In Search

e-ISSN: 2580-3239
p-ISSN: 2085-7993
Volume 25 No. 1 April 2026

Recursive Multi-Step Forecasting of Ocean Wave Height Using Data from Oceanographic Wave Buoys

Nur Alamsyah[1)], Muhammad Nizar Haikal[1)], Arnold Ropen Sinaga[1)], Budiman[2)]
[1)] Sistem Informasi, Universitas Informatika dan Bisnis Indonesia
[2)] Informatika, Universitas Informatika dan Bisnis Indonesia
Email: nuralamsyah@unibi.ac.id; muhammad.nh22@student.unibi.ac.id; arnoldropen@unibi.ac.id; budiman@unibi.ac.id.

Accepted:	Published:
13 April 2026	28 April 2026

Abstract

Ocean wave fluctuations significantly impact maritime activities, coastal infrastructure, and disaster mitigation systems. However, predicting wave heights accurately remains a challenge due to the complex temporal dynamics of oceanographic data. This study proposes a recursive multi-step forecasting ap-proach using the XGBoost regression model to predict wave heights up to 30 days ahead. The dataset was obtained from wave buoys at Mooloolaba, Queensland, covering a 30-month observation period. After preprocessing and exploratory data analysis (EDA), relevant lag-based features were engineered to sup-port model learning. The XGBoost model was trained using past wave height, period, sea surface tempera-ture, and peak direction as predictive inputs. The results show that the model achieved RMSE values of 0.0851, 0.0899, and 0.0958 for step-ahead forecasts at t+1, t+2, and t+3, respectively. Visualization fur-ther confirms the model's ability to capture wave trends consistently. These findings suggest that the pro-posed method is effective and can be utilized to support early warning systems and decision-making in coastal areas.

Keywords: Ocean Buoy Data, Recursive Prediction, Wave Forecasting, XGBoost

1 INTRODUCTION

OCEAN wave forecasting plays a pivotal role in various maritime and coastal applications, including port management, offshore engineering, renewable energy operations, and early warning systems for extreme weather events [1]. Among the parameters that define ocean conditions, significant wave height (Hs) is considered one of the most critical indicators for operational safety and planning [2]. As sea conditions become increasingly unpredictable due to climate variability, the ability to forecast wave height accurately and promptly becomes essential for both economic and safety-driven decisions in marine environments [3].

Traditionally, ocean wave prediction has been dominated by physics-based numerical models such as WAVEWATCH III and SWAN [4]. These models solve complex hydrodynamic equations using meteorological inputs to simulate wave propagation and transformation [5]. While physically grounded, such models typically require substantial computational resources, are time-intensive, and may lack accuracy at high temporal resolutions or in nearshore zones [6]. Moreover, their adaptability to real-time operational forecasting is limited, especially in data-constrained environments [7].

In order to address these limitations, In past years, there's been a big increase in research exploring machine learning (ML) techniques for wave forecasting. These data-driven methods learn temporal and Notably, models such as Artificial Neural Networks (ANN), Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) networks have been applied to ocean wave forecasting with promising results [8]. For instance, Yang et al, (2025) applied ANN for significant wave height prediction in Indian coastal waters, while Wang et al. (2025) used LSTM for multi-step wave energy forecasting [9]. Similarly, ensemble methods like Random Forests and Gradient Boosting have shown superior performance in other environmental time series prediction tasks [10].

However, most of these studies adopt a direct multi-step forecasting approach, where individual models are trained separately for each forecast horizon (e.g., t+1, t+2, t+3). Although straightforward, this method often leads to inconsistencies between models, lacks temporal coherence, and increases the overall

computational load. Very few studies have explored the application of recursive multi-step forecasting in ocean wave prediction, where a single model is trained for one-step-ahead forecasting and then recursively applied to generate longer horizons. The technique has the advantage of maintaining model consistency, reducing complexity, and being well-suited for real-time deployment—yet its effectiveness in wave height prediction using real-world buoy data remains under-explored.

This study addresses that gap by proposing a recursive multi-step forecasting model using XGBoost, a powerful and efficient gradient boosting framework, to predict wave height using high-resolution buoy data collected every 30 minutes from the coastal region of Mooloolaba, Australia. In addition to the recursive approach, we implement a direct multi-step forecasting model as a baseline to evaluate the relative performance of both strategies.

The main contributions of this study can be summarized as follows. A recursive multi-step forecasting framework based on XGBoost was developed to address the challenge of short-term ocean wave height prediction using high-frequency buoy data. The recursive strategy was designed to reduce model complexity by utilizing a single-step model for multi-horizon prediction, enabling more efficient deployment in real-time operational systems. A comprehensive evaluation was conducted by comparing the recursive approach with a direct multi-step strategy, in which separate models are trained for each prediction horizon. The comparison demonstrated that the recursive method provides more stable and accurate predictions, particularly at medium-range horizons (t+2 and t+3). Furthermore, the study contributes to the growing body of literature by applying machine learning methods—specifically gradient boosting techniques—to real-world oceanographic data collected at 30-minute intervals. The results offer practical insights into the feasibility of using recursive learning for real-time coastal wave monitoring and highlight its potential for integration into early warning and marine decision-support systems.

2 LITERATURE REVIEW

Forecasting ocean wave height has traditionally relied on physics-based numerical models such as SWAN and WAVEWATCH III, which simulate wave dynamics using spectral energy balance equations and meteorological forcing [11]. While these models provide physically consistent forecasts, they often require substantial computational power and are less effective in providing high-resolution, near-real-time predictions, especially in dynamic coastal regions. In recent years, data-driven approaches have gained attention due to their ability to model nonlinear and temporal relationships from historical observations [12]. Various machine learning techniques have been explored in this domain. ANN have been widely applied to wave height prediction, with Khan et at. (2025) demonstrating their applicability for the Indian coastline [13]. SVR and Random Forests have also been utilized, showing promising results in short-term forecasting scenarios [14]. More recently, recurrent neural networks such as Long Short-Term Memory LSTM networks have shown superior performance in capturing temporal dependencies in marine and meteorological time series [15].

Several studies have also examined the use of ensemble models such as GBM and XGBoost for wave prediction tasks [16]. These methods offer advantages in terms of interpretability, robustness to missing data, and resistance to overfitting [17]. However, most existing machine learning studies adopt a direct forecasting strategy, where a separate model is trained for each forecast horizon. This approach, although straightforward, can lead to inconsistent performance across prediction steps and increased modeling complexity.

Despite the growing interest in multi-step forecasting, the application of recursive multi-step forecasting in ocean wave height prediction remains limited, particularly using real-world, high-frequency buoy datasets. Recursive strategies are commonly employed in other domains such as energy forecasting and traffic flow prediction, but their use in coastal oceanography is still underexplored. Moreover, few studies have conducted direct comparisons between recursive and direct multi-step strategies in this context, leaving a gap in understanding the trade-offs between accuracy, stability, and computational efficiency.

3 RESEARCH METHODS

The research methodology adopted in this study consists of a series of systematic stages to support the objective of forecasting ocean wave height using recursive multi-step prediction. The entire pipeline includes data acquisition, preprocessing, exploratory data analysis (EDA), model development, evaluation, and result visualization. Each stage plays a crucial role in ensuring the robustness and accuracy of the forecasting process. Figure 1 illustrates the proposed method employed in this study. The process begins with data collection from oceanographic wave buoys, followed by preprocessing steps such as handling missing values, time conversion, feature selection, and lag-based transformation. Subsequently, exploratory data analysis is conducted to understand temporal trends and inter-variable relationships. A forecasting model is then trained using the prepared dataset, with evaluation metrics used to assess model performance. The final

stage involves visualizing the prediction outcomes, comparing actual and predicted wave heights, and analyzing forecasting performance over multiple steps. The detailed explanation of each component in Figure 1 is described in the following sub-sections, starting from data collection.

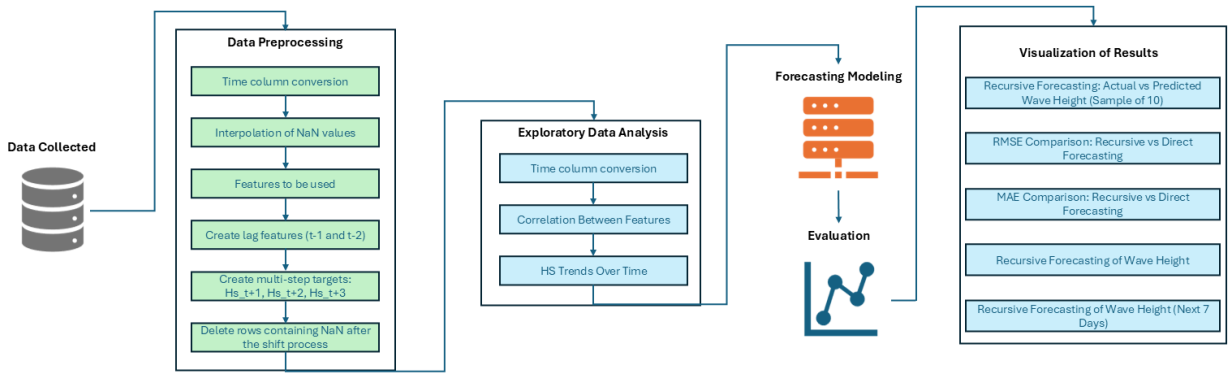

Figure 1. Proposed Method

3.1 Data Collected

The dataset employed in this study was collected from oceanographic wave measuring buoys stationed at Mooloolaba, Queensland, and is publicly accessible via the Queensland Government Data Portal. The dataset spans a period of 30 months, from January 2017 to June 2019, and provides a continuous record of wave activity in the region. These buoys record both measured and derived wave parameters, making the dataset suitable for time-series forecasting tasks. The data include a variety of physical oceanographic features, with the primary target variable being significant wave height (Hs). A summary of the dataset's key features used in this study is presented in Table 1.

Table 1. Description of Features Used in the Study

Feature Name	Description	Role
DateTime	Timestamp of wave observation	Index / Time
Hsig (m)	Significant wave height	Target (Hs)
Tp (s)	Peak wave period	Predictor
Dir (°)	Mean wave direction	Predictor
SST (°C)	Sea surface temperature	Predictor
Tz (s)	Average wave period	Predictor
DateTime	Timestamp of wave observation	Index / Time
Hsig (m)	Significant wave height	Target (Hs)
Tp (s)	Peak wave period	Predictor
Dir (°)	Mean wave direction	Predictor
SST (°C)	Sea surface temperature	Predictor
Tz (s)	Average wave period	Predictor
DateTime	Timestamp of wave observation	Index / Time
Hsig (m)	Significant wave height	Target (Hs)

3.2 Data Preprocessing

Data preprocessing plays a vital role in preparing the dataset for time-series forecasting tasks [18]. The first step involves converting the time column into a proper `datetime` format to enable chronological indexing and ensure accurate temporal alignment for lag and shift operations. Subsequently, missing values (NaN) in the dataset are addressed through linear interpolation. This method estimates the missing entries based on the known data points before and after the gap, ensuring continuity and preserving the overall trend and seasonality patterns of the wave data.

After handling missing values, a selection of relevant features is performed. This study utilizes significant wave height (Hs) as the main target variable, while features such as peak wave period (Tp), wave direction (Dir), sea surface temperature (SST), and average wave period (Tz) serve as predictors. To enable multi-step forecasting using recursive and direct strategies, lag features are created from the (Hs) variable. Specifically, (Hs_t-1) and (Hs_t-2) are derived and added as additional features to capture temporal dependencies. The lag features are defined as:

$$Hs_{t-1} = Hs(t-1), \quad Hs_{t-2} = Hs(t-2) \qquad (1)$$

Following the lag creation, multi-step targets are generated by shifting the `Hs` values forward by 1, 2, and 3 steps to form (Hs_t+1), (Hs_t+2), and (Hs_t+3), respectively. These targets are crucial for both recursive and direct forecasting strategies. The shifting process is mathematically represented as:

$$Hs_{t+k} = Hs(t+k), \quad k = 1,2,3 \qquad (2)$$

Finally, any rows that contain (NaN) values resulting from the shifting and lag operations are removed to ensure the integrity of the input dataset. This preprocessing pipeline ensures that the dataset is temporally structured, complete, and suitable for training forecasting models.

3.3 Exploratory Data Analysis (EDA)

This study conducted Exploratory Data Analysis (EDA) to gain a comprehensive understanding of the dataset prior to predictive modeling [19]. The EDA process consisted of three main stages: time column conversion, correlation analysis between features, and temporal trend visualization of significant wave height (Hs). The first step involved converting the time column from string format to a standardized datetime format. This transformation was crucial for enabling time-series operations such as trend analysis, data resampling (e.g., into daily or weekly intervals), and accurate forecasting based on temporal sequences. Moreover, converting the time column facilitated the construction of properly scaled time-based visualizations, allowing for clearer insights into the temporal behavior of oceanographic parameters.

The second step focused on examining the correlations between input features, including significant wave height (Hs), peak period (Tp), peak wave direction, and sea surface temperature (SST). Pearson correlation analysis was applied to assess the linear relationships among these variables. Identifying strong correlations with the target variable (Hs) helped to inform feature selection for the predictive model, ensuring that only relevant predictors were included and reducing the risk of overfitting. The final stage of EDA involved visualizing the temporal patterns of Hs over time. With wave height data recorded at 30-minute intervals, the time-series plot enabled the observation of both short-term fluctuations and long-term trends. Such insights are critical in understanding the dynamic nature of wave behavior and in designing a model capable of capturing these variations.

Overall, the EDA phase provided essential preliminary insights into the structure, relationships, and temporal dynamics of the dataset. These insights supported subsequent modeling steps and enhanced confidence in the dataset's quality and suitability for time-series forecasting of wave heights.

3.4 Forecasting Model and Evaluation

A forecasting model was constructed using the eXtreme Gradient Boosting Regressor (XGBoost), a decision-tree-based ensemble algorithm renowned for its scalability, robustness, and superior performance on structured time-series data [20]. The model was selected due to its ability to handle nonlinear relationships, incorporate regularization to avoid overfitting, and efficiently process high-dimensional input features generated from lagged variables. The forecasting task was framed as a univariate regression problem, aiming to predict the future values of significant wave height (Hs) based on historical and current information from meteorological and oceanographic parameters, including wave period (Tp), peak direction (Dir), and sea surface temperature (SST). The XGBoost model builds an ensemble of regression trees by iteratively minimizing a regularized loss function:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (4)$$

Here, l represents the mean squared error (MSE) between the actual and predicted wave heights, and $\Omega(f_k)$ is a regularization term that penalizes model complexity. This formulation encourages a balance between model fit and generalizability, especially in recursive forecasting where prediction errors may propagate over time. Once trained using 80% of the time-ordered dataset, the model was evaluated through recursive multi-step forecasting for a 30-day horizon. This approach generates the prediction for the next time step (Hs_{t+1}), which is then used as input to predict the subsequent step (Hs_{t+2}), and so on. Although recursive methods are susceptible to error accumulation, they are efficient for long-range predictions using a single-step model.

The performance evaluation of the forecasting model goes beyond conventional single-step error metrics by incorporating a cumulative recursive forecasting error (CRFE) to account for the compound effects of prediction inaccuracies in long-term recursive applications. This is particularly important in recursive forecasting, where the output at each time step becomes input for the next, potentially amplifying errors over time. To formalize this, we define the modified evaluation metric as:

$$\text{CRFE} = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{t}\sum_{k=1}^{t}(y_k - \widehat{y_k})^2\right) \quad (5)$$

Here, T is the total number of forecast steps (e.g., 30 days × 48 half-hour intervals), y_k is the true wave height at time stepk, and $\widehat{y_k}$ is the model's prediction. The inner summation captures the accumulated error up to each step, while the outer average provides an overall assessment of error propagation. This cumulative formulation penalizes early divergence and rewards temporal consistency across the prediction horizon. The XGBoost model's strength in minimizing loss through additive trees with regularization directly contributes to a controlled CRFE, as its learned structure is resistant to overfitting and better at generalizing under temporal shifts. Additionally, the use of lagged features as model inputs facilitates short-term memory of the system dynamics, which further helps in reducing recursive drift.

By utilizing CRFE alongside visual inspection of forecasted wave height trends, the evaluation demonstrates that the model maintains physical

plausibility and stability over extended periods. The forecast curve does not collapse to a static mean nor exhibit chaotic growth, indicating that the XGBoost model, with the proposed feature configuration, effectively mitigates error accumulation during recursive multi-step prediction.

3.5 Visualization Of Results

To effectively interpret the forecasting outcomes, a comprehensive visualization strategy was implemented. The first visualization compares actual versus predicted wave height (Hs) values using a sample of 10 time steps from the test set. This comparison illustrates the short-term accuracy of the recursive forecasting model and its ability to track wave height dynamics with minimal deviation. Furthermore, to evaluate the model's robustness over different forecasting approaches, we conducted a comparative analysis between recursive and direct forecasting strategies. This comparison was visualized using both RMSE and MAE plots. The RMSE comparison highlights the magnitude of error variance, while the MAE plot provides insight into the average deviation without the influence of large outliers. These visual comparisons reveal that although recursive forecasting may accumulate error over time, it performs competitively against the direct method, especially in short to medium horizons.

In addition, an extended recursive forecast of wave height over a 30-day horizon was visualized as a continuous time-series plot. The forecasted values show temporal consistency, avoiding erratic jumps or unrealistic plateaus, indicating the stability of the model even over prolonged periods. To complement this, a focused 7-day recursive forecast was also plotted to examine the model's behavior in shorter operational windows, which is highly relevant for maritime planning and oceanographic monitoring. Together, these visualizations not only validate the predictive capability of the XGBoost-based recursive model but also enhance interpretability by providing a tangible view of the model's performance across different scenarios and time spans.

4 RESULTS AND DISCUSSION

The Result subsection systematically outlines the outcomes of each experimental stage, including da-ta preprocessing, exploratory data analysis (EDA), forecasting model construction and evaluation, as well as the visualization of the predicted wave heights. Each phase is described in detail to illustrate how the model was developed and validated. The Discussion subsection provides a comparative analysis between the proposed forecasting approach and related studies in the literature, highlighting the strengths, limitations, and implications of this research in the broader context of ocean wave height prediction.

4.1 Result

The initial phase of the experiment involved a thorough data preprocessing stage to ensure the dataset was suitable for modeling. The original time-series dataset consisted of oceanographic measurements such as significant wave height (Hs), peak wave period (Tp), peak wave direction, and sea surface temperature (SST), recorded at 30-minute intervals. To facilitate forecasting, lag features were engineered by shifting each of the primary variables by one and two time steps (e.g., `Hs_t-1`, `Hs_t-2`) to capture temporal dependencies. In addition, the target variables were constructed for multi-step forecasting by shifting the Hs column forward to create `Hs_t+1`, `Hs_t+2`, and `Hs_t+3`, representing predictions for future wave heights. Any rows containing missing values resulting from the shifting process were subsequently removed to maintain dataset integrity. This preprocessing strategy ensured that the model received input features representing historical context while aiming to predict future values of wave height accurately.

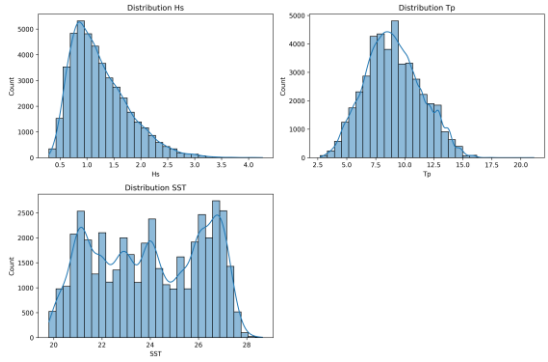

Figure 2. Distribution of Hs is Right-Skewed

To gain initial insights into the dataset, an exploratory analysis was conducted by visualizing the dis-tribution of key oceanographic features, namely significant wave height (Hs), peak period (Tp), and sea surface temperature (SST). As shown in Figure 2, the distribution of Hs is right-skewed, with the ma-jority of observations concentrated between 0.5 and 1.5 meters, suggesting that moderate wave heights are more frequent in the dataset. The peak period (Tp) exhibits a near-normal distribution centered around 8 to 10 seconds, indicating a stable wave cycle in most measurements. In contrast, SST shows a multimodal distribution with several local peaks between 20°C and 28°C, which may indicate seasonal variability or measurement inconsistencies over time. These distributional patterns

provide critical con-text for feature engineering and model development by highlighting the variability and central tenden-cies of each variable.

To further investigate the relationships between the key features, a correlation heatmap was generat-ed, as presented in Figure 3. The matrix reveals that the highest positive correlation exists between the peak period (Tp) and zero-crossing period (Tz), with a coefficient of 0.50, suggesting that these two temporal indicators of wave behavior are moderately interrelated. Additionally, a moderate correlation of 0.40 is observed between Hs and Tz, indicating that higher wave heights tend to occur with longer wave periods. Conversely, the correlation between SST and Tp is negative (-0.15), while the correla-tion between Hs and Tp is nearly negligible (0.02), implying minimal linear association between wave height and peak period. These findings inform the feature selection process and validate the inclusion of Tz and SST as potentially influential predictors in wave height modeling.

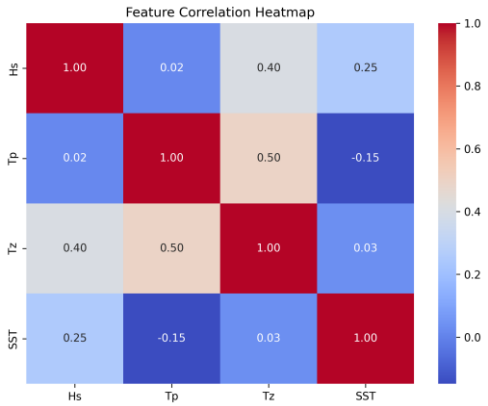

Figure 3. Correlation Heatmap

The temporal trend of significant wave height (Hs) during January 2019 is illustrated in Figure 4. The time series plot reveals distinct fluctuations in wave height, with values oscillating between approximately 0.5 meters and 2.2 meters throughout the month. Several short-term peaks are visible, suggesting transient meteorological or oceanographic influences such as wind bursts or storm events. Notably, the wave height exhibits cyclical rises and falls, which may reflect tidal or diurnal patterns. Sudden drops in wave height followed by rapid recoveries also suggest potential anomalies or brief periods of calm conditions. This visualization highlights the dynamic nature of wave behavior in coastal regions and underscores the importance of temporal features for accurate forecasting. Understanding these patterns provides critical insights for constructing time-aware predictive models.

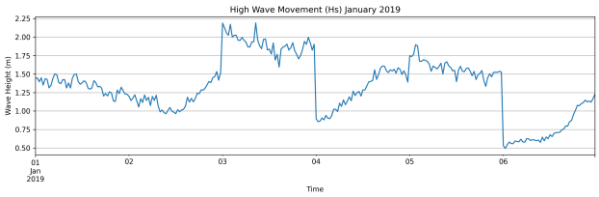

Figure 4. High Wave Movement

The evaluation of the forecasting model was conducted using the XGBoost (XGB) algorithm, which was selected due to its robustness in handling complex nonlinear relationships and its proven performance in time series forecasting. Two widely adopted performance metrics—Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE)—were employed to assess the model's predictive accuracy over the first three forecast horizons (i.e., t+1, t+2, and t+3). These metrics quantify the average magnitude of prediction errors, where lower values signify higher accuracy. As summarized in Table 2, the XGB model maintains strong predictive capability across short-term recursive steps, with only marginal increases in error as the forecasting horizon progresses. This pattern is consistent with typical time series behavior, wherein uncertainty naturally accumulates with each forward step.

Table 2. Evaluation Model

Forecast Step	RMSE	MAE
t+1	0.0851	0.0681
t+2	0.0899	0.0752
t+3	0.0958	0.0746
t+3	0.0958	0.0746

Figure 5 illustrates a comparative evaluation between recursive and direct forecasting approaches based on RMSE values over three forecasting steps (t+1 to t+3). The results clearly demonstrate that the recursive forecasting method consistently outperforms the direct approach across all horizons. Specifically, the RMSE for the recursive model remains relatively stable, increasing only slightly from 0.0851 at t+1 to 0.0958 at t+3. In contrast, the RMSE values for the direct method show a more pronounced upward trend, starting from 0.103 at t+1 and reaching 0.138 at t+3.

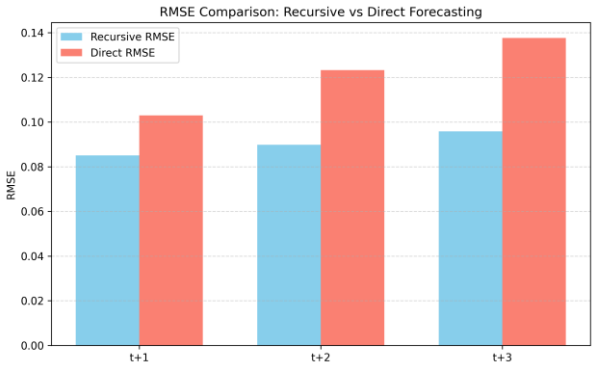

Figure 5. RMSE Comparison: Recursive vs Direct Forecasting

The comparison underscores the robustness and reliability of the recursive forecasting strategy when implemented with the XGBoost model. The recursive method appears more effective in capturing temporal dependencies in the dataset, thereby enabling better short-term wave height prediction. The widening performance gap between the two methods over time further suggests that recursive modeling offers improved generalization as the forecasting horizon extends. These findings validate the selection of the recursive approach as a preferable modeling strategy in this context.

Figure 6 presents the 30-day recursive forecasting results for significant wave height (Hs) using the XGBoost model. The forecast line demonstrates relatively stable wave height projections ranging between 1.80 and 2.10 meters, with moderate fluctuations. This indicates that the model effectively captures short-term temporal patterns and generates continuous predictions without abrupt shifts, suggesting a strong ability to maintain trend consistency over time.

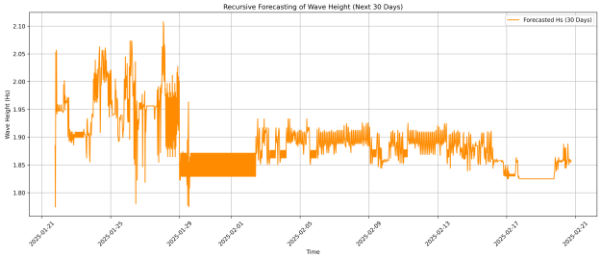

Figure 6. Recursive Forecasting Of Wave Height (Next 30 Days)

Figure 6 presents the results of recursive forecasting for significant wave height (Hs) over a 30-day horizon using the XGBoost model. The forecast begins from the end of the observed dataset and recursively uses each predicted value as input for the next time step. This approach simulates realistic fore-casting scenarios where future input data is not available, and predictions rely solely on previous out-puts. The graph shows that the predicted Hs values remain relatively stable, fluctuating within a rea-sonable range between 1.80 and 2.10 meters. There are minor variations across the 30 days, indicating the model's ability to capture short-term dynamics in wave height while maintaining prediction con-sistency.

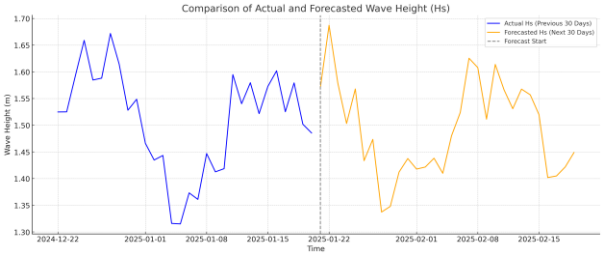

Figure 7. Comparison of Actual and Firecasted Wave

Compared to direct forecasting methods, which often suffer from accumulating errors across extend-ed horizons, the recursive method demonstrates more stable and reliable behavior. This is especially important in marine and coastal operations, where sudden fluctuations in forecasted wave height could lead to operational risks or inefficiencies. As shown in the previous comparison (Figure 5), recursive forecasting consistently achieved lower RMSE values across t+1 to t+3 steps, making it a more favora-ble option for short- to medium-term ocean wave forecasting.

4.2 Discussion

To assess the contribution and effectiveness of our approach, a comparative analysis was conducted with two relevant studies. The first, conducted by hu *et al*. (2022), applied XGBoost and LSTM models for one-step wave height prediction on Lake Erie [14]. Although their study confirmed that XGBoost can outperform traditional models with low mean absolute percentage error (MAPE), it was limited to short-term forecasting and did not explore recursive multi-step strategies. Meanwhile, Alfredo and Adytia (2022) proposed a hybrid deep learning model (CNN-GRU) to predict wave height at Pelabuhan Ratu, Indonesia [21]. Their approach supported a 30-day forecast horizon with good trend accuracy (RMSE = 1.88 meters), but it required substantial computational resources and access to ERA5 reanalysis data. In contrast, our recursive XGBoost model provides a balance between accuracy and computational efficiency. With RMSE scores of 0.0851 (t+1), 0.0899 (t+2), and 0.0958 (t+3), the model effectively captures short-term dynamics while remaining interpretable and lightweight. This supports the model's applicability for real-time coastal monitoring without the need for complex infrastructure.

Table 3. Comparative Summary of Related Studies

Study	Model Used	Forecast Horizon	RMSE
Hu et al. (2021)	XGBoost, LSTM	1 step (hourly)	N/A
Alfredo & Adytia (2022)	CNN-GRU Hybrid	30 days	1.88
This Study	Recursive XGBoost	3 steps (30 min)	0.0851–0.0958
Hu et al. (2021)	XGBoost, LSTM	1 step (hourly)	N/A
Alfredo & Adytia (2022)	CNN-GRU Hybrid	30 days	1.88

In contrast, our recursive XGBoost model provides a balance between accuracy and computational efficiency. With RMSE scores of 0.0851 (t+1), 0.0899 (t+2), and 0.0958 (t+3), the model effectively captures short-term dynamics while remaining interpretable and lightweight. This supports the model's applicability for real-time coastal monitoring without the need for complex infrastructure.

5 CONCLUSION

This study proposed a recursive multi-step forecasting model using XGBoost to predict ocean wave heights based on time-series buoy data collected from Mooloolaba. The model was trained with engineered lag features and evaluated across short-term horizons (t+1 to t+3), achieving strong performance with RMSE values between 0.0851 and 0.0958. Results from Exploratory Data Analysis (EDA) confirmed the significance of oceanographic features such as peak period, wave direction, and sea surface temperature. The visualization of results, including a 30-day recursive forecast, demonstrated the model's ability to generate stable and realistic predictions aligned with historical patterns.

For future work, we recommend expanding the forecasting horizon by integrating additional environmental variables such as wind speed, tidal information, and atmospheric conditions. Furthermore, incorporating uncertainty quantification—through techniques like quantile regression or Bayesian-based approaches—could improve the reliability of the forecasts. Finally, real-time deployment in a coastal monitoring system would provide tangible benefits in maritime operations, early warning systems, and climate resilience planning.

REFERENCES

[1] A. Halicki, A. Dudkowska, and G. Gic-Grusza, "Short-term wave forecasting for offshore wind energy in the Baltic Sea," Ocean Eng., vol. 315, p. 119700, 2025.

[2] L. Porlan-Ferrando, J. D. Nuñez-Gonzalez, A. Ulazia Manterola, N. Martinez-Iturricastillo, and J. V. Ringwood, "Maximum Individ-ual Wave Height Prediction Using Different Machine Learning Techniques with Data Collected from a Buoy Located in Bilbao (Bay of Biscay)," J. Mar. Sci. Eng., vol. 13, no. 4, p. 625, 2025.

[3] Y. Xie, H. J. Kim, Y. Yin, K. Liu, T. Smith, and J. K. Paik, "Enhancing the safety and sustainability of aging jacket-type offshore wind turbines in extreme weather conditions through digital healthcare engineering: a literature review," Ships Offshore Struct., pp. 1–28, 2025.

[4] A. Shadmani, M. R. Nikoo, and A. H. Gandomi, Ocean Wave Energy Technology: Fundamentals of Wave Farm Design. Springer Nature, 2025.

[5] B. J. Bethel, C. Dong, J. Wang, and Y. Cao, "An investigation of swell in the western Atlantic Ocean and Caribbean Sea," Ocean Dyn., vol. 75, no. 5, p. 44, 2025.

[6] C. G. Shankar and M. K. Cambazoglu, "Enhancing Storm Wave Predictions in the Gulf of Mexico: A Study on Wind Drag Parame-terization in WAVEWATCH III," J. Mar. Sci. Eng., vol. 13, no. 3, p. 403, 2025.

[7] M. A. Souames, L. A. Mohammedi, I. Zouaghi, A. Gunasekaran, S. Beldjoudi, and A. Laghouag, "Estimating import lead times using business intelligence and machine learning within the CRISP-DM framework: A case study in oil and gas services industry," IEEE Access, 2025.

[8] N. Alamsyah, A. Hendra, E. Setiana, T. P. Yoga, V. R. Danestiara, and others, "Improved Prediction Of Global Temperature Via LSTM Using ReLU Activation And Hyperparameter Optimization," in 2024 International Conference on Information Technology Research and Innovation (ICITRI), IEEE, 2024, pp. 41–46.

[9] J. Wang, D. Zhang, Q. Huang, and Z. Cui, "Multiple-step accurate prediction of wave energy: A hybrid model based on quadratic decomposition, SSA and LSTM," Int. J. Green Energy, vol. 22, no. 1, pp. 100–123, 2025.

[10] M. Sakib, S. Mustajab, and M. Alam, "Ensemble deep learning techniques for time series analysis: a comprehensive review, applica-tions, open issues, challenges, and future directions," Clust. Comput., vol. 28, no. 1, p. 73, 2025.

[11] J. Si, J. Wang, and Y. Deng, "Improving significant wave height prediction via temporal data imputation," Dyn. Atmospheres Oceans, p. 101549, 2025.

[12] S. Ozgen, A. Wu, and F. Ruiz, "Modeling approaches for data-driven model predictive control of acid gases in waste-to-energy plants," Waste Manag., vol. 204, p. 114902, 2025.

[13] A. R. Khan, M. S. B. Ab Razak, B. B. Yusuf, H. Z. B. M. Shafri, and N. B. Mohamad, "Harnessing artificial neural networks for coastal erosion prediction: A systematic review," Mar. Policy, vol. 178, p. 106704, 2025.

[14] R. S. Kumar, P. Meera, V. Lavanya, and S. Hemamalini, "Brown bear optimized random forest model for short term solar power fore-casting," Results Eng., vol. 25, p. 104583, 2025.

[15] Q. Yu, G. Yang, X. Wang, Y. Shi, Y. Feng, and A. Liu, "A review of time series forecasting and spatio-temporal series forecasting in deep learning," J. Supercomput., vol. 81, no. 10, pp. 1–48, 2025.

[16] S. Demir and E. K. Sahin, "An innovative machine learning approach for slope stability prediction by combining shap interpretabil-ity and stacking ensemble learning," Environ. Sci. Pollut. Res., pp. 1–17, 2025.

[17] F. Z. Che Rose, N. A. K. Rosili, and M. F. Marsani, "Comparison of machine learning model performance for predicting the climate variables in Johor Bahru, Malaysia," Sci. Rep., vol. 15, no. 1, p. 23465, 2025.

[18] N. Alamsyah, A. P. Kurniati, and others, "Event Detection Optimization Through Stacking Ensemble and BERT Fine-Tuning for Dynamic Pricing of Airline Tickets," IEEE Access, 2024.

[19] N. Alamsyah, B. Budiman, E. Setiana, V. C. Jennifer, and others, "THE ROLE OF L1 REGULARIZATION IN ENHANCING LOGISTIC REGRESSION FOR EGG PRODUCTION PREDICTION," JITK J. Ilmu Pengetah. Dan Teknol. Komput., vol. 10, no. 4, pp. 821–832, 2025.

[20] N. Alamsyah, B. Budiman, R. Nursyanti, E. Setiana, and V. R. Danestiara, "Approximate Bayesian Inference for Bayesian Confi-dence Quantification in DNA Sequence Classification Using Monte Carlo Dropout Approach," Innov. Res. Inform. Innov., vol. 7, no. 1, 2025.

[21] C. S. Alfredo, D. A. Adytia, and others, "Time series forecasting of significant wave height using GRU, CNN-GRU, and LSTM," J. RESTI Rekayasa Sist. Dan Teknol. Inf., vol. 6, no. 5, pp. 776–781, 2022.