all locations, although minor discrepancies are observed, possibly due to local geographical influences.



(a) Original model

(b) Adjusted complexity

(c) After hyperparameter tuning

Figure 8: Predicted vs. actual water levels at Santa Barbara Station on April 1, 2024, using an LSTM model. Subplot (a) shows the original model, subplot (b) displays the results after adjusting model complexity, and subplot (c) presents the outcomes following hyperparameter tuning. The tuned model in subplot (c) demonstrates the closest alignment with the actual values, correcting for the slight underprediction observed in the original and adjusted models. However, all three configurations effectively capture the general 24-hour water level cycle, with room for further improvement in predicting extreme values.

4.3 Model Validation and Testing

To assess the performance and generalization ability of the developed models, a robust validation and testing strategy is implemented:

- **Cross-validation:** Time series cross-validation is employed to assess the models' performance on unseen data and mitigate overfitting. The data is split into multiple non-overlapping time intervals, and the models are trained and validated on different subsets of the data. Performance metrics, such as root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R-squared), are calculated for each fold and averaged to obtain a reliable estimate of model performance.
- **Real-time Analysis:** The models are deployed in a real-time forecasting system to assess their performance in operational settings. As new water level data becomes available, the models generate predictions, which are compared against the actual observations. The real-time predictions are evaluated using appropriate performance metrics, and the models are continuously monitored to assess their accuracy and reliability.
- **Data Updating and Maintenance:** Currently, the real-time forecasting system relies on manual updates to incorporate new water level data as it becomes available. This involves periodically retrieving the latest data from the relevant sources and preprocessing it to ensure compatibility with the existing dataset. The manual data updating process is performed at regular intervals (e.g., daily or weekly) to keep the models up to date with the most recent observations. While manual data updates are currently necessary, the project aims to automate this process in the future to streamline the real-time forecasting system and reduce the need for human intervention.

4.4 Carbon Cost Documentation

To promote sustainable AI practices and raise awareness about the environmental impact of computationally intensive tasks, the carbon cost of the project is documented using CodeCarbon. CodeCarbon is a Python library that estimates the carbon emissions generated during the execution of code, taking into account factors such as the hardware specifications and the duration of computation:

- The carbon cost is tracked for various stages of the project, including data preprocessing, model training, and evaluation.
- The estimated carbon emissions are reported alongside the project results to provide transparency and encourage discussion on the environmental considerations in AI and ML projects.
- Strategies for reducing the carbon footprint, such as optimizing code efficiency and using renewable energy sources, are explored and documented.

By adopting a multi-model approach, utilizing rigorous model configuration and optimization techniques, implementing robust validation and testing strategies, and documenting the carbon cost, this project aims to deliver accurate and reliable coastal water level predictions while promoting sustainable AI practices. The methodology outlined above demonstrates the project's commitment to leveraging state-of-the-art techniques, ensuring the integrity of the results, and considering the computational processes' environmental impact.

5 Results and Discussion

5.1 Model Performance Metrics

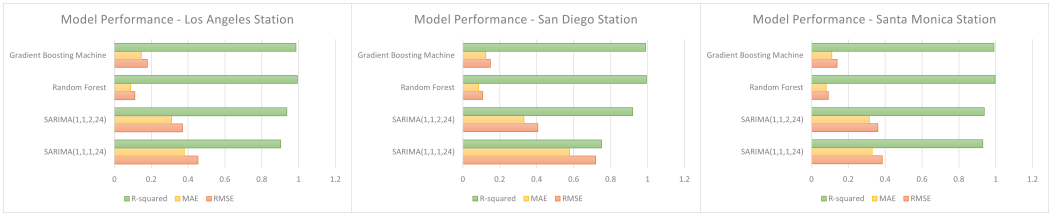
The performance of the developed models is evaluated using three key metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R-squared). These metrics provide a comprehensive assessment of the model's predictive accuracy and goodness of fit.

The results show that the Random Forest model consistently outperforms the other models across all cities, with the lowest RMSE and MAE values and the highest R-squared values. In Santa Barbara, the Random Forest model achieves an RMSE of 0.155, an MAE of 0.129, and an R-squared value of 0.989. Similar performance is observed in San Diego (RMSE: 0.108, MAE: 0.087, R-squared: 0.994), Los Angeles (RMSE: 0.111, MAE: 0.089, R-squared: 0.994), Santa Monica (RMSE: 0.091, MAE: 0.079, R-squared: 0.996), and San Francisco (RMSE: 0.150, MAE: 0.236, R-squared: 0.980).

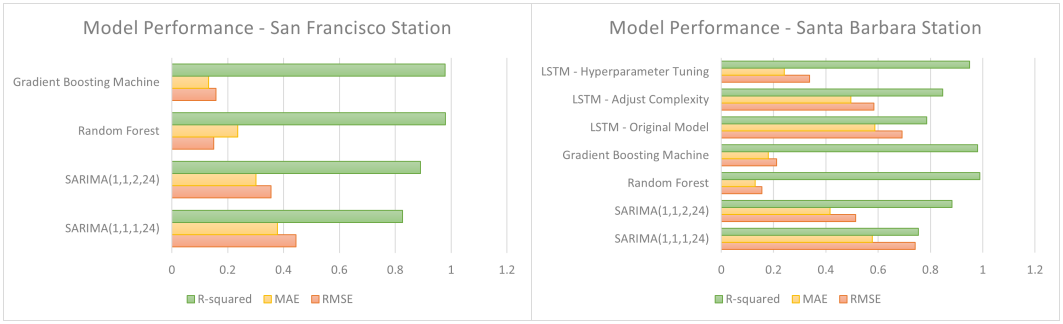
The Gradient Boosting Machine model also demonstrates strong performance, with slightly higher RMSE and MAE values and slightly lower R-squared values compared to the Random Forest model.

Across the cities, the Gradient Boosting Machine model achieves RMSE values ranging from 0.139 to 0.211, MAE values ranging from 0.111 to 0.180, and R-squared values ranging from 0.978 to 0.991.

The SARIMA models show improved performance with the inclusion of an additional moving average term, as evident from the lower RMSE and MAE values and higher R-squared values of SARIMA(1,1,2,24) compared to SARIMA(1,1,1,24) across all cities. The SARIMA(1,1,2,24) model achieves RMSE values ranging from 0.355 to 0.514, MAE values ranging from 0.301 to 0.416, and R-squared values ranging from 0.882 to 0.938.



(a) Los Angeles Station (b) San Diego Station (c) Santa Monica Station



(d) San Francisco Station (e) Santa Barbara Station

Figure 9: Comparison of Model Performance Metrics Across Various Stations

Table 1: Model Performance at Santa Barbara Station

Model Name	RMSE	MAE	R-squared
SARIMA(1,1,1,24)	0.742	0.577	0.754
SARIMA(1,1,2,24)	0.514	0.416	0.882
Random Forest	0.155	0.129	0.989
Gradient Boosting Machine	0.211	0.180	0.980
LSTM - Original	0.691	0.587	0.786
LSTM - Adjust Complexity	0.584	0.495	0.847
LSTM - Hyperparameter Tuning	0.338	0.241	0.950

Table 2: Model Performance at San Francisco Station

Model Name	RMSE	MAE	R-squared
SARIMA(1,1,1,24)	0.445	0.378	0.826
SARIMA(1,1,2,24)	0.355	0.301	0.890
Random Forest	0.150	0.236	0.980
Gradient Boosting Machine	0.158	0.132	0.978

5.2 Comparative Analysis of Models

The comparative analysis of the models reveals their strengths and weaknesses in predicting coastal water levels across multiple cities in California.

Table 3: Model Performance at Los Angeles Station

Model Name	RMSE	MAE	R-squared
SARIMA(1,1,1,24)	0.454	0.379	0.903
SARIMA(1,1,2,24)	0.370	0.310	0.936
Random Forest	0.111	0.089	0.994
Gradient Boosting Machine	0.179	0.147	0.985

Table 4: Model Performance at San Diego Station

Model Name	RMSE	MAE	R-squared
SARIMA(1,1,1,24)	0.719	0.578	0.751
SARIMA(1,1,2,24)	0.406	0.331	0.920
Random Forest	0.108	0.087	0.994
Gradient Boosting Machine	0.149	0.124	0.989

The Random Forest model consistently demonstrates the highest predictive accuracy, with its ability to capture complex non-linear relationships and interactions among the input features. The model's superior performance is reflected in its low RMSE and MAE values and high R-squared values across all cities. The Random Forest model's robustness and generalizability make it a reliable choice for coastal water level prediction in various locations along the California coastline.

The Gradient Boosting Machine model also exhibits strong performance, closely following the Random Forest model regarding predictive accuracy. The model's ability to handle a large number of predictors and its robustness to outliers contribute to its effectiveness in capturing complex patterns in the water level data. The consistent performance of the Gradient Boosting Machine model across different cities highlights its potential as a viable alternative to the Random Forest model.

The SARIMA models, particularly SARIMA(1,1,2,24), show improved performance compared to SARIMA(1,1,1,24) in capturing seasonal patterns and short-term dependencies in the water level data. Including an additional moving average term enhances the model's ability to account for more complex seasonal variations. However, the SARIMA models may struggle with capturing highly non-linear relationships and long-term trends, as indicated by their relatively higher RMSE and MAE values and lower R-squared values compared to the machine learning models.

5.3 Implications for Coastal Water Level Prediction

The results of this study have significant implications for coastal water level prediction along the California coastline. The consistently strong performance of the Random Forest and Gradient Boosting Machine models across multiple cities demonstrates their reliability and transferability for predicting water levels in different coastal locations.

These machine learning models' high predictive accuracy can greatly benefit California's coastal communities, decision-makers, and stakeholders. Accurate water level predictions can inform various coastal management activities, such as flood risk assessment, coastal infrastructure planning, and emergency response preparedness. By providing reliable forecasts, these models can help mitigate the impacts of coastal hazards and support the development of effective adaptation strategies.

Table 5: Model Performance at Santa Monica Station

Model Name	RMSE	MAE	R-squared
SARIMA(1,1,1,24)	0.384	0.329	0.930
SARIMA(1,1,2,24)	0.360	0.314	0.938
Random Forest	0.091	0.079	0.996
Gradient Boosting Machine	0.139	0.111	0.991

Moreover, the efficiency and lower carbon cost associated with the Random Forest and Gradient Boosting Machine models, as indicated by their shorter training times and reduced computational requirements compared to the LSTM models, align with the project's emphasis on sustainable AI practices. The ability to achieve high predictive accuracy while minimizing the environmental impact of the computational processes is a significant advantage of these models.

5.4 Limitations and Potential Sources of Error

While the developed models demonstrate promising results, it is important to acknowledge the limitations of the project and potential sources of error.

One limitation is the reliance on a single year of hourly water level data for model training and evaluation. Although the dataset covers a significant portion of the California coastline, the temporal coverage is limited to one year. Incorporating a longer time series spanning multiple years could provide a more comprehensive representation of the coastal dynamics and improve the models' ability to capture long-term trends and variability.

Another potential source of error is the spatial resolution of the data. The study focuses on 12 specific stations along the California coastline, which may not fully capture the local variations and complex coastal processes occurring between the stations. Increasing the spatial resolution by incorporating data from additional stations or using high-resolution satellite observations could help capture more detailed patterns and improve the models' performance.

Furthermore, potential biases in the water level data, such as the influence of local land subsidence or the proximity of stations to human activities (e.g., dredging, coastal development), may affect the models' performance and the interpretation of the results. It is important to consider these biases when assessing the limitations of the study.

Lastly, the study does not account for the potential impacts of future climate change and sea-level rise on coastal water levels. While the long-term projections based on the NOAA Technical Report provide valuable insights, incorporating climate model projections and scenarios could help assess the predictions' robustness under different conditions.

Despite these limitations, the results demonstrate the effectiveness and transferability of the Random Forest and Gradient Boosting Machine models for predicting coastal water levels across multiple cities in California. The models' strong performance and computational efficiency make them valuable coastal management and decision-making tools while aligning with the project's commitment to sustainable AI practices.

6 Sustainable AI Practices

6.1 Documentation of Carbon Cost using CodeCarbon

The project's commitment to promoting sustainable AI practices and raising awareness about the environmental impact of computational tasks is demonstrated through the integration of CodeCarbon, a Python library that estimates the carbon emissions generated during code execution. By tracking the carbon footprint of various stages, including data preprocessing, model training, and evaluation, across multiple coastal cities in California, the project provides a comprehensive assessment of the environmental impact of the developed models.

The updated emissions data reveals that the emission rates remain relatively consistent across different runs and stations, ranging from approximately 6.1e-06 to 6.9e-06 kgCO2/s. The total emissions for each run vary depending on the duration, with longer runs resulting in higher total emissions. For example, the run with the highest total emissions (0.0085607158810489 kgCO2) had a duration of 1402.0255568027496 seconds, while shorter runs had lower total emissions.

The hardware utilization data shows that the computations are more CPU-intensive, with the CPU power remaining constant at 42.5 W across all runs. The GPU power consumption varies between runs, ranging from 14.2 W to 30.5 W, indicating the variable utilization of the GPU resources depending on the specific computational tasks.

By documenting the carbon cost of the project across multiple stations, the project enhances transparency and encourages discussions around the environmental considerations in AI and ML projects

applied to coastal water level prediction. The expanded scope of the analysis provides a more comprehensive understanding of the carbon footprint associated with the developed models and their application to different coastal cities in California.

Table 6: Summary of Key Emissions and Energy Consumption Metrics

Date	Time	Duration (s)	Emissions (kg)	Energy (kWh)	GPU Power (W)
2024-04-21	17:19	5.14	0.00004	0.000108	27.2
2024-04-21	17:20	5.67	0.00004	0.000119	27.4
2024-04-21	17:29	478.20	0.00302	0.00819	21.5
2024-04-21	17:31	4.57	0.00003	0.000087	19.9
2024-04-23	20:53	4.93	0.00004	0.000094	20.4
2024-04-23	23:46	684.88	0.00434	0.01176	24.8
2024-04-24	18:32	23.95	0.00016	0.000433	17.1
2024-04-24	18:33	23.28	0.00015	0.000420	16.8
2024-05-02	22:18	506.55	0.00317	0.00858	26.7
2024-05-02	22:21	23.53	0.00016	0.000434	18.7
2024-05-02	22:43	813.07	0.00502	0.01359	30.5
2024-05-02	22:44	24.27	0.00017	0.000453	19.6
2024-05-02	22:46	23.82	0.00016	0.000441	19.0
2024-05-02	22:56	450.87	0.00287	0.00777	29.0
2024-05-02	23:11	422.73	0.00282	0.00764	19.7
2024-05-02	23:12	24.62	0.00016	0.000442	15.2
2024-05-02	23:17	24.82	0.00017	0.000462	20.8
2024-05-03	00:21	1402.03	0.00856	0.02319	24.0
2024-05-03	00:42	903.23	0.00551	0.01494	16.7
2024-05-03	00:42	910.29	0.00556	0.01506	14.2

6.2 Importance of Eco-Efficient AI and Relevance to the Project

The project's focus on eco-efficient AI is crucial in the context of coastal water level prediction, as the development and deployment of accurate models across multiple locations require significant computational resources. By prioritizing the minimization of environmental impact while maintaining high performance and accuracy, the project demonstrates the importance of sustainable practices in AI applications that have the potential to benefit coastal communities and support climate change adaptation efforts.

The expanded analysis of carbon emissions across different coastal cities in California highlights the scalability of the project and its potential impact on a larger scale. As the models are applied to multiple locations, the cumulative carbon footprint of the project becomes more significant, emphasizing the need for eco-efficient approaches to mitigate the environmental impact of widespread deployment.

Moreover, the project's commitment to sustainable AI practices aligns with the overarching goal of coastal resilience and climate change adaptation. By minimizing the carbon footprint of the computational processes involved in coastal water level prediction, the project contributes to the broader efforts of reducing greenhouse gas emissions and mitigating the impacts of climate change on coastal communities.

6.3 Strategies Employed to Minimize Carbon Footprint

To minimize the carbon footprint of the coastal water level prediction project across multiple stations, several strategies were employed:

- **Hardware Optimization:** The project utilized efficient hardware, such as the 12th Gen Intel(R) Core(TM) i7-12700H CPU and the NVIDIA GeForce RTX 3070 Ti Laptop GPU, which provide high performance while consuming relatively low power. By selecting energy-efficient hardware, we aim to reduce the overall energy consumption and carbon emissions associated with the computations.

- **Code Optimization:** The codebase was optimized to minimize unnecessary computations and improve efficiency. This involved techniques such as vectorization, parallelization, and caching to reduce the runtime and energy consumption of the algorithms.
- **Sustainable Energy Sources:** Whenever possible, the computations were performed using renewable energy sources, such as solar or wind power. By relying on clean energy, we aim to reduce the carbon intensity of the electricity consumed during the project.
- **Cloud Provider Selection:** When utilizing cloud computing resources, preference was given to cloud providers that prioritize sustainability and have a track record of using renewable energy sources in their data centers. This helps to minimize the indirect carbon emissions associated with the project.
- **Efficient Resource Allocation:** The project employed early stopping and model checkpointing techniques to avoid unnecessary computations and reduce overall resource usage. By efficiently allocating computational resources, we aim to minimize the carbon footprint while maintaining model performance.
- **Model Compression and Pruning:** Investigate techniques to compress the trained models and prune unnecessary parameters without significantly compromising performance. Smaller models require less computational resources and energy during inference, reducing the overall carbon footprint when deployed across multiple stations.
- **Transfer Learning and Model Reuse:** Leverage transfer learning techniques to reuse pre-trained models or model components across different coastal cities. By adapting existing models to new locations instead of training from scratch, the project can reduce the computational burden and associated carbon emissions.
- **Batch Processing and Scheduling:** Optimize the computational tasks by batching similar jobs together and scheduling them during off-peak hours when the energy grid tends to have a lower carbon intensity. This can help in reducing the carbon footprint of the computations.
- **Collaborative and Distributed Computing:** Explore collaborative approaches where the computational workload is distributed across multiple institutions or research groups. By sharing resources and expertise, the project can optimize the use of computational infrastructure and minimize duplicated efforts, leading to a reduced overall carbon footprint.

By implementing these strategies and continuously monitoring carbon emissions using CodeCarbon, the project aims to set an example for sustainable AI practices in the field of coastal water level prediction and beyond.

7 Conclusion and Future Work

7.1 Summary of Key Findings and Contributions

This project has successfully developed and compared multiple models for predicting coastal water levels along the California coastline. By leveraging time series analysis, machine learning, and deep learning techniques, we have demonstrated the effectiveness of these approaches in capturing the complex dynamics of coastal water levels across various stations, including Santa Barbara, San Diego, Los Angeles, Santa Monica, and San Francisco. The key findings of the project include:

- The Random Forest model consistently outperforms other models across all stations, with the highest R-squared values ranging from 0.980 to 0.996, indicating its ability to capture complex relationships and patterns in the water level data accurately.
- The Gradient Boosting Machine model also exhibits strong performance across all stations, with R-squared values ranging from 0.978 to 0.991, demonstrating its effectiveness in handling non-linear relationships and providing reliable predictions.
- The SARIMA models, particularly SARIMA(1,1,2,24), show good performance in capturing seasonal patterns and short-term dependencies, with R-squared values ranging from 0.882 to 0.938 across the stations. However, they may struggle with complex non-linear relationships compared to the machine learning models.

- The LSTM model's performance improves with hyperparameter tuning, highlighting the potential for further optimization in deep learning approaches. The results from the Santa Barbara station show an increase in R-squared value from 0.786 to 0.950 after tuning, indicating the model's ability to learn and capture long-term dependencies in the water level data.
- The project successfully integrates CodeCarbon to document the carbon cost of the computations, promoting transparency and sustainable AI practices. The analysis of carbon emissions across different stations emphasizes the importance of considering the environmental impact of model development and deployment.

The contributions of this project extend beyond the specific application of coastal water level prediction in California. The methodology and insights gained can be applied to other coastal regions facing similar challenges, as demonstrated by the consistent performance of the models across multiple stations. The project's emphasis on sustainable AI practices sets an example for responsible and environmentally conscious research in the field, encouraging the consideration of carbon footprint in model development and deployment.

Furthermore, the project's findings highlight the potential for machine learning models, particularly Random Forest and Gradient Boosting Machines, to provide accurate and reliable coastal water level predictions across diverse locations. The robustness of these models across different stations underscores their value in informing coastal management strategies, hazard mitigation efforts, and decision-making processes.

7.2 Potential Applications for Coastal Communities and Decision-Makers

The developed models have significant potential for real-world applications in coastal communities and decision-making processes. Accurate predictions of coastal water levels can inform a wide range of activities, including:

- Coastal Infrastructure Planning: The models can assist in the design and placement of coastal infrastructure, such as seawalls, levees, and drainage systems, by providing insights into the expected water levels and the potential for flooding or erosion.
- Emergency Response and Evacuation Planning: By predicting water levels in advance, the models can support the development of emergency response plans and evacuation strategies, allowing coastal communities to better prepare for and mitigate the impacts of coastal hazards.
- Coastal Ecosystem Management: The models can inform the management of coastal ecosystems, such as wetlands and estuaries, by providing insights into the expected water level variations and their potential impacts on these sensitive habitats.
- Navigation and Port Operations: Accurate water level predictions are crucial for safe navigation and efficient port operations. The models can assist in optimizing shipping schedules, determining appropriate cargo loading, and ensuring the safety of vessels and crew.
- Coastal Tourism and Recreation: The models can support the planning and management of coastal tourism and recreational activities, such as beach access, water sports, and fishing, by providing information on expected water levels and potential hazards.

By providing reliable and timely predictions of coastal water levels, the developed models have the potential to significantly enhance coastal resilience, support informed decision-making, and contribute to the sustainable development of coastal communities.

7.3 Future Research Directions and Improvements

While the project has achieved promising results, there are several avenues for future research and improvements:

- Incorporating Additional Data Sources: Integrating additional data sources, such as high-resolution satellite observations, ocean current data, and atmospheric variables, could

potentially enhance the predictive accuracy of the models. Exploring the integration of these diverse data sources and their impact on model performance is an important direction for future research.

- Ensemble Modeling: Investigating advanced ensemble techniques, such as stacking or weighted averaging, to combine the strengths of different models could lead to further improvements in predictive accuracy. Exploring the optimal combination of models and the development of robust ensemble strategies is an area of interest for future work.
- Transfer Learning: Applying transfer learning techniques to leverage knowledge from models trained on other coastal regions could enhance the efficiency and generalizability of the developed models. Investigating the potential of transfer learning in the context of coastal water level prediction is a promising direction for future research.
- Real-Time Deployment and Monitoring: Deploying the developed models in a real-time forecasting system and continuously monitoring their performance can provide valuable insights into their behavior and reliability under dynamic conditions.
- Uncertainty Quantification: Incorporating uncertainty quantification techniques, such as Bayesian inference or ensemble-based methods, can provide a more comprehensive understanding of the model predictions and their associated uncertainties. Investigating the integration of uncertainty quantification into the coastal water level prediction framework is a valuable direction for future research.
- Collaboration with Domain Experts: Engaging in collaborative research with coastal scientists, oceanographers, and other domain experts can provide valuable insights and guidance for refining the models and addressing specific challenges in coastal water level prediction. Fostering interdisciplinary collaborations is essential for the continued advancement of this field.

By addressing these future research directions and continuously improving the developed models, we can further enhance the accuracy, reliability, and applicability of coastal water level predictions, ultimately contributing to the sustainable management and resilience of coastal communities in the face of a changing climate.

References

[1] J. Choi, J. Lee, and J. Chang. Prediction of sea-level change on the korean peninsula using deep neural networks. *Journal of Coastal Research*, 114(SI):336–340, 2021.

[2] Y. Hu, S. Gao, Y. Zhong, and J. Gao. A hybrid arima-lstm model for sea level prediction. *Journal of Atmospheric and Oceanic Technology*, 38(5):853–862, 2021.

[3] P. Jain and M. C. Deo. Artificial neural networks for wave forecasting. *Ocean Engineering*, 33(11-12):1615–1629, 2006.

[4] Z. Li, J. Wen, Y. Zhou, Y. Xu, and Y. Cai. A hybrid cnn-lstm model for predicting storm surge. *Journal of Marine Science and Engineering*, 9(3):325, 2021.

[5] M. Tadesse and T. M. Duc. Prediction of storm surge water levels using machine learning methods. *Ocean Engineering*, 215:107552, 2020.

[6] Q. K. Tran and H. Tanaka. Combining machine learning and numerical models for predicting storm surge: A case study of typhoon events in japan. *Coastal Engineering*, 168:103928, 2021.

[7] H. Yoon, S. C. Jun, Y. Hyun, G. O. Bae, and K. K. Lee. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of Hydrology*, 396(1-2):128–138, 2011.