ENHANCING SALINITY PREDICTION IN THE NEUSE RIVER ESTUARY VIA MACHINE LEARNING MODELS

Mina Gachloo

A Thesis Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science

University of North Carolina Wilmington

2024

Approved by

Advisory Committee

Karl Ricanek	Qianqian Liu
Leo Emre Gokce	Yang Song Chair

Accepted by

Dean, Graduate School

TABLE OF CONTENTS

ABSTRACT

The escalating threat to ocean water quality places strain on essential marine water resources, fishery habitats, and ecosystems. Salinity is one of the key indicators of water quality, offering valuable information about the exchange between coastal seas, rivers, and watersheds. The main object of this research is to predict water salinity and identify its influencing factors in estuarine and coastal waters, taking the Neuse River Estuary (NRE) in North Carolina as an example. This study was conducted at 11 mid-river sampling stations and involved comparing three machine learning models: Random Forest, Multiple Linear Regression, and Multi-Output Regressor with Gradient Boosting Regression (MOGBR), to predict salinity at various depths. The input predictors to our prediction models include aggregated river discharge, aggregated sea level, and aggregated wind based on eight directions. By prioritizing the most significant predictors, we streamlined the model-building process and developed a hindcast system covering the years 1994 to 2024. The methodology was divided into two phases: the first phase involved feature engineering and model selection, identifying MOGBR as the most effective model for predicting salinity across multiple depths and stations. In the second phase, the selected model was applied to predict seasonal salinity variations, enabling a comparative analysis of ground truth inputs, actual measurements, and predicted values. Results showed that the MOGBR model effectively captured spatial and seasonal salinity trends, with improved R^2 values across depths and stations. These findings demonstrate the utility of machine learning models in advancing salinity prediction and supporting coastal water management efforts.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

In many parts of the world, rivers and estuaries provide the principal sources of water for irrigation, human consumption, and the habitats of aquatic species [1,2]. In addition, these water sources are one of the important pathways for the transport and distribution of substances, such as dissolved nutrients or pollutants between surrounding lands and coastal seas. As an essential marker of these processes, salinity fluctuates based on the combined effects of riverine inputs, meteorological conditions, and tidal movements [3]. The average salinity of seawater is around 35 g/L, but this value can be significantly different in various ocean regions. Both low and high salinity levels can impact water quality and estuary ecosystem health; therefore, forecasting salinity can enhance ecosystem restoration and management to alleviate the adverse effects caused by these changes [4,5].

Various models, including physical or mathematical like QUAL2K and MOUSE, have been developed to assist in predicting and managing salinity, but using these models can be challenging due to their complexity and time-consuming process, especially in developing countries where such resources and expertise are often limited [1]. Statistical water quality models have effectively utilized linear and non-linear relationships to analyze features such as salinity, dissolved oxygen, and flow rates. However, in coastal environments where water quality is influenced by multivariate factors like tidal dynamics, riverine inputs, and anthropogenic pressures, these models often fail to represent the underlying complexity. This limitation is particularly significant for estuarine systems, where interactions between physical, chemical, and biological processes exhibit strong non-linear behavior that challenges traditional modeling techniques [6,7].

Recently, machine learning has been increasingly recognized, particularly for its impressive performance in modeling the non-linear processes of engineering systems. In groundwater studies, many researchers have adopted machine learning methods due to their efficiency, requiring fewer input data and computational resources while achieving results comparable to traditional approaches [8]. In this paper, we aim to enhance the accuracy of salinity predictions in the Neuse River Estuary (NRE) in North Carolina, USA, by evaluating the performance of machine learning models at various depths.

We conduct a comprehensive comparison of three machine learning models: Random Forest, Multiple Linear Regression (MLR), and a combination of Multi-Output and Gradient Boosting Regression (MOGBR), applied at 11 sampling stations across different depths.

To develop machine learning models, we explored multiple combinations of input features derived from datasets from the NRE Modeling and Monitoring program (ModMon [9]), river discharge data, and NOAA NDBC meteorological observations [10]. We utilized hindcast models to select the most relevant aggregated features and validate the salinity prediction models. The hindcast approach relied on real observed data, ensuring accurate feature selection and reliable model validation. For real-time forecasts, model-predicted meteorological and river discharge would be utilized.

This research aims to uncover the complexities of the NRE, offering predictive solutions for salinity and identifying key factors affecting water quality. In Chapter 2, we discuss related work. Chapter 3 presents a comprehensive explanation of the datasets, data processing techniques, feature engineering, and model selection based on results and

accuracy to guide the next steps of our project. Chapter 4 focuses on implementing the best model for seasonal salinity forecasting.

CHAPTER 2: RELATED WORK

Salinity, a fundamental oceanographic parameter, measures the total concentration of dissolved salts in seawater and plays a crucial role in understanding coastal and marine ecosystems. Salinity levels are shaped by different factors, including freshwater inputs, tidal mixing, evaporation, and anthropogenic influences, making it an essential indicator of environmental health and water quality [11-12]. The NRE is a sub-estuary within the Pamlico Sound estuarine system, recognized as the second-largest estuarine complex in the United States [13]. As one of the largest estuarine systems in the U.S., the Pamlico Sound supports nearly half of the nursery areas for commercially significant fish species along the East Coast [14].

The NRE, a critical fisheries habitat and recreational resource, requires precise salinity predictions to ensure ecological balance and water quality. Predicting salinity is particularly important because shifts in salinity can lead to stress on aquatic organisms, including economically important fish species [14]. The NRE often experiences significant salinity differences between surface and bottom waters, as freshwater from rivers flows above denser, saltier seawater from Pamlico Sound [15]. This information supports decision-making by resource managers and fisheries authorities [13]. Ignoring depth in salinity predictions can result in inaccuracies, as surface and bottom salinity levels can vary greatly due to factors like freshwater inflow, tidal mixing, and stratification [16-17]. Addressing these depth-related salinity variations is crucial for improving the accuracy of predictions and supporting the effective management of this essential estuarine system [18].

In 2003, Wool et al. [19] utilized a three-dimensional modeling approach (EFDC) to investigate salinity levels at varying depths within the NRE. In this project, by using the

EFDC hydrodynamic model, they analyzed how freshwater inflows reduce salinity in surface layers, while denser saltwater influences deeper levels. The model incorporated observational data collected between 1998 and 2000 from various sources, including the USGS, the University of North Carolina (MODMON project), and North Carolina State University, which included surface and bottom salinity, water surface elevation, temperature profiles, and nutrient concentrations. It also accounted for wind-driven mixing and tidal forces, which redistribute salinity between surface and bottom layers. This depth-specific approach, calibrated for 1998 and validated using data from 1999 to 2000, provided valuable insights into how salinity levels vary throughout the estuary, supporting efforts to manage water quality and predict ecological impacts.

Traditional models like EFDC, while effective in simulating physical processes, often require extensive computational resources and detailed input data, which can limit their applicability in certain scenarios [20]. In recent years, machine learning techniques have been increasingly applied to salinity prediction, offering powerful tools for modeling salinity variations across spatial and temporal scales. For instance, Ahmad et al. [21] investigated the use of machine learning models, demonstrating their accuracy in predicting water quality for drinking purposes. El Bilali et al. [22] utilized machine learning models, including ANN, RF, and AdaBoost, to predict 10 irrigation water quality (IWQ) parameters, such as SAR and TDS, in Morocco's Bouregreg watershed using electrical conductivity (EC) and pH as input features. Their research demonstrated that most models achieved high accuracy during training and validation.

In 2022, Tran et al. [23] evaluated the performance of machine learning models for predicting saltwater intrusion using metrics such as the Nash-Sutcliffe efficiency coefficient,

Mean Absolute Error, and Root Mean Square Error. Similarly, in 2023, Guillou et al. [24] conducted a comprehensive study that applied machine learning algorithms to predict sea surface salinity in the Elorn estuary, located within the Bay of Brest in northwestern Europe. The study aimed to address the challenges of modeling salinity, a critical parameter for understanding estuarine and coastal ecosystem dynamics. This estuary is characterized by complex interactions between tidal advection, riverine inputs, and meteorological forcings, making it a compelling case study for predictive modeling.

The research by Guillou et al. utilized six years of in-situ salinity observations from 2015 to 2021 collected at the mouth of the estuary and tested several machine-learning techniques, including Multi-Layer Perceptron (MLP), Support Vector Regression (SVR), and Random Forest (RF). These models were trained on a range of input parameters, such as tidal free-surface elevation, river discharge, and wind velocity, to mimic the non-linear relationships governing salinity variations. To evaluate model performance, the authors used metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Normalized Root Mean Square Error (NRMSE), and R^2, providing a comprehensive assessment of predictive accuracy.

The SVR model was better than the other models and numerical models, particularly during challenging winter periods marked by high freshwater inflows and tidal variability, as shown in Table 1. Their findings presented that machine learning models were capable of capturing both seasonal cycles and tidal modulations of salinity.

Table 1. Comparison of Machine Learning Models in Guillou's paper

Model	MAE	RMSE	NRMSE	R^2
MLR	2.46	3.48	11.7%	0.29
Multiple Polynomial Regression (MPR)	2.33	3.14	10.5%	0.49
MultiLayer Perceptron (MLP)	2.42	3.26	10.9%	0.48
Support Vector Regression (SVR)	2.26	3.16	10.6%	0.51
Random forest	2.44	3.32	11.1%	0.46
Model for Application at Regional Scale (MARS)	2.29	3.73	12.5%	-2.52

The review of related work in salinity prediction demonstrates the suitability of machine learning models for addressing the inherent complexities of environmental systems. Their ability to effectively capture nonlinear interactions and adapt to diverse datasets establishes them as a powerful tool for improving the precision of salinity forecasts in estuarine and coastal regions, where accurate predictions are essential for sustainable management.

CHAPTER 3: FEATURE ENGINEERING AND MODEL SELECTION

3.1 Study Area

The NRE, as shown in Figure 1, is formed by the Neuse River, which drains the fourth-largest basin in North Carolina. The basin spans urban centers like Raleigh and Durham in the Piedmont and highly productive agricultural regions on the coastal plain, highlighting the interplay between urbanization and agriculture in shaping the estuary's water quality. The NRE, a drowned river valley about 70 kilometers long with an average depth of 3.5 meters, provides critical habitat for fish and wildlife in the region. Sustainable management of the NRE-Pamlico Sound ecosystems plays a crucial role in preserving the environment while also supporting local economic growth [2]. Salinity levels in the NRE fluctuate significantly due to natural processes, including variations in river discharge, tidal dynamics, and sea levels. These fluctuations influence key habitats and recreational areas, reinforcing the importance of accurate salinity forecasts for long-term planning and resource management.

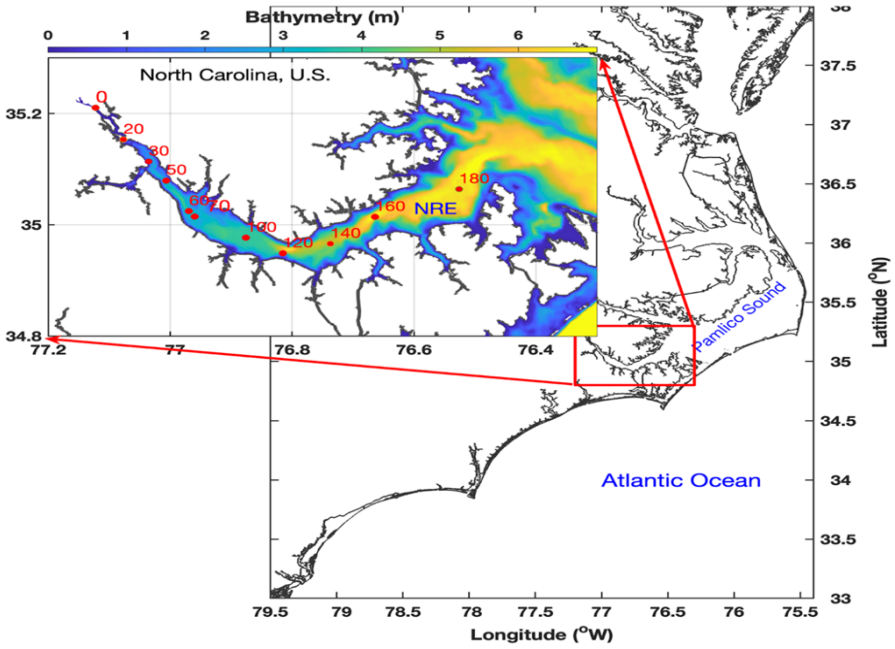
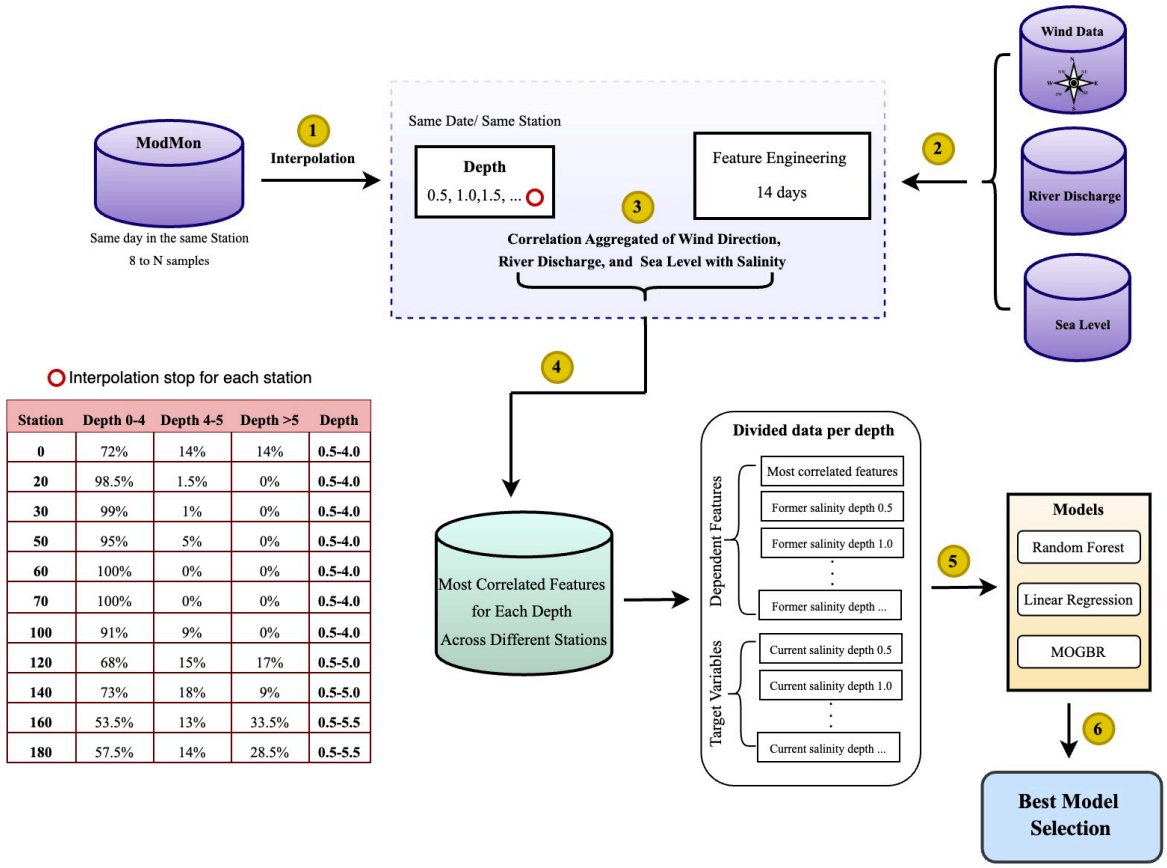


Figure 1. Location and bathymetry of the NRE, with ModMon sampling sites represented by red dots.

3.2 Data and Feature Engineering

The ModMon program has monitored salinity and other critical ecological and biogeochemical parameters at 11 mid-river sampling stations along the NRE, from its head to Pamlico Sound, as part of its water quality management efforts since 1994. The program collects bi-weekly hydrographic, chemical, and ecological data at multiple depths, depending on each station's depth range, throughout the year. Figure 2 provides a structured overview of our methodology for salinity prediction. The six-step process includes data preparation, feature selection, and model evaluation, applying machine learning techniques such as Random Forest, MOGBR, and Multiple Linear Regression to identify the best-performing model for further applications.



Station	Depth 0-4	Depth 4-5	Depth >5	Depth
0	72%	14%	14%	0.5-4.0
20	98.5%	1.5%	0%	0.5-4.0
30	99%	1%	0%	0.5-4.0
50	95%	5%	0%	0.5-4.0
60	100%	0%	0%	0.5-4.0
70	100%	0%	0%	0.5-4.0
100	91%	9%	0%	0.5-4.0
120	68%	15%	17%	0.5-5.0
140	73%	18%	9%	0.5-5.0
160	53.5%	13%	33.5%	0.5-5.5
180	57.5%	14%	28.5%	0.5-5.5

Figure 2. Workflow Overview: Initial Steps

We collected data from multiple sources to ensure a comprehensive dataset for salinity prediction. The primary sources included the ModMon dataset, river discharge, and meteorological data. From the ModMon dataset, we used a limited number of input data and focuned on three critical features: salinity, depth, and station, which provide essential context for understanding salinity levels. Data collection occurred at 11 mid-river stations (stations 0, 20, 30, 50, 60, 70, 100, 120, 140, 160, and 180; Figure 1) from 1994 to 2024.

It is worth mentioning that MODMON data has multiple samples on the same station and the same date, with some days featuring between 8 and N samples, depending on the station and sampling schedule. Daily river discharge for the Neuse River was acquired from the U.S. Geological Survey (USGS) at site 02091814 (https://waterdata.usgs.gov/nwis/;

accessed on 26 Jun 2024). Meteorological features, including wind speed, wind direction, air pressure, air temperature, and sea level, were obtained from NOAA's NDBC station CLKN7 near Cape Lookout Bight, NC (https://www.ndbc.noaa.gov; accessed on 26 Jun 2024).

In the first phase of our workflow, we separated the dataset by station and analyzed the proportion of observations within three depth ranges: shallow (0–4 meters), intermediate (4–5 meters), and deep (greater than 5 meters). Using this categorization, we applied linear interpolation starting at 0.5 meters to produce salinity values at regular 0.5-meter intervals (e.g., 0.5, 1.0, 1.5 meters). The interpolation was customized for each station to accommodate the variability in data density, ensuring that the depth profiles were appropriately represented.

In the next phase of our analysis, we included the manipulated wind, river discharge, and sea level as input features and used feature selection to identify the most important predictors. Aggregated (summed) wind speeds, calculated over 1- to 14-day periods before the prediction time, were divided into eight directional sectors (N-NE, NE-E, E-SE, SE-S, S-SW, SW-W, W-NW, NW-N) to reflect the cumulative and directional influence of wind on ecological and physical systems. Considering winds mainly affect salinity through mixing processes, and the influence by river discharge is through regulating the upstream salt intrusion, we consider up to 14 days to aggregate the features for this part.

In addition to three features from the ModMon datasets, we analyzed the correlation between the aggregated wind data across directional sectors and salinity levels at various depths for each station (Figure 3 uses station 100 as an example, and the rest of the feature

selection results are provided in Table 2). At Station 100, we identified 18 features, detailed in Table 2, that demonstrated the strongest correlations with salinity at each depth.

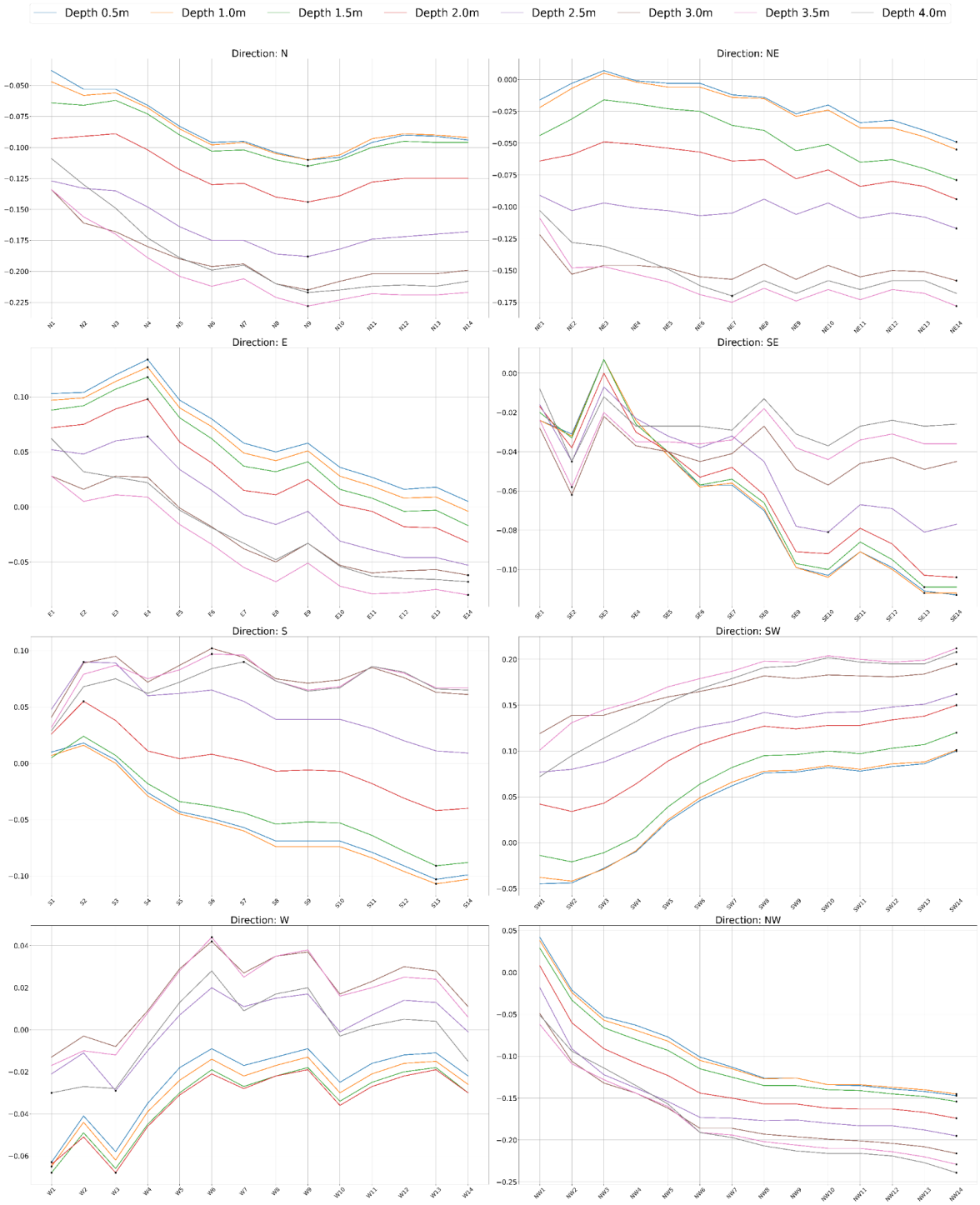


Figure 3. Correlation of aggregated wind over time with Salinity levels at Station 100

Table 2. Summary of Features Used for Training and Testing at different Stations

Station	Featured selected	Salinity samples	Number of observations (Training)	Number of observations (Testing)
0	N6, NW7, E7, S14, NW13, NE12, N14, NE2, SE14, E8, SW14, NE6, SW1, W4, S13, N7, W5, E6, NW6, Acc_flow14, Acc_flow1, MSL 14, MSL6, MSL1, MSL7, Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	455	195
20	NW12, N4, W13, S8, E7, E10, S14, SW3, SW2, N2, N1, NE3, W1, N14, S13, NW6, W9,SE9, E8, NW10, NE2, NW9, N3, SW1, E4, Acc_flow1, Acc_flow14, MSL4, MSL1, MSL5, Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	455	195
30	E12, W9, SE10, E4, NW9, NW10, E14, NE3, E8, SW3, S4, S13, SE7, SE9, NE11, NE9, SW2, SE14, SW1, NW1, W1, NE6, N13, SE6, N2, S12, Acc_flow14, Acc_flow1, MSL1, MSL5, Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	473	204
50	E4, NE5, SW14, NW10, N11, E1, SE14, E14, SE9, NW1, NW14, N9, NE6, S4, S12, NW8, W9, W1, NE11, SE13, SE7, N10, NE14, N2, SW2, S13, Acc_flow1, Acc_flow14, MSL5, MSL1, Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	439	189
60	NE9, SW1, N10, SW14, S13, E8, N9, SE13, NW14, S4, SE14, SE7, NE6, N14, E4, E14, NE14, W4, SW3, NW12, W1, Acc_flow14, MSL5, MSL1,Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	436	188
70	SW14, NE11, NW14, NE3, S4, N10, W4, SW1, N9, SE14, S14, NW12, S11, S13, W14, W1, SE9, NE9, E12, E8, SE6, E4, Acc_flow14, MSL4, MSL5, MSL1, Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	474	204
100	SW14, SE13, SE14, SE2, S2, S13, E14, N9, E4, S6, W6, NW14, SE10, W3, NE7, W1, NE14, S7, Acc_flow14, MSL12, MSL1, MSL5, Salinity former 0.5, …, salinity former 4.0	[0.5, 4.0]	415	178
120	S13, E4, S6, SE14, SE3, NW14, N10, NE7, SE9, S2, N9, NE14, S7, SE10, W14, SE2, W1, NE2, E12, SW14, Acc_flow14, MSL1, MSL11, MSL12, MSL5, Salinity former 0.5, …, salinity former 5.0	[0.5, 5.0]	471	203
140	S3, SE5, N9, N13, NE6, E1, E14, E4, W7, NW14, NE3, N10, W1, SE3, NE14, SE10, SW14, SE13, Acc_flow14, MSL5, MSL1, Salinity former 0.5, …, salinity former 5.0	[0.5, 5.0]	405	174
160	N8, NW14, N9, W7, E4, NE3, SE13, SE3, N10, S3, E1, SW14, W1, NE2, NE14, Accl_flow14, MSL1, MSL5, MSL4, Salinity former 0.5, …, salinity former 5.5	[0.5, 5.5]	406	174
180	S13, NE2, N9, NW14, S3, N8, E1, S2, SE10, W10, NE8, SW14, E3, SE3, Acc_flow14, MSL4, MSL12,Salinity former 0.5, …, salinity former 5.5	[0.5, 5.5]	364	157

We calculated cumulative discharge from the Neuse River over 1- to 14-day periods before the prediction time (named ACC_Flow below) at different depths across stations. By integrating this data with the three ModMon features, we observed strong correlations, with the 14-day aggregated discharge showing the most significant relationship with salinity levels at all depths (Figure 4, station 100 as an example). Strong correlations are typically observed near -1 or 1, reflecting a significant relationship between features. As illustrated in Figure 4, the deeper depths (e.g., 4.0 and 3.5 m) show lower values close to -0.5, which indicates a strong negative correlation compared to shallower depths.

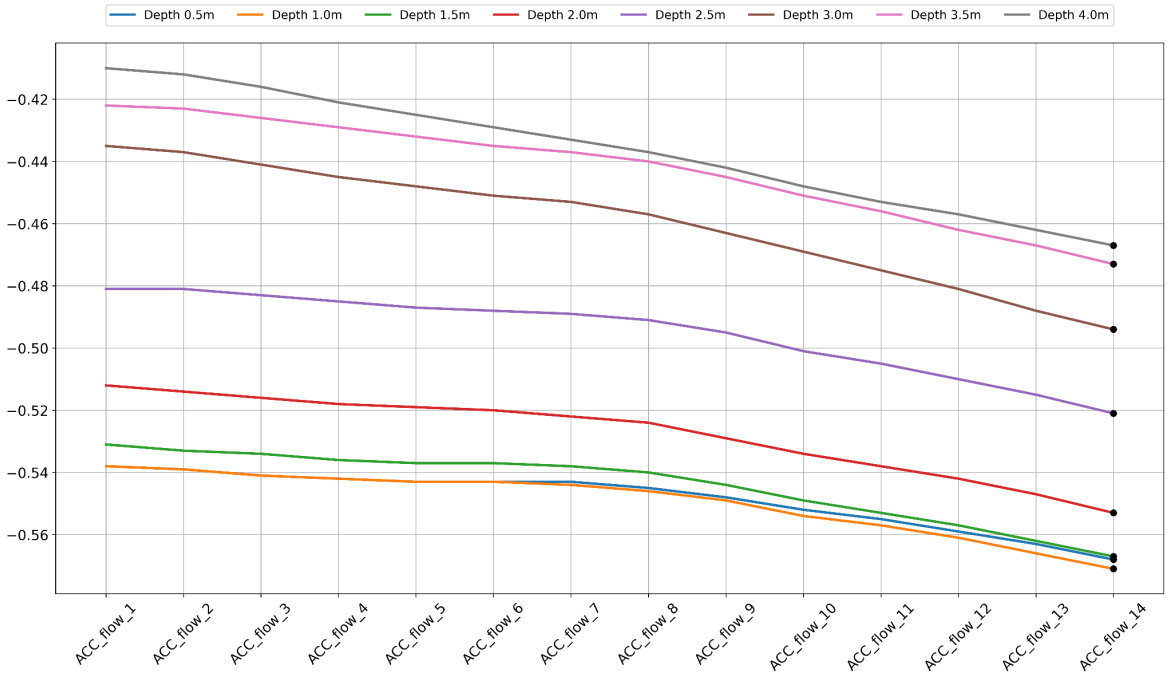


Figure 4. Correlation of aggregated river discharge with Salinity levels at Station 100

In the last part of the aggregation and correlation analysis, we examined mean sea level data aggregated over 1- to 14-day periods preceding the prediction time (named MSL below) across varying depths for each station. Combined with the three ModMon features, the analysis identified strong correlations, with sea level aggregations over 1-, 5-, and 12-day

intervals showing the most significant associations with salinity levels at all depths (Figure 5, using station 100 as an example).

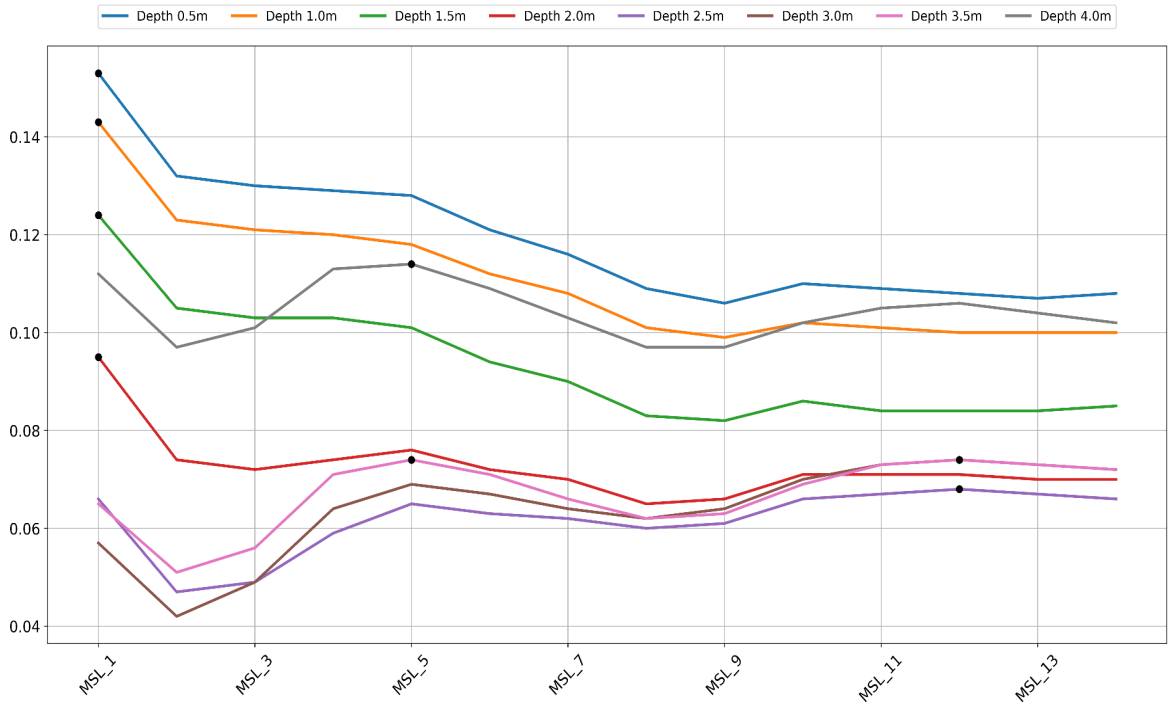


Figure 5. Correlation of aggregated sea level with Salinity levels at Station 100

The dataset was organized to predict salinity at specific depths by defining targets and input features. Targets were set as current salinity values at depths ranging from 0.5 to 4 meters, resulting in eight distinct targets for station 100. The input features consisted of former salinity values from previous time steps and the top correlated features. This setup enabled a comprehensive analysis of how past salinity and environmental conditions influence present salinity levels.

3.3 Machine Learning Models

We employed models to predict salinity levels, including random forest, Multi-Output and Gradient Boosting Regression (MOGBR), and multiple linear regression techniques. Each model was chosen for its unique ability to capture and learn from temporal and non-linear patterns within the data.

We used four standard evaluation metrics to assess the performance of our models: R-squared (R^2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Normalized Root Mean Squared Error (NRMSE). These metrics help evaluate the accuracy and variance explained by the models, with smaller MAE, RMSE, and NRMSE values and larger R^2 values indicating better performance. Below are the definitions and formulas for each metric:

R-squared (R^2):

R^2 represents the proportion of variance in the observed data that is explained by the model. A value closer to 1 indicates better predictive performance. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \underline{y})^2}$$

Where y_i is the observed value, $\hat{y}_i$ is the predicted value, $\underline{y}$ is the mean of observed values, and n is the number of data points.

Mean Absolute Error (MAE):

MAE measures the average magnitude of prediction errors without considering their direction. Smaller values show higher model accuracy. It is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE):

RMSE quantifies the standard deviation of prediction errors, penalizing larger errors more heavily, it is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}$$

Normalized Root Mean Squared Error (NRMSE):

NRMSE scaled RMSE to make it unitless and comparable across datasets. We calculate NRMSE using this formula:

$$NRMSE = \frac{RMSE}{\underline{y}}$$

3.3.1 Random Forest

First described by Breiman [15] in 2001, Random Forest is a machine learning algorithm that builds predictive models by combining multiple decision trees. This model, as a flexible and efficient algorithm, is designed to solve both classification and regression problems. Random Forest is recognized as one of the most powerful methods for data analysis and has been widely implemented in numerous research studies [26]. Notably, This it has been extensively applied in salinity prediction, particularly for addressing the complex challenges of modeling surface water quality parameters. For example, Khan et al. [27] effectively utilized this approach in their research on water quality. Xu et al. [28] implemented a random Forest-based framework that achieved an impressive 92.94% accuracy in predicting nearshore seawater salinity. Their work demonstrated the algorithm's ability to effectively

model complex, non-linear relationships, further validating its suitability for the present study.

This model was calculated as:

$$\hat{y}_i = \frac{1}{N_t} \sum_{t=1}^{N_t} b_t(x)_i$$

Where $\hat{y}_i$ is the predicted value for the i-th output variable, N_t is the total number of trees in model, and $b_t(x)_i$ is the prediction for the i-th output variable from the t-th tree, given input x.

3.3.2 Multi-Output with Gradient Boosting Regression

One of the powerful machine learning algorithms is the Gradient Boosting Regression (GBR) [29], which builds predictive models by sequentially combining weak learners, typically decision trees, to minimize the error in predictions. The algorithm works iteratively, with each new tree correcting the errors of the previous ones, making it highly effective for modeling complex, non-linear relationships in data. Multi-Output Regressor (MOR) [30] extends the capabilities of GBR by enabling it to handle multiple target features simultaneously.

In MOR, a separate GBR model is trained for each target variable, allowing the algorithm to model distinct dynamics for each output while leveraging the shared features across all targets. The combination of MOR and GBR is ideal for tackling problems with *interconnected targets*, like predicting salinity levels at various depths. By creating separate models for each target, the MOGBR approach can effectively capture the unique traits of each target while also accounting for their shared connections to common predictors, making it a highly adaptable solution for multi-target regression challenges.

3.3.3 Multiple Linear Regression

Multiple Linear Regression (MLR) techniques [31-32] is one of the simplest and most commonly used models for predictive tasks in machine learning. It establishes a linear relationship between input features and the target variable by fitting a line to the data [33]. This model has been used extensively in water quality studies, as shown by the work of Guillou [24], Charulatha [34] and Yildiz and Degirmenci [35], who demonstrated its utility in characterizing key relationships within datasets. Its simplicity and reliability ensure that it remains a fundamental choice for various predictive tasks.

Seeboonruang et al. [36] applied a time-lagged multiple linear regression model to forecast groundwater levels and salinity fluctuations in a saline-irrigated region in Northeastern Thailand. This study used lagged features, including rainfall, river stages, and regulating gate outflow, to capture the influence of prior surface water regulation and precipitation on groundwater salinity. By integrating sinusoidal components with regression techniques, the model demonstrated superior alignment with observed data compared to traditional linear regression approaches. This approach provides a practical and cost-effective alternative for predicting and managing groundwater salinity, particularly in scenarios where more complex numerical models are not viable.

3.4 Model Application Process

With our datasets prepared, we partitioned the data into 70% training and 30% testing sets for model development and validation. Table 2 presents the complete list of features used for this analysis for all stations. This split ensured both the thorough training of our models and a rigorous evaluation of their predictive performance. Based on the segmentation

of the stations, we obtained 11 distinct data subsets comprising both past and present salinity values (the past salinity values are used as input feature to the model to be trained). Each model was trained independently for each station and according to each station trained independently for each depth. Considering the difference in the water depth, geometry, and distance to the river mouth and the coastal ocean, the different stations will respond differently to the same environmental factors. The value of the salinity is different by station. For example, the MOGBR model for station 20 was trained independently compared to station 100, so those two models had different parameters despite both models using the same MOGBR structure. By training the models independently for each station, we can enhance the effectiveness and adaptability of the models.

3.5 Results from Feature Engineering and Model Selection

3.5.1 Model Comparison Across Depths and Stations

We compared the performance of three models in salinity prediction, as shown in Tables 3 to 13, which present the metrics MAE, RMSE, NRMSE, and R^2. We put the experiment results from station 100 first since we have used this station for the feature engineering discussion. As indicated in Table 3, MOGBR consistently demonstrated the best performance across most depths, achieving the lowest MAE and NRMSE, along with the highest R^2 values, which indicate strong predictive accuracy and model fit. For instance, at a depth of 0.5 meters, MOGBR achieved an MAE of 1.49, an NRMSE of 10.2%, and an R^2 of 0.85, outperforming both Random Forest (MAE: 1.64, NRMSE: 11.5%, R^2: 0.81) and MLR (MAE: 1.94, NRMSE: 13.0%, R^2: 0.76).

While Random Forest performed slightly better than MLR at shallower depths, its performance declined for deeper depths, where MOGBR maintained higher accuracy and stability. These results highlight MOGBR's effectiveness in capturing complex relationships between predictors and salinity levels, making it the most suitable model for this study.

Table 3. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 100

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	1.64	1.55	1.47	1.52	1.69	2.17	2.17	2.31
	RMSE	2.24	2.10	1.99	2.07	2.27	2.80	2.80	2.95
	NRMSE	11.5%	10.6%	10.06%	10.3%	11.1%	13.2%	13.1%	13.0%
	R^2	0.81	0.83	0.85	0.85	0.82	0.76	0.75	0.74
MLR	**MAE**	1.94	1.88	1.82	1.82	1.93	2.25	2.33	2.24
	RMSE	2.54	2.47	2.36	2.38	2.53	2.89	2.97	2.88
	NRMSE	13.0%	12.5%	11.9%	11.9%	12.3%	13.7%	13.1%	12.7%
	R^2	0.76	0.77	0.79	0.80	0.78	0.75	0.74	0.75
MOGBR	**MAE**	1.49	1.43	1.44	1.43	1.72	2.14	2.38	2.26
	RMSE	1.98	1.91	1.95	2.03	2.26	2.80	2.81	2.83
	NRMSE	10.2%	9.7%	9.8	10.1%	11.0%	13.2%	13.2%	12.4%
	R^2	**0.85**	**0.86**	**0.86**	0.85	0.82	**0.76**	**0.75**	**0.76**

Table 4. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 0

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	0.02	0.02	0.04	0.08	0.19	0.29	0.35	0.39
	RMSE	0.04	0.09	0.25	0.36	0.73	1.12	1.29	1.39
	NRMSE	7.5%	5.7%	5.4%	5.3%	9.9%	11.3%	10.9%	11.0%
	R^2	0.47	0.46	**0.44**	0.54	**0.46**	**0.34**	**0.33**	**0.33**
MLR	**MAE**	0.03	0.04	0.10	0.19	0.37	0.50	0.61	0.69
	RMSE	0.06	0.12	0.34	0.57	0.95	1.31	1.54	1.72
	NRMSE	10.2%	7.7%	7.6%	8.3%	13.0%	13.1%	13.0%	13.6%
	R^2	0.02	0.03	0.00	0.00	0.08	0.11	0.05	0.00
MOGBR	**MAE**	0.02	0.02	.05	0.08	0.19	0.28	0.35	0.41
	RMSE	0.04	0.08	0.27	0.34	0.74	1.14	1.33	1.44
	NRMSE	7.3%	5.5%	6.0%	4.9%	10.1%	11.5%	11.2%	11.4%
	R^2	**0.50**	**0.51**	0.33	**0.61**	0.44	0.32	0.29	0.29

Table 5. Performance Comparison of Three Models for Salinity Prediction at Different Depths - Station20

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	0.38	0.60	0.90	1.13	1.31	1.51	1.71	1.78
	RMSE	0.80	1.35	2.12	2.44	2.66	2.84	3.00	3.13
	NRMSE	10.3%	11.1%	13.3%	13.8%	14.1%	15.1%	16.0%	16.6%
	R^2	0.56	0.54	0.46	0.48	0.50	0.55	0.58	0.56
MLR	**MAE**	0.55	0.84	1.22	1.51	1.79	2.09	2.35	2.41
	RMSE	0.91	1.53	2.30	2.67	2.92	3.16	3.32	3.37
	NRMSE	11.6%	12.6%	14.5%	15.2%	15.5%	16.8%	17.6%	17.9%
	R^2	0.44	0.41	0.36	0.37	0.39	0.45	0.49	0.49
MOGBR	**MAE**	0.43	0.60	0.95	1.11	1.33	1.48	1.80	1.82
	RMSE	0.80	1.28	2.16	2.37	2.64	2.64	2.95	3.02
	NRMSE	10.3%	10.6%	13.6%	13.5%	14.0%	14.0%	15.7%	16.0%
	R^2	**0.56**	**0.58**	0.43	**0.51**	0.50	**0.61**	**0.60**	**0.59**

Table 6. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 30

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	0.99	1.20	1.98	2.76	3.15	3.22	3.17	3.16
	RMSE	1.48	1.84	3.18	3.95	4.20	4.23	4.20	4.19
	NRMSE	13.1%	12.9%	15.7%	18.2%	18.8%	18.7%	18.6%	18.6%
	R^2	0.67	0.65	0.55	0.54	0.51	0.50	0.50	0.50
MLR	**MAE**	1.30	1.51	2.27	2.74	2,85	2.88	2.86	2.85
	RMSE	1.83	2.14	3.27	3.72	3.80	3.81	3.80	3.80
	NRMSE	16.2%	15.0%	16.1%	17.1%	17.0%	16.9%	16.8%	16.8%
	R^2	0.50	0.53	0.53	0.59	0.60	0.59	0.59	0.590.90
MOGBR	**MAE**	0.90	1.12	1.91	2.63	2.90	3.00	2.94	2.95
	RMSE	1.35	1.76	3.07	3.75	3.90	3.94	3.93	3.93
	NRMSE	12.0%	12.3%	13.1%	17.3%	17.5%	17.4%	17.4%	17.4%
	R^2	**0.72**	**0.68**	**0.58**	**0.58**	**0.57**	**0.57**	**0.57**	**0.57**

Table 7. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 50

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	1.18	1.38	2.14	2.55	2.78	2.92	2.96	2.94
	RMSE	1.66	1.97	2.99	3.45	3.71	3.79	3.83	3.80
	NRMSE	13.5%	12.7%	16.1%	18.1%	18.0%	18.0%	18.0%	17.9%
	R^2	0.74	0.72	0.58	0.58	0.57	0.56	0.54	0.53
MLR	**MAE**	1.60	1.76	2.19	2.60	2.77	2.87	2.89	2.87
	RMSE	2.08	2.31	2.97	3.36	3.65	3.74	3.86	3.83
	NRMSE	16.9%	14.9%	16.1%	17.6%	17.7%	17.7%	18.1%	18.0%
	R^2	0.60	0.62	0.59	0.61	0.59	0.57	0.53	0.53
MOGBR	**MAE**	1.09	1.33	2.12	2.46	2.70	2.75	2.84	2.75
	RMSE	1.57	1.87	2.92	3.31	3.57	3.61	3.74	3.62
	NRMSE	12.8%	12.1%	15.8%	17.4%	17.3%	17.1%	17.6%	17.0%
	R^2	**0.77**	**0.75**	**0.60**	**0.62**	**0.60**	**0.60**	**0.56**	**0.58**

Table 8. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 60

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	1.63	1.58	1.61	1.98	2.33	2.43	2.62	2.62
	RMSE	2.25	2.20	2.28	2.80	3.11	3.13	3.28	3.28
	NRMSE	14.6%	13.8%	12.8%	11.6%	12.6%	12.6%	13.3%	13.2%
	R^2	0.74	0.77	0.77	0.71	0.69	0.71	0.68	0.67
MLR	**MAE**	1.89	1.76	1.76	2.07	2.37	2.41	2.56	2.56
	RMSE	2.48	2.36	2.38	2.88	3.14	3.09	3.24	3.24
	NRMSE	16.1%	14.8%	13.3%	11.9%	12.7%	12.5%	13.1%	13.1%
	R^2	0.69	0.74	0.75	0.69	0.69	0.71	0.69	0.69
MOGBR	**MAE**	1.64	1.49	1.60	1.96	2.26	2.42	2.58	2.59
	RMSE	2.23	2.04	2.27	2.78	3.02	3.13	3.26	3.28
	NRMSE	14.5%	12.8%	12.8%	11.5%	12.2%	12.6%	13.2%	13.2%
	R^2	**0.75**	**0.80**	**0.78**	**0.72**	**0.71**	**0.71**	**0.69**	**0.68**

Table 9. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 70

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0
Random Forest	**MAE**	1.63	1.53	1.62	1.79	2.23	2.41	2.52	2.52
	RMSE	2.27	2.10	2.18	2.46	2.99	3.15	3.20	3.20
	NRMSE	12.1%	11.2%	11.6%	13.1%	14.0%	13.4%	13.6%	13.6%
	R^2	0.77	0.81	0.80	0.77	0.72	0.69	0.69	0.69
MLR	**MAE**	1.93	1.86	1.90	1.94	2.26	2.39	2.42	2.43
	RMSE	2.65	2.51	2.51	2.65	3.00	3.01	3.06	3.06
	NRMSE	14.2%	13.4%	13.3%	14.1%	14.0%	12.8%	13.0%	13.0%
	R^2	0.69	0.73	0.74	0.73	0.71	0.72	0.71	0.71
MOGBR	**MAE**	1.64	1.53	1.55	1.76	2.15	2.27	2.40	2.40
	RMSE	2.33	2.14	2.11	2.38	2.83	2.88	3.00	3.00
	NRMSE	12.4%	11.2	11.2%	12.6%	13.2%	12.3%	12.8%	12.8%
	R^2	**0.77**	**0.81**	**0.81**	**0.78**	**0.74**	**0.74**	**0.72**	**0.72**

Table 10. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 120

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0	Salinity depth 4.5	Salinity depth 5.0
Random Forest	**MAE**	1.73	1.74	1.69	1.71	1.68	1.91	2.14	2.21	2.20	2.25
	RMSE	2.27	2.26	2.19	2.19	2.18	2.55	2.73	2.78	2.81	2.88
	NRMSE	10.2%	10.1%	9.5%	9.3%	9.2%	10.7%	11.3	11.4%	11.2%	11.5%
	R^2	0.81	0.81	0.82	0.82	0.83	0.79	0.77	0.76	0.75	0.74
MLR	**MAE**	1.74	1.73	1.69	1.67	1.72	1.87	1.93	1.89	1.90	1.92
	RMSE	2.30	2.28	2.19	2.18	2.22	2.47	2.48	2.42	2.44	2.50
	NRMSE	10.4%	10.2%	9.5%	9.2%	9.3%	10.3%	10.3%	9.9%	9.7%	9.9%
	R^2	0.80	0.80	0.80	0.83	0.83	0.80	0.81	0.82	0.81	0.80
MOGBR	**MAE**	1.54	1.60	1.60	1.59	1.71	1.84	1.91	1.94	1.87	1.90
	RMSE	2.05	2.07	2.05	2.06	2.12	2.38	2.42	2.44	2.36	2.42
	NRMSE	9.3%	9.3%	8.9%	8.7%	8.9	10.0%	10.0%	10.0%	9.4%	9.6%
	R^2	**0.84**	**0.84**	**0.84**	**0.84**	**0.84**	**0.82**	**0.82**	**0.82**	**0.82**	**0.82**

Table 11. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 140

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0	Salinity depth 4.5	Salinity depth 5.0
Random Forest	**MAE**	1.44	1.40	1.39	1.48	1.62	1.79	1.80	1.96	2.11	1.44
	RMSE	1.85	1.80	1.81	1.93	2.17	2.33	2.36	2.47	2.66	1.85
	NRMSE	8.6%	8.4%	8.4%	8.7%	9.6%	10.2%	9.6%	9.2%	9.5%	8.6%
	R^2	0.86	0.87	0.87	0.85	0.82	0.80	0.79	0.77	0.74	0.86
MLR	**MAE**	1.60	1.52	1.47	1.53	1.70	1.78	1.79	1.93	2.12	1.60
	RMSE	2.04	1.98	1.94	2.01	2.30	2.38	2.34	2.43	2.68	2.04
	NRMSE	9.5%	9.2%	9.0%	9.0%	10.2%	10.5%	9.5%	9.1%	9.6%	9.5%
	R^2	0.83	0.84	0.85	9.0	0.80	0.79	0.80	0.78	0.74	0.83
MOGBR	**MAE**	1.42	1.37	1.43	1.46	1.62	1.71	1.69	1.78	1.86	1.42
	RMSE	1.77	1.72	1.81	1.87	2.16	2.23	2.23	2.29	2.44	1.77
	NRMSE	8.3%	8.0%	8.4%	8.4%	9.6%	9.85%	9.1%	8.5%	8.7%	8.3%
	R^2	**0.87**	**0.88**	**0.87**	**0.86**	**0.82**	**0.81**	**0.82**	**0.81**	**0.78**	**0.87**

Table 12. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 160

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0	Salinity depth 4.5	Salinity depth 5.0	Salinity depth 5.5
Random Forest	**MAE**	1.50	1.44	0.86	1.23	1.20	1.26	1.31	1.36	1.42	1.56	1.81
	RMSE	1.99	1.93	0.86	1.62	1.55	1.68	1.76	1.85	1.95	2.08	2.46
	NRMSE	8.9%	8.6%	8.0%	7.2%	6.9%	6.8%	7.2%	7.6%	7.9%	8.3%	9.8%
	R^2	0.85	0.86	0.88	0.90	0.90	0.89	0.87	0.86	0.84	0.82	0.76
MLR	**MAE**	1.50	1.43	1.34	1.25	1.18	1.22	1.23	1.32	1.38	1.54	1.72
	RMSE	2.11	2.01	1.88	1.70	1.63	1.73	1.77	1.89	1.90	2.02	2.39
	NRMSE	9.4%	9.0%	8.4%	7.6%	7.3%	7.3%	7.3%	7.7%	7.7%	8.0%	9.5%
	R^2	0.84	0.85	0.87	0.89	0.89	0.88	0.87	0.85	0.85	0.83	0.78
MOGBR	**MAE**	1.49	1.40	1.32	1.15	1.10	1.23	1.24	1.28	1.36	1.44	1.59
	RMSE	1.94	1.87	1.76	1.52	1.45	1.66	1.69	1.73	1.75	1.82	2.13
	NRMSE	8.7%	8.3%	7.9%	6.8%	6.4%	7.0%	6.9%	7.1%	7.1%	7.2%	8.5%
	R^2	**0.86**	**0.87**	**0.88**	**0.91**	**0.91**	**0.89**	**0.88**	**0.87**	**0.87**	0.86	0.82

Table 13. Performance Comparison of Three Models for Salinity Prediction at Different Depths – Station 180

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0	Salinity depth 4.5	Salinity depth 5.0	Salinity depth 5.5
Random Forest	**MAE**	1.42	1.40	1.36	1.33	1.26	1.20	1.14	1.11	1.15	1.34	1.57
	RMSE	2.00	1.95	1.87	1.80	1.75	1.69	1.66	1.58	1.64	1.88	2.12
	NRMSE	8.2%	8.2%	8.3%	8.3%	8.0%	7.8%	7.8%	7.6%	7.9%	8.5%	9.3%
	R^2	0.84	0.84	0.85	0.86	0.87	0.87	0.87	0.88	0.87	0.82	0.78
MLR	**MAE**	1.38	1.36	1.33	1.30	1.26	1.24	1.20	1.19	1.23	1.37	1.50
	RMSE	1.98	1.94	1.88	1.83	1.82	1.78	1.75	1.70	1.74	2.04	2.11
	NRMSE	8.1%	8.2%	8.4%	8.4%	8.4%	8.2%	8.2%	8.2%	8.3%	9.2%	9.2%
	R^2	0.84	0.84	0.85	0.86	0.86	0.86	0.86	0.86	0.85	0.79	0.78
MOGBR	**MAE**	1.31	1.29	1.26	1.25	1.18	1.16	1.16	1.09	1.13	1.30	1.44
	RMSE	1.81	1.77	1.67	1.63	1.60	1.59	1.61	1.56	1.64	1.85	1.95
	NRMSE	7.4%	7.5%	7.4%	7.5%	7.4%	7.4%	7.5%	7.5%	7.9%	8.4%	8.5%
	R^2	**0.86**	**0.87**	**0.88**	**0.88**	**0.89**	**0.89**	**0.88**	**0.89**	**0.87**	**0.83**	**0.81**

CHAPTER 4: SEASONAL SALINITY PREDICTION

4.1 Data and Best Model

Our methodology for seasonal salinity prediction, as illustrated in Figure 6, begins with selecting the MOGBR as the optimal model. This model was chosen for its capability of handling multi-output predictions effectively and its strong performance during the initial stages of our project to predict salinity levels across different depths. We utilized the most significant features from meteorological data, sea level, and river discharge (based on feature engineering results from Table 2) influencing salinity, ensuring the model captured the critical features affecting seasonal dynamics. As the first step of the prediction process, the initial recorded salinity values for each depth and season (Current Salinity 0.5, Current Salinity 1.0, Current Salinity 1.5, etc.) were used as ground truth inputs (purple triangles in Figure 6), providing a reliable foundation for subsequent predictions.

In the first step of the prediction process, initial recorded salinity values for each depth and season (e.g., Current Salinity 0.5, Current Salinity 1.0, Current Salinity 1.5, etc.) were used as ground truth inputs (represented by purple triangles in Figure 6), forming the foundation for subsequent predictions. These initial values were then treated as "Former Salinity" inputs for the next day following the sampling gap of MODMON. In Step 2, the MOGBR model was applied to predict the "Current Salinity" values (illustrated as orange circles in Figure 6) for each depth. Once these predictions were generated, they were used as inputs (Former Salinity) for the following day in Step 3, where the MOGBR model was again applied to forecast salinity values for each depth (represented by orange squares in Figure 6).

This iterative prediction process enabled the model to forecast salinity levels for all subsequent days within a season, using its previous predictions as inputs for the next day following the sampling gap of MODMON. By maintaining temporal consistency, the model successfully captured the dynamic evolution of salinity throughout each season. This iterative framework not only incorporated temporal variability but also accounted for depth-specific changes, allowing the model to represent the complex and evolving patterns of salinity over time. By emphasizing seasonal salinity predictions, this methodology offers valuable insights into both temporal and depth-related salinity levels, making it an essential tool for understanding and forecasting salinity trends. These insights are crucial for effective water quality management in estuarine systems.

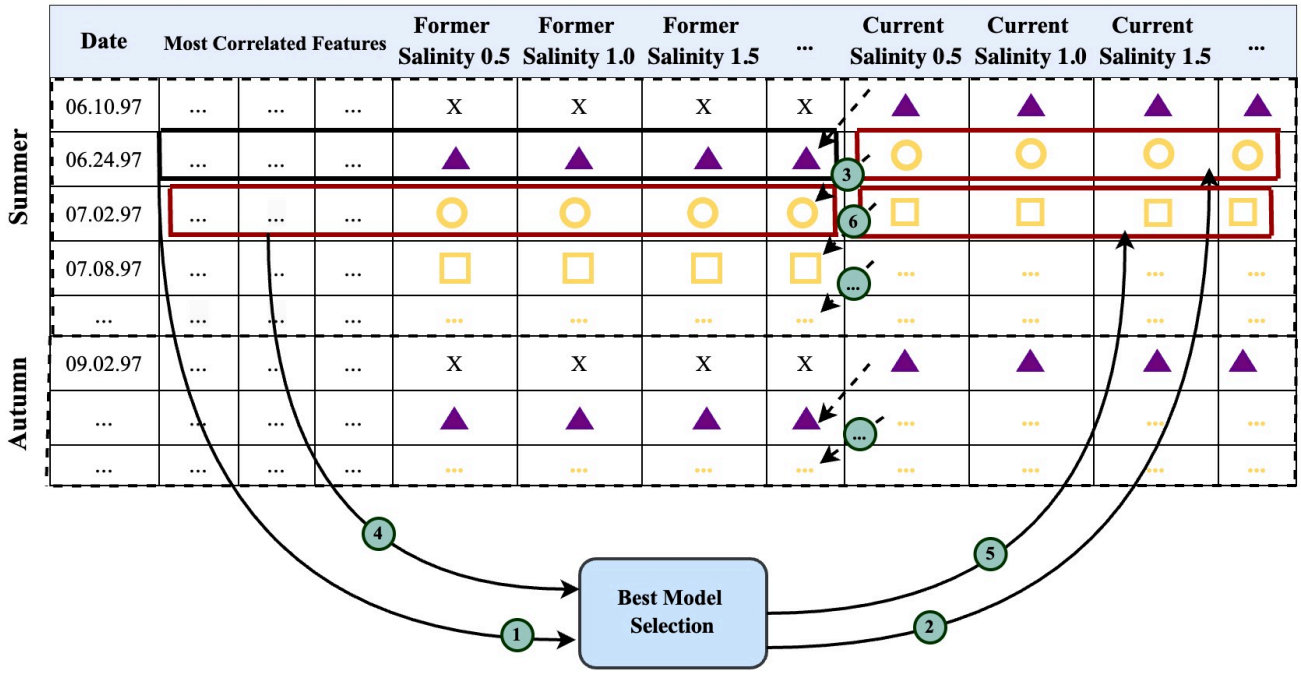


Figure 6. Workflow Overview: Final Steps

4.2 Result from Best Model

Table 14 provides a detailed evaluation of the MOGBR model's performance in predicting salinity levels across various seasons and depths. The model exhibits strong predictive capabilities, particularly at shallower depths, with R^2 values reaching up to 0.93 at 0.5 meters in Station 160. This indicates that the model can effectively capture surface salinity variations. However, even at deeper depths, such as 4.0 meters or beyond, the model maintains robust performance with R^2 values above 0.8 across most stations, reflecting its ability to adapt to the increasing complexity of salinity levels in deeper layers.

The MAE and RMSE metrics further highlight the model's accuracy and sensitivity to depth-specific variability. For instance, at Station 0, the RMSE increases from 0.03 at 0.5 meters to 0.77 at 4.0 meters, a natural reflection of the higher variability and uncertainty inherent in deeper salinity measurements. Similarly, at Station 100, MAE values are lowest at shallow depths (1.26 at 0.5 meters) and gradually increase at greater depths, reaching 1.79 at 4.0 meters. The NRMSE values, remaining under 10% across most cases, confirm that the model maintains high efficiency and accuracy across a wide range of depths.

Figure 7 complements these findings by providing a visual comparison of actual salinity values, predicted salinity values, and ground truth inputs for Station 100. The figure reveals that the MOGBR model excels at capturing seasonal trends in salinity, with predicted values closely following actual measurements at shallow depths such as 0.5 and 1.0 meters. At deeper levels, such as 3.5 and 4.0 meters, the alignment between predicted and actual values weakens slightly, reflecting the greater variability and complexity of salinity levels in those layers. The ground truth inputs, denoted by purple triangles in the figure, serve as

reinitialization points during the prediction process, ensuring that the model remains temporally consistent and accurately captures seasonal levels.

Table 14. Performance of MOGBR for seasonal Salinity Prediction at Different Depths in each Station

Station		Salinity depth 0.5	Salinity depth 1.0	Salinity depth 1.5	Salinity depth 2.0	Salinity depth 2.5	Salinity depth 3.0	Salinity depth 3.5	Salinity depth 4.0	Salinity depth 4.5	Salinity depth 5.0	Salinity depth 5.5
0	**MAE**	0.01	0.01	0.03	0.06	0.11	0.15	0.19	0.21	-	-	-
	RMSE	0.03	0.06	0.15	0.27	0.41	0.54	0.67	0.77	-	-	-
	NRMSE	2.9%	4.1%	3.4%	3.9%	5.3%	6.2%	7.4%	7.5%	-	-	-
	R^2	0.74	0.63	0.52	0.60	0.66	0.62	0.58	0.56	-	-	-
20	**MAE**	0.26	0.38	0.56	0.75	0.91	1.08	1.48	1.52	-	-	-
	RMSE	0.56	0.89	1.38	1.61	1.82	1.96	2.33	2.39	-	-	-
	NRMSE	7.30%	7.3%	8.6%	9.1%	9.7%	10.3%	11.4%	11.7%	-	-	-
	R^2	0.74	0.73	0.66	0.69	0.70	0.75	0.71	0.71	-	-	-
30	**MAE**	0.65	0.83	1.47	2.10	2.38	2.40	2.39	2.42	-	-	-
	RMSE	1.09	1.41	2.46	3.03	3.18	3.20	3.19	3.22	-	-	-
	NRMSE	8.1%	7.7%	11.3%	13.7%	14.2%	14.2%	14.1%	14.2%	-	-	-
	R^2	0.80	0.76	0.68	0.69	0.69	0.70	0.70	0.70	-	-	-
50	**MAE**	0.88	1.11	1.60	1.97	2.17	2.25	2.31	2.28	-	-	-
	RMSE	1.32	1.65	2.28	2.70	2.95	3.02	3.09	3.02	-	-	-
	NRMSE	8.5%	8.67%	10.5%	11.8%	12.6%	12.9%	13.0%	12.7%	-	-	-
	R^2	0.84	0.81	0.77	0.75	0.74	0.74	0.72	0.73	-	-	-
60	**MAE**	1.21	1.19	1.28	1.55	1.82	1.91	1.90	1.90	-	-	-
	RMSE	1.74	1.70	1.82	2.17	2.44	2.51	2.52	2.53	-	-	-
	NRMSE	9.0%	8.79%	9.3%	9.0%	9.9%	10.2%	10.2%	10.2%	-	-	-
	R^2	0.84	0.85	0.85	0.83	0.81	0.81	0.80	0.80	-	-	-
70	**MAE**	1.25	1.22	1.29	1.53	1.70	1.82	1.86	1.86	-	-	-
	RMSE	1.79	1.70	1.77	2.07	2.25	2.36	2.42	2.42	-	-	-
	NRMSE	9.1%	8.6%	8.8%	9.3%	9.2%	9.5%	9.7%	9.7%	-	-	-
	R^2	0.85	0.87	0.87	0.84	0.84	0.83	0.82	0.82	-	-	-
100	**MAE**	1.26	1.24	1.23	1.29	1.48	1.75	1.85	1.79	-	-	-
	RMSE	1.75	1.71	1.69	1.85	2.01	2.34	2.46	2.38	-	-	-
	NRMSE	8.5%	8.3%	8.2%	9.0%	8.8%	9.5%	9.8%	9.4%	-	-	-
	R^2	0.88	0.88	0.89	0.87	0.86	0.83	0.81	0.82	-	-	-
120	**MAE**	1.41	1.39	1.39	1.47	1.53	1.64	1.66	1.66	1.68	1.71	-
	RMSE	1.88	1.84	1.83	1.91	1.97	2.11	2.12	2.14	2.15	2.18	-
	NRMSE	8.5%	8.3%	7.9%	8.1%	8.3%	8.8%	8.5%	8.6%	8.5%	8.4%	-
	R^2	0.86	0.87	0.87	0.86	0.86	0.85	0.85	0.85	0.85	0.84	-
140	**MAE**	1.31	1.28	1.26	1.27	1.24	1.37	1.45	1.47	1.51	1.59	-
	RMSE	1.71	1.67	1.64	1.67	1.62	1.84	1.88	1.91	1.96	2.07	-
	NRMSE	7.4%	7.2%	7.1%	7.2%	7.0%	7.5%	7.5%	7.4%	7.3%	7.2%	-
	R^2	0.89	0.90	0.90	0.90	0.90	0.88	0.87	0.87	0.86	0.85	-
160	**MAE**	1.16	1.13	1.11	1.01	1.00	1.03	1.03	1.04	1.10	1.12	1.24
	RMSE	1.60	1.54	1.50	1.36	1.34	1.40	1.41	1.43	1.47	1.50	1.68
	NRMSE	6.6%	6.3%	6.0%	5.5%	5.4%	5.7%	5.7%	5.7%	5.7%	6.0%	6.7%
	R^2	0.90	0.91	0.91	0.93	0.93	0.92	0.92	0.91	0.91	0.90	0.88
180	**MAE**	1.06	1.04	1.03	1.02	0.98	0.98	0.98	1.00	1.04	1.12	1.23
	RMSE	1.42	1.40	1.37	1.35	1.30	1.30	1.30	1.35	1.41	1.50	1.65
	NRMSE	4.3%	4.3%	4.3%	4.4%	4.2%	4.2%	4.2%	4.4%	4.6%	4.9%	5.4%
	R^2	0.92	0.92	0.92	0.92	0.93	0.93	0.92	0.92	0.91	0.90	0.88

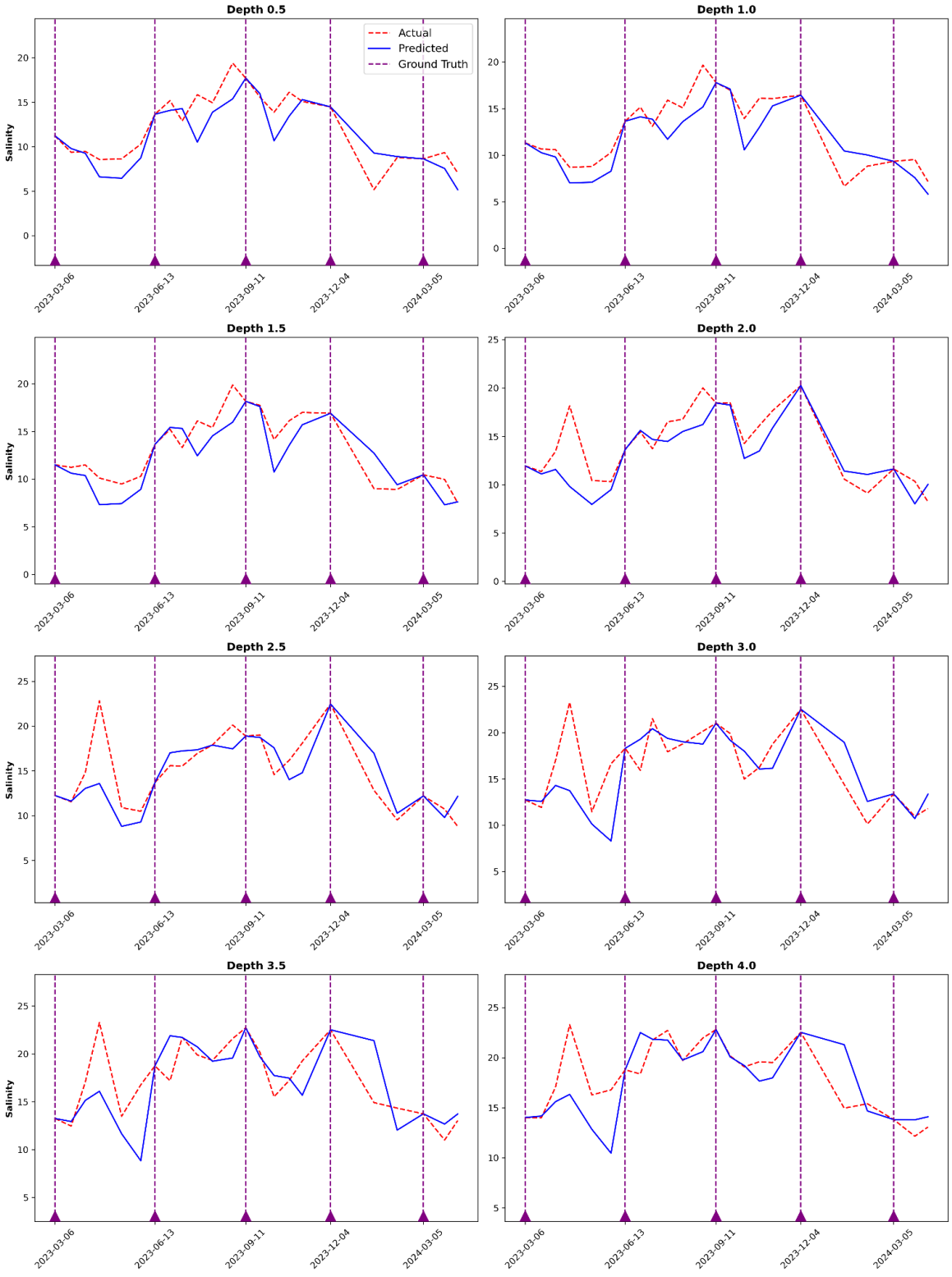


Figure 7. Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 100

CHAPTER 5: DISCUSSION AND CONCLUSION

This project aimed to predict salinity in the NRE at 11 mid-river sampling stations using machine learning models, with a particular focus on depth-specific predictions. Initially, we refined the dataset to prioritize critical features, including salinity, depth, and station from the ModMon dataset, as well as river discharge, wind direction (eight directions), and sea level data. This project was carried out in two main steps.

Step 1 involved refining the dataset to focus on critical features, including salinity, depth, and station from the ModMon dataset, along with river discharge, wind direction, and sea level data. Depth was a key consideration, as salinity can vary significantly at different depths in estuarine systems. After cleaning the data and performing feature engineering, we evaluated three machine learning models, including random forest, MOGBR, and MLR. While both random forest and MOGBR performed well, MOGBR demonstrated superior accuracy across depths, making it the best choice for salinity prediction in this study. The performance of different models is summarized in Figure 2.

Step 2, as shown in Figure 6, focused on seasonal predictions, where the data was organized into four seasons per year to account for temporal variability. For each depth, the first salinity value of each season was retained as a ground truth input, and MOGBR was applied to predict salinity for subsequent days across different depths iteratively. The iterative framework allowed the model to incorporate temporal consistency by using the predicted salinity values of one day as inputs for the next day following the sampling gap of MODMON. By integrating depth-specific and seasonal variability, this approach captured the dynamic progression of salinity throughout the season. The two-step methodology

provides a reliable and cost-effective framework for salinity management and offers actionable insights for conserving estuarine systems.

The comparison between the two phases of the MOGBR model's performance, as highlighted in Tables 2 and the Table 14, demonstrates the evolution of prediction accuracy across different approaches. In the initial step, focused on bi-weekly predictions, the MOGBR model performed well at Station 100, achieving R^2 values ranging from 0.85 at shallower depths (0.5 meters) to 0.76 at deeper depths (4.0 meters). However, the MAE and RMSE metrics reveal an increase in prediction error with depth—MAE rose from 1.49 at 0.5 meters to 2.26 at 4.0 meters, and RMSE increased from 1.98 to 2.83. These results suggest that the model captures salinity levels effectively at shallower depths but encounters greater challenges in accurately predicting salinity at deeper layers. However, our experimental results are comparable to the reported results from hydrodynamic models done by Chen at el. [37] at the Danshui River estuarine system, which is similar to NRE.

In contrast, Table 14, which focuses on seasonal predictions, shows an overall improvement in prediction accuracy across all depths, reflecting the model's ability to address temporal variability better. At Station 100, R^2 values remain consistently above 0.8 across depths, indicating consistently strong performance for the seasonal approach compared to bi-weekly predictions. The integration of initial ground truth inputs for each season and the iterative prediction framework contributed to this improvement. Additionally, NRMSE values in the seasonal phase are more consistent, with most depths showing values under 10%, highlighting the model's adaptability to seasonal trends. While errors such as MAE and RMSE still increase with depth in both approaches, the seasonal predictions show

reduced variability and greater consistency, emphasizing the advantages of incorporating seasonal patterns into the prediction process.

By comparing these two phases, it becomes evident that the seasonal framework not only improves prediction accuracy but also enhances the model's ability to generalize across depths and temporal scales. This makes it a more reliable tool for predicting salinity levels and supporting effective water quality management in the NRE.

For future studies, we will extend the aggregation period for sensitivity experiments to examine whether the change of aggregation period will significantly improve the model’s performance.

APPENDIX

In this section, we present the feature engineering process, including correlation analyses related to wind direction, river discharge, and sea level. Additionally, we compare the actual and predicted salinity levels across various stations, including Stations 0, 20, 30, 50, 60, 70, 120, 140, 160, and 180, as detailed below.

Station 0

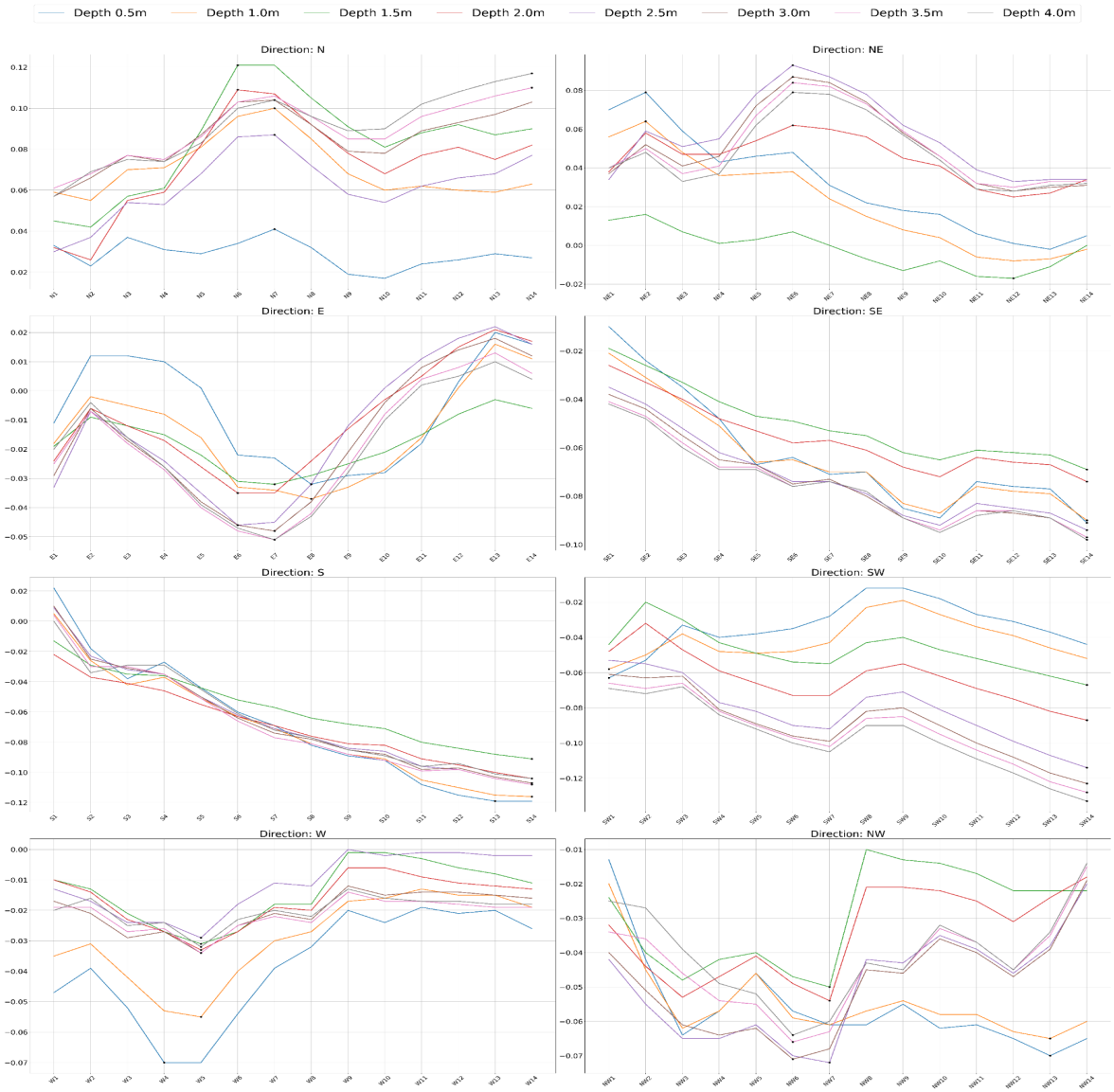


Figure 8. Correlation of aggregated wind over time with Salinity levels at Station 0

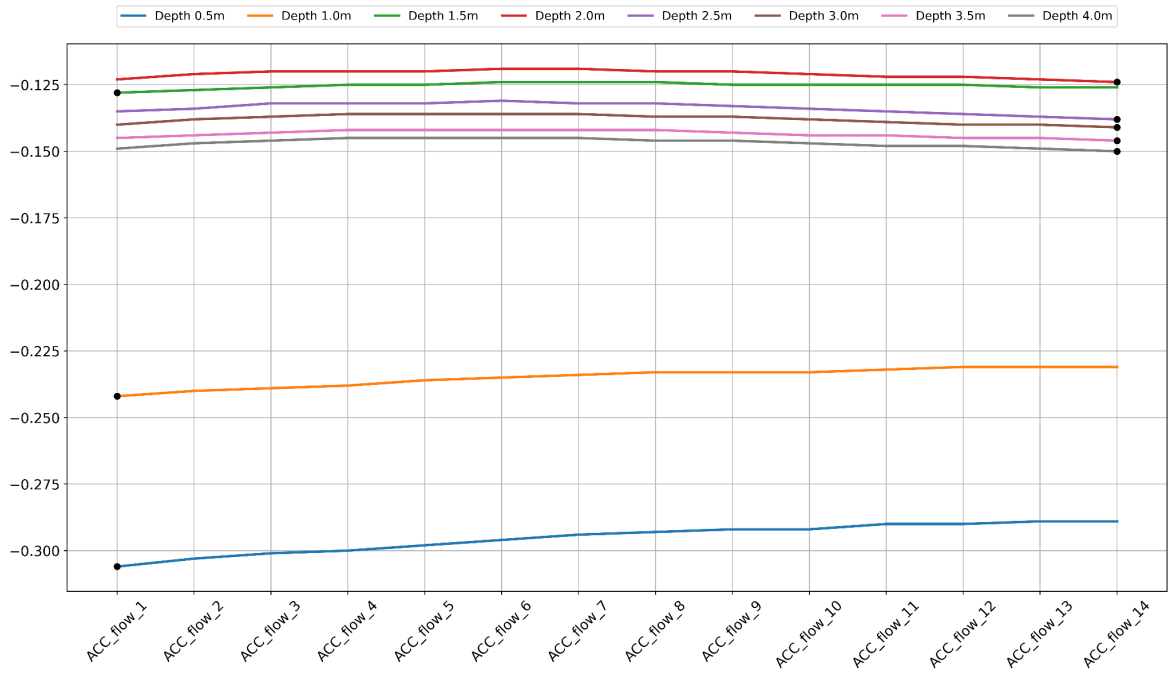


Figure 9. Correlation of aggregated river discharge with Salinity levels at Station 0

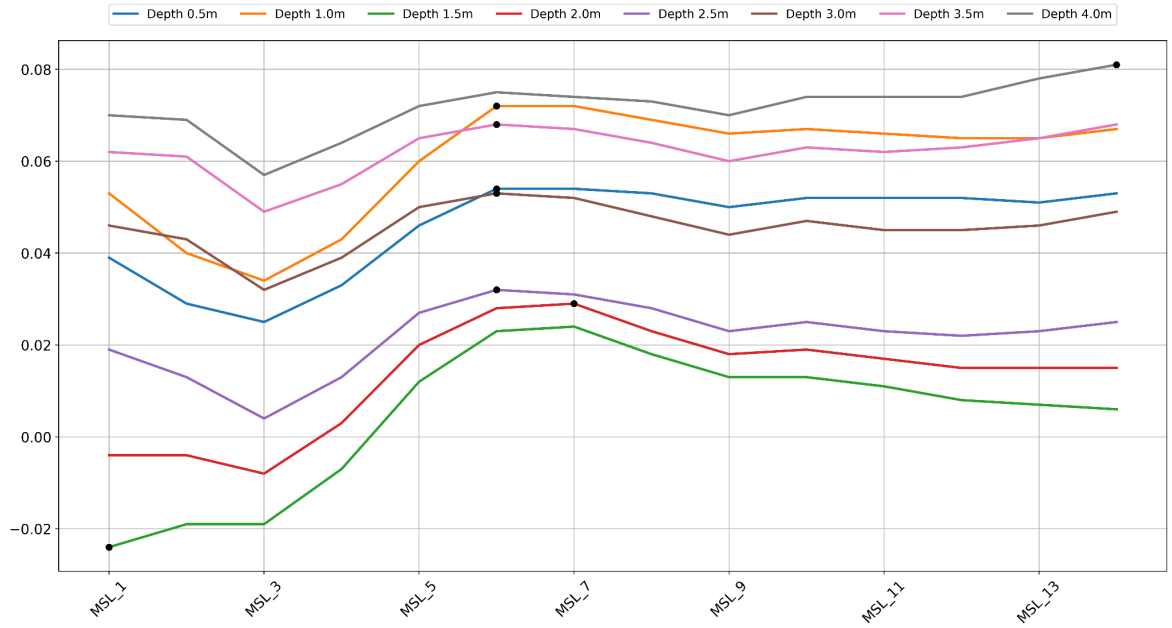


Figure 10. Correlation of aggregated sea level with Salinity levels at Station 0

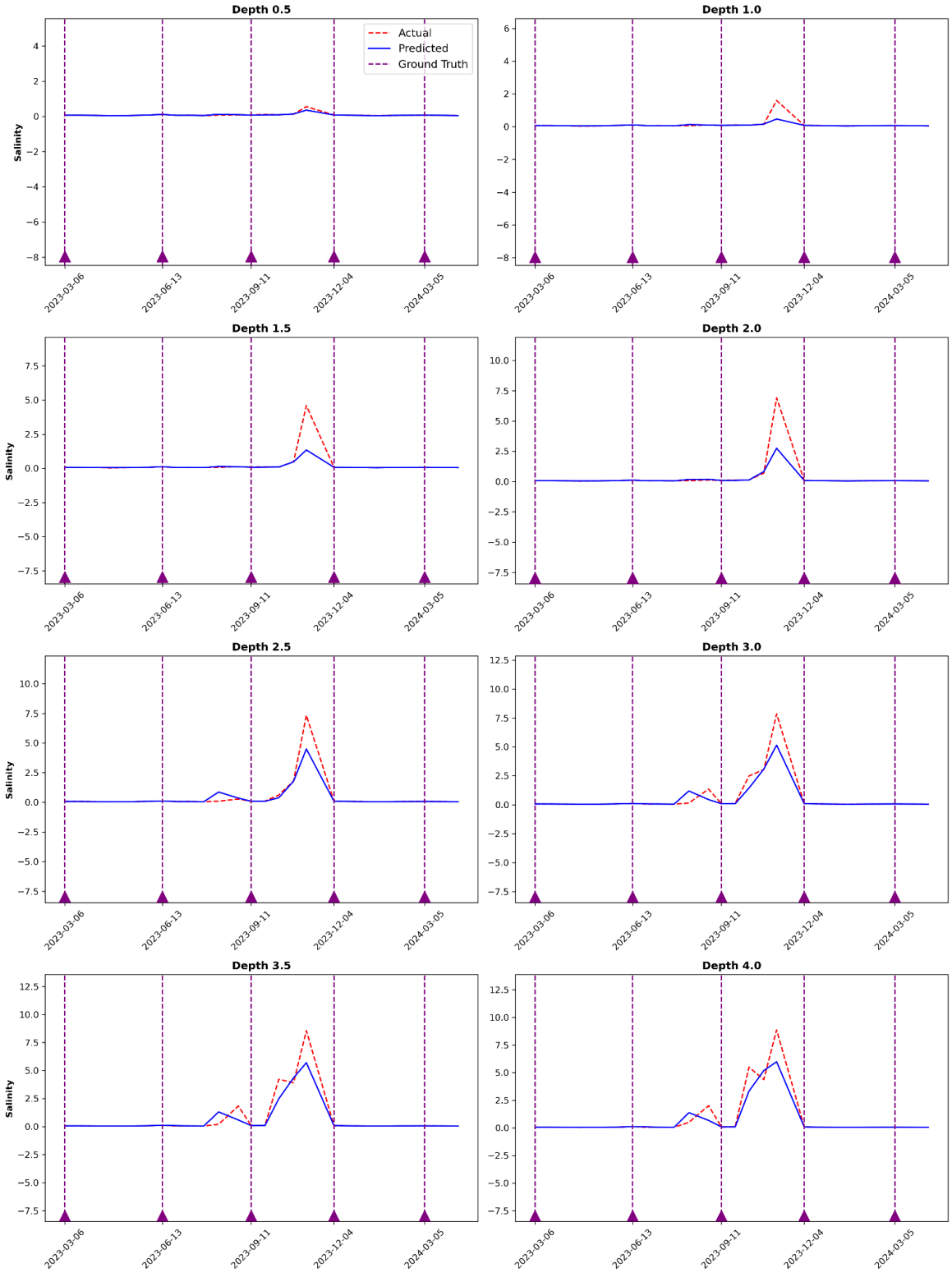


Figure 11 .Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 0

Station 20

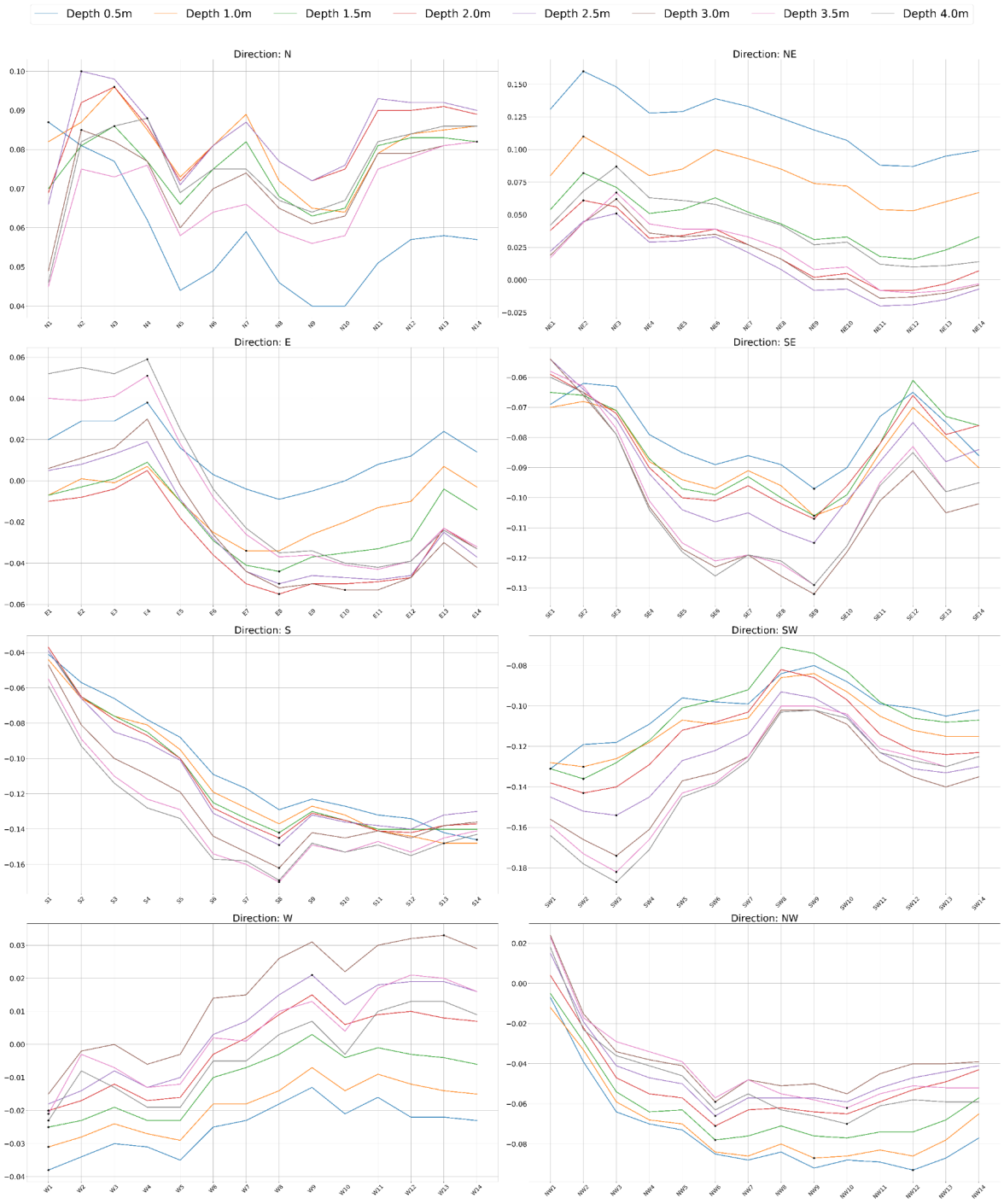


Figure 12 . Correlation of aggregated wind over time with Salinity levels at Station 20

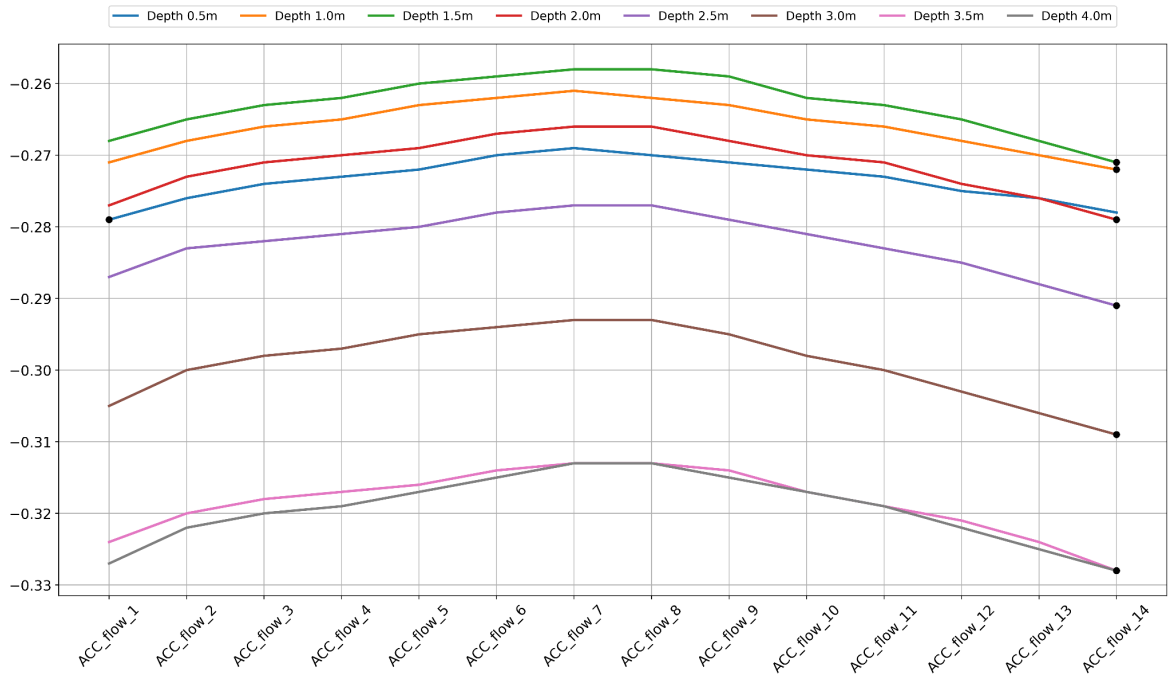


Figure 13. Correlation of aggregated river discharge with Salinity levels at Station 20

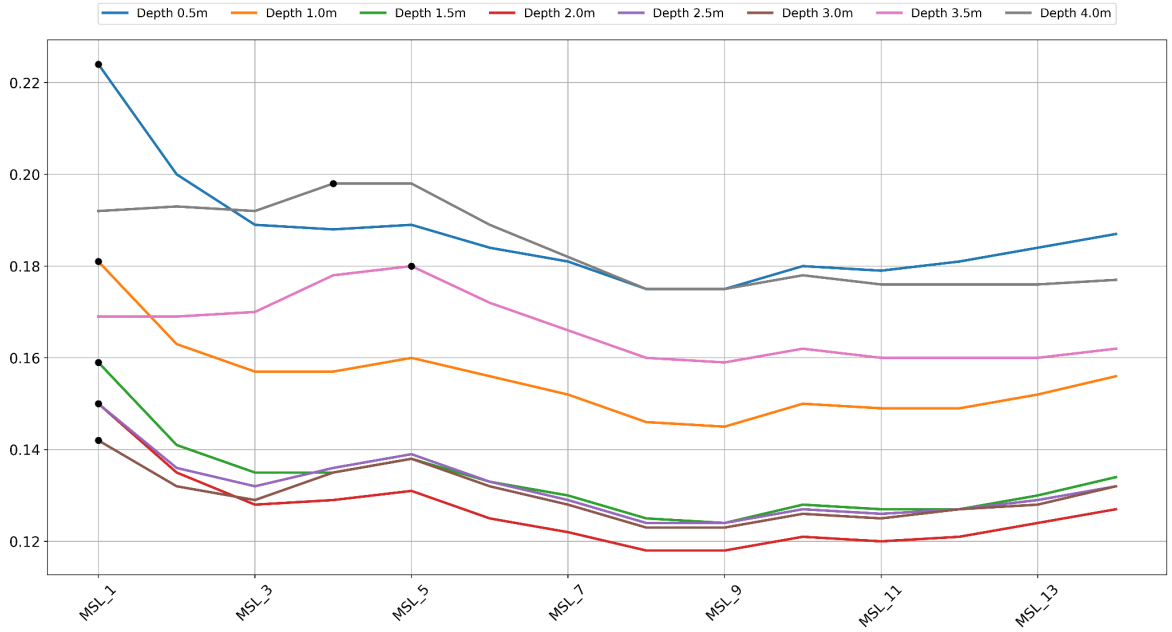


Figure 14. Correlation of aggregated sea level with Salinity levels at Station 20

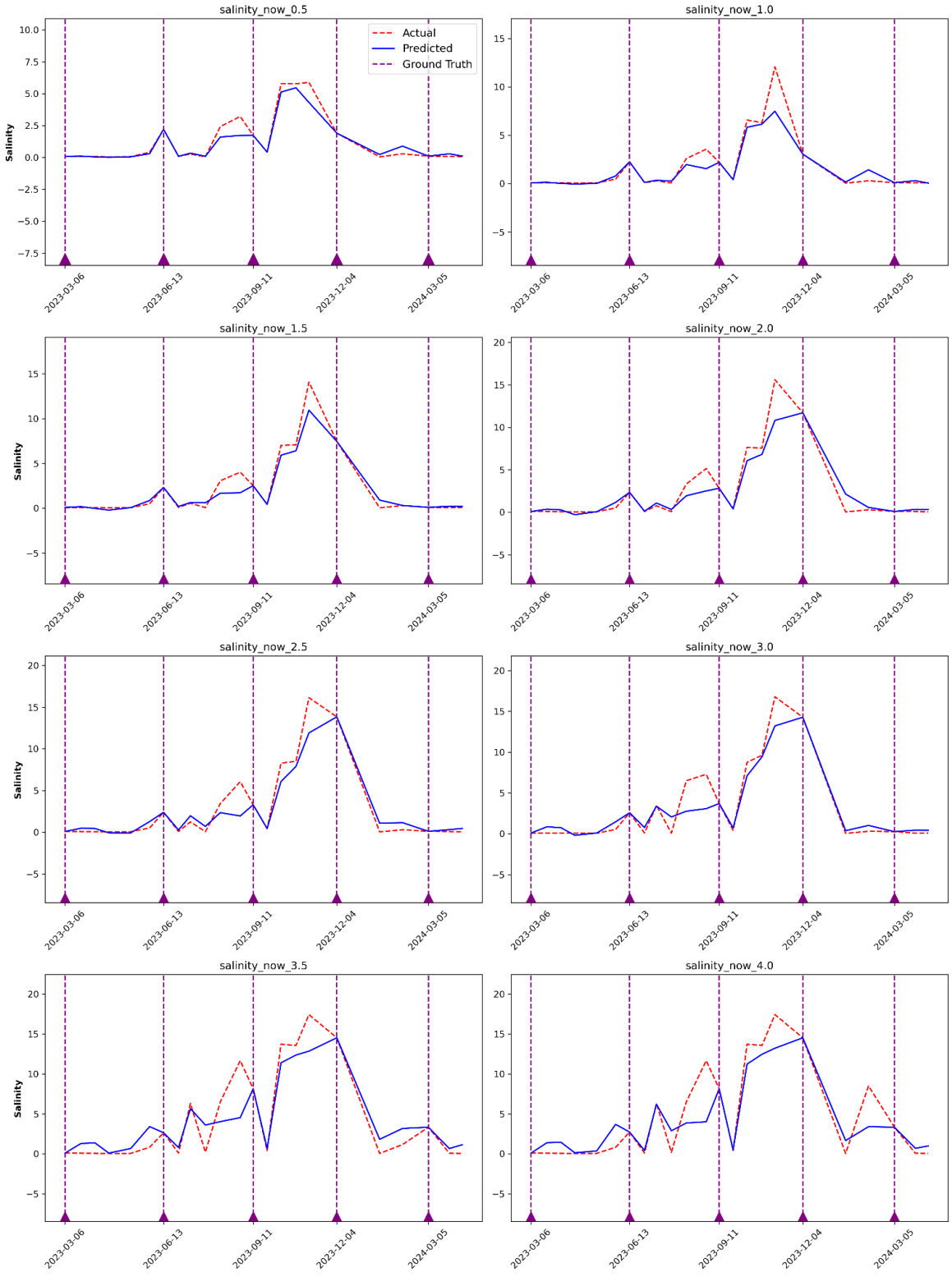


Figure 15.Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 20

Station 30

Figure 16 . Correlation of aggregated wind over time with Salinity levels at Station 30

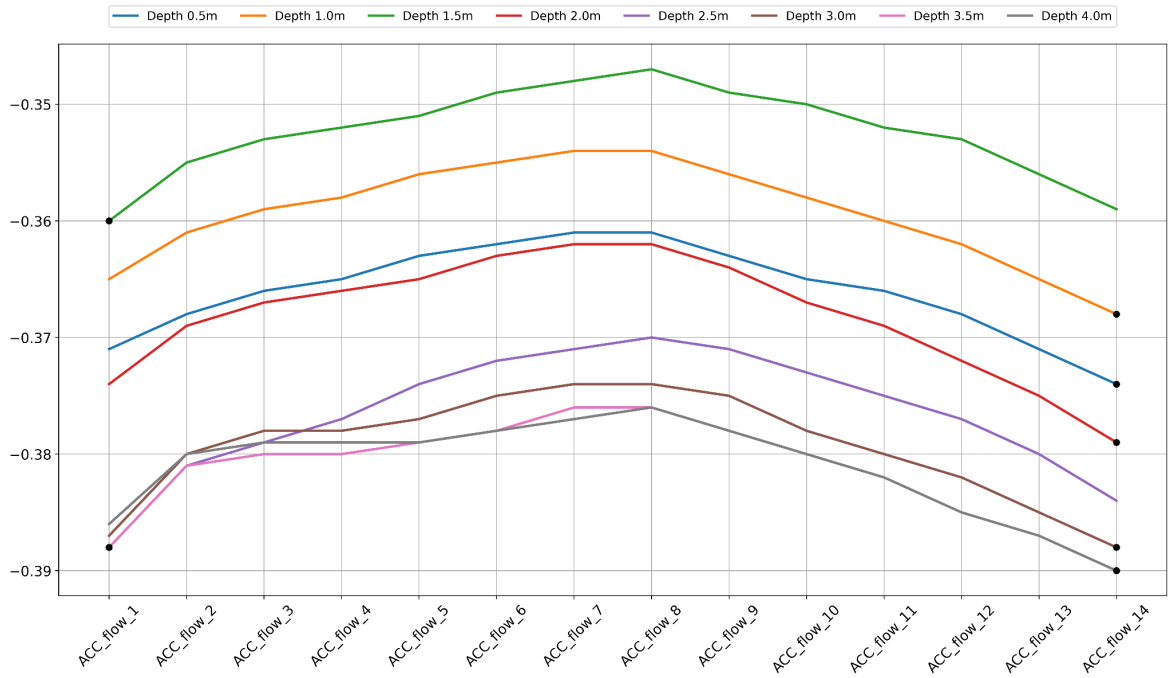


Figure 17. Correlation of aggregated river discharge with Salinity levels at Station 30

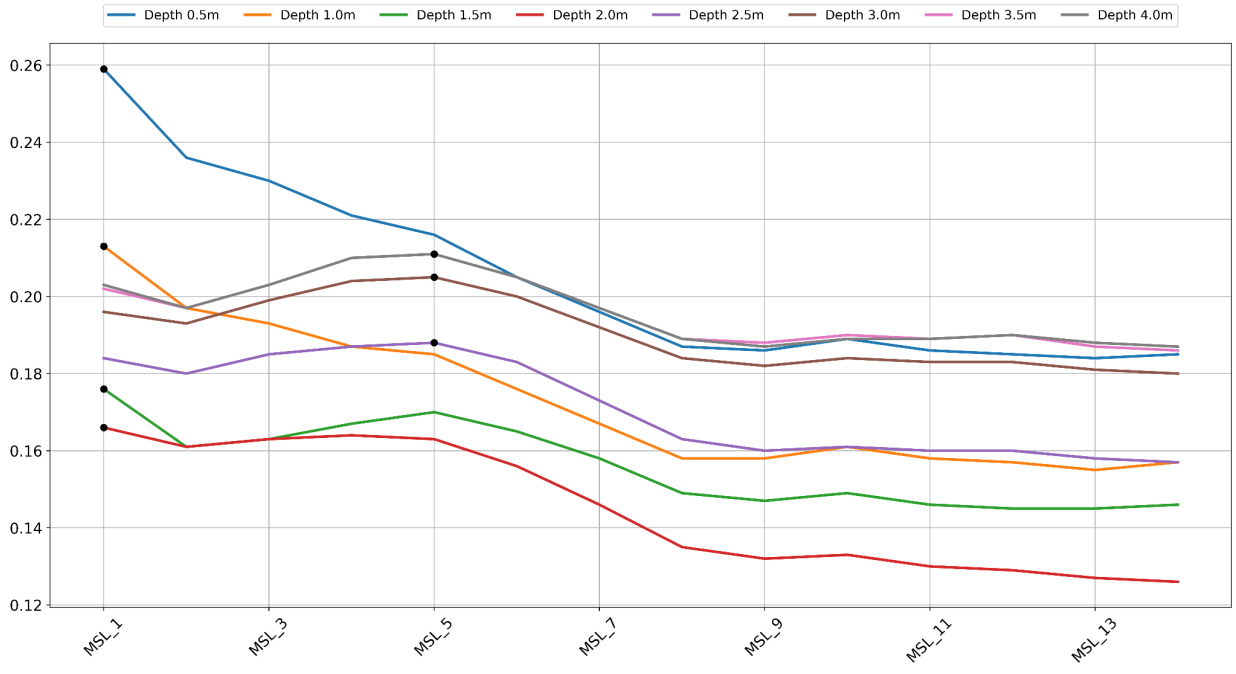


Figure 18. Correlation of aggregated sea level with Salinity levels at Station 30

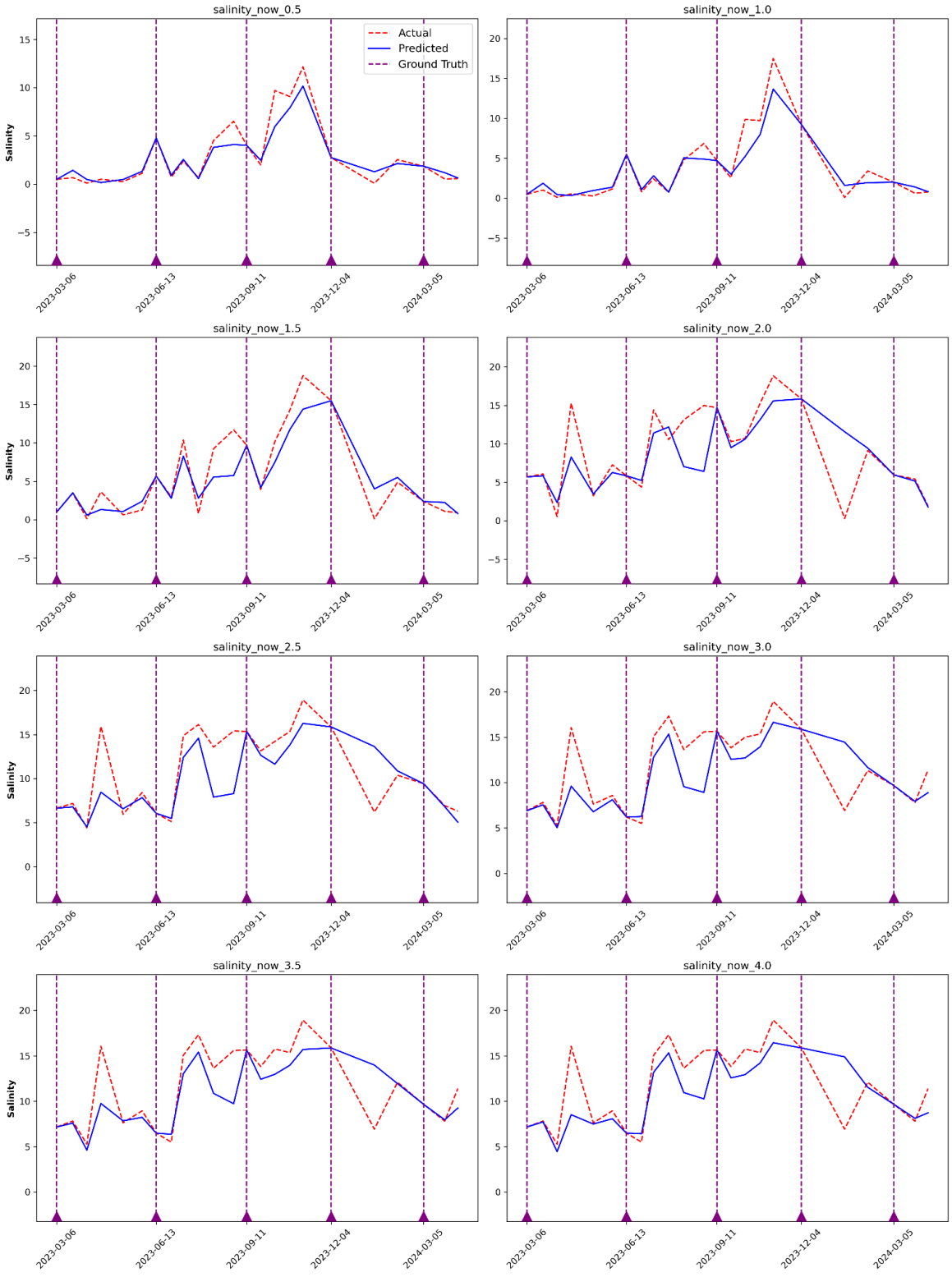


Figure 19 .Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 30

Station 50

Figure 20 . Correlation of aggregated wind over time with Salinity levels at Station 50

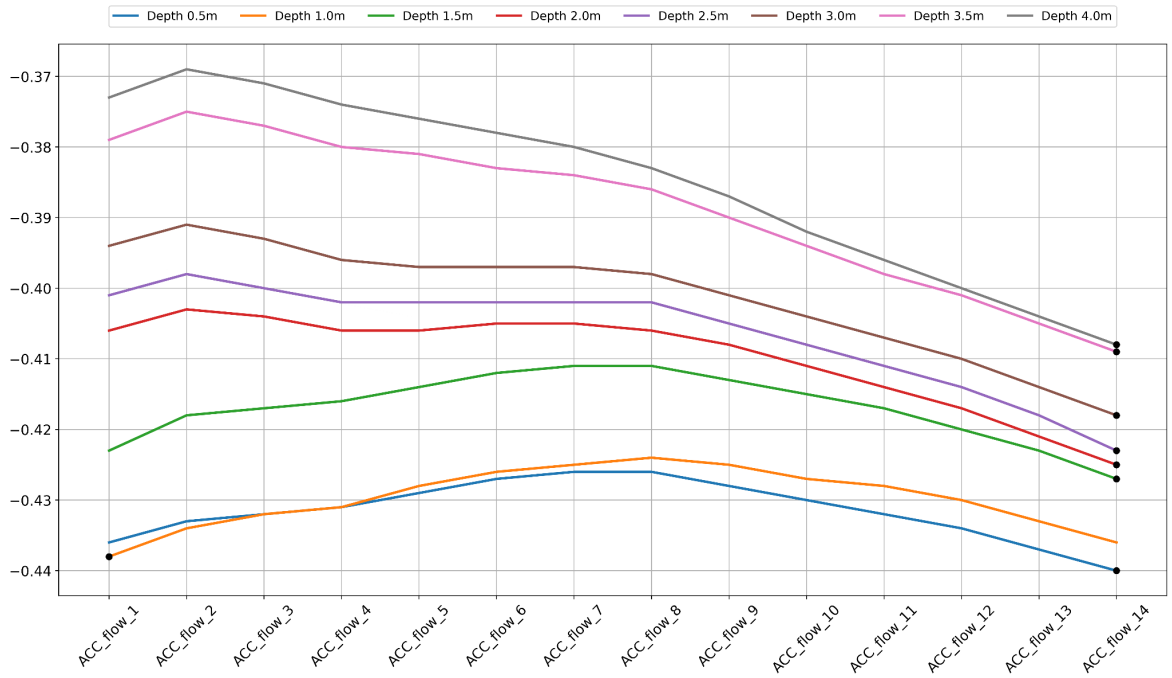


Figure 21 . Correlation of aggregated river discharge with Salinity levels at Station 50

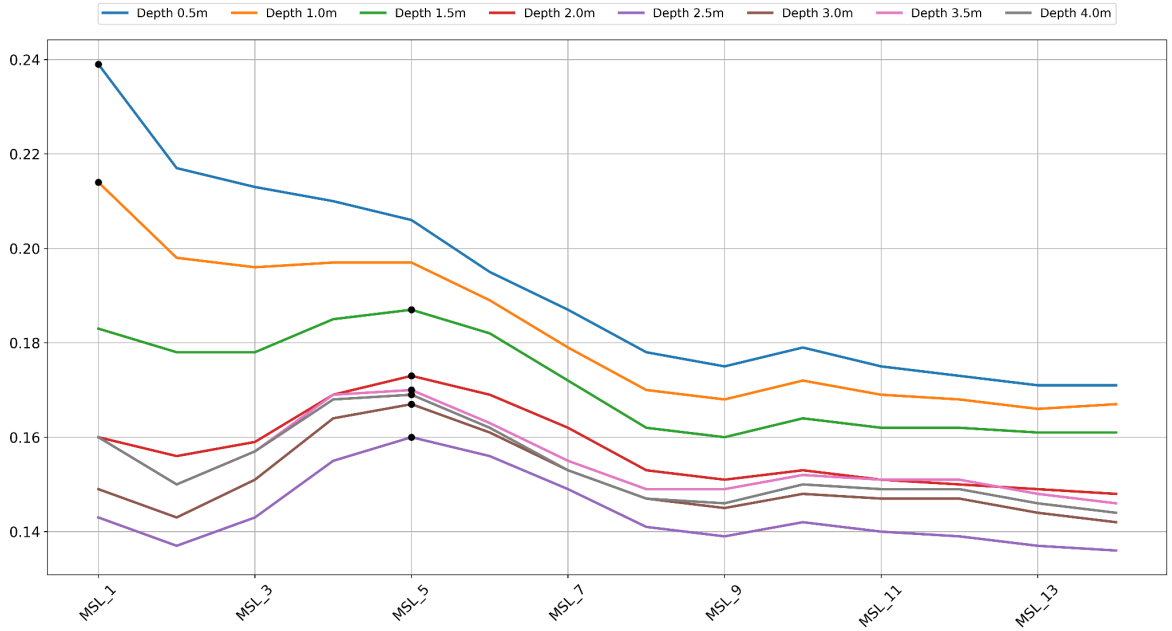


Figure 22. Correlation of aggregated sea level with Salinity levels at Station 50

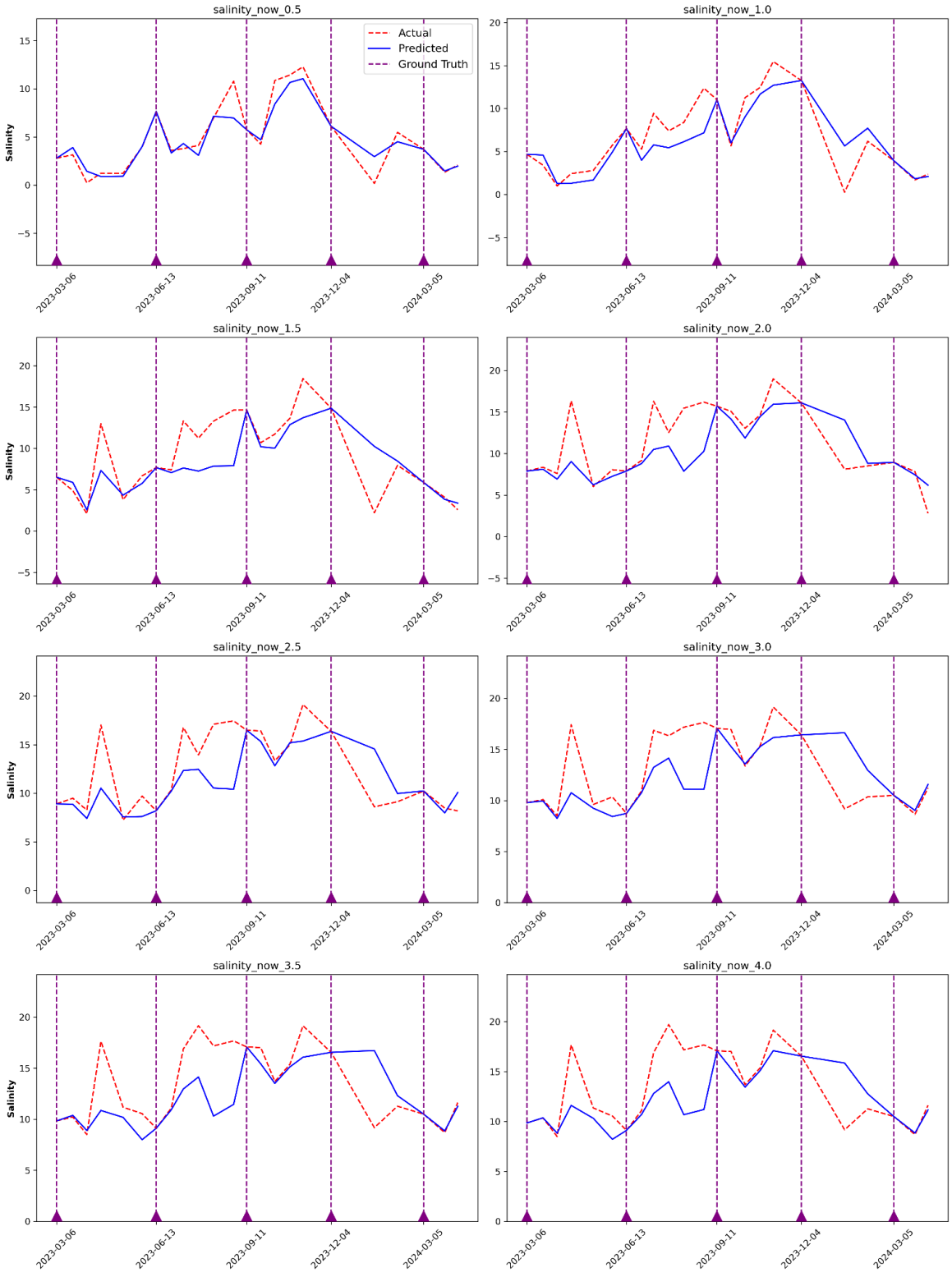


Figure 23.Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 50

Station 60

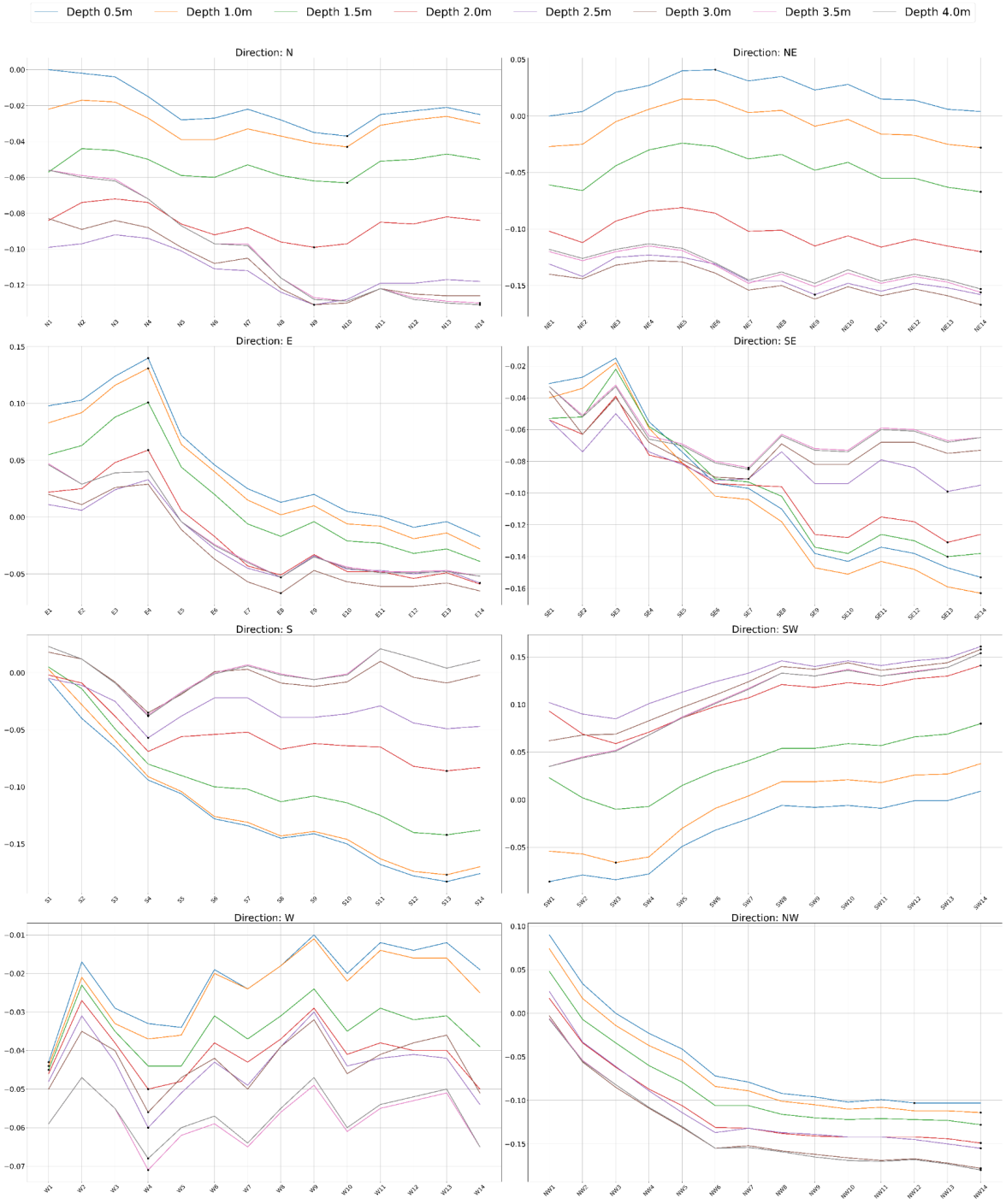


Figure 24 . Correlation of aggregated wind over time with Salinity levels at Station 60

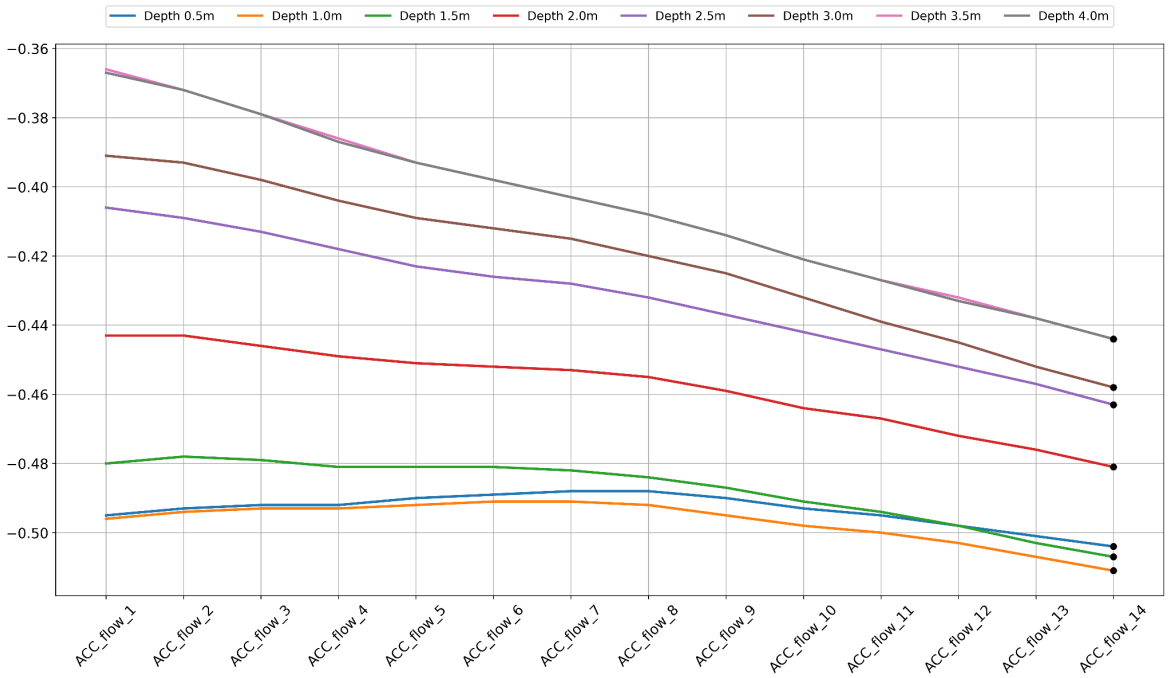


Figure 25 . Correlation of aggregated river discharge with Salinity levels at Station 60

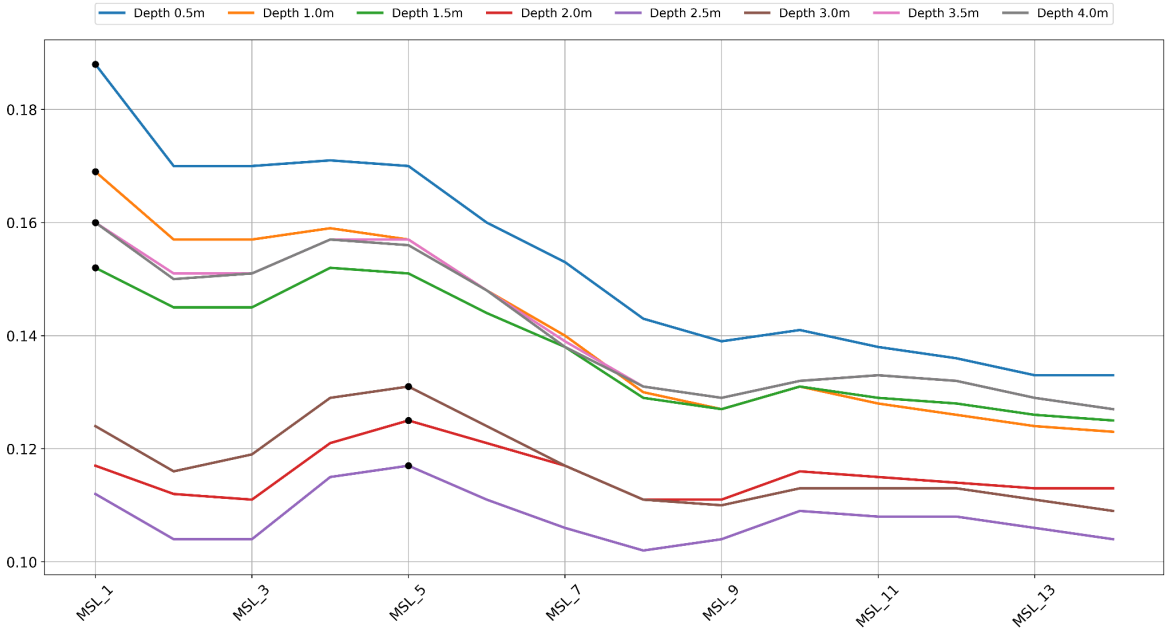


Figure 26. Correlation of aggregated sea level with Salinity levels at Station 60

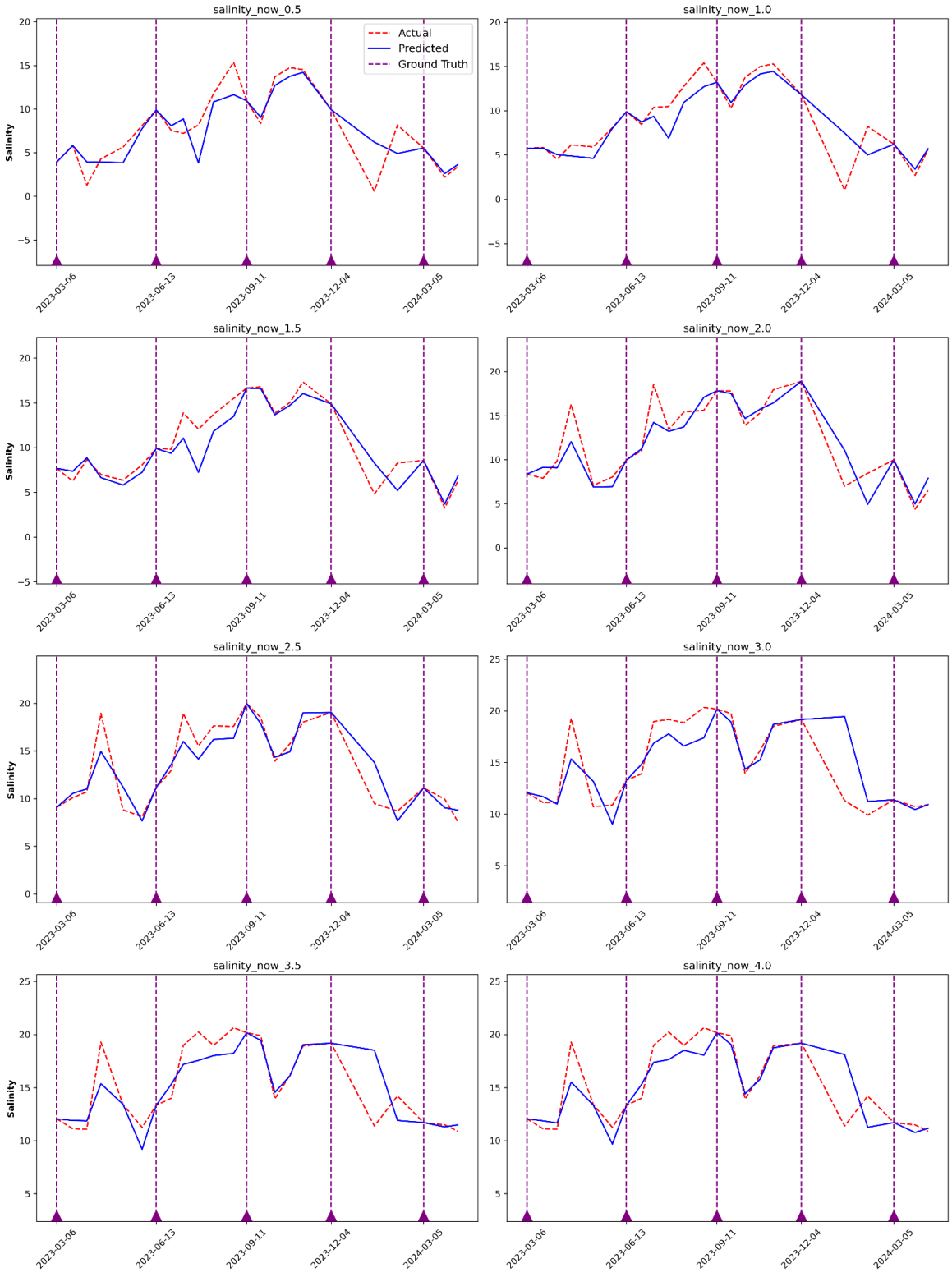


Figure 27 .Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 60

Station 70

Figure 28 . Correlation of aggregated wind over time with Salinity levels at Station 70

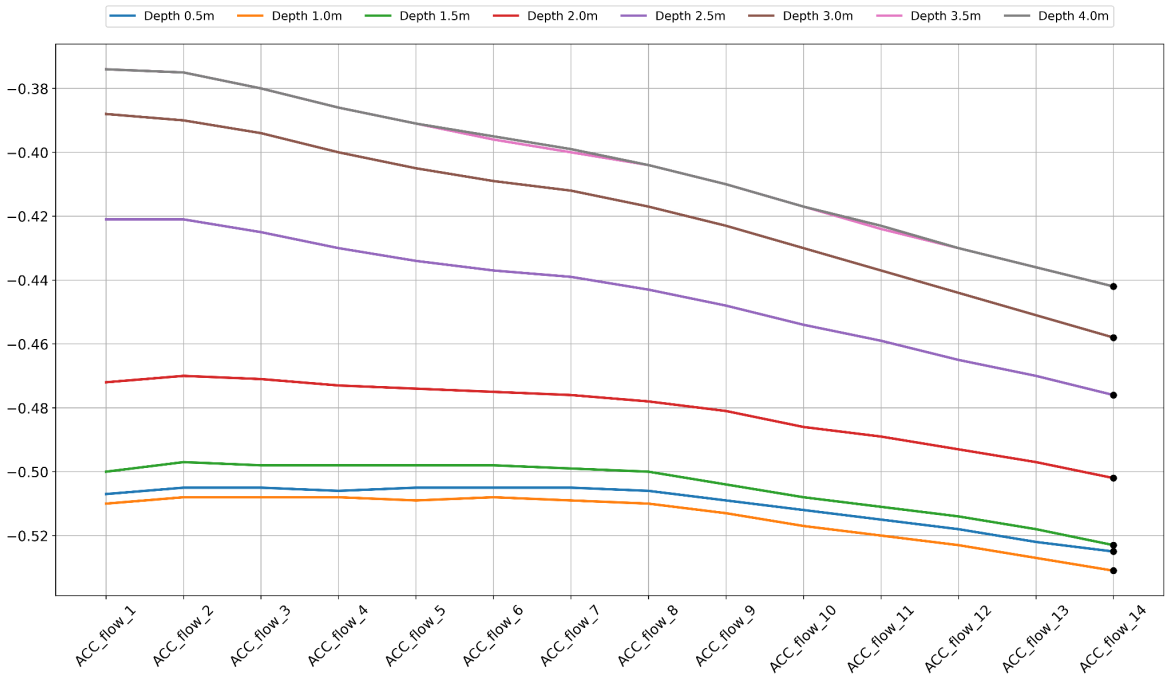


Figure 29 . Correlation of aggregated river discharge with Salinity levels at Station 70

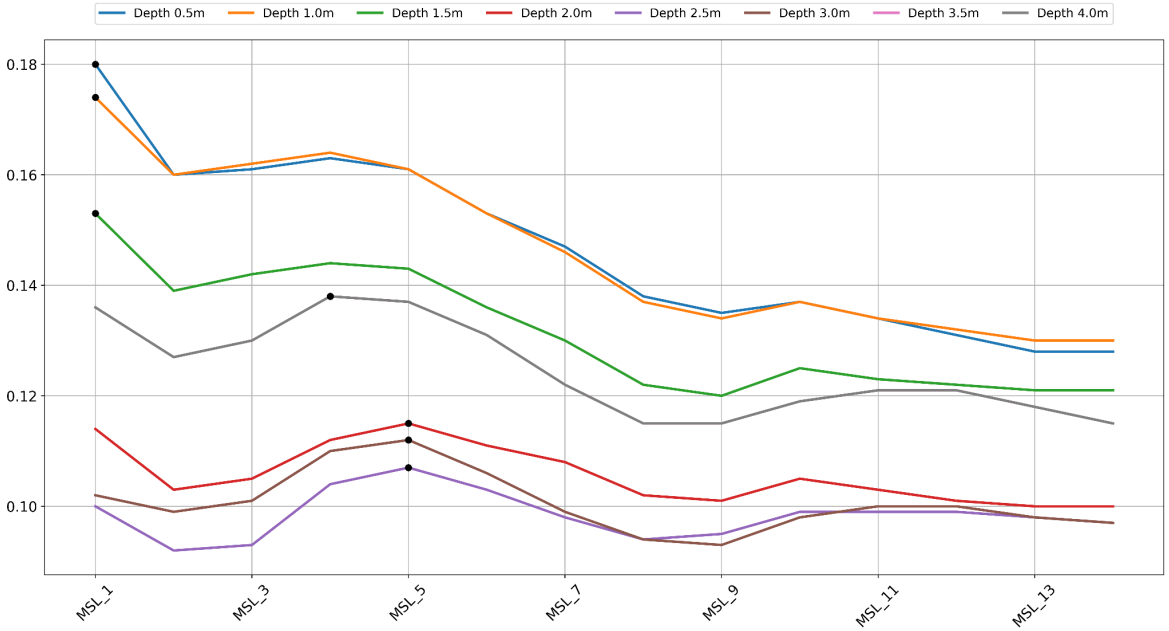


Figure 30. Correlation of aggregated sea level with Salinity levels at Station 70

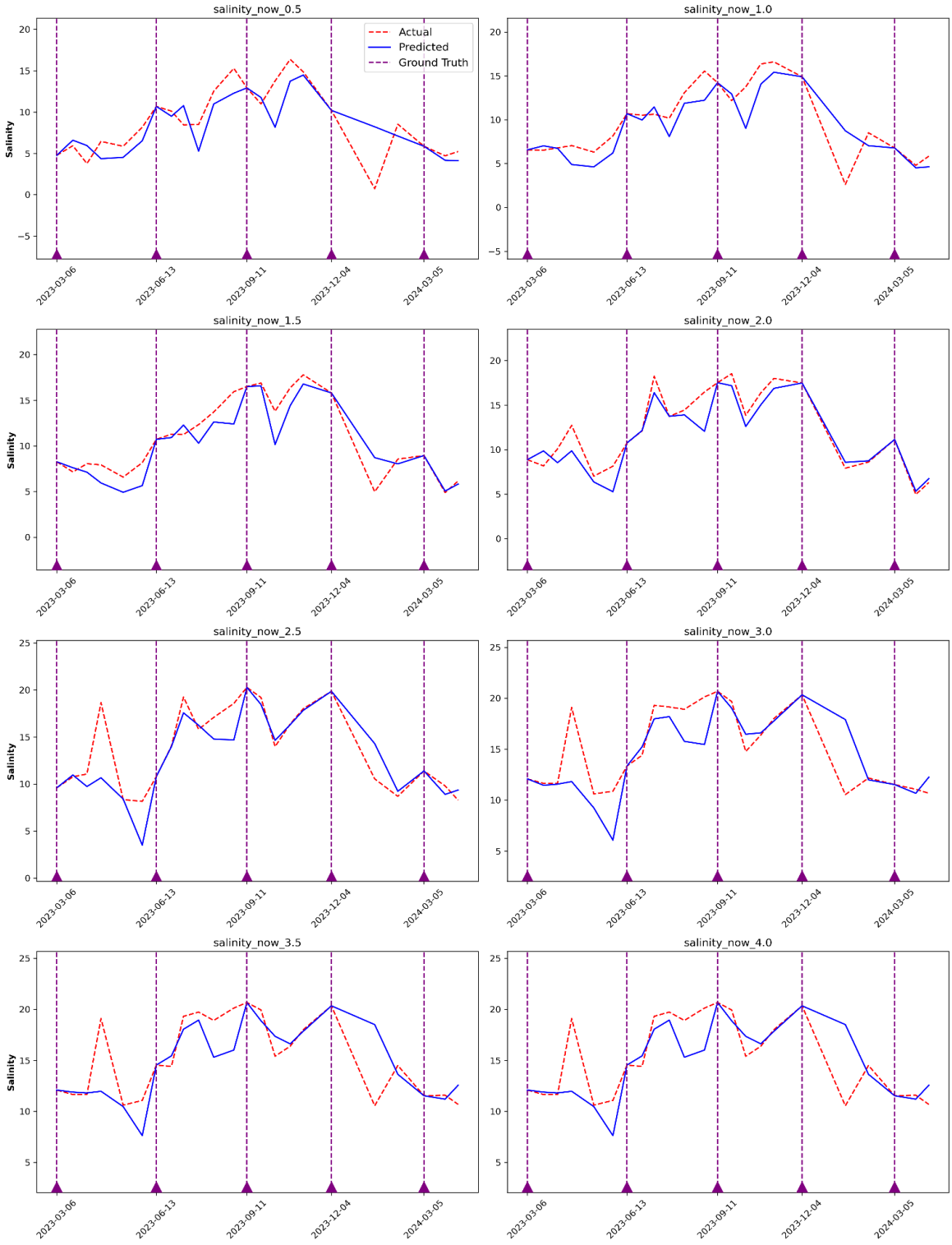


Figure 31.Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 70

Station 120

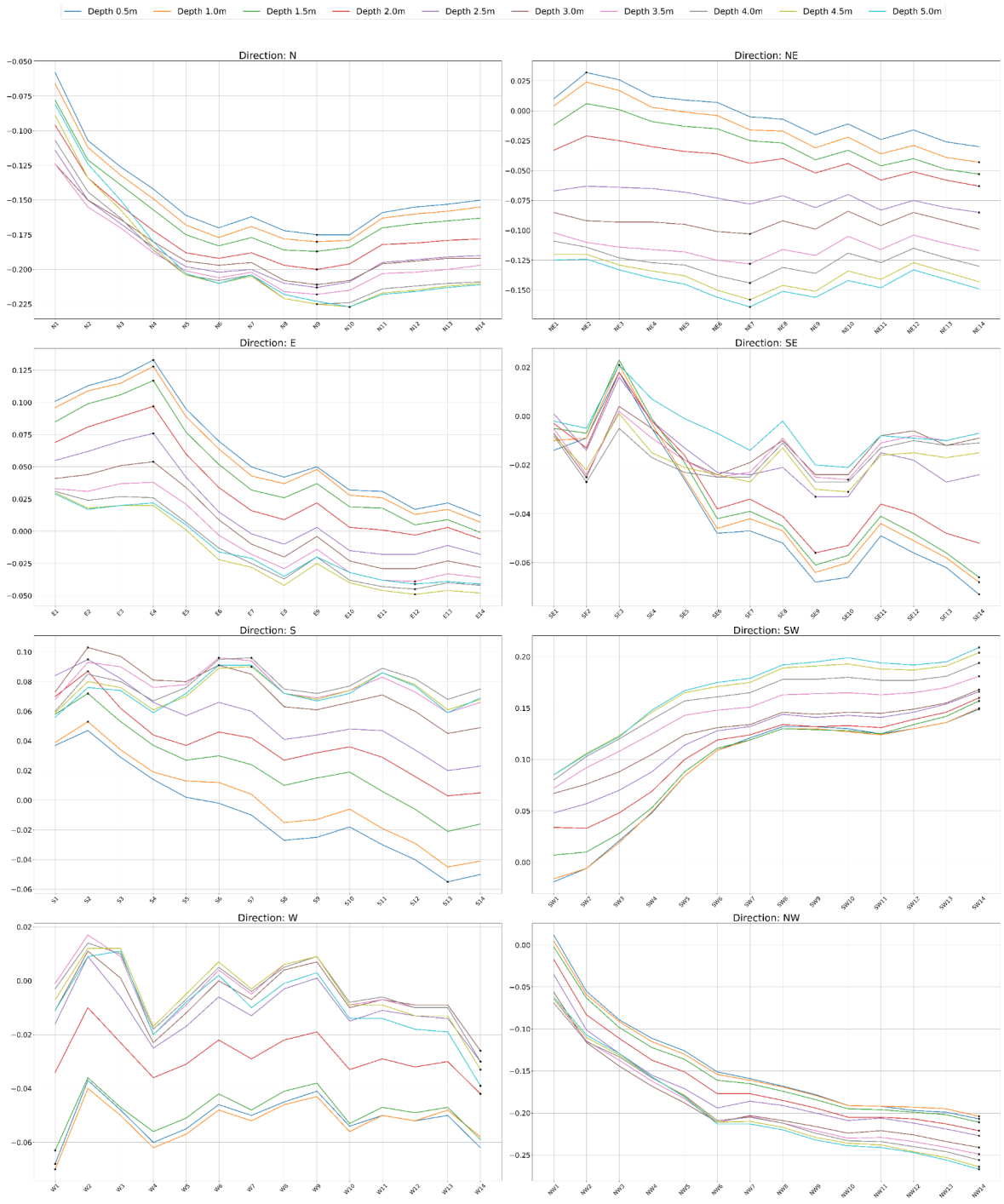


Figure 32 . Correlation of aggregated wind over time with Salinity levels at Station 120

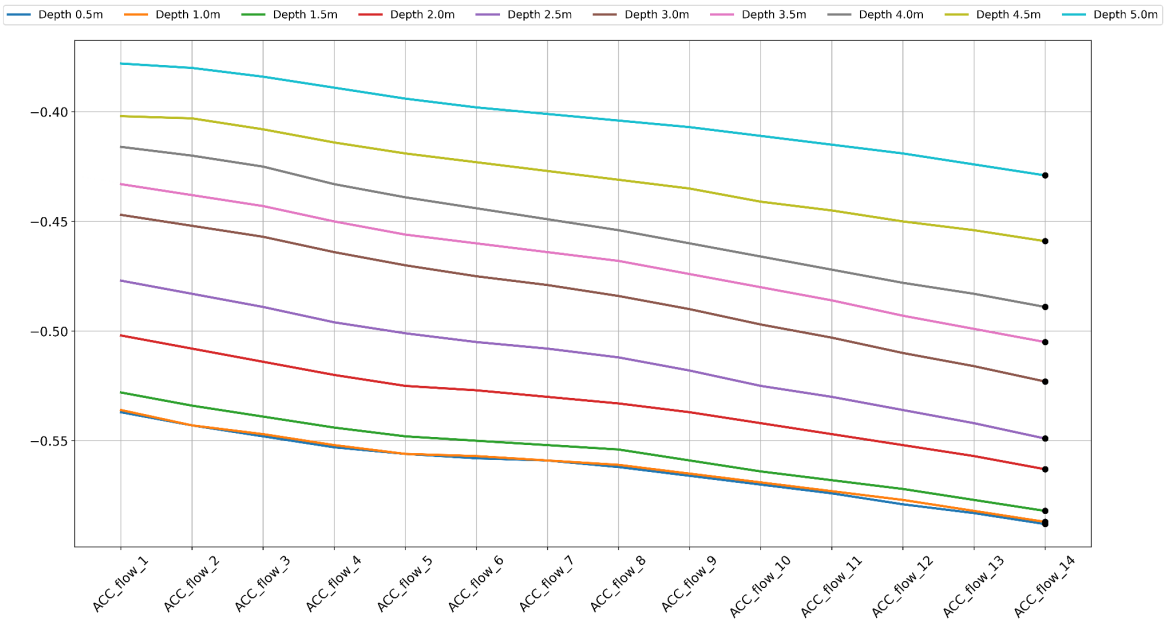


Figure 33 . Correlation of aggregated river discharge with Salinity levels at Station 120

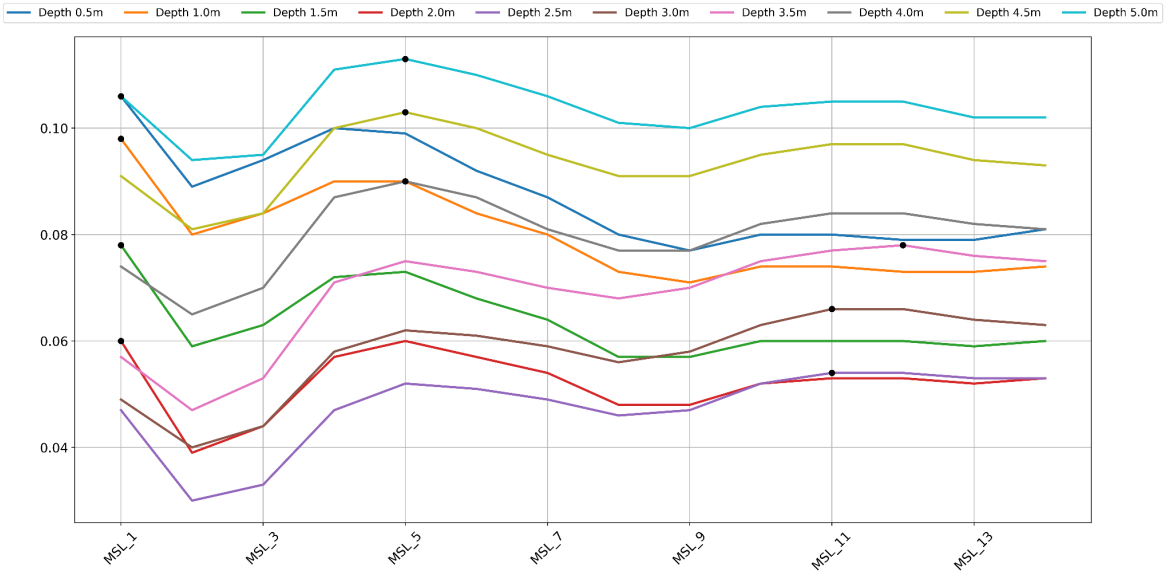


Figure 34 . Correlation of aggregated sea level with Salinity levels at Station 120

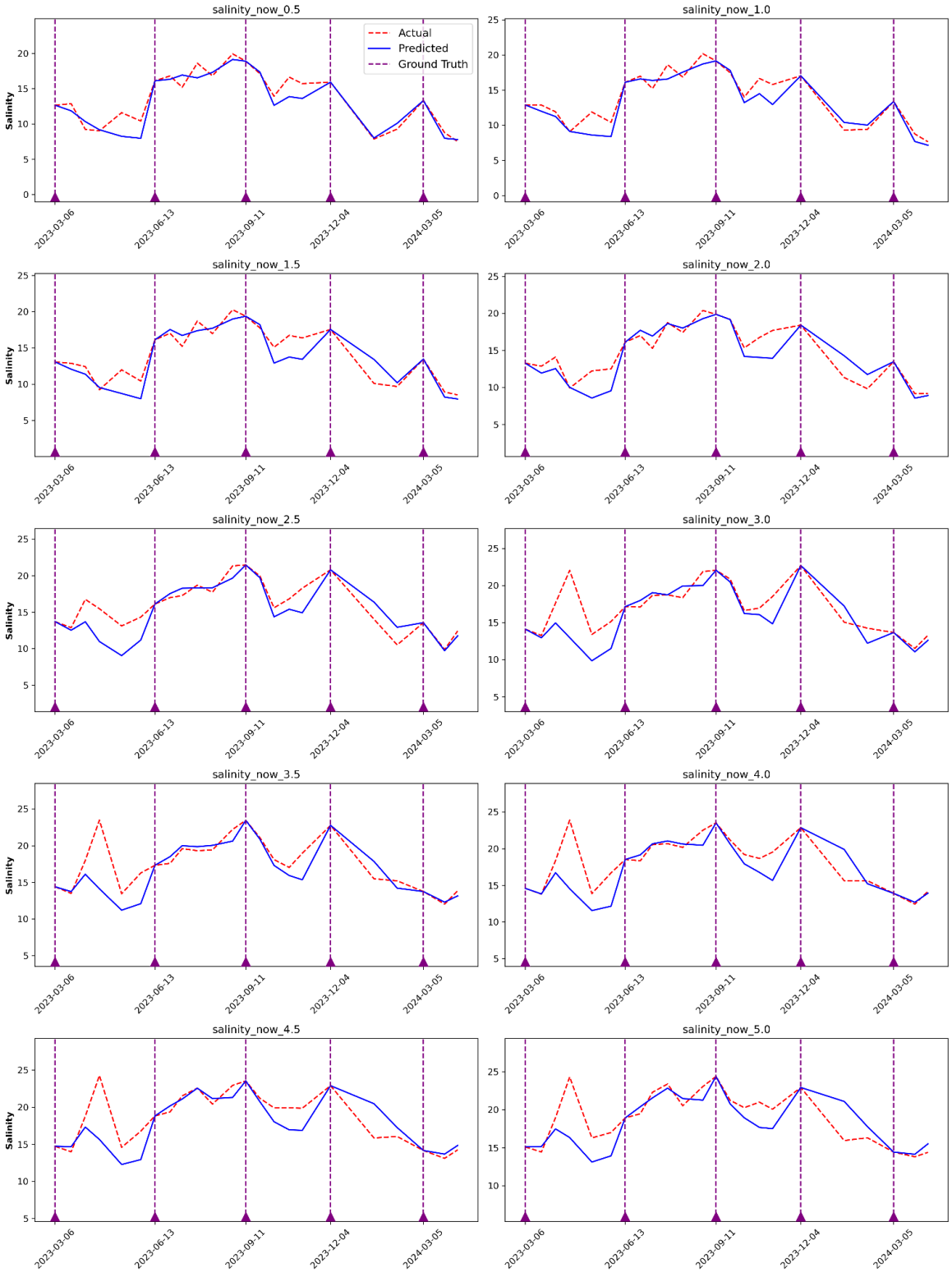


Figure 35 .Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 120

Station 140

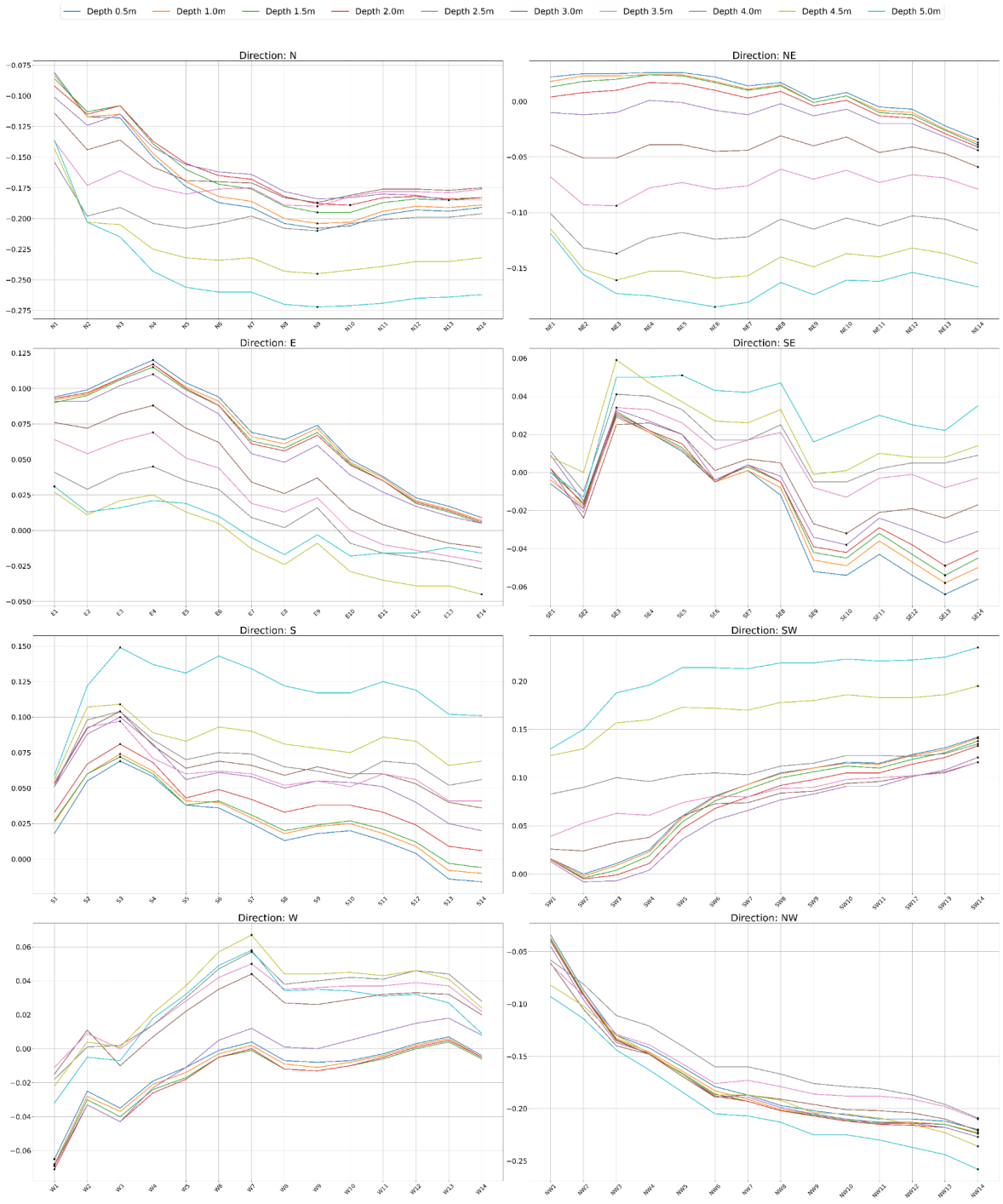


Figure 36 . Correlation of aggregated wind over time with Salinity levels at Station 140

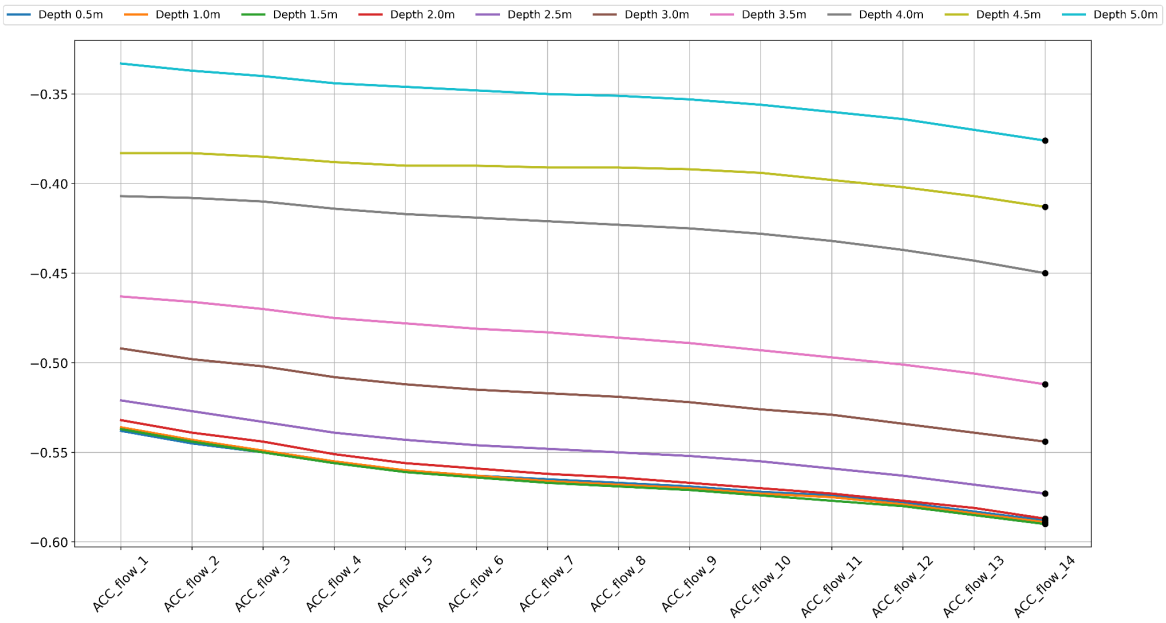


Figure 37 . Correlation of aggregated river discharge with Salinity levels at Station 140

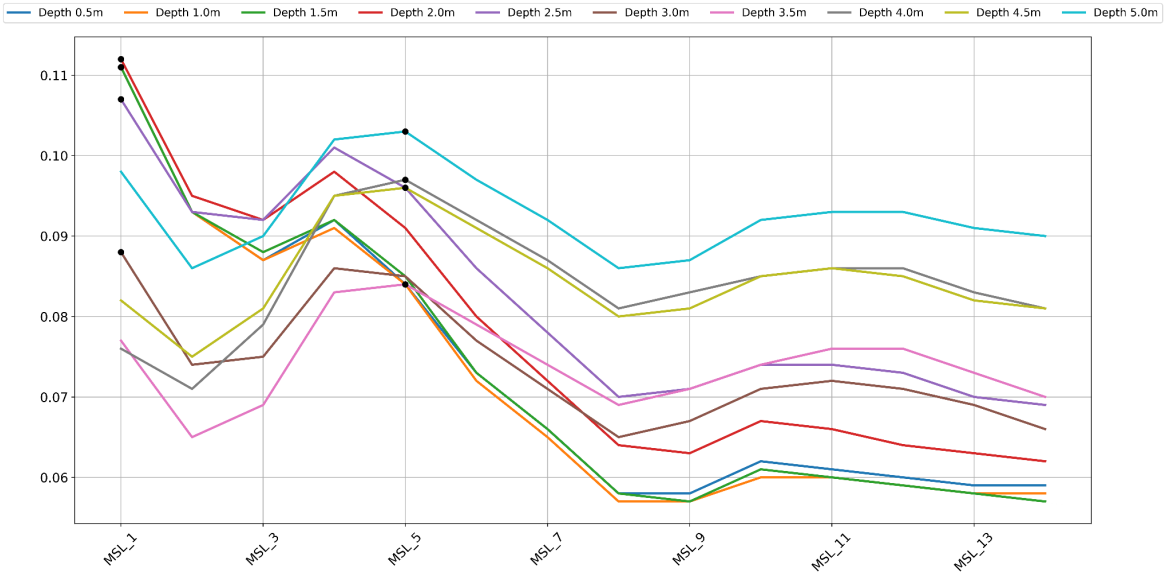


Figure 38. Correlation of aggregated sea level with Salinity levels at Station 140

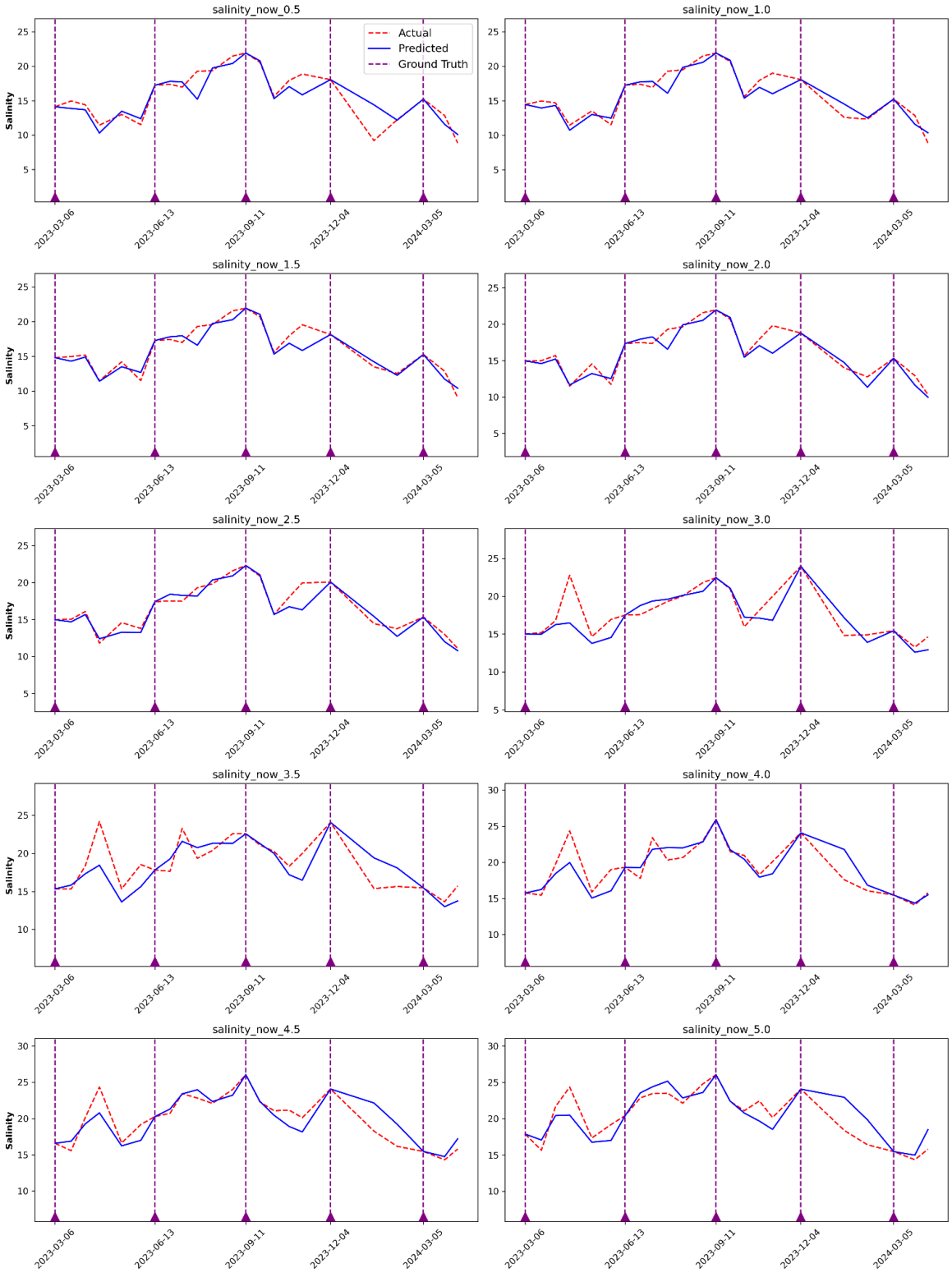


Figure 39.Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 140

Station 160

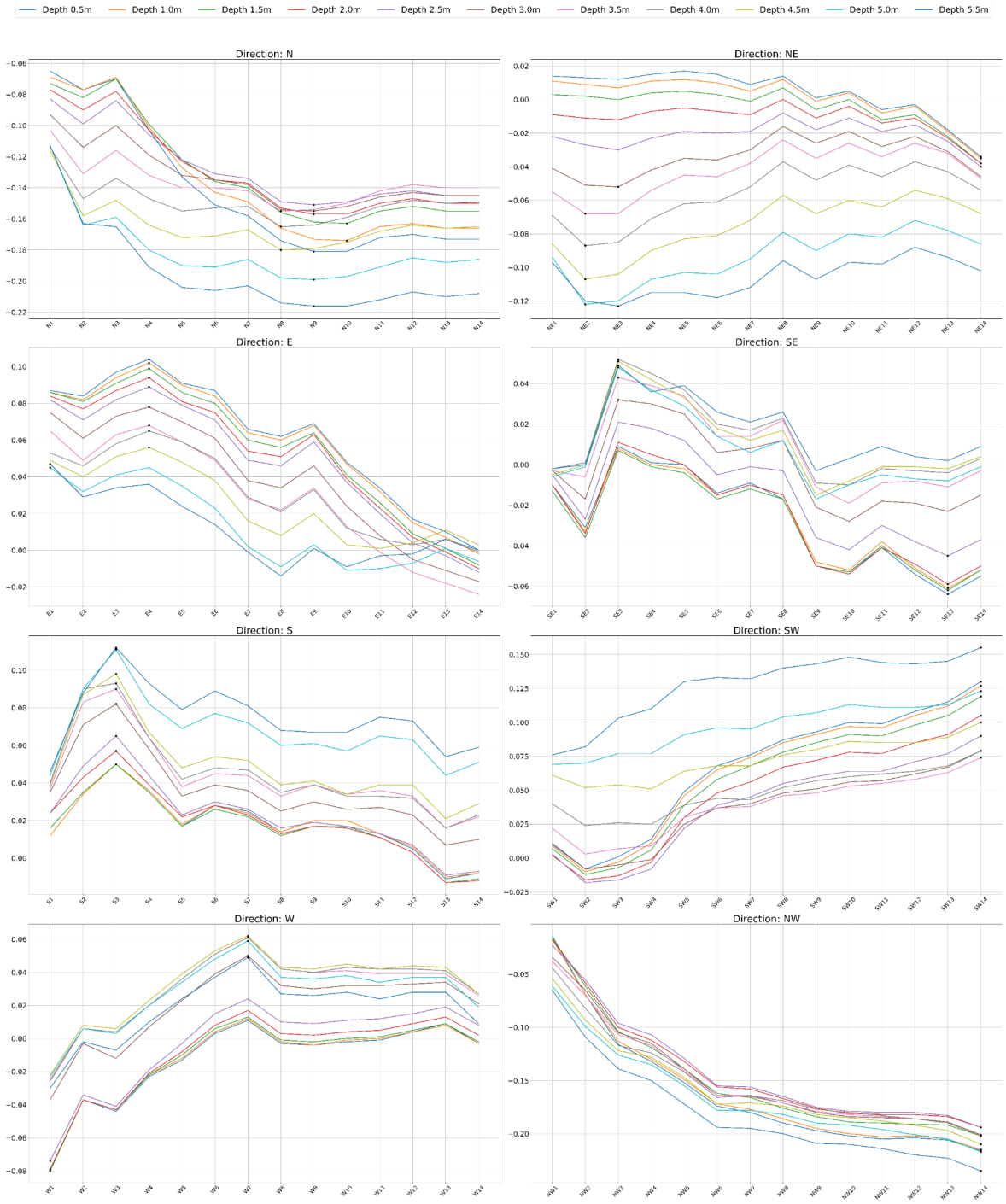


Figure 40 . Correlation of aggregated wind over time with Salinity levels at Station 160

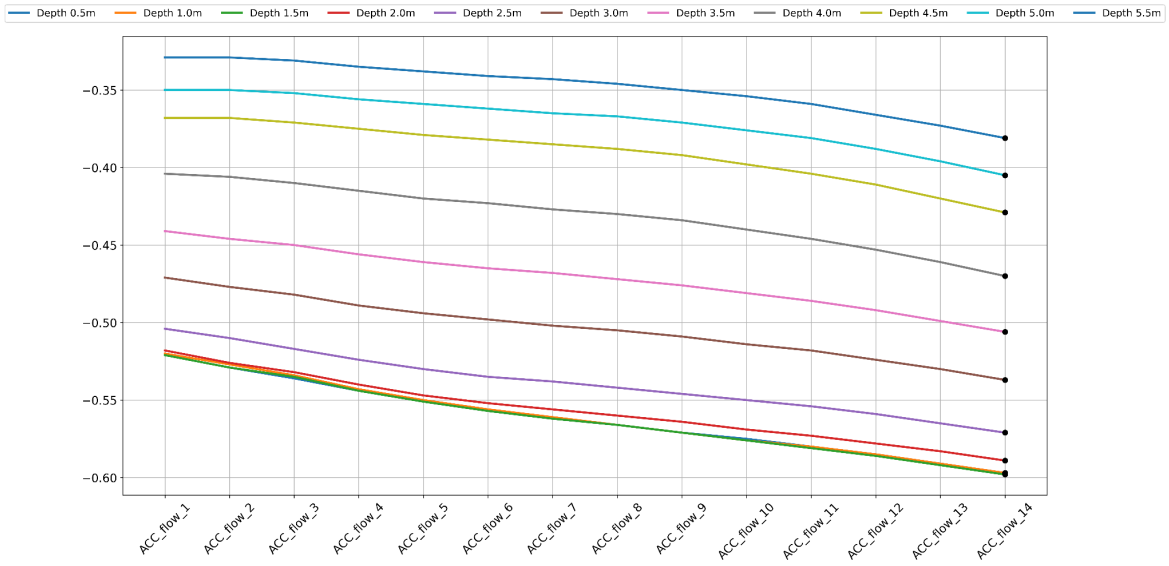


Figure 41 . Correlation of aggregated river discharge with Salinity levels at Station 160

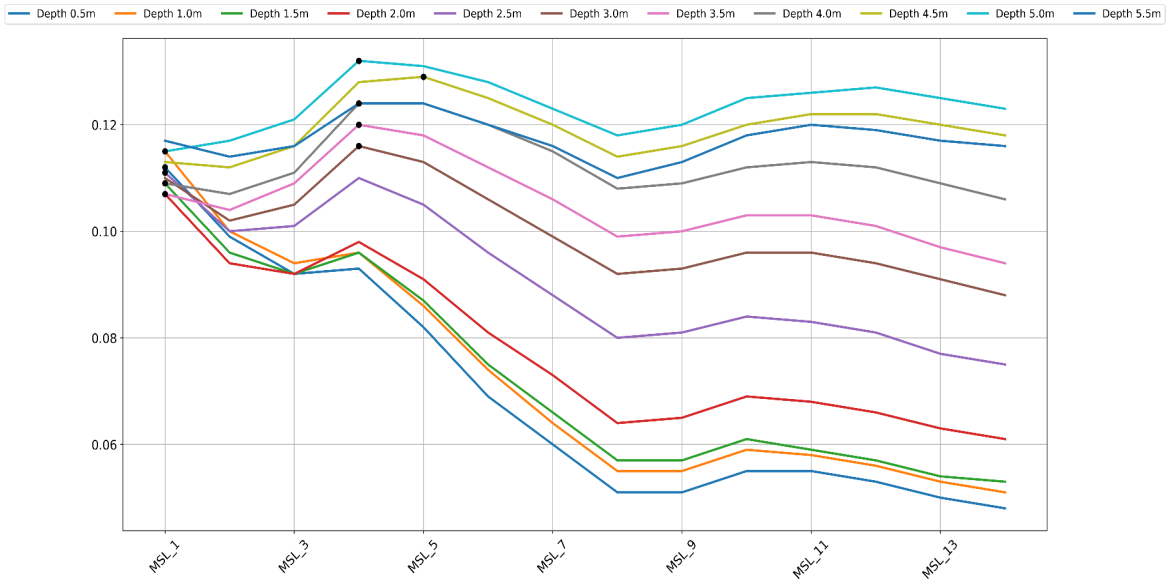


Figure 42. Correlation of aggregated sea level with Salinity levels at Station 160

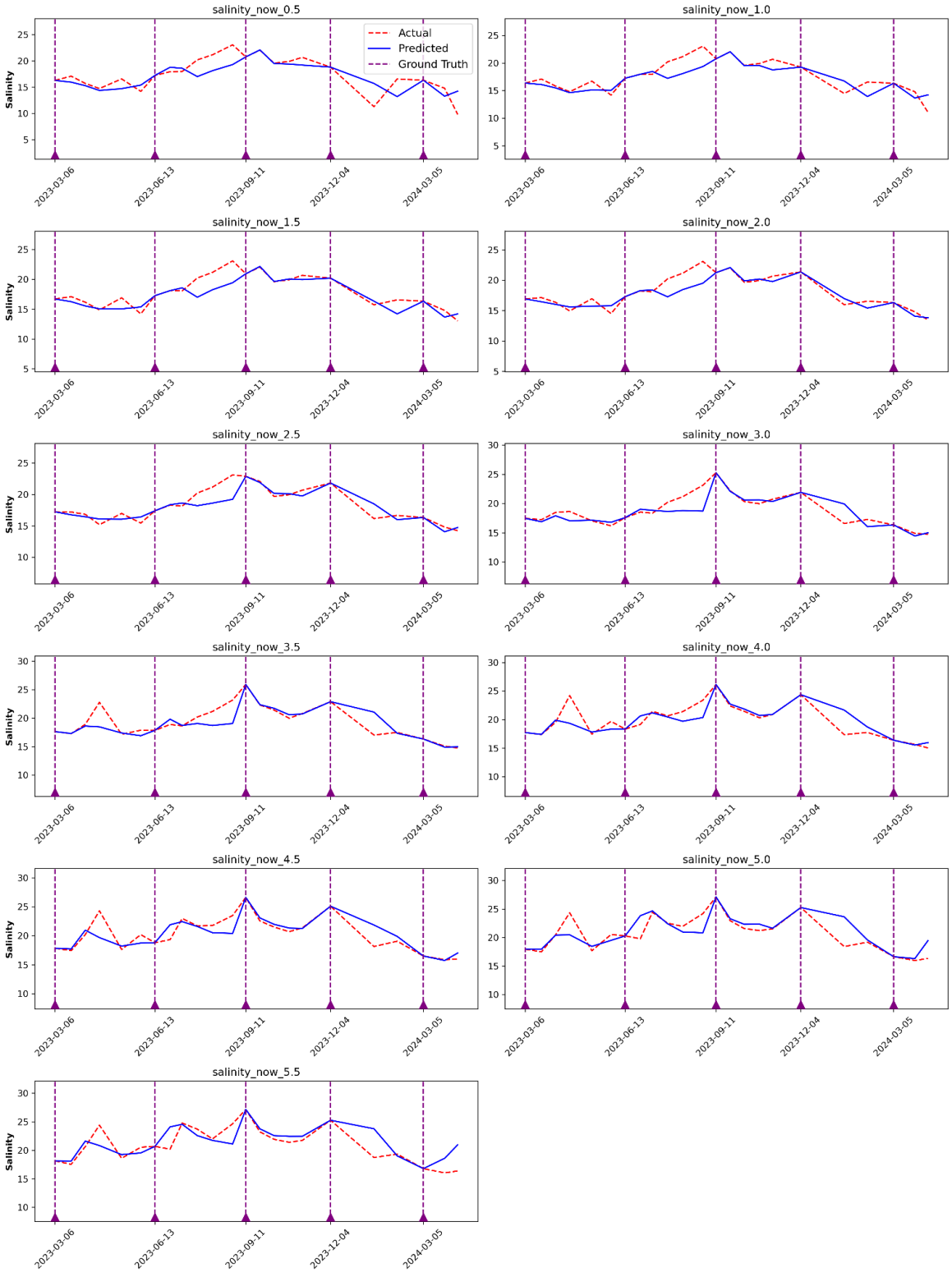


Figure 43. Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 160

Station 180

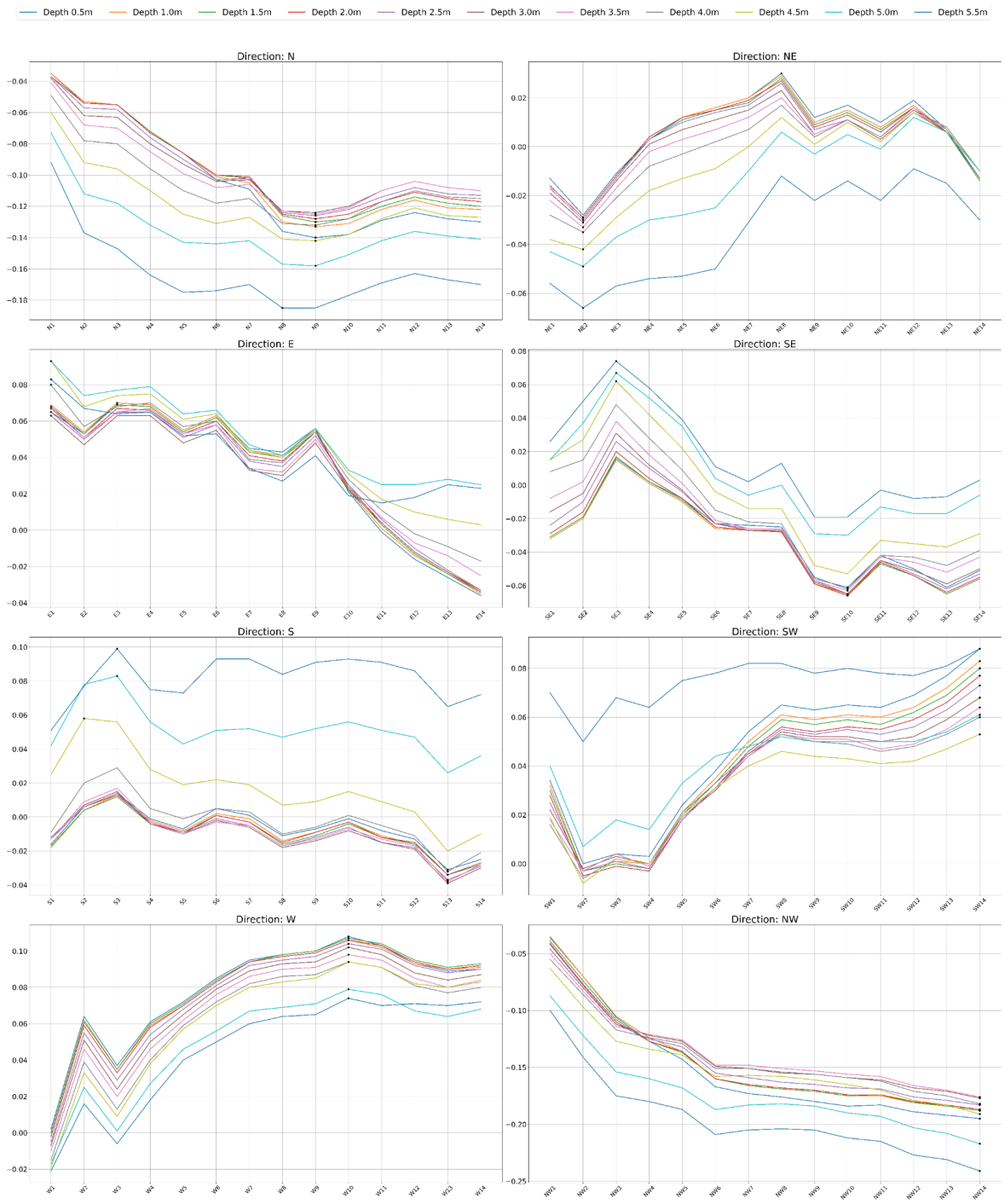


Figure 44. Correlation of aggregated wind over time with Salinity levels at Station 180

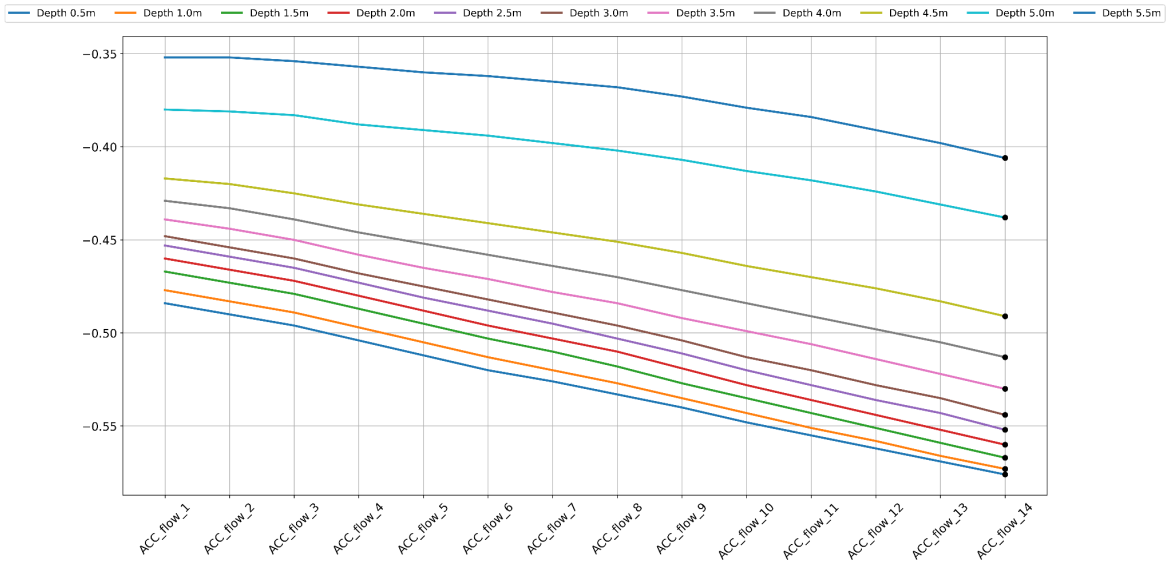


Figure 45. Correlation of aggregated river discharge with Salinity levels at Station 180

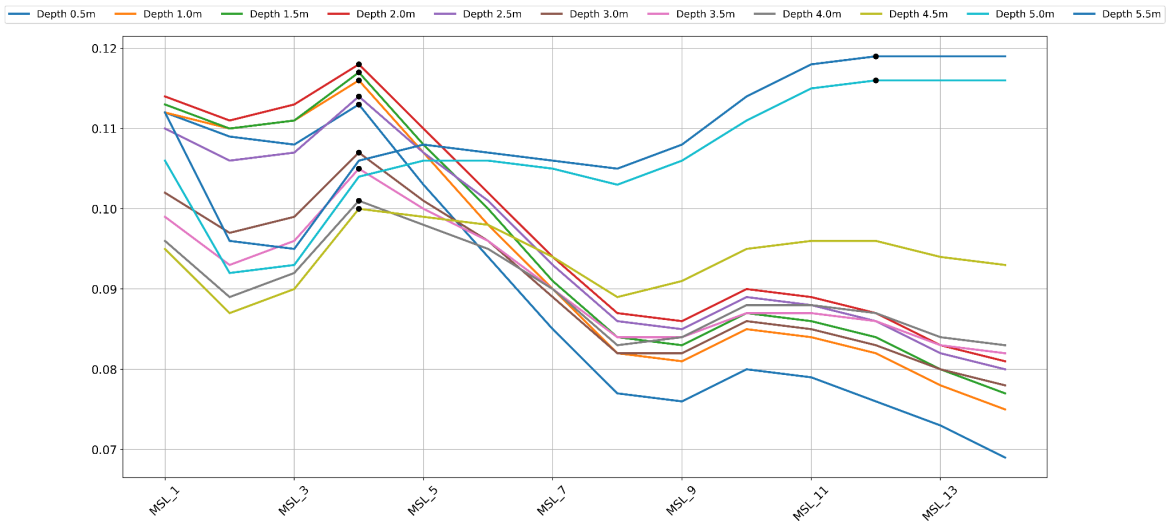


Figure 46. Correlation of aggregated sea level with Salinity levels at Station 180

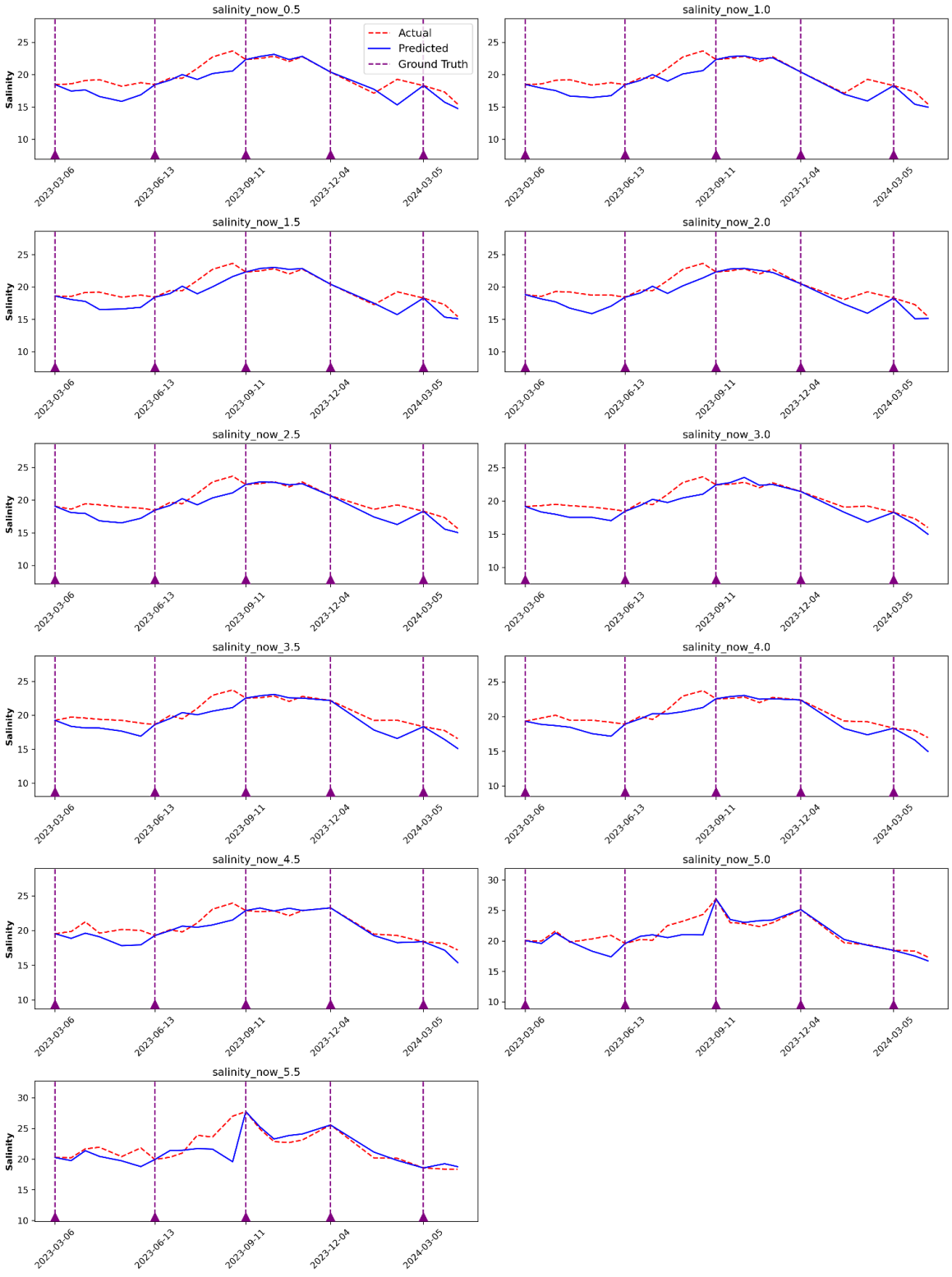


Figure 47 . Comparison of Actual, Predicted, and Ground Truth Salinity Across Depths at Station 180

REFERENCES

[1] Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River water salinity prediction using hybrid machine learning models. *Water*, *12*(10), 2951.

[2] Gachloo, M., Liu, Q., Song, Y., Wang, G., Zhang, S., & Hall, N. (2024). Using Machine Learning Models for Short-Term Prediction of Dissolved Oxygen in a Microtidal Estuary. *Water*, *16*(14), 1998.

[3] Guillou, N., Chapalain, G., & Petton, S. (2023). Predicting sea surface salinity in a tidal estuary with machine learning. *Oceanologia*, *65*(2), 318-332.

[4] Borovskaya, R., Krivoguz, D., Chernyi, S., Kozhurin, E., Khorosheltseva, V., & Zinchenko, E. (2022). Surface water salinity evaluation and identification for using remote sensing data and machine learning approach. *Journal of Marine Science and Engineering*, *10*(2), 257.

[5] Hu, J., Liu, B., & Peng, S. (2019). Forecasting salinity time series using RF and ELM approaches coupled with decomposition techniques. *Stochastic Environmental Research and Risk Assessment*, *33*, 1117-1135.

[6] Tiwari, M. K., & Adamowski, J. F. (2015). Medium-term urban water demand forecasting with limited data using an ensemble wavelet–bootstrap machine-learning approach. *Journal of Water Resources Planning and Management*, *141*(2), 04014053.

[7] Tiwari, M. K., & Adamowski, J. F. (2017). An ensemble wavelet bootstrap machine learning approach to water demand forecasting: a case study in the city of Calgary, Canada. *Urban Water Journal*, *14*(2), 185-201.

[8] Tran, D. A., Tsujimura, M., Ha, N. T., Van Binh, D., Dang, T. D., Doan, Q. V., ... & Pham, T. D. (2021). Evaluating the predictive power of different machine learning algorithms for groundwater salinity prediction of multi-layer coastal aquifers in the Mekong Delta, Vietnam. *Ecological Indicators*, *127*, 107790.

[9] https://paerllab.web.unc.edu/modmon/

[10] station CLKN7; https://www.ndbc.noaa.gov

[11] Pawlowicz, R. (2013). Key physical variables in the ocean: temperature, salinity, and density. Nature Education Knowledge, 4(4), 13.

[12] Gibson, R. N., Barnes, M., & Atkinson, R. J. A. (2002). Impact of changes in flow of freshwater on estuarine and open coastal habitats and the associated organisms. Oceanography and marine biology: an annual review, 40, 233.

[13] Lin, J., Liu, Q., Song, Y., Liu, J., Yin, Y., & Hall, N. S. (2023). Temporal prediction of coastal water quality based on environmental factors with machine learning. Journal of Marine Science and Engineering, 11(8), 1608.

[14] Burkholder, J., Eggleston, D., Glasgow, H., Brownie, C., Reed, R., Janowitz, G., ... & Springer, J. (2004). Comparative impacts of two major hurricane seasons on the Neuse River and western Pamlico Sound ecosystems. Proceedings of the National Academy of Sciences, 101(25), 9291-9296.

[15] Robbins, J. C., & Bales, J. (1995). Simulation of hydrodynamics and solute transport in the Neuse River Estuary (No. 94-511). US Geological Survey; Earth Science Information Center, Open-File Reports Section.

[16] https://en.wikipedia.org/wiki/Estuarine_water_circulation?utm_source=chatgpt.com

[17] MacCready, P. (1999). Estuarine adjustment to changes in river flow and tidal mixing. Journal of Physical Oceanography, 29(4), 708-726.

[18] https://paerllab.web.unc.edu/2020/10/14/neuse-river-estuary-conditions/

[19] Wool, T. A., Davie, S. R., & Rodriguez, H. N. (2003). Development of three-dimensional hydrodynamic and water quality models to support total maximum daily load decision process for the Neuse River Estuary, North Carolina. Journal of Water Resources Planning and Management, 129(4), 295-306.

[20] Yan, X., Zhang, T., Du, W., Meng, Q., Xu, X., & Zhao, X. (2024). A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years. Journal of Marine Science and

[21] Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. Journal of Hydrology, 578, 124084.

[22] El Bilali, A., & Taleb, A. (2020). Prediction of irrigation water quality parameters using machine learning models in a semi-arid environment. Journal of the Saudi Society of Agricultural Sciences, 19(7), 439-451.

[23] Tran, T. T., Pham, N. H., Pham, Q. B., Pham, T. L., Ngo, X. Q., Nguyen, D. L., ... & Veettil, B. K. (2022). Performances of different machine learning algorithms for predicting saltwater intrusion in the vietnamese mekong delta using limited input data: a study from Ham Luong River. Water Resources, 49(3), 391-401.

[24] Guillou, N., Chapalain, G., & Petton, S. (2023). Predicting sea surface salinity in a tidal estuary with machine learning. Oceanologia, 65(2), 318-332.

[25] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[26] Hidayat, F., & Astsauri, T. M. S. (2022). Applied random forest for parameter sensitivity of low salinity water Injection (LSWI) implementation on carbonate reservoir. Alexandria Engineering Journal, 61(3), 2408-2417.

[27] Khan, M. A., Shah, M. I., Javed, M. F., Khan, M. I., Rasheed, S., El-Shorbagy, M. A., ... & Malik, M. Y. (2022). Application of random forest for modelling of surface water salinity. Ain Shams Engineering Journal, 13(4), 101635.

[28] Xu, J., Xu, Z., Kuang, J., Lin, C., Xiao, L., Huang, X., & Zhang, Y. (2021). An alternative to laboratory testing: Random forest-based water quality prediction framework for inland and nearshore water bodies. *Water*, *13*(22), 3262.

[29] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*, 1937-1967.

[30] Borchani, H., Varando, G., Bielza, C., & Larranaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(5), 216-233.

[31] Yates, S. R., Zhang, R., Shouse, P. J., & Van Genuchten, M. T. (1993). Use of geostatistics in the description of salt-affected lands. In Water Flow and Solute Transport in Soils: Developments and Applications In Memoriam Eshel Bresler (1930–1991) (pp. 283-304). Berlin, Heidelberg: Springer Berlin Heidelberg.

[32] Lesch, S. M., Rhoades, J. D., Lund, L. J., & Corwin, D. L. (1992). Mapping soil salinity using calibrated electromagnetic measurements. Soil Science Society of America Journal, 56(2), 540-548.

[33] Forkuor, G., Hounkpatin, O. K., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PloS one*, *12*(1), e0170478.

[34] Charulatha, G., Srinivasalu, S., Uma Maheswari, O., Venugopal, T., & Giridharan, L. (2017). Evaluation of ground water quality contaminants using linear regression and artificial neural network models. *Arabian Journal of Geosciences*, *10*, 1-9.

[35] Yildiz, S. A. Y. I. T. E. R., & Degirmenci, M. U. S. T. A. F. A. (2015). Estimation of oxygen exchange during treatment sludge composting through multiple regression and artificial neural networks (estimation of oxygen exchange during composting). *International Journal of Environmental Research*, *9*(4), 1173-1182.

[36] Seeboonruang, U. (2015). An application of time-lag regression technique for assessment of groundwater fluctuations in a regulated river basin: a case study in Northeastern Thailand. Environmental Earth Sciences, 73, 6511-6523.

[37] Chen, W., Liu, W., Huang, W., & Liu, H. (2016). Prediction of salinity variations in a tidal estuary using artificial neural network and three-dimensional hydrodynamic models. *Computational Water, Energy, and Environmental Engineering*, *6*(1), 107-128.