Wave Height Estimation using Atmospheric Meteorological Data

Weston Brousseau, Andrew Bryceland, Precious Morenikeji, Jacob Munoz
University of Delaware
Department of Mechanical Engineering
{westwas, abryce, pmore, jmunoz}@udel.edu

December 15, 2025

This paper covers a variety of methods for using machine learning to predict significant wave height. Models trained and tested include linear regression, support vector regression, random forest, and a long-short term memory recurrent neural network. The dataset used for training and testing was from a buoy off the coast of Virginia Beach that records both wave and atmospheric data. The goal of this project was to develop a learned model capable of using atmospheric measurements to predict significant wave high so that buoys do not need expensive and unreliable inertial measurement units to monitor significant wave height.

1 Introduction

Accurate wave height measurements are critical for marine engineering, research, commerce, and recreation. When seagoing vessels, commercial or research, leave port they need to know whether their ship will be able to withstand whatever conditions await them out at sea. Accurate wave data is also valuable for offshore structures such as wind turbines and oil drilling platforms. These structures require constant maintenance and monitoring to make sure they can safely function at peak efficiency in this often destructive environment. Large hurricanes and tropical storms frequently pose a threat to swimmers, surfers, and fishermen during the late summer and early autumn storm season on the east coast. Accurate wave data is critical for local authorities to assess the severity of these storms and warn people to stay out of the water.

1.1 Background

The goal of our models in this paper is to predict the significant wave height which is defined as the average height of the tallest third of the waves in a wave field and the height of the wave is measured from trough to crest. The primary driver behind waves is the wind. When wind blows over the ocean high pressure regions are created behind preexisting ripples in the surface, turning ripples to waves. This pressure difference results wind constantly pushing the wave crest froward, transferring energy from the wind to the wave. The most simplified approximation of this effect is that the significant wave height is proportional to the wind speed ten meters above the surface, squared then divided by the acceleration of gravity [1].

While wind speed is the primary driver of wave height it is not the only driver. The rate at which energy from the wind is dissipates into the water surface is also dependent on the difference between the water surface temperature and air temperature [2]. When the water surface temperature is less than the air temperature a cooled laminar layer of air forms above the water's surface. This laminar layer is resistant to the vertical movement from the wind's turbulent flow, reducing the effectiveness of momentum transfer from wind to water. Conversely when the atmosphere is well-mixed and when the air temperature is less than the water temperature energy transfer from wind to water is more efficient and wind has an even bigger impact on significant wave height.

Other factors that are more specific to location and impact wave height in non-linear ways include depth of the water, bottom friction, current effects, and bathymetry features (i.e. canyons, ridges, sandbars, etc) [3].

1.2 Approach

During this project our goal was to find the best machine learning approach to estimate significant wave height based off atmospheric meteorological data. All of our data comes from a National Oceanic and Atmospheric Administration (NOAA) buoy of the coast of Virginia Beach which collects both wave and atmospheric data. The inputs to each of our models are wind speed, air temperature, water surface temperature, air pressure, day of year, and time of day (day of year and time of day are in sine/cosine function representations). We trained a total of four models a linear regression model, support vector regression (SVR) model, random forest model, and a long-short term memory recursive neural network (LSTM-RNN). After all models were trained our evaluation showed that the random forest model performed better on every error metric.

2 Related Work

Historically significant wave height prediction has been done using global physics-spectral models based models. Recently modern breakthroughs in machine learning have lead to the adaptation of data driven approaches to augment or replace traditional physics based models.

2.1 Physics Based Models

The standard in oceanographic and meteorological modeling since the 1980's has been the third generation wave model. Third-generation wave models are characterized by their ability to represent the wave spectrum without making *a priori* assumptions about the shape. All third generation models numerically solve the spectral wave action balance equation which tracks the change of wave action density across all of the Earth's oceans. The general form of the spectral wave action balance equation is

$$\frac{\partial N}{\partial t} + \frac{1}{\cos\phi}\frac{\partial}{\partial\lambda}(\dot{\lambda}N\cos\phi) + \frac{\partial}{\partial\phi}(\dot{\phi}N) + \frac{\partial}{\partial\sigma}(\dot{\sigma}N) + \frac{\partial}{\partial\theta}(\dot{\theta}N) = \frac{S}{\sigma}$$

Where N is the spectral wave action density, t is time, λ is longitude, ϕ is latitude, σ is relative angular frequency, θ is the wave propagation direction, and S is the net source term. The net source term represents all physical processes that effect wave energy and is calculated by the equation

$$S = S_{in} + S_{diss} + S_{nl}$$

Where S_{in} is the wind input, S_{diss} is the dissipated energy, and S_{nl} is the non-linear wave interactions [3].

The first widely used third-generation wave model was called the Wave Model (WAM) [4]. WAM set the foundation for models to come with its three source energy function and lack of assumptions relating to the wave spectrum shape.

Currently, the most widely used third-generation wave model is Wavewatch III (WW3) [5]. WW3 added global and basin scale forecasting and compared to earlier versions of WAM, WW3 updated model parameters for better accuracy at large ocean scales.

While Wavewatch III has become the standard for deep water wave modeling it struggles in nearshore environments due to additional shallow water effects. For shallow water modeling, Simulating Waves Nearshore (SWAN) [6] has been the standard since the 1990's.

2.2 Data Driven Models

The recent boom in machine learning has found its way to oceanography and significant wave height prediction models. Many research scientists have been exploring shifting from historically dominant physical models to data driven, machine learning models or augmenting physical models with machine learning to reduce bias from assumptions and simplifications.

2.2.1 Linear Regression

Despite their simplicity, linear regression models can serve an essential role as a baseline model that is easy to interpret compared to more complex structures like deep learning neural networks. A standard linear regression model uses a straight line equation to fit the data, assuming a linear relationship between input and output data. This may seem like an oversimplification, but a linear regression model can often provide a decent first-order estimate of wave dynamics. Previous work has also shown that using multiple linear regression with a variety of meteorological and time-lagged data can implicitly capture some of the non-linear effects that impact significant wave height [7].

2.2.2 Simple Vector Regression

Support vector machines (SVMs) were originally developed to solve the classification problem but a similar concept can be applied to numerical estimation. This technique which combines regression with support vector methods is termed support vector regression (SVR). SVRs are known to be robust in high-dimensional spaces making them a possible alternative to neural networks. Manhjoobi and Mosabbeb used buoy data from Lake Michigan to predict significant wave height [8]. Their results showed that the SVR model performed similarly to artificial neural networks with much lower computational demand. The results were close enough that with better parameter selection or in a different environment, in our case the ocean, a SVR model could outperform more advanced neural networks.

2.2.3 Tree Models

In a 2009 paper Etemad-Shahidi and Mahjoobi used an M5' model tree to predict significant wave height in Lake Superior [9]. In their study they found that the model tree was marginally more accurate than an artificial neural network (three-layer feed-forward network). In addition to being marginally more accurate the model tree approach is also more computationally efficient. The model tree was also able to explicitly represent the non-linear relationship between wind speed and wave height once the wind speed is fast enough. The M5 model tree uses linear regression to predict continuous values at each of its leaf nodes. In this report we use a random forest model instead of an M5 model. The random forest model uses bootstrapping to take random samples of the dataset and generate multiple trees. The random sampling of random forest generates a similar effect to dropout in neural networks, reducing model overfitting. This different tree-based approach could provide even better results with our dataset.The random sampling of random forest generates a similar effect to dropout in neural networks, reducing model overfitting.

2.2.4 Neural Networks

Most recent work on applying machine learning to wave forecasting has been centered around using deep learning neural network architectures. Dogan et al. trained a hybrid RNN-LSTM neural network to predict wave measurements from buoy data collected off the coast of Willington, NC [10]. Their prediction model showed an average validation accuracy of 81-83%. We are doing instantaneous estimation not prediction but a recent trend in the meteorological community towards using LSTM-RNNs led to us including our own in our model testing. Zang et al. used another hybrid LSTM this time combined with a convolutional neural network (CNN) to form a CNN-LSTM [11]. Their work used convolutional layers to extract spatial features and LSTM layers for temporal memory, resulting in a 70-73% reduction in RMSE compared to MLP, SVM, and random forest models.

3 Methods

To address the problem of Wave estimation, we leveraged the cyclic effect of seasonal weather on wave height by modeling time as a sinusoidal function [12].

Also cite the actual dataset?

3.1 Data Acquisition

All data used in this study were obtained from the National Oceanic and Atmospheric Administration (NOAA) National Data Buoy Center (NDBC) [13]. Measurements were collected from a buoy located off the coast of Virginia Beach, Virginia, which records both atmospheric and wave-related variables at regular time intervals.

The target variable for all models was significant wave height, defined as the average height of the highest one-third of waves in a given wave field. Input features consisted solely of atmospheric and temporal measurements, including wind speed, air temperature, sea surface temperature, atmospheric pressure, day of year, and time of day. Temporal variables were encoded using sine and cosine transformations to preserve periodicity and avoid artificial discontinuities at cycle boundaries [12].

The dataset spans multiple years of observations, providing a diverse range of environmental conditions suitable for training and evaluating data-driven wave height estimation models.

3.2 Data Preprocessing

Prior to model training, the raw buoy data were preprocessed to ensure consistency and numerical stability across all learning algorithms. Records containing missing or invalid measurements were removed. All remaining variables were converted to numerical formats suitable for machine learning models.

Continuous input features were standardized using z-score normalization to produce zero-mean, unit-variance inputs. Feature scaling is particularly important for kernel-based models such as SVR and distance-based learning methods, as unscaled variables can dominate similarity computations.

Temporal features representing day of year and time of day were transformed into sine and cosine components to maintain continuity across periodic boundaries. After preprocessing, the dataset was partitioned into training and testing subsets.

3.3 Linear Regression

The simplest model we used was a linear regression model. This model used both polynomial and multivariate linear regression to model the relationship between our data and wave heights. Additionally, our model utilized a time-based window of moving averages to infer more information about the history of our state. Because we can use the linear equation to run the linear regression model very quickly, we performed an ablation study of linear regression based on the size of our moving-average window and the highest order of our system. For clarity, the windows that were used were a moving average over a specific amount of time. The mini ablation studies can be seen in Figures 1 and 2.

Window Size	RMSE	MAE	R^2
1	0.52	0.39	0.50
6	0.48	0.37	0.58
12	0.47	0.36	0.60
24	0.48	0.37	0.57

Table 1: Ablation for linear regression 1st order only

Highest Order	RMSE	MAE	R^2
1	0.47	0.36	0.60
2	0.44	0.35	0.64
3	$1e^{17}$	$8e^{15}$	$-2e^{34}$

Table 2: Ablation based on order size for linear regression. Testing the highest order of the polynomial

Our mini ablation study shows us that the linear regression was most effective when it was considering a period over the last 12 hours with a 2nd degree relationship. This actually makes sense as the common relationship between wind speed and wave height is quadratic [**?**].

3.4 Long Short-Term Memory Network (LSTM)

To capture temporal dependencies in wave evolution, we trained a Long Short-Term Memory (LSTM) recurrent neural network. The model uses a fixed input window of the previous 6 hours of observations, producing a prediction of the current significant wave height. The input at each time step consists of the engineered atmospheric and time features described in Sections 3.1–3.2.

The network architecture includes two stacked LSTM layers with hidden size 32, followed by a fully connected head with two layers of sizes 64 and 1, respectively. A dropout rate of 0.3 was applied to reduce overfitting. Training was performed using mean squared error (MSE) loss and the Adam optimizer with a learning rate of 0.001 and batch size 32. The network was trained for 350 epochs, and the best-performing checkpoint on the validation/test evaluation during training was saved for final comparison in Section 4.

3.5 Random Forest Model

A Random Forest regressor was trained as a nonlinear, ensemble-based method for estimating significant wave height from atmospheric inputs. Random Forest models reduce variance relative to a single decision tree by averaging predictions across many trees trained on bootstrapped samples of the dataset.

The Random Forest was configured with 300 trees and a maximum depth of 15 per tree. Bootstrapping was enabled, and the minimum number of samples required to split an internal node was set to 5, with a minimum of 2 samples per leaf node. At each split, the number of candidate features considered was set to the square root of the total number of input features, a common heuristic for regression forests. Training was parallelized across available CPU cores, and a fixed random seed was used to ensure reproducibility. The final Random Forest configuration was evaluated against the other models using the metrics defined in Section 4.

3.6 Support Vector Regression

Support Vector Regression (SVR) was used as a nonlinear baseline to estimate significant wave height from atmospheric measurements. SVR extends the maximum-margin concept of support vector machines to continuous outputs by fitting a function that is as “flat” as possible while allowing deviations up to a specified tolerance. This formulation is well-suited for buoy measurements, which contain sensor noise and occasional outliers.

Given training pairs $\{(x_i, y_i)\}_{i=1}^{N}$, SVR estimates

$$f(x) = w^T \phi(x) + b, \tag{1}$$

where $\phi(\cdot)$ maps the input vector into a higher-dimensional feature space. The ε-insensitive loss introduces a tube of width 2ε around the regression function, within which errors are not penalized. The primal optimization problem is

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} w^2 + C \sum_{i=1}^{N} (\xi_i + \xi_i^*), \tag{2}$$

subject to $\mathrm{y}_i - f(x_i) \leq \varepsilon + \xi_i$,
$f(x_i) - y_i \leq \varepsilon + \xi_i^*$,
$\xi_i, \xi_i^* \geq 0$. Here, C controls the trade-off between model complexity (flatness) and the penalty on errors outside the ε-tube. Larger C prioritizes minimizing training error but can increase overfitting, while smaller C increases regularization. The parameter ε sets the tolerated error margin and influences sparsity: larger ε generally yields fewer support vectors and smoother predictions.

To capture nonlinear interactions between wind forcing, thermodynamic effects, and wave response, an RBF kernel was selected:

$$K(x_i, x_j) = \exp\left(-\gamma x_i - x_j^2\right), \tag{3}$$

The parameter γ controls the spatial scale of the kernel, with larger values emphasizing local structure and smaller values producing smoother global behavior. SVR performance is sensitive to feature scaling; therefore, all continuous input features were standardized prior to training to ensure that no single variable dominates the kernel distance computation. In this study, the regularization parameter was set to $C = 10$ and the ε-insensitive loss width was set to $\varepsilon = 0.1$, providing a balance between model smoothness and sensitivity to wave height variability. Unlike the linear regression model, the SVR formulation does not explicitly incorporate temporal windowing or moving-average features. Instead, SVR operates on instantaneous atmospheric inputs, relying on nonlinear

Model	RMSE	MAE	R^2	Train Time
Linear Reg.	0.44	0.35	0.64	0
SVR	0.54	0.62	0.47	1min 20 sec
Rand Forest	0.41	0.35	0.70	14 sec
LSTM	0.44	0.32	0.65	3 min 2 sec

Table 3: Evaluation of Models

similarity in feature space to implicitly capture relationships between environmental conditions and wave height. This design choice simplifies model structure but may limit the ability of SVR to leverage longer-term temporal dependencies present in wave evolution.

4 Experiments

While several evaluation metrics were considered, we ultimately selected Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), as these are well-suited for regression problems. Additionally, we include the training time of our model to show how the computation cost affects each model. The parameters differed, as shown in the individual sections for each model. We also provide an ablation study of all of the models to compare them in Table 3.

5 Conclusion

In conclusion, we were able to find multiple different models that performed moderately well with our data. However, due to the inherently non-linear relationships of fluid dynamics and weather, we were unable to see high confidence in our models. However, this is to be expected as weather is an inherently complex and chaotic system. That said, our best model, the random forest model, performed very well with an R^2 score of 0.70 and trained very quickly despite the size of our data set. Following the random forest model, the LSTM model was the next most accurate model; however, it scored the worst in terms of time to train. While this isn't an issue for such a small dataset, if we wanted to use any of these models at scale, LSTM would have the greatest issues, however. Following LSTM, our linear regression model performed closely in terms of performance, while the linear regression model used the normal equation, and so the training time was almost instantaneous. Lastly, the SVR exhibited lower predictive performance, possibly due to the inherent temporal nature of the data, as that model did not utilize a moving average like the linear regression model. All this data points to the importance of temporal data in our dataset.

Possible improvements to our models could look at this temporal connection as a way to improve efficacy. For example, one straightforward improvement we could make is to utilize 1-dimensional convolutions prior to the models. There are also numerous other avenues we could pursue, such as different architectures, but the convolution method is the most straightforward in terms of applicability.

References

[1] Thomas, T. J., and Dwarakish, G., 2015. "Numerical wave modelling–a review". *Aquatic procedia,* **4**, pp. 443–448.

[2] Kahma, K. K., and Calkoen, C. J., 1992. "Reconciling discrepancies in the observed growth of wind-generated waves". *Journal of Physical Oceanography,* **22**(12), pp. 1389–1405.

[3] Van der Westhuysen, A., 2012. "Modeling nearshore wave processes". In ECWMF Workshop on Ocean Waves, European Centre for medium-range weather forecasts, Reading, Vol. 1, pp. 50–61.

[4] Group, T. W., 1988. "The wam model—a third generation ocean wave prediction model". *Journal of physical oceanography,* **18**(12), pp. 1775–1810.

[5] Tolman, H. L., et al., 2009. "User manual and system documentation of wavewatch iii tm version 3.14". *Technical note, MMAB contribution,* **276**(220).

[6] Booij, N., Holthuijsen, L., and Ris, R., 1996. "The" swan" wave model for shallow water". In *Coastal engineering 1996*. ASCE, pp. 668–676.

[7] Ali, M., Prasad, R., Xiang, Y., and Deo, R. C., 2020. "Near real-time significant wave height forecasting with hybridized multiple linear regression algorithms". *Renewable and Sustainable Energy Reviews,* **132**, p. 110003.

[8] Mahjoobi, J., and Mosabbeb, E. A., 2009. "Prediction of significant wave height using regressive support vector machines". *Ocean Engineering,* **36**(5), pp. 339–347.

[9] Etemad-Shahidi, A., and Mahjoobi, J., 2009. "Comparison between m5 model tree and neural networks for prediction of significant wave height in lake superior". *Ocean Engineering,* **36**(15-16), pp. 1175–1181.

[10] Dogan, G., Ford, M., and James, S., 2021. "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), IEEE, pp. 1–6.

[11] Zhang, J., Luo, F., Quan, X., Wang, Y., Shi, J., Shen, C., and Zhang, C., 2024. "Improving wave height prediction accuracy with deep learning". *Ocean Modelling,* **188**, p. 102312.

[12] Young, I. R., 1999. "Seasonal variability of the global ocean wind and wave climate". *International Journal of Climatology: A Journal of the Royal Meteorological Society,* **19**(9), pp. 931–950.

[13] National Oceanic and Atmospheric Administration, 2025. National data buoy center (ndbc). `https://www.ndbc.noaa.gov`. Accessed December 2025.