Department of Technology Management and Innovation

Capstone Project Report

Water Quality Forecasting in the Biscayne Bay

Building Time Series and Machine learning models for predicting and testing quality on predefined standards

Team members : Anna Cui, Juanita Jacobraja, Keya Keya, Priya Kundu and Yifan Wu

Sponsored by WaterVue

Guidance: Prof. Shrutivandana Sharma, New York University

Date: May 8. 2024

Special Thanks to
Jose Sucre, Founder and CEO, WaterVue
Aliza Karim, Science and Research Manager, Miami Waterkeeper

Table of Contents

Abstract

The degradation of water quality in Biscayne Bay has become an urgent concern, exacerbated by the rapid pace of urbanization, agricultural activities, and the looming impacts of climate change. These factors collectively pose significant risks to the delicate balance of marine life and the well-being of local communities reliant on the bay's resources.

Recognizing the gravity of these challenges, our project endeavors to pioneer a comprehensive approach to address the complex dynamics of water quality in Biscayne Bay. Our primary objective is to develop and implement a robust predictive framework capable of accurately forecasting key water quality parameters, including chlorophyll levels, dissolved oxygen concentrations, and nutrient levels. By harnessing the power of advanced machine learning models such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Exponential Smoothing (SES), and Long Short-Term Memory (LSTM) networks, in conjunction with traditional time series analysis techniques, we aim to gain deeper insights into the intricate patterns governing water quality fluctuations.

Yet, our ambition extends beyond mere prediction; we aspire to empower strategic environmental management through the deployment of a sophisticated recommendation system. By leveraging the wealth of insights garnered from our predictive framework, this system will categorize distinct areas of the bay into clusters based on predefined water quality standards. Armed with this nuanced understanding, stakeholders will be equipped with targeted recommendations tailored to specific clusters, facilitating informed decision-making aimed at enhancing ecological resilience and safeguarding the long-term sustainability of Biscayne Bay's invaluable resources.

In essence, our project represents a proactive and forward-thinking approach to tackle the multifaceted challenges facing Biscayne Bay. By fusing cutting-edge technology with a steadfast commitment to environmental stewardship, we endeavor to pave the way towards a future where the bay thrives as a vibrant ecosystem, resilient to the pressures of urbanization and climate change, while continuing to enrich the lives of all who depend on its bounty.

1. Introduction

Background of the project

Biscayne Bay, nestled along the southeastern coast of Florida, serves as a vital ecosystem supporting diverse marine life and providing recreational and economic opportunities for surrounding communities. However, in recent years, the bay has faced escalating challenges due to deteriorating water quality, attributed to various factors including urbanization, agricultural runoff, and climate change impacts. These challenges have necessitated the implementation of robust strategies for monitoring and managing water quality to safeguard the health and sustainability of the bay ecosystem.

Importance of water quality forecasting and management in Biscayne Bay

Water quality forecasting and management play pivotal roles in ensuring the preservation and restoration of Biscayne Bay's ecological balance and functionality. Accurate predictions of water quality parameters such as chlorophyll levels, dissolved oxygen concentrations, salinity, and nutrient levels are essential for understanding ecosystem dynamics, identifying pollution sources, and implementing targeted interventions to mitigate adverse impacts.

By establishing a comprehensive understanding of water quality trends and dynamics, stakeholders can proactively implement measures to prevent and mitigate harmful algal blooms, hypoxic conditions, and other detrimental phenomena that pose threats to aquatic life and human well-being. Additionally, effective water quality management enhances the resilience of Biscayne Bay's ecosystems to environmental stressors, thereby safeguarding its ecological integrity and supporting the diverse array of species that depend on its habitats for survival.

Objectives of the project

The primary objective of this project is to develop a robust machine learning model for Biscayne Bay that pioneers precise water quality forecasting and management, laying the groundwork for scalable technology solutions. Specifically, the project aims to achieve the following objectives:

- Data Acquisition and Preprocessing: Collect comprehensive water quality data spanning multiple years and locations within Biscayne Bay, and preprocess the data to ensure its quality and reliability for subsequent analysis.
- Model Development and Training: Explore and compare various machine learning models, including ARIMA, SES, and LSTM, to develop an accurate predictive model capable of forecasting water quality parameters over time.
- Recommendation System for Better Water Quality: Utilize the insights gained from the predictive model to develop a recommendation system that identifies areas within Biscayne Bay requiring targeted interventions to improve water quality, thereby facilitating more effective management strategies.

By achieving these objectives, the project endeavors to provide stakeholders with actionable insights and tools to support informed decision-making and proactive management of Biscayne Bay's precious aquatic resources.

2. Data Acquisition and Preprocessing

Overview of the data sources and parameters

The dataset utilized in this project was provided by the sponsoring company, which comprises quarterly water testing results from 2006 to 2022 for 60 different locations within Biscayne Bay with about 19,400 rows . The dataset includes measurements of seven key water quality parameters: Chlorophyll A, Dissolved Oxygen, Salinity, Specific Conductance, Total Nitrogen, Total Phosphorus, and Turbidity. These parameters are crucial indicators of the bay's environmental health and serve as vital metrics for assessing water quality conditions.

	Site #	Location	Sample Date	Analysis code	Analysis	DQC	Result	Units	Detection Limit
0	1.0	#1 HILLSBORO CANAL US 1	2/22/06	Chl a	Chlorophyll a	NaN	12.100	mg/m3	0.955
1	1.0	#1 HILLSBORO CANAL US 1	2/22/06	Conductivity	Specific Conductance	NaN	31300.000	umho/cm	1.000
2	1.0	#1 HILLSBORO CANAL US 1	2/22/06	DO	Dissolved Oxygen	NaN	6.980	mg/L	0.050
3	1.0	#1 HILLSBORO CANAL US 1	2/22/06	Sal	Salinity	NaN	19.400	ppt	0.025
4	1.0	#1 HILLSBORO CANAL US 1	2/22/06	TN	Total Nitrogen	NaN	0.830	mg/L	0.107
...	...	...	...	...	...	...	...	...	...
19395	129.0	#129 Taft Street	11/15/21	Sal	Salinity	NaN	1.230	ppt	0.025
19396	129.0	#129 Taft Street	11/15/21	TP	Total Phosphorus	I	0.035	mg/L	0.013
19397	129.0	#129 Taft Street	11/15/21	Turbidity	Turbidity	NaN	1.400	NTU	0.080
19398	129.0	#129 Taft Street	11/15/21	TN	Total Nitrogen	NaN	0.821	mg/L	0.049
19399	129.0	#129 Taft Street	11/15/21	Chl a	Chlorophyll a	NaN	11.300	ug/L	0.125

19400 rows × 9 columns

Process of data acquisition

The sponsoring company supplied the dataset, which was obtained through extensive monitoring efforts conducted over the past decade. The data collection process involved regular sampling and testing of water quality parameters at designated locations throughout Biscayne Bay, utilizing state-of-the-art instrumentation and methodologies to ensure accuracy and reliability.

Cleaning and preprocessing steps:

- Outlier removal: To enhance the integrity of the dataset, outliers were identified and removed using statistical techniques to mitigate their potential impact on subsequent analyses and modeling efforts.
- We also did Data transformation and cleaning wherever needed, for example: logarithmic transformation to reduce skewness of values
- Resampling: Given the quarterly nature of the original data, resampling was conducted to aggregate the data into half-yearly intervals. This resampling approach helped preserve as much temporal information as possible while reducing the computational complexity of the analysis.
- Location selection: To maintain data consistency and reliability, locations with the most comprehensive and consistent data records were prioritized for inclusion in the analysis. Consequently, the initial set of 60 locations was narrowed down to 26 locations for further analysis.
- Handling missing values: Missing data points were addressed through interpolation techniques, wherein missing values were estimated based on the values of neighboring data points in time. This approach helped minimize data gaps and ensure the continuity of the dataset for subsequent analyses.

```
Maximum value of Chlorophyll A: 99.2
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 61.5
Minimum value of Dissolved Oxygen: 0.32
Maximum value of Salinity: 72.4
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 427119.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.5149999999999997
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.649
Minimum value of Total Phosphorus: -0.147
Maximum value of Turbidity: 23.0
Minimum value of Turbidity: 0.0
```

Range of the Parameters
Pre - Removal of
Outliers

```
Maximum value of Chlorophyll A: 14.6
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 9.39
Minimum value of Dissolved Oxygen: 0.82
Maximum value of Salinity: 37.3
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 58800.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.38
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.147
Minimum value of Total Phosphorus: -0.047
Maximum value of Turbidity: 3.4
Minimum value of Turbidity: 0.0
```

Range of the Parameters
Post - Removal of
Outliers

```
Maximum value of Chlorophyll A: 99.2
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 61.5
Minimum value of Dissolved Oxygen: 0.32
Maximum value of Salinity: 72.4
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 427119.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.5149999999999997
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.649
Minimum value of Total Phosphorus: -0.147
Maximum value of Turbidity: 23.0
Minimum value of Turbidity: 0.0
```

Range of the Parameter
Pre Log Transformation

```
Maximum value of Chlorophyll A: 14.6
Minimum value of Chlorophyll A: -0.048
Maximum value of Dissolved Oxygen: 9.39
Minimum value of Dissolved Oxygen: 0.82
Maximum value of Salinity: 37.3
Minimum value of Salinity: 0.0
Maximum value of Specific Conductance: 58800.0
Minimum value of Specific Conductance: 37.5
Maximum value of Total Nitrogen: 2.38
Minimum value of Total Nitrogen: 0.0
Maximum value of Total Phosphorus: 0.147
Minimum value of Total Phosphorus: -0.047
Maximum value of Turbidity: 3.4
Minimum value of Turbidity: 0.0
```

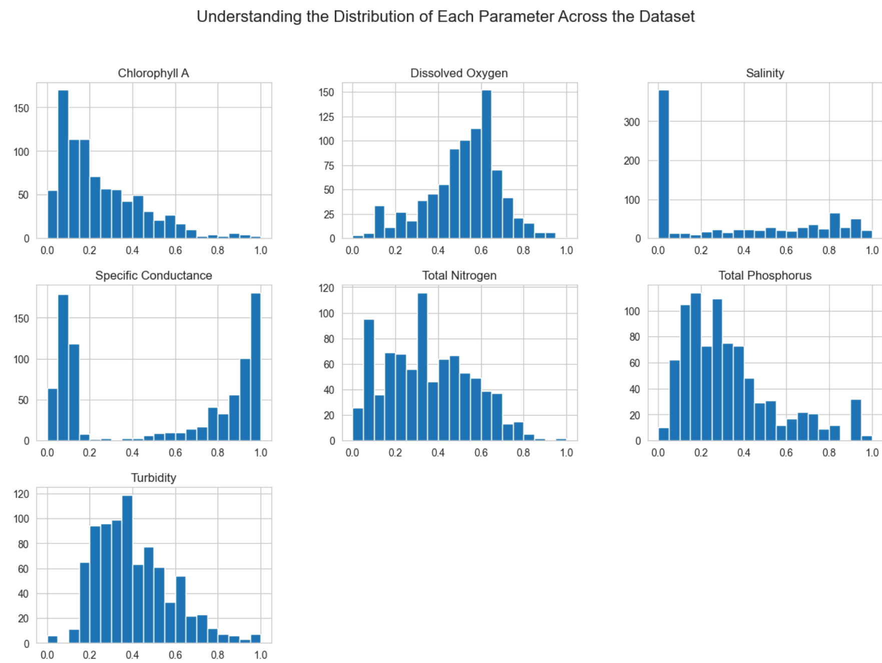
Range of the Parameter
Post Log Transformation

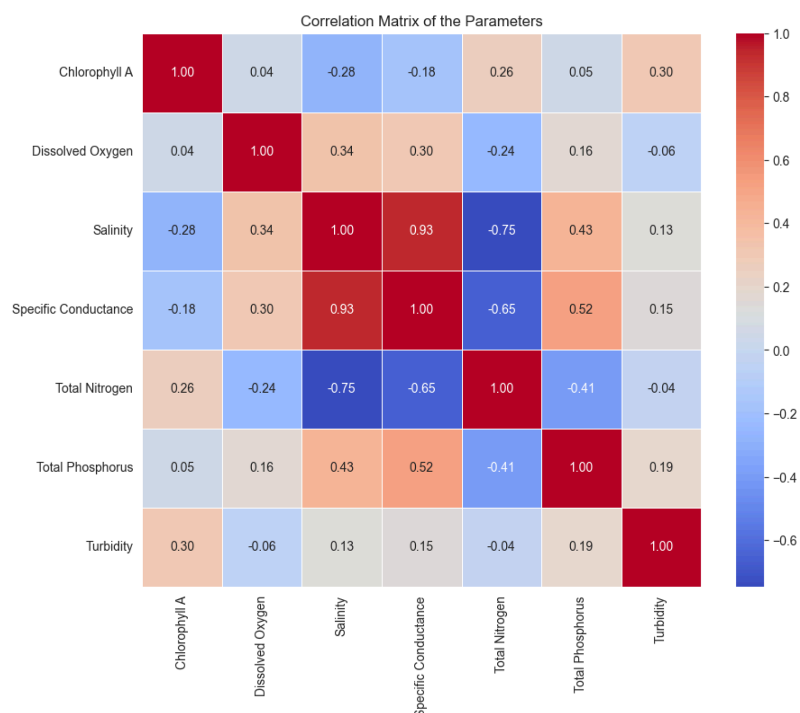
	Location	Sample Date	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
0	#1 HILLSBORO CANAL US 1	2006-02-28	12.100000	6.980000	19.400000	10.351373	0.830000	0.086000	2.400000
1	#1 HILLSBORO CANAL US 1	2006-08-31	4.430000	5.540000	15.750000	10.165844	0.981000	0.109000	1.400000
2	#1 HILLSBORO CANAL US 1	2007-02-28	2.605000	6.195000	23.200000	10.501905	0.754000	0.083500	1.650000
3	#1 HILLSBORO CANAL US 1	2007-08-31	4.890000	4.730000	31.100000	10.774781	0.777000	0.094000	2.300000
4	#1 HILLSBORO CANAL US 1	2008-02-29	5.925000	5.770000	12.500000	9.893361	1.440000	0.102000	2.050000
...	...	...	...	...	...	...	...	...	...
1559	#89 NOB HILL RD POMPANO CANAL	2020-02-29	3.933333	5.703333	0.310000	6.451930	1.097133	0.014000	0.750000
1560	#89 NOB HILL RD POMPANO CANAL	2020-08-31	1.570000	7.760000	0.250000	6.265301	0.910000	0.011000	0.000000
1561	#89 NOB HILL RD POMPANO CANAL	2021-02-28	2.966667	5.833333	0.233333	6.153383	0.984033	0.003667	0.566667
1562	#89 NOB HILL RD POMPANO CANAL	2021-08-31	5.510000	5.370000	0.270000	6.324359	1.410500	0.049000	0.775000
1563	#89 NOB HILL RD POMPANO CANAL	2022-02-28	2.900000	4.940000	0.290000	6.388561	1.040000	0.008000	0.550000

858 rows × 9 columns

Final Dataset for Model
Training and Testing

- Data exploration through visualization techniques: Exploratory data analysis techniques, including data visualization and graphical representations, were employed to gain insights into the nature and patterns of the dataset. Visualizations such as time series plots, histograms, and scatter plots were utilized to identify trends, correlations, and anomalies within the data, providing valuable insights for subsequent modeling and analysis phases. We also clustered every parameter by location for analyzing the levels in the water.





By implementing these preprocessing steps, the dataset was refined and prepared for further analysis, laying the foundation for the development of predictive models and recommendation systems aimed at enhancing water quality forecasting and management in Biscayne Bay.

3. Model Development and Training

Explanation of the modeling approach

In this phase of the project, various machine learning models were explored and trained to develop accurate predictive models for forecasting water quality parameters in Biscayne Bay. The modeling approach involved leveraging historical water quality data to train the models, which would then be used to predict future trends and fluctuations in key parameters. The ultimate goal was to develop models capable of providing reliable forecasts to support proactive water quality management strategies.

Description of the models explored:

- ARIMA (Autoregressive Integrated Moving Average): ARIMA is a time series forecasting method that combines autoregression, differencing, and moving average components to model and predict future values based on past observations. This model is particularly well-suited for capturing temporal dependencies and trends in sequential data, making it a valuable tool for forecasting water quality parameters over time.
- SES (Simple Exponential Smoothing): SES is a basic exponential smoothing method that assigns exponentially decreasing weights to past observations, with more recent data points receiving higher weights. This model is straightforward and computationally efficient, making it suitable for simple time series forecasting tasks where the data exhibit constant or slowly changing trends.
- LSTM (Long Short-Term Memory): LSTM is a type of recurrent neural network (RNN) architecture designed to capture long-term dependencies and patterns in sequential data. Unlike traditional feedforward neural networks, LSTM networks incorporate memory cells and gates that enable them to retain and propagate information over extended time intervals. This makes LSTM well-suited for modeling complex temporal dynamics and non-linear relationships in time series data, making it a powerful tool for water quality forecasting tasks.
- Training methodology

Training the machine learning models:

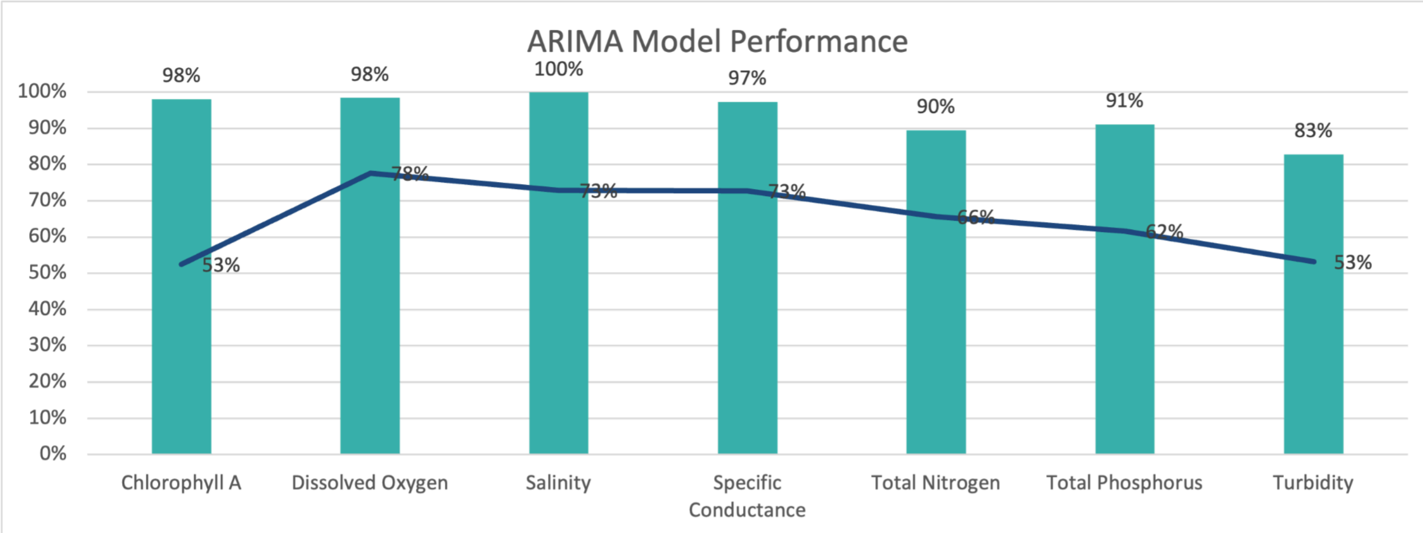
- Data splitting: The dataset was divided into training and validation sets, with the training set containing historical data used to train the models, and the validation set reserved for evaluating model performance.
- Model training: Each machine learning model (ARIMA, SES, LSTM) was trained using the training data, wherein the model parameters were optimized to minimize prediction errors and maximize predictive accuracy. Training involved iterative optimization techniques, such as gradient descent or backpropagation, depending on the specific model architecture.
- Model evaluation: After training, the performance of each model was evaluated using the validation set, wherein model predictions were compared against actual observed values. Evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) were computed to assess the accuracy and reliability of the models.
- Hyperparameter tuning: To further optimize model performance, hyperparameter tuning techniques were employed to fine-tune model parameters and configurations. This involved experimenting with different parameter values and model architectures to identify the optimal configuration that yielded the best predictive performance.
- By following this training methodology, the machine learning models were trained and optimized to develop accurate predictive models capable of forecasting water quality parameters in Biscayne Bay, thereby providing valuable insights for proactive water quality management and decision-making.

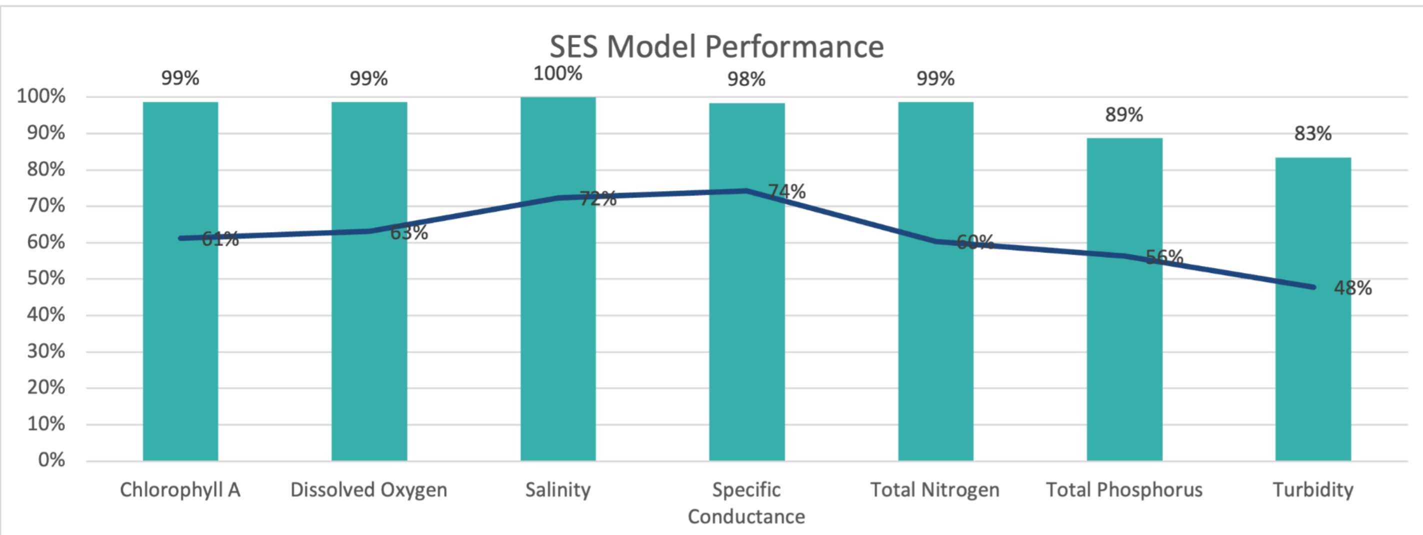
Model Performance:

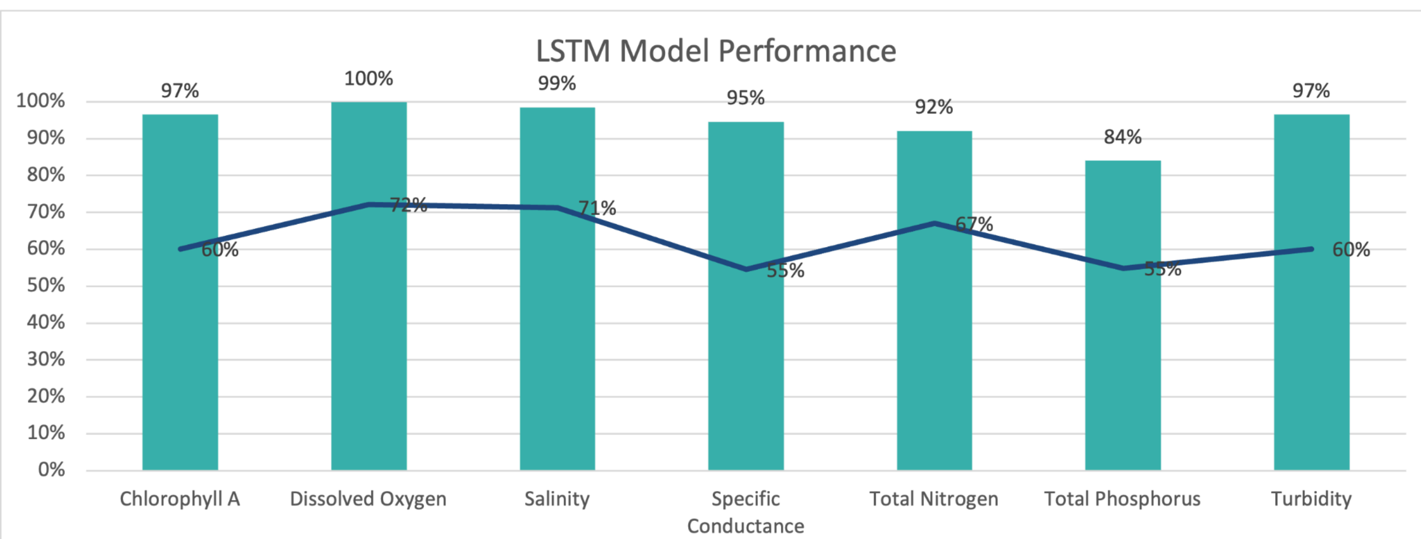
ARIMA							
	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
mean	**53%**	**78%**	**73%**	**73%**	**66%**	**62%**	**53%**
std	27%	21%	32%	30%	20%	26%	22%
max	98%	98%	100%	97%	90%	91%	83%
SES							
	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
mean	**61%**	**63%**	**72%**	**74%**	**60%**	**56%**	**48%**
std	25%	25%	33%	31%	25%	28%	22%
max	99%	99%	100%	98%	99%	89%	83%
LSTM							
	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
mean	**60%**	**72%**	**71%**	**55%**	**67%**	**55%**	**60%**
std	24%	30%	29%	26%	24%	22%	24%
max	97%	100%	99%	95%	92%	84%	97%

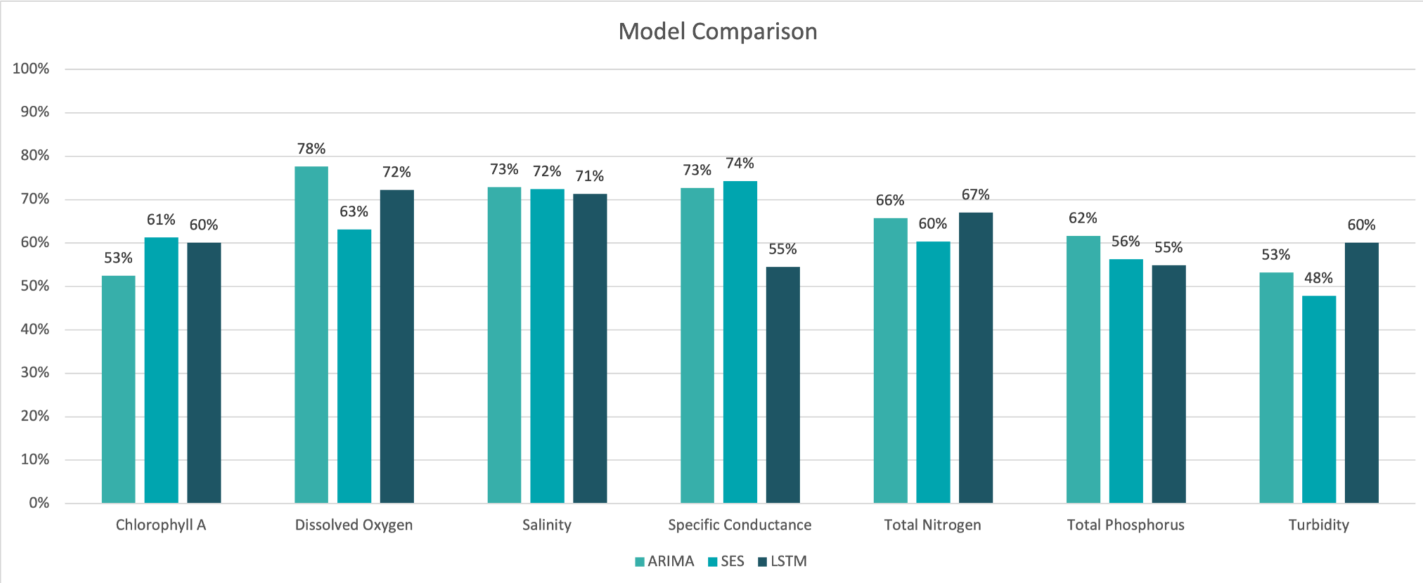
Model Comparison:

	Chlorophyll A	Dissolved Oxygen	Salinity	Specific Conductance	Total Nitrogen	Total Phosphorus	Turbidity
ARIMA	**53%**	**<u>78%</u>**	**<u>73%</u>**	**73%**	**66%**	**<u>62%</u>**	**53%**
SES	<u>61%</u>	63%	72%	<u>74%</u>	60%	56%	48%
LSTM	60%	72%	71%	55%	<u>67%</u>	55%	<u>60%</u>


ARIMA Model Performance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
98%
98%
100%
97%
90%
91%
83%
53%
78%
73%
73%
66%
62%
53%
Chlorophyll A
Dissolved Oxygen
Salinity
Specific Conductance
Total Nitrogen
Total Phosphorus
Turbidity


SES Model Performance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
99%
99%
100%
98%
99%
89%
83%
61%
63%
72%
74%
60%
56%
48%
Chlorophyll A
Dissolved Oxygen
Salinity
Specific Conductance
Total Nitrogen
Total Phosphorus
Turbidity


LSTM Model Performance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
97%
100%
99%
95%
92%
84%
97%
60%
72%
71%
55%
67%
55%
60%
Chlorophyll A
Dissolved Oxygen
Salinity
Specific Conductance
Total Nitrogen
Total Phosphorus
Turbidity



Insights Gained from the Models:

Consideration of standard deviation alongside mean values provides a more comprehensive understanding of the model performance. While ARIMA often exhibited lower mean values, it frequently demonstrated lower standard deviations, indicating greater consistency in its predictions. Conversely, SES and LSTM occasionally achieved higher mean values but showed higher variability in their predictions, suggesting less stability in their forecasts. Therefore, the choice of model should be made based not only on mean performance but also on the desired level of prediction consistency and variability.

- ARIMA: ARIMA consistently demonstrated competitive performance across multiple water quality parameters, particularly in predicting Dissolved Oxygen and Total Phosphorus. Its reliance on historical data and time series analysis techniques proved effective in capturing temporal patterns and trends, making it well-suited for forecasting tasks.
- SES: SES exhibited mixed performance across different parameters, with notable strengths in predicting Chlorophyll A and Specific Conductance. Its simplicity and computational efficiency make it an attractive option for straightforward forecasting tasks, although it may lack the complexity to capture nuanced relationships in the data.
- LSTM: LSTM showed promising results, especially in predicting Total Nitrogen and Turbidity. Its ability to capture long-term dependencies and nonlinear relationships in sequential data proved advantageous for modeling complex water quality dynamics. However, LSTM's performance varied across parameters, indicating the need for further refinement and optimization.

Overall, each model offered unique strengths and limitations, highlighting the importance of selecting the most appropriate model based on the specific characteristics of the data and the forecasting task at hand. Further analysis and refinement of the models could enhance their predictive capabilities and contribute to more accurate water quality forecasting and management strategies in Biscayne Bay.

4. Recommendation System for Better Water Quality

Overview of the Recommendation System:

The recommendation system aims to enhance water quality management in Biscayne Bay by providing targeted interventions and strategies for improvement based on the analysis of predicted Dissolved Oxygen levels at various locations. By clustering locations into distinct categories based on predefined water quality standards and considering the mean error of prediction, stakeholders can prioritize resources and implement measures to address specific challenges and improve overall water quality in the bay effectively.

Research on Water Quality Standards:

The recommendation system is informed by established water quality standards, which serve as benchmarks for assessing the health and condition of aquatic ecosystems. In this context, standards for Dissolved Oxygen levels are utilized to categorize locations into different groups based on their compliance with designated ranges:

- Anoxic range (0.0-0.5 mg/L): Indicates severely depleted Dissolved Oxygen levels, posing significant risks to aquatic life and ecosystem health.
- Hypoxic range (0.6-2.0 mg/L): Represents suboptimal Dissolved Oxygen levels that may lead to physiological stress and reduced habitat suitability for aquatic organisms.
- Biological Stress range (2.1-5.0 mg/L): Signifies conditions where Dissolved Oxygen levels may support basic biological functions but may limit growth and reproduction for some species.
- Optimal range (5.0+ mg/L): Indicates healthy Dissolved Oxygen levels capable of supporting diverse aquatic life and ecosystem functions.

Additionally, the mean error of prediction is considered to evaluate the reliability of predictions and refine recommendations for improvement.

Creation of Location Groups Based on Water Quality Categories:

Utilizing the predicted mean Dissolved Oxygen values and mean error of prediction for each location, the recommendation system categorizes locations into groups based on their alignment with predefined water quality standards. Categories include:

- **Biological Stress:** Locations with Dissolved Oxygen levels falling within the Biological Stress range (2.1-5.0 mg/L) and a relatively low mean error of prediction, indicating suboptimal conditions for aquatic life but with predictions that are more reliable.
- **Optimal**: Locations with Dissolved Oxygen levels within the Optimal range (5.0+ mg/L) and low mean error of prediction, indicating healthy conditions for aquatic organisms and reliable predictions.

Clustering Methodology:

Clustering techniques are employed to group locations with similar Dissolved Oxygen profiles and mean errors of prediction together. This approach facilitates targeted interventions tailored to the specific needs of each group, considering both predicted water quality conditions and the reliability of predictions.

Results after Clustering based on the pre-defined company standards:

Locations	DO Range	Mean DO	Category	Mean Error
#1 HILLSBORO CANAL US 1	(4.66, 4.95)	4.853333	Biological Stress	0.172349
#10 MIDDLE RIVER E SUNRISE	(5.247, 5.35)	5.281111	Optimal	0.22339
#110 POMPANO CANAL AT DIXIE AN	(3.66, 5.0)	4.508889	Biological Stress	0.738609
#111 S. FORK MID R. @ N.E. 15	(4.15, 5.645)	4.691667	Biological Stress	0.911607
#12 MIDDLE RIVER NW 31ST AVE	(4.535, 6.345)	5.353333	Optimal	1.268651
#15 NEW RIVER ANDREWS AVE	(3.847, 5.84)	5.118889	Optimal	0.894132
#17 PLANTATION CANAL @ S-33	(3.7, 4.53)	3.976667	Biological Stress	1.883363
#19 NEW RIVER RIVER REACH	(3.173, 4.42)	3.904444	Biological Stress	0.770735
#22 N NEW RIVER SW 125 AVE	(0.9, 3.62)	2.636667	Biological Stress	1.394191
#24 DANIA CUT-OFF US-1	(3.69, 4.895)	4.231667	Biological Stress	0.62353
#25 HOLLYWOOD CANAL STIRLING	(2.925, 4.78)	3.858333	Biological Stress	0.949167
#28 S NEW RIVER CANAL FLAMINGO	(2.617, 5.17)	3.467778	Biological Stress	1.432569
#31 SNAKE CRK CANAL FLAMINGO	(2.03, 3.377)	2.645556	Biological Stress	0.90474
#32 C-9 CANAL US 27	(1.59, 3.98)	2.604444	Biological Stress	4.130915
#36 ICW COMMERCIAL BLVD	(5.76, 6.053)	5.955556	Optimal	0.137776
#37 ICW SUNRISE BLVD	(5.56, 5.73)	5.673333	Optimal	0.080323
#38 ICW 17TH ST CAUSEWAY	(6.033, 6.08)	6.048889	Optimal	0.064447
#39 ICW MARKER 35	(4.9, 5.2)	5.1	Optimal	0.449001
#4 HILLSBORO CANAL SE GROWERS	(2.63, 3.82)	3.056667	Biological Stress	2.139645
#40 ICW SHERIDAN ST	(5.32, 5.59)	5.5	Optimal	0.103221
#41 ICW HALLANDALE BCH BLVD	(5.47, 6.78)	6.343333	Optimal	0.780887
#5 POMPANO CANAL US1	(4.94, 6.24)	5.59	Optimal	0.984356
#6 CYPRESS CREEK DIXIE HWY	(5.33, 6.25)	5.83	Optimal	0.557735
#7 CYPRESS CREEK S. PALM AIRE	(4.58, 6.025)	5.445	Optimal	0.875108

#8 POMPANO CANAL US 441	(3.57, 6.05)	5.218889	Optimal	1.08835
#89 NOB HILL RD POMPANO CANAL	(4.94, 5.833)	5.381111	Optimal	0.458105

Discussion on Recommendations for Improvement:

Emphasizing the mean error of prediction alongside Dissolved Oxygen levels allows for more nuanced recommendations:

Biological Stress Locations:

- Prioritize interventions in areas with both suboptimal Dissolved Oxygen levels and low mean error of prediction to address immediate water quality concerns effectively.
- Implement management strategies to improve water quality and ecosystem health, considering the reliability of predictions to guide decision-making.

Optimal Locations:

- Ensure continued monitoring and maintenance of water quality parameters to sustain favorable conditions and prevent potential degradation, utilizing reliable predictions to inform management practices.
- Implement proactive measures to preserve optimal water quality conditions and prevent potential deterioration, considering the reliability of predictions for long-term planning.

By integrating the mean error of prediction into the recommendation system, stakeholders can make more informed decisions and allocate resources effectively to improve water quality in Biscayne Bay, ultimately supporting the health and resilience of the bay's ecosystems and benefiting both aquatic life and human communities.

5. Conclusion

Summary of Key Findings:
Throughout this project, a comprehensive approach to water quality forecasting and management in Biscayne Bay has been undertaken, utilizing machine learning models, clustering techniques, and adherence to established water quality standards. Key findings from the analysis include:

- Model Performance: Comparative analysis of ARIMA, SES, and LSTM models revealed varying degrees of performance across different water quality parameters. While each model showed strengths in certain areas, their effectiveness varied, emphasizing the importance of selecting appropriate models tailored to specific forecasting tasks.
- Clustering and Categorization: Clustering locations based on predicted Dissolved Oxygen levels and mean errors of prediction allowed for the creation of distinct groups representing varying water quality conditions. This approach facilitated targeted interventions and management strategies to address specific challenges and improve overall water quality in the bay.
- Recommendations for Improvement: The recommendation system highlighted areas requiring immediate attention, particularly locations experiencing suboptimal Dissolved Oxygen levels and low mean errors of prediction. By prioritizing interventions in these areas and considering the reliability of predictions, stakeholders can make informed decisions to enhance water quality management efforts.

Outcomes of the Project:

- Robust Predictive Models: Development and evaluation of ARIMA, SES, and LSTM models provided valuable insights into water quality dynamics in Biscayne Bay, laying the foundation for more informed decision-making and proactive management strategies.
- Clustering and Recommendation System: Implementation of clustering techniques and the recommendation system allowed for the identification of priority areas and the formulation of targeted interventions to address water quality challenges effectively.
- Integration of Water Quality Standards: Incorporation of established water quality standards into the analysis provided a framework for assessing and categorizing locations based on their compliance with predefined Dissolved Oxygen ranges, guiding management actions and resource allocation.

Recommendations for Future Work:

- Model Refinement: Continued refinement and optimization of machine learning models to improve predictive accuracy and reliability, considering additional variables and features that may influence water quality dynamics.
- Enhanced Data Collection: Expansion of the dataset to include more comprehensive and diverse sources of water quality data, incorporating real-time monitoring and sensor technologies to capture temporal and spatial variability.
- Integration of Stakeholder Feedback: Engagement with local stakeholders and community members to incorporate their insights and priorities into water quality management efforts, fostering collaboration and collective action for the preservation of Biscayne Bay's ecosystems.

In conclusion, this project has provided valuable insights and tools to support proactive water quality management in Biscayne Bay, laying the groundwork for continued efforts to safeguard the health and resilience of this vital aquatic ecosystem. By leveraging advanced modeling techniques, clustering methodologies, and adherence to established standards, stakeholders can work together to address water quality challenges and ensure the long-term sustainability of Biscayne Bay for future generations.