CLUSTERING-BASED GLOBAL FORECASTING MODELS FOR SIGNIFICANT WAVE HEIGHT PREDICTION

by

Rohini Chandrala

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
August 2023

I dedicate this work in loving memory of my academic advisor Dr. Luis Torgo

Table of Contents

List of Tables

List of Figures

Abstract

Accurate wave predictions safeguard maritime operations, coastal communities, and marine ecosystems. Significant wave height, an average height of the highest one-third of the waves recorded during the sampling period, plays a crucial role in analyzing wave conditions and assessing coastal hazards among various wave fields. Forecasting significant wave height for various future timeframes, starting from 0.5 hours ahead, is vital for estimating coastal storm surges, issuing weather warnings, and preventing coastal disasters, especially during imminent large waves. Numerical methods are commonly used for wave forecasting; however, due to their computational intensity, they often require more time. In emergency situations, data-driven models offer faster wave predictions while maintaining accuracy, for shorter timeframes into the future. Data-driven forecasting models often treat data reported by buoys individually and forecast significant wave height based on the historical data of the respective buoy. Models trained on data from multiple buoys might leverage combined insights. However, training a single model on all different buoys may reduce forecasting accuracy when the data is from buoys in different environments. This study proposes a two-step approach to improve significant wave height predictions on a set of Environment and Climate Change Canada (ECCC) buoy data. First, we cluster buoys with similar data, enabling the formation of clusters with similar environmental conditions. Second, we train a global forecasting model on each cluster and predict significant wave height for individual buoys. We evaluate our proposed approach for significant wave height forecasting using data collected by 28 ECCC buoys distributed across the Atlantic, Pacific, and Great Lakes regions of Canada. Our results demonstrate that the clustering-based forecasting models, which leverage the shared patterns and relationships among multiple related buoy data, show competitive performance compared to the data-driven models trained on individual buoy data or universal model trained on all buoy data, in extreme events where wave height exceeds 6 meters.

Acknowledgements

I express my sincere gratitude towards my academic advisor, Dr. Luis Torgo, who had been my pillar of support for my thesis. Without his guidance and assistance, this achievement would not have been possible. I am most deeply indebted to my fellow researcher, Dr. Vitor Cerqueira, who has been a constant source of support and encouragement. Starting from brainstorming ideas to finalizing the research topic and completing the thesis, he has patiently helped me at every step, providing invaluable support, guidance, and feedback. I also appreciate his patience in answering the countless questions I asked.

I am sincerely grateful to Dr. Micheal McAllister for his invaluable help, support, and guidance on the thesis document in the absence of Dr. Luis Torgo. I would also like to extend my heartfelt thanks to my teachers, Dr. Stan Matwin, Dr. Sageev Oore, and Dr. Fernando Paulovich, who have imparted their knowledge of Machine Learning with Big Data, Machine Learning & Deep Learning, and Visual Analytics in the most effective way. The insights and understanding that I gained from their lectures have helped me to approach my research in a more informed and thoughtful manner and indirectly contributed to its success.

I would like to express my gratitude to my parents for their constant support and unwavering love. I would also like to express my gratitude to my brother, who not only instilled in me the importance of higher education but also supported me through my ups and downs. He always believed in me more than I did in myself, and his unwavering support has been invaluable. I am grateful to my guru, Sitaram, for imparting valuable lessons that have helped me grow both personally and professionally. I will always be indebted to him.

Finally, I am grateful to all my friends, relatives, and former colleagues for their constant support and encouragement throughout my masters program. Constant cheering and motivation from everyone have helped me stay focused and achieve this milestone. Without the support of all these people, I would not have made it to this end.

Chapter 1

Introduction

Ocean waves play a fundamental role in shaping coastal environments, ecosystems, and a wide range of human activities (Young and Babanin [2020]). Accurate wave predictions are vital for safeguarding maritime operations, coastal communities, and marine ecosystems (Davidson-Arnott et al. [2019], Barange et al. [2010]). Among various wave fields, significant wave height holds particular importance. It represents the average height (in meters) of the highest one-third of waves during a sampling period, typically lasting 20 minutes or more, depending on the measurement principles of the measuring devices. This measurement is essential for analyzing wave conditions, marine ecosystem assessment (Holthuijsen [2007]), and evaluating coastal hazards (Tucker [1991]).

Forecasting significant wave height helps to estimate coastal storm surges and issue coastal weather warnings (Finkl and Makowski [2013]). Forecasting imminent large waves, specifically extreme significant wave heights, is crucial in protecting wave energy converters (Li et al. [2012]). Forecasting significant wave heights is thus a critical area of study, essential for assessing the risk or potential impact on aforementioned activities. Smart buoys, also called marine environmental monitoring systems, are one of the sources for collecting information on significant wave height. Environment and Climate Change Canada (ECCC) maintains several such buoys along coastal and marine regions of Canada, effectively monitoring and collecting data related to wave conditions, climate, and environmental factors.

Numerical methods are the most commonly used approach for forecasting waves. Researchers created initial wave models in the late 1960s and early 1970s that employed numerical techniques to mimic wave behavior. These numerical models encapsulate wave propagation in mathematical differential equations that consider factors like wind patterns, currents, depths, and coastal features. Environmental and Climate

Change Canada(ECCC), relies on third-generation model numerical model, WAVEWATCH III (NOAA National Centers for Environmental Prediction [2005]). Typically, these numerical models require a high-performance computing infrastructure to solve the equations, coupled with long time periods (Yoon et al. [2011]).

In case of emergencies arising in the ocean, where rapid wave height predictions are crucial, faster and dependable forecasting techniques become essential. The challenges and limitations associated with traditional numerical methods used for predicting waves lead to the idea of using machine learning methods that require less computational resources while still being able to provide accurate predictions for predicting sea wave behavior in the short term. This new approach is considered interesting because it has the potential to solve the computational difficulties while still achieving accurate predictions about wave behavior for shorter lead-times (as discussed by Wang et al. [2018], Berbić et al. [2017]).

Various data-driven models, including statistical, machine learning, and hybrid approaches are often used for forecasting significant wave height. For instance, Emmanouil et al. [2020] used Bayesian Networks to forecast significant wave heights in Liverpool Bay, situated in the eastern part of the Irish Sea. Deka and Prahlada [2012] utilized Artificial Neural Networks with wavelet transformation to forecast significant wave height near Mangalore, India, up to 48 hours ahead. Ali et al. [2020b] employed a multiple linear regression (MLR) model optimized with the covariance-weighted least squares (CWLS) estimation algorithm to predict near real-time significant wave height using climate and oceanic inputs.

Existing statistical, machine learning, or hybrid models used for forecasting significant wave heights rely on data from individual buoys during training. These models can only predict based on the historical data collected by that specific buoy. Training a forecasting model with data from multiple buoys would enable the forecasting model to learn how significant wave height responds to different environmental conditions or weather events across various regions, leading to better predictions at multiple buoy locations. Also, during extreme events in coastal regions, certain buoys may have observed unique conditions that other buoys have not yet encountered. A data-driven model trained on the data with such extreme events could improve predictions for unseen conditions that other buoys have not yet encountered.

Across various domains, such as sales forecasting, researchers are exploring the development of forecasting models that leverage the collective information from multiple time series. However, this approach has not yet been extended to the specialized field of significant wave height forecasting. In the context of sales forecasting, Trapero et al. [2014] employed a pooled regression model by aggregating related time series, resulting in a 30 percent reduction in observed forecast error compared to forecasts provided by human experts for reliable promotional predictions in the absence of historical sales data. Hartmann et al. [2015] propose a cross-sectional regression model for sets of related time series, aiming to address missing values and rapidly attain accurate forecasting results at diverse aggregation levels, enhancing the model forecasting performance. In recent times, following the M4 (Semenoglou et al. [2021]) and M5 (Makridakis et al. [2022b]) competitions, there has been a growing emphasis on leveraging cross-series information to enhance forecasting accuracy. This trend reflects the recognition of the potential benefits of considering multiple time series simultaneously in forecasting tasks. However, training forecasting models across disparate time series may reduce overall accuracy (Bandara et al. [2020]).

Adapting the idea of leveraging cross-series information in the context of significant wave height forecasting, we hypothesize that within the ECCC buoy dataset collection, grouping the buoys with similar data, and training a forecasting model on each group, will lead to improved significant wave height predictions for each buoy compared to the predictions generated by forecasting models trained on the data from individual buoy.

We believe training the forecasting model on related buoy data can exploit collective information from multiple buoy data within each group and thereby enhance the accuracy of significant wave height forecasts for regular daily data and during extreme events where significant wave heights exceed 6 meters. These enhancements in significant wave height forecasts are valuable for making maritime activities safer, improving offshore operations, and reducing the potential damage from unexpected and extreme waves. Reliable significant wave height forecasts provide valuable information for navigation, shipping, coastal management, and disaster preparedness, contributing to safer and more efficient activities in marine environments.

1.1 Research Objective

Our primary research objective is to enhance significant wave height forecasting for individual buoy data of the ECCC dataset collection. By leveraging historical data from related buoys, we seek to provide improved forecasts in coastal and marine regions during regular and extreme events with significant wave heights exceeding 6 meters.

1.2 Research Methodology and Findings

To achieve our research objective, we propose a two-step approach. First, we identify similar buoy data through clustering techniques. Second, we train a global forecasting model on each cluster and use the corresponding clustering-based model to predict significant wave height for each buoy.

Each buoy records data over time as a sequence of data points, with each data point containing values of multiple fields measured or reported by the buoy. So each buoy data can be considered a multivariate time series (Tsay [2013]), leading us to frame the clustering task as a multivariate time series clustering problem. However, traditional clustering algorithms, such as centroid-based, hierarchical-based, and density-based algorithms, are not well-suited for multivariate time series for various reasons, including high-dimensionality and irregular or unequal lengths (Liao [2005]). For this reason, we use a feature-based clustering approach where we extract temporal features from each buoy data and apply traditional clustering algorithms to the extracted features.

Various clustering algorithms, such as K-means, Affinity, DBSCAN, OPTICS, and Agglomerative, are tested on the extracted temporal features to select the most suitable algorithm for grouping buoy data. After clustering the buoy data, we train a forecasting model on each cluster using the LightGBM algorithm. This forecasting model considers the 24 most recent measurements of significant wave height and provides forecasts for forecasting horizons at 1hr, 6hr, and 12hr for each buoy during regular environmental conditions.

Predicting imminent large waves is useful for coastal disaster prevention and protecting wave energy converters (Li et al. [2012]). While forecasting significant wave

height for extended lead times during extreme weather events, such as severe storms or hurricanes, is crucial for making decisions related to evacuations, resource allocation, and risk mitigation strategies, shorter-term forecasts for significant wave height, such as 1-hour predictions, also hold significance in specific contexts. For example, these 1-hour forecasts help in optimizing ocean wave energy converters. By enabling operators to make real-time adjustments to converter systems, they maximize energy extraction from strong waves while prioritizing safety. Moreover, these forecasts support preventive maintenance efforts, allowing operators to proactively safeguard converters from damage during extreme wave conditions. Additionally, they contribute to personnel safety by assisting operators in anticipating hazardous wave intensities and implementing essential safety measures.

Hence, we evaluate the forecasting performance of the clustering-based forecasting models in extreme events where the significant wave heights exceed 6 meters, considering a forecasting horizon of 1 hour. This evaluation helps assess the ability of clustering-based forecasting models to accurately predict significant wave heights during severe weather events such as storms or hurricanes. We conducted all the experiments using data from 28 active ECCC buoys across Canada.

In our study, we employ silhouette score (Rousseeuw [1987]) to evaluate the effectiveness of the formed clusters. The silhouette score evaluates clustering quality by measuring how well data points group within each cluster and how distinct the clusters are from each other. Our goal is to use plausible clusters of related buoys for creating models. We leverage the silhouette score as a metric to assess the effectiveness of clustering algorithms. We are particularly interested in observing whether the clustering algorithm yielding the highest silhouette score corresponds to improved forecasting results.

Common statistical metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) are widely used in research to assess the performance of data-driven models for predicting significant wave heights. We use these metrics to compare the output of trained models with the actual target values. To evaluate our clustering-based forecasting models, we focus on MAE, which treats all errors equally. MAE treats both overestimations and underestimations equally during the evaluation process.

We compare the MAE values of clustering-based forecasting models with the MAE values of models trained on each buoy data (referred to as local models) and with the MAE values of a model trained using data from all buoys (referred to as the universal model). By comparing MAEs across the ECCC buoy dataset, we can assess how well our approach performs for each buoy. To assess whether the observed differences in improvements resulting from the comparison between clustering-based models, local models, and a universal model are statistically significant or are due to random chance, we employ a statistical test called the Wilcoxon signed-rank test(Scheff [2016]).

In our experiments, K-means yielded the highest silhouette score of 0.417 among the tested clustering algorithms, indicating better cluster separation and coherence than other clustering algorithms. However, the forecasting results based on K-means clustering did not consistently exhibit the lowest MAE compared to other clustering approaches such as Affinity, DBSCAN, OPTICS, or Agglomerative. This indicates that our hypothesis, associating better silhouette scores with enhanced forecasting results, is not supported.

Examining the forecasting results across 28 buoys, we find that under regular conditions, clustering-based models consistently reported equal or lower MAE (with an average MAE of 0.11 meters) in comparison to local models for 1-hour forecasts, with a single exception. In the case of 6-hour and 12-hour forecasts, performance varied depending on the clustering algorithm used. For 6-hour forecasts, the clustering-based forecasting models exhibited equal MAE for 12 buoys, lower MAE for 11 buoys, and higher MAE for 5 buoys compared to local models. For 12-hour forecasts, the clustering-based forecasting models exhibited equal MAE for 10 buoys, lower MAE for 11 buoys, and higher MAE for 7 buoys compared to local models. In the context of 1-hour forecasts during extreme events, out of 17 buoys, the clustering-based forecasting models exhibited equal MAE for 6 buoys, lower MAE for 10 buoys, and higher MAE for 1 buoy compared to local models.

Comparing the results of clustering-based forecasting models and the universal model across 28 buoys, we find that clustering-based models exhibited either equal or lower MAE (for 4 buoys) compared to universal models for 1-hour forecasts. For 6-hour forecasts, the clustering-based forecasting models exhibited equal MAE for 17 buoys, lower MAE for 9 buoys, and higher MAE for 1 buoy compared to the universal

model. For 12-hour forecasts, the clustering-based forecasting models exhibited equal MAE for 10 buoys, lower MAE for 13 buoys, and higher MAE for 5 buoys compared to the universal model. In the context of 1-hour forecasts during extreme events, out of 17 buoys, the clustering-based forecasting models exhibited equal MAE for 4 buoys, lower MAE for 11 buoys, and higher MAE for 2 buoys compared to universal model.

In summary, our hypothesis is valid during extreme events when significant wave heights exceed 6 meters. For the 1-hour forecasting horizon, clustering-based models exhibited significant performance improvements over both local models and the universal model by an average of 11 centimeters and 2 centimeters respectively, and with p-values of 0.02 and 0.01. Across 1-hour, 6-hour, and 12-hour forecasts under regular conditions, clustering-based models reported lower MAE, achieving an average improvement of 2 centimeters compared to the universal model across 28 buoys. The corresponding p-value was 0.02. Compared to local models, the p-values are above 0.05, indicating that there is no significant statistical difference between the clustering-based models and the local models in terms of forecasting significant wave heights.

1.3 Outline

The rest of the thesis is structured as follows. Chapter 2 reviews the related work on significant wave height forecasting models, global or cross-learning models, and time-series clustering, including literature and architectures of models used in our study. It also discusses feature extraction techniques, clustering algorithms, algorithmic frameworks, and architectures. Chapter 3 defines the problem and discusses the limitations of current data-driven forecasting models trained on individual buoy data. Additionally, we describe the problem of forecasting significant wave height from a time series perspective and provide information on the data source and details about the buoy data collection.

Furthermore, Chapter 4 presents our research methodology, outlining the approach and providing the workflow of our proposed clustering-based forecasting models. We also explain the evaluation process used to verify the effectiveness of the clustering-based approach. Chapter 5 details the exploratory data analysis conducted on the

ECCC buoy dataset collection to understand the relation between buoy data fields and feasibility assessment for clustering. The experiments conducted and their results, along with an analysis of the results, are presented in Chapter 6. Finally, Chapter 7 concludes the thesis by providing key findings, limitations, and recommendations for future research.

Chapter 2

Related Work

In this chapter, we will explore the evolution of significant wave height forecasting models, beginning with traditional numerical models and progressing to the latest advancements in the field. We will also discuss the concept of global models for time series forecasting and time series clustering techniques.

Section 2.1 provides an overview of existing numerical models used for predicting significant wave height, followed by a discussion of data-driven models, including statistical and machine learning approaches. In Section 2.2, we delve into recent research on the development of global models for time series forecasting, which leverages information from multiple related time series. We also present the architecture of the global forecasting model used in our study. Lastly, in Section 2.3, we review the techniques employed in the clustering of time series data and present the algorithmic architecture of the clustering algorithms used in our study

2.1 Significant Wave Height Forecasting : Literature Review

2.1.1 Background on Numerical Models

Wave forecasting numerical models are computer-based simulations used to predict the behavior of ocean waves, including the significant wave height. These models are designed to mimic the complex dynamics of waves by using mathematical equations that consider various factors such as wind patterns, ocean currents, water depths, and coastal features (Komen et al. [1964]). The concept of wave prediction was first introduced in the 1960s and 1970s by Komen et al. [1964]. They developed the initial wave models by representing physical processes through mathematical relations that approximate the underlying physical laws. However, these early models had limitations as they did not fully calculate the wave spectrum from the energy balance equation (SWAMP [1985]). Consequently, they overestimated the influence of wind

and neglected nonlinear transfer in the energy balance equation. Due to these shortcomings, the first-generation wave models had challenges in accurately representing the development of waves.

After conducting extensive wave growth experiments, researchers identified the relative significance of nonlinear transfer and wind input. Subsequently, Hasselmann et al. [1985] worked on the development of second-generation wave models over a period of seven years. According to SWAMP [1985], second-generation wave models encountered challenges in accurately representing wave characteristics. Specifically, they struggled to properly simulate complex wave fields generated by rapidly changing winds, such as those seen during hurricanes, small-scale cyclones, or fronts. This limitation hindered their ability to provide accurate forecasts in such dynamic and intense weather conditions. To address the shortcomings of both first and second-generation wave models, the Sea Wave Modelling Project (SWAMP) was initiated. The SWAMP project aimed to compare and evaluate ten different wave prediction models in-depth, including the second-generation models. The project extensively discussed the limitations observed in these early models and aimed to drive improvements in wave forecasting techniques.

Following the limitations identified in first and second-generation wave models, Tolman [1999] developed a third-generation model, WAVEWATCH III, with WAve Model (WAM) as the baseline, to address the shortcomings and improve the accuracy of wave forecasting. The WAVEWATCH III model computes the wave spectrum by integrating the energy balance equation without any preconceived restrictions on the spectral shape. Over time, the model has been adopted by various institutions worldwide, with slight variations in its implementation. The latest version of WAVEWATCH III, 6.07 (Tolman et al. [2019]), includes the latest scientific advancements making it more accurate and capable of providing better predictions and simulations of wave behavior in various oceanic and coastal conditions. Additionally, Booij et al. [1999] developed another third-generation numerical wave model known as Simulating Waves Nearshore (SWAN) as an alternative to WAVEWATCH III (WAM) for specific applications. SWAN was specifically designed to compute random short-crested waves in coastal regions with shallow water and ambient currents. It addresses some limitations of the WAM model in scenarios where the water depth is less than 20-30 meters,

making it more suitable for simulating wave behavior in nearshore environments with complex coastal features and varying water depths.

2.1.2 Data-Driven models

One of the alternative approaches for the aforementioned physically-based models is data-driven models. As highlighted in the work by Berbić et al. [2017], numerical models excel in estimating wave characteristics across an entire geographical region, while the data-driven approach is typically applied to specific locations, particularly those equipped with buoys. Numerical models rely on measurements from particular sites to validate their results, whereas data-driven models, including machine learning methods, necessitate measured data that includes input variables like wave height, period, wind velocities, fetch, air pressures, and temperatures along with desired output like wave height or wave period. In scenarios with time limitations, data-driven models offer efficiency advantages, making them suitable for swift short-term predictions of sea waves. Researchers have employed various data-driven models, encompassing statistical and machine learning approaches, to forecast significant wave height. We delve into these models in the subsequent sections.

Statistical Models

Traditional forecasting models, like ETS, ARIMA, or Theta, have been widely in time series forecasting (Hyndman and Khandakar [2008]). However, statistical models like Seasonal Autoregressive Integrated Moving Average (SARIMA) (Box et al. [1994]) were used for forecasting significant wave height. Yang et al. [2019] have used SARIMA models to predict long-term wave height in specific regions, such as the South China Sea and Adjacent Waters, using third-generation wave model WAVEWATCH-III simulated data and observed Root Mean Squared Error (RMSE) of 0.339m for 12-step prediction.

SARIMA models are built on the assumption that time series data can be divided into trend, seasonality, and random components (Hyndman and Athanasopoulos [2018]). The trend component represents long-term behavior, the seasonality component captures periodic patterns, and the random component accounts for short-term variability or noise. By separating these components, the SARIMA model can

effectively model the relationships between variables. However, it is essential to note that SARIMA models handle data with regular and predictable seasonal patterns ((Wang et al. [2021])), while significant wave height in oceans often exhibits irregular and unpredictable fluctuations (Woolf et al. [2002]).

Machine Learning Models

Machine learning models are alternatives to traditional statistical models and they possess the ability to capture relationships, trends, or structures within the data. They have the capabilities to account for temporal dependencies and handle variable-length sequences, which is useful in modeling time series data (Längkvist et al. [2014]).

In the context of significant wave height forecasting, various machine learning techniques have been applied. For instance, James et al. [2018] trained a multi-layer perceptron model as a surrogate for the physics-based SWAN model to simulate the wave field in Monterey Bay. Their study revealed that the multi-layer perceptron model exhibited superior performance compared to the physics-based SWAN model in terms of Root Mean Squared Error (RMSE), resulting in up to 80 percent reduction in errors. Similarly, Berbić et al. [2017] used Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) to predict significant wave height at two different locations in the Adriatic Sea. They compared the results with the predictions of numerical models for an 11-step forecast. Their study found that SVM performed better overall than the neural networks and numerical models, achieving an average MAE of 0.137 meters.

In another study, Nikoo and Kerachian [2017] developed an Artificial Immune Recognition System (AIRS) for predicting significant wave height with different time lags in Lake Superior, North America. They compared the results with five other models, including artificial neural networks, support vector regression, bayesian networks, and rough set theory. The results showed that both the AIRS and artificial neural network models outperformed the other data-driven models. AIRS performed exceptionally well in predicting significant wave heights specifically during extreme weather events, with a Root Mean Squared Error (RMSE) of 0.139 meters.

Similarly, Shamshirband et al. [2020] used three models, namely artificial neural networks (ANNs), extreme learning machines (ELM), and support vector regression

(SVR), to predict wave heights at Bushehr and Assaluye ports. Comparing the results of the different machine learning-based models indicated that ANNs, ELM, and SVR models provided similar predictions for both stations. However, the ELM model slightly outperformed the others, achieving a Mean Absolute Error (MAE) of 0.21 meters.

Deep learning models, such as Recurrent Neural Networks (RNNs) and Long-Short Term Memory Networks (LSTMs), have proven to be effective in capturing non-linear and hierarchical dependencies within data. For instance, Sadeghifar et al. [2017] employed Nonlinear Autoregressive eXogenous inputs (NARX) with RNNs to predict coastal wave height in the South Caspian Sea, achieving an RMSE of 0.38 meters for a 12-step prediction. Similarly, Fan et al. [2020] utilized LSTMs for significant wave height prediction, covering a forecasting range from 1 hour to 3 days. Additionally, Song et al. [2022] used the deep learning method, Convolutional Long-Short Term Memory Networks(ConvLSTM) with a masking technique to predict significant wave height across the entire Beibu Gulf, showing promising results compared to other ConvLSTM variants with different inputs.

Researchers have also explored variants of machine learning models and hybrid approaches, combining various data-driven models to forecast significant wave height at different locations (Ali et al. [2021], Londhe et al. [2016], Ali et al. [2020a], Ali et al. [2020b], Dixit et al. [2015], Dogan et al. [2021]). However, a common limitation of these studies is their focus on training and forecasting for a single location or specific time series data. The potential of leveraging cross-series information across various locations and related buoy datasets remains untapped in these studies.

Model Evaluation and Comparison in Significant Wave Height Forecasting Studies

In the aforementioned studies, statistical metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2) were used to evaluate machine learning models for significant wave height forecasting with varying inputs and forecasting horizons. These metrics provide valuable insights into the performance of the models, with lower values indicating higher accuracy and better agreement between the prediction of the models and observed data. The studies compared their

proposed model metrics with those of various other model metrics to determine which one exhibits superior performance.

However, a notable observation in these studies is that none of them explicitly defines how much difference in MAEs is considered substantial enough to prefer one model over another. For instance, it is demonstrated that SVM improves accuracy, for some time steps, in the study by Berbić et al. [2017], where SVM reported an MAE of 0.137 meters and ANNs reported an MAE of 0.142 meters. Shamshirband et al. [2020] considered ELM as a better model with an MAE of 0.21 meters than ANNs with an MAE of 0.22 meters. Despite this lack of a clear benchmark, these studies consistently prioritized the reduction of prediction errors, with low MAE or RMSE being a key factor in evaluating model performance.

2.2 Global or Cross-Learning Models

2.2.1 Literature review

Although cross-series information across various locations remains untapped in wave height forecasting, it has been actively explored in other domains of time series forecasting, particularly after the M4 competition. The main principle behind this approach is to develop global models (Januschowski et al. [2020]) that leverage information from multiple time series simultaneously, rather than creating separate models for each series (referred to as local models in our study).

A few studies have incorporated cross-series learning into their forecasting models using deep neural networks. For instance, the winning solution of the M4 competition (Smyl [2020]) approach combining exponential smoothing methods, machine learning techniques, and hierarchical forecasting to achieve the best performance over a wide range of domains, including sales, finance, inventory, and more. Similarly, Salinas et al. [2020] used DeepAR to produce accurate probabilistic forecasts, based on training an auto-regressive recurrent network model on a large number of related time series, specifically focusing on electricity and traffic data. Bandara et al. [2020] employed a clustering-based approach for forecasting across time series databases using recurrent neural networks. The results showed that their approach outperformed the winning model of the CIF2016 forecasting competition (Štěpnička and Burda [2017])

in terms of forecasting accuracy.

More recently, following the M5 competition (Makridakis et al. [2022b]), LightGBM (Ke et al. [2017]), a decision tree-based machine learning approach, has gained significant recognition as a superior cross-learning forecasting model. The M5 competition focused on accurate predictions for 42,840 hierarchical time series representing Walmart sales. Participants were required to submit 30,490 point forecasts for various levels of aggregation. According to Makridakis et al. [2022a], LightGBM was employed by almost all of the top 50 competitors, and the top 5 winning teams achieved accuracy improvements greater than 20 percent compared to previous competitions. This success highlights the effectiveness of LightGBM in handling multiple related series, making it an excellent choice as a global forecasting model.

2.2.2 LightGBM for global forecasting models

The success of LightGBM in various forecasting problems across different domains highlights its robustness and ability to handle complex data interactions and missing values. For instance, researchers in Deng et al. [2021] achieved significant forecasting improvements by using LightGBM to predict daily sales for an online retail platform, surpassing the performance of LSTM and XGBoost models. Likewise, in the domain of energy consumption, Di Persio and Fraccarolo [2023] demonstrated the superiority of LightGBM in accurately forecasting hourly electricity consumption in commercial buildings compared to models like ARIMA and random forest. Additionally, LightGBM has exhibited its effectiveness in cryptocurrency forecasting, as evidenced by Sun et al. [2020], where it outperformed LSTM and ARIMA models in predicting the prices of three cryptocurrencies. Given the extensive evidence of the superior performance of LightGBM in diverse forecasting scenarios, we have selected it as the forecasting model for our study.

LightGBM: LightGBM (Light Gradient Boosting Machine) is an open-source gradient boosting framework that uses tree-based learning algorithms. It is developed by Ke et al. [2017] and it is designed to be efficient in terms of memory usage and training speed while maintaining high accuracy. This algorithm is based on the Gradient Boosting Decision Tree (GBDT) algorithm. However, it differs from other

GBDT algorithms in several ways, such as its handling of categorical features, the use of the histogram-based approach for binning, and the way it handles missing data. LightGBM also employs a novel algorithm called Gradient-based One-Side Sampling (GOSS) that reduces the training time and memory usage by sampling the data instances based on their gradients.

The following is a brief overview of LightGBM

1. Decision Trees: LightGBM builds decision trees that partition the data into smaller subsets based on a set of splitting criteria. Each tree consists of nodes and leaves, where the nodes are the splitting points and the leaves are the terminal points that contain the predicted output.

2. Gradient Boosting: This algorithm uses a gradient boosting approach to combine multiple decision trees into an ensemble. The idea is to add new trees to the ensemble that improve the accuracy of the current prediction. The gradient descent algorithm is used to optimize the loss function.

3. Objective Function: The objective function in LightGBM is a combination of the loss function and a regularization term. It is defined as follows:

$$\text{Obj}(\phi) = \sum_{i=1}^{n} l(y_i, f(x_i; \phi)) + \sum_{k=1}^{K} \Omega(z_k)$$

   where ϕ is the model parameters, l is the loss function, f is the prediction function, x_i is the input data, y_i is the target output, K is the number of trees in the ensemble, and $\Omega(z_k)$ is the regularization term.

4. Splitting Criteria: This algorithm uses a histogram-based approach to select the best-split points for each node in the decision tree. The histogram-based approach groups the continuous features into discrete bins and then selects the best-split point based on the distribution of data within each bin.

5. Leaf-wise Tree Growth: This algorithm uses a leaf-wise tree growth algorithm, where each new split is made on the leaf that will result in the largest reduction in the loss function. This approach lead to faster training times and better accuracy compared to other algorithms that use level-wise tree growth.

6. Categorical Features: It has built-in support for categorical features, which are usually represented as integer values. The algorithm can automatically handle categorical features by splitting them based on the values of the integers.

7. GPU Acceleration: It supports GPU acceleration for training and prediction, which can significantly speed up the process and handle larger datasets.

2.3 Time Series Clustering

Time-series data are dynamic, with feature values changing over time. Clustering time series data presents unique challenges, such as selecting appropriate similarity measures, handling varying lengths of samples, and representing time series with suitable features (Liao [2005]). Various algorithms have been developed for clustering different types of time series data, depending on the type of application. The survey by Liao [2005], described three main approaches to whole time-series clustering: a raw-data-based approach where clustering is directly based on the distance calculated on the raw data points, a feature-based approach where features are extracted from the raw data and clustering algorithms are applied on the extracted feature vectors, and model-based approaches where model parameters are extracted from the raw data and then clustering algorithms are applied on the extracted model parameters.

In the raw-data-based clustering approach, performance is greatly influenced by the distance metric used. Aghabozorgi et al. [2015] discuss distance measures for time series clustering and highlight the challenges of identifying a suitable distance metric for raw time series data, especially when the data has noise, different lengths, and different dynamics. Similarly, Fraley and Raftery [2002] highlights that model-based approaches can be sensitive to the choice of initial starting values, particularly for complex models with many parameters. In contrast, feature-based clustering techniques do not rely on a distance metric to capture the similarity of point values and instead use sets of global features obtained from a time series to summarize and describe the salient information of the time series (Fulcher [2018]). Therefore, feature-based approaches do not suffer from the challenges of identifying a suitable distance metric for raw time series data or the sensitivity to the choice of initial starting values that model-based approaches face.

Feature-based clustering involves extracting global features from data before applying clustering algorithms to group similar data points based on these extracted features. The global features capture complex temporal patterns produced by various underlying mechanisms on different timescales and represent them as low-dimensional vectors, providing valuable insights into the generative processes behind the time series (Fulcher [2018]). Given these advantages, we choose to focus on feature extraction techniques followed by traditional clustering for our use case.

2.3.1 Feature Extraction Techniques: Literature

Time series feature extraction has a broad literature in various fields. This approach can be more interpretable and more resilient to missing and noisy data. At first, many researchers started extracting basic features like max, min, skewness, and generic patterns such as peaks but later on, researchers in various fields analyzed time series and explored specialized features. For instance, Mierswa and Morik [2005] extracted features related to peak sounds that are helpful to classify audio data, Yen and Lin [2000] extracted wavelet-based features to monitor vibrations, Fulcher and Jones [2014] collected more than 9000 features from 1000 different feature-generating algorithms that are discussed in fields such as medicine, astrophysics, finance, mathematics, climate science, industrial applications.

In the literature, several feature extraction packages are available to analyze time series data. Notable examples include FATS (Nun et al. [2015]), designed initially for astronomical light curve data but applicable to various applications, CESIUM (Naul et al. [2016]), which offers an end-to-end time series analysis framework with a Python library and web front-end interface, and HCTSA (Fulcher and Jones [2017]), enabling extensive feature extraction and comparison of over 7,700 features from interdisciplinary time-series analysis literature. Additionally, TSFRESH (Christ et al. [2018]) and TSFEL (Barandas et al. [2020]) are feature extraction packages that focus on statistical hypothesis tests and comprehensive analysis of temporal complexity. TSFRESH, in particular, provides an automatic configuration of statistical tests based on the machine learning problem and feature type, while TSFEL classifies features into temporal, statistical, and spectral domains, expanding its scope for in-depth temporal feature analysis. These packages offer a diverse set of tools for extracting

relevant features from time series data, catering to various applications and machine learning models.

The success of TSFEL in various forecasting problems across different domains highlights its effectiveness and versatility as a feature extraction algorithm. Its ability to handle big data and extract meaningful features from time series data has made it a valuable tool in diverse applications of time series analysis tasks (Kurian et al. [2021], Tlachac et al. [2021], Bhattacharyya et al. [2021]). Given the presence of sensor noise and missing readings, TSFELs robustness and efficiency in feature extraction in these cases with less computation time (Henderson and Fulcher [2021]) make it a suitable choice for our study.

2.3.2 Clustering Algorithms

Clustering is a fundamental technique in machine learning and data analysis used to group similar data points into distinct clusters based on their similarities. The primary goal of clustering is to identify inherent patterns, structures, or natural groupings in the data without the need for predefined labels or categories. The process involves assigning each data point to a cluster in such a way that points within a cluster are more similar to each other than to points in other clusters. Xu and Wunsch [2005] provides a survey on clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in various problems. Also, Berkhin [2006] gives an overview of different clustering methods. In this study, we apply conventional clustering algorithms to the extracted features to find the optimal groupings between the buoy data. We test different kinds of clustering algorithms, including K-means, Affinity, DBSCAN, OPTICS, and Agglomerative algorithms, to assess the robustness of the proposed framework.

Clustering Algorithmic Frameworks and Architectures

K-means: K-means was originally proposed by MacQueen [1967] and is a widely-used and straightforward clustering technique that is often employed to address clustering problems. The method involves partitioning the given dataset into k clusters, where k is determined by the user. The main idea behind K-means is to identify k groups of data, with each group represented by a centroid at the center of the data.

The objective function J is given as follows

$$Minimize\ J = \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

Here, k represents the number of clusters, C_i represents the i^{th} cluster, x represents the data points in the cluster, and μ_i represents the centroid of the i^{th} cluster. The objective function quantifies the sum of squared distances between data points and their assigned centroids. By minimizing the objective function, K-means aims to find the optimal positions for the cluster centroids such that the within-cluster variation is minimized. The algorithm iteratively updates the centroid positions until convergence, resulting in the final partitioning of the data into k clusters.

The procedure of the K-means algorithm is composed of the following steps

1. Initialization: Suppose we decide to form k-clusters for any given dataset. Now take k distinct random points. These points represent the initial group of centroids. As these centroids changes after each iteration before clusters are fixed, these can be chosen randomly.

2. Assign each object to the group that has the closest centroid.

3. When all objects have been assigned, recalculate the positions of the k centroids.

4. Repeat Steps 2 and 3 until the centroids no longer move within a reasonable threshold. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The only difference between K-means and K-medoids is that K-means uses the mean of all data points within each group to define the center of a cluster, and K-medoids use an actual data point in the cluster that is closest to all other points. From a performance perspective, we have conducted experiments with K-means.

Affinity Propagation: Affinity Propagation proposed by Frey and Dueck [2007] is a clustering algorithm that operates based on the concept of "message passing" between data points. Unlike other clustering algorithms, Affinity Propagation does not require a predefined number of clusters. Instead, it determines the number of

clusters based on the data itself. The algorithm starts by assigning each data point as its own exemplar. It then iteratively updates "responsibility" and "availability" values for each pair of data points. The responsibility value reflects how well-suited a point is to be the exemplar for another point, considering the similarity between their features. The availability value represents how well-suited a point is to be assigned to a cluster based on the exemplars of other points. During each iteration, the responsibility and availability values are updated based on messages exchanged between data points. The algorithm continues to iterate until convergence is reached. At convergence, the exemplars are determined based on the responsibility and availability values. Each data point is then assigned to the cluster represented by its corresponding exemplar.

Following steps outline the main procedure of the Affinity Propagation algorithm.

1. Similarity Matrix: Calculate the similarity between all pairs of data points using a similarity function. The similarity function, s, could be any function that measures the similarity or dissimilarity between data points. Let the similarity between data points i and j be represented by $s(i, j)$.

2. Responsibility: Initialize the responsibility matrix $r(i, j)$ to θ for all data points i and j. The responsibility matrix represents the amount of responsibility that point i assigns to point j to be its exemplar.

3. Availability: Initialize the availability matrix $a(i, j)$ to θ for all data points i and j. The availability matrix represents the amount of availability that point j has for being an exemplar.

4. Message Passing: For each iteration t, update the responsibility and availability matrices using the following equations:
   Responsibility Update: $r(i, j) = s(i, j) - \max_{k \neq j}(a(i, k) + s(i, k))$
   Availability Update: $a(i, j) = \min\left(0, r(j, j) + \sum_{k \neq i, k \neq j}(\max(0, r(k, j)))\right)$

5. Exemplars: After convergence, the exemplars are the data points that have the highest value of the sum of the responsibility and availability matrices:

$$e(i) = \text{argmax}_j(r(i, j) + a(i, j))$$

where $e(i)$ is the exemplar of data point i.

6. Clustering: Assign each data point to its corresponding exemplar to form clusters. The algorithm can be formulated mathematically using the following equations:
   $s(i, j)$ - the similarity between data points i and j
   $r(i, j)$ - the responsibility of point i to point j
   $a(i, j)$ - the availability of point j for being an exemplar
   $e(i)$ - the exemplar of data point i
   Initialization: $R(i, j) = 0, \quad A(i, j) = 0$
   Responsibility Update: $r(i, j)^{t+1} = s(i, j) - \max_{k \neq j}(a(i, k) + s(i, k))^t$
   Availability Update: $a(i, j)^{t+1} = \min 0, r(j, j)^t + \sum_{k \neq i, k \neq j} \max 0, r(k, j)^t$
   Exemplars: $e(i) = \text{argmax}_j(r(i, j) + a(i, j))$

After assigning data points to their corresponding exemplars, the algorithm forms clusters by grouping together data points with the same exemplar. Overall, the AP algorithm aims to find a set of exemplars that represent the input data set and to group similar data points together into clusters based on their similarity.

DBSCAN: The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm introduced by Ester et al. [1996], depends on a density-based notion of clusters. The algorithm groups together the points that are closely packed together and classify them as a cluster. It defines a neighborhood around each point and then looks for densely populated neighborhoods. Points that are not within these dense neighborhoods are classified as noise or outliers. The algorithm has two main parameters: epsilon (ϵ), which defines the radius of the neighborhood around each point, and minPts, which defines the minimum number of points required to form a dense neighborhood.

The procedure of the DBSCAN algorithm is composed of the following steps

1. Initialize the algorithm by selecting an arbitrary point from the dataset and checking its ϵ-neighborhood, which consists of all points that are within a distance ϵ from the selected point.

2. If the ϵ-neighborhood contains at least the minimum number of points (minPts), then the selected point is marked as a core point, and all points in its ϵ-neighborhood are added to its cluster.

3. Repeat the process for all points in the core points ϵ-neighborhood until there are no points to add added to the cluster.

4. If the ϵ-neighborhood does not contain enough points, the selected point is marked as a border point and added to a cluster if it is in the ϵ-neighborhood of a core point.

5. If the selected point is neither a core nor a border point, it is marked as a noise point and excluded from any cluster.

6. Steps 1 to 5 are repeated until all points have been assigned to a cluster or marked as noise.

In summary, DBSCAN starts with an arbitrary point and iteratively expands clusters by adding points that are close to core points. Border points are assigned to clusters, while noise points are discarded.

OPTICS: OPTICS (Ordering Points To Identify the Clustering Structure) algorithm proposed by Ankerst et al. [1999] is a powerful clustering method that can identify clusters of various shapes and sizes. It can handle noise and outliers effectively, and it can detect clusters of different densities and shapes. Unlike k-means or hierarchical clustering algorithms, OPTICS does not require the number of clusters to be specified beforehand. The algorithm utilizes two parameters: ϵ, that describes the maximum distance (radius) to consider, and MinPts, that describes the number of points required to form a cluster.

In OPTICS algorigthm a point p is a core point if at least MinPts points are found within its ϵ-neighborhood $H_\epsilon(p)$ (including point p itself). Each core point is assigned a core distance, which represents the distance to its MinPtsth closest point. This implies:

$$core - dist_{\epsilon,MinPts}(p) = \begin{cases} \text{undefined}, & \text{if } |H_\epsilon(p)| < \ MinPts \\ \min_{q \in N_\epsilon(p), q \neq p} d(p,q), & \text{otherwise.} \end{cases}$$

The reachability distance(r-dist) of another point o from a point p is either the distance between o and p, or the core distance of p, whichever is bigger. This implies:

$$r - dist_{\epsilon,MinPts}(o,p) = \begin{cases} \text{undefined}, & \text{if } |H_\epsilon(p)| < \ MinPts \\ \max(\text{core-dist}_{\epsilon,MinPts}(p), d(p,o)), & \text{otherwise.} \end{cases}$$

If p and o are nearest neighbors, this is the ϵ'$< \epsilon$ we need to assume to have p and o belong to the same cluster.

The procedure of the OPTICS algorithm is composed of the following steps

1. Calculate the pairwise distances between all points in the dataset.
2. Choose a distance threshold (ϵ) and a minimum number of points (minPts) for the clustering.
3. For each point, calculate its local density by counting the number of points within a distance of ϵ.
4. For each point, calculate its reachability distance, which is the maximum distance to a core point along a path of points with increasing density.
5. Order the points based on their reachability distance, with the lowest values first.
6. Traverse the ordered list of points and update the reachability distances of their neighbors as necessary.
7. Identify clusters by extracting the local maxima in the reachability distance plot.

In summary, the OPTICS algorithm offers a flexible and robust approach for identifying clusters of various shapes and sizes, providing a valuable tool for data analysis and pattern recognition.

Agglomerative: Agglomerative hierarchical clustering is a bottom-up clustering method where each observation begins in its own cluster, and pairs of clusters are merged based on their similarity. This process continues until all observations belong to a single cluster, resulting in a dendrogram that illustrates the hierarchical

relationships between the clusters. The algorithm can be divided into two phases: merging and cutting. In the merging phase, the two most similar clusters are iteratively merged, while in the cutting phase, the dendrogram is cut at a certain level to obtain the desired number of clusters. The procedure of Agglomerative hierarchical clustering can be summarized as follows:

1. Assign each data point to its own cluster.
2. Compute the proximity matrix that contains the distances between each pair of clusters.
3. Merge the two closest clusters into a single cluster.
4. Update the proximity matrix by computing the distances between the new cluster and each of the remaining clusters.
5. Repeat steps 3-4 until all data points are in a single cluster, or until the desired number of clusters is reached.
6. Construct a dendrogram to represent the hierarchical structure of the clustering.

Here, the proximity between two clusters can be measured using various distance metrics, such as Euclidean distance or cosine similarity. Different linkage criteria, such as single linkage (minimum distance), complete linkage (maximum distance), or average linkage (average distance), can be employed to determine the distance between two clusters. The choice of distance metric and linkage criteria can significantly impact the resulting clusters and dendrogram, providing different perspectives on the data grouping patterns and hierarchical structure.

Chapter 3

Problem Definition and Buoy Data Description

In this Chapter, we define the problem and highlight the limitations of existing data-driven forecasting models trained on individual buoy data in Section 3.1. In Section 3.2, we describe the data source for ECCC buoy data collection and a description of the buoy data. Lastly, in Section 3.3, we explain our predictive goal and the measure of success in our study.

3.1 Problem Definition

In the context of machine learning, the task of predicting significant wave height is approached as a time series problem. This involves transforming the accumulated historical data of significant wave heights over time into a univariate time series. Various machine learning models, as listed in Section 2.1.2, are commonly used for forecasting the significant wave height of a buoy at a specific global location.

This study focuses on increasing the accuracy (or minimizing prediction errors) of significant wave height for each buoy within 28 ECCC buoys located across the Pacific, Atlantic, and Great Lakes of Canada. We forecast significant wave heights with 1hr, 6hr, and 12hr forecasting horizons under normal conditions, providing a single value prediction for each hour in the case of 6hr and 12hr forecasts. During extreme events characterized by wave heights exceeding 6 meters, we focus on 1-hour forecasts.

3.2 Dataset Collection and Description

The buoy dataset collection (ECCC) contains the data collected from 28 buoys across Canada. These buoys actively report data every hour, each with their unique wave sampling period start time. Among the 28 buoys, 16 are positioned along the Pacific Ocean coast, 8 are situated within the Great Lakes and St. Lawrence Seaway, and

Figure 3.1: Buoy locations across Canada

4 are placed along the Atlantic Ocean coast. These buoys are maintained by Environment and Climate Change Canada (ECCC), and their geographic locations are shown in Figure 3.1. The data collected by the buoys are made available to the public by the Marine Environmental Data Service (MEDS), a division of the Department of Fisheries and Oceans Canada (Fisheries and Oceans [2019]).

Each buoy data encompasses wave fields such as significant wave height, maximum wave height, wave period, and other Meteorological and Oceanographic fields, along with buoy station ID and reporting time. Table 3.1 describes the fields in the buoy dataset. The buoy data is recorded at 1-hour intervals. Each buoy independently collects its data. Buoys may have different sampling periods, like 20 minutes or 40 minutes, depending on the measurement principles of the devices and distinct wave acquisition start times. To ensure a consistent time frame across all buoys, we standardize the data by grouping it hourly.

Buoy data is collected from sensors, so it is common to encounter missing values for certain fields and noisy data. In the context of buoy data, noisy data refers to information that contains errors or inconsistencies, typically arising from issues like sensor inaccuracies or incorrect measurements. The percentage of missing data for

Fields related to Wave Height	
VCAR	Characteristic significant wave height (calculated by MEDS) (m)
VWH$	Characteristic significant wave height (reported by the buoy) (m)
VCMX	Maximum zero crossing wave height (reported by the buoy) (m)
Fields related to Wave Period	
VTPK	Wave spectrum peak period (calculated by MEDS) (s)
VTP$	Wave spectrum peak period (reported by the buoy) (s)
Meteorological & Oceanographic fields	
WDIR	Direction from which the wind is blowing (° True)
WSPD	Horizontal wind speed (m/s)
WSS$	Horizontal scalar wind speed (m/s)
GSPD	Gust wind speed (m/s)
ATMS	Atmospheric pressure at sea level (mbar)
DRYT	Dry bulb temperature (°C)
SSTP	Sea surface temperature (°C)

Table 3.1: Various fields reported by buoy

each buoy is given in Appendix A.2. The median percentage of missing data across all buoys in the dataset ranges from 11 percent to 48 percent.

From Table 3.1, we see that there are duplicate fields for characteristic significant wave height (VCAR) and wave spectrum peak period (VTPK). Of these duplicate fields, the buoy reports one, and the other variable is recomputed from the spectra by MEDS. However, since 2 April 2020, MEDS discontinued recomputing significant wave height and peak period from the spectra. Although the rationale for this recalculation by MEDS has not been explicitly stated, we observed an average difference of 2 centimeters between the recalculated values and those directly reported by the sensors on the buoys.

3.3 Predictive Goal and Measure of Success

The main objective of our study is to improve the accuracy of significant wave height forecasting across all the ECCC buoys. We measure the success of the forecasting models, each trained on a group of related buoy data, by assessing the reduction in error achieved by the proposed model under both regular conditions. As per NOAA [2005], during extreme events, the average wave height of the highest 10 percent of

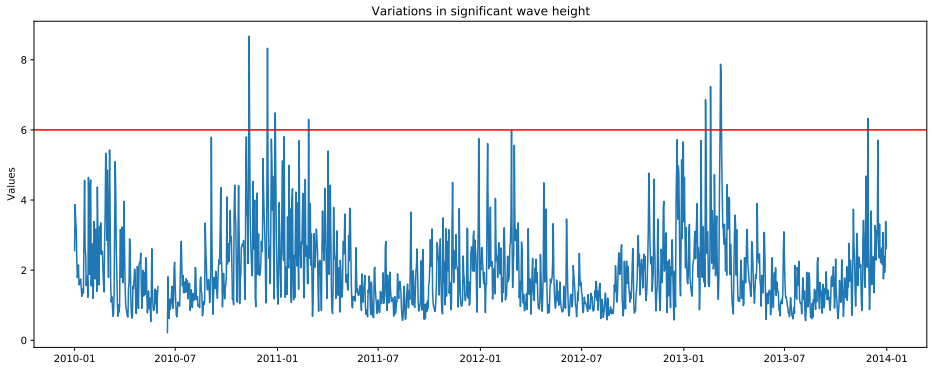

Figure 3.2: Significant Wave Height Time Series measurements for Buoy C44150. Redline marks wave heights above 6 meters.

the wave will be around 7 meters (which are rare as shown in Figure 3.2). We have considered the significant wave heights exceeding 6 meters, as this would consider the records close to 7 meters as well.

To assess whether the forecasting models trained on groups of related buoy data have improved the forecasting performance of significant wave height across all ECCC buoys, we first collect the baseline level of performance. In the existing literature, only one research study by Fasuyi et al. [2020] measured the Mean Absolute Error (MAE) for one of the 28 ECCC buoys using the Random Forest model. Since this study focuses on multiple buoys, we establish a baseline for comparison across all 28 ECCC buoys.

To create this baseline, we train individual forecasting models for each of the 28 buoys. For our comparison, we adopt the same model architecture (LightGBM) for training the individual buoy-based models and the proposed forecasting models that leverage cross-series information. Furthermore, we assess whether our proposed forecasting models improve performance compared to a single model trained on all 28 buoys.

As highlighted in subsection 2.1.2, the literature commonly uses statistical metrics, such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R2), to evaluate the performance of data-driven forecasting models, employing either one or a combination of these metrics. Also, in the literature, the model with the lowest error is typically considered more accurate for significant wave height

forecasting. Our study uses the Mean Absolute Error (MAE), which treats all errors with equal weight, as our evaluation metric. During our comparative analysis for each buoy, we compare the MAE of our proposed forecasting models with the MAE of individual buoy-based models and the model trained on all buoys separately. We also prioritize the model that demonstrates the lower error, following the common practice in the literature.

We use the Wilcoxon signed-rank test (Scheff [2016]), to assess whether a significant difference in improvement is achieved (across the entire ECCC buoy dataset) by the proposed forecasting models compared to the baseline models. Subsection 4.3.2 details the Wilcoxon signed-rank test and how it helps with decision-making.

Chapter 4

Methodology

In this chapter, we describe our research methodology. In Section 4.1, we outline the proposed approach. In Section 4.2, we present the workflow of our proposed approach, covering the data preparation, the process of clustering, and training forecasting models, along with the methods used. In Section 4.3, we explain the evaluation process used to verify the effectiveness of both the clustering algorithms and forecasting models.

4.1 Proposed Approach

This study aims to enhance the forecasting performance of significant wave height for each buoy within the ECCC buoy dataset collection by leveraging historical data from multiple related buoys. To achieve this, we propose a two-phase approach. First, we cluster the buoy data based on their wave patterns and environmental conditions to identify related buoys. Second, we train a global forecasting model on each cluster of related buoys, which we refer to as clustering-based models. This approach is depicted in Figure 4.1.

Before delving into clustering the buoy data, let us look at the buoy data from a time series perspective. To illustrate this, we will use one of the buoy datasets (C44137) as an example. This understanding helps formulate our approach for clustering the buoys and training a forecasting model for each cluster.

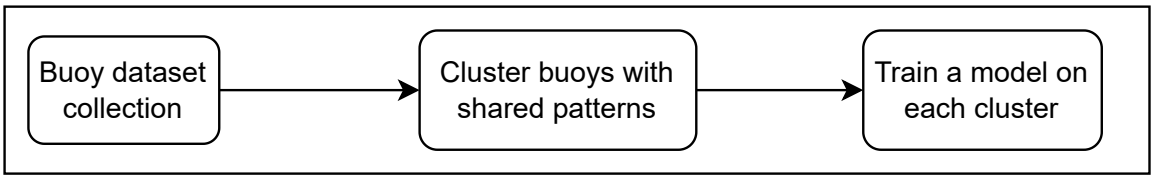

Figure 4.1: Proposed Approach for Significant Wave Height Prediction

STN_ID	DATE	VCAR	VTPK	VWH$	VCMX	VTP$	WDIR	WSPD	WSS$	GSPD	ATMS	DRYT	SSTP
C44137	05/16/2013 19:20	2.15	10.24	2.20	4.40	10.20	238.00	8.90		10.80	1006.80	10.20	9.40
C44137	05/16/2013 20:20	2.08	10.24	2.10	3.60	10.20	228.00	7.70		9.10	1007.30	9.50	9.50
C44137	05/16/2013 21:20	2.53	9.48	2.50	4.50	9.50	222.00	7.60		9.40	1007.50	9.60	9.80
C44137	05/16/2013 22:20	2.64	9.48	2.70	4.60	9.50	220.00	6.60		8.70	1007.80	10.10	9.80
C44137	05/16/2013 23:20	2.96	9.48	3.00	4.30	9.50	214.00	6.70		8.20	1007.90	10.10	9.80
C44137	05/17/2013 00:20	2.71	9.85	2.70	4.00	9.80	231.00	6.20		7.60	1008.40	10.30	9.80
C44137	05/17/2013 01:20	3.38	11.13	3.40	5.20	11.10	227.00	5.60		7.00	1008.20	10.50	9.60
C44137	05/17/2013 02:20	3.16	10.67	3.20	5.10	10.70	258.00	5.90		7.30	1008.30	9.40	9.50
C44137	05/17/2013 03:20	3.24	10.67	3.30	4.90	10.70	243.00	4.80		6.50	1009.20	9.90	9.50

Figure 4.2: Sample Data points of the buoy with station id C44137

4.1.1 Buoy DataSet Collection from Time series Perspective

A univariate time series involves only one variable observed over time, while a multivariate time series involves multiple variables observed over time, each forming its own time series (Wilson [2016]). Each buoy data is a multivariate time series because it contains multiple variables or measurements recorded for each observation at regular intervals. For example, consider the buoy with station id C44137, which collects ten distinct Meteorological and Oceanographic measurements, apart from sampling collection start time and buoy station id, every hour (Figure 4.2). The collected data for this buoy forms a time series, where each data point consists of values for all ten distinct fields at each reported timestamp.

4.1.2 Clustering Buoy Data

Various approaches can be employed to identify related buoy data, such as manual selection based on domain knowledge or automated techniques like clustering. We use clustering techniques to group similar buoy data based on shared wave patterns and environmental conditions.

As mentioned in Section 4.1.1, each buoy data is a multivariate time series, leading us to frame the clustering task as a multivariate time series clustering problem. We employ time series clustering to group buoy data, specifically adopting a feature-based approach. In this approach, we first extract temporal features from each field for each buoy data. These temporal features then become inputs for traditional clustering algorithms, enabling us to form clusters of related buoys based on their

wave patterns and environmental conditions. Therefore, this step consists of two sub-steps: Feature Extraction and Clustering.

Feature Extraction

We have data collection from n buoys. Each buoy data B_i, is depicted as a d-dimensional vector, with each dimension of the vector being a time series. The length of the time series differs across buoys. Given the quantity of data for each buoy and the inconsistency in the length of each time series, it is difficult to apply traditional clustering algorithms to the collection of buoy data. Consequently, we want to reduce the data into a common length vector of values for each buoy.

For each buoy B_i, we construct a feature vector E_i as follows. First, we summarize each d time series by 18 characteristics. These characteristics are given in Table 4.1. For each time series, for example, VWH$, we compute each characteristic and produce a vector in $\mathbb{R}^{18}$. The feature vector E_i is the catenation of the d vectors from $\mathbb{R}^{18}$. Thus, E_i is a vector from $\mathbb{R}^{18d}$. We use the set of n feature vectors $E = E_1, E_2, \ldots, E_n$ to cluster.

Clustering

Given the set of feature vectors that describe the buoy data, the next step is to cluster them using traditional clustering algorithms. Clustering can be formulated as defining a function $g(E)$ which takes the set of feature vectors and produces a set of clusters $C = C_1, C_2, \ldots, C_k$. Here, k is the number of clusters formed. Each cluster C_i is a subset of E. Upon completing the clustering phase, we save the resulting cluster information, which will be utilized in the subsequent training of the forecasting models.

4.1.3 Training Global Forecasting Models

After the clustering step, we obtain clusters of buoy data that coarsely share similar wave patterns and environmental conditions. We now proceed to train a forecasting model on each cluster. We train each forecasting model on all the buoy data of each cluster sequentially. This process results in a set of forecasting models denoted as

$M_1, M_2, ..., M_k$ where k represents the number of clusters of related buoys, and the value of k is dependent on the clustering algorithm employed.

4.2 Workflow of Proposed Approach

The workflow of our proposed approach involves two key steps: feature-based clustering and training forecasting models. We employ various methods in each of these steps to implement our approach. Also, before proceeding with any of these tasks, we first clean the data to ensure its quality and consistency.

4.2.1 Data Preparation

We preprocess the buoy data to manage records containing missing or noisy data. Our study focuses on the dataset from January 1, 2010, to December 31, 2021. We apply data processing steps to the buoy data collected during this timeframe to prepare it for downstream tasks. To handle missing or noisy data of significant wave height (VWH$), we replaced any noisy or missing VWH$ values occurring before April 2, 2020, with the corresponding calculated values (VCAR-denoting calculated values), if available. Similarly, for the wave spectrum peak period (VTP$), we replaced noisy or missing values with the corresponding calculated values (VTPK). We chose this replacement method because we believe that substituting with calculated values was a more reliable approach compared to interpolation, where the data is estimated based on neighboring values or outright removal, given that the calculated values are in close agreement (upto 2 centimeters) with the original data points. After replacement, we remove data records with missing or noisy data from the dataset.

For the variables VCAR, VCMX, and VTPK, any values surpassing 20 meters, 20 meters, and 60 seconds were treated as noise. The maximum wave height recorded in North Atlantic as of 2016, is of 19 meters (WMO [2016]). So, we assumed the wave measurements, VCAR and VCMX, above 20 meters as potentially be outliers or the result of sensor inaccuracies. For VTPK the value range, as we check all the buoy data, is usually below 60 unless it is noise or sudden change in value. We remove the corresponding data records from the preparation dataset. We observe no noisy values for other buoy data fields during the period considered for training and testing. Removing noisy records ensures the overall quality and reliability of the

dataset used in our analyses. Considering the timeframe we utilized for training and testing, excluding a relatively small number of noisy records is unlikely to impact the overall process.

The sampling periods for wave collection at each buoy have distinct start times and hence the data collected from various buoys is not uniformly aligned. To address this, we standardized the time frame across all buoys by calculating hourly averages, resulting in consistent one-hour intervals. This ensures that buoy behavior can be more easily compared over time. Consolidating the buoy dataset collection, we normalized the values to a range between 0 and 1 (for feature extraction and clustering). This preprocessing step ensures consistency and comparability of the data as it brings all the values to a common scale. We save the prepared data for further use in clustering and forecasting steps.

We split the preprocessed data into training and test sets during feature extraction, clustering, or model training. We divided the dataset into two segments, allocating 75 percent of the data for training and the remaining 25 percent for testing. This deviation from the traditional 80-20 rule was due to missing data for an extended period in some buoys during 2020. Thus training set includes data from January 1, 2010, to December 31, 2018, while the test set covers the period from January 1, 2019, to December 31, 2021.

We use the training data for feature extraction, clustering, and model training. Additionally, as part of data preparation for model training, the input data undergoes time delay embedding (Takens [1981]), called the auto-regression process. This involves constructing input features by considering the past twenty-four values of significant wave height; a choice arrived after exploratory data analysis and experiments detailed in Chapter 5. Consequently, we train the forecasting model using the preceding 24 values, also called lags, of significant wave height to accurately predict the target variable at any given moment. Furthermore, the model predicts the value for the next one-hour interval when making target predictions. We use this prediction as input for subsequent predictions in 6-hour and 12-hour forecasts.

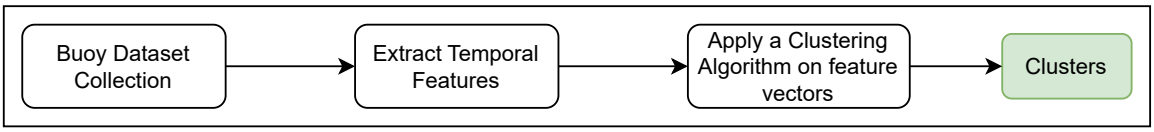

Figure 4.3: Feature Extraction and Clustering Workflow

4.2.2 Feature-Based Clustering

Identification of clusters of buoys with similar wave patterns and environmental conditions plays a crucial role in the overall proposed approach as this clustering decides on what data forecasting models get trained on and hence directly affects the performance of the forecasting results. In this step, we extract features from the buoy dataset collection and apply clustering algorithms (one at a time) to determine possible clusters. The feature extraction and clustering process are illustrated in Figure 4.3, and the steps involved in this phase are as follows:

1. Extract temporal features from cleaned data resulting in a feature vector for each buoy.

2. Apply a clustering algorithm on the feature vectors to identify buoys with similar wave patterns and environmental conditions.

3. Store the cluster labels of individual buoys for the downstream task of training forecasting models.

Feature Extraction

Each ECCC buoy reports 14 data fields, including timestamp, station id, and 12 fields listed in Table 3.1. In our study, we deliberately chose specific fields that we considered relevant to waves and the environmental factors influencing them. We focus on specific fields: VCAR, VTP, VCMX, WSPD, GSPD, ATMS, and SSTP. For each of these selected fields, we extract a set of features that capture the overall characteristics of the data related to that specific field.

Analyzing all the available libraries listed in subsection 2.3.1 and the kind of features extracted by each of these different libraries, we found TSFEL(Barandas et al. [2020]) as the most suitable choice for our dataset for two reasons. First, the temporal features extracted by TSFEL are highly robust to noise (asserted by its

developers), which is particularly important for buoy data that often contains sensor noise (if any is available after cleaning the data). Second, these features are unaffected by the varying lengths of buoy data, ensuring consistent and reliable feature extraction across all buoys in our dataset. Table 4.1 presents the list of features extracted for each field of every buoy data. Clustering algorithms then use these extracted features to identify related buoys.

Clustering

To identify clusters of buoys exhibiting similar wave and environmental patterns, we test a variety of clustering algorithms such as K-means, Affinity Propagation, DBSCAN, OPTICS, and Agglomerative. Each of these algorithms operates on the extracted feature vectors of the buoy dataset collection. In each of these clustering algorithms, the number of clusters is determined based on the specific clustering algorithm employed. The parameter settings for each algorithm are given in Chapter 6.1.

4.2.3 Training Global Forecasting Models

After identifying clusters, we train a global forecasting model on each cluster. We employ a holdout estimation approach to divide the buoy dataset collection into separate training and test datasets for each buoy of the buoy dataset collection. The training dataset consists of data from January 1, 2010, to December 31, 2018, while the testing dataset covers the period from January 1, 2019, to December 31, 2021.

Training process for a single cluster is illustrated in Figure 4.4, and the steps involved in this phase are as follows:

1. From buoy dataset collection choose the buoys that are part of same cluster.
2. Split each buoy data into training and test datasets
3. Combine training datasets of all the buoys in the cluster.
4. Train a forecasting model (referred as clustering-based model) using LightGBM on the aggregated training datasets.

Feature Name	Feature Description
Absolute energy	Sum of the squared magnitudes of the signal values.
Area under the curve	Integral of the time series over a specified time interval.
Autocorrelation	Measure of how closely a time series is related to itself over time.
Centroid	Center of mass of the time series and represents the average value of the signal.
Entropy	Measure of the randomness or unpredictability of the signal values.
Mean absolute diff	Average absolute difference between adjacent signal values.
Mean diff	Average difference between adjacent signal values.
Median absolute diff	Median absolute difference between adjacent signal values.
Median diff	Median difference between adjacent signal values.
Negative turning points	Points in a time series where the signal value changes from positive to negative.
Neighborhood peaks	Peaks in a time series that are above a certain threshold and occur within a time interval.
Peak to peak distance	Distance between the highest and lowest points of a signal.
Positive turning points	Points in a time series where the signal value changes from negative to positive.
Signal distance	Euclidean distance between two signals.
Slope	Average rate of change of the signal values over time.
Sum absolute diff	Sum of the absolute differences between adjacent signal values.
Total energy	Sum of the squared magnitudes of the signal values.
Zero crossing rate	Number of times the signal crosses the zero axis per unit time.

Table 4.1: Features extracted for each variable reported by buoy

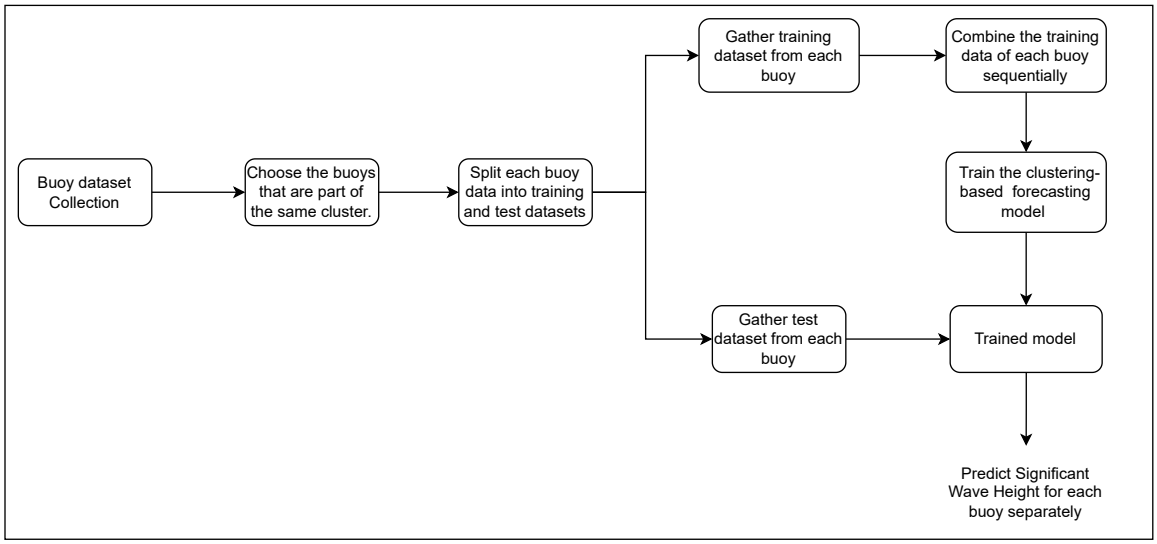

Figure 4.4: Forecasting workflow for single cluster

5. Pass each test dataset of the buoys to predict significant wave height.

We repeat this procedure for each cluster identified by the clustering algorithm, resulting in a number of models corresponding to the clusters determined by the algorithm. Each clustering algorithm generates distinct clusters.

Method

The global forecasting model utilized in our study is LightGBM, as detailed in subsection 2.2.2. A few reasons to select LightGBM as a forecasting model for our experiments is that LightGBM offers a range of useful features and capabilities for time series forecasting, such as regularization techniques, efficient gradient boosting, and scalability (Ke et al. [2017]). These features can help improve the generalization ability of the model, handle overfitting, and efficiently use computing resources. Also, LightGBM has been shown to perform well compared to other popular forecasting models in various benchmark tests (Makridakis et al. [2022b]). This gives us confidence that the model will provide accurate and reliable forecasts for each time series in our analysis. The parameter settings for LightGBM can be found in Chapter 6.1.

4.3 Evaluation

4.3.1 Cluster Evaluation

The clustering process applied to the feature-extracted buoy data is an unsupervised approach, indicating that no predefined ground truth clusters are available for comparison. We rely on intrinsic measures to assess the quality of the clusters. Specifically, we use the silhouette score (Rousseeuw [1987]). The silhouette score is calculated for each data point and represents how similar the data point is to its cluster compared to other neighboring clusters. It ranges from -1 to 1.

A silhouette score close to -1 indicates that the data point is incorrectly clustered and would be better placed in a neighboring cluster. Conversely, a score of 1 indicates the opposite. A silhouette score 0 indicates that the data point is close to the decision boundary between two clusters. It implies that the data point could belong to either cluster. Our primary goal is not to determine the best clustering algorithm for segregating and finding related buoy data. We use the silhouette score to see how well each clustering algorithm separates the data into distinct groups. We test our workflow using the clustering algorithms: K-means, Affinity Propagation, DBSCAN, OPTICS, and Agglomerative.

4.3.2 Forecasting model Evaluation

Evaluation Procedure

To assess the effectiveness of clustering-based models in improving forecasting results, we initially need to establish the baseline level of performance of existing models. Since we do not have baseline results of any model trained on each of ECCC buoy dataset collections (except one by Fasuyi et al. [2020]) to compare the clustering-based forecasting results, we consider two alternative forecasting models to assess the performance of our proposed model: local forecasting models and the universal forecasting model. Local models are trained individually for each buoy, using only its training data (resulting in 28 models for 28 buoys). These individual models then forecast the corresponding test set of the buoy. On the other hand, we train the universal forecasting model using all the available training data from all buoys (resulting in 1 model). This one universal model then forecasts the test set of individual buoys. By

examining the performance of the clustering-based models against these local models and the universal model separately, we can gain insights into the effectiveness and potential advantages of our proposed approach.

Evaluation metric

To assess the accuracy of the predictions on the buoy dataset, we compare the Mean Absolute Error (MAE) of the clustering-based forecasting models with that of the local models and universal model for each buoy. We employ the Wilcoxon signed-rank test (Scheff [2016]) to assess whether a significant difference in improvement exists between the compared models: clustering-based vs. local and clustering-based vs. universal.

The Wilcoxon signed-rank test is a non-parametric statistical test. It is used to evaluate whether a significant difference exists between paired observations, where each data point in one dataset corresponds to a data point in another dataset for comparison. We opt for this test because it is less influenced by outliers and is non-parametric, which avoids assumptions about the specific distribution of the data (Scheff [2016]). Also, Demšar [2006] recommends this test to compare pairs of predictive models.

This test gives us information about the p-value. The p-value is a measurement of how strong the evidence is against the idea that there is no real difference between the two sets of data we are comparing. A small p-value suggests that the observed difference is likely genuine and not due to chance fluctuations. P-values of 0.05 and 0.01 are commonly used as a standard threshold (Moore et al. [2012]), and we adopt the threshold of 0.05 for our study. We input the actual MAE differences between the compared models in this test to obtain the p-value.

Once the overall procedure is in place, for any new buoy data, we extract the features and identify the cluster that particular buoy data belongs to and then use the corresponding clustering-based forecasting model to train on the new buoy data and generate forecasts.

Chapter 5

Exploratory Data Analysis of ECCC Buoy Data

In this Chapter, we explore the ECCC buoy dataset to understand its characteristics and assess the feasibility of clustering buoys with similar wave patterns and environmental conditions based on raw data examination. Additionally, we also explore the relationships between different fields of buoy data.

5.1 Feasibility Assessment for Buoy Clustering

In our study, one of the fundamental steps involves clustering buoys based on their wave patterns and environmental conditions. To achieve this, we have opted for a feature-based clustering approach. However, before proceeding with this method, it is essential to understand the raw data to determine if clusters are feasible. For instance, if each buoy reports distinct wave patterns, it may result in the absence of meaningful clusters. Therefore, this initial data exploration will help us assess the feasibility and to obtain a rough estimation of potential clusters within the buoy data based on wave characteristics.

Figures 5.1 and 5.2 offer insights into the variations in significant wave heights observed in the Pacific and Atlantic oceans. The figures depict that wave heights in the Pacific and Atlantic range from 0.1 meters to around 9 meters. However, the Pacific coast exhibits a higher frequency of observations in the range above 3 meters, in contrast to the Atlantic. Additionally, this observation acknowledges that the Pacific coast tends to experience higher waves than the Atlantic coast (Thompson et al. [1972]). Furthermore, Figure 5.3 provides an overview of wave heights in the Great Lakes, indicating that most observations fall within the range of 0.1 meters to 1 meter, with very few instances exceeding 1 meter.

In the Pacific and Atlantic oceans, examining the geographic distribution of buoys reveals that the buoys are positioned across a range of depths along the coasts of both oceans. For example, Figure 5.4, focusing on the Pacific buoys, illustrates the

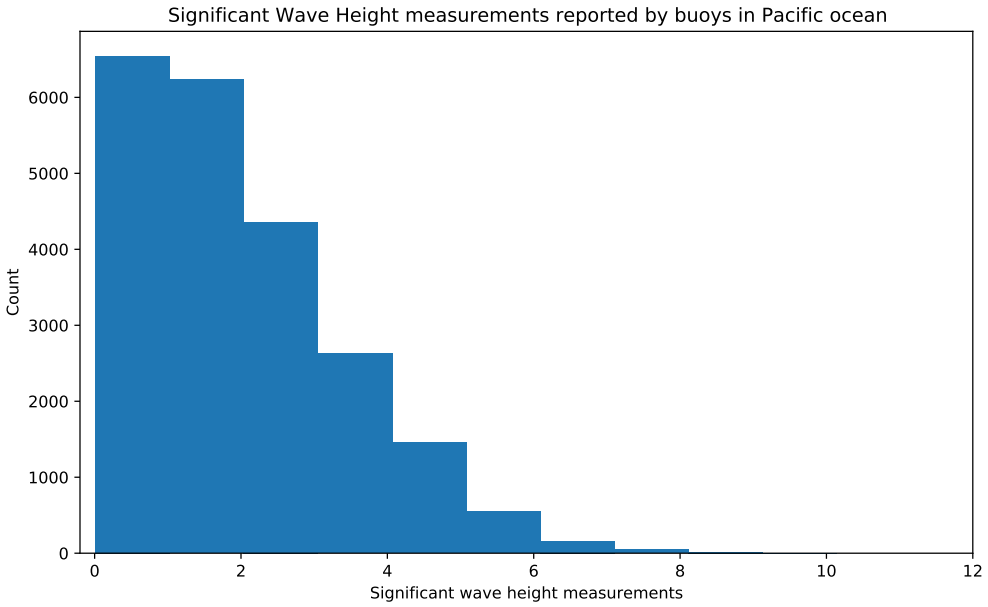

Figure 5.1: Histogram of significant wave heights reported by Pacific Ocean buoys between January 2010 and December 2013.

varying distances of these buoys from the shore, spanning from deep ocean locations to intermediate and nearshore areas. Based on the location of the buoys, we estimate that buoys near the shore may experience shorter wave heights than those in the deep ocean.

To check the wave height ranges of the buoys positioned across a range of depths in the Pacific, we examined the data from three buoys positioned at different depths: one from the deep end, another from an intermediate position, and a third from the shallow end. The significant wave heights observed for these buoys are depicted in line graphs in Figures 5.5, 5.6, and 5.7. These figures illustrate that the significant wave height ranges differ based on the position of the buoy.

Similarly, we examined the Atlantic buoys at varying distances from the shore. The data includes one buoy close to the shore and the other three in the deep ocean. These details are included in Appendix A.

The insights derived from examining wave height variations and geographic distribution provide valuable cues for the rough estimation of clusters in our study. The significant differences in wave height frequencies between the Pacific and Atlantic

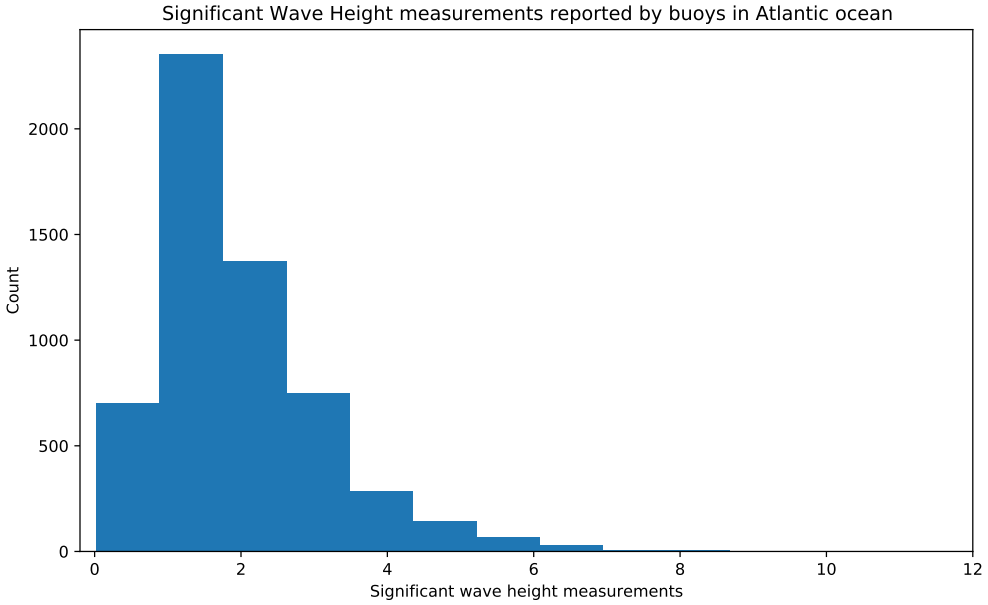

Figure 5.2: Histogram of significant wave heights reported by Atlantic Ocean buoys between January 2010 and December 2013.

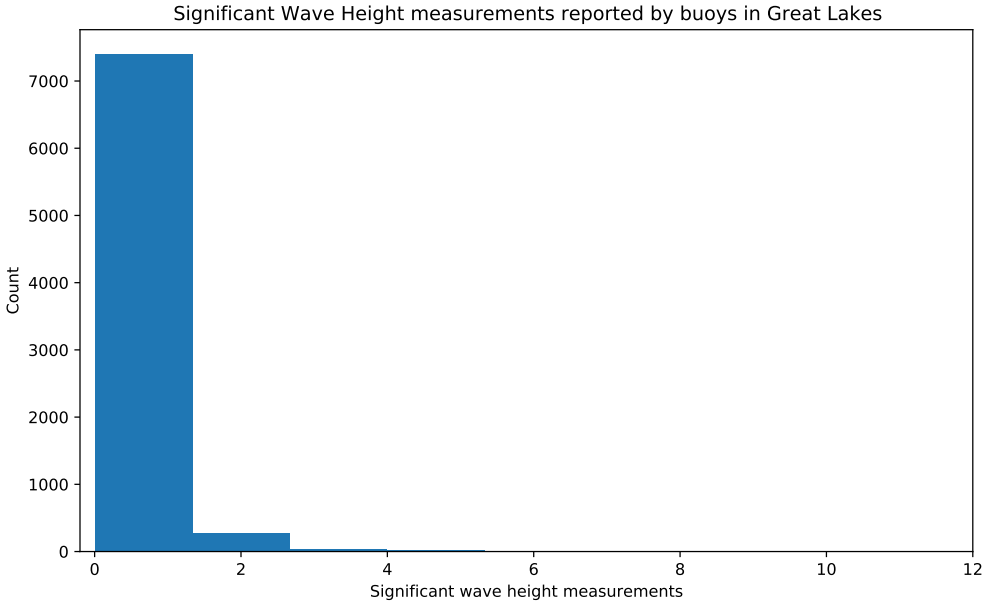

Figure 5.3: Histogram of significant wave heights reported by Great Lakes and Seaway buoys from January 2010 to December 2013.

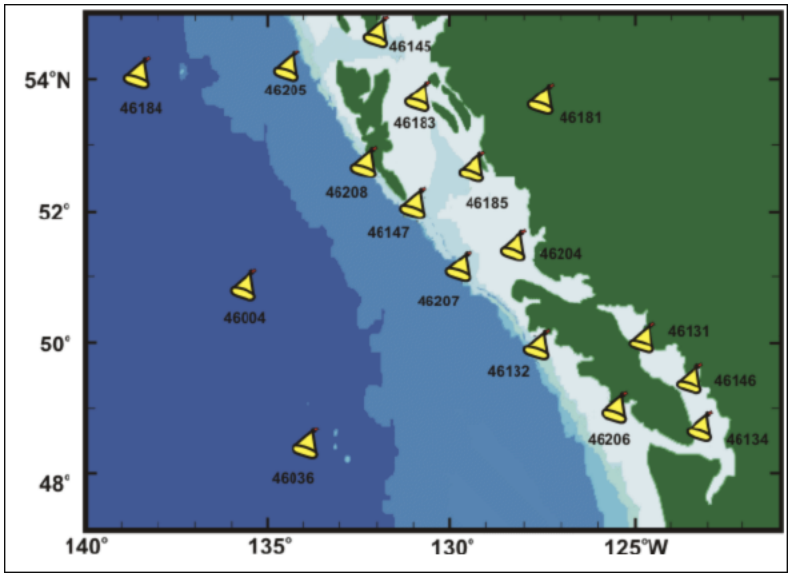

Figure 5.4: Buoys in Pacific Ocean

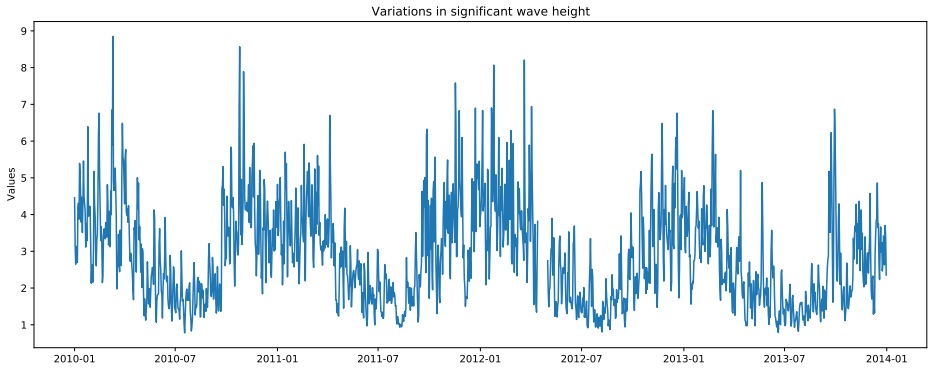

Figure 5.5: Line graph illustrating 4-year significant wave height variations for buoy C46036 in the deep Pacific Ocean.

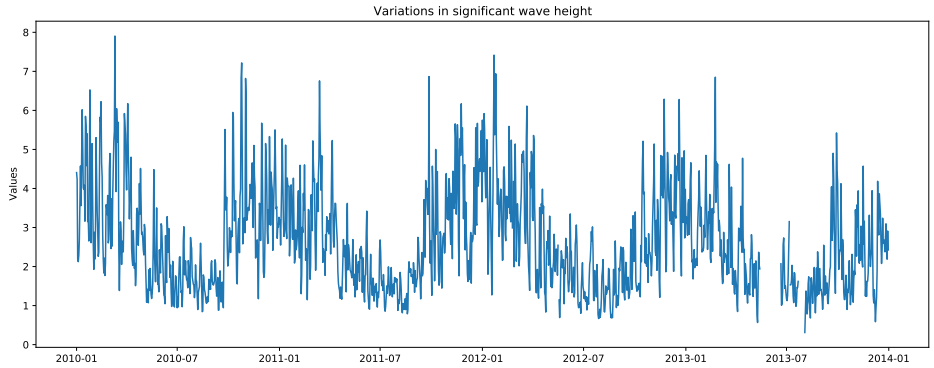

Figure 5.6: Line graph showing 4-year significant wave height changes for buoy C46132 in the mid-Pacific Ocean.

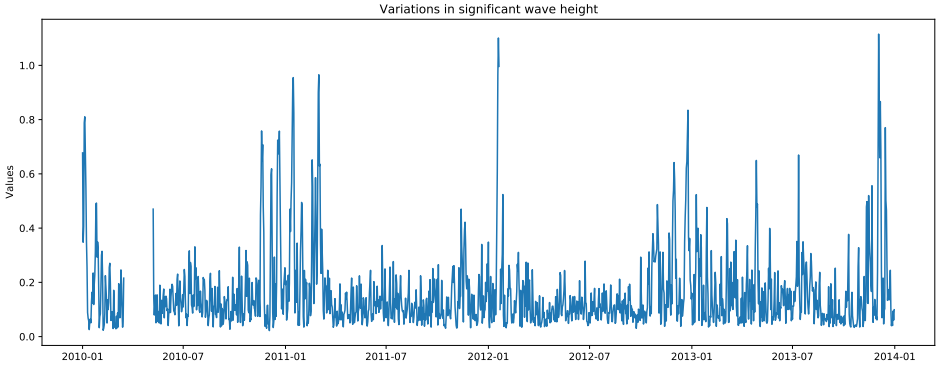

Figure 5.7: Line graph depicting 4-year significant wave height changes for buoy C46181 in the shallow Pacific Ocean.

coasts suggest that buoys in each region might exhibit distinct patterns and characteristics, potentially forming separate clusters. Furthermore, the observed variations in the wave height range between buoys near the shore and buoys in mid to deep ocean areas in both oceans propose that these buoys may form distinct clusters based on their proximity to the shore.

5.2 Understanding Relationship between Buoy Data Fields

One of the initial steps in our analysis is to identify fields that exhibit a strong correlation with the target variable, significant wave height, as these fields are likely to influence the prediction or forecasting of significant wave height. We calculated a correlation matrix to examine the relationship between significant wave height and other fields. The correlation matrix aids in feature selection by identifying fields that strongly correlate with a target variable, significant wave height, making them potentially important for prediction or forecasting tasks. Figure 5.8 presents the correlation between all buoy data fields. It is important to note that this correlation is calculated using the combined data from all 28 buoys.

Correlation coefficient values between 0.7 and 1.0 (or -0.7 and -1.0) indicate a strong positive (or negative) correlation. Values between 0.3 and 0.7 (or -0.3 and -0.7) typically indicate a moderate positive (or negative) correlation. Values below 0.3 (or above -0.3) are commonly regarded as weak or negligible correlations (Cohen [1988]). Analyzing the correlation coefficients from Figure 5.8, we find that significant wave height (VCAR) demonstrates a strong correlation with maximum wave height (VCMX), with a coefficient value of 0.98. This result aligns with our expectations, as significant wave height is the average of the one-third highest waves, and therefore, it is influenced by VCMX. Additionally, we observe that wind speed and gust speed exhibit a moderate correlation, with coefficient values of 0.54 each.

Based on the analysis of the correlation matrix, the significant wave height has a strong positive correlation with VCMX, as indicated by a correlation coefficient of 0.98, suggesting a high level of association between these two fields. However, including VCMX as an input feature may introduce unnecessary noise to the model as it is the highest wave height reported during the sampling period, whereas significant wave height is the average of one-third of the highest wave heights. Additionally, in

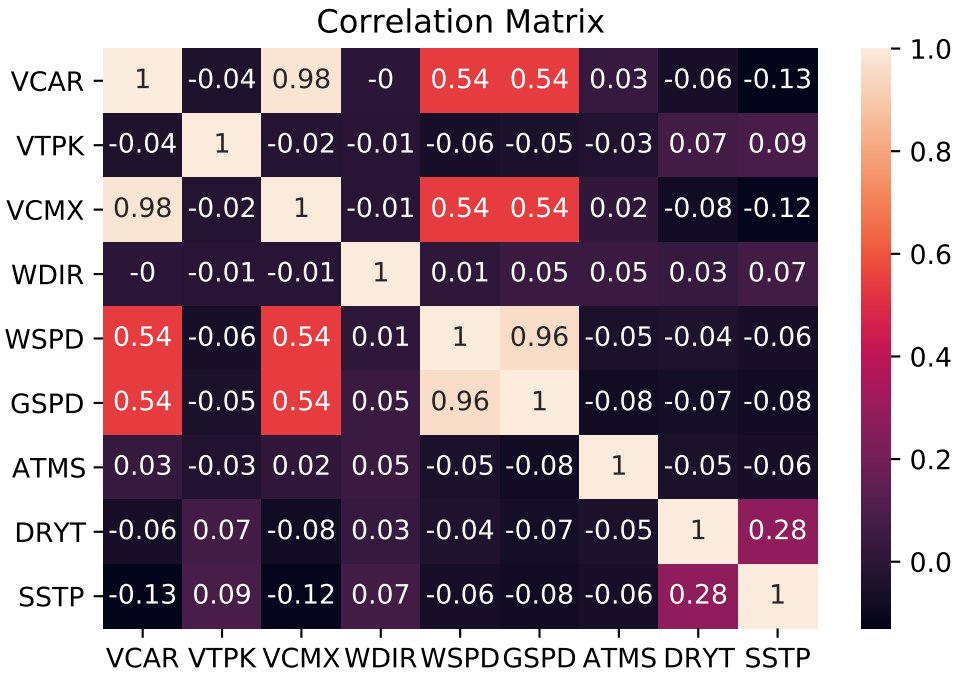

Figure 5.8: Correlation between all buoy data fields

real-time, we never ask the model to forecast the significant wave height given the real values of the VCMX, as both are related to wave measurements that buoy reports at a time. Also, if we can get the VCAR, we can approximate the VCMX as, in most cases, VCMX is approximately 1.86 times the significant wave height.

In addition to examining the correlation between significant wave height and other fields, we also explore the Autocorrelation Function (ACF) specifically for significant wave height. The ACF provides valuable insights into the relationship between a variable and its past values, allowing us to assess the presence of any temporal dependencies or patterns. By analyzing the ACF plot of significant wave height, we can observe the decay in autocorrelation at different lags.

Figure 5.9 illustrates the ACF of significant wave height (VCAR), revealing important information about the temporal dynamics of the significant wave height. We can see a slower decay in autocorrelation, indicating that the values of the series are influenced by their past values, even at long lags. This suggests the presence of a strong temporal dependency in the significant wave height, where each observation

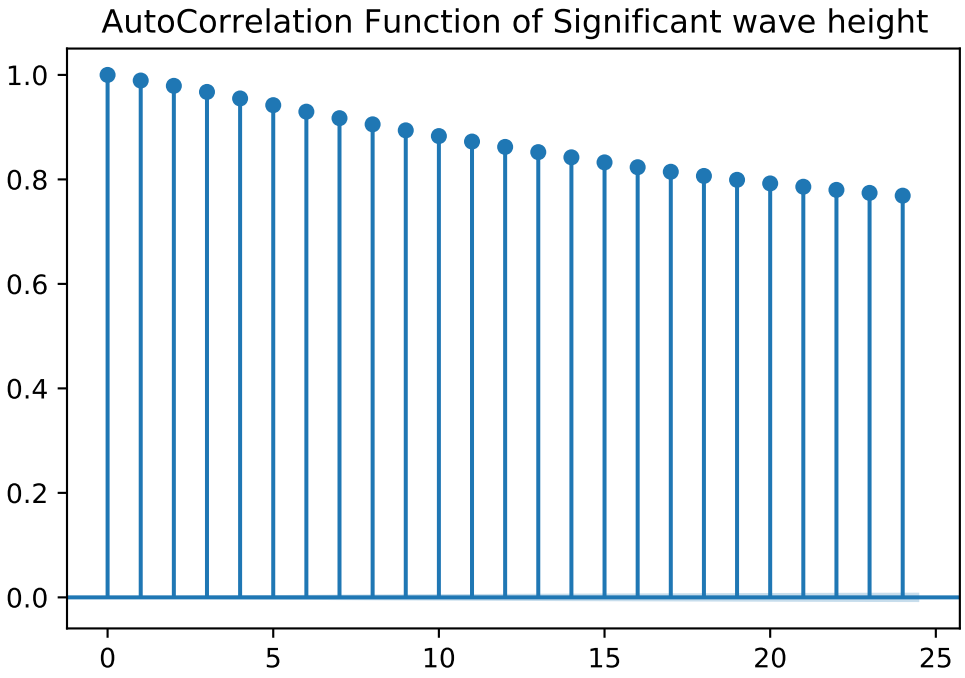

Figure 5.9: AutoCorrelation Function of Significant Wave Height

is dependent on its historical values. This indicates that incorporating the lagged values of significant wave height as features in our forecasting models can effectively capture the temporal dependencies.

Our forecasting approach aims to capture temporal dependencies and patterns in significant wave height data. In the autocorrelation function (ACF) plot, we observe a slower decay in autocorrelation, indicating that past values strongly influence the current values of significant wave height. This suggests that historical data holds crucial information for accurately predicting future significant wave heights.

To effectively capture these temporal patterns and dependencies specific to significant wave height, we used only lagged values of significant wave height in our forecasting models. By doing so, the model can directly incorporate the relevant historical data to make focused and precise predictions for this particular variable. We conducted experiments using 12, 24, and 36 lag values for one of the buoys to determine the optimal number of lag values. We observed only marginal differences between using 12 and 24 lags. However, considering that employing 24 lags allows

the model to consider the effects of previous cycles within the whole day, we decided to use 24 lags to capture the time dependencies effectively.

Chapter 6

Experiments and Results

This chapter focuses on the experimental setup, execution, and results. Section 6.1 provides an overview of the experimental setup, including infrastructure details and parameter settings. We present the conducted experiments in the subsequent Section 6.2. Finally, we present an analysis and insights from the study in Section 6.3.

6.1 Experimental Setup and Configuration

Infrastructure Details

The experiments were conducted on a system with the following configuration and no other processes running in the background: The machine featured a Quad-Core Intel Core i5 processor with a clock speed of 2 GHz. The system had a single processor with four cores, each with 512 KB of L2 cache and 6 MB of L3 cache. Hyper-Threading Technology was enabled, allowing the processor to handle multiple threads simultaneously. The system was equipped with 16 GB of memory. The operating system used was macOS Monterey, with version 12.5.1.

Algorithm Parameter Settings

The specific parameters employed in each of the algorithms used in our study are:

1. K-means: To determine the optimal number of clusters, we employed the elbow method. Applying the elbow method (Bishop [2006]) to our data, we found that the optimal number of clusters for the given feature vectors is 3.

2. Affinity Propogation: During our experimentation, we explored various damping values ranging from 0.5 to 0.8 to determine their impact on the clustering results. We observed that the silhouette score remained consistent across all damping values within this range. However, when we increased the damping

value to 0.9, the silhouette score exhibited a minor improvement of 0.01. Despite this improvement, we noticed that the forecasting results on the test set deteriorated compared to the clustering performed with a damping value of 0.8. As a result, use a damping value of 0.8 in our analysis, as it produced more favorable forecasting outcomes.

3. DBSCAN: To determine the optimal value for the epsilon (eps) parameters in the DBSCAN algorithm, we employed the use of a k-distance graph. By analyzing the k-distance graph, we found appropriate eps value to be 2.75. Additionally, we set the min_samples parameter to 2, specifying that a minimum of two points should be present within the eps radius to form a dense region.

4. OPTICS: To determine the appropriate epsilon (eps) value, we examined the reachability plot. We observed that with an epsilon value of 0.75, the reachability plot displayed significant changes in density-reachability distances. Additionally, we set the min_samples parameter to 4 after experimenting with few values and checking the silhouette score.

5. Agglomerative: To determine the optimal number of clusters, we performed experiments and evaluated the silhouette score and subsequent forecasting results. After testing different number of clusters, we found that setting the number of clusters to 4 yielded lower MAE compared to other cluster configurations for this clustering.

6. LightGBM: We have chosen to train the LightGBM model for regression tasks, with a relatively small number of leaves (15) to control the complexity of the model. The learning rate was set to 0.05 to balance the speed of convergence and the accuracy of the model. The performance of the models was evaluated using the Mean Squared Error and Mean Absolute Error metrics. We also enabled linear trees, which offer interpretability and competitive performance in forecasting significant wave height.

6.2 Experiments

In our study, the overall workflow consists of two steps: feature-based clustering and forecasting. We tested our workflow using K-means, Affinity, DBSCAN, OPTICS, and Agglomerative clustering algorithms. We use a single clustering algorithm for each run to create clusters of related buoy data. Subsequently, we trained separate forecasting models on each cluster and predicted significant wave height for the next 1 hour for each buoy. This approach allows us to assess how the forecasting results may vary based on the specific clustering algorithm. We used LightGBM, with the same hyperparameters, for training all the forecasting models.

We conducted experiments with forecasting horizons of 1-hour, 6-hour, and 12-hour predictions under regular conditions. We predicted significant wave height with a 1-hour forecasting horizon during extreme events. These extreme events are defined by significant wave heights exceeding 6 meters.

We repeated the experiments twice to test for any differences in the forecasting results, considering the variability of regression problems. We observed that the clustering results were the same for the second run. The MAE of forecasting results (measured in meters) did not change for up to 3 decimal places. The code used in this study is available online.[1]

When considering the clustering approach, a natural consideration would be to cluster based on the major regions where the buoys are located. Therefore, we organized the buoys into three clusters: Pacific, Atlantic, and Great Lakes, and trained a forecasting model on each cluster. The forecasting results for each cluster, for forecasting horizons of 1hr, 6hr, and 12hr under regular conditions and 1hr under extreme events are detailed in Section 6.3.

6.3 Results

Figures 6.1 through 6.5 depicts the results obtained for each tested clustering algorithm. K-means and Affinity Propagation, centroid-based clustering algorithms, grouped the buoy data into 3 and 4 clusters, respectively. DBSCAN and OPTICS, density-based clustering algorithms, grouped the buoy data into 3 and 2 clusters,

[1] https://git.cs.dal.ca/chandrala/buoyclusterforecast.git

Figure 6.1: K-means Clustering

respectively. Agglomerative clustering, a hierarchical-based algorithm, grouped the buoy data into 4 clusters.

Table 6.1 presents the clusters identified by each clustering algorithm tested. We observe that K-means achieves the highest silhouette score among the tested clustering algorithms, indicating better cluster separation and coherence than other clustering algorithms.

As described in Section 3.3, we establish local models and a universal model as baseline models. To assess the effectiveness of our proposed approach, we compare each baseline model (local and universal) with the clustering-based forecasting model results. Hence, we will begin by presenting the forecasting results of the local models and the universal model. Tables 6.2 and 6.3 present the forecasting results of the local models and the universal model for forecasting horizons of 1-hr, 6-hr, 12-hr in regular conditions and for 1-hr under extreme events, respectively. In Table 6.2 and 6.3, N/A indicates no data points in the test dataset where the significant wave height exceeds 6 meters. All reported errors are in meters, rounded to two decimal places

Figure 6.2: Affinity Propagation Clustering

Figure 6.3: DBSCAN Clustering

Figure 6.4: OPTICS Clustering

Figure 6.5: Agglomerative Clustering

Algorithm	# Clusters	Cluster ID	Buoys Associated	Silhouette Score
K-means	3	1	3	0.417
		2	9	
		3	16	
Affinity	4	1	3	0.240
		2	7	
		3	9	
		4	9	
DBSCAN	3	1	3	0.341
		2	8	
		3	16	
		outlier	1	
OPTICS	2	1	11	0.398
		2	17	
Agglomerative	4	1	3	0.315
		2	4	
		3	8	
		4	13	

Table 6.1: Clustering results for the buoy dataset.

(centimeters), in alignment with the reporting standard used by MEDS, from which we collected the ECCC buoy dataset.

Clustering-Based Forecasting Results: Regular Conditions (1-hr, 6-hr, 12-hr)

The resulting MAEs for each clustering-based forecasting model are summarized in Table 6.4 for 1-hour predictions, Table 6.5 for 6-hour predictions, and Table 6.6 for 12-hour predictions. For the 6-hour and 12-hour predictions, we used the previous predictions as inputs for the subsequent forecasts sequentially. All reported errors are measured in meters and rounded to two decimal places (centimeters).

BuoyId	1-hr	6-hr	12-hr	1-hr extreme events
C44137	0.14	0.29	0.43	0.58
C44139	0.13	0.28	0.42	0.48
C44150	0.13	0.27	0.39	0.61
C44258	0.10	0.21	0.30	1.50
C45132	0.06	0.14	0.19	N/A
C45136	0.05	0.10	0.15	N/A
C45139	0.04	0.09	0.12	N/A
C45143	0.05	0.12	0.17	N/A
C45149	0.06	0.14	0.19	N/A
C45151	0.06	0.11	0.13	N/A
C45154	0.06	0.11	0.15	N/A
C45159	0.05	0.10	0.14	N/A
C46004	0.17	0.35	0.49	0.64
C46036	0.16	0.30	0.41	0.53
C46131	0.07	0.15	0.20	N/A
C46132	0.15	0.28	0.40	0.48
C46145	0.12	0.23	0.31	0.85
C46146	0.06	0.12	0.15	N/A
C46147	0.17	0.33	0.46	0.60
C46181	0.04	0.08	0.09	N/A
C46183	0.27	0.51	0.67	2.43
C46184	0.16	0.30	0.43	0.52
C46185	0.12	0.28	0.40	0.77
C46204	0.13	0.28	0.38	0.56
C46205	0.15	0.29	0.41	0.56
C46206	0.14	0.28	0.41	0.56
C46207	0.16	0.31	0.44	0.49
C46208	0.15	0.29	0.41	0.46

Table 6.2: MAEs of the Local forecasting models for the ECCC buoy dataset collection for forecasting horizons of 1-hr, 6-hr, 12-hr and for 1-hr under extreme events.

BuoyId	1-hr	6-hr	12-hr	1-hr extreme events
C44137	0.14	0.29	0.43	0.58
C44139	0.13	0.27	0.41	0.46
C44150	0.13	0.27	0.39	0.61
C44258	0.10	0.19	0.28	1.20
C45132	0.06	0.15	0.22	N/A
C45136	0.05	0.12	0.18	N/A
C45139	0.05	0.10	0.14	N/A
C45143	0.06	0.13	0.19	N/A
C45149	0.06	0.15	0.21	N/A
C45151	0.06	0.11	0.14	N/A
C45154	0.06	0.12	0.17	N/A
C45159	0.05	0.12	0.17	N/A
C46004	0.17	0.33	0.47	0.64
C46036	0.16	0.29	0.40	0.54
C46131	0.07	0.16	0.23	N/A
C46132	0.15	0.28	0.40	0.47
C46145	0.12	0.23	0.31	0.79
C46146	0.06	0.13	0.17	N/A
C46147	0.17	0.33	0.44	0.59
C46181	0.05	0.09	0.12	N/A
C46183	0.25	0.47	0.63	1.91
C46184	0.16	0.30	0.42	0.53
C46185	0.12	0.28	0.41	0.81
C46204	0.13	0.27	0.37	0.55
C46205	0.15	0.29	0.39	0.56
C46206	0.13	0.26	0.36	0.50
C46207	0.16	0.30	0.43	0.50
C46208	0.15	0.28	0.40	0.47

Table 6.3: MAEs of the Universal forecasting models for the ECCC buoy dataset collection for forecasting horizons of 1-hr, 6-hr, 12-hr and for 1-hr under extreme events.

BuoyId	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
C44137	0.14	0.14	0.14	0.14	0.14
C44139	0.13	0.13	0.13	0.13	0.13
C44150	0.13	0.13	0.13	0.13	0.13
C44258	0.10	0.10	0.10	0.10	0.10
C45132	0.06	0.06	0.06	0.06	0.06
C45136	0.05	0.05	0.05	0.05	0.05
C45139	0.04	0.04	0.04	0.04	0.04
C45143	0.05	0.05	0.05	0.05	0.05
C45149	0.06	0.06	0.06	0.06	0.06
C45151	0.06	0.06	0.06	0.06	0.06
C45154	0.06	0.06	0.06	0.06	0.06
C45159	0.05	0.05	0.05	0.05	0.05
C46004	0.17	0.17	0.17	0.17	0.17
C46036	0.16	0.16	0.16	0.16	0.16
C46131	0.07	0.07	0.07	0.07	0.07
C46132	0.14	0.14	0.14	0.14	0.14
C46145	0.12	0.12	0.12	0.12	0.12
C46146	0.06	0.06	0.06	0.06	0.06
C46147	0.17	0.17	0.17	0.17	0.17
C46181	0.05	0.05	0.05	0.05	0.05
C46183	0.25	0.24	0.25	0.24	0.24
C46184	0.16	0.16	0.16	0.16	0.16
C46185	0.12	0.12	0.12	0.12	0.12
C46204	0.13	0.13	0.13	0.13	0.13
C46205	0.15	0.15	0.15	0.15	0.15
C46206	0.13	0.13	0.13	0.13	0.13
C46207	0.16	0.16	0.16	0.16	0.15
C46208	0.15	0.15	0.15	0.15	0.15

Table 6.4: MAEs of clustering-based forecasting models for the ECCC buoy dataset collection in the context of 1-hour prediction.

BuoyId	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
C44137	0.29	0.29	0.29	0.29	0.29
C44139	0.27	0.27	0.27	0.27	0.27
C44150	0.27	0.27	0.27	0.27	0.27
C44258	0.19	0.19	0.21	0.20	0.21
C45132	0.15	0.15	0.14	0.14	0.14
C45136	0.11	0.11	0.11	0.11	0.11
C45139	0.10	0.10	0.09	0.09	0.09
C45143	0.12	0.12	0.12	0.12	0.12
C45149	0.14	0.14	0.14	0.14	0.14
C45151	0.10	0.10	0.10	0.10	0.10
C45154	0.12	0.12	0.12	0.12	0.12
C45159	0.11	0.11	0.11	0.11	0.11
C46004	0.33	0.34	0.33	0.33	0.33
C46036	0.29	0.29	0.29	0.29	0.29
C46131	0.15	0.15	0.15	0.15	0.15
C46132	0.28	0.28	0.28	0.28	0.28
C46145	0.24	0.23	0.24	0.24	0.23
C46146	0.12	0.12	0.12	0.12	0.12
C46147	0.32	0.32	0.32	0.32	0.32
C46181	0.09	0.09	0.09	0.09	0.09
C46183	0.47	0.47	0.47	0.47	0.47
C46184	0.30	0.30	0.30	0.30	0.30
C46185	0.28	0.28	0.28	0.28	0.28
C46204	0.27	0.27	0.27	0.27	0.27
C46205	0.29	0.29	0.29	0.29	0.29
C46206	0.26	0.26	0.26	0.26	0.26
C46207	0.30	0.30	0.30	0.30	0.30
C46208	0.28	0.28	0.28	0.28	0.28

Table 6.5: MAEs of clustering-based forecasting models for the ECCC buoy dataset collection in the context of 6-hour prediction.

BuoyId	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
C44137	0.43	0.44	0.43	0.43	0.42
C44139	0.41	0.41	0.41	0.41	0.40
C44150	0.40	0.40	0.40	0.39	0.39
C44258	0.26	0.26	0.30	0.30	0.31
C45132	0.20	0.20	0.20	0.20	0.20
C45136	0.16	0.16	0.15	0.16	0.15
C45139	0.13	0.13	0.12	0.12	0.12
C45143	0.17	0.17	0.17	0.17	0.17
C45149	0.19	0.19	0.19	0.18	0.19
C45151	0.13	0.13	0.12	0.13	0.12
C45154	0.16	0.16	0.15	0.15	0.15
C45159	0.15	0.15	0.15	0.15	0.15
C46004	0.47	0.47	0.47	0.47	0.47
C46036	0.40	0.40	0.40	0.40	0.40
C46131	0.20	0.20	0.20	0.21	0.20
C46132	0.40	0.40	0.40	0.40	0.40
C46145	0.32	0.31	0.32	0.32	0.32
C46146	0.15	0.15	0.15	0.16	0.15
C46147	0.44	0.44	0.44	0.44	0.45
C46181	0.11	0.11	0.11	0.11	0.11
C46183	0.64	0.63	0.64	0.64	0.64
C46184	0.42	0.42	0.42	0.42	0.42
C46185	0.42	0.40	0.42	0.41	0.42
C46204	0.38	0.37	0.38	0.38	0.38
C46205	0.40	0.40	0.40	0.39	0.40
C46206	0.37	0.36	0.37	0.37	0.37
C46207	0.43	0.43	0.43	0.43	0.43
C46208	0.40	0.40	0.40	0.40	0.40

Table 6.6: MAEs of clustering-based forecasting models for the ECCC buoy dataset collection in the context of 12-hour prediction.

BuoyId	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
C44137	0.57	0.56	0.57	0.57	0.57
C44139	0.46	0.45	0.46	0.46	0.47
C44150	0.61	0.60	0.61	0.61	0.61
C44258	1.49	1.49	1.50	1.22	1.17
C46004	0.64	0.64	0.64	0.64	0.64
C46036	0.52	0.52	0.52	0.53	0.53
C46132	0.46	0.46	0.46	0.46	0.46
C46145	0.77	0.80	0.77	0.78	0.79
C46147	0.59	0.60	0.59	0.59	0.59
C46183	1.90	1.84	1.90	1.87	1.82
C46184	0.52	0.52	0.52	0.52	0.52
C46185	0.79	0.82	0.79	0.79	0.81
C46204	0.53	0.55	0.53	0.53	0.53
C46205	0.55	0.55	0.55	0.55	0.55
C46206	0.48	0.50	0.48	0.48	0.48
C46207	0.49	0.50	0.49	0.49	0.49
C46208	0.46	0.46	0.46	0.46	0.45

Table 6.7: MAEs of clustering-based forecasting models for the ECCC buoy dataset collection in the context of 1-hour prediction under extreme events.

Clustering-based Forecasting Results: Extreme Events (1-hr)

Table 6.7 shows the clustering-based forecasting model results for significant wave height under extreme events with a forecasting horizon of 1-hour. For evaluation under extreme events, we selected data points from the test set where the target variable, significant wave height, is greater than 6 meters for each buoy. These data points were used to assess the performance of the forecasting models.

Region-Based Forecasting Results: Regular(1-hr,6-hr,12-hr) and Extreme Events(1-hr)

Table 6.8 present the forecasting results of the local models and the universal model for forecasting horizons of 1-hr, 6-hr, 12-hr in regular conditions and for 1-hr under extreme events, respectively. In Table 6.8, N/A indicates no data points in the test dataset where the significant wave height exceeds 6 meters. All reported errors are in meters, rounded to two decimal places (centimeters), in alignment with the reporting standard used by MEDS, from which we collected the ECCC buoy dataset.

6.3.1 Observations from Clustering Results

All the clustering algorithms had some common patterns for certain buoys across the clustering algorithms. Notably, all algorithms grouped the buoys in the Great Lakes as one cluster, while those located near the shallow end of the Pacific Ocean formed another cluster. K-means, OPTICS, and DBSCAN algorithms clustered the buoys in the medium to deep ends of the Pacific Ocean and the buoys in the deep ends of the Atlantic into a single cluster. The clustering results of Agglomerative almost matched our assessment done in exploratory data analysis (refer to Section 5.1). In that section, we noted similar wave patterns among buoys in the shallow Pacific Ocean regions. The buoys positioned in the medium to deep end of the Pacific Ocean exhibited similar wave patterns. We also noticed differences in wave ranges between buoys in the Pacific and Atlantic Oceans.

6.3.2 Observations and Analysis from Forecasting Model Results

Comparison of local and clustering-based models in regular conditions

We compared the results of the local forecasting models with those of clustering-based forecasting models obtained from each clustering algorithm tested. We created summary tables for each forecasting horizon summarizing the number of cases in which a particular clustering-based forecasting model yielded lower MAE, higher MAE, or equal MAE compared to the local models. Data from all 28 buoys were available for testing when testing the models under regular conditions.

Examining the data in the summary Table 6.9, we see that for 1-hour forecasts under regular conditions, clustering-based forecasting models showed equal performance for 24 buoys and reported lower MAE for three buoys compared to the local models. The clustering-based models exhibited a lower performance in a single case than local models.

Considering these results, we cannot draw a conclusion about the superiority or inferiority of clustering-based models compared to local models, given the equal

BuoyId	1-hr	6-hr	12-hr	1-hr extreme events
C44137	0.14	0.29	0.42	0.57
C44139	0.13	0.27	0.40	0.47
C44150	0.13	0.27	0.39	0.61
C44258	0.10	0.21	0.31	1.17
C45132	0.06	0.14	0.20	N/A
C45136	0.05	0.11	0.15	N/A
C45139	0.04	0.09	0.12	N/A
C45143	0.05	0.12	0.17	N/A
C45149	0.06	0.14	0.19	N/A
C45151	0.06	0.10	0.12	N/A
C45154	0.06	0.12	0.15	N/A
C45159	0.05	0.11	0.15	N/A
C46004	0.17	0.33	0.47	0.64
C46036	0.16	0.29	0.40	0.53
C46131	0.07	0.17	0.24	N/A
C46132	0.14	0.28	0.40	0.46
C46145	0.12	0.23	0.31	0.80
C46146	0.06	0.13	0.18	N/A
C46147	0.17	0.32	0.44	0.59
C46181	0.05	0.10	0.12	N/A
C46183	0.24	0.46	0.62	1.80
C46184	0.16	0.30	0.42	0.52
C46185	0.12	0.28	0.40	0.81
C46204	0.13	0.27	0.38	0.54
C46205	0.15	0.28	0.39	0.56
C46206	0.13	0.26	0.36	0.48
C46207	0.16	0.30	0.42	0.49
C46208	0.15	0.28	0.40	0.46

Table 6.8: MAEs of the region-based clustering forecasting models for the ECCC buoy dataset collection for forecasting horizons of 1-hr, 6-hr, 12-hr and for 1-hr under extreme events.

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	3	3	3	3	3
# with higher MAE	1	1	1	1	1
# with equal MAE	24	24	24	24	24

Table 6.9: Summary of Clustering-Based Forecasting Models: Comparing MAE with Local Model for 1-hr Prediction.

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	11	11	10	11	10
# with higher MAE	7	6	5	5	4
# with equal MAE	10	11	13	12	14

Table 6.10: Summary of Clustering-Based Forecasting Models: Comparing MAE with Local Model for 6-hr Prediction.

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	10	12	11	11	12
# with higher MAE	9	8	6	7	6
# with equal MAE	9	8	11	10	10

Table 6.11: Summary of Clustering-Based Forecasting Models: Comparing MAE with Local Model for 12-hr Prediction.

MAE reported for 24 buoys. However, we will further investigate this improvement through statistical testing to find whether this observed improvement significantly deviates from equivalent performance. Although it is not our main focus in terms of measuring success, it is worth mentioning that the performance achieved by 28 local models is attained by a maximum of 4 (dependent on the clustering method employed) clustering-based forecasting models, with only one exception.

From the summary Tables 6.10 and 6.11, we observe that the performance of the clustering-based forecasting models for 6-hour and 12-hour forecasts under regular conditions varies depending on the chosen clustering algorithm. Three clustering-based models reported lower Mean Absolute Error (MAE) than local models in 11 out of 28 instances for 6-hour forecasts. In the case of 12-hour forecasts, two clustering-based models reported lower MAE in 12 out of 28 cases.

While MAE represents the absolute error magnitude, let us look at the distribution of the difference in MAEs between clustering-based and local forecasting models using a boxplot for 6-hour and 12-hour forecasts. The boxplot visualizes the spread of improvements or deterioration in forecasting performance achieved by the clustering-based models compared to the local models. To compute this difference, we subtract the MAE of the local model from the MAE of clustering-based models. A greater

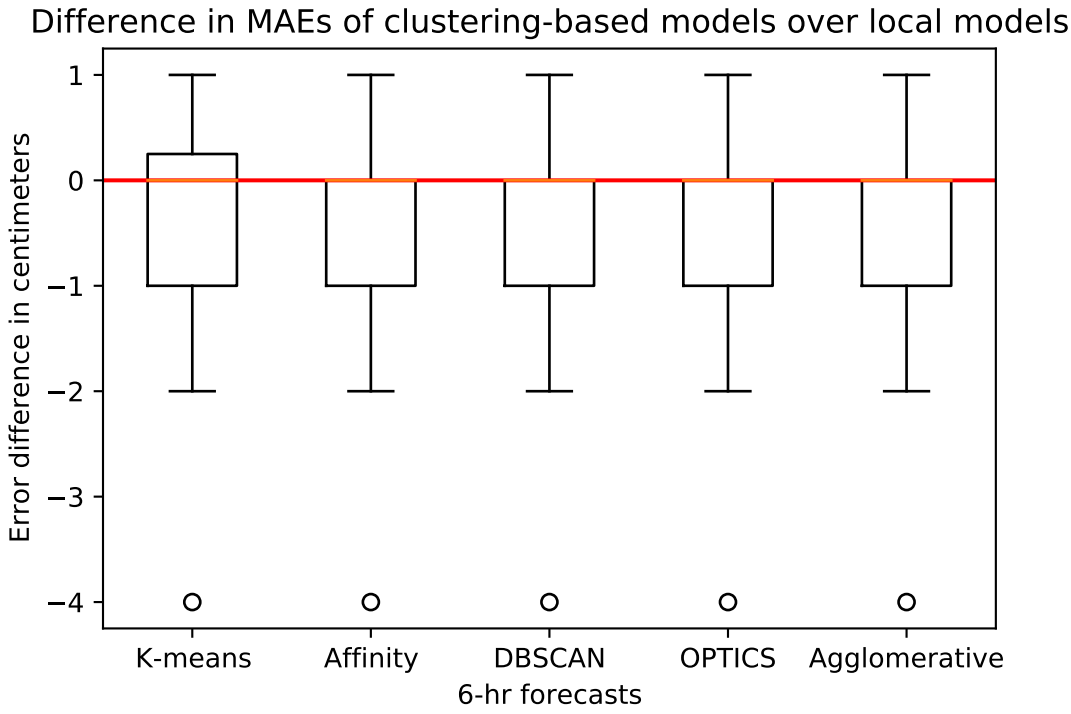

Figure 6.6: MAE Difference Boxplot: Clustering vs. Local Models (6-hr). Negative values indicate clustering-based model superiority.

concentration of values towards the negative side indicates improved performance of the clustering-based models (in terms of MAE). We used a centimeter scale for the boxplots since all the differences are within the range of 1 meter, indicating that the variations in errors are relatively small. Also, the dots in the boxplots represent data points that exhibit unusually high differences in MAE between the clustering-based and local forecasting models.

From the boxplot shown in Figure 6.6, the presence of data points below the median (whiskers line indicated in orange) indicates instances where the clustering-based models reported MAE and the data points above the whisker line indicates instances where clustering-based models reported higher MAE compared to local models. We observe that clustering-based forecasting models display a maximum error difference of 1 centimeter compared to local models, while the improvements by clustering-based models over local models span up to 2 centimeters, with an outlier reaching 4 centimeters.

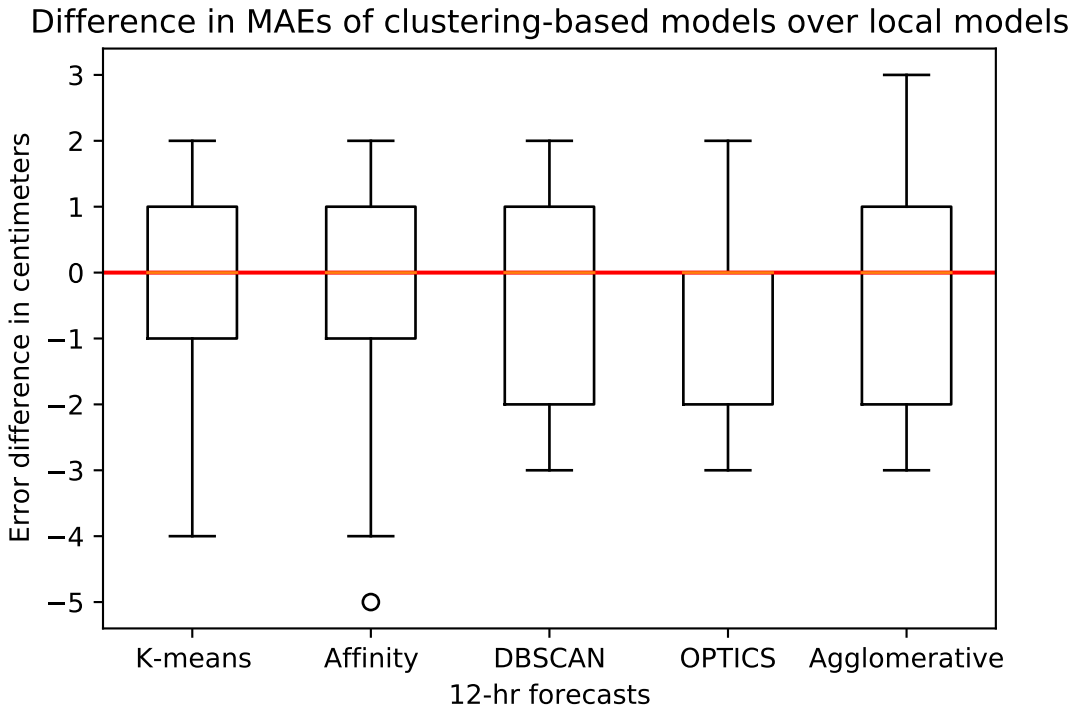

Figure 6.7: MAE Difference Boxplot: Clustering vs. Local Models (12-hr). Negative values indicate clustering-based model superiority.

From the boxplot shown in Figure 6.7, we can observe that clustering-based forecasting models exhibit a maximum error difference of up to 2 centimeters compared to local models. In the case of the agglomerative clustering algorithm, the difference reaches up to 3 centimeters. The improvements made by clustering-based models over local models span from 1 to 4 centimeters. Additionally, there is an outlier of 5 centimeters in one of the cases for the Affinity clustering algorithm.

Clustering-based model	p-value(1-hr)	p-value(6-hr)	p-value(12-hr)
K-means	0.26	0.13	0.33
Affinity	0.26	0.11	0.15
DBSCAN	0.26	0.09	0.21
OPTICS	0.26	0.07	0.20
Agglomerative	0.26	0.06	0.16

Table 6.12: Comparing Clustering-Based Models with Local Models: Wilcoxon Signed-Rank Test Results

While box plots and summary tables provide visual and numerical insights into the distribution and trends of MAE differences, they do not explicitly determine whether these differences can be considered significant. To assess the significance of the improvement between clustering-based models and local models, we turn to the Wilcoxon signed-rank test for each of these forecasting horizons. We hypothesize that clustering-based and local forecasting models perform equally in this test. If the p-value is below 0.05, we reject this hypothesis; otherwise, we accept the hypothesis.

Analyzing the Wilcoxon test results (in Table 6.12), we observe that p-values are above 0.05. This indicates that we accept the hypothesis of equivalent performance between clustering-based and local models. From the summary tables, we see that the performance differences between clustering-based models and local models are varied. In some cases, clustering-based models showed improvements over local models, and in others, it is the opposite.

Analyzing the instances where clustering-based models showed lower accuracy than local models, we identified four buoys for which clustering-based models consistently reported higher MAEs. Among these buoys, three are located in distinct lakes, and we observed that all clustering algorithms clustered these buoys together. This might indicate that the wave patterns of these buoys in these distinct lakes may not align well with those of other Great Lakes buoys when clustered together. A similar observation was made for a buoy in the Pacific region. This clustering of buoys with diverse patterns might have contributed to the reduced forecasting accuracy in these specific cases. Overall, based on the forecasting results and the Wilcoxon test results, we conclude that employing clustering-based models for predicting significant wave height under regular conditions, across various lead times (1, 6, and 12 hours), did not exhibit improved performance across the 28 ECCC buoys compared to local models.

Comparison of local and clustering-based models in extreme events

Table 6.13 summarizes the cases in which a particular clustering-based forecasting model yields lower MAE, higher MAE, or equal MAE compared to the local models for extreme events. When evaluating all the forecasting models under extreme events, only 17 buoys had test data with significant wave heights exceeding 6 meters.

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	11	11	10	10	10
# with higher MAE	1	1	1	1	1
# with equal MAE	5	5	6	6	6

Table 6.13: Summary of Clustering-Based Forecasting Models: Comparing MAE with Local Model for 1-hr Prediction under extreme events.

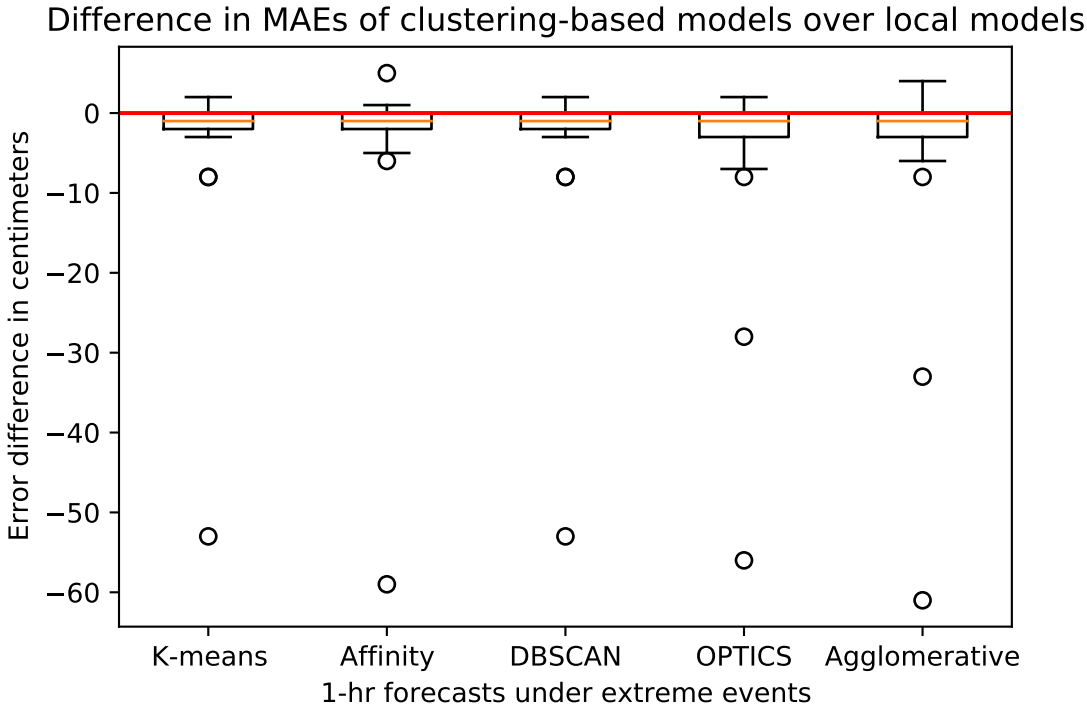

Figure 6.8: MAE Difference Boxplot: Clustering vs. Local Models (1-hr, Extreme Events). Negative values indicate clustering-based model superiority.

In the case of extreme events, for 1-hr forecasts, comparing the values from Tables 6.7, and 6.2, we observe the results are better (in terms of less MAE) than under regular conditions. From Table 6.13 clustering-based models have reported lower MAE for 10 out of 17 cases compared to local models.

We applied the same approach to visualize the distribution of MAE differences between clustering-based and local forecasting models as we did when comparing them under regular conditions. From the boxplot presented in Figure 6.8, we observe that the extent of error difference in instances of outliers for clustering-based

Clustering-based model	p-value(1-hr)
K-means	0.01
Affinity	0.03
DBSCAN	0.02
OPTICS	0.01
Agglomerative	0.01

Table 6.14: Comparing Clustering-Based Models with Local Models under Extreme Events: Wilcoxon Signed-Rank Test Results

forecasting models, irrespective of the clustering algorithm used, implies better performance compared to the local models for certain buoys. The maximum error where clustering-based models showed degraded performance is limited to 5 centimeters (as indicated by the actual data) compared to the local models.

To evaluate whether a significant difference in improvement exists between clustering-based forecasting models and local models during extreme events, we turn to the Wilcoxon signed-rank test. Examining the results of the Wilcoxon test displayed in Table 6.14, we note that the p-values are consistently below the threshold of 0.05. This finding indicates a statistically significant distinction in improvement between the clustering-based forecasting models and the local models during extreme events.

Upon comparing the clustering-based and local forecasting models and considering the insights derived from the forecasting results, summary tables, and Wilcoxon test results, we can conclude that leveraging the collective information from multiple related buoys proves to be significantly advantageous during extreme events.

Comparison of universal and clustering-based models in regular conditions

Using a similar procedure as employed in the comparison between local models and clustering-based models, we proceeded to evaluate the performance of the universal forecasting model against clustering-based forecasting models generated by each tested clustering algorithm. We summarize the number of cases in which a particular clustering-based forecasting model yielded lower MAE, higher MAE, or equal MAE

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	3	4	3	4	5
# with higher MAE	0	0	0	0	0
# with equal MAE	25	24	25	24	23

Table 6.15: Summary of Clustering-Based Forecasting Models: Comparing MAE with Universal Model for 1-hr Prediction.

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	8	8	10	10	10
# with higher MAE	1	1	2	2	1
# with equal MAE	19	19	16	16	17

Table 6.16: Summary of Clustering-Based Forecasting Models: Comparing MAE with Universal Model for 6-hr Prediction.

compared to the universal model. Data from all 28 buoys were available for testing when testing the models under regular conditions.

Examining the data presented in the summary Table 6.15, we see that, in the context of 1-hour forecasts under regular conditions, clustering-based forecasting models reported equal MAE in 24 out of 28 instances and reported lower MAE in all other 4 cases, compared to the universal model.

From summary Tables 6.16, 6.17 for 6-hr and 12-hr forecasts under regular conditions, the performance of clustering-based forecasting models varies based on the clustering algorithm used. We find that in 9 out of 28 instances for 6-hour forecasts and 13 out of 28 instances for 12-hour forecasts, the clustering-based models reported

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	13	14	12	12	14
# with higher MAE	6	3	7	4	7
# with equal MAE	9	11	9	12	7

Table 6.17: Summary of Clustering-Based Forecasting Models: Comparing MAE with Universal Model for 12-hr Prediction.

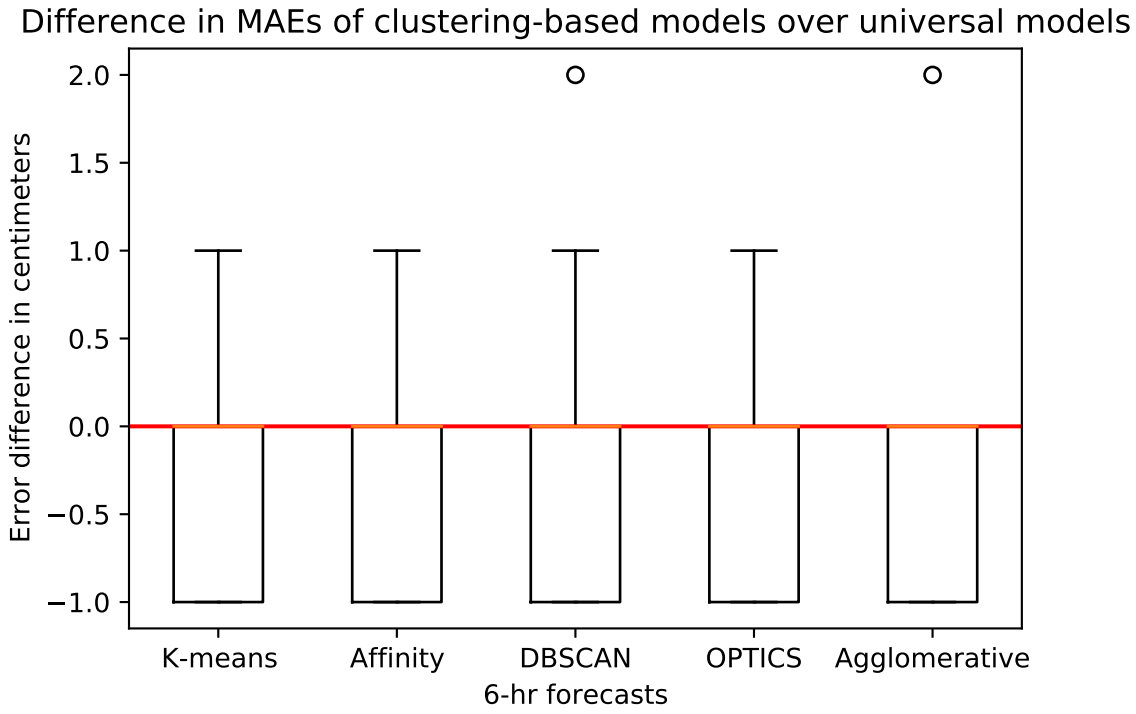

Figure 6.9: MAE Difference Boxplot: Clustering vs. Universal Model (6-hr). Negative values signify clustering-based model superiority

lower MAE, compared to the universal model. Also, the Agglomerative clustering-based models reported only one instance with higher error than the universal model.

We use boxplots for visualizing the distribution of improvements or deterioration in forecasting performance achieved by the clustering-based models compared to the universal model. To calculate this difference, we applied the same process we used for comparing local models with clustering-based models.

From the boxplot depicted in Figure 6.9, it is noticeable that most clustering-based forecasting models exhibit a variation in their performance improvements or deteriorations within a range of +1 to -1 centimeters. However, two specific clustering-based models, DBSCAN and Agglomerative, show a deviation in their performance deterioration, with a difference of 2 centimeters in one of the cases.

From the boxplot depicted in Figure 6.10, the error differences between affinity clustering-based forecasting models and the universal model are notable. The maximum error difference for affinity clustering-based models compared to the universal model is 1 centimeter. The improvements achieved by affinity clustering-based models

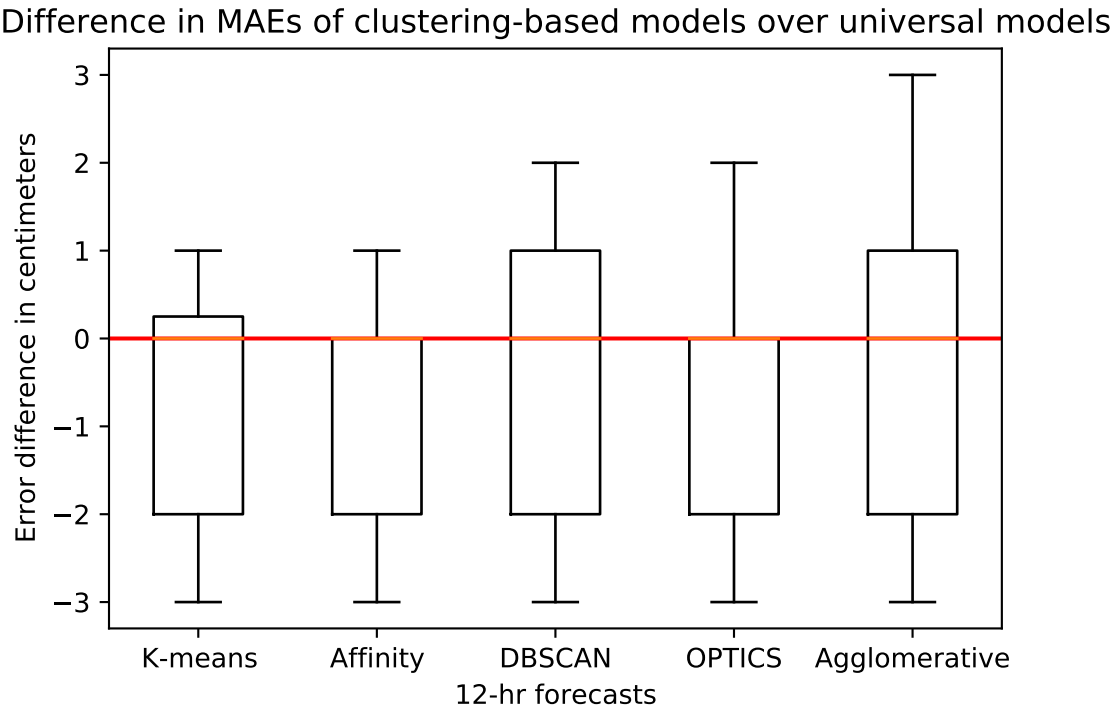

Figure 6.10: MAE Difference Boxplot: Clustering vs. Universal Model (12-hr). Negative values signify clustering-based model superiority

over the universal model are within a range of 1 to 3 centimeters.

While box plots and summary tables offer visual and numerical insights into the distribution and trends of MAE differences, they do not directly indicate whether these differences hold statistical significance. To determine whether there exists a significant difference in improvement between clustering-based forecasting models and the universal model, we will turn to the Wilcoxon signed-rank test for each forecasting horizon.

From the Wilcoxon test results shown in Table 6.18, we observe that the p-values are below 0.05 in 12 out of 15 cases, indicating a significant difference in improvement between the clustering-based models compared to the universal model. However, there are a few exceptions, such as K-means and DBSCAN for 1-hour predictions and DBSCAN for 6-hour predictions, where the p-values exceed 0.05. The differences in these cases was not statistically significant.

Clustering-based model	p-value(1-hr)	p-value(6-hr)	p-value(12-hr)
K-means	0.083	**0.019**	**0.028**
Affinity	**0.045**	**0.019**	**0.003**
DBSCAN	0.083	0.071	**0.037**
OPTICS	**0.045**	**0.02**	**0.031**
Agglomerative	**0.025**	**0.032**	**0.043**

Table 6.18: Comparing Clustering-Based Models with Universal Model: Wilcoxon Signed-Rank Test Results

When examining the actual Mean Absolute Errors (MAEs) generated by clustering-based models with the MAEs of the universal model, we observed that each clustering-based model exhibited either lower or equal MAEs across all Great Lakes buoys for 6-hour forecasts, and lower MAEs for 12-hour forecasts for the same buoys. This observation particularly highlights that training forecasting models on unrelated series leads to reduced overall accuracy. This is evident from the forecasting outcomes of the universal model for these specific buoys.

In the context of 6-hour forecasts, we observe that DBSCAN clustering reported higher error for one of the Atlantic buoys (C44258), and reported the same MAEs as OPTICS clustering-based models in other scenarios. In the case of DBSCAN, this specific buoy was classified as an outlier, resulting in the clustering-based model relying solely on its information for forecasting, while the universal model incorporates information from multiple buoys. This particular scenario highlights the value of utilizing cross-series information from multiple related buoys to improve the accuracy of predictions.

Overall, from the Wilcoxon test results, we can conclude that three clustering-based models demonstrate a significant improvement difference compared to the universal model under regular conditions. However, two clustering-based forecasting models deviate from this trend.

Comparison of universal and clustering-based models in extreme events

We created the summary Table 6.19 summarizing the number of cases in which a particular clustering-based forecasting model yielded lower MAE, higher MAE, or

	K-means	Affinity	DBSCAN	OPTICS	Agglomerative
# with lower MAE	12	8	11	12	11
# with higher MAE	2	5	2	2	1
# with equal MAE	3	4	4	3	5

Table 6.19: Summary of Clustering-Based Forecasting Models: Comparing MAE with Universal Model for 1-hr Prediction under extreme events.

equal MAE compared to the universal model under extreme events. When evaluating the models under extreme events, only 17 buoys had test data with significant wave heights exceeding 6 meters.

In the case of extreme events, for 1-hr forecasts, from Table 6.7, and Table 6.3, we observe the results are better (in terms of less MAE) than under regular conditions. From Table 6.19, clustering-based models reported lower MAE for 11 out of 17 cases compared to the universal model.

We visualize the spread of the difference in MAEs between clustering-based and universal forecasting models using a boxplot for 1-hour forecasts. To compute this difference, we subtract the MAE of the universal model from the MAE of clustering-based models. A concentration of values toward the negative side indicates the improved performance of the clustering-based models (in terms of MAE).

The boxplot shown in Figure 6.11 displays the extent of improvement, as measured by MAE, achieved by clustering-based models compared to the universal model. We observe that this improvement ranges from 1 to 5 centimeters, with a few outliers exceeding 5 centimeters for Affinity and Agglomerative clustering-based models.

Outliers at 30 centimeters for K-means, Affinity, and DBSCAN indicate a specific scenario where the universal model exhibited significantly improved forecasting performance compared to the clustering-based models. The reason is that K-means and Affinity clustered a particular buoy within the Great Lakes region. As the wave ranges in the Great Lakes are typically less than 5 meters (confirmed by actual data and shown in Histogram 5.3), the model did not have sufficient data on extreme events beyond the data collected from this specific buoy. For the OPTICS and Agglomerative clustering-based models, this specific buoy is grouped with other buoys with data about extreme events, leading to lower errors. This observation further highlights

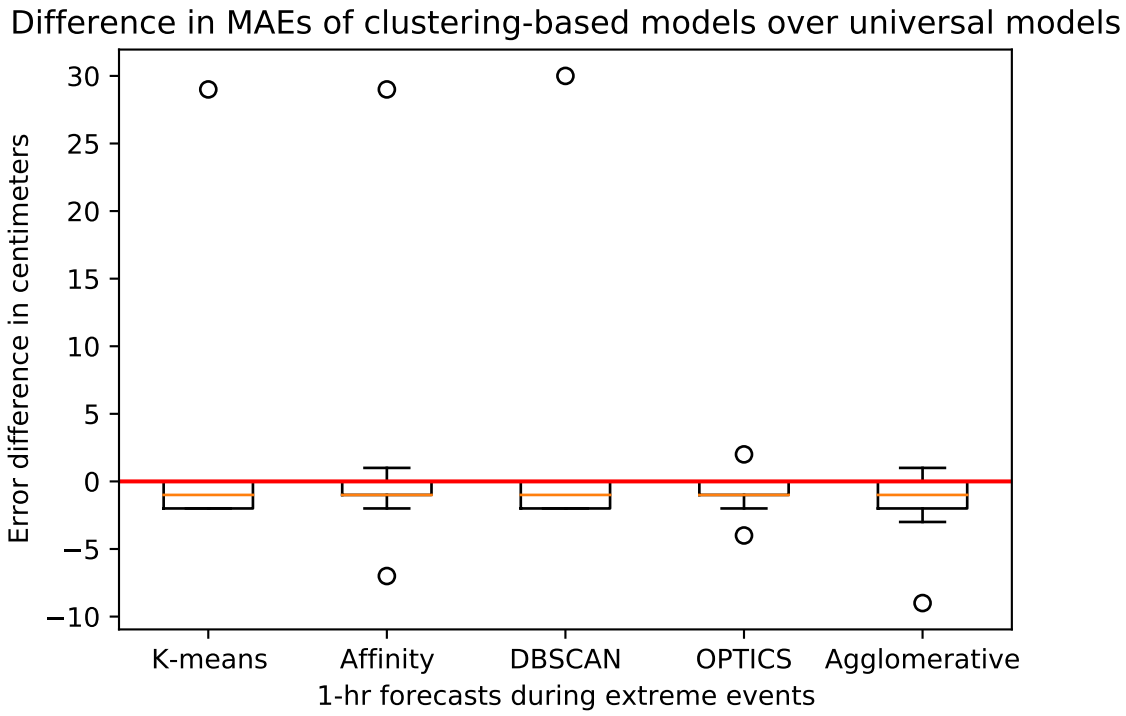

Figure 6.11: MAE Difference Boxplot: Clustering vs. Universal Model (1-hr, Extreme Events). Negative values indicate clustering-based model superiority.

Clustering-based model	p-value(1-hr)
K-means	0.02
Affinity	0.203
DBSCAN	0.02
OPTICS	0.01
Agglomerative	0.004

Table 6.20: Comparing Clustering-Based Models with Universal Model in extreme events: Wilcoxon Signed-Rank Test Results

the importance of using data from multiple related buoys. We also see that the improvements in forecasting performance by the Agglomerative clustering-based models extend up to 10 centimeters, and the maximum error where clustering-based models showed deteriorated performance is up to 1 centimeter compared to the universal models.

To determine whether there is a significant difference in the improvement achieved by clustering-based forecasting models over the universal model during extreme events, we will examine the results of the Wilcoxon signed-rank test.

From the Wilcoxon test results shown in Table 6.20, we observe that the p-values are below 0.05 in all cases, except for Affinity, indicating a significant difference in improvement between the clustering-based models compared to the universal model.

In summary, leveraging collective information from multiple related buoys is beneficial during extreme events in comparison to both local models and the universal model. In comparison with the universal model, the improvements achieved by clustering-based forecasting models hold statistical significance in both regular and extreme conditions, except for cases like DBSCAN clustering under regular conditions and Affinity clustering under extreme conditions. Clustering-based models, when compared to local models under normal conditions, did not exhibit improved forecasting accuracy.

BuoyId	Agg.1-hr	RB.1-hr	Agg.6-hr	RB.6-hr	Agg.12-hr	RB.12-hr
C46004	0.17	0.17	0.33	0.33	0.47	0.47
C46036	0.16	0.16	0.29	0.29	0.40	0.40
C46131	0.07	0.07	**0.15**	0.17	**0.20**	0.24
C46132	0.14	0.14	0.28	0.28	0.40	0.40
C46145	0.12	0.12	0.23	0.23	0.32	0.31
C46146	0.06	0.06	**0.12**	0.13	**0.15**	0.18
C46147	0.17	0.17	0.32	0.32	0.45	0.44
C46181	0.05	0.05	**0.09**	0.10	**0.11**	0.12
C46183	0.24	0.24	0.47	**0.46**	0.64	**0.62**
C46184	0.16	0.16	0.30	0.30	0.42	0.42
C46185	0.12	0.12	0.28	0.28	0.42	**0.40**
C46204	0.13	0.13	0.27	0.27	0.38	0.38
C46205	0.15	0.15	0.29	0.28	0.40	**0.39**
C46206	0.13	0.13	0.26	0.26	0.37	**0.36**
C46207	0.15	0.16	0.30	0.30	0.43	**0.42**
C46208	0.15	0.15	0.28	0.28	0.40	0.40

Table 6.21: Comparing Agglomerative Clustering-Based Models with Region-Based Models for Pacific Buoys under regular conditions

Comparison of automated clustering-based and region-based clustering models in regular and extreme conditions

The forecasting results of region-based clustering models, which are clustered manually according to the region of buoy location, are similar to those of Agglomerative clustering forecasting models. This alignment is anticipated since Agglomerative clustering has clustered the buoys in a similar manner as the region-based approach, except for the Pacific region where Agglomerative has created two clusters: one for shallow-end buoys and the other for remaining Pacific buoys.

Given the similarity in clustering patterns between Agglomerative and region-based clustering, we concentrate our analysis of forecasting results on Pacific buoys and compare the Agglomerative clustering-based models with region-based clustering models. This aids in determining whether the automated clustering procedure offers any advantages compared to the intuitive approach of grouping clusters based on geographical regions.

From Table 6.21, we observe that the 1-hour forecasting results for both Agglomerative clustering-based models and region-based clustering models are identical. Looking at the 6-hour forecasts, we observe that Agglomerative clustering-based models reported lower MAEs than region-based clustering models for 3 buoys (highlighted in bold). These 3 buoys are the shallow end buoys, grouped together by Agglomerative clustering. The 12-hour forecasts also show that the shallow-end buoys have lower MAEs. In the 6-hour and 12-hour forecasts, there are instances where region-based clustering models reported lower MAEs. From these observations, we can deduce that for the shallow-end buoys, grouping them separately from other Pacific buoys yields improved forecasting accuracy by 1 or 2 centimeters. The cases where region-based clustering models reported lower MAE (by 1 or 2 centimeters) are the buoys in the mid-range Pacific. It is plausible that data from the shallow-end buoys in these scenarios, particularly in dealing with shorter wave conditions where these shallow-end buoys provide more data on such waves.

Chapter 7

Conclusion and Future work

7.1 Conclusion

Our study aims to enhance the forecasting performance of significant wave height for each buoy within the ECCC buoy dataset collection. We hypothesized that leveraging cross-series information from multiple related buoys can improve the accuracy of the forecasting models. To test our hypothesis, we took an approach that involved clustering buoys with similar data. Subsequently, we trained a forecasting model on each group of related buoys (sequentially training the model on each buoy within the group).

Based on our research findings, our hypothesis remains valid during extreme events when the significant wave height exceeds 6 meters. This holds for the 1-hour forecasting horizon, where clustering-based showed significantly improved performance compared to both local models and the universal model. Under regular conditions, on 28 buoys, and across forecasting horizons of 1 hour, 6 hours, and 12 hours, the clustering-based models exhibit significant improvements over the universal model. The Wilcoxon test results accepted the null hypothesis of equivalent performance, indicating there is no significant difference in improvements of clustering-based models compared to the local models.

In conclusion, our study serves as an initial exploration into the potential benefits of using cross-series information to enhance the forecasting performance of significant wave height predictions. While our approach did not improve all 28 ECCC buoys datasets, it did demonstrate the value of cross-series information in predicting rare events more accurately. Under regular conditions, while there are instances where clustering-based models exhibited lower MAE values, the extent of improvement is relatively small. To identify buoys with related buoy data, we employed feature-based clustering and tested five distinct clustering algorithms to identify clusters. This method is found advantageous for forecasting only during extreme events.

7.2 Limitations

Our study has certain limitations at every step of the process. The TSFEL library offered the capability to extract features across temporal, statistical, and spectral domains, but we only utilized features from the temporal domain due to variations in the lengths of the buoy data. This limitation might have hindered any informative features from statistical and spectral domains which capture different characteristics of the time series and thereby lead to different clustering.

While the buoy data comprises ten distinct fields, excluding duplicates, buoy station id, and reporting time, we selectively focused on seven fields that we believed to influence wave patterns and environmental conditions. Although this selection was made based on our understanding of the buoy data, it is plausible that other unexplored fields could also contribute valuable information in the clustering process.

During the training of the forecasting models, we relied solely on lags of significant wave height as input features. While this approach has been widely used, considering additional relevant features in the training process could potentially enhance the forecasting performance. Also, we only tested one regression model and the results could be different from other models. Our forecasting horizons were focused on 1 hour, 6 hours, and 12 hours during regular conditions and for 1 hour during extreme events. The results could be different for other forecasting horizons. During extreme events, the dataset contained a relatively small number of observations due to the infrequency of such events. It becomes challenging to extend these results to other scenarios or purposes given the rare occurrence of these events. Also, we did not compare our results to numerical models.

We have specifically tested our approach on 28 ECCC buoy data collection. The results could be different for other datasets. Although significant in some cases, the MAE differences were small for certain buoys and hence it should be checked with the domain experts on whether they would make a difference in real-time usage.

Modifying any of the configurations used in our study could yield diverse outcomes, as adjustments to one aspect may trigger a cascade of effects. For example, altering the features extracted using the TSFEL library could lead to distinct feature vectors, subsequently affecting the formation of clusters. The input data for forecasting models would change with different clusters, ultimately influencing the results.

7.3 Possible Future Directions

In our study, clustering plays an important role as it determines the data on which forecasting models are trained, directly impacting the subsequent forecasting outcomes. Although this approach has shown some enhancements in forecasting results, it is worth considering alternate methods for grouping similar buoy data. Exploring such alternatives could potentially amplify the effectiveness of the clustering process and subsequently improve the accuracy of forecasting results.

One alternative approach worth considering is the application of spatial clustering, where buoy locations and geographical characteristics are considered to form clusters. This spatial clustering may provide valuable insights into the regional similarities and differences in wave patterns and environmental conditions. Another approach to explore is using deep learning techniques like graph neural networks, which can directly learn the similarity between buoys from the raw data, bypassing the feature extraction step.

Forecasting over extended horizons, such as beyond 1 hour for extreme events and beyond 12 hours for regular conditions, helps in a better understanding of predictive capabilities. Also, comparing the forecasting results with those of numerical models aids in validating the reliability of our clustering-based approach. Also exploring a diverse range of regression models beyond LightGBM could be beneficial.

Future research can also focus on models that use clustering information as an input feature for forecasting models. This exploration can help determine if leveraging clustering information more effectively improves prediction results. Additionally, it would be valuable to explore the feasibility of forecasting the next 12-hour values at the same time, rather than predicting one-hour intervals independently.

Expanding this research to other domains and other buoy datasets would provide valuable insights into the generalizability and broader applicability of the proposed approach. However, to apply the same approach, the dataset should exhibit a wide range of patterns, trends, or behaviors, and the time series should be related based on some underlying similarity.

Bibliography

Nov 2016. URL https://public.wmo.int/en/media/press-release/19-meter-wave-sets-new-record-highest-significant-wave-height-measured-buoy.

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series clustering – a decade review. *Information Systems*, 53:16–38, 2015. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2015.04.007.

Ahmed Ali, Ahmed Fathalla, Ahmad Salah, Mahmoud Bekhit, Esraa Eldesouky, et al. Marine data prediction: an evaluation of machine learning, deep learning, and statistical predictive models. *Computational Intelligence and Neuroscience*, 2021, 2021.

Mumtaz Ali, Ramendra Prasad, Yong Xiang, and Ravinesh C. Deo. Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms. *Renewable and Sustainable Energy Reviews*, 132:110003, 2020a. ISSN 1364-0321. doi: https://doi.org/10.1016/j.rser.2020.110003. URL https://www.sciencedirect.com/science/article/pii/S136403212030294X.

Mumtaz Ali, Ramendra Prasad, Yong Xiang, and Ravinesh C Deo. Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms. *Renewable and Sustainable Energy Reviews*, 132:110003, 2020b.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2): 49–60, 1999.

Kasun Bandara, Christoph Bergmeir, and Slawek Smyl. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert systems with applications*, 140:112896, 2020.

Marília Barandas, Duarte Folgado, Letícia Fernandes, Sara Santos, Mariana Abreu, Patrícia Bota, Hui Liu, Tanja Schultz, and Hugo Gamboa. Tsfel: Time series feature extraction library. *SoftwareX*, 11:100456, 2020. ISSN 2352-7110. doi: https://doi.org/10.1016/j.softx.2020.100456.

Manuel Barange, John G. Field, Roger P. Harris, Eileen E. Hofmann, R. Ian Perry, and Francisco Werner. *Marine Ecosystems and Global Change*. Oxford University Press, 02 2010. ISBN 9780199558025. doi: 10.1093/acprof:oso/9780199558025.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199558025.001.0001.

Jadran Berbić, Eva Ocvirk, Dalibor Carević, and Goran Lončar. Application of neural networks and support vector machine for significant wave height prediction.

Oceanologia, 59(3):331–349, 2017. ISSN 0078-3234. doi: https://doi.org/10.1016/j.oceano.2017.03.007. URL https://www.sciencedirect.com/science/article/pii/S0078323417300271.

P. Berkhin. *A Survey of Clustering Data Mining Techniques*, pages 25–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-28349-2. doi: 10.1007/3-540-28349-8_2.

Shreya Bhattacharyya, Souvik Majumder, Papiya Debnath, and Manash Chanda. Arrhythmic heartbeat classification using ensemble of random forest and support vector machine algorithm. *IEEE Transactions on Artificial Intelligence*, 2(3):260–268, 2021. doi: 10.1109/TAI.2021.3083689.

C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 9780387310732. URL https://books.google.ca/books?id=kTNoQgAACAAJ.

N. Booij, R. C. Ris, and L. H. Holthuijsen. A third-generation wave model for coastal regions: 1. model description and validation. *Journal of Geophysical Research: Oceans*, 104(C4):7649–7666, 1999. doi: https://doi.org/10.1029/98JC02622. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98JC02622.

G. E. P. Box, Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, 1994.

Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing*, 307:72–77, 2018.

Jacob Cohen. edition 2. statistical power analysis for the behavioral sciences, 1988.

Robin Davidson-Arnott, Bernard Bauer, and Chris Houser. *Introduction to Coastal Processes and Geomorphology*. Cambridge University Press, 2 edition, 2019. doi: 10.1017/9781108546126.

Paresh Chandra Deka and R Prahlada. Discrete wavelet neural network approach in significant wave height forecasting for multistep lead time. *Ocean Engineering*, 43:32–42, 2012.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

Tingyan Deng, Yu Zhao, Shunxian Wang, and Hongjun Yu. Sales forecasting based on lightgbm. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 383–386, 2021. doi: 10.1109/ICCECE51280.2021.9342445.

Luca Di Persio and Nicola Fraccarolo. Energy consumption forecasts by gradient boosting regression trees. *Mathematics*, 11(5), 2023. ISSN 2227-7390. doi: 10.3390/math11051068. URL https://www.mdpi.com/2227-7390/11/5/1068.

Pradnya Dixit, Shreenivas Londhe, and Yogesh Dandawate. Removing prediction lag in wave height forecasting using neuro-wavelet modeling technique. *Ocean Engineering*, 93:74–83, 2015.

Gulustan Dogan, Meghan Ford, and Scott James. Predicting ocean-wave conditions using buoy data supplied to a hybrid rnn-lstm neural network and machine learning models. In *2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 1–6, 2021. doi: 10.1109/ICMLANT53170.2021.9690528.

ECCC. Eccc buoy data collection. [Online; accessed 2023-08-05].

Stergios Emmanouil, Sandra Gaytan Aguilar, Gabriela F. Nane, and Jan-Joost Schouten. Statistical models for improving significant wave height predictions in offshore operations. *Ocean Engineering*, 206:107249, 2020. ISSN 0029-8018. doi: https://doi.org/10.1016/j.oceaneng.2020.107249. URL https://www.sciencedirect.com/science/article/pii/S0029801820302997.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

Fan, Shuntao, Xiao, Nianhao, Dong, and Sheng. A novel model to predict significant wave height based on long short-term memory network. *Ocean Engineering*, 205:107298, 2020.

Jesuseyi Will Fasuyi, Jason Newport, and Chris Whidden. A machine learning redundancy model for the herring cove smart buoy. *Journal of Ocean Technology*, 15 (3), 2020.

Charles W. Finkl and Christopher Makowski. Coastal hazards, 2013. URL https://link.springer.com/book/10.1007/978-94-007-5234-4.

Fisheries and Oceans. Government of canada, Sep 2019. URL https://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/waves-vagues/index-eng.htm.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. doi: 10.1198/016214502760047131. URL https://doi.org/10.1198/016214502760047131.

Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

Ben D Fulcher. Feature-based time-series analysis. In *Feature engineering for machine learning and data analytics*, pages 87–116. CRC press, 2018.

Ben D. Fulcher and Nick S. Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12): 3026–3037, 2014. doi: 10.1109/TKDE.2014.2316504.

Ben D Fulcher and Nick S Jones. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell systems*, 5(5):527–531, 2017.

Claudio Hartmann, Martin Hahmann, Wolfgang Lehner, and Frank Rosenthal. Exploiting big data in time series forecasting: A cross-sectional approach. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2015. doi: 10.1109/DSAA.2015.7344786.

S. Hasselmann, K. Hasselmann, J. H. Allender, and T. P. Barnett. Computations and parameterizations of the nonlinear energy transfer in a gravity-wave specturm. part ii: Parameterizations of the nonlinear energy transfer for application in wave models. *Journal of Physical Oceanography*, 15(11):1378 – 1391, 1985. doi: https://doi.org/10.1175/1520-0485(1985)015<1378:CAPOTN>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/phoc/15/11/1520-0485_1985_015_1378_capotn_2_0_co_2.xml.

Trent Henderson and Ben D. Fulcher. An empirical evaluation of time-series feature sets. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 1032–1038, 2021. doi: 10.1109/ICDMW53433.2021.00134.

Leo H. Holthuijsen. *Waves in Oceanic and Coastal Waters*. Cambridge University Press, 2007. doi: 10.1017/CBO9780511618536.

Rob J Hyndman and Yeasmin Khandakar. Automatic time series forecasting: the forecast package for r. *Journal of statistical software*, 27:1–22, 2008.

Robin John Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Australia, 2nd edition, 2018.

James, Zhang Y, and O'Donncha F. A machine learning framework to forecast wave conditions. *Coastal Engineering*, 137:1–10, 2018. ISSN 0378-3839. doi: https://doi.org/10.1016/j.coastaleng.2018.03.004. URL https://www.sciencedirect.com/science/article/pii/S0378383917304969.

Tim Januschowski, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1):167–177, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

G. J. Komen, L. Cavaleri, M. Donelan, K. Hasselmann, S. Hasselmann, and P. A. E. M. Janssen. Dynamic and modelling of ocean waves., 1964.

John Joy Kurian, Marcel Dix, Ido Amihai, Glenn Ceusters, and Ajinkya Prabhune. Boat: A bayesian optimization automl time-series framework for industrial applications. In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 17–24. IEEE, 2021.

Guang Li, George Weiss, Markus Mueller, Stuart Townley, and Mike R Belmont. Wave energy converter control by wave prediction and dynamic programming. *Renewable energy*, 48:392–403, 2012.

T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11): 1857–1874, 2005.

SN Londhe, Shalaka Shah, PR Dixit, TM Balakrishnan Nair, P Sirisha, and Rohit Jain. A coupled numerical and artificial neural network model for improving location specific wave forecast. *Applied Ocean Research*, 59:483–491, 2016.

Martin Längkvist, Lars Karlsson, and Amy Loutfi. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognition Letters*, 42, 06 2014. doi: 10.1016/j.patrec.2014.01.008.

J MacQueen. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA, 1967.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022a. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2021.11.013. URL https://www.sciencedirect.com/science/article/pii/S0169207021001874. Special Issue: M5 competition.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m5 competition: Background, organization, and implementation. *International Journal of Forecasting*, 38(4):1325–1336, 2022b.

Ingo Mierswa and Katharina Morik. Automatic feature extraction for classifying audio data. *Machine learning*, 58(2):127–149, 2005.

D.S. Moore, G.P. McCabe, and B.A. Craig. *Introduction to the Practice of Statistics*. W.H. Freeman, 2012. ISBN 9781429286640. URL https://books.google.ca/books?id=nFEPKQEACAAJ.

Brett Naul, Stéfan van der Walt, Arien Crellin-Quick, Joshua S Bloom, and Fernando Pérez. cesium: Open-source platform for time-series inference. *arXiv preprint arXiv:1609.04504*, 2016.

Mohammad Reza Nikoo and Reza Kerachian. Wave height prediction using artificial immune recognition systems (airs) and some other data mining techniques. *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, 41:329–344, 2017.

NOAA. Noaa, 2005. URL https://www.noaa.gov/jetstream/ocean/waves.

EMC NOAA National Centers for Environmental Prediction. Noaa national centers for environmental prediction, wavewatch iii, Oct 2005. URL https://polar.ncep.noaa.gov/waves/wavewatch/.

Isadora Nun, Pavlos Protopapas, Brandon Sim, Ming Zhu, Rahul Dave, Nicolas Castro, and Karim Pichara. Fats: Feature analysis for time series. *arXiv preprint arXiv:1506.00010*, 2015.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

Tayeb Sadeghifar, Maryam Nouri Motlagh, Massoud Torabi Azad, and Mahdi Mohammad Mahdizadeh. Coastal wave height prediction using recurrent neural networks (rnns) in the south caspian sea. *Marine Geodesy*, 40(6):454–465, 2017. doi: 10.1080/01490419.2017.1359220. URL https://doi.org/10.1080/01490419.2017.1359220.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. Deepar: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020.

Stephen W. Scheff. Chapter 8 - nonparametric statistics. In Stephen W. Scheff, editor, *Fundamental Statistical Principles for the Neurobiologist*, pages 157–182. Academic Press, 2016. ISBN 978-0-12-804753-8. doi: https://doi.org/10.1016/B978-0-12-804753-8.00008-7. URL https://www.sciencedirect.com/science/article/pii/B9780128047538000087.

Artemios-Anargyros Semenoglou, Evangelos Spiliotis, Spyros Makridakis, and Vassilios Assimakopoulos. Investigating the accuracy of cross-learning time series forecasting methods. *International Journal of Forecasting*, 37(3):1072–1084, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2020.11.009. URL https://www.sciencedirect.com/science/article/pii/S0169207020301850.

Shahaboddin Shamshirband, Amir Mosavi, Timon Rabczuk, Narjes Nabipour, and Kwok-wing Chau. Prediction of significant wave height; comparison between nested grid numerical model, and machine learning models of artificial neural networks, extreme learning and support vector machines. *Engineering Applications of Computational Fluid Mechanics*, 14(1):805–817, 2020.

Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.

Tao Song, Runsheng Han, Fan Meng, Jiarong Wang, Wei Wei, and Shiqiu Peng. A significant wave height prediction method based on deep learning combining the correlation between wind and wind waves. *Frontiers in Marine Science*, 9, 2022. ISSN 2296-7745. doi: 10.3389/fmars.2022.983007. URL https://www.frontiersin.org/articles/10.3389/fmars.2022.983007.

Martin Štěpnička and Michal Burda. On the results and observations of the time series forecasting competition cif 2016. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6. IEEE, 2017.

Xiaolei Sun, Mingxi Liu, and Zeqian Sima. A novel cryptocurrency price trend forecasting model based on lightgbm. *Finance Research Letters*, 32:101084, 2020. ISSN 1544-6123. doi: https://doi.org/10.1016/j.frl.2018.12.032. URL https://www.sciencedirect.com/science/article/pii/S1544612318307918.

SWAMP. An intercomparison study of wind wave prediction models, part 1: Principal results and conclusions., 1985.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer Berlin Heidelberg, 1981. doi: 10.1007/978-3-642-93370-3_32. URL https://doi.org/10.1007/978-3-642-93370-3_32.

E.F. Thompson, D.L. Harris, and Coastal Engineering Research Center (U.S.). *A Wave Climatology for U.S. Coastal Waters*. Reprint (Coastal Engineering Research Center (U.S.))). U.S. Army, Corps of Engineers, Coastal Engineering Research Center, 1972. URL https://books.google.ca/books?id=I_16yPcGzlgC.

ML Tlachac, Veronica Melican, Miranda Reisch, and Elke Rundensteiner. Mobile depression screening with time series of text logs and call logs. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4, 2021. doi: 10.1109/BHI50953.2021.9508582.

Tolman. User manual and system documentation of wavewatch-iii version 1.18., 1999.

Tolman, Hendrik, et al. *User manual and system documentation of WAVEWATCH III (R) version 6.07*. 03 2019.

Juan Trapero, Nikolaos Kourentzes, and Robert Fildes. On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*, 66:299–307, 01 2014. doi: 10.1057/jors.2013.174.

Ruey S Tsay. *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, 2013.

M.J. Tucker. *Waves in Ocean Engineering: Measurement, Analysis, Interpretation*. Ellis Horwood Series in Applied Science and Industrial Techn. E. Horwood, 1991. ISBN 9780139329555. URL https://books.google.ca/books?id=mv5RAAAAMAAJ.

Shixiong Wang, Chongshou Li, and Andrew Lim. Why are the arima and sarima not sufficient, 2021.

Wenxu Wang, Ruichun Tang, Cheng Li, Peishun Liu, and Liang Luo. A bp neural network model optimized by mind evolutionary algorithm for predicting the ocean wave heights. *Ocean Engineering*, 162:98–107, 2018.

Granville Tunnicliffe Wilson. Time series analysis: Forecasting and control, 5th edition, by george e. p.nbsp;box, gwilym m.nbsp;jenkins, gregory c.nbsp;reinsel and greta m.nbsp;ljung, 2015. published by john wiley and sons inc., hoboken, new jersey, pp. 712. isbn: 978-1-118-67502-1. *Journal of Time Series Analysis*, 37(5): 709–711, 2016. doi: 10.1111/jtsa.12194.

David K Woolf, PG Challenor, and PD Cotton. Variability and predictability of the north atlantic wave climate. *Journal of Geophysical Research: Oceans*, 107(C10): 9–1, 2002.

Rui Xu and D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005. doi: 10.1109/TNN.2005.845141.

Shaobo Yang, Zhenquan Zhang, Linlin Fan, Tianliang Xia, Shanhua Duan, Chongwei Zheng, Xingfei Li, and Hongyu Li. Long-term prediction of significant wave height based on sarima model in the south china sea and adjacent waters. *IEEE Access*, 7:88082–88092, 2019. doi: 10.1109/ACCESS.2019.2925107.

G.G. Yen and K.-C. Lin. Wavelet packet feature extraction for vibration monitoring. *IEEE Transactions on Industrial Electronics*, 47(3):650–667, 2000. doi: 10.1109/41.847906.

Heesung Yoon, Seong-Chun Jun, Yunjung Hyun, Gwang-Ok Bae, and Kang-Kun Lee. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *Journal of hydrology*, 396(1-2): 128–138, 2011.

I. Young and A. Babanin. *Ocean Wave Dynamics*. World Scientific Publishing Company, 2020. ISBN 9789811208683. URL https://books.google.ca/books?id=8GLcDwAAQBAJ.

Nomenclature

B	Single Buoy Data
C	Cluster of buoy data
d	Dimensionality of the vector representing the data for buoy
E	Feature vector
g	function that applies a clustering algorithm on a set of feature vectors
k	number of clusters
M	Model trained on a cluster
n	Number of ECCC buoys
SWH	Average height of the highest one-third of the waves collected during a sampling period.

Appendix A

Further details on Buoy Dataset Collection

A.1 Exploration of Atlantic Buoy Data

In the Atlantic Ocean, the buoys are located at varying distances from the shore, with one buoy situated close to the shore and the other three in the deep ocean. Line graphs representing the significant wave heights recorded at two of the Atlantic buoys, one near the shore and the other in deep waters, can be found in Figures A.1 and A.2. Again, it is evident that the buoy positioned towards the deep end of the Atlantic reports a larger variation in wave height range compared to the one near the shore.

A.2 Information on Missing Buoy Data

The buoy data collected from sensors may contain missing values due to various factors, such as buoy damage or maintenance. Before proceeding with data preparation, we conducted an assessment of the missing data by calculating the percentage of missing values for each buoy. We considered an ideal scenario where the buoy reports data at regular 1-hour intervals. The results, along with the total number of data points and the corresponding percentage of missing data, are presented in Table A.1.

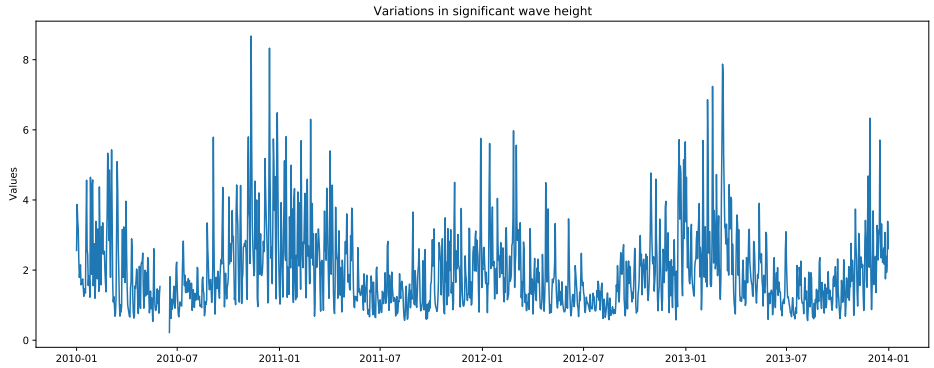

Figure A.1: Line graph showing significant wave height variations over 4 years for buoy C44150 in the deep Atlantic Ocean.

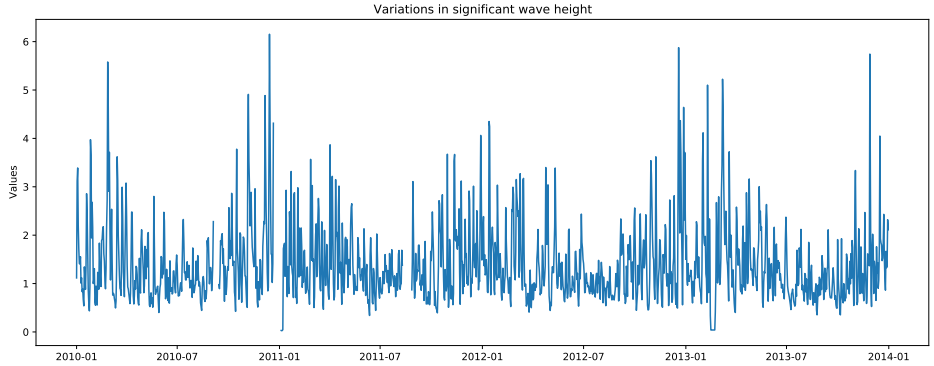

Figure A.2: Line graph displaying significant wave height variations over 4 years for buoy C44258 in the shallow Atlantic Ocean.

Buoy ID	DataPoints	MissingData(%)
c44137	59957	24
c44139	59746	24
c44150	56785	28
c44258	40629	48
c45132	55163	30
c45136	47367	40
c45139	54772	31
c45143	52006	34
c45149	45634	42
c45151	48090	39
c45154	44530	44
c45159	48761	38
c46004	55384	30
c46036	59300	25
c46131	70126	11
c46132	52467	33
c46145	68150	14
c46146	64715	18
c46147	64342	18
c46181	66405	16
c46183	63801	19
c46184	63877	19
c46185	63973	19
c46204	62668	21
c46205	60065	24
c46206	59356	25
c46207	64076	19
c46208	56600	28

Table A.1: Table describing the characteristics of each buoy including Buoy ID, current DataPoints, missing data %.