# Comparative Analysis of Models for Predicting Coastal Water Levels in California
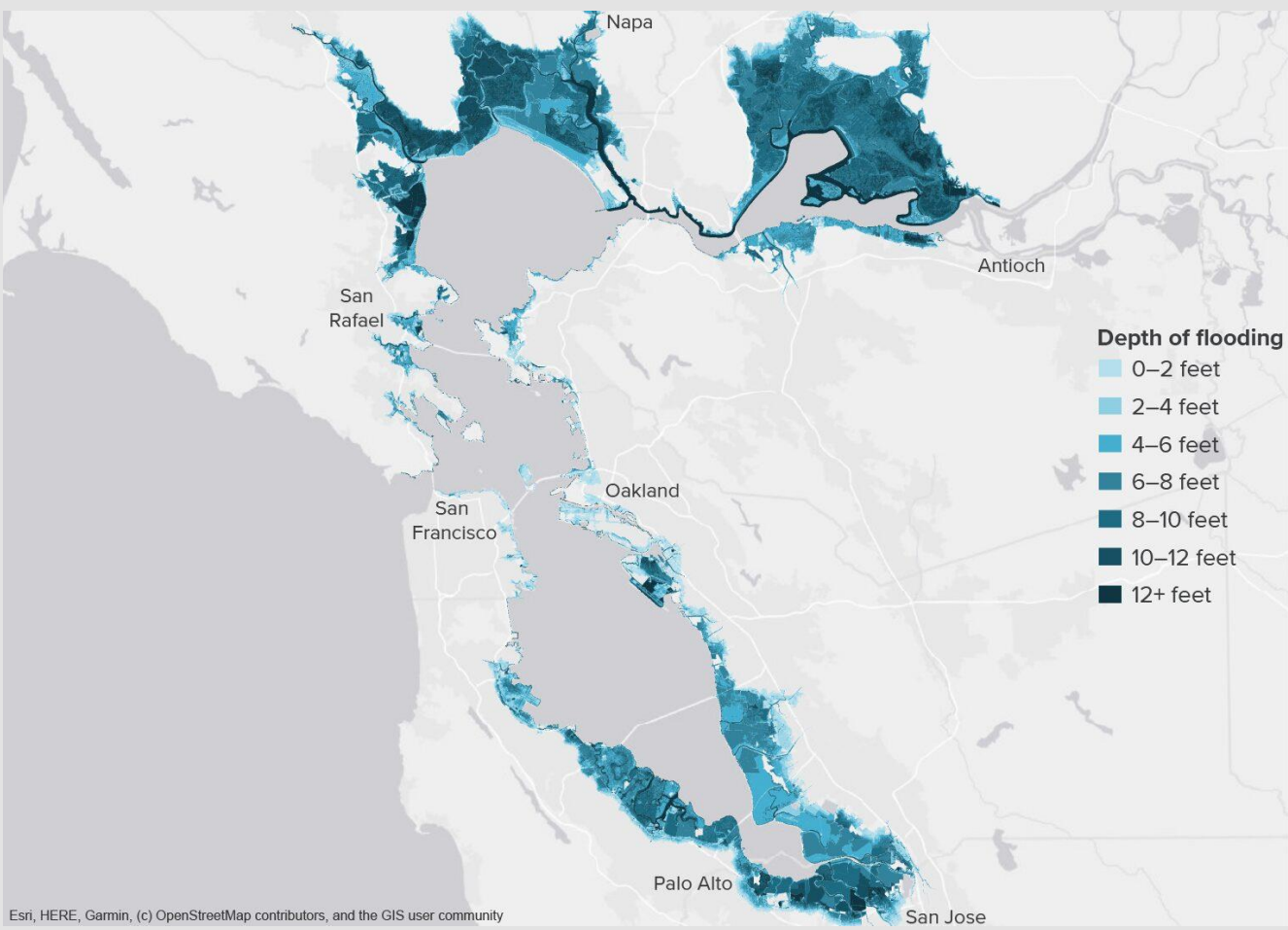
Hongxin Wu

Georgetown University

GEORGETOWN UNIVERSITY
1789

## 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 the strengths of each, 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 and aids in the planning and execution of long-term adaptation strategies. This initiative not only advances our understanding of coastal dynamics under changing climatic conditions but also contributes to the broader field of environmental science by implementing ethical AI practices and promoting the sustainable use of technological resources.



Depth of flooding
- 0–2 feet
- 2–4 feet
- 4–6 feet
- 6–8 feet
- 8–10 feet
- 10–12 feet
- 12+ feet

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.
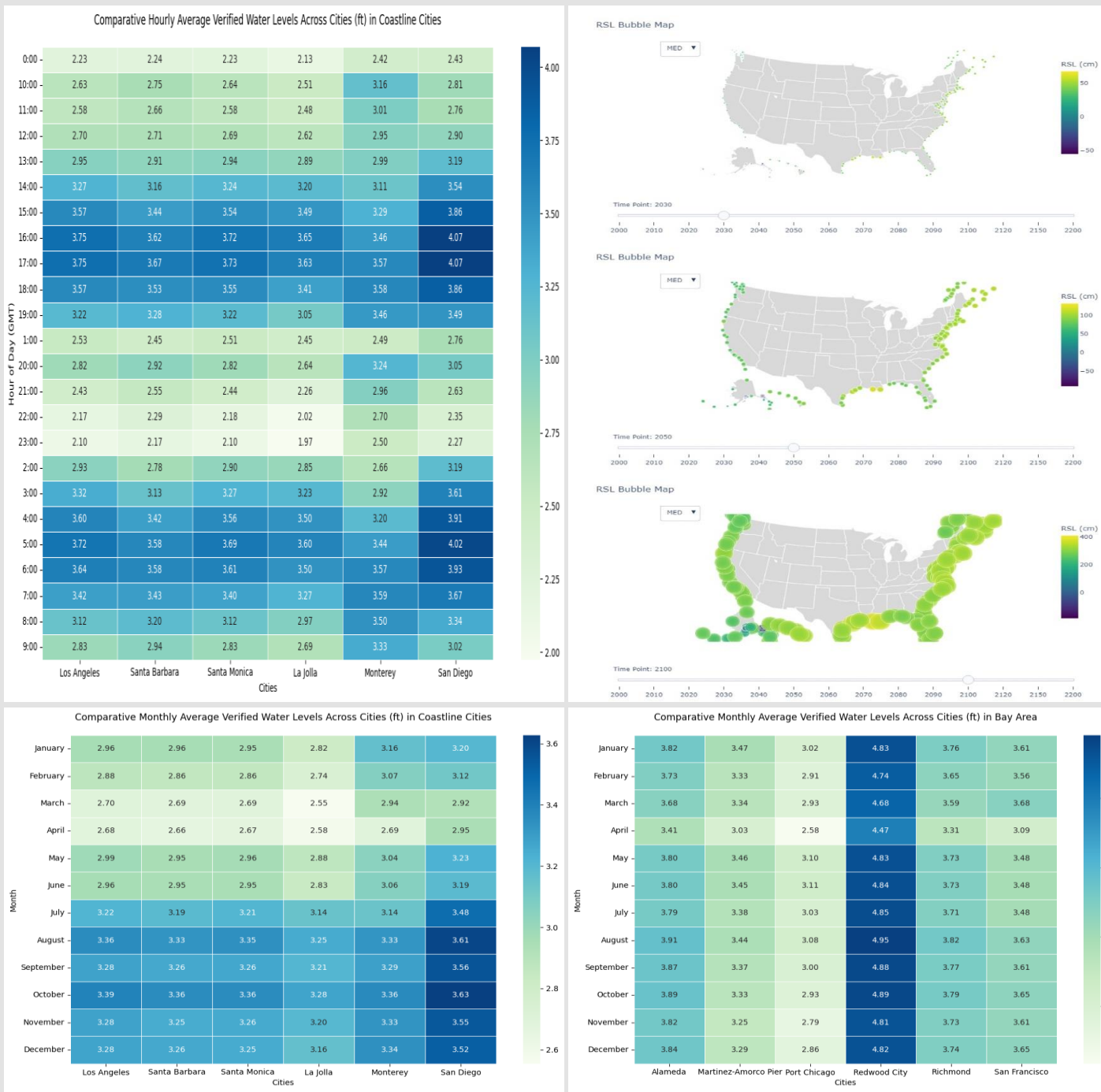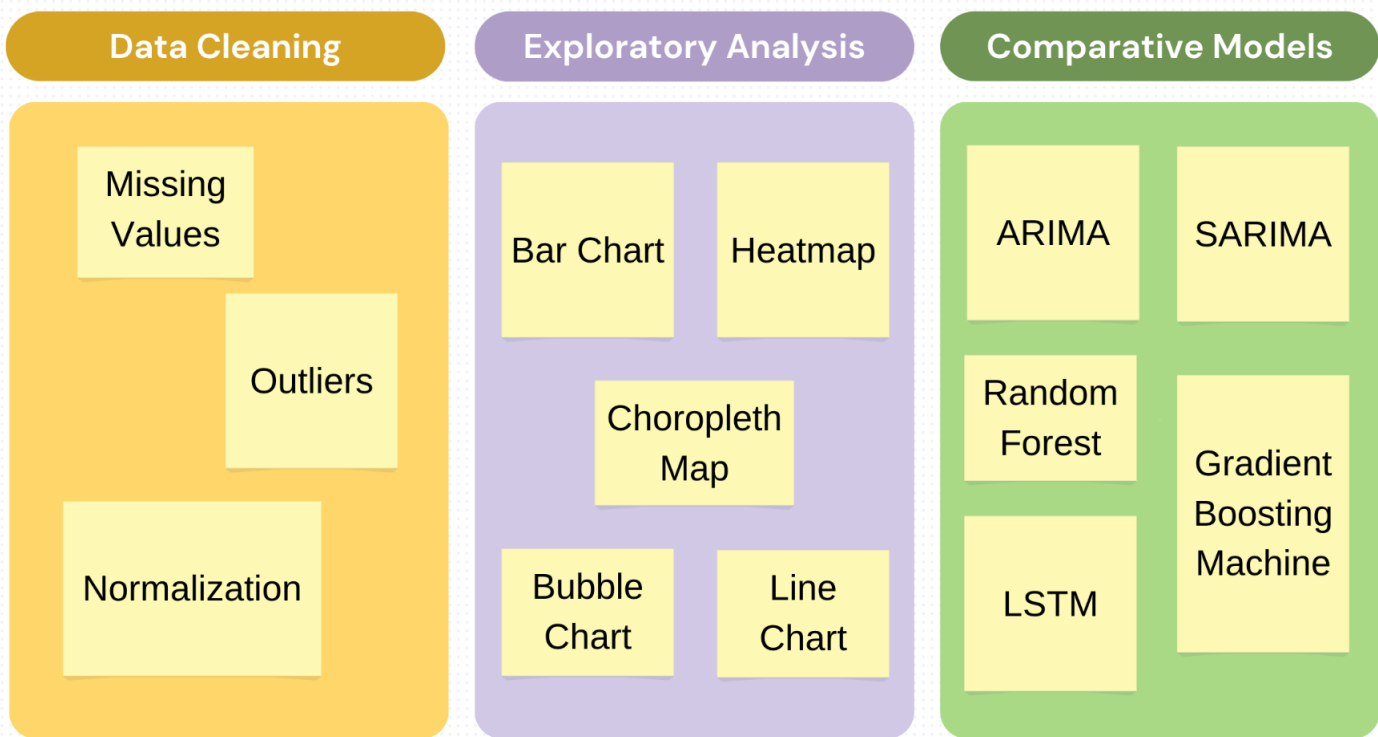
### Data Overview

| 137088 | 12 | 456 |
|---|---|---|
| Data Points | Stations | Days |

## Methodology

| Data Cleaning | Exploratory Analysis | Comparative Models |
|---|---|---|
| Missing Values | Bar Chart / Heatmap | ARIMA / SARIMA |
| Outliers | Choropleth Map | Random Forest / Gradient Boosting Machine |
| Normalization | Bubble Chart / Line Chart | LSTM |



## 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.
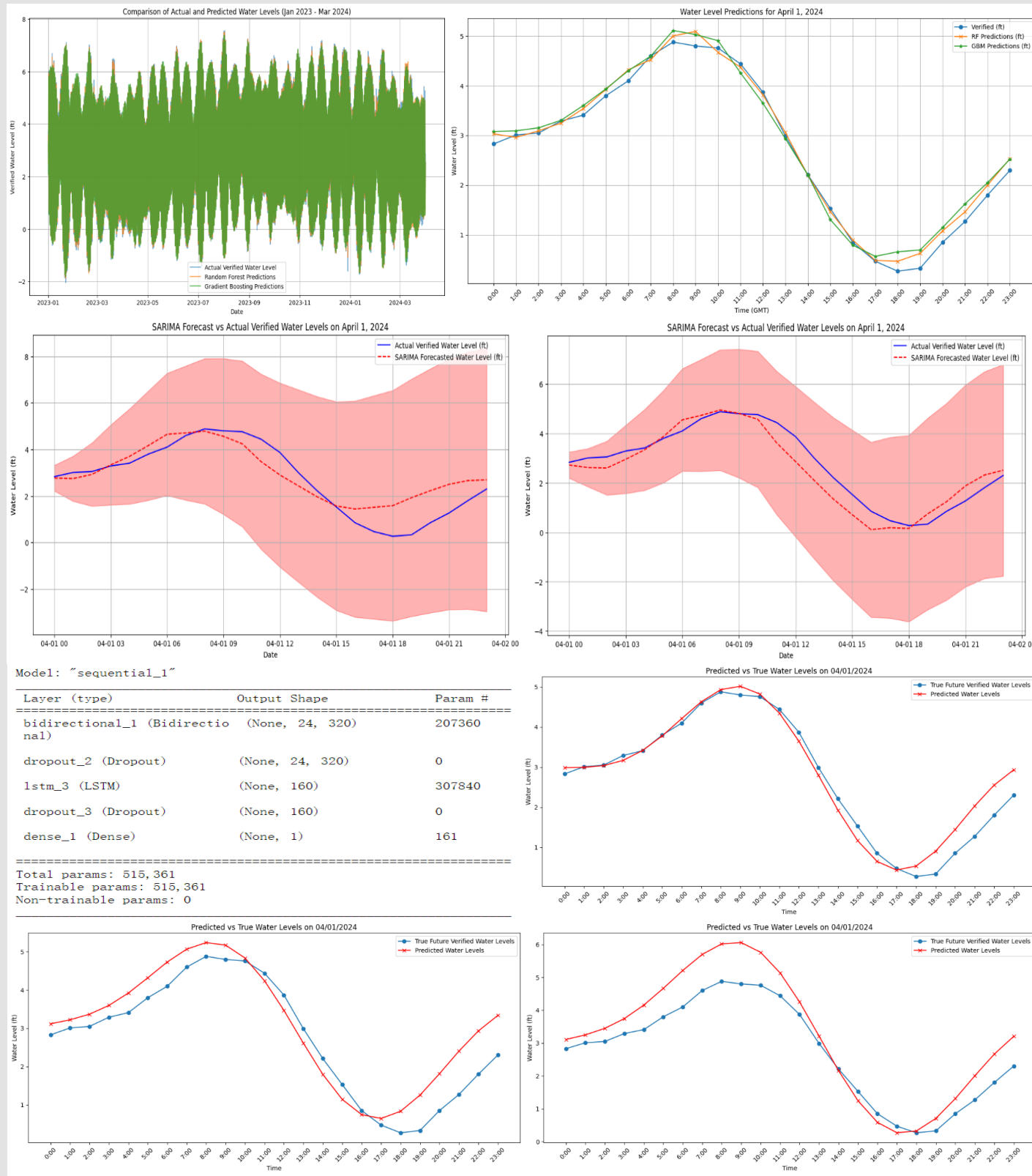


| Model Name | RMSE | MAE | R-squared |
|---|---|---|---|
| SARIMA(1,1,1,24) | 0.742 | 0.577 | 0.754 |
| SARIMA(1,1,2,24) | 0.514 | 0.416 | 0.882 |
| Random Forest | 0.155 | 0.129 | 0.989 |
| Gradient Boosting Machine | 0.211 | 0.180 | 0.980 |
| LSTM - Original Model | 0.691 | 0.587 | 0.786 |
| LSTM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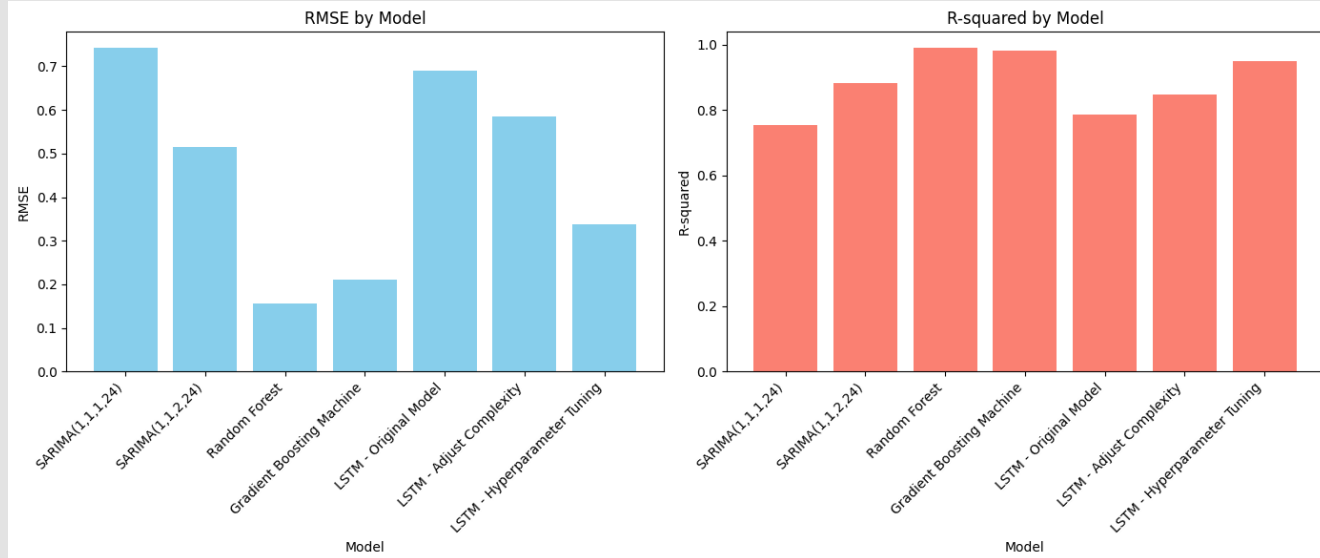


## Conclusion

### 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