TALLINN UNIVERSITY OF TECHNOLOGY

School of Information Technologies

Ilja Lastovko 213100IAIB

# STATISTICAL MACHINE LEARNING TECHNIQUES FOR WAVE SPECTRE ESTIMATION IN COASTAL SEAS

Bachelor's Thesis

Supervisor: Prof. Sven Nõmm
Ph.D.

Co-supervisor: Sander Rikka
Ph.D.

Co-supervisor: Mihhail Daniljuk
M.Sc.

Tallinn 2025

TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Ilja Lastovko 213100IAIB

# STATISTILISED MASINÕPPE MEETODID RANNIKUMERE LAINESPEKTRI HINDAMISEKS

Bakalaureusetöö

Juhendaja: Prof. Sven Nõmm
Ph.D.

Kaasjuhendaja: Sander Rikka
Ph.D.

Kaasjuhendaja: Mihhail Daniljuk
M.Sc.

Tallinn 2025

# Author’s Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Ilja Lastovko

15.01.2025

# Abstract

This thesis investigates the use of statistical machine learning methods and shallow neural network models to estimate the wave spectrum in coastal regions. Satellite-based synthetic aperture radar (SAR) imagery has been effectively utilized to estimate ocean wave spectra. However, the analytical techniques developed for long ocean waves have proven ineffective for much shorter wind-driven waves that dominate coastal areas.

While several deep learning approaches have been recently adapted for this task, there has been limited exploration into the potential of simpler statistical machine learning models and shallow neural networks.

This study evaluates the effectiveness of polynomial regression, regression trees, and regression forests, as well as their boosted variants, compared to shallow neural network models, for estimating the wave spectrum in the Baltic Sea using SAR imagery.

The findings of this research clearly indicate that boosted models and basic multilayer perceptron networks are the most accurate, achieving the lowest mean square error (below $0.5m$) and the highest Pearson correlation coefficient (up to $0.8$) between the estimated and observed wave spectra for certain frequencies.

The thesis is written in English and is 25 pages long, including 6 chapters and 10 figures.

# Annotatsioon

## Statistilised masinõppe meetodid rannikumere lainespektri hindamiseks

Käesolev bakalaureusetöö keskendub statistiliste masinõppe meetodite ja madala taseme närvivõrkude rakendamisele rannikumere lainete spektri hindamiseks. Satelliidipõhine sünteetilise apertuuri radar (SAR) pildistamine on osutunud tõhusaks ookeaniliste lainete spektri hindamiseks. Siiski on pika perioodiga ookeanilainetele arendatud analüütilised meetodid osutunud ebaefektiivseks palju lühemate tuulelainete suhtes, mis domineerivad rannikumere piirkondades.

Kuigi hiljuti on mitmed sügava õppe meetodid kohandatud selle ülesande jaoks, on lihtsamate statistiliste masinõppe mudelite ja madala taseme närvivõrkude potentsiaali uuritud vähem.

Käesolev uuring hindab polünoomregressiooni, regressioonipuude ja regressioonimetsade, samuti nende võimendatud versioonide tõhusust, võrreldes neid madala taseme närvivõrkude mudelitega, et hinnata Läänemere lainete spektrit SAR-piltide põhjal.

Uuringu tulemused näitavad selgelt, et võimendatud mudelid ja lihtsad multilayer perceptron närvivõrgud on kõige täpsemad, saavutades madalaima ruutkeskmise vea (alla $0.5m$) ja kõrgeima Pearsoni korrelatsioonikordaja (kuni $0.8$) hinnatud ja mõõdetud laine spektrite vahel teatud sagedustel.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 25 leheküljel, 6 peatükki ja 10 joonist.

# List of Abbreviations and Terms

| | |
|---|---|
| DT | Decision Tree |
| FFT | Fast Fourier Transform |
| GB | Gradient Boosting |
| HS | Significant Wave Height |
| ISP | Image Spectra |
| LR | Linear Regression |
| MLP | Multilayer Perceptron |
| NORA3 | Nordic wave hindcast model |
| PR | Polynomial Regression |
| RF | Random Forest |
| S1 | Sentinel-1 |
| SAR | Synthetic Aperture Radar |
| SVM | Support Vector Machine |
| Tm02 | Mean Wave Period |
| VH | Vertical transmit, Horizontal receive |
| VV | Vertical transmit, Vertical receive |
| WAM | Wave Model |
| XGB | Extreme Gradient Boosting |

# Table of Contents

# List of Figures

# 1 Introduction

Understanding wave spectra is essential for various applications in coastal and marine engineering, such as the design and operation of offshore wind turbines, harbors, and ensuring maritime navigation safety. Although buoy systems can provide highly accurate measurements, deploying and maintaining a sufficient number of them is both technologically challenging and financially burdensome. Additionally, their presence may interfere with navigation. In contrast, satellite-borne Synthetic Aperture Radar (SAR) offers a reliable method to capture sea surface data in all weather and lighting conditions, making it a practical alternative.

## 1.1 Background

For open ocean waves, established frameworks exist for estimating wave spectra. The fundamental mathematical description of SAR-based ocean wave imaging was first introduced by Alpers and Rufenach [1]. Following the MARSEN experiment [2], a generalized model was developed, providing analytical tools to map ocean wave spectra using SAR data [3]. Subsequent advances incorporated additional processes into inversion algorithms, often relying on image cross-spectra to determine wave direction [4], [5]. However, these methods are generally ineffective for the short, steep wind waves that dominate coastal regions [6].

Recent years have seen growing interest in applying deep learning techniques to estimate wave spectra. Long Short-Term Memory (LSTM) models [7], for example, have achieved Pearson correlation coefficients of $0.85$ between SAR-estimated and buoy-measured spectra [8], [9]. Similarly, transformer-based models [10] have demonstrated promising results, achieving correlation coefficients as high as $0.95$ for certain wave ranges [11], [12]. Despite these successes, deep learning models often involve high computational costs, require significant resources for training and operation, and lack interpretability.

## 1.2 Research Objectives

The complexity and high resource requirements of deep learning methods make it increasingly important to explore alternative approaches that are simpler, more computationally efficient, and easier to interpret and implement in practice. This research is dedicated to two main objectives:

1. **Evaluating Simpler Models:** Assess the applicability of statistical machine learning methods and shallow neural networks to estimate wave spectra. Determine the levels of accuracy these models can achieve compared to deep learning approaches.
2. **Understanding Spectral Influence:** Identify the specific components of the SAR image spectrum that have the most significant impact on predicting wave spectra.

# 2 Formal Problem Statement

The wave spectrum is typically represented as a sequence of numerical values. The prediction task can be framed as a sequence-to-sequence transformation problem, where the input sequence (SAR image spectrum) is transformed into the output sequence (in situ wave spectrum). This formalization opens opportunities for the use of advanced computational and statistical methods to address the problem efficiently.

## 2.1 Strategies for Spectrum Prediction

Two primary strategies can be used to address this problem, depending on the requirements for resolution, computational resources, and the level of interpretability desired.

### 2.1.1 Unified Model Approach

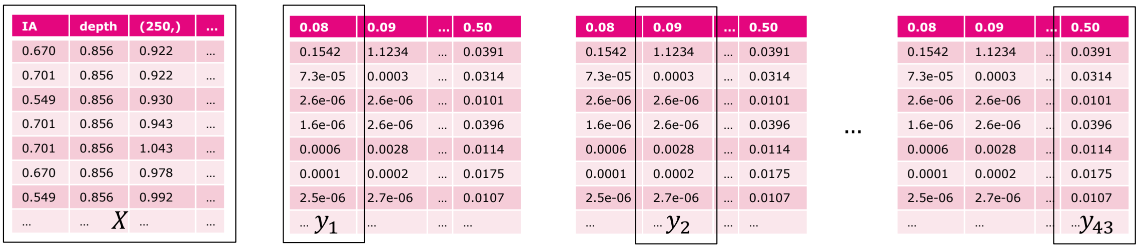
One approach involves training a single comprehensive model capable of predicting the entire in situ wave spectrum in a single step. This method is particularly advantageous for scenarios that require high-resolution spectral predictions or when large-scale datasets are available. Deep learning models, such as recurrent neural networks (RNNs) or transformers, are well suited for this purpose, because of their ability to handle complex, high-dimensional data.

### 2.1.2 Modular Model Approach

Alternatively, a modular approach involves training multiple models, each responsible for predicting a specific element of the spectrum. This strategy divides the task into smaller, more manageable units, reducing computational demands and increasing flexibility.

In this research, the in situ spectrum comprises sequences of $43$ numerical values. Each value can be predicted independently using separate models, as illustrated in Figure 1. Although this method may be less computationally efficient and lacks the holistic approach of deep learning, it offers distinct advantages:

- Smaller models require less computational power.
- Training models for individual spectral components provides insight into their unique significance.



| IA | depth | (250,) | ... |
|---|---|---|---|
| 0.670 | 0.856 | 0.922 | ... |
| 0.701 | 0.856 | 0.922 | ... |
| 0.549 | 0.856 | 0.930 | ... |
| 0.701 | 0.856 | 0.943 | ... |
| 0.701 | 0.856 | 1.043 | ... |
| 0.670 | 0.856 | 0.978 | ... |
| 0.549 | 0.856 | 0.992 | ... |
| ... | ... | ... | ... |

| 0.08 | 0.09 | ... | 0.50 |
|---|---|---|---|
| 0.1542 | 1.1234 | ... | 0.0391 |
| 7.3e-05 | 0.0003 | ... | 0.0314 |
| 2.6e-06 | 2.6e-06 | ... | 0.0101 |
| 1.6e-06 | 2.6e-06 | ... | 0.0396 |
| 0.0006 | 0.0028 | ... | 0.0114 |
| 0.0001 | 0.0002 | ... | 0.0175 |
| 2.5e-06 | 2.7e-06 | ... | 0.0107 |
| ... | ... | ... | ... |

| 0.08 | 0.09 | ... | 0.50 |
|---|---|---|---|
| 0.1542 | 1.1234 | ... | 0.0391 |
| 7.3e-05 | 0.0003 | ... | 0.0314 |
| 2.6e-06 | 2.6e-06 | ... | 0.0101 |
| 1.6e-06 | 2.6e-06 | ... | 0.0396 |
| 0.0006 | 0.0028 | ... | 0.0114 |
| 0.0001 | 0.0002 | ... | 0.0175 |
| 2.5e-06 | 2.7e-06 | ... | 0.0107 |
| ... | ... | ... | ... |

| 0.08 | 0.09 | ... | 0.50 |
|---|---|---|---|
| 0.1542 | 1.1234 | ... | 0.0391 |
| 7.3e-05 | 0.0003 | ... | 0.0314 |
| 2.6e-06 | 2.6e-06 | ... | 0.0101 |
| 1.6e-06 | 2.6e-06 | ... | 0.0396 |
| 0.0006 | 0.0028 | ... | 0.0114 |
| 0.0001 | 0.0002 | ... | 0.0175 |
| 2.5e-06 | 2.7e-06 | ... | 0.0107 |
| ... | ... | ... | ... |

Figure 1. Input and output data example.

# 3 Data and Methods

This study combines data from two key sources: SAR imagery from Sentinel-1 and the NORA3 wave hindcast. SAR provides detailed sea surface observations, while NORA3 provides reliable wave spectra for training and validating machine learning models.

## 3.1 Data

The ground truth data utilized in this study is derived from the NORA3 wave hindcast, which is based on the WAM wave model [13], [14]. The model spectra are represented by 30 frequencies, logarithmically spaced between $0.0345$ Hz and $0.5476$ Hz, along with 24 directions uniformly distributed over a full circular range. NORA3 data are generated using atmospheric input from HARMONIE-AROME, ice concentration data from ARC-MFC, and boundary wave spectra from ERA-5, ensuring high-resolution and reliable wave hindcasting. The dataset spans from 1964 to the present and is continuously updated with a delay of 4–5 months.

The Sentinel-1 (S1) Interferometric Wide (IW) swath Single Look Complex (SLC) subimages underwent calibration and noise filtering while preserving their radar projection to avoid data loss. The image spectra ($ISP$) were calculated using the fast Fourier transform (FFT) [15], [16] from two polarization configurations: vertically transmitted and received ($VV$), and vertically transmitted but horizontally received ($VH$). Additional metadata, such as satellite heading ($PASS$), incidence angle ($IA$), water depth, and image texture, were also preserved for subsequent analysis. The SAR data, which span from early 2015 to the end of 2021, were matched with the model spectra from the corresponding locations. Figure 2 illustrates the distribution of $H_S$ in the mean wave propagation direction used in this study.

For input to the models, the division of the polarization spectra was applied. This data handling approach is supported by two primary considerations. First, it ensures consistency with the methodologies established in previous studies, such as [8]. Second, certain statistical learning methods employed in this research are not inherently designed to process multidimensional input, necessitating such pre-processing.

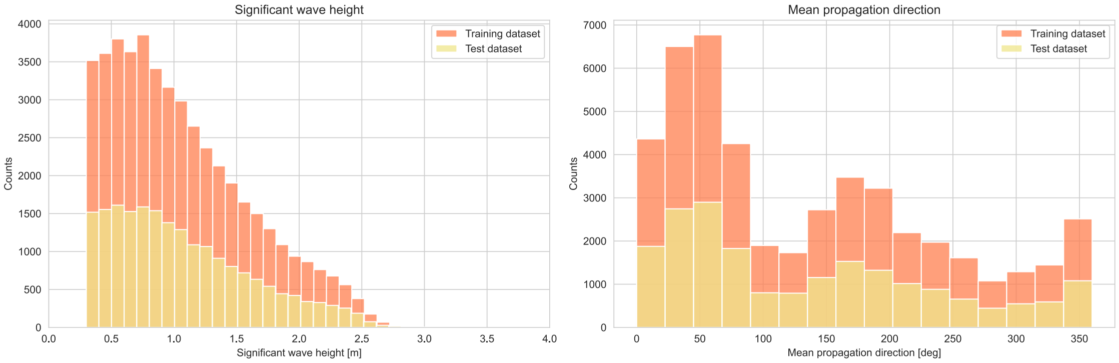


Figure 2. Distribution of significant wave height ($H_S$) and mean wave propagation direction for train and test dataset.

## 3.2 Methods and Their Application

The selection of models for evaluation was guided by their popularity and availability within the scikit-learn library [17]. This process resulted in the inclusion of several models: linear regression, polynomial regression, decision tree regression, random forest regression, gradient boosting regression, support vector regression and XGBoost regression. Most of these techniques are well documented in [18] and [19], while XGBoost, a more recent development, is specifically detailed in [20].

In addition, shallow neural network models were incorporated into the study. These models were constrained to a maximum of three layers and up to 150 neurons per layer. Unlike deep learning models, the architecture and functionality of these layers remained consistent throughout. It is also worth noting that combinations of different types of machine learning models were not explored in this study. Instead, a single type of model was trained and evaluated for each value in the wave spectra sequence, resulting in $n$ independent models of the same type (e.g., linear regression).

## 3.3 Models Training Process Details

To ensure a fair comparison among different models, all algorithms were trained and evaluated using a consistent train/test split of 70% for training and 30% for testing. Hyperparameter optimization was performed using grid search or randomized search methods, depending on the computational feasibility of each model. An important note here is that the hyperparameter grids were defined at the model type level. This means that all $n$ individual models of the same type (one for each element of the wave spectrum) were tuned using the same parameter grid.

### 3.3.1 Linear Regression

For linear regression, recursive feature elimination (RFE) was used for feature selection. The number of features to select was varied as follows:

- `n_features_to_select`: [5, 10, 15, 20] - How many features to keep for training.

### 3.3.2 Polynomial Regression

Similarly, polynomial regression employed RFE with the same range for feature selection. Additionally, the degree of polynomial was varied:

- `n_features_to_select`: [5, 10, 15, 20] - Number of features selected.
- `degree`: [2, 3, 5] - Degree of the polynomial.

### 3.3.3 Decision Tree Regression

For the decision tree regressor, a grid search was conducted on the following parameter grid:

- `max_depth`: [3, 5, 8, 11, 14, 17, 20] - How deep the tree can grow.
- `max_features`: [5, 10, 15, 20, 25, 30, 35, 40] - Maximum number of features to use for splits.

### 3.3.4 Random Forest Regression

The random forest regression model was tuned using randomized search over the following parameter distributions:

- `n_estimators`: [100, 200, 300, 400, 500] - Number of trees in the forest.
- `max_depth`: [None, 5, 10, 20, 30, 40, 50] - Maximum depth of each tree.
- `min_samples_split`: [2, 5, 10] - Minimum samples required to split a node.
- `min_samples_leaf`: [1, 2, 4] - Minimum samples required at a leaf.
- `max_features`: ['sqrt', 'log2', None] - Number of features to consider for splits.

### 3.3.5 Support Vector Machine

The support vector regressor was tuned using grid search over the following parameters:

- `C`: [0.1, 1, 10] - Controls how flexible the model is. Higher values mean less regularization.
- `kernel`: ['linear', 'rbf'] - Type of kernel used.

### 3.3.6 Gradient Boosting

The gradient boosting regressor was optimized using randomized search with the following parameter distributions:

- `n_estimators`: [50, 100, 200, 300] - Number of boosting stages (trees).
- `max_depth`: [3, 4, 5, 6, None] - Maximum depth of each tree.
- `min_samples_split`: [2, 5, 10] - Minimum samples required to split a node.
- `learning_rate`: [0.01, 0.05, 0.1, 0.2] - Step size for updates.
- `subsample`: [0.6, 0.8, 1.0] - Fraction of samples used for training.
- `max_features`: ['sqrt', 'log2', None] - Number of features to consider for splits.

### 3.3.7 XGBoost

The hyperparameters of the XGBoost regressor were optimized by a randomized search with these parameter distributions:

- `n_estimators`: [100, 200, 300, 400, 500] - Number of trees.
- `max_depth`: [3, 5, 7, 9, 12] - Maximum depth of trees.
- `learning_rate`: [0.01, 0.05, 0.1, 0.2] - How much each tree contributes to the prediction.

### 3.3.8 Multilayer Perceptron

The MLP regressor was tuned using grid search on a comprehensive parameter grid:

- `hidden_layer_sizes`: Configurations like (3,) or (150, 125, 100) - Number of neurons in each layer.
- `activation`: ['tanh', 'relu'] - Function used to calculate neuron output.
- `solver`: ['adam'] - Optimization algorithm.
- `alpha`: [0.0001, 0.001] - Regularization term to avoid overfitting.
- `learning_rate`: ['constant'] - How quickly the model updates weights.
- `learning_rate_init`: [0.001] - Initial learning rate value.
- `max_iter`: [1000, 2000] - Maximum number of training iterations.
- `batch_size`: ['auto', 64] - Number of samples processed at once.

# 4 Main Results

To comprehensively assess the performance of the models, four evaluation methods were applied. First, each frequency was individually analyzed to calculate $n$ comparable metrics, including MSE loss and correlation coefficients between predicted and actual values, as illustrated in Figure 3. Second, the predictions from all models of the same type were concatenated to form complete wave spectrum predictions. These predictions were then evaluated using predefined metrics at the spectrum level to better reflect the practical and realistic performance of the models, as shown in Figures 4 and 5. Third, the integrated wave parameters, such as the significant wave height ($H_s$) and the wave period ($Tm_{02}$), were analyzed, providing a higher-level understanding of the effectiveness of the models. Fourth, the data set was divided into two subsets based on wave height ($H_s \leq 1$ m and $H_s > 1$ m) to evaluate the performance of the model in different wave height ranges, addressing the distinct characteristics of smaller and larger waves.

The metrics used to evaluate model performance at the spectrum level include:

- Correlation coefficient between actual and predicted spectra.
- Difference between actual and predicted peak values of the spectra.
- Difference between actual and predicted peak locations of the spectra.

To summarize these metrics numerically, absolute values were used for metrics 2 and 3, ensuring that errors in either direction contributed to the final evaluation score.

Additionally, the influence of individual SAR image spectrum components on wave spectrum predictions was analyzed. This assessment highlighted which input features most significantly impact the accuracy of the models, particularly for different frequency ranges and wavelength groups. These insights help to clarify the relationships between SAR spectral features and model performance.

## 4.1 Frequency-Level Evaluation

In Figure 3, it is evident that all models, except the SVM, exhibit similar behavior: Higher frequencies are associated with higher correlation coefficients and lower MSE values. This trend could be due to more complex nonlinear relationships in the wave spectra data and smaller variation at lower frequencies, which make accurate predictions more challenging for the models.

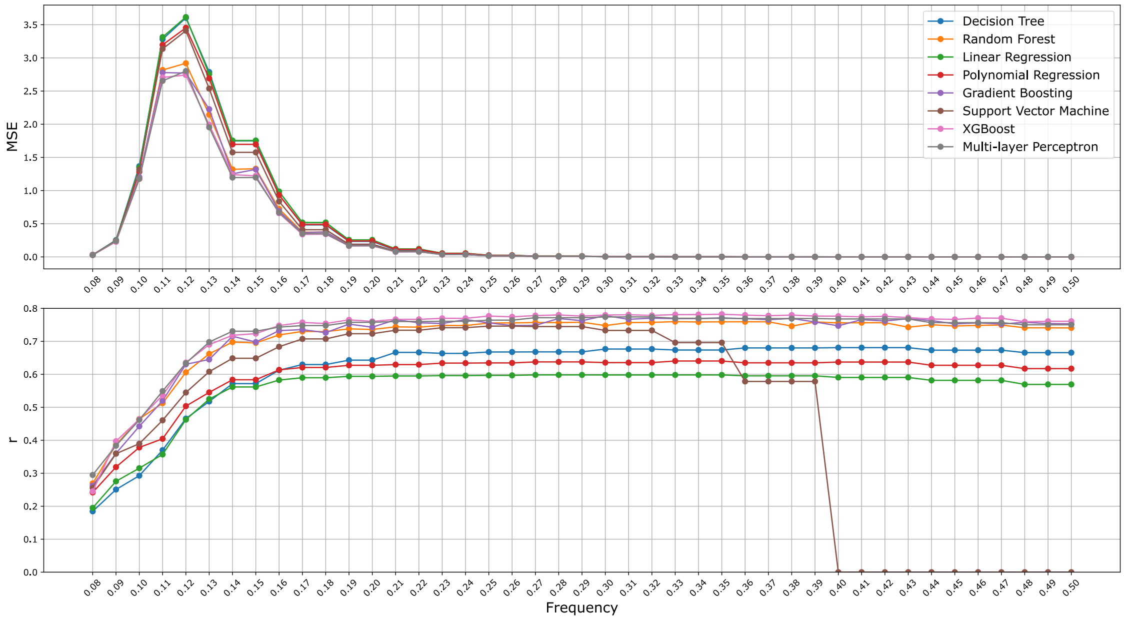


Figure 3. Models' performances across frequency domain. Upper plot shows mean squared error and bottom plot shows correlation.

It can also be observed that the MSE error distributions across different frequencies are quite similar among the models. However, the distributions of the correlation coefficients reveal notable differences. As anticipated, simpler regression models, such as linear regression, decision trees, and polynomial regression, demonstrate lower predictive performance. This is reflected in the green, blue, and red lines that represent these models. The similarity between polynomial regression and linear regression can be explained by the fact that the best-fitting polynomial for the data, determined through an exhaustive search, was of degree 2. The slightly better performance of the decision tree is probably due to its ability to capture more complex nonlinear patterns in the data.

Ensemble methods, including random forest, gradient boosting, and XGBoost, deliver significantly better performance compared to simpler models, with higher correlation coefficients and lower MSE values. These methods benefit from the combination of multiple weak learners, which improves predictive accuracy. Among these, XGBoost stands out as the statistical model that performs best, achieving the best correlation coefficients and the lowest MSE values in most frequencies. Gradient boosting ranks slightly below XGBoost, followed by random forest, which still performs competitively, but falls behind in terms of accuracy.

The best overall performance is achieved by XGBoost, closely followed by multilayer perceptron (MLP) networks. MLPs outperform other models in many cases, but fall slightly behind XGBoost in terms of correlation coefficients. This is expected as MLPs are highly complex and parameter-intensive models. Gradient boosting and random forest follow in the third and fourth place, respectively, offering a balance between performance and computational efficiency. Considering the marginal performance difference and the computational cost of training MLPs, XGBoost emerges as the most practical and effective solution for this problem.

## 4.2 Spectra-Level Performance

Moving to a more practical evaluation of models within the spectra domain, the results become less straightforward. Figures 4 and 5 illustrate the distributions of performance metrics for spectral estimation. These figures also show the values of the 25th percentile (Q1), 50th percentile (Q2 or median), and 75th percentile (Q3). In Figure 4, simpler statistical models are compared, revealing that linear regression consistently performs the worst across all metrics, as expected due to its purely linear nature. Polynomial regression of degree 2 performs slightly better, benefiting from its limited ability to model non-linearity. The decision tree demonstrates a substantial improvement in all metrics, particularly in the correlation distribution and the peak value location error. Although the SVM model shows high correlation coefficient quartiles, its reliability is questionable, which aligns with the earlier observation that the SVM often predicts constant values across multiple frequencies.

For the more advanced models, including ensemble methods and neural networks, the ensemble methods exhibit very similar performance, with XGBoost slightly outperforming the others in all metrics. The close results reflect the shared principles behind these methods, although minor differences arise due to their distinct underlying algorithms.

The MLP, being the largest and most complex model, produces mixed results. It is not the best model in terms of the correlation distribution or the quartile values. However, in particular, it achieves the lowest mean peak-value error, indicating its strength in accurately capturing peak magnitudes. However, it has one of the highest peak-value location errors, suggesting that the model can replicate the overall shape of the spectrum, but with a potential shift. This behavior also contributes to its relatively lower correlation coefficient distribution.

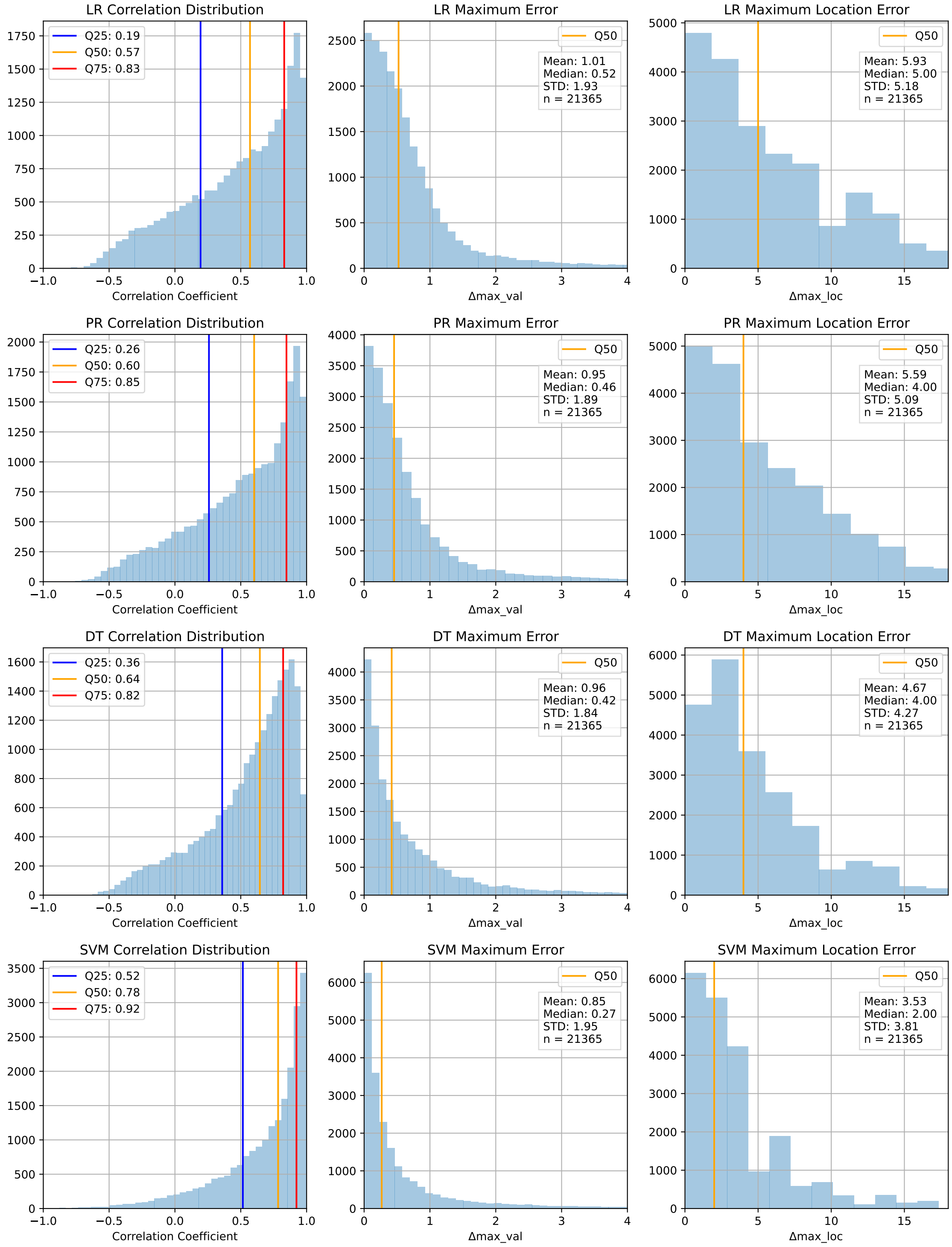


Figure 4. Correlation and error distribution across the test set for linear regression, polynomial regression, decision tree regression and SVM regression models.

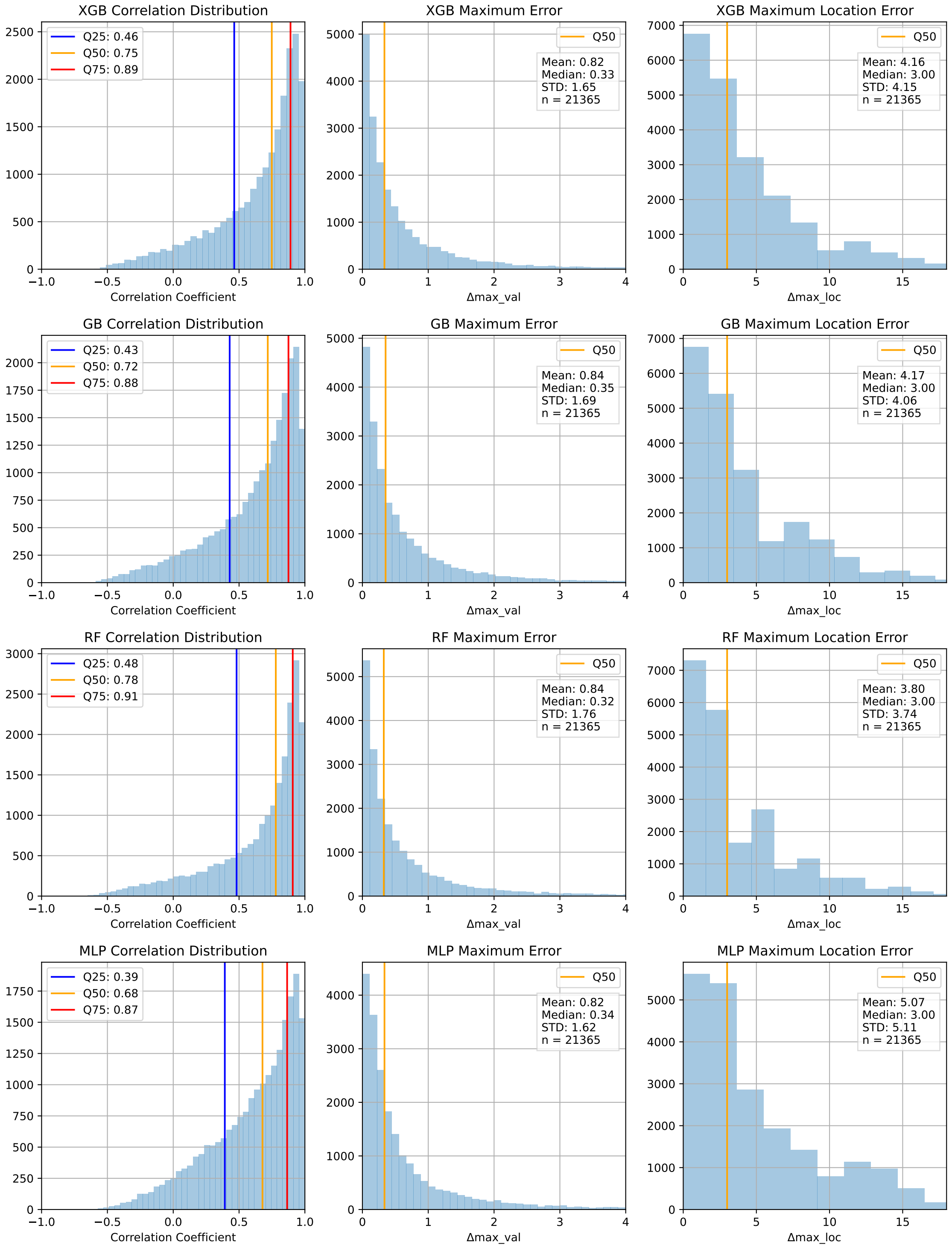


Figure 5. Correlation and error distribution across the test set for XGB regression, GB, Random forest regression and MLP models.

## 4.3 Integrated Parameters Comparison

Another practical comparison involves analyzing the integrated wave parameters, specifically the significant wave height ($H_s$) and wave period ($Tm_{02}$). Figures 6 and 7 display scatter plots comparing predicted values to actual values for all models, with the $y = x$ line representing perfect alignment.

For $H_s$, simpler models such as linear regression and polynomial regression perform the weakest due to their limited ability to model nonlinear relationships in the data. The decision tree shows a noticeable improvement, while the SVM demonstrates competitive results among the simpler models. This is an interesting contrast to the weaker performance of the SVM at higher frequencies when evaluating the full-wave spectrum. A possible explanation is that errors at higher frequencies contribute less to overall performance for $H_s$, allowing SVM to achieve better results. The ensemble methods, including random forest, gradient boosting, and XGBoost, perform consistently well, with XGBoost standing out as the best ensemble method. The MLP model achieves the best overall performance for $H_s$, effectively capturing nonlinear dependencies in the data.

For $Tm_{02}$, the performance varies more significantly across the models. Linear and polynomial regression again rank among the weakest, while the decision tree provides moderate improvements. The SVM achieves the strongest correlation and the lowest errors among the models. Ensemble methods, including XGBoost, gradient boosting, and random forest, deliver reliable performance but lag slightly behind the SVM. The MLP performs comparably to the SVM in terms of correlation but shows slightly higher errors, suggesting some difficulty in capturing extreme cases for $Tm_{02}$.

In general, the results highlight the strengths and limitations of different modeling approaches. The MLP model is highly effective for $H_s$, while the SVM demonstrates its strength for $Tm_{02}$.

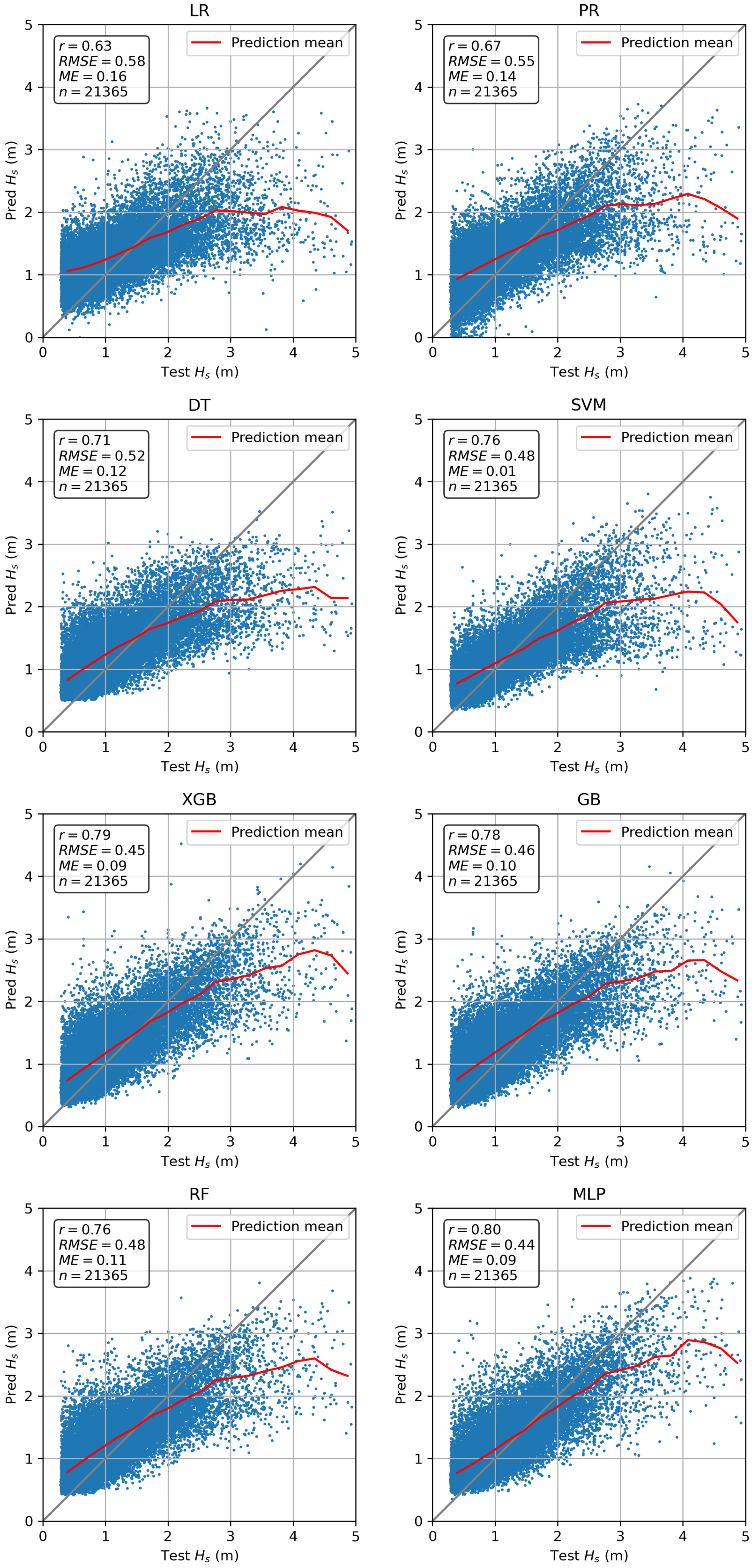


Figure 6. Scatter plot of predicted vs. test significant wave height ($H_s$) for all models.

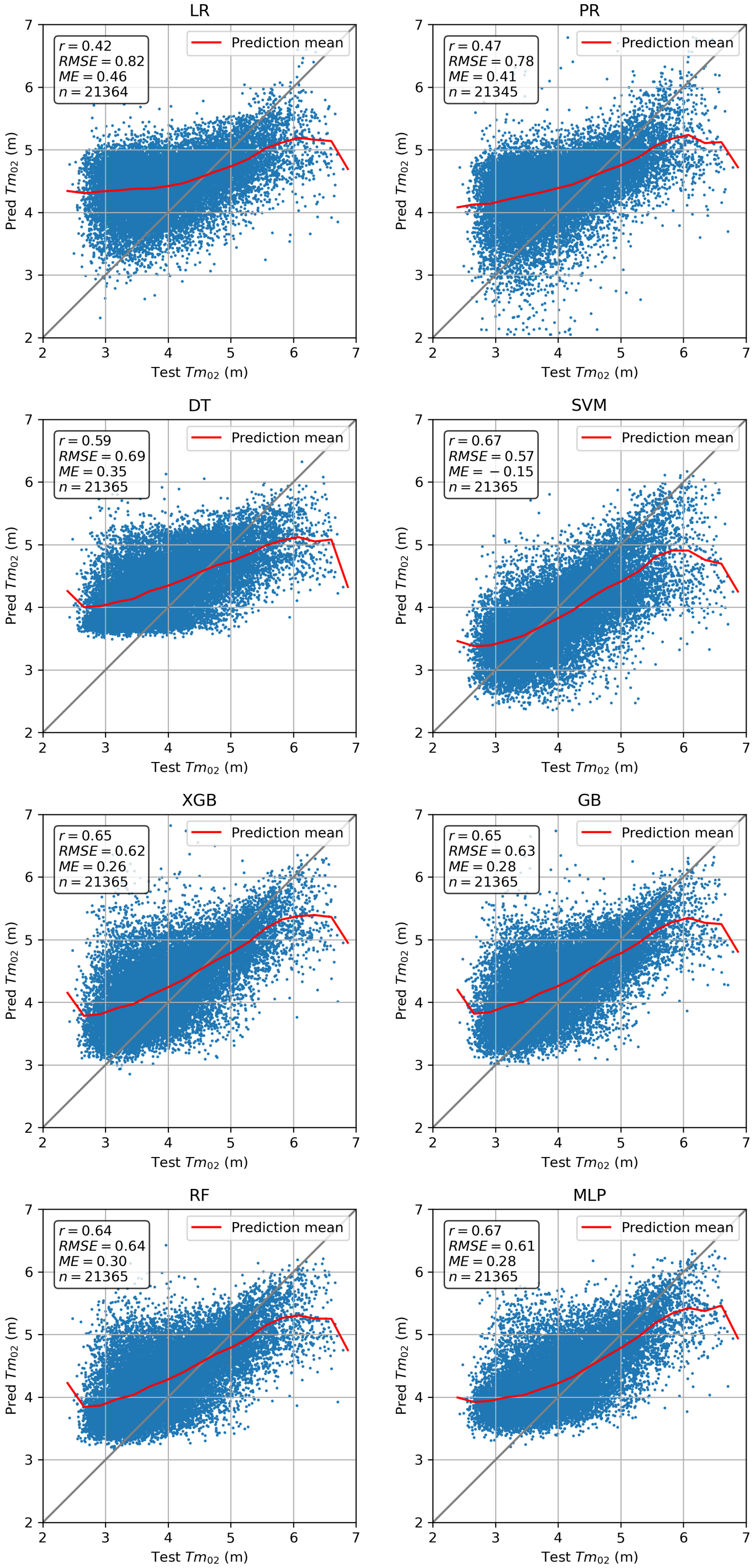


Figure 7. Scatter plot of predicted vs. test wave period $(Tm_{02})$ for all models.

## 4.4 Performance Comparison Based on Wave Height Ranges

Furthermore, to evaluate the effectiveness of the models, the data set was divided into two subsets based on wave height: waves with a significant wave height ($H_s$) $\leq 1$ m and waves with $H_s > 1$ m. This segmentation was motivated by the observation that the distribution of wave characteristics differs significantly between smaller and larger waves.

### 4.4.1 Waves Lower Than or Equal to One Meter

For waves with $H_s \leq 1$ m, all models demonstrated reduced performance compared to the results in the entire data set (Figure 8). The correlation coefficients peaked at $0.58$, indicating a noticeable decline in predictive accuracy across the board. This decline can likely be attributed to the narrower variability in wave spectra within this range, which limits the ability of models to distinguish key patterns effectively.

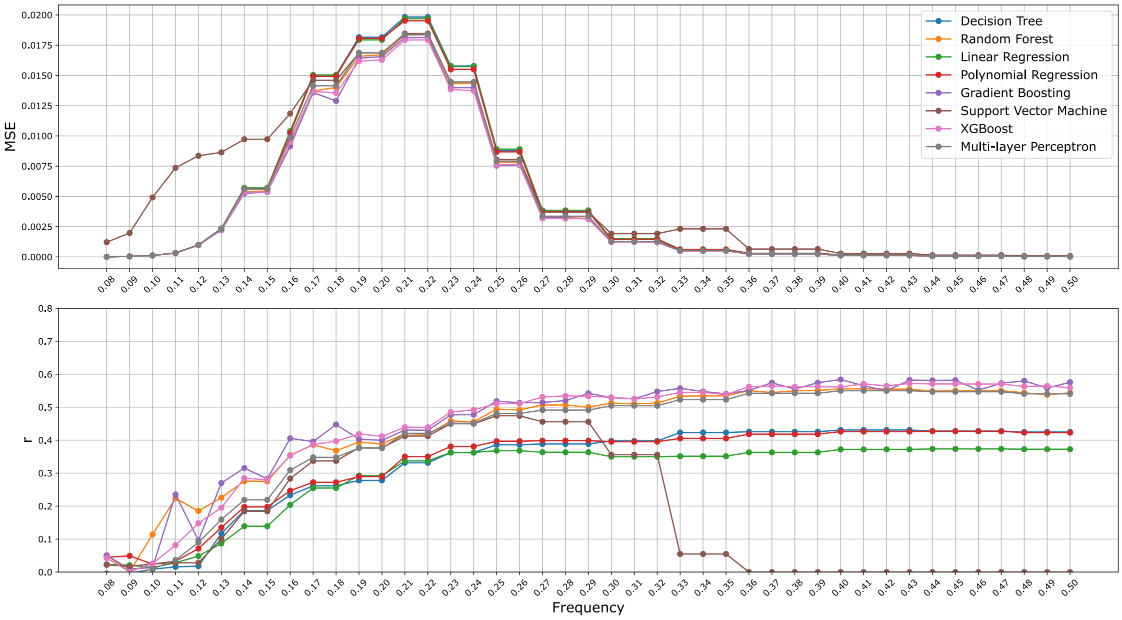


Figure 8. Models' performance across frequency domain for waves less than or equal to 1 meter. Upper plot shows mean squared error and bottom plot shows correlation.

Despite the overall reduction in performance, the relative ranking of the models remained consistent. Ensemble methods, such as random forest and XGBoost, and the multilayer perceptron continued to outperform simpler models like linear regression and polynomial regression. XGB and MLP once again shared the top positions, achieving the highest correlation coefficients and the lowest mean squared errors among all evaluated models.

### 4.4.2 Waves Higher Than One Meter

For waves with $H_s > 1$ m, predictive performance improved compared to the range $H_s \leq 1$ m, with correlation coefficients that reach $0.7$ (Figure 9). However, the results were still slightly worse than those obtained when all waves were included in the data set.

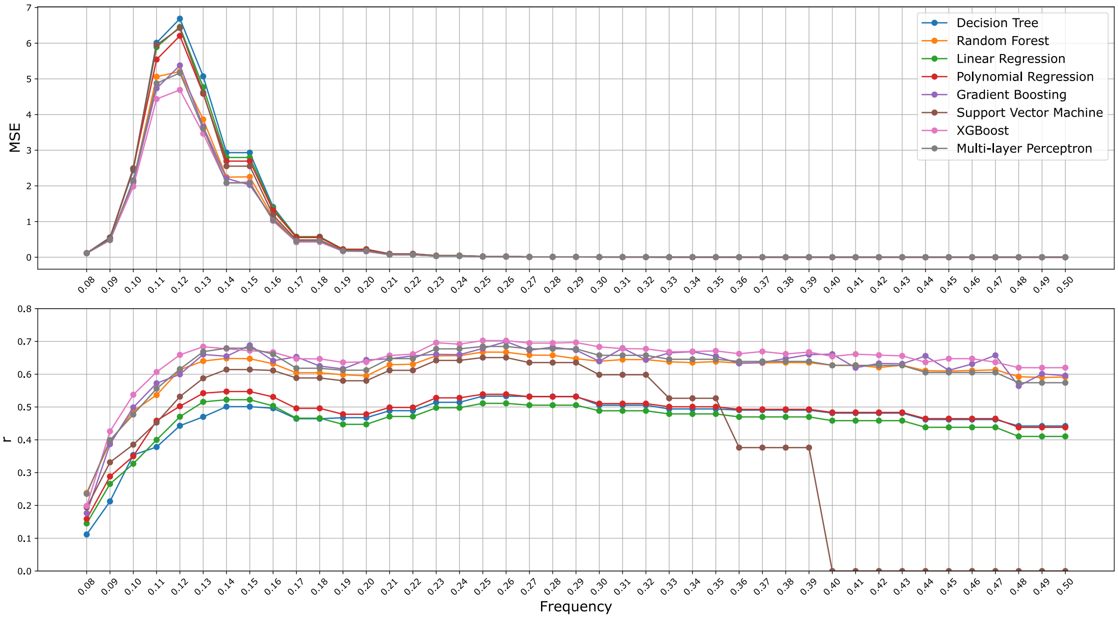


Figure 9. Models' performance across frequency domain for waves greater than 1 meter. Upper plot shows mean squared error and bottom plot shows correlation.

The ensemble methods continued to outperform simpler approaches. XGB retained its position as the model with the highest correlation coefficients and the lowest mean squared errors.

## 4.5 Analysis of Spectrum Components

Analysis of the importance values of the input features (SAR image spectra values) highlights key patterns of model performance in estimating the wave spectra (Figure 10).

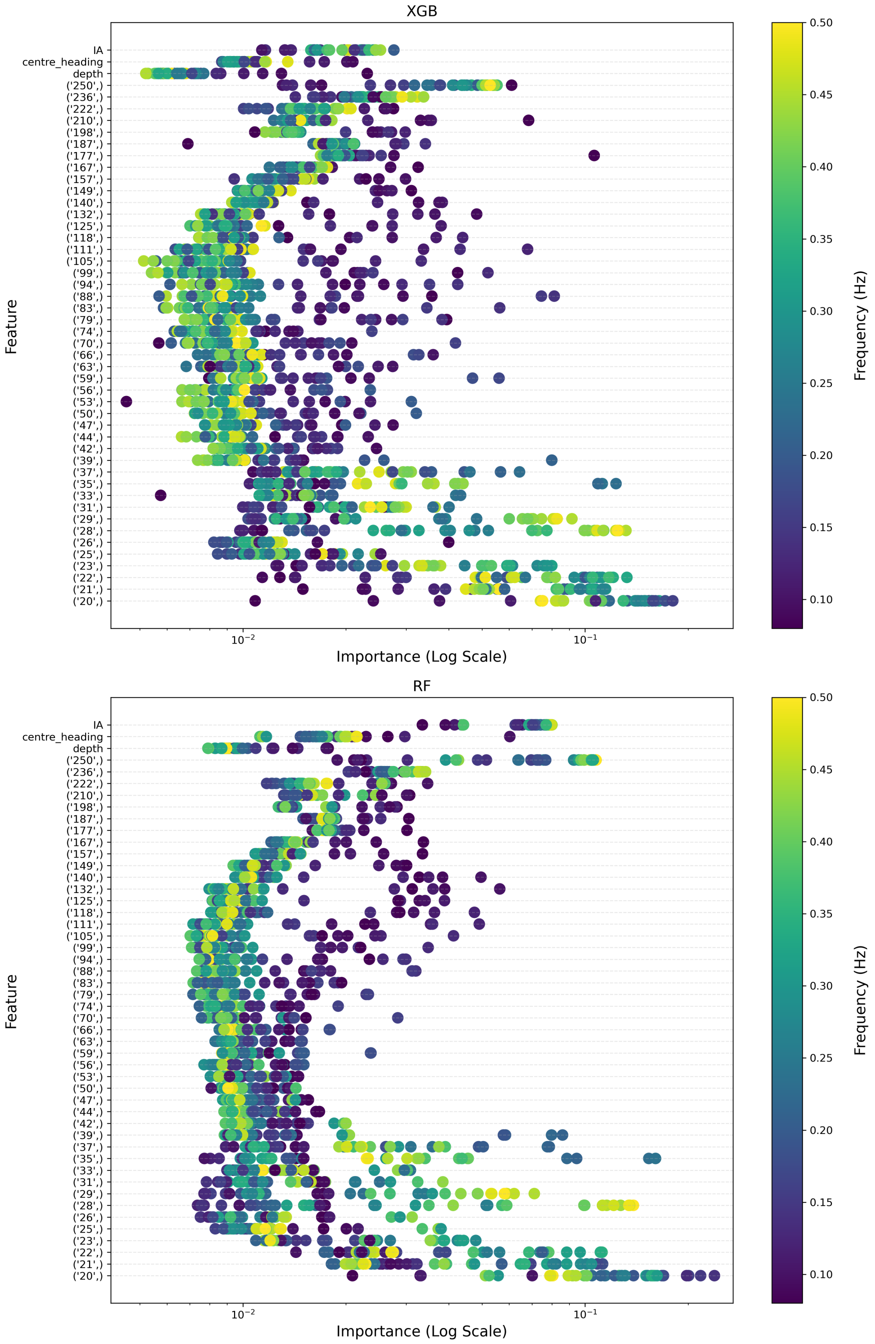


Figure 10. Feature importance distributions for XGBoost and Random Forest models.

For short wavelengths ($20 - 40$ meters) and long wavelengths ($> 150$ meters), the importance distributions for higher frequencies (approximately above $0.2$) are more scattered. This scattering likely correlates with improved accuracy in value estimation, suggesting that higher wave frequencies are better predicted as a result of the variability in feature significance across these wavelength ranges. However, a notable observation is the low importance assigned to medium wavelengths ($40 - 150$ meters).

The higher wave frequencies appear to depend on both the head and tail of the SAR image spectrum. For lower wave frequencies, the importance values are distributed relatively evenly across wavelengths. Despite this uniform distribution, these components do not significantly improve the value estimation.

# 5 Discussion

The overall performance of the models aligns well with expectations, apart from some exceptions, such as the unusual behavior of the support vector regression at specific frequencies. Although the SVM model demonstrates high performance across all metrics, its reliability is questionable due to an earlier observation that it predicts constant values across multiple frequencies, requiring further investigation. The linear regression model appears too simplistic to fully capture the underlying relationships in the data. For polynomial regression, further tuning of hyperparameters might improve performance, but the authors intentionally avoided aggressive adjustments to preserve the "vanilla" nature of the model. The relatively small performance gap between neural network-based models and boosted ensemble models highlights the complexity of the relationship between SAR imagery spectra and in situ wave spectra.

The difficulty in accurately estimating values at low frequencies (Fig. 3) remains somewhat unclear. Previous studies suggest that SAR should theoretically perform better for lower frequencies, corresponding to longer wavelengths. However, in the Baltic Sea, long-period waves are rare, leading to minimal representation in the data set [21], [22], [23], [24]. Furthermore, the wave model used may not accurately estimate these low frequencies, as it is primarily tuned for open ocean conditions [13]. Even when such waves are present, their low energy levels make them difficult to estimate accurately. Moreover, SAR imaging of wind waves can introduce noise or clutter in the imagery, potentially obscuring low-frequency signals.

Interestingly, the predicted average $H_S$ (Fig. 6) aligns closely with the line $y = x$ around the typical average wave height in the Baltic Sea. This suggests that the models perform well in estimating the mean wave height. However, a significant underestimation is observed for values of $H_S$ greater than approximately 2.5 m, probably due to the scarcity of such data in the training set.

For the estimation of the wave period (Fig. 7), there is a significant overestimation for shorter periods. Although the MLP model performs better in estimating these low periods and occasionally aligns closely with the minimum values of the dataset, it still falls short of the precision achieved by deep learning approaches, as shown in previous studies [9].

Notably, shorter wavelengths in the SAR image spectrum capture finer surface roughness and rapid changes in the wave field, which are sensitive to local wind conditions and other high-frequency phenomena. However, the mid-range wavelength band may fail to capture large- or small-scale dynamics effectively, making the information from this range potentially redundant for the output.

In many natural spectra, the energy distribution at extreme regions often exhibits clearer and more stable relationships with specific physical variables (e.g., wave height). Midrange wavelengths, however, may represent a "mixed zone" where the spectral energy density does not correlate well with the output, thus providing less predictive utility.

Exploring new strategies for model development could address some of these limitations. One potential avenue for future research is the combination of different machine learning models. Although this study focused on evaluating single models for each value in the wave spectra sequence, ensemble approaches that integrate the strengths of various models could potentially improve predictive accuracy.

# 6 Summary

In addressing the two research questions, it can be concluded that, in addition to the support vector regression - whose performance, though promising, is inconsistent - and linear regression - the remaining models demonstrate satisfactory performance for frequencies above $0.14$.

The analysis of spectrum components revealed that short and long wavelengths contribute significantly to the accuracy of the model, especially for higher frequencies (above $0.2$), due to their scattered importance distributions. In contrast, medium wavelengths exhibit low importance, suggesting limited predictive utility in this range.

However, both polynomial regression and decision tree regression are likely to require more extensive hyperparameter tuning or additional data preprocessing to achieve better results. These models show potential, but currently fall short in comparison to more robust methods.

XGBoost stands out as the model that performs the best, closely followed by multilayer perceptron and other boosting algorithms, with the random forest ranking slightly lower. This performance hierarchy confirms that the research objectives have been successfully met.

The poor performance of all models for frequencies below $0.14$ remains an open question. This issue highlights a key challenge to be addressed in future studies, in particular to better understand and overcome the limitations in this frequency range.

The preliminary results of this study have been submitted for review at a scientific conference [25].

# References

[1] W Alpers and CL Rufenach. “The effect of orbital motions on synthetic aperture radar imagery of ocean waves”. In: *IEEE transactions on Antennas and Propagation* 27.5 (1979), pp. 685–690.

[2] Klaus Hasselmann et al. “Theory of synthetic aperture radar ocean imaging: A MARSEN view”. In: *Journal of Geophysical Research: Oceans* 90.C3 (1985), pp. 4659–4686.

[3] Klaus Hasselmann and Susanne Hasselmann. “On the nonlinear mapping of an ocean wave spectrum into a synthetic aperture radar image spectrum and its inversion”. In: *Journal of Geophysical Research: Oceans* 96.C6 (1991), pp. 10713–10729.

[4] Geir Engen and Harald Johnsen. “SAR-ocean wave inversion using image cross spectra”. In: *IEEE transactions on geoscience and remote sensing* 33.4 (1995), pp. 1047–1056.

[5] Susanne Hasselmann, C Brüning, Klaus Hasselmann, and Patrick Heimbach. “An improved algorithm for the retrieval of ocean wave spectra from synthetic aperture radar image spectra”. In: *Journal of Geophysical Research: Oceans* 101.C7 (1996), pp. 16615–16629.

[6] AL Pleskachevsky, Wolfgang Rosenthal, and Susanne Lehner. “Meteo-marine parameters for highly variable environment in coastal regions from satellite radar images”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 119 (2016), pp. 464–484.

[7] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[8] Martin Simon, Sander Rikka, Sven Nomm, and Victor Alari. "Application of the LSTM Models for Baltic Sea Wave Spectra Estimation". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), pp. 72–77.

[9] Sander Rikka et al. "Wave Density Spectra Estimation with LSTM from Sentinel-1 SAR in the Baltic Sea". In: *2024 IEEE/OES Thirteenth Current, Waves and Turbulence Measurement (CWTM)*. IEEE. 2024, pp. 1–5.

[10] A Vaswani. "Attention is all you need". In: *Advances in Neural Information Processing Systems* (2017).

[11] Kevin Daniel and Didit Adytia. "A Significant Wave Height and Peak Wave Period Prediction with Transformer and LSTM Approach in Cilacap, Indonesia". In: *2023 International Conference on Data Science and Its Applications (ICoDSA)*. 2023, pp. 344–349. DOI: 10.1109/ICoDSA58501.2023.10276753.

[12] Mihhail Daniljuk, Sander Rikka, and Sven Nõmm. "Adaptation of transformer model for numeric case". In: *Proceedings 23rd IEEE International Conference on Machine Learning and Applications, ICMLA 2024, December 18-20, Miami, Florida. IEEE [Accepted]*. 2024.

[13] Øyvind Breivik et al. "The Impact of a Reduced High-Wind Charnock Parameter on Wave Growth With Application to the North Sea, the Norwegian Sea, and the Arctic Ocean". In: *Journal of Geophysical Research: Oceans* 127.3 (Mar. 2022). DOI: 10.1029/2021jc018196. URL: https://doi.org/10.1029/2021jc018196.

[14] *Norwegian Meteorological Institute, spectra database*. Accessed: 2022. URL: https://thredds.met.no/thredds/catalog/windsurfer/mywavewam3km_spectra/catalog.html.

[15] M. Heideman, D. Johnson, and C. Burrus. "Gauss and the history of the fast fourier transform". In: *IEEE ASSP Magazine* 1.4 (1984), pp. 14–21. DOI: 10.1109/MASSP.1984.1162257.

[16] Charles Van Loan. *Computational frameworks for the fast Fourier transform*. SIAM, 1992.

[17] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[18] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2001. ISBN: 9780387952840.

[19] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012. ISBN: 9780262018029.

[20] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

[21] Laura Tuomi, Kimmo K. Kahma, and Heidi Pettersson. "Wave hindcast statistics in the seasonally ice-covered Baltic Sea". In: *Boreal Environ. Res.* 16.6 (2011), pp. 451–472. ISSN: 12396095.

[22] Jan-Victor Björkqvist et al. "Comparing a 41-year model hindcast with decades of wave measurements from the Baltic Sea". In: *Ocean Engineering* 152 (2018), pp. 57–71.

[23] Jan-Victor Björkqvist et al. "Wave height return periods from combined measurement–model data: a Baltic Sea case study". In: *Natural Hazards and Earth System Sciences* 20.12 (2020), pp. 3593–3609.

[24] J.-V. Björkqvist et al. "Swell hindcast statistics for the Baltic Sea". In: *Ocean Science* 17.6 (2021), pp. 1815–1829. DOI: 10.5194/os-17-1815-2021. URL: https://os.copernicus.org/articles/17/1815/2021/.

[25] Mihhail Daniljuk, Ilja Lastovko, Sander Rikka, and Sven Nõmm. "Statistical Machine Learning Techniques for Wave Spectre Estimation in Coastal Seas". In: *Proceedings of the 17th Asian Conference on Intelligent Information and Database Systems (ACIIDS) [Submitted for review]*. 2025.

# Appendix 1 – Non-Exclusive License for Reproduction and Publication of a Graduation Thesis[1]

I Ilja Lastovko

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis “Statistical Machine Learning Techniques for Wave Spectre Estimation in Coastal Seas”, supervised by  Prof. Sven Nõmm, Sander Rikka and Mihhail Daniljuk
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons’ intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

15.01.2025

[1] The non-exclusive licence is not valid during the validity of access restriction indicated in the student’s application for restriction on access to the graduation thesis that has been signed by the school’s dean, except in case of the university’s right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.