Comparative Analysis of Models for Predicting Coastal Water Levels in California

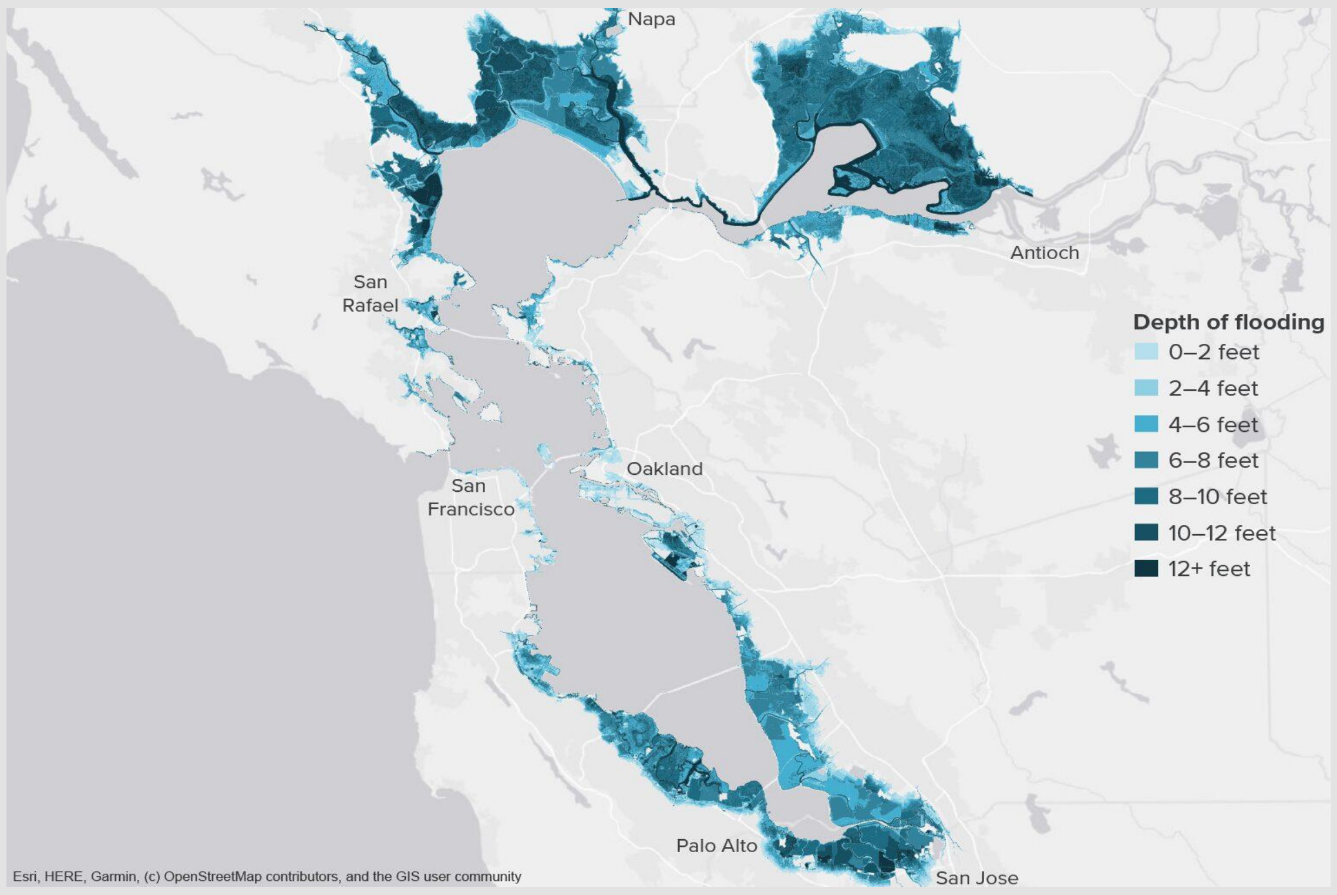
Hongxin Wu
Georgetown University

Abstract

This project undertakes the development and comparative analysis of multiple predictive models to forecast coastal water levels along California's coastline with a focus on enhancing accuracy and reliability. Employing time series analysis (ARIMA, SARIMA), machine learning techniques (Random Forest, Gradient Boosting Machines), and deep learning methodologies (LSTM), the study delivers forecasts on two scales: short-term (hourly for the next 24 hours) and long-term (looking ahead to 2050, 2100, and 2200). Data sourced from tide gauges and advanced climate models form the basis of our refined input. Our results underscore the benefits of integrating various modeling approaches, highlighting superior performance in both temporal scales. Additionally, the project champions sustainable scientific practices by rigorously documenting its carbon footprint via CodeCarbon, setting a benchmark for responsible AI usage in environmental research. The findings are expected to significantly influence future coastal resilience strategies, aiding policymakers and community leaders in proactive decision-making.

Introduction

Globally, coastal communities are increasingly vulnerable to the adverse effects of climate change, with rising sea levels threatening both human and ecological systems. In California, where vast populations reside along the coast, the stakes are particularly high, necessitating precise and actionable forecasts of water levels. This project aims to address these urgent needs by synthesizing traditional and contemporary predictive methodologies to create a robust framework for water level forecasting. Our approach combines time series analysis, machine learning, and deep learning to harness each of their strengths, thus optimizing accuracy and reliability across different time frames. By providing detailed hourly forecasts up to 24 hours and strategic long-term projections for the years 2050, 2100, and 2200, the project facilitates immediate emergency responses, such as coordinating timely evacuations or adjusting harbor activities to current water levels. And aids in the planning and execution of long-term adaptation strategies. This initiative not only deepens our comprehension of coastal dynamics amid a shifting climate but also furthers environmental science. It embodies a commitment to ethical AI, utilizing technology that supports our planet's sustainability and equips communities to navigate the future with greater certainty.



Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

Materials

Data Sources

- **Tide Gauge Data** - Use historical and current sea level records from NOAA's National Water Level Observation Network. These records provide ground-truth measurements that are essential for training and testing the predictive models.
- **Meteorological Data** - Include wind speed, barometric pressure, and air temperature data, which are significant drivers of sea level changes and are used to enhance the predictive accuracy of the models.

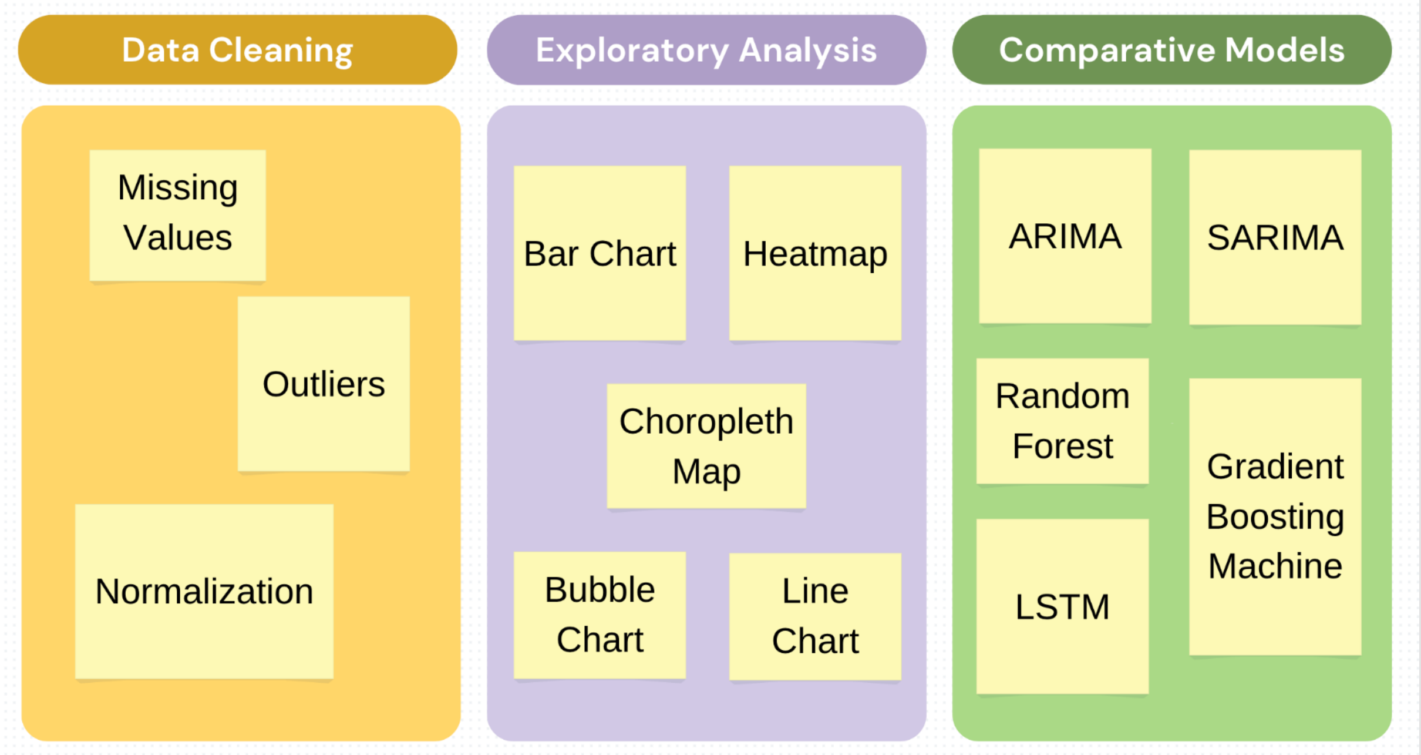
Data Overview

137088
Data Points

12
Stations

456
Days

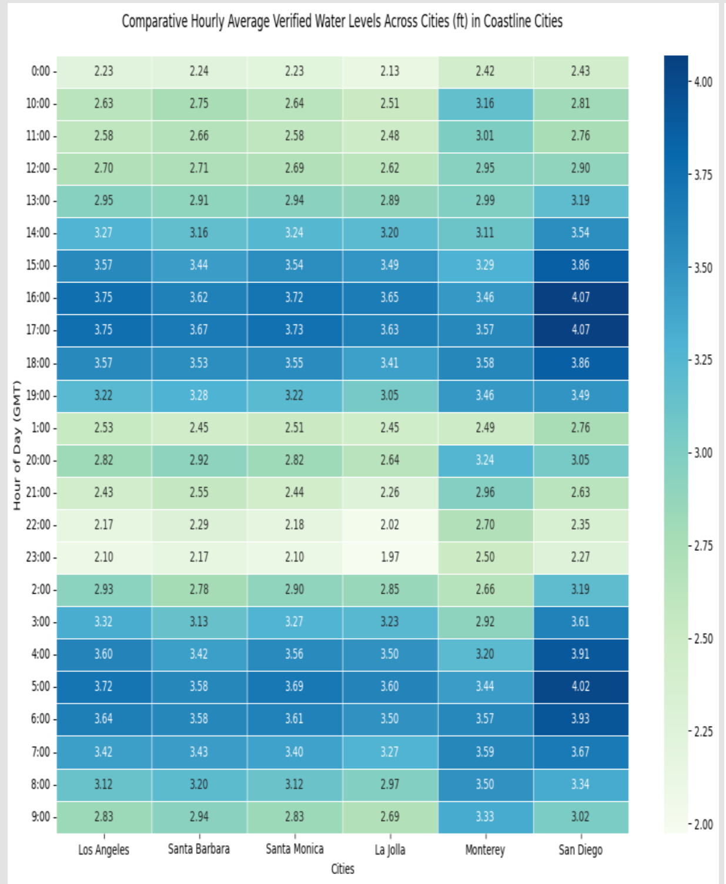
Methodology



Data Cleaning: Missing Values; Outliers; Normalization

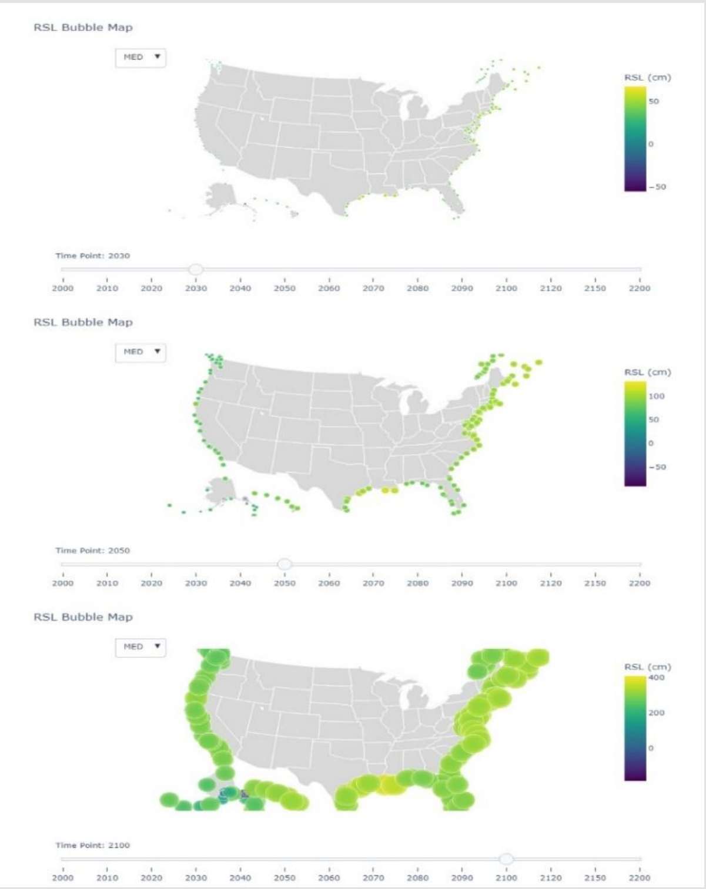
Exploratory Analysis: Bar Chart; Heatmap; Choropleth Map; Bubble Chart; Line Chart

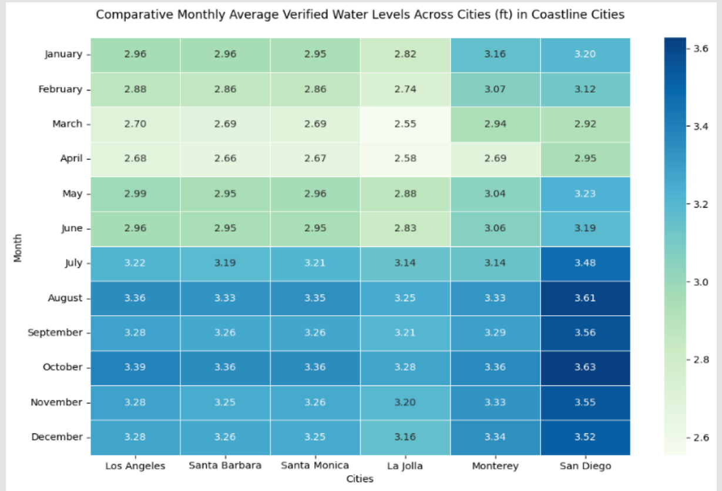
Comparative Models: ARIMA; SARIMA; Random Forest; Gradient Boosting Machine; LSTM



Comparative Hourly Average Verified Water Levels Across Cities (ft) in Coastline Cities

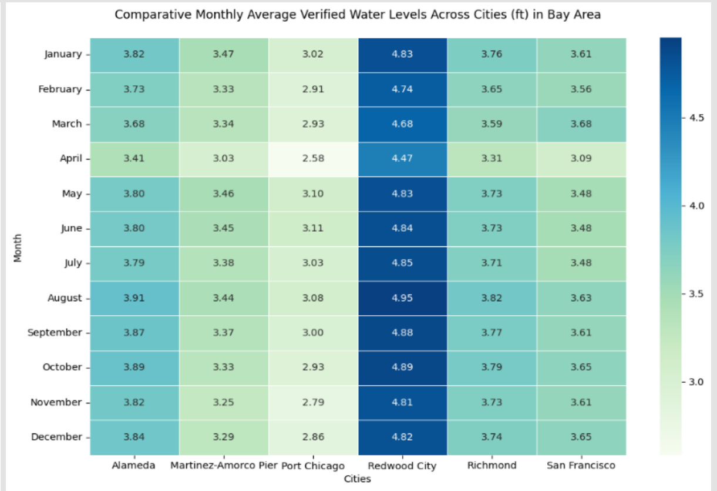
Hour of Day (GMT)	Los Angeles	Santa Barbara	Santa Monica	La Jolla	Monterey	San Diego
0:00	2.23	2.24	2.23	2.13	2.42	2.43
10:00	2.63	2.75	2.64	2.51	3.16	2.81
11:00	2.58	2.66	2.58	2.48	3.01	2.76
12:00	2.70	2.71	2.69	2.62	2.95	2.90
13:00	2.95	2.91	2.94	2.89	2.99	3.19
14:00	3.27	3.16	3.24	3.20	3.11	3.54
15:00	3.57	3.44	3.54	3.49	3.29	3.86
16:00	3.75	3.62	3.72	3.65	3.46	4.07
17:00	3.75	3.67	3.73	3.63	3.57	4.07
18:00	3.57	3.53	3.55	3.41	3.58	3.86
19:00	3.22	3.28	3.22	3.05	3.46	3.49
1:00	2.53	2.45	2.51	2.45	2.49	2.76
20:00	2.82	2.92	2.82	2.64	3.24	3.05
21:00	2.43	2.55	2.44	2.26	2.96	2.63
22:00	2.17	2.29	2.18	2.02	2.70	2.35
23:00	2.10	2.17	2.10	1.97	2.50	2.27
2:00	2.93	2.78	2.90	2.85	2.66	3.19
3:00	3.32	3.13	3.27	3.23	2.92	3.61
4:00	3.60	3.42	3.56	3.50	3.20	3.91
5:00	3.72	3.58	3.69	3.60	3.44	4.02
6:00	3.64	3.58	3.61	3.50	3.57	3.93
7:00	3.42	3.43	3.40	3.27	3.59	3.67
8:00	3.12	3.20	3.12	2.97	3.50	3.34
9:00	2.83	2.94	2.83	2.69	3.33	3.02





Comparative Monthly Average Verified Water Levels Across Cities (ft) in Coastline Cities

Month	Los Angeles	Santa Barbara	Santa Monica	La Jolla	Monterey	San Diego
January	2.96	2.96	2.95	2.82	3.16	3.20
February	2.88	2.86	2.86	2.74	3.07	3.12
March	2.70	2.69	2.69	2.55	2.94	2.92
April	2.68	2.66	2.67	2.58	2.69	2.95
May	2.99	2.95	2.96	2.88	3.04	3.23
June	2.96	2.95	2.95	2.83	3.06	3.19
July	3.22	3.19	3.21	3.14	3.14	3.48
August	3.36	3.33	3.35	3.25	3.33	3.61
September	3.28	3.26	3.26	3.21	3.29	3.56
October	3.39	3.36	3.36	3.28	3.36	3.63
November	3.28	3.25	3.26	3.20	3.33	3.55
December	3.28	3.26	3.25	3.16	3.34	3.52



Comparative Monthly Average Verified Water Levels Across Cities (ft) in Bay Area

Month	Alameda	Martinez-Amorco Pier	Port Chicago	Redwood City	Richmond	San Francisco
January	3.82	3.47	3.02	4.83	3.76	3.61
February	3.73	3.33	2.91	4.74	3.65	3.56
March	3.68	3.34	2.93	4.68	3.59	3.68
April	3.41	3.03	2.58	4.47	3.31	3.09
May	3.80	3.46	3.10	4.83	3.73	3.48
June	3.80	3.45	3.11	4.84	3.73	3.48
July	3.79	3.38	3.03	4.85	3.71	3.48
August	3.91	3.44	3.08	4.95	3.82	3.63
September	3.87	3.37	3.00	4.88	3.77	3.61
October	3.89	3.33	2.93	4.89	3.79	3.65
November	3.82	3.25	2.79	4.81	3.73	3.61
December	3.84	3.29	2.86	4.82	3.74	3.65

Modeling and Results

Dataset Description

The dataset includes hourly time-series data from 12 NOAA Tides and Currents stations along the California coastline, with 6 stations in coastal cities and 6 in the Bay Area covering the period from *January 1, 2023*, to *March 31, 2024.*

Data Splitting

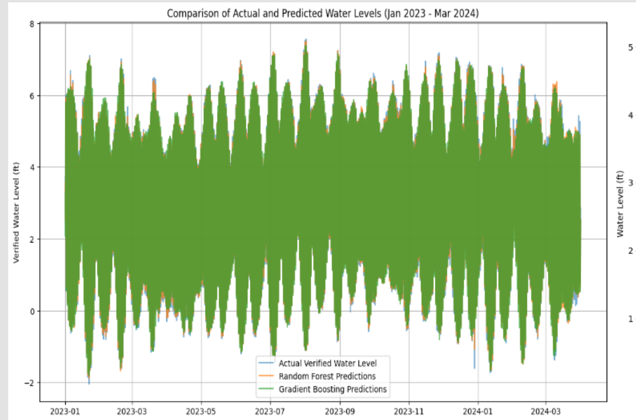
- **LSTM Model** - 90% of the data was used for training and 10% for testing, aligning with the sequential nature of time-series data.
- **RF and GBM Models** - Training was conducted using data from January 1, 2023, to February 29, 2024, with testing on data from March 1, 2024, to March 31, 2024, simulating real-world forecasting scenarios.

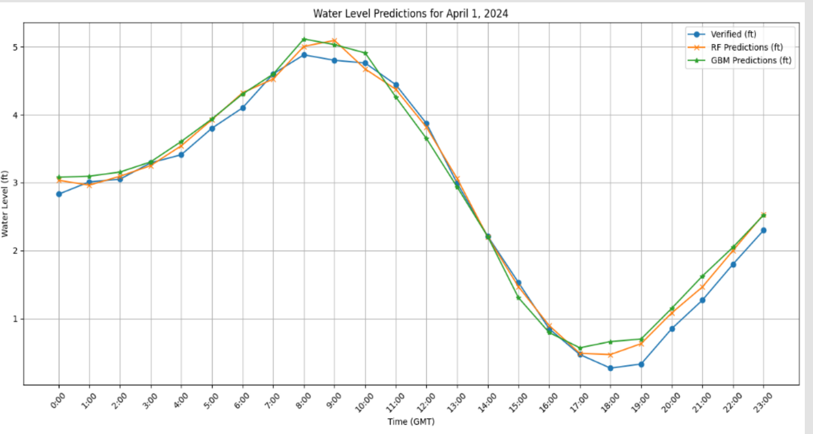
Model Selection and Rationale

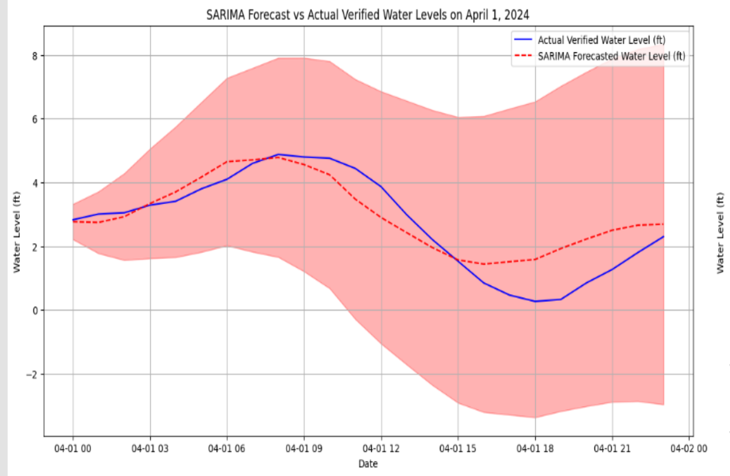
- **SARIMA** - Chosen for its effectiveness in modeling seasonal time-series data, which is critical given the daily tidal patterns in our dataset.
- **Random Forest & Gradient Boosting Machine** - These models were selected for their ability to handle nonlinear patterns and complex interactions, useful in environmental data analysis.
- **LSTM** - Ideal for sequential data like ours, LSTM models excel in capturing long-term dependencies necessary for predicting future water levels.

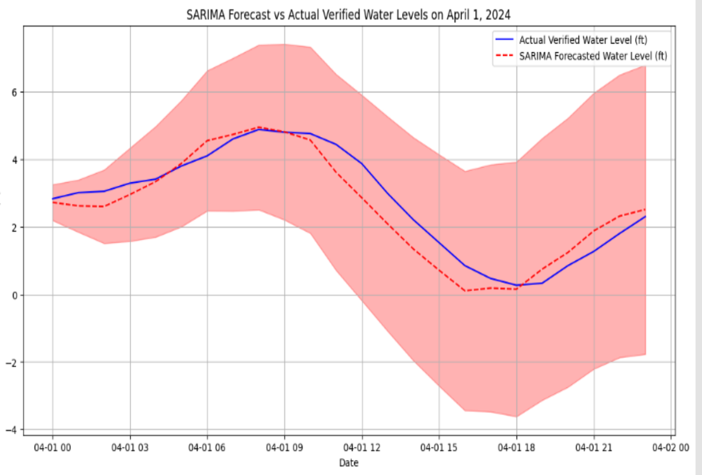
Model Configuration

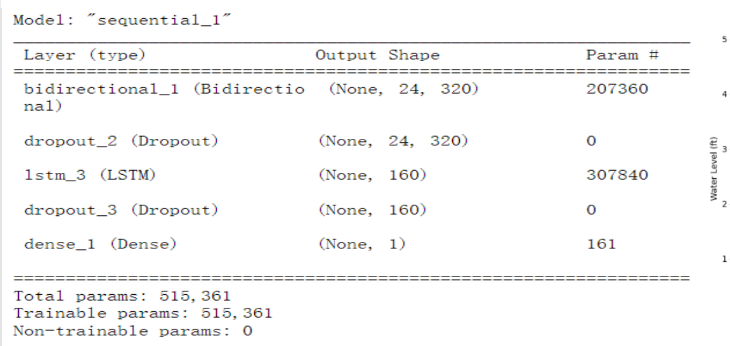
- **SARIMA Configurations** - Based on ACF and PACF analysis, SARIMA models were configured as ARIMA(1,1,1) and ARIMA(1,1,2), extended to SARIMA(1,1,1,24) and SARIMA(1,1,2,24) to account for 24-hour seasonal cycles.



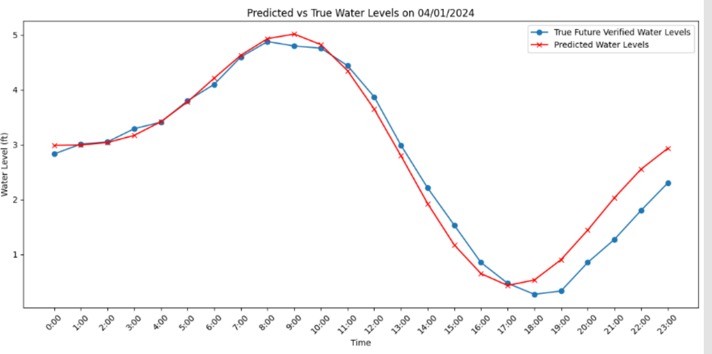


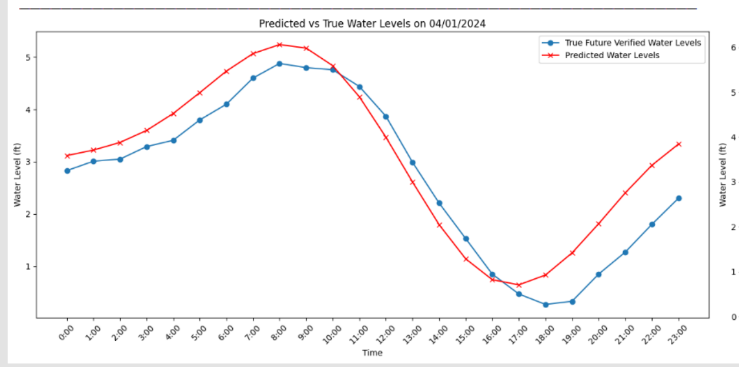


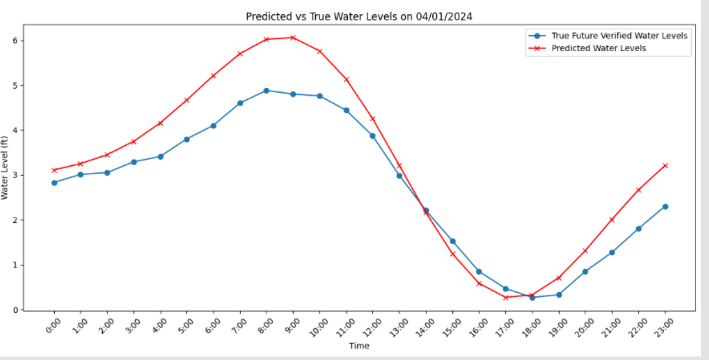




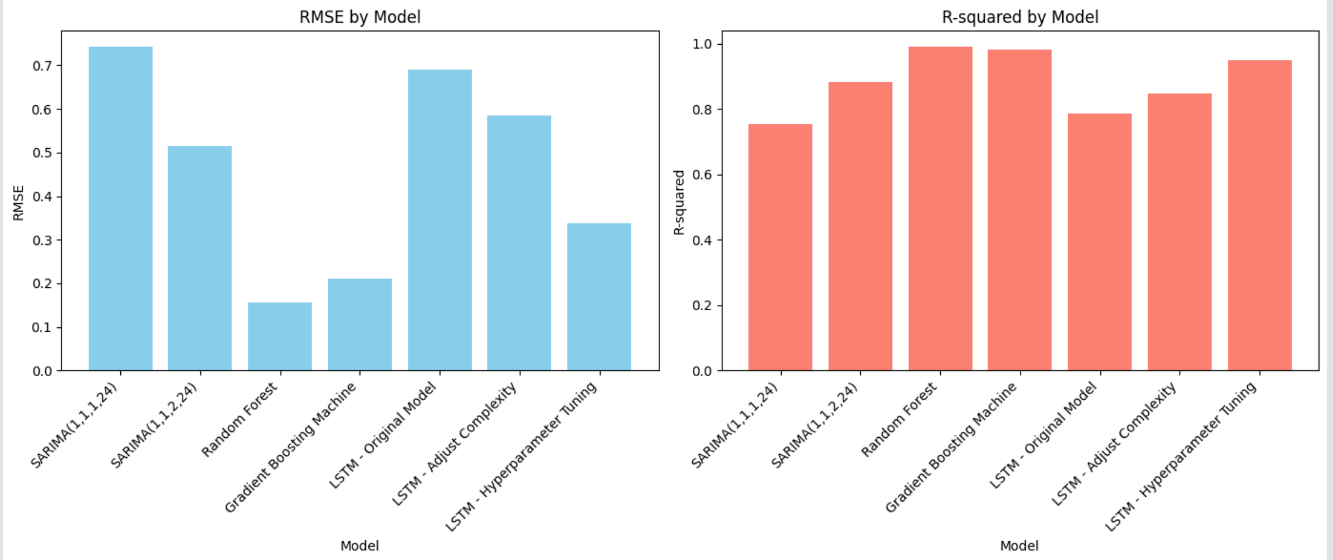
```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bidirectional_1 (Bidirectio  (None, 24, 320)          207360
 nal)

 dropout_2 (Dropout)         (None, 24, 320)           0

 lstm_3 (LSTM)               (None, 160)               307840

 dropout_3 (Dropout)         (None, 160)               0

 dense_1 (Dense)             (None, 1)                 161

=================================================================
Total params: 515,361
Trainable params: 515,361
Non-trainable params: 0
_________________________________________________________________
```







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 Model	0.691	0.587	0.786
LSTM - Adjust Complexity	0.584	0.495	0.847
LSTM - Hyperparameter Tuning	0.338	0.241	0.950



Conclusion

Performance and Efficiency

- *Random Forest* and *Gradient Boosting Machine* not only excel in predictive accuracy with the **highest R-squared** values and the **lowest RMSE** but also demonstrate superior energy efficiency, indicating their viability for time-series forecasting tasks where both performance and computational sustainability are considered.

Model Comparison

- The *SARIMA* model is a good performer, particularly with an additional MA term; however, it's less carbon-efficient than *RF* and *GBM.* Meanwhile, the *LSTM* shows notable performance gains with complexity and tuning, but its long computation times suggest a potentially high carbon cost, highlighting the importance of efficiency in model architecture.

Future Work

- **Ensemble Methods** - Use ensembles of RF, GBM, and LSTM to capitalize on each model's predictive strengths.
- **Cross-validation** - Test models across different temporal and geographical datasets to ensure robustness and generalizability.
- **Real-time Analysis** - Deploy models in a real-time forecasting system to assess performance in operational settings.
- **Eco-Efficient AI** - Streamline energy consumption across models to minimize carbon footprint and promote 'Green AI' practices for sustainable and effective modeling.

References

- NOAA Tides and Currents