Enhancing Ocean Monitoring for Coastal Communities Using AI

Erika Spiteri Bailey

Supervisor: Dr Kristian Guillaumier

Co-Supervisor: Dr Adam Gauci

May 2025

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Artificial Intelligence.

L-Università ta' Malta
Faculty of Information &
Communication Technology

Abstract

This research addresses the challenge of estimating significant wave height (SWH) using seismic data; a task beneficial for coastal safety, maritime operations, and environmental monitoring. Given that the livelihoods of over three billion people depend on coastal and marine resources, accurate and accessible models for SWH prediction are essential for informed decision-making and disaster preparedness. This study revisits and improves upon existing methods for the prediction of SWH using seismic data, specifically focusing on the limitations of data quality and computational feasibility in resource-constrained settings.

The problem was tackled by first formulating a baseline method, followed by the identification and implementation of key innovations. These included the use of a longest-stretch algorithm for more accurate seismic data, and hyperparameter tuning tailored to the local characteristics of each seismic station. The seven final models were trained and evaluated on consumer-grade hardware, ensuring their accessibility for deployment in areas with limited resources. These models were rigorously evaluated against existing baselines, with performance metrics including the coefficient of determination (R^2) and mean absolute error (MAE). The seven final models achieved a mean R^2 of 0.82978 (minimum: 0.60686, maximum: 0.92060) and a mean MAE of 0.13476 (minimum: 0.10066, maximum: 0.18243).

The results demonstrate significant improvements over the baseline models, particularly in terms of predictive accuracy and model efficiency. Specifically, the final models achieved an increase in R^2 of up to 0.13316 and a reduction in MAE by 0.02625 m over the baseline models, considering the average performance across all stations. However, challenges such as accurately predicting extreme weather still remain, due to the limited existence of such instances within the data.

The primary contribution of this research is the development of computationally efficient, locally optimised models for SWH prediction using seismic data, covering a region of interest around Sicily and Malta, that can be deployed on consumer-grade hardware. This expands the accessibility of artificial intelligence in low-resource settings. Future work could focus on advanced gap-filling techniques, data augmentation for extreme weather scenarios, and alternative model architectures for better handling extreme values.

Acknowledgements

On my toughest days, may this dissertation serve as a reminder that with God's grace and the unwavering support of my loved ones, I am capable of overcoming any challenge. This journey has shown me that with faith, love, and determination, I can rise above even the hardest moments and keep moving forward.

I would also like to extend my sincere gratitude to the University of Malta, my supervisor, Dr Kristian Guillaumier, my co-supervisor, Dr Adam Gauci, and the Malta Digital Innovation Authority (MDIA) for their support through the scholarship awarded for the Master of Science in Artificial Intelligence.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial intelligence.

ANN Artificial neural network.

CMEMS Copernicus Marine Environment Monitoring Service.

CNN Convolutional neural network.

DL Deep learning.

EIDA European Integrated Data Archive.

GPU Graphics Processing Unit.

HF High frequency.

KNN K-nearest neighbours.

LGB Light gradient boosting.

lidar light detection and ranging.

LSTM Long short-term memory.

MAE Mean absolute error.

MARE Mean average relative error.

MCMC Markov chain Monte Carlo.

ML Machine learning.

MSE Mean squared error.

PM Primary microseism.

PSD Power spectral density.

RF Random forest.

RMS Root mean square.

RMSE Root mean squared error.

RNN Recurrent neural network.

SM Secondary microseism.

SMOTE Synthetic minority over-sampling technique.

SPSM Short period secondary microseism.

SVM Support vector machine.

SWH Significant wave height.

XGBoost Extreme Gradient Boosting.

YOLO You Only Look Once.

1 Introduction

The seas and oceans are the lungs of the Earth; they regulate the climate, and absorb all but 10% of the excess heat that is produced by greenhouse gas emissions [1]. Recognising their importance, the United Nations has identified the conservation and sustainable use of oceans, seas and marine resources for sustainable development as the 14th Sustainable Development Goal [2]. The livelihoods of over three billion people are dependent on marine and coastal resources; however, despite their critical role in sustaining life and global economies, the understanding of the ocean remains remarkably limited [3]. One area of interest is the relationship between seismic activity and sea state parameters, such as significant wave height.

Current wave estimation methods have limited reliability, and real-time measurements are prone to communication interruption. Wave buoys are one such instrument of wave estimation but they are costly and are prone to bio-fouling. In this context, bio-fouling refers to a process whereby sea-life build-up impacts the instrument accuracy and efficiency, sometimes rendering the buoys unusable. An example of a bio-fouled wave buoy, whereby the fouling material added 23% to the wave buoy mass is shown in Figure 1.1 [4]. Weather buoys are also prone to being sent adrift in stormy weather, being damaged by moving ships, or being caught in fishing equipment [5].

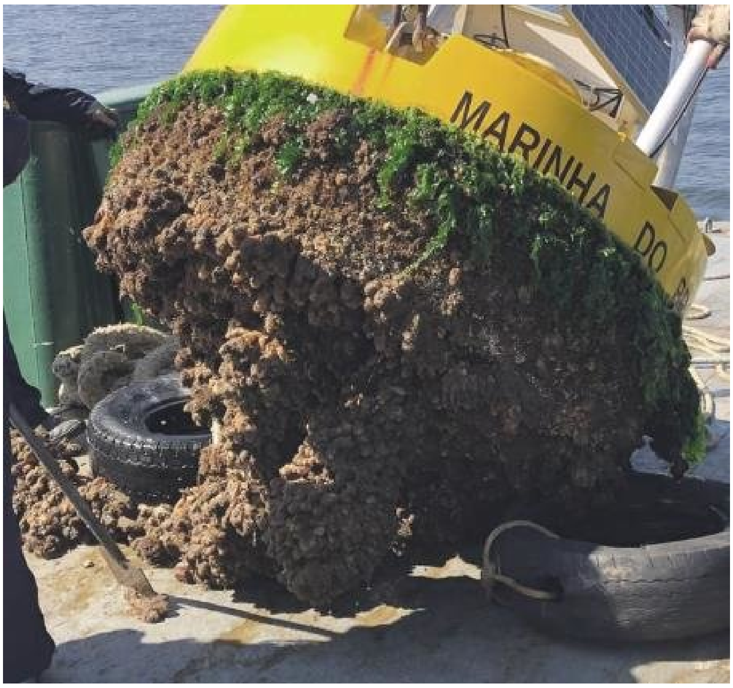


Figure 1.1 Photo of a bio-fouled wave buoy in Rio de Janeiro, Brazil in 2019 [4].

An alternative to weather buoys for measuring sea wave conditions is weather satellites. Although they are widely used and are not prone to the same risks as weather buoys, their fuel is limited and runs out over time and eventually, they become part of what is commonly referred to as ‘space junk’; orbiting debris polluting space and increasing complexities for future space missions. In 2022, the material orbiting the Earth at high speeds exceeded 9500 metric tonnes [6].

This places the seismometer at an advantage. It is placed on land, making it easily accessible for maintenance and repairs. In the event of a replacement, it can be carefully and ethically disposed of. Communication delays are non-existent, as measurements are made in real-time and onshore, and the signals are largely continuous. The 12th Sustainable Development Goal targets sustainable consumption and production patterns [2]. With a well-established relationship between seismic signals and sea conditions, the sea state would be known through a pre-existing resource: the seismometer. This would provide an additional, ethical way of measuring the sea state. Subsequently, the three billion people whose livelihoods depend upon the sea have a sustainable method for obtaining the information that they need, while preserving the sea, protecting it from unnecessary pollution, in line with the 14th Sustainable Development Goal [2].

1.1 Literature Gap

Despite the critical importance of monitoring ocean conditions for climate modelling, maritime safety, and coastal management, current methodologies for estimating wave height rely predominantly on direct oceanographic measurements such as satellite altimetry or buoys. While effective, these approaches can be resource-intensive, spatially limited, or introduce delays in delivering real-time data. At the same time, seismic noise, which is often considered a by-product in geophysical studies, has shown potential as a proxy for oceanic wave activity, particularly through microseism analysis. However, the relationship between seismic activity and sea state, especially SWH, remains underexplored and poorly understood in the literature. More critically, few studies have attempted to leverage artificial intelligence (AI) techniques to systematically investigate and model this relationship. This represents a notable gap in the field and provides an opportunity for innovation in ocean sensing and environmental monitoring.

To use seismic data, it must first be transformed and feature-engineered to condense the information from a quasi-continuous signal into a computationally feasible data source. Previous work suggests that the correlation between seismic data and wave height decreases over distance between the two points [7]. However, existing AI-based solutions fail to focus on local elements, choosing a large region of interest that undoubtedly diminishes the features of this relationship. Most reviewed methods also lacked variety and detail in the evaluation metrics they were challenged upon, providing limited options for baseline comparison.

Additionally, the employed methodologies were not robust; while it is common knowledge that the sea state can change within a few hours, previous work made use of simple linear interpolation to fill hourly gaps, accumulating to over 200 days long [7]. The data was over-engineered during pre-processing, aggravated by poor model selection that lacks justifiable reasoning. For example, skewness correction was applied to the data prior to training three different models – two of which are known to be resilient to skewness [7].

The above critiques are explored in greater detail in the following sections. However, what becomes immediately evident is a significant gap in the current body of literature – namely, the absence of a robust methodological baseline upon which further research in this area can reliably build. While preliminary studies have suggested a potential relationship between seismic signals and significant wave height, this relationship remains limited in scope and underexplored in depth. Without a consistent, data-driven foundation, the field lacks the clarity and direction needed to fully harness the potential of this connection.

1.2 Aims and Objectives

The aim of this study is to investigate the relationship between lower-frequency seismic amplitude and SWH, with a particular focus on the coastal regions of Sicily and Malta. The central objective is to establish a foundational baseline for future research in this domain. This will be achieved by critically examining the goals and limitations of prior studies, refining their methodologies, and demonstrating that the relationship between seismic signals and sea state not only exists, but can be effectively modelled using AI techniques.

Given the preliminary nature of research in this area, there is currently no compelling justification for employing highly complex AI architectures. As will be discussed in later sections, traditional machine learning models, specifically RF regressors, are shown to perform well when aligned with a coherent methodology, compared with existing approaches, in both generalisation and error. This study presents a recreated baseline, against which the performance of these models with an improved methodology is compared, offering evidence that complexity does not always equate to performance in this context.

Discrepancies between the literature, the recreated baseline, and the final models will be rigorously analysed, with particular attention paid to earlier AI-based efforts that predate this dissertation. By optimising hyperparameters for each selected location, the computational efficiency of the models will be enhanced, reducing the reliance on advanced hardware typically required in related studies. Moreover, the pipeline's efficiency will be improved through methodological clarity and the use of minimal synthetic data, foregoing more elaborate gap-filling techniques in favour of practical simplicity.

The dataset used spans a broad geographical area that encapsulates the region of interest. Only seismic stations with sufficient data coverage – defined here as at least one full year – will be included, ensuring the models are trained on data that captures seasonal variability.

Model performance will be assessed using a diverse set of evaluation metrics, providing a meaningful basis for benchmarking in future studies. By addressing key shortcomings in existing work and presenting a robust, reproducible methodology, this research aims to meaningfully advance our understanding of the connection between seismic activity and sea state conditions. To meet the aforementioned aims, the following research question will be answered:

1.2.1 Primary Research Question

Can a robust and reliable relationship between seismic signals and sea conditions be established using AI under varying environmental and geophysical conditions?

1.2.2 Objectives

To meet the aims and answer the primary research question, the following objectives shall be addressed:

Objective 1: Baseline Relationship

To establish a reliable starting point for future investigation, this study aims to recreate and evaluate baseline results for the relationship between hourly seismic RMS amplitude and SWH. Given the limited scope and clarity in existing literature, the recreated baseline will mirror previous methodologies while introducing carefully selected evaluation metrics to enable a fair and comprehensive comparison. This will allow for verification of prior findings, improvement through contextual enhancement, and an expanded performance review across multiple evaluation criteria.

Objective 2: Cost-Effective Solution

This study aims to develop a cost-effective modelling approach that enhances the accessibility and deployability of AI-driven solutions in regions with limited technical or financial resources. Existing literature often assumes the availability of advanced, high-performance computing infrastructure, which can pose a significant barrier to adoption, particularly in communities operating with constrained budgets. By leveraging consumer-grade hardware and efficient models, this research challenges the notion that complex problems always require complex solutions.

Beyond cost and practicality, there are important ethical considerations. Prioritising lightweight and efficient solutions over deep networks ensures that computational resources are allocated responsibly, avoiding the unnecessary environmental and economic costs associated with energy-intensive models. In doing so, this research contributes not only to scientific advancement but to the ethical deployment of AI in environmental monitoring.

Objective 3: Efficient and Deployable Pipeline

By designing an efficient data processing pipeline that applies only the minimal necessary transformations, this study aims to evaluate the true capability of subsequent machine learning (ML) models in a realistic context. This approach ensures that any preprocessing steps used during training are not only computationally lightweight, but are also feasible to implement during real-time inference in practical deployments. Minimising data transformations reduces system complexity and supports long-term goals of computational efficiency, particularly in low-resource environments. This further reinforces the study's commitment to scalable and deployable AI solutions.

Objective 4: Location-Specific Hyperparameter Tuning

In the context of efficient pipelines, simplified model architectures, and the use of real-world observational data, hyperparameter tuning emerges as a critical objective. Given that models are trained on data specific to individual geographic locations, it is essential to optimise them accordingly. Regional variations in underwater topography, sea currents, and atmospheric conditions can significantly influence the relationship between seismic activity and sea state. Therefore, tailored hyperparameter configurations are required to ensure robust model generalisation and performance. This objective reinforces the need for location-aware model development, further enhancing the relevance and adaptability of the proposed approach in diverse marine environments.

Objective 5: Prioritising Real-World Data

Gap-filling techniques, while often necessary, introduce uncertainty and potential bias into datasets, especially when used to train machine learning models that may propagate and amplify these errors. To mitigate this risk, this study adopts a *less-is-more* and *quality-over-quantity* philosophy, favouring high-integrity, real-world data with minimal preprocessing or interpolation. Only seismic stations with at least one full year of data and minimal missing values will be included in the modelling process. This approach ensures that the models are trained on data that closely reflects actual environmental conditions, thereby improving reliability, interpretability, and generalisability of the results.

1.3 Proposed Solution

The proposed solution is structured into several key stages, each aimed at addressing the research objectives and refining the methods for establishing the relationship between seismic signals and sea conditions. These stages are outlined below:

- Replicate existing methods:

  The first stage involves replicating the methodologies from existing literature as accurately as possible. This includes performing extensive data preprocessing, applying necessary transformations, and verifying the results reported in previous studies. By employing a broader set of evaluation metrics, this stage ensures that the relationship between seismic signals and sea state can be validated, laying the groundwork for subsequent improvements. The results of these initial recreations serve as a baseline for comparison.

- Improving existing methods:

  Once the baseline is established, the next stage focuses on improving the original methodologies. This involves identifying best practices, eliminating redundant or unnecessary data transformations, and selecting the most appropriate model for the dataset. To achieve these improvements, exploratory data analysis will be conducted, coupled with a critical review of the techniques employed in the literature. This stage seeks to optimise the approach and lay the foundation for more robust and efficient modelling.

- Data selection and minimising interpolation:

  A crucial step in ensuring data quality involves identifying the longest continuous stretch of data from each of the 14 seismic stations. To maintain the integrity of the data and minimise potential errors from interpolation, only stations with sufficient continuous data of at least one year will be included in the analysis. Stations with insufficient data will be excluded to ensure that models are trained on realistic, high-quality datasets that accurately reflect the variability of marine environments.

- Hyperparameter tuning:

  In this stage, hyperparameter tuning will be conducted for each included location, local to the stations across Sicily and Malta. Target variables will be identified, including both raw and engineered data, to evaluate the optimal model performance. The tuning process will verify that the best hyperparameters are selected for each station, contributing to improved model accuracy and performance.

- K-fold cross validation:

  To assess the robustness of the models and ensure consistent performance across different data subsets, K-fold cross-validation will be performed using the optimal set of hyperparameters (with $k = 5$). This method helps to analyse the variance in model performance across different training and test data partitions, ensuring that all data points contribute to both model training and testing.

- Error analysis:

  Finally, a detailed error analysis will be conducted to identify areas of recurring underperformance. This stage involves investigating potential causes for errors and inconsistencies in model predictions. By pinpointing the sources of these errors, adjustments could be made to improve the models and the overall methodology, ensuring greater reliability and robustness in subsequent research.

1.4 Contributions

By meeting the aims and objectives, the following contributions were established:

- A scientifically sound baseline was established, based on careful methodological replication and validation against existing literature. This baseline serves as a reliable reference for future research and enables a clear comparison of improvements in methodology, performance, and model generalisation. The models developed for establishing the SWH through seismic signals are suitable for day-to-day use.

- In line with the goal of making AI technology accessible and cost-effective, models were successfully trained and deployed on consumer-grade hardware. This ensures that the developed solution is deployable in environments with limited resources, contributing to the ethical deployment of AI in marine monitoring.

- Through the process of hyperparameter tuning, the optimal configurations for each station were identified. This step enhances model accuracy and ensures that each location's unique environmental conditions are adequately reflected in the predictive models, thus increasing their reliability and relevance. Furthermore, an algorithm was developed to identify the longest continuous stretch of data for each station, minimising the need for complex gap-filling techniques. By limiting preprocessing and focusing on high-quality real-world data, the models were trained on datasets that better reflect actual environmental conditions. The RF regressors trained using this approach outperformed baseline models by up to 0.2566 in R^2, demonstrating the benefits of minimal data manipulation.

- Through a thorough analysis of the datasets, particularly those from station AIO and CSLB, problematic data sources were identified, providing valuable insights into areas where model performance could be improved. This analysis highlighted inconsistencies that might have otherwise been overlooked, contributing to a deeper understanding of data quality in seismic-wave height research within this region.

- A significant bias in the dataset was uncovered, revealing an under-representation of periods characterised by extreme sea conditions. This discovery suggests that further work is needed to ensure that training data is representative of all sea states, particularly extreme conditions, which are critical for model generalisation in real-world applications.

1.5 Dissertation Structure

Chapter 2 introduces the fundamental concepts of seismic noise and the impact of sea waves on seismic signals. The chapter aims to provide the reader with a clear overview of the relevant background material, without delving into excessive technical detail. This context sets the stage for the more detailed analysis that follows in subsequent chapters.

Chapter 3 reviews existing literature in the field, focusing on both traditional numerical methods and more recent AI-based approaches. In addition to directly related studies, this chapter explores parallel research areas that, while addressing different research questions, offer valuable insights and methods that can help fill the gaps identified in current literature.

Chapter 4 presents the methodology adopted for both the recreated baseline and the final proposed solution. It covers the entire process, from data collection and preprocessing to exploratory data analysis, model selection, experimental setup, and implementation details. This chapter outlines the steps taken to ensure that the methodology is both robust and transparent, establishing the foundation for the subsequent experimental work.

Chapter 5 provides a comprehensive presentation of the experimental results. This includes an evaluation of the incremental improvements made to the baseline model, along with a detailed assessment of the proposed solution's performance, particularly with regard to hyperparameter tuning. The results of the K-fold cross-validation for each station are presented, accompanied by waveforms and examples of predicted versus actual sea conditions. A direct comparison to the baseline is made, highlighting areas of consistency as well as improvements. The chapter also includes a critical evaluation of the methodology, with particular attention to error analysis and an exploration of the underlying causes of recurring underperformance. Finally, key findings are summarised, with their implications tied back to the original research objectives.

Chapter 6 concludes the dissertation by summarising the methods and results, assessing the contributions made to the field, and discussing the limitations of the study. Based on the findings, recommendations for future research are also provided, suggesting avenues for further exploration to build on the work presented here.

2 Background

This chapter provides essential background information to familiarise readers with key concepts and technicalities relevant to this work's geo-scientific and AI context; it provides a general understanding of the domain-specific aspects of this research. Understanding concepts such as microseisms and their significance in geophysical studies is crucial for appreciating the broader objectives and methodologies discussed in the ensuing dissertation.

2.1 Wave-Coast Interactions

Ocean waves can be caused by various disturbances to the sea's equilibrium state, such as the weather, earthquakes, and areas with varying atmospheric pressure which result in a pressure gradient [8]. Placing a lens on the Mediterranean, and their central waters in the vicinity of Malta as the region of interest for this dissertation, it is common knowledge that extreme weather events such as typhoons, hurricanes, and tornados are uncommon. The higher-impact storms observed in the Mediterranean are referred to as *Medicanes* (Mediterranean Hurricanes), which tend to be less intense and are smaller in diameter than a typical Atlantic hurricane [9]. They are also not considered regular occurrences. Other phenomena that may result in waves such as tides are not relevant since the Mediterranean exhibits a low tide amplitude [10].

In parallel, sensitive instrumentation provides granular insight into seismic activity, allowing for research to be conducted on even the smallest ground movements, known as microseisms. Commonly used seismometers gather data on velocity, displacement and acceleration of ground motion. Figure 2.1 shows the three components of seismic data and the interchangeable ways they are referred to.

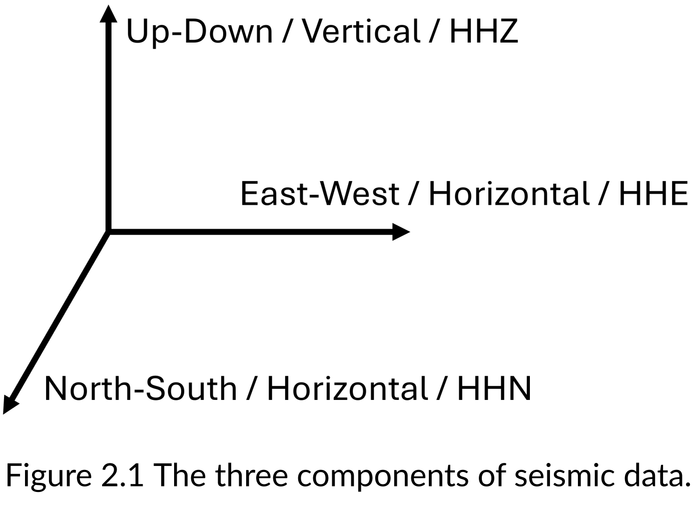


Figure 2.1 The three components of seismic data.

2.2 Seismic Noise

The term “microseism” refers to ambient seismic activity having a period in the range $[2, 20]$ seconds [11]. Such microseisms have been extensively researched to understand the origin of this continuous background noise that is observed in seismic monitoring. Before modern, sensitive instrumentation became available, these signals could only be classified as noise, however, recent research on these granular signals uses microseismic information to provide information on ocean and lake processes, and to study changes in stress and strain for volumes in the Earth’s crust, among other topics [12].

Evidence suggests that microseisms are largely caused by ocean waves. When these ocean waves reach the coast, they induce ground motion, generating seismic signals. These signals are mostly continuous and are recorded everywhere on Earth [13]. Their properties make seismic data an abundant source of information, making it highly appealing for AI projects, which are typically data-intensive tasks.

Lower frequency seismic noise is associated with natural activities such as ocean and meteorological activity [14]. Ocean gravity waves create pressure variations that result in energy transfer to the solid Earth, producing microseism seismic noise [15]. This research will focus on these lower frequencies, including the microseism range, to determine their relationship with the height of the waves which result in these signals.

Various researchers have proposed further subdivisions within the extra-low frequency microseism range. However, there is no universally accepted standard for the exact frequencies of these subdivisions, as different studies suggest slightly varying frequency ranges for each category.

There are two distinguishable categories of microseism – the primary microseism, and the secondary microseism. The primary microseism has the same frequency as the generating ocean waves. It is created by direct variations in ocean wave pressure on the ocean bottom [16]. Since the amplitude of pressure fluctuations decays exponentially from the free surface to the sea bottom, the primary microseism can become undetectable outside shallow water. This implies that sources near coastal regions generate primary microseisms [17]. While Ferretti et al. [18] state that the frequency content of the primary microseism ranges from 0.05 to 0.1 Hz, Borzi et al. [19] suggest a $[13, 20]$ seconds period, which translates to frequencies in the range of 0.05 to 0.077 Hz by the formula $f = \frac{1}{T}$, where f is the frequency, and T is the period.

The secondary microseism features frequencies approximately twice that of ocean waves, referred to as double-frequency. According to Longuet-Higgins [20], this phenomenon arises from the interaction of two equal wavelength ocean waves travelling in opposite directions, which produces microseismic oscillations on the ocean floor with double-frequency. These waves propagate with very low attenuation and eventually convert into microseismic energy. The secondary microseism spans frequencies from 0.1 to 0.5 Hz according to Ferretti et al. [18]. On the other hand, Borzi et al. [19] indicate a $[5, 10]$ seconds period, which translates to a frequency range of 0.1 to 0.2 Hz.

The secondary microseism can be further subdivided, since distant and local winds cause two microseism peaks within this double-frequency vibration. Short-period secondary microseisms have a frequency range of 0.2 to 0.5 Hz and are caused by local winds. Long-period secondary microseisms have a lower frequency range of 0.085 to 0.2 Hz, and are caused by sources further away from the coast, such as the swell of distant storms [21].

A key parameter widely used in the context of sea data is the significant wave height. It is an indicator of sea-state severity [22]. The SWH is defined as the average height of the highest one-third of waves in a given sea state, over a specified time [23]. It is calculated from the zero moment (m_0) of a non-directional wave spectrum (the wave elevation variance), and is given by Equation 2.1 [18].

$$H_{\frac{1}{3}} = 4\sqrt{m_0}, \tag{2.1}$$

where m_0 is given by

$$m_0 = \int_{f_{\text{min}}}^{f_{\text{max}}} \frac{2}{T}|S(f)|^2 \, df \tag{2.2}$$

in which $\frac{2}{T}|S(f)|^2$ is the power spectral density (PSD) of sea waves, and f_{min} and f_{max} are the minimum and maximum frequencies of integration.

2.3 Regression Models

In the context of supervised learning, regression models aim to predict a continuous output variable based on input features. This subsection outlines the architectures and working principles of four widely-used regression algorithms: RF, ANN, KNN, and LGB. Each model is described with a focus on its suitability and mechanism for regression tasks. For a more detailed explanation, we refer the reader to [24] and [25].

2.3.1 Random Forest Regression

RF is an ensemble learning method that constructs a multitude of decision trees during training and outputs the average prediction of the individual trees for regression tasks. Each tree in the forest is trained on a random bootstrap sample of the data, and at each split in the tree, a random subset of features is considered, introducing diversity among the trees and reducing overfitting [26].

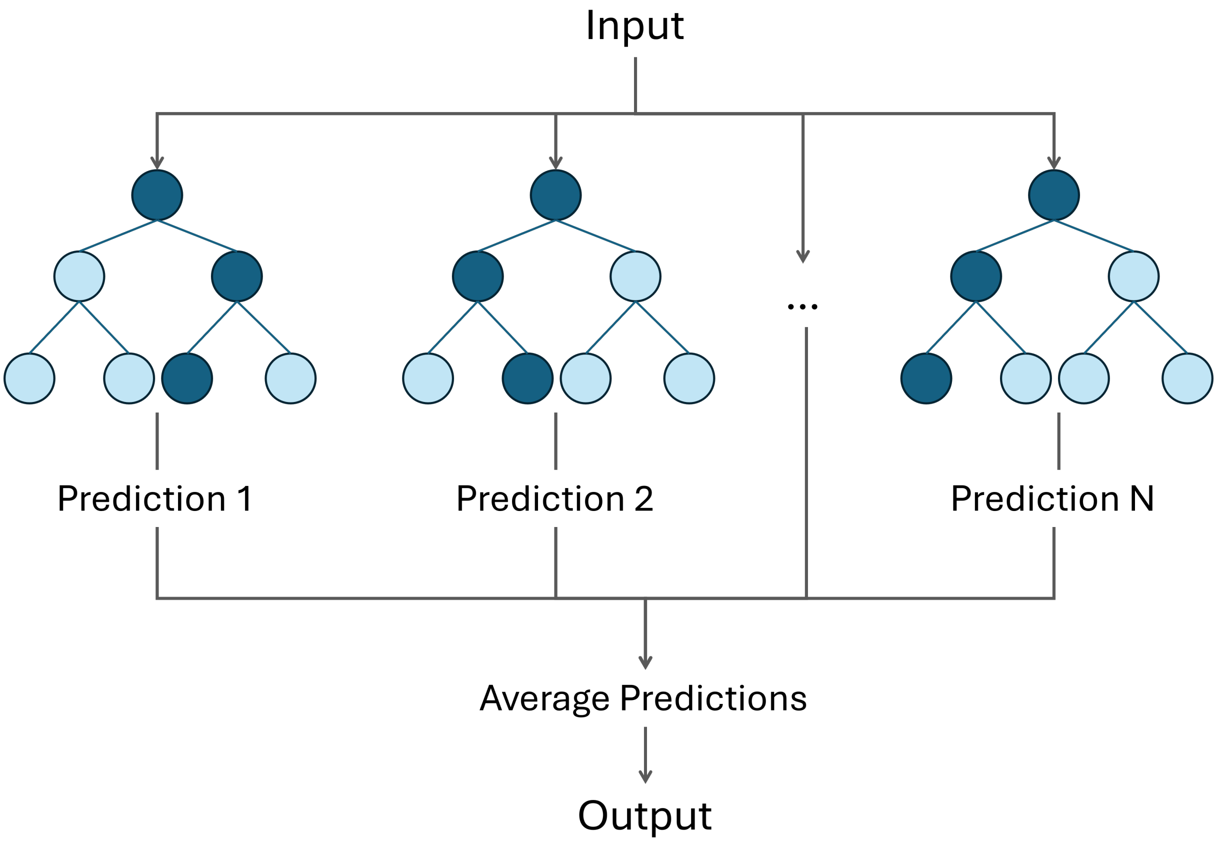


Figure 2.2 Sketch of the RF regressor's architecture.

For regression, each decision tree outputs a numerical prediction, and the final output is computed as the mean of all tree predictions, as shown in Figure 2.2. This averaging process helps mitigate variance and improve generalisation. RF are particularly robust to outliers and can handle high-dimensional data with minimal preprocessing. However, they can be computationally expensive for large datasets and less interpretable than simpler models.

2.3.2 Artificial Neural Networks

ANNs consist of layers of interconnected nodes (neurons), where each neuron applies a weighted sum of its inputs followed by a non-linear activation function. For regression, a typical feedforward ANN architecture is shown in Figure 2.3, and includes an input layer, one or more hidden layers with non-linear activations such as ReLu, and a single output neuron with a linear activation function to produce continuous-valued outputs [27].

Training is performed using backpropagation and an optimisation algorithm such as stochastic gradient descent, with a loss function suited to regression tasks such as MAE. ANNs are highly flexible and can model complex non-linear relationships, but

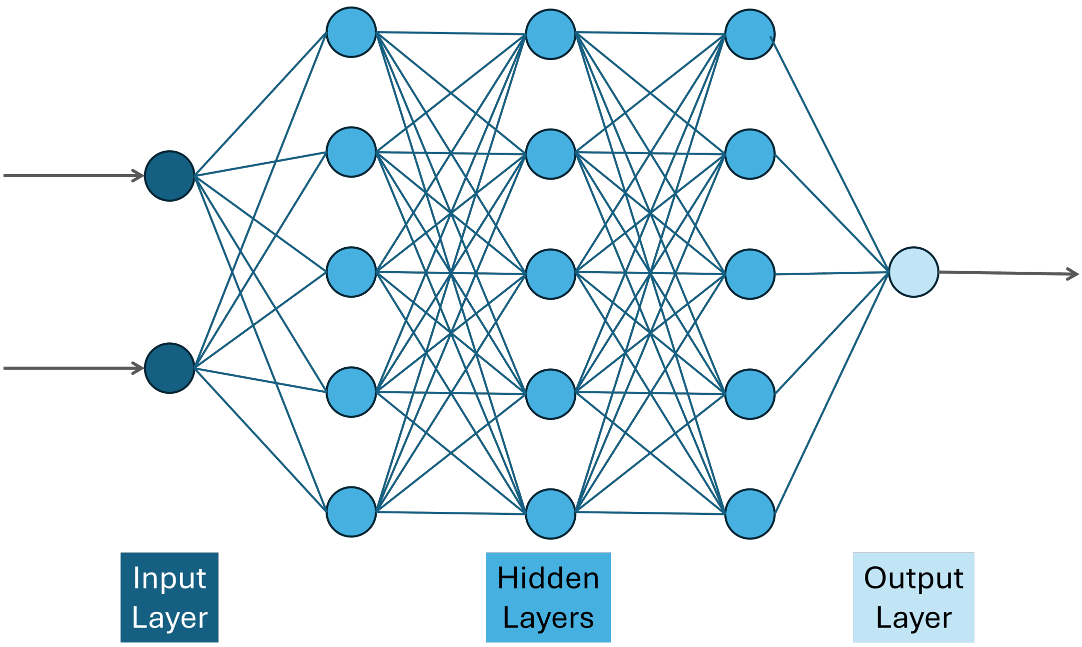


Figure 2.3 Sketch of the ANN regressor's architecture.

require careful tuning of hyperparameters and are susceptible to overfitting if not properly regularised [28].

2.3.3 K-Nearest Neighbours Regression

KNN is a non-parametric, instance-based learning algorithm that predicts the output for a query point by averaging the outputs of the k most similar training instances, where similarity is typically measured using a distance metric such as Euclidean distance. In regression tasks, the predicted value is the mean or weighted mean of the target values of the k nearest neighbours [29]. A sketch of the KNN regression architecture is shown in Figure 2.4.

The strength of KNN lies in its simplicity and lack of training phase; however, it suffers from poor scalability with large datasets due to the need to compute distances to all training samples at prediction time. Additionally, KNN regression can be sensitive to the choice of k and the presence of irrelevant or noisy features and skewed data, which can distort distance calculations.

2.3.4 Light Gradient Boosting Machine Regression

LGB is a high-performance gradient boosting framework that builds ensembles of decision trees in a sequential manner. Unlike traditional boosting algorithms, LightGBM uses a histogram-based algorithm and leaf-wise tree growth with depth constraints, enabling faster training and lower memory usage.

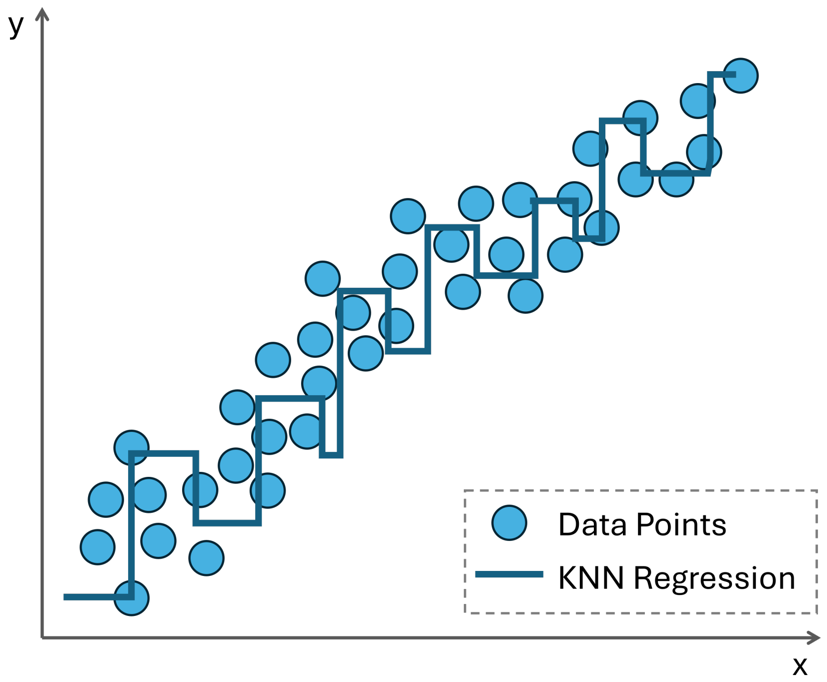


Figure 2.4 Sketch of the KNN regressor's architecture.

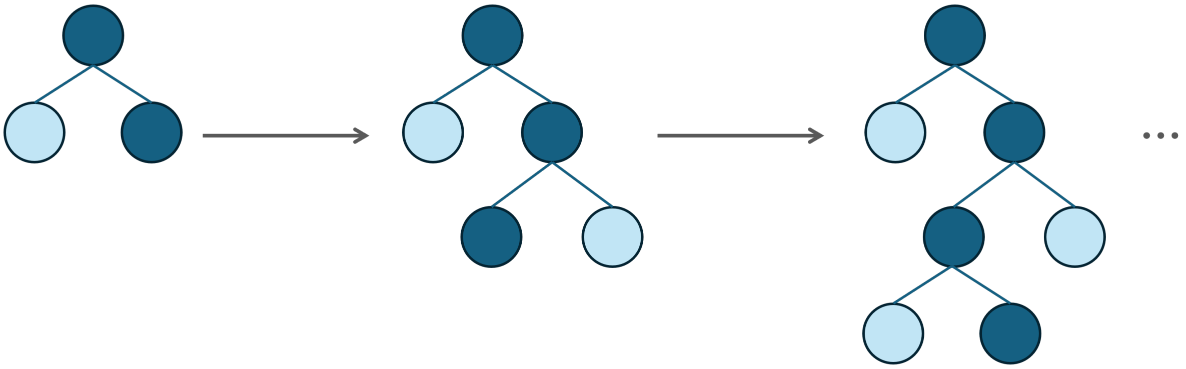


Figure 2.5 Sketch of the LGB regressor's architecture.

In regression, LightGBM minimises a differentiable loss function (typically MSE or MAE) by adding new trees that correct the residual errors of existing ones, as shown in Figure 2.5. Its regularisation techniques help prevent overfitting. LightGBM is particularly effective on large-scale and high-dimensional datasets, due to the balance of accuracy, efficiency, and interpretability compared to other gradient boosting methods [30].

2.4 Gap-Filling Techniques for Time-Continuous Signals

In time-continuous signal processing, particularly in domains such as seismology, missing data can severely compromise the quality of analysis and ML model

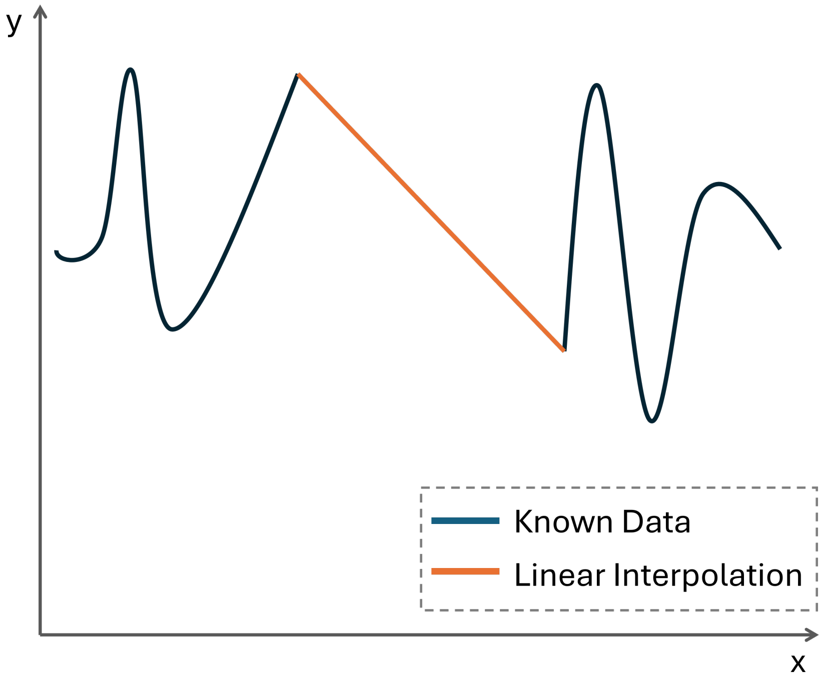


Figure 2.6 Sketch of linear interpolation to impute missing data in a non-stationary signal.

performance. Accurate gap-filling, or imputation, increases data availability while preserving the temporal and statistical integrity of the signal.

2.4.1 Linear Interpolation

Linear interpolation is one of the simplest gap-filling techniques. It estimates missing values by connecting the last and the next known data point with a straight line, shown in Figure 2.6. While it may be appropriate for short, sporadic gaps in a relatively smooth signal, it is not suitable for long temporal gaps in complex and non-stationary signals.

Seismic data typically exhibit non-linear and transient characteristics, including abrupt events, frequency-dependent patterns, and temporal correlations [31]. Applying linear interpolation across extended gaps disregards the signal's inherent dynamics, and can lead to oversimplified reconstructions, suppression of important features, and artificial smoothing. This introduces biases and may distort ML model training or inference.

2.4.2 Deep Learning Based Signal Reconstruction

The inherent complexity and non-stationary nature of seismic signals, coupled with frequent data loss during acquisition, have motivated extensive research into reconstruction methods. Traditional approaches such as those based on wave equation modelling or classical signal processing have been widely employed [32, 33]. However, these techniques often rely on strong assumptions about the signal and can become computationally impractical when applied to large-scale, high-resolution seismic datasets.

To overcome these limitations, deep learning (DL) has emerged as a powerful alternative, capable of learning complex, non-linear patterns directly from data. Deep neural networks can exploit both spatial and temporal structures in seismic signals, making them well-suited for signal reconstruction tasks.

Recent work has explored a variety of architectures. Encoder-decoder style convolutional neural networks (CNNs), such as U-Net, have been adapted to interpolate missing seismic traces by learning hierarchical spatial features and contextual relationships [34]. These models benefit from their ability to reconstruct both local textures and global structures.

In parallel, recurrent neural networks (RNNs), including long short-term memory (LSTM) variants, have been applied to time-series interpolation by capturing temporal dependencies across sequential data [35]. These models are particularly effective when missing values occur over time-continuous intervals, as they can model long-range temporal correlations.

Gap-filling in the context of seismic data presents a non-trivial challenge due to the signal's complex, non-linear, and non-stationary characteristics. Effective preprocessing is critical, as inappropriate interpolation techniques can distort key signal features, leading to the generation of low-fidelity synthetic data. ML models trained on such compromised data may appear to perform well during evaluation, particularly if the validation data shares the same synthetic characteristics. However, this performance is often illusive; when deployed in real-world settings where the data distribution differs significantly, these models tend to generalise poorly. Careful handling of known gaps in seismic data ensures robustness and reliability of ML models applied to seismic data.

2.5 Narrow Artificial Intelligence

In various industries and use-cases, AI models are not a one-size-fits-all solution. On the contrary, research supports the idea that AI models designed for specific sub-categories, contexts, or geographies, often outperform generalised models. This type of AI is commonly referred to as "narrow AI"; it is limited in scope, with the objective of boosting performance. Narrow AI is increasingly effective when diversity within the data or the environment can significantly affect model performance. The following literature highlights such cases and examines the rationale behind localising models, challenging the effectiveness of singular models trained on diverse data.

Samson and Aweda [36] proposed a method for forecasting rainfall across six Nigerian cities, employing separate models for each location. Their approach utilised both RF and LSTM architectures. A key advantage of the LSTM model in this context was its ability to capture long-term dependencies, selectively retaining or discarding information during training. However, the RF regressors demonstrated superior performance, with the MAE improving by 111–127mm across all six cities, compared with the LSTM models.

In a parallel research area, self-driving cars typically comprise multiple narrow AI modules, each designed to execute a specific function within a broader autonomous driving system. Unlike general artificial intelligence, these modules focus on specialised tasks such as detecting obstacles, interpreting traffic signals, or planning routes, making them highly efficient and reliable within their limited domains. This segregation of tasks contributes to safe navigation, particularly in complex urban environments.

- Sensing: understanding the immediate environment through sensor data. Often, the sensors are cameras or light detection and ranging (lidar) sensors, typically making this element a computer vision problem, solved through deep-learning models such as You Only Look Once (YOLO) or CNNs.
  - Obstacle detection: localisation of static and dynamic objects, such as pedestrians, vehicles and roadblocks.
  - Traffic signal detection: recognises and interprets traffic lights and signs.
- Planning: determination of the best course of action to reach the destination. Algorithms such as Dijkstra's are frequently employed to determine the shortest distance between two points.
  - Route planning: computing the high-level path between the current location, and the final destination.
  - Motion planning: creating a safe and feasible trajectory for a vehicle to follow.
- Control: translating the trajectory into physical vehicle behaviour, interfacing directly with the actuators that control the direction, speed, and start and stop motion of the car. Fuzzy logic and deep reinforcement learning-based controllers are commonly used.
  - Motion behaviour: real-time adjustments to follow the planned trajectory while adapting to dynamic changes such as merging vehicles.

Each of these modules operates as a narrow AI system, finely tuned for a specific task within the larger autonomous driving pipeline [37]. Their integration allows the vehicle to perceive its surroundings, make informed decisions, and act accordingly. The modular nature also facilitates independent improvements and testing.

Although the objective of the work included in this section is adjacent to the focus of this dissertation, it aligns with broader trends in AI research. The observed performance gains from using category-specific models highlight that for complex problems, the solution does not necessarily lie in new, more complex architectures, but possibly, in existing, established architectures used wisely.

In summary, this background initiates the notion that microseisms are linked to ocean waves, while providing important background context on the subject. This foundational understanding paves the way for the subsequent sections of this dissertation, where the intricacies of this relationship and innovative methodologies to bridge this gap are explored.

3 Literature Review

This chapter will provide an overview of the current technologies used for estimating wave parameters, including both AI-based methods and other approaches. Adjacent applications, handling similar types of problems are also reviewed since they can serve as inspiration for an innovative proposed solution. While surveying existing technologies for estimating wave parameters, areas for further improvement will also be identified.

3.1 Numerical Methods Approaches

As reviewed hereunder, Ferretti et al. [18] demonstrate that tasks of this nature have been successfully addressed using computational techniques such as Markov chain Monte Carlo (MCMC), showcasing the feasibility and effectiveness of approaches that do not employ AI. Such work is included in the literature as it establishes a foundation for exploring whether AI-based methods can further advance these efforts. Through increased scalability and flexibility, AI holds promise for achieving comparable or superior results in solving similar tasks [38].

Ferretti et al. [18] studied the relation between sea wave height and microseism recordings, following sea storms that caused significant damage in the Ligurian Sea in 2008. They developed an empirical law which establishes the SWH as a function of the PSD of the vertical component of the microseism. MCMC method was used to solve the problem and to calibrate this law for the Ligurian Sea. The equation was derived from Equation 2.1, shown in Equation 3.1.

$$H_{\frac{1}{3}}^{\text{calc}} = 4c\sqrt{\int_{f_{\text{min}}}^{f_{\text{max}}} \frac{2}{T}|S_m(f)|^2 df}, \tag{3.1}$$

where c represents the scaling coefficient that adjusts the seismic signal to the sea wave height, and $\frac{2}{T}|S_m(f)|^2$ is the PSD of microseism.

MCMC is a powerful computational technique used for sampling from complex probability distributions. It is a practical solution for high-dimensional spaces where traditional analytical methods are not feasible. MCMC methods are primarily used for calculating numerical approximations of multi-dimensional integrals. It has proven successful in areas such as parameter estimation and probabilistic modelling [39–41].

Ferretti et al. made use of data gathered between 25 October and 19 November 2008, which encapsulates the aforementioned sea storm that occurred on 30 and 31 October of the same year. An extended data set, consisting of data from January to December 2011 was used for calibration. The microseism data used consisted of three-component velocimetric recordings, with a sampling frequency of 100 Hz.

The pre-processing steps that were applied involved instrumental correction to eliminate any effects introduced by the instrument during recording, ensuring that the recorded signal accurately reflected the ground motion. This was followed by resampling of the signal to a 2 Hz frequency.

Any offsets and linear trends in the data were removed, eliminating long-term drifts and ensuring a stable baseline [42]. The continuous seismic signal was then divided into 1-hour windows, aligning with the sampling intervals of buoy data.

The Fourier transform was subsequently applied to each 1-hour window, converting the time-domain signals into the frequency domain and revealing the signal's spectral components [43]. Spectrograms were generated to visualise the variation of the signal's frequency content over time. Finally, the polarisation of the microseism signal was estimated. Polarisation analysis helps understand the directionality and nature of the seismic waves; a process not relevant to this research topic.

Cross-correlation analysis revealed a two-hour delay between a microseism and the corresponding change in sea wave height, which may be attributed to the sea basin characteristics and transmission delays. Such insights are interesting since the theoretical background provided earlier suggests that the sea influences the microseism and not vice versa.

The numerical model was then constructed to estimate the SWH. In this work, Ferretti et al. [18] make use of MCMC to search for three constants – $f_{\text{min}}, f_{\text{max}}$ and c – as represented in Equation 3.1. A set of 2,500 solutions were obtained from 100,000 iterations of this method. The best-fitting solution was then chosen for testing.

The first model fit resulted in an average difference between observed and computed wave heights of around 0.26 m. To improve this error, the model was modified since the SWH followed a lognormal distribution, proven via a Kolmogorov-Smirnov test. The updated formula is shown in Equation 3.2, and MCMC is used to estimate a, b, f_{min}, and f_{max}.

$$H_{\frac{1}{3}}^{\text{calc}} = \exp\left(a + b \ln\left(\sqrt{\int_{f_{\text{min}}}^{f_{\text{max}}} \frac{2}{T}|S_m(f)|^2 df}\right)\right), \tag{3.2}$$

where a, b, f_{min}, and f_{max} are the four unknowns of the model.

From the obtained solutions, the best-fitting model was again selected. This method yielded a cross-correlation of 93% between measured sea wave heights and wave heights estimated by the model, with a mean difference of 0.19 m between the observed and computed wave heights. The model did occasionally overestimate observed values by up to 1.75 m, and underestimated wave height by less than 0.5 m [18]. This showed a notable improvement over the 0.26 m error produced by the

Table 3.1 Correlation between measured and predicted sea wave heights for various frequency bands [18].

$f_{min}(Hz)$	$f_{max}(Hz)$	Correlation
0.01	0.15	64%
0.15	0.4	87%
0.24	0.78	93%

first model.

The cross-correlation between measured sea wave heights and predicted sea wave heights revealed a higher correlation in higher frequency bands when compared with the correlation observed in lower frequency bands, as shown in Table 3.1. However, Ferretti et al. suggest that the lower degree of correlation obtained in the lowest frequency band indicates that the low-frequency microseism may originate from a source area outside the Ligurian Sea [18].

The work carried out by Ferretti et al. suggests that establishing a relationship between the microseism and SWH is non-trivial. The clear correlation between the two suggests that it is worth pursuing and investigating. While AI was not employed as part of the proposed solution, the method revealed valuable statistical findings, data processing techniques, as well as a baseline through the error metrics, which can be used to compare numerical approaches with AI-based techniques.

Borzi et al. investigated another extreme weather event in the Mediterranean – storm Helios – which caused plenty of damage in February 2023 [19]. Malta was not spared of Medicane Helios' damage, attracting much press coverage locally [44]. The research carried out by Borzi et al. combines several sources of information to evaluate the relationship between the Medicane and the features of microseism. In doing so, various relevant data processing techniques were applied, leading to pertinent findings. The findings are particularly relevant as they focus on the same geographical area that will be used in the method proposed in this research.

In this work, Borzi et al. used data from 105 seismic stations to perform spectral analysis and localisation analysis by grid search, and obtain the seismic signature of the event.

Spectral analysis entails the transformation of a seismic signal from the time domain to the frequency domain. The existence of computers has facilitated the computations required for this process, making it a widely used method for analysis. Since geophysical phenomena are typically expressed in a frequency-dependent form, such analysis is advantageous. An added benefit of spectral analysis is that it makes use of all of the shape of the signal, in contrast with time-domain analysis, which refers to point measurements such as amplitude or direction of displacement of the signal at various points in time [45]. The spectral analysis performed by Borzi et al. reveals a trend in the amplitude of the microseism signal in the three main frequency bands of microseisms (primary microseism, secondary microseism and short period secondary microseism), which increases further during the Medicane Helios. This suggests that a relationship between the seismic information and the SWH exists.

The term “seismic signature” is commonly used in the geosciences domain to describe distinctive characteristics of a seismic waveform, including its shape, polarity, amplitude, frequency, or phase [46].

To obtain the spatiotemporal distribution of the SWH, wave period and direction, Borzi et al. used data provided by the Copernicus Marine Environment Monitoring Service (CMEMS) through the “MED_SEA_HINDCAST_WAV_006_012” product [47]. The cyclone was spatially and temporally tracked through satellite images (SEVIRI). Additionally, SWH, period and direction of the waves were obtained through High Frequency (HF) radars and again through the wavemeter buoy.

A correlation analysis was also carried out between the microseism RMS amplitude time series and SWH. In contrast with the correlation carried out by Ferretti et al. [18] described earlier, Borzi et al. carry out the correlation analysis for each cell of the hindcast maps during the period under review. In doing so, information is obtained about the spatial variability of the correlation coefficients. The analysis is carried out for different frequency bands, as follows:

1. Primary microseism (PM): 0.05-0.07 Hz
2. Secondary microseism (SM): 0.1-0.2 Hz
3. Short Period Secondary microseism (SPSM): 0.2-0.4 Hz

Figure 3.1 indicates that the correlation observations between sea wave height and microseism over different frequencies follow a similar pattern to the observations shown in Table 3.1 [18]. The correlation is greatest in the higher frequency bands, in both cases. Moreover, Borzi et al. confirm that the SWH from the various sources is aligned with the seismic RMS amplitude trends.

3.2 Artificial Intelligence Solutions

In a comprehensive assessment of contemporary sea measurement technologies, Ardhuin et al. [48] positioned seismic data as a particularly promising and innovative modality for the observation and analysis of oceanic conditions. Their work highlighted the emerging potential of leveraging seismic information to enhance the accuracy and resolution of sea state estimation. However, they also acknowledged significant

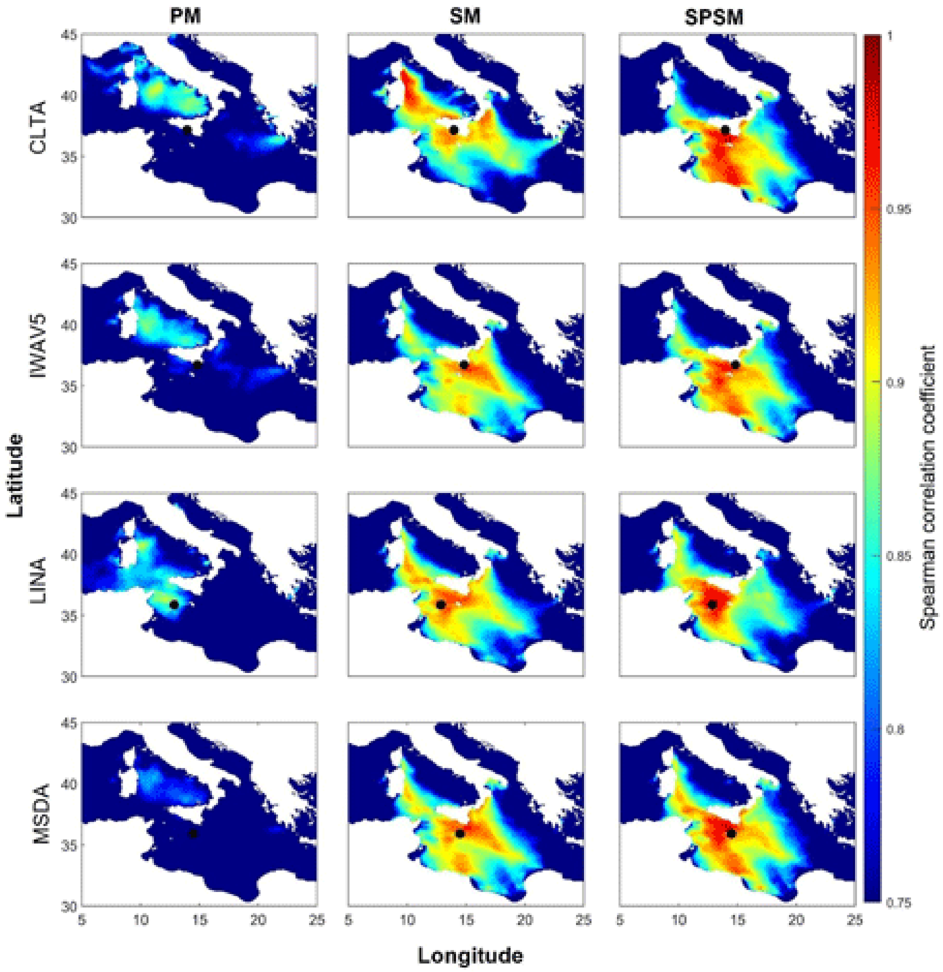


Figure 3.1 Correlation maps for the vertical components of seismic stations Malta (MSDA), Linosa (LINA) and southern Sicilian coast (CLTA and IWAV5) for the PM, SM and SPSM frequency bands over the investigated period [19].

limitations in the integration of such data streams: at the time of their study, robust methodologies to effectively fuse seismic data with other observational sources, such as satellite imagery or in-situ buoy measurements, remained underdeveloped, insufficiently validated, and largely experimental.

A particularly salient point raised by Ardhuin et al. pertains to the geolocation of wave sources within the ocean, which they refer to as an “important question” in the context of estimating key sea parameters [48]. Ocean waves can originate from a wide range of causes, including wind patterns, seismic events, atmospheric pressure shifts, and anthropogenic (human) activities. Disentangling these overlapping signals to accurately attribute wave origins poses a complex, non-trivial challenge which potentially warrants a dedicated research agenda of its own.

From an AI and ML perspective, this raises an intriguing research question: can AI models implicitly learn to generalise across spatial variability without such structured, categorised input? If this holds true, it could suggest a significant departure from traditional physical modelling approaches, potentially streamlining data processing pipelines while maintaining predictive accuracy. This direction opens a valuable avenue for investigation, where ML techniques could be evaluated for their capacity to internalise complex physical patterns from raw or minimally processed data.

In an early exploration of the relationship between ocean microseisms and SWH, Cannata et al. [49] proposed the use of a RF regression model to estimate SWH from seismic data. Although the implementation details were described only briefly, the core methodology involved training the model using the RMS amplitude time series derived from seismic recordings as input features, with hindcast maps of SWH from the same time period serving as the target variable. The study employed K-fold cross-validation to assess model performance and generalisability.

One of the most striking results reported in the study is the model's ability to achieve MAE values as low as approximately 0.1 m, particularly along the Sicilian coastline. This suggests a potentially strong correlation between seismic activity and SWH in this region, possibly influenced by regional geological or oceanographic characteristics.

While promising, the study also reflects a broader trend in early data-driven ocean monitoring research: the use of relatively simple machine learning models with minimal feature engineering or physical domain integration. From an AI standpoint, this research provides foundational groundwork for more advanced modelling efforts. Future work could benefit from exploring this relationship with an improved methodology, providing insight into hyperparameters and formalised evaluation.

On the same theme, Minio et al. propose a method for monitoring the sea state based on microseism using ML [7]. The research question shares numerous similarities to the research of this paper. To this end, it will be a central part of the literature review.

The three models studied by Minio et al. are RF, KNN, and LGB. An RF model is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. It is particularly effective for classification and regression tasks due to its robustness against noise and capacity to handle high-dimensional data. RF works by selecting random subsets of features and data samples for each tree, thereby promoting diversity among the trees and enhancing generalisation [26].

LightGBM is a gradient-boosting framework optimised for speed and efficiency. Unlike traditional boosting methods, it employs a histogram-based algorithm and a leaf-wise growth strategy, allowing it to process large-scale datasets more efficiently while maintaining high predictive accuracy. LightGBM is typically preferred due to its ability to handle categorical features, sparse data, and its superior performance on structured datasets [50].

KNN is a simple yet powerful instance-based learning algorithm used for both classification and regression. It operates by measuring the distance between a point and its K-nearest neighbours in the feature space, assigning the majority label (for classification), or averaging the values (for regression). KNN is non-parametric and highly flexible, making it useful for pattern recognition tasks [51]. However, its performance is heavily influenced by the choice of K and the distance metric, and it can become computationally expensive as the dataset size increases [52].

Minio et al. make use of publicly available raw data. Like Borzi et al., sea state data is obtained through CMEMS [47]. The seismic data was obtained from European Integrated Data Archive (EIDA). Additionally, an earthquake catalogue was used to obtain periods where significant tectonic activities occurred [53].

This method utilised the largest dataset in terms of time period among the methods reviewed, owing to the substantial data requirements of ML techniques. Specifically, four years of data (2018–2021) were analysed; collected from 14 stations. All three components of the data were included: up-down, north-south, and east-west, demonstrated in Figure 2.1. The choice of these stations was based on their short distance from the coastline and minimal gaps during the time period. Moreover, these stations employ broadband sensors, sensitive enough to record the entire microseism band. The importance of utilising data from stations close to the coastline is noted and is logical, as a close distance reduces the risk of introducing secondary noise into the signals. This approach is supported by research indicating that seismic stations near shorelines are more effective in capturing accurate seismic data due to reduced noise interference [54].

In analysing temporal variability of the seismic data, the daily RMS amplitude of the seismic signal was calculated for 14 frequency ranges, four of which correspond to the microseism band.

The region of interest specified in the software implementation[1] was found to exclude a significant portion of the intended study area, particularly when compared to the locations of the seismic stations. This discrepancy is visually represented in Figure 3.2, which highlights a notable lack of sea data in proximity to the MSRU, MPNC, MUCR, and SOLUN seismic stations. Given that the data is publicly available and users have the flexibility to define their own region of interest, there is no apparent justification for this exclusion. The omission of these areas raises concerns about the model's ability to accurately capture the relationship between seismic activity and SWH at the respective locations.

The data underwent thorough pre-processing to prepare it for use in ML models. The hourly RMS of seismic signals was computed for each of the frequency bands, stations and components. This produced 588 features (14 frequency bands × 14 stations × 3 components). Input features with more than 5,000 missing data points

[1] `https://://github.com/VittorioMinio93/shwpredict`, last accessed 3 March 2025.

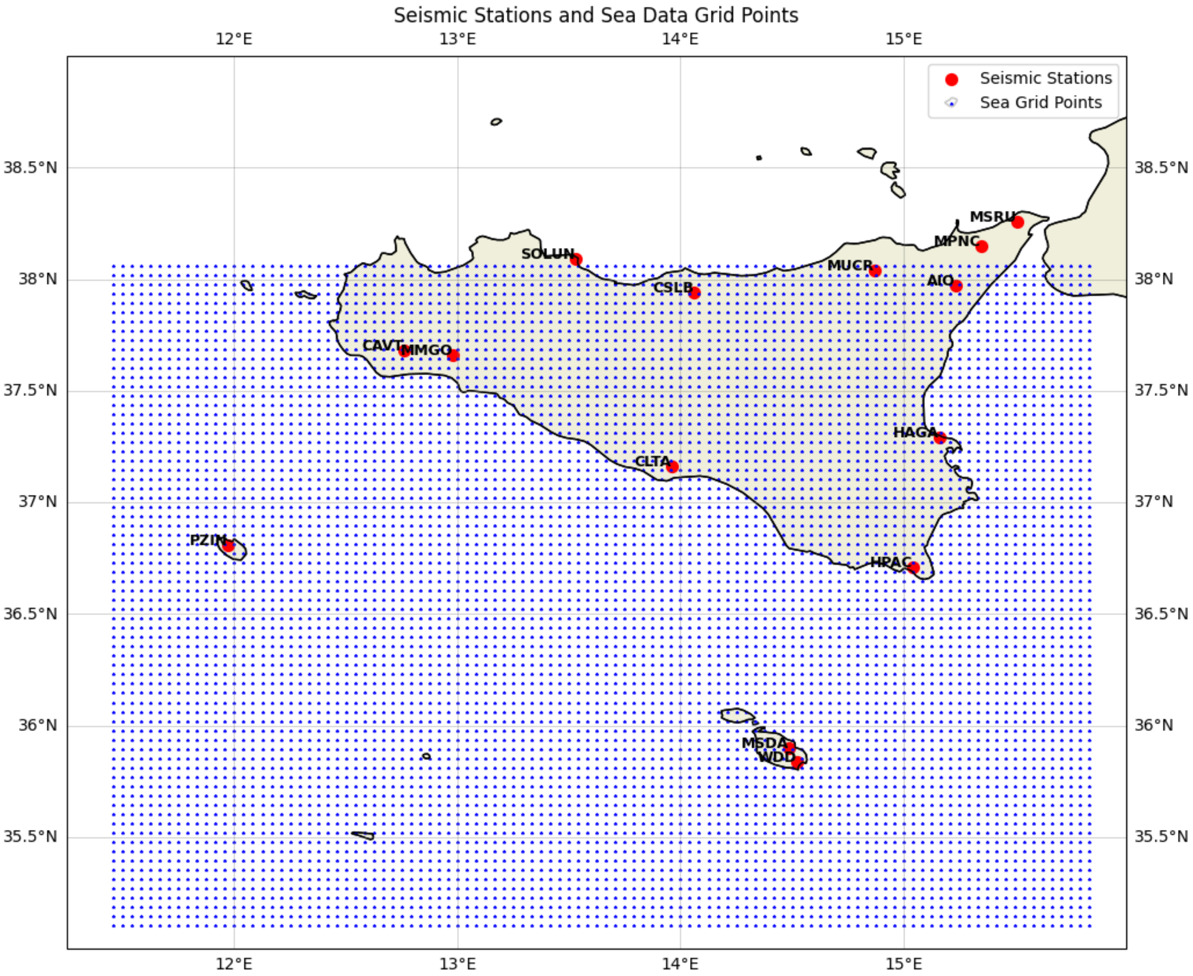


Figure 3.2 The region of interest as defined in Minio et al.'s software implementation [7].

(approximately 14% of all the data) were deleted, otherwise, linear interpolation was applied to fill in the missing information. Instances of significant seismic activity were excluded through the earthquake catalogue, since these include some frequencies similar to microseism signals [55]. To improve the accuracy of the regressors and facilitate the training of ML non-parametric models, Box-Cox transformation was applied to features showing high values of skewness. This was applied to bring the data similar to a normal distribution [7]. Additionally, min-max normalisation was applied.

Box-Cox transformation is applied to highly skewed data sets with the objective of bringing their distribution closer to a normal distribution [56]. The transformed value of a variable y_i is given by Equation 3.3, where λ can be fixed or automatically determined by an optimiser to bring the distribution as close to normal as possible.

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{(\lambda)}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y_i) & \text{if } \lambda = 0 \end{cases} \tag{3.3}$$

where $y_i^{(\lambda)}$ is the transformed observation of y_i.

Min-max normalisation is also a well-known data pre-processing technique, whereby the training data is scaled to the range $[0, 1]$. This ensures that the models are not influenced by differences in orders of magnitude [57]. The min-max normalisation is given by Equation 3.4.

$$y_{i\,\text{scaled}} = \frac{y_i - y_{\text{min}}}{y_{\text{max}} - y_{\text{min}}} \tag{3.4}$$

where y_i is one observation in the data set, and y_{min} and y_{max} are the minimum and maximum values of the training set respectively.

Some discrepancies were observed between the published academic paper and the referenced software implementation. One notable inconsistency was the number of frequency bands considered – while the paper listed 14 ranges, the software implementation considered only 13.

The decision to interpolate missing data when fewer than 5,000 data points are absent in a column is considered risky and problematic. To put this into perspective, should the 5,000 missing data points occur consecutively, this would correspond to approximately 208.33 days – over half a year. The fundamental hypothesis of the study is that SWH can be estimated from the hourly RMS of the seismic signal. However, attempting to infer over 200 days of missing data based on this principle seems highly questionable, especially given the dynamic nature of weather conditions. It would be logically reasonable to linearly interpolate gaps spanning only a few hours, but such a blanket threshold of 5,000 data points appears too broad. Since SWH is influenced by rapidly changing meteorological factors, even small gaps in the dataset could introduce uncertainty. Linearly interpolating such an extensive period disregards the inherent variability of oceanic and atmospheric systems, potentially compromising the accuracy and reliability of the model's predictions. In fact, several studies specifically look into gap-filling of seismic data; it is considered not trivial and merits research in and of itself [58, 59].

The use of Box-Cox transformation on features with high skewness aims to improve the accuracy of the linear regressors and facilitate the training of non-parametric ML models. However, particularly given the choice of models used for testing: LGB, RF, and KNN, this is questionable. While these models can perform well in many cases, LGB and RF are more resistant to skewness, while KNN tends to suffer more when the data is skewed. This is because LGB and RF are based on ensemble learning, building trees by splitting the data based on feature values, while KNN relies on the distance between points to learn, through which a skewed distribution might cause bias [60]. Instead of applying a potentially redundant transformation, it would be more resourceful to select models that are inherently robust to skewness, thus eliminating the need for this preprocessing step.

Similarly, incorporating an earthquake catalogue into the methodology may seem like a logical choice, as it helps account for periods of significant seismic activity occurring outside the region of interest. However, this decision introduced a fundamental limitation within the developed methods. By excluding time periods influenced by seismic events originating outside the region of interest, the system was effectively designed to disregard such occurrences rather than adapt to them. This created a vulnerability, as the model is susceptible to mishandling real-world scenarios where external seismic activity may impact the data.

A more robust approach would have been to retain these periods and focus on identifying optimal models or fine-tuning hyperparameters to account for the effects of external seismic disturbances. A system robust to external influences improves its adaptability and resilience in practical applications. The exclusion of these events, while simplifying the analysis, ultimately restricts the model's ability to function effectively in dynamic, real-world environments where external seismic activity is an unavoidable factor.

The above is further substantiated since adding preprocessing steps increases the overall complexity of the system. If these models were to be used in practice, the steps applied to correct skewness would need to be replicated in real-time for incoming data, adding to the computational expense and processing time, and ultimately adding to the inference time. Moreover, as an added dependency, one would need to receive data locally from the respective stations, as well as internationally, since in the case of significant seismic activity elsewhere, the system outputs would need to be suspended. Such measures could decrease the overall efficiency of the system, which is a critical factor when timely results are required. Therefore, reconsidering the need for skewness correction; opting for models that do not require such transformations, and eliminating the consideration of external seismic events could help streamline the process and improve system performance.

Data segmentation was carried out strategically, due to the temporal nature of the data. The input (seismic) and target (sea wave) data were randomly and temporally divided into several non-consecutive chunks for training and testing [7]. The input and target variables were defined as time series with low autocorrelation decay, meaning that the correlation between values over time decreased slowly over time (with increasing lag). Essentially, observations remain significantly correlated over long distances or time intervals [61]. With such data, there is the risk that the training data is temporally too close to the testing data, which could compromise the model's ability to generalise on unseen data, increasing the risk of overfitting in training [62]. By introducing this element of random shuffling, the risk of overfitting was reduced.

Minio et al. reported using 70% of data for training, 20% for testing and 10% for validation. However, the corresponding software indicates extracting the training set (70%), and subsequently obtaining the testing set by removing the points forming part of the training set from the data set (30%). This adds to the noted inconsistencies between the published work and the software implementation. Supervised learning was used; RF, LGB and KNN regressors were trained upon this data. Grid search technique was employed during training to establish the best hyperparameters, evaluated against MAE on the test sets. No information was provided on the values of the hyperparameters which formed part of the grid search technique. K-fold cross-validation was also applied to avoid overfitting.

The research demonstrates that models were trained using seismic data from all available seismic stations as input, with the SWH at each grid cell serving as the target variable. However, this approach raises concerns regarding both efficiency and effectiveness, as it suggests that 588 features were utilised to estimate the SWH across thousands of grid cells (the region of interest contained 7,632 grid cells, covering land and sea). A primary issue is the potential lack of correlation. Minio et al. themselves establish, through the Spearman correlation coefficient, that the correlation between seismic RMS amplitudes and frequency declines as the distance between the seismic station and the corresponding SWH grid cell increases, as shown in Figure 3.3 for station CAVT. Consequently, it remains unclear whether, by way of example, seismic data from a station in the northwest of Sicily can meaningfully contribute to estimating the SWH at a relatively distant location, such as the southeastern coast of Malta.

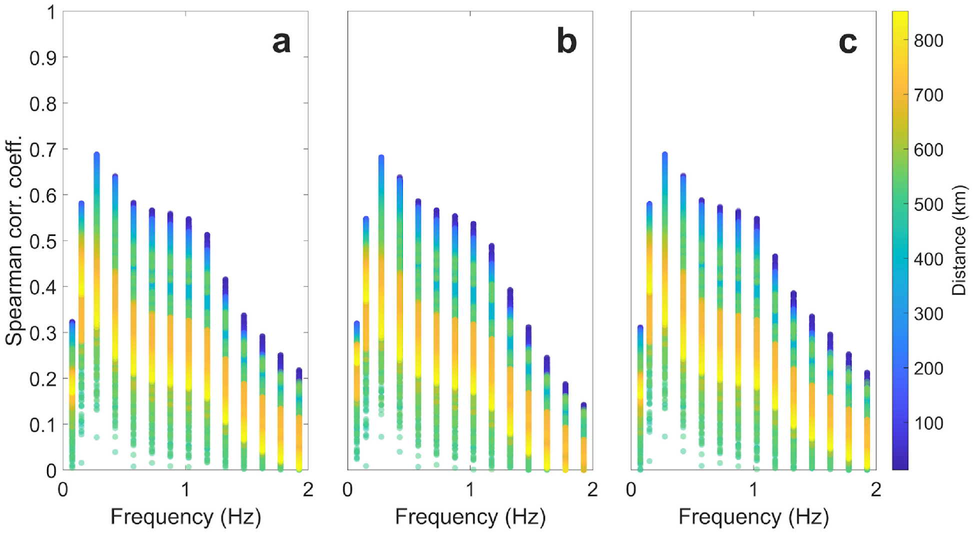


Figure 3.3 Cross-plots showing the Spearman correlation coefficient between SWH and seismic RMS amplitudes for the east-west (a), up-down (b), and north-south (c) components of the CAVT station, in relation to station-grid cell distance and analysis frequencies [7].

Furthermore, Minio et al. describe the utilisation of high-performance computing hardware, explicitly designed for computationally intensive tasks. The system operates on Windows 11 Pro 64-bit and features two AMD EPYC 7713 processors (64 cores), 640 GB DDR4 3200M Hz RAM (expandable to 8 TB), and two NVIDIA A100 40 GB GPUs (PCIe Gen 4). While these resources are undoubtedly valuable, their use in this context may be inefficient. The models' disregard for the limitations imposed by long distances between seismic stations and the target grid cells undermines its own predictive reliability, potentially leading to a misallocation of computational resources.

Minio et al. report evaluating the models separately for normal sea, and extreme sea conditions, where the former's target variable (SWH) is lower than their 99th percentile, while the latter's target variable is greater than their 99th percentile. The reported evaluation metrics are:

- Absolute and relative errors (99th percentile),
- R^2 – coefficient of determination,
- The mean absolute error (MAE) – average absolute model performance error,
- The mean average relative error (MARE) – ratio of MAE to the actual value.

A major benefit of using the R^2 coefficient of determination as an evaluation metric is that as a standalone metric, it provides insight into the model performance. A similar quick analysis cannot be carried out using MAE, since it is a form of aggregation of volume of error. For example, an MAE of 0.65 m does not shed light on whether the model has performed well, or otherwise [63]. The benefit of MAE is that it measures in the same terms as the target variable, and as such, provides high interpretability [64].

The results of the correlation analysis between seismic and sea state data, shown in Figure 3.4, again reveal that the highest correlation in the SPSM band. This affirms the similar findings of Ferretti et al. [18] and Borzi et al. [19].

Minio et al. concluded that, among KNN, RF, and LGB, the RF algorithm was the most effective method for producing reliable predictions based on the R^2 coefficient and error metrics. This superior performance was attributed to several key factors. Firstly, RF is less sensitive to parameter selection compared to other machine learning models, reducing the need for extensive hyperparameter tuning [65]. Secondly, it demonstrates a high degree of robustness to outliers and noise, which is particularly beneficial when working with real-world environmental data that may contain inconsistencies or unexpected variations [26].

Additionally, RF is well-suited for capturing non-linear relationships between input and output variables. This characteristic is especially relevant in the context of predicting SWH, as the relationship between microseism amplitude and wave height is inherently complex and non-linear [66]. The ability of RF to model such interactions without requiring explicit assumptions makes it a strong choice for this application.

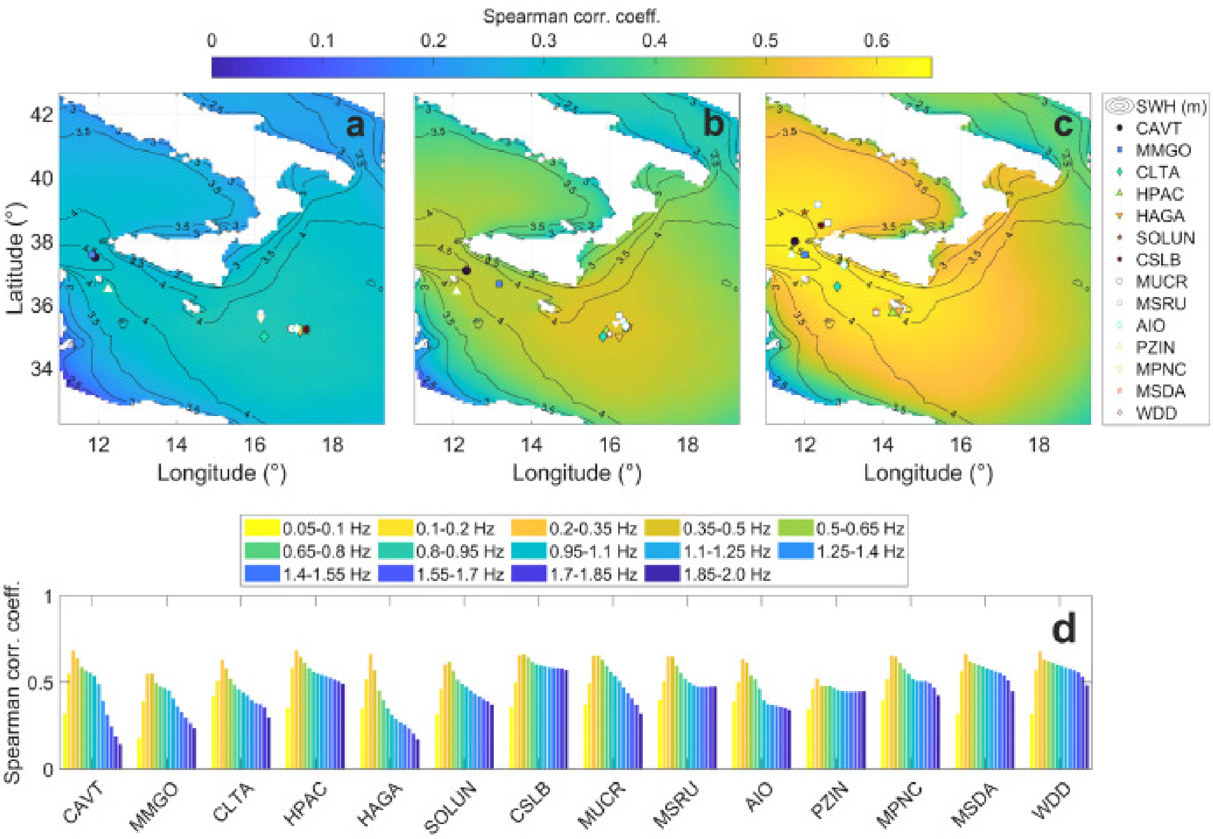


Figure 3.4 Correlation maps for the vertical components of all the stations, averaged, for the (a) PM, (b) SM and (c) SPSM frequency bands over the investigated period, alongside (d) the maximum Spearman correlation coefficients computed between SWH and the hourly seismic RMS for the vertical component of each station and frequency band [7].

Moreover, RF employs an ensemble learning approach, aggregating multiple decision trees to improve generalisation and reduce the risk of overfitting [65]. This ensemble strategy ensures that the model does not overly rely on any single subset of data, leading to more stable and reliable predictions across different conditions.

It is worth noting that Minio et al. assessed these metrics primarily from a geophysical perspective, using visual representations such as maps to illustrate the spatial distribution of errors. However, their study does not contain sufficient emphasis on concrete evaluation metrics, which are typically expected when assessing an AI model. The available baseline is shown in Table 3.2. It was assumed that the unspecified hyperparameters were left as the default within grid search and the software implementation. Evaluation metrics for the LGB and KNN models are not provided in the published work [7].

Furthermore, several challenges were encountered in executing the software as provided, due to the various errors that it raised. Despite significant efforts to troubleshoot and adapt the implementation, certain portions of the code required extended processing times even on a small extract of the dataset. These computational

Table 3.2 Hyperparameters for the RF model trained by Minio et al, and corresponding evaluation metrics [7].

Hyperparameters	Results
RF number of trees: 200 RF number of features: 40 RF maximum tree depth: 15	$R^2 = 0.89$ Mean prediction error = 0.21±0.23 m

constraints, combined with the discrepancies mentioned, posed substantial obstacles to a proper, applicable replication.

In the concluding weeks of this research, a significant new contribution to the domain was made by Baranbooei et al. [67], who presented a case study from the northeast Atlantic Ocean, specifically near the Irish coastline. Their work explored the relationship between the secondary microseism and the SWH from continuous seismic recordings.

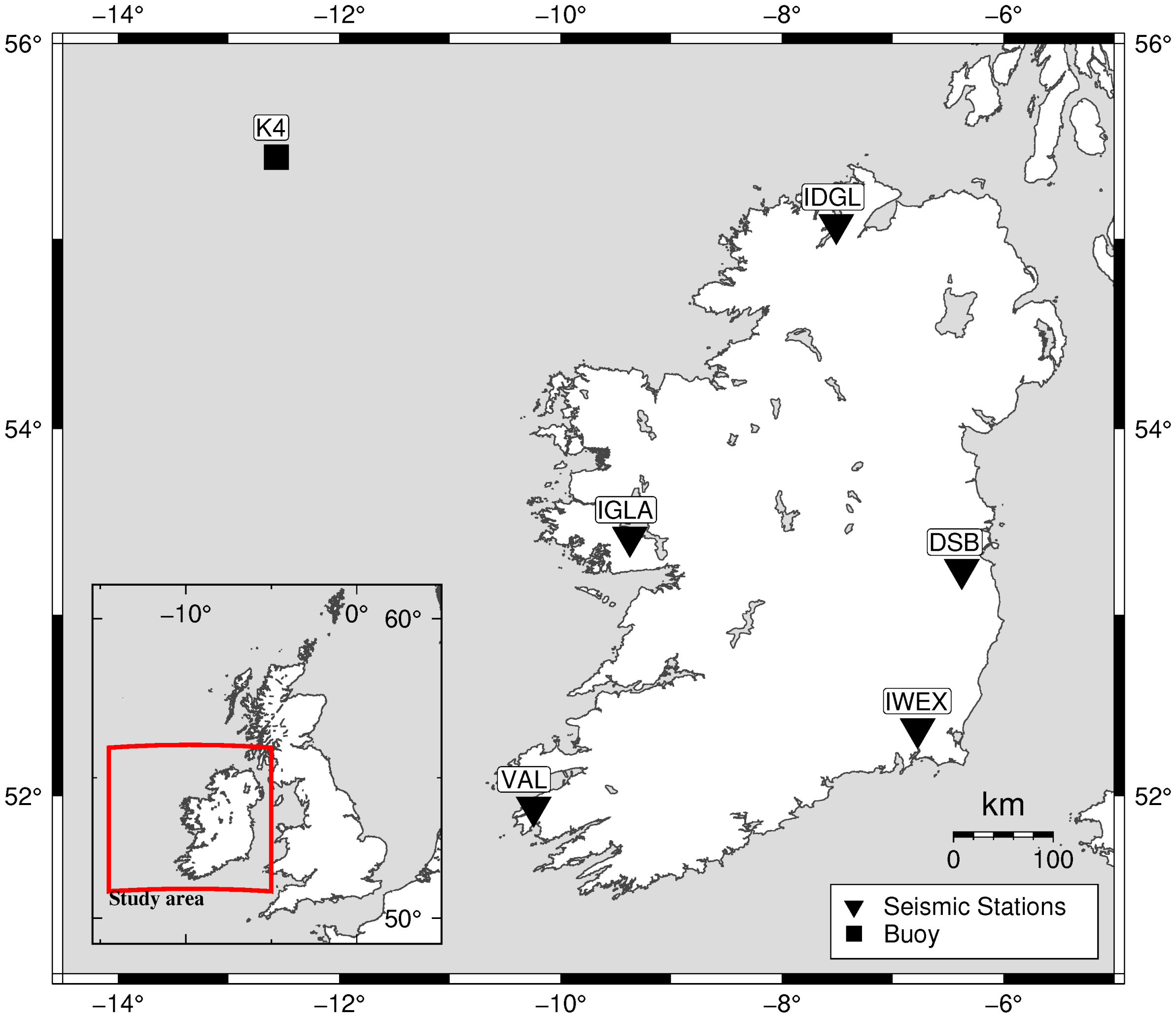


Figure 3.5 The region of interest considered by Baranbooei et al. [67].

The region of interest examined by Baranbooei et al. [67] is illustrated in Figure 3.5, with the buoy location denoted as K4, from which sea state data were collected. Unlike the approach of Minio et al. [7], who incorporated buoy data from multiple locations, Baranbooei et al. relied on a single buoy measurement point. Nevertheless, their deployment of seismic stations around the Irish coastline bears a resemblance to the distributed seismic network used by Minio et al., with stations positioned across several islands within the study area.

Although the precise distances between the K4 buoy and the surrounding seismic stations were not explicitly stated, the spatial distribution depicted in the map suggests that these distances are likely non-negligible. Such separation introduces uncertainty regarding the representativeness of the buoy-derived sea state data in relation to the seismic observations. This spatial decoupling is particularly significant in light of findings from adjacent research, which highlight the multifaceted influences on observed microseisms, stemming from both local and regional sources [68]. As discussed previously, these influences, especially those arising from variations in local bathymetry and geological heterogeneity, can differ considerably across stations. Consequently, models trained under such conditions may struggle to generalise effectively, raising questions about their ability to accurately capture the spatial variability inherent in microseismic–wave height relationships.

Microseismic amplitude data were used as the input signals. These were first corrected for instrument response, and, following the procedures outlined by Ferretti et al. and Minio et al., mean and long-term trend removal was applied to suppress instrument-induced noise. A bandpass filter was then employed to retain only the signals in the $[2, 10]$ second period band (0.1-0.5 Hz). The filtered signals were temporally resampled to match the sampling rate of the corresponding sea state data. The significant microseism amplitude was subsequently calculated as four times the root mean square of the filtered signal. In alignment with the methodology of Minio et al., periods associated with local and regional seismic events were excluded, thereby removing potentially confounding signals from the dataset. Finally, the square root of the significant microseism amplitude was computed to produce the input features used in the subsequent ANN models.

As previously discussed, the exclusion of time windows influenced by seismic events introduces a methodological limitation. While this step aims to reduce noise and improve signal quality, it also results in the removal of data that reflects realistic and potentially informative geophysical conditions. Consequently, models trained on this curated dataset may exhibit reduced robustness when exposed to earthquake-affected signals in real-world applications. This introduces a vulnerability, as the models are effectively deprived of learning from extreme but plausible scenarios, potentially limiting their generalisability and operational reliability.

It is known that ANNs require a complete exposure to different scenarios in training, as they have a limited capability to extrapolate beyond the training set [69]. To this end, the initial dataset comprised approximately ten years of continuous recordings; however, it contained multiple gaps due to intermittent station downtimes and data acquisition issues. To ensure temporal consistency and spatial completeness, only time periods during which all five seismic stations and the K4 buoy were simultaneously operational were retained. Following the removal of earthquake-influenced intervals and other identified outliers, the dataset was further reduced. The final cleaned dataset spanned roughly four years, amounting to approximately 42,500 valid data points.

The study employed artificial neural network (ANN) models comprising five hidden layers, each containing five neurons. Given the non-linear and complex nature of the microseismic data, a hyperbolic tangent sigmoid (tansig) activation function was applied within the hidden layers to allow for non-linear transformations. To improve generalisation and mitigate overfitting, Bayesian regularisation was implemented, optimising the model weights over ten training epochs. This approach offered a practical trade-off between computational efficiency and predictive performance.

The modelling framework was structured around two principal scenarios. Both utilised the same ANN architecture and shared the same input - the processed significant microseism amplitude. In the first scenario, the SWH measured by the K4 buoy was used as the target variable. In the second, the target was the SWH obtained from a numerical wave model hindcast, at the same geographical location as the K4 buoy. This dual-scenario setup enabled an evaluation of the ANN's performance against two distinct but related sources of sea state data, offering insight into its robustness across observational and model-derived outputs.

The results yielded comparable findings across both modelling scenarios. Evaluation metrics for both were computed with respect to the buoy-measured significant wave height, despite acknowledged discrepancies between the buoy data and the hindcast model outputs due to the differing data acquisition methods. A summary of the results is presented in Table 3.3. The first scenario – using buoy data as the target variable – demonstrated marginally superior performance relative to the second scenario, as evidenced by lower error metrics and a R^2. Baranbooei et al. [67] described model performance as generally reliable for wave heights below approximately 10 m. However, they also highlighted that regional variations in bathymetry, wave climate, and seismic wave propagation paths may influence the generalisability and consistency of model performance across different geographic settings.

Table 3.3 Evaluation metrics obtained by Baranbooei et al. [67] for the ANNs trained under both scenarios, with respect to the buoy data.

	ANN Scenario 1	ANN Scenario 2
RMSE	0.8780	0.9505
MAE	0.6132	0.6816
R^2	0.8363	0.8059

Table 3.4 Baseline performance – from literature review.

		MAE (m)	RMSE (m)	R^2
Numerical Methods				
Ferretti et al.		0.19	–	–
AI-Based Solutions				
Cannata et al.		~0.1	–	–
Minio et al.		0.21±0.23	–	0.89
Baranbooei et al.	Scenario 1	0.6132	0.8780	0.8363
Baranbooei et al.	Scenario 2	0.6816	0.9505	0.8059

3.3 Summary of Literature Gaps

The above literature highlights both the strengths and limitations of existing efforts to model the relationship between seismic data and sea SWH. Notably, research in this area remains sparse, particularly from an AI-driven perspective.

Across the reviewed methods, performance was not unreasonable, reporting error magnitudes of around 0.7 m or less, and R^2 values in the range of 0.8 and 0.9, where applicable, as shown in Table 3.4. However, a common limitation was the absence of standardised benchmarks for evaluation, especially within the context of AI methodologies. While certain methods incorporated preprocessing steps such as computing the hourly RMS of seismic signals, offering effective dimensionality reduction without significant information loss, they relied heavily on complex transformations whose necessity remains unclear. In addition, the full geographic range surrounding seismic stations was not considered, and questionable gap-filling techniques were employed.

Collectively, these observations remark the existence of a broader issue in the field: a lack of methodological consistency and benchmarking standards in AI-based approaches. For the domain to evolve further, there is a clear need for more robust evaluations and methodological clarity to establish a foundation for future work.

4 Methodology

This study aims to establish a relationship between seismic data and SWH using AI, building upon and addressing the limitations identified in Minio et al.'s method [7]. The research initiated with an attempt to reproduce their work, however, challenges in replicability, computational feasibility, and method justification led to a series of necessary methodological improvements.

4.1 Overview

The following sections detail the above implementation, describing the implemented process end-to-end; from querying the data to evaluation metrics. The following paragraphs provide a brief overview, introducing the main concepts of the proposed methodology.

A major refinement involved optimising the preprocessing pipeline to make it feasible on standard computer hardware. During the replication attempts, it was noted that Minio et al.'s method was computationally demanding at the preprocessing stage. Moreover, this study shifts towards a generalisable, day-to-day application of the method, rather than focusing on specific extreme weather events. This makes it more practical for real-world use.

To enhance data reliability, raw seismic and oceanographic data were obtained directly from external sources, and preprocessing was performed from scratch. The region of interest was corrected over Minio et al.'s method, since this excluded the SWH from a crucial area near six stations to the north of Sicily. A "longest stretch" algorithm was developed and implemented to reduce the dependency on data interpolation for gaps in the dataset, ensuring that the model relies primarily on real-world observations.

Moreover, a "nearest grid cells" algorithm was developed and implemented to identify the nearest grid cells corresponding to each seismic station, reducing computational load and improving the relevance between the input and target variables. Unnecessary preprocessing steps described by Minio et al. [7] were removed to streamline efficiency for real-world deployment, and test the limits of the subsequent ML models that were implemented.

For each station, 243 combinations of hyperparameters were trained, each for 7 target variables, resulting in a total of 1,701 RF regressors trained for each station. The target variables were the SWH at the five individual grid cells closest to each station, along with the mean and median of the SWH at the same grid cells. The extensive hyperparameter search was conducted to identify the optimal configuration for each station. Further to performance analysis, K-fold cross-validation was applied to all stations, analysing the evidence of model robustness and generalisability.

4.2 Data Collection and Preprocessing

To improve upon the methodology proposed by Minio et al. (2023) in a scientifically rigorous manner, the starting point was to reproduce their approach using the same data sources. This ensures a fair and direct comparison, to the best of our ability, given the limited details available regarding their evaluation.

4.2.1 Data Collection

Raw seismic and oceanographic data were obtained directly from the original sources, covering the period from 1 January 2018 to 31 December 2021. The data is publicly available; however, due to significant gaps in data availability, additional seismic data for Malta were sourced through the Department of Geosciences at the University of Malta, specifically for the WDD (Għar Dalam) and MSDA (Msida) stations.

While the study could have been conducted solely using data from Sicily and Pantelleria, Maltese seismic data was incorporated. This addition was motivated by both scientific and contextual relevance since it increased the region of interest of the research. This effort was also motivated by the research being carried out in Malta through the University of Malta.

Seismic data is freely available from the European Integrated Data Archive (EIDA)[1], through the Istituto Nazionale di Geofisica e Vulcanologia (INGV) network [70]. Sea state data was obtained from CMEMS, a satellite network [47].

The included seismic stations were: AIO, CAVT, CLTA, CSLB, HAGA, HPAC, MMGO, MPNC, MSDA, MSRU, MUCR, PZIN, SOLUN and WDD. The stations MSDA and WDD are located in Malta, PZIN is located in Pantelleria, and the remaining are in Sicily. These locations are shown in Figure 3.2.

For a recreated baseline as alike to Minio et al.'s method as possible, an Earthquake Catalogue was also used to identify periods when significant earthquake activity external to the region of interest may impact the ability of models to perform well [53].

4.2.2 Preprocessing

Like Minio et al., preprocessing began with reading the data for a single station, channel and day. Two detrend processes were applied to the signals. The first was mean

[1] EIDA: `https://www.orfeus-eu.org/data/eida/`, last accessed: 31 March 2025

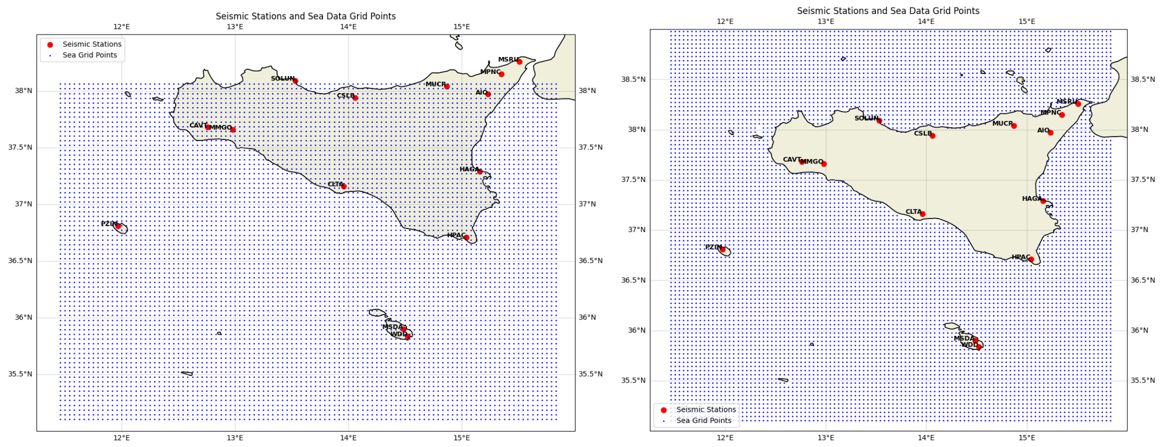
removal, which removed the average of the signal. This eliminated any constant offset that could impact amplitude analysis, which would have been introduced due to sensor drift or baseline offsets [71]. Additionally, linear trend removal was applied to remove any slow, systematic drift that would generally be caused by long-term ground movement or instrument noise [72]. These filters contributed to increased consistency and, thus, comparability across different stations. This preprocessing step was implemented by Minio et al. [7] who implemented an AI-based solution, and by Ferretti et al. [18], who used MCMC to investigate the relationship between the microseism and SWH, as described in Chapter 3.

A bandpass filter was then applied to filter the trace in 13 frequency ranges: (0.05-0.2) Hz, (0.2-0.35) Hz, (0.35-0.5) Hz, (0.5-0.65) Hz, (0.65-0.8) Hz, (0.8-0.95) Hz, (0.95-1.1) Hz, (1.1-1.25) Hz, (1.25-1.4) Hz, (1.4-1.55) Hz, (1.55-1.7) Hz, (1.7-1.85) Hz and (1.85-2.0) Hz. These frequency ranges correspond to the 13 ranges implemented in Minio et al.'s code, as discussed in Chapter 3. The signals were then adjusted to the sensitivity of the respective stations.

The hourly RMS of the seismic signals was then computed. To improve upon the method employed by Minio et al. [7], a simple yet computationally intensive calculation through ObsPy, a seismic signal processing library, was replaced with a lengthier yet efficient implementation of the RMS in NumPy. A sliding window method was employed, padding the signal with zeros at the edge of the trace, where the window exceeded the signal. The data was resampled using the median to obtain the hourly data.

The preprocessing of the SWH data was also modified for improved efficiency. One of the most significant improvements was cropping to the region of interest prior to transforming the data. This eliminated unnecessary transformations, alleviating the computational load on this task. While the same choice of stations as Minio et al. were selected, for sea state data, an initial region of interest for latitudes from 35.10°N to 38.09°N, and longitudes from 11.45°E to 15.87°E, was considered. The region of interest was widened slightly; a conservative region was preferred over a region too small to ensure that all the required data is accessible in further analysis. The improved region of interest over Minio et al. is visually depicted in the map in Figure 4.1.

Chapter 3 discussed how, using seismic information from all the stations as input data, and sea wave height across the entire region of interest, was a questionable approach. The argument was mainly based on the claim of Minio et al. themselves, where correlation between seismic RMS amplitudes and frequency declines as the distance between the seismic station and the corresponding SWH grid cell increases, as shown in Figure 3.3 for station CAVT. Therefore, the proposed method considered multiple models; one for each station. This entailed narrowing down the region of interest per station to the grid cells closest to the respective station. The closest five



(a) The region of interest for sea state data and seismic stations adopted by Minio et al. [7]. No cleaning was applied to remove the grid cells that falling upon land in the image.

(b) The region of interest for sea state data and seismic stations adopted in this approach. The grid cells falling upon land were removed.

Figure 4.1 Maps showing the region of interest for sea state data and the seismic stations.

grid points were established by measuring the distance between each grid cell that lies on the sea, and the station, using the ObsPy `gps2dist_azimuth` function. To reduce computational effort, the region over which the distance was calculated was cropped to 0.5° north and south, and 0.5° east and west of the station. This reduced the area to 7.4% of the total region of interest. The resulting nearest grid cells are shown in Figure 4.2. Without the initial expansion of the region of interest, the nearest grid cells to stations MPNC, MSRU, MUCR, and SOLUN would not have been correctly captured.

By using the five nearest grid cells to each station, only the most relevant information to each station was retained. Figure 4.3 illustrates how the Spearman correlation between the average hourly seismic RMS across all frequency bands and SWH changes with increasing distance between the station and the respective grid cell. The plot reveals that at stations such as HAGA and CSLB, the correlation decreased significantly within the five nearest grid cells. This supports the decision to limit the analysis to these five grid cells, as incorporating additional cells further from the station would introduce data that is more detrimental than beneficial.

In Chapter 3, one of the key critiques of Minio et al.'s methodology was their use of interpolation for up to 5,000 data points, corresponding to approximately 208 days. Any frequency band with missing data (NaNs) exceeding this threshold was excluded from their analysis. However, rather than applying such a rigid cut-off, in this research, an algorithm was developed to identify the longest continuous stretch of data for each station while minimising interpolation.

To ensure high-quality data for subsequent analysis, an algorithm was developed to identify the longest continuous segment of valid data within each seismic station's

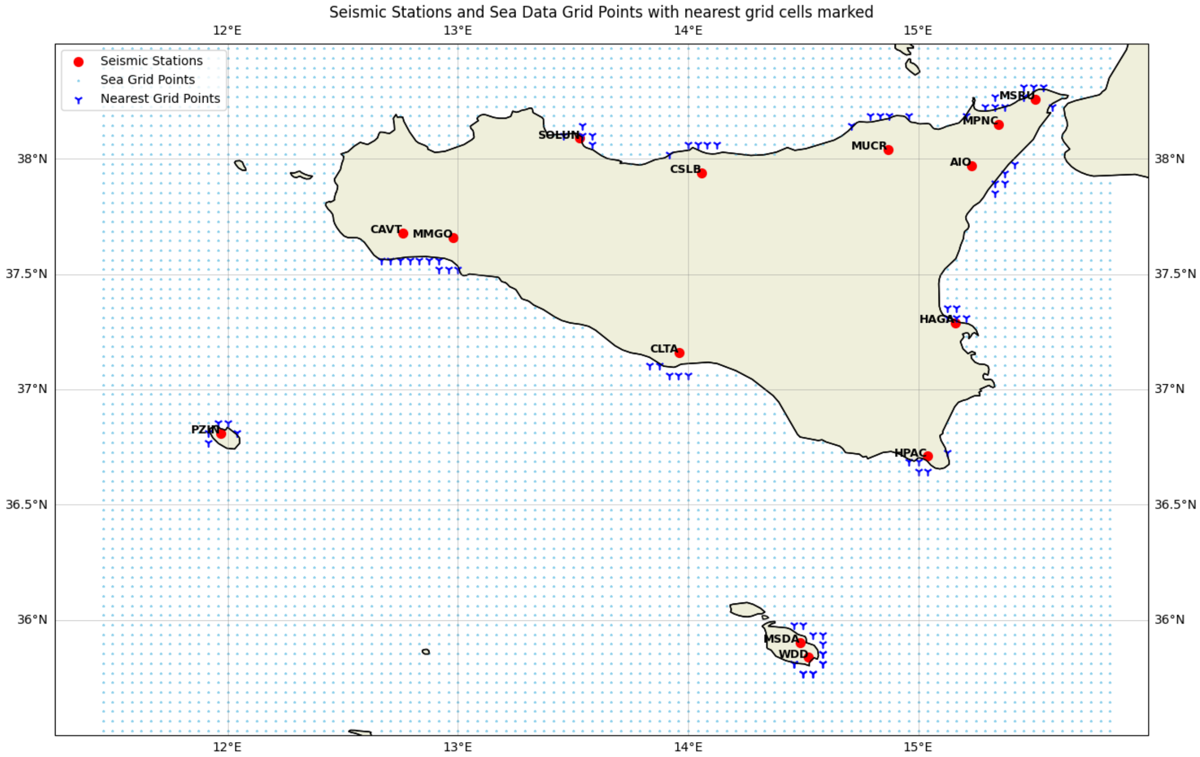


Figure 4.2 The nearest grid cells obtained for each station.

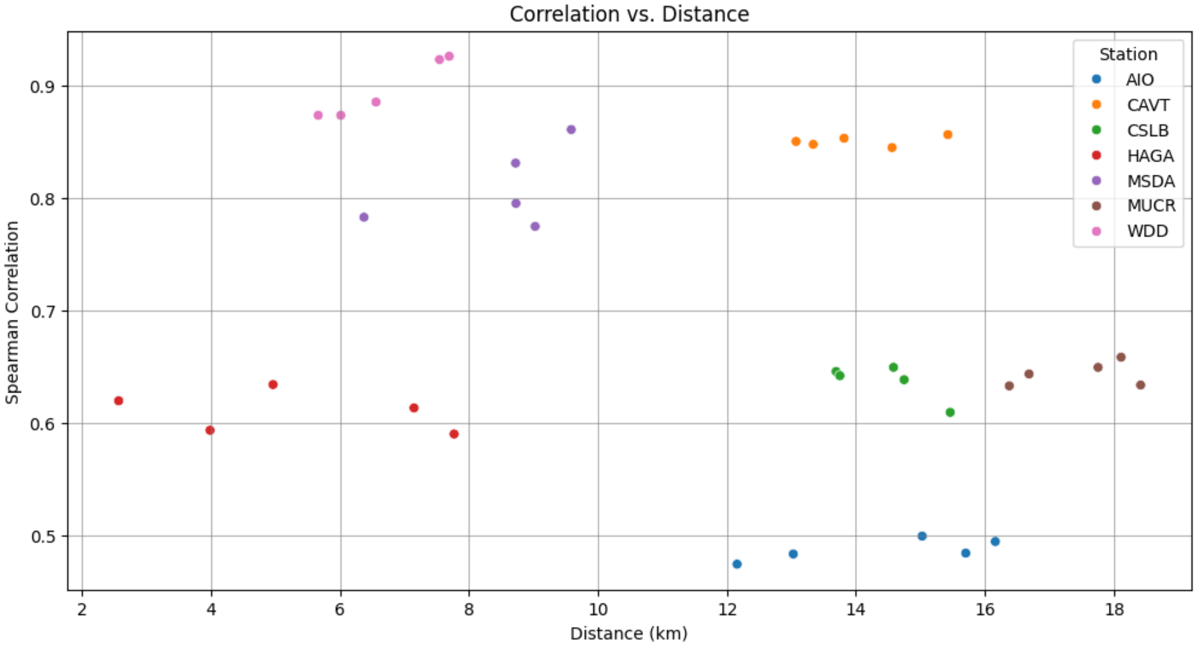


Figure 4.3 The variation of the Spearman correlation coefficient with distance for each station and grid cell considered.

time series, while allowing for small, tolerable gaps. This process is visualised in Figure 4.4. The corresponding pseudocode can be found in Algorithm 1.

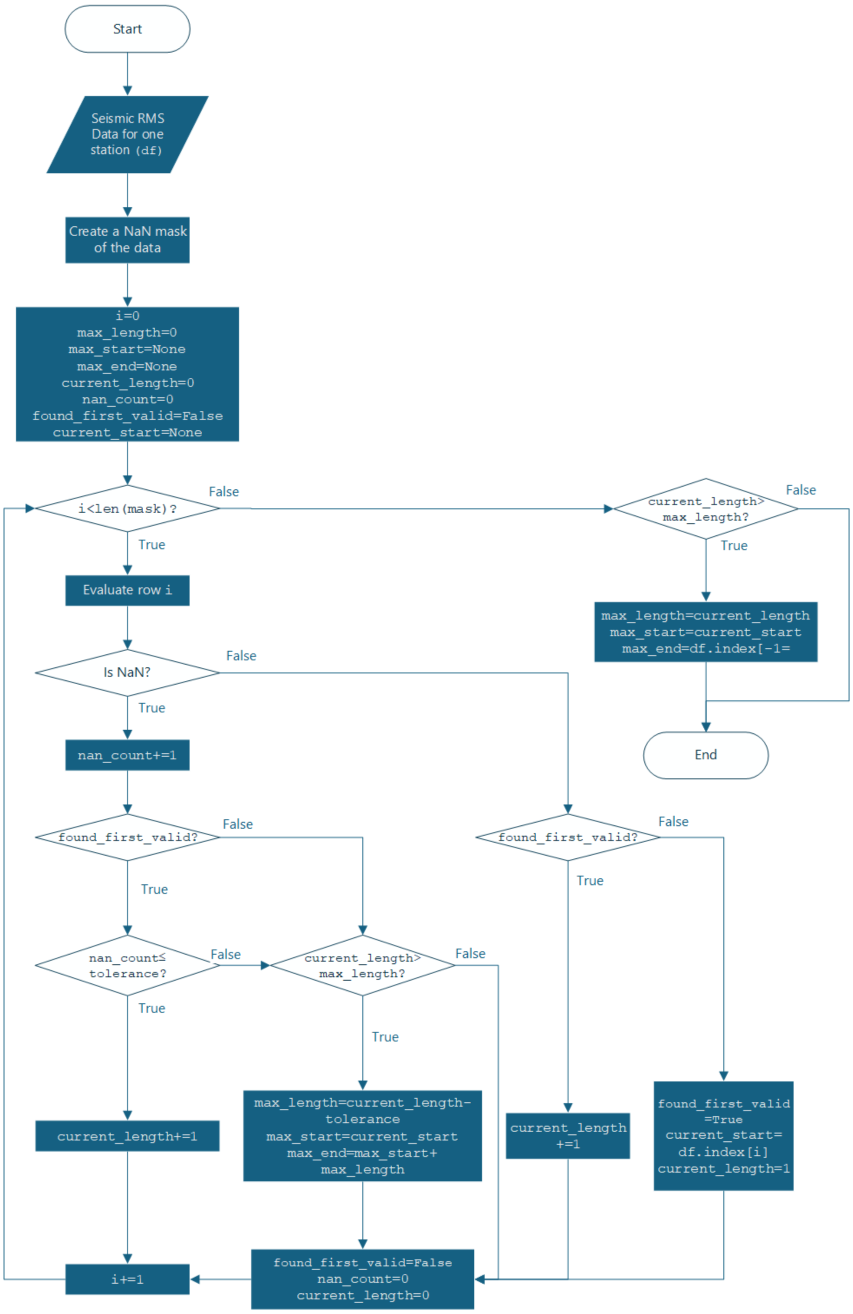


Figure 4.4 Flow chart showing how the longest stretch of data was found for each station for a set tolerance of NaNs.

The idea was to traverse the time series data of each station efficiently (in time linear to the size of the dataset) and locate the segment with the maximum length of consecutive non-missing (non-NaN) data, allowing for occasional small gaps. A tolerance threshold was defined beforehand, representing the maximum number of consecutive NaNs that were to be tolerated within an otherwise continuous segment. For instance, a tolerance of four allowed up to four consecutive NaNs to be treated as

Algorithm 1 Longest stretch algorithm

```
Require: DataFrame $df$, tolerance $t$, column index $c$
Ensure: $(max\_start, max\_end, max\_length)$
$mask \leftarrow$ NaN indicators of column $c$ in $df$        ▷ 1 if NaN, else 0
$max\_length \leftarrow 0$, $max\_start \leftarrow$ None, $max\_end \leftarrow$ None
$current\_length \leftarrow 0$, $current\_start \leftarrow$ None
$nan\_count \leftarrow 0$, $found\_first\_valid \leftarrow$ False
for $i \leftarrow 0$ to $len(mask) - 1$ do
    if $mask[i] = 0$ then        ▷ Non-NaN value
        if not $found\_first\_valid$ then
            $found\_first\_valid \leftarrow$ True
            $current\_start \leftarrow df.index[i]$
            $current\_length \leftarrow 1$
        else
            $current\_length \leftarrow current\_length + 1$
        end if
        $nan\_count \leftarrow 0$
    else        ▷ NaN encountered
        $nan\_count \leftarrow nan\_count + 1$
        if $found\_first\_valid$ and $nan\_count \leq t$ then
            $current\_length \leftarrow current\_length + 1$
        else
            if $current\_length > max\_length$ then
                $max\_length \leftarrow current\_length - t$
                $max\_start \leftarrow current\_start$
                $max\_start\_idx \leftarrow df.index.get\_loc(max\_start)$
                $max\_end \leftarrow df.index[max\_start\_idx + max\_length - 1]$
            end if
            $found\_first\_valid \leftarrow$ False
            $current\_length \leftarrow 0$
            $nan\_count \leftarrow 0$
        end if
    end if
end for
if $current\_length > max\_length$ then        ▷ Handle end-of-data case
    $max\_length \leftarrow current\_length$
    $max\_start \leftarrow current\_start$
    $max\_end \leftarrow df.index[-1]$
end if
return $(max\_start, max\_end, max\_length)$
```

part of a continuous valid segment. The algorithm operated as follows:

1. Initialisation: counters were initialised to track the length of the current valid segment, the length of the longest segment, the current number of consecutive NaNs, and the indices marking the start and end of the segment. A flag was also set to indicate when the first valid data point had been encountered.
2. Iterative looping: the data was scanned point by point. For each data point -
    - If the value was not a NaN, the current segment length was incremented. If this was the first non-NaN value encountered, the start of a new segment was recorded, and the flag was raised.
    - If the value was a NaN, the NaN counter was incremented.
3. Tolerance check: if the value was a NaN, and after the first valid value had been found, the algorithm checks whether the number of consecutive NaNs exceeded the pre-defined tolerance -
    - If within tolerance, the algorithm continued, treating the missing data as part of the ongoing segment.
    - If the tolerance was exceeded, the algorithm compared the current segment length with the longest valid segment recorded so far. If the current segment was longer, it updated the stored start and end indices for the longest segment. In either scenario, the current segment counters were reset.
4. Final comparison: after iterating through the entire data set, the final check confirmed whether the end of the longest segment was the end of the data set, recording it accordingly.

The above algorithm was implemented for several interpolation tolerances/ thresholds. The selection of an interpolation threshold involved a trade-off between two key considerations. On one hand, it was important to minimise the number of interpolated hours, as weather conditions – and by extension, SWH – can vary considerably even over short periods. On the other hand, the approach needed to retain as many stations as possible to ensure a broad region of interest and analysis. By optimising this balance, the resulting models became minimally dependent on interpolated data, thereby enhancing their reliability and alignment with real-world conditions.

Figure 4.5 illustrates how the number of available data points for each station increased with higher interpolation thresholds. For example, by interpolating just eight hours of data, the number of data points for station CAVT increases from 7,142 to

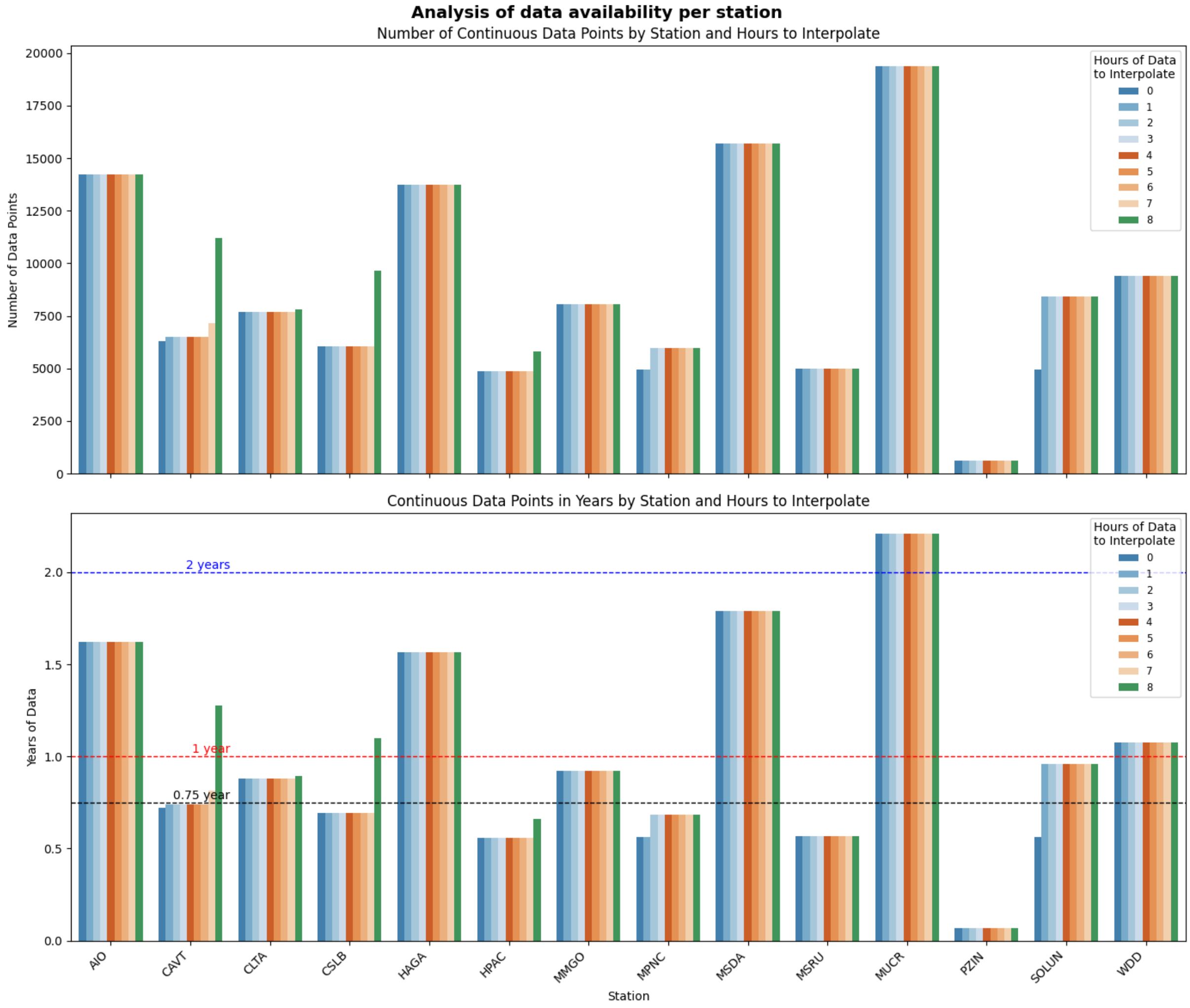


Figure 4.5 The number of data points and corresponding years of data available for varying interpolation thresholds, for each station.

11,201, surpassing the one-year threshold required for inclusion in subsequent analysis. Similarly, station CSLB met the inclusion criteria after eight hours of interpolation, with its data count rising from 6,058 to 9,636 points.

The second subplot in Figure 4.5 shows that more than one year of continuous data was available for stations AIO, HAGA, MSDA, MUCR, and WDD, without requiring any interpolation. By interpolating just eight hours worth of data, stations CAVT and CSLB met the eligibility threshold of one year, and were included in the study, further diversifying the results. Based on this analysis, models were trained on these seven stations, allowing for up to eight hours of interpolation where necessary.

The results of this algorithm were presented within this section for continuity's sake, as all subsequent analysis was carried out from the seven stations: AIO, CAVT, CSLB, HAGA, MSDA, MUCR and WDD. As explained earlier, the algorithm resulted in a dynamic start and end date for each station, as shown in Table 4.1.

Table 4.1 Details on the final data set chosen for each station.

Station	Start Date and Time	End Date and Time	Number of Data Points
AIO	2019-04-24 04:00	2020-12-06 12:00	14,217
CAVT	2019-06-24 08:00	2020-10-03 00:00	11,201
CSLB	2019-08-09 08:00	2020-09-13 19:00	9,636
HAGA	2019-02-18 01:00	2020-09-11 16:00	13,720
MSDA	2019-12-15 22:00	2021-09-29 11:00	15,686
MUCR	2019-05-17 02:00	2021-08-01 09:00	19,376
WDD	2018-05-21 12:00	2019-06-18 00:00	9,421

4.3 Exploratory Data Analysis

The baseline hypothesis of this dissertation is that low-frequency seismic activity recorded near coastlines carries within it the signature of the surrounding ocean's behaviour; specifically, that SWH can be estimated directly from seismic data using AI. This idea seems ambitious, but is grounded in the physical interaction between the ocean and the Earth's crust – a subtle yet measurable relationship that, if captured correctly, could offer a new way to monitor ocean conditions with land-based sensors.

The choice of the stations followed Minio et al.'s selection, which was logical, as they are positioned relatively close to the coastline, an important characteristic when considering the relationship between ground movement and the sea. The distance between the seismic station and the coast was as short as 2.6 km, from station HAGA to its nearest grid cell, and as long as 18.4 km, from station MUCR to its farthest grid cell. The distances between each station and their respective nearest grid cells, along with the average distance across all five grid cells per station is shown in Figure 4.6. The chart shows that the distance is quite consistent across the grid cells nearest to each station, with the largest variance occurring at station HAGA; where the nearest grid cell is roughly 5 km closer to station HAGA than the furthest grid cell. However, since this station also has the lowest average distance, it is not considered to be a concerning factor.

The first step towards validating this hypothesis was to uncover the underlying patterns within the data. This started with exploring the correlation between seismic signals and SWH measurements. Minio et al. use the Spearman correlation coefficient to understand the non-linear dependence between SWH and the seismic RMS amplitude [7]. The Spearman correlation coefficient was calculated between the average SWH at the nearest five grid cells for each station, and the corresponding seismic RMS for each frequency band and channel.

The Spearman correlation coefficients calculated on the data used for training

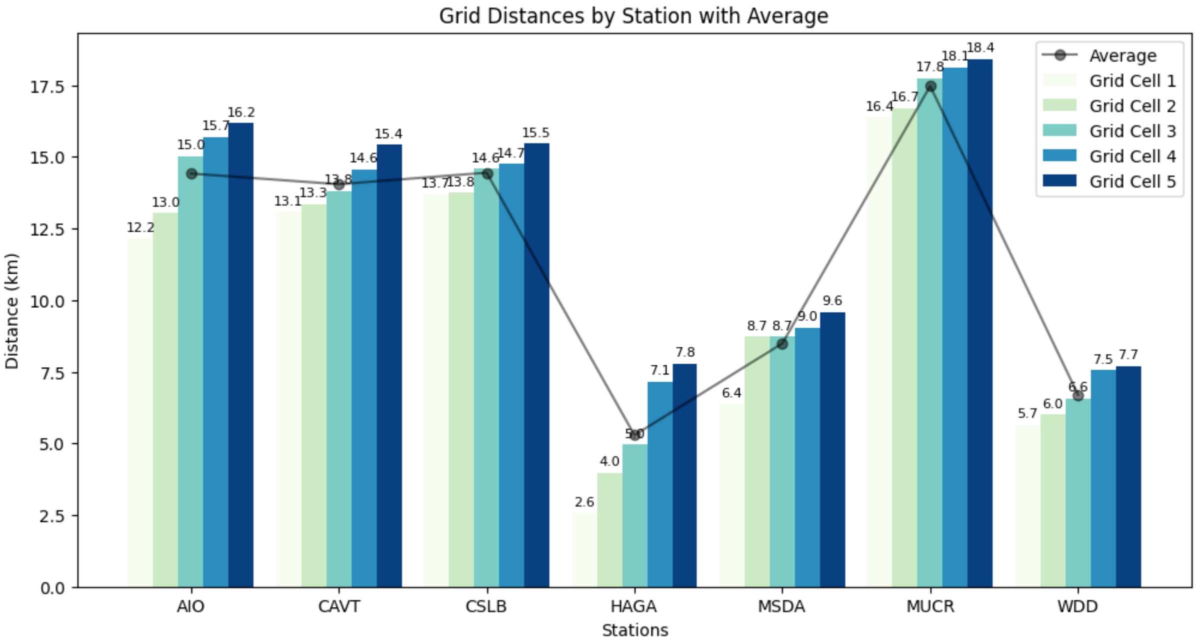
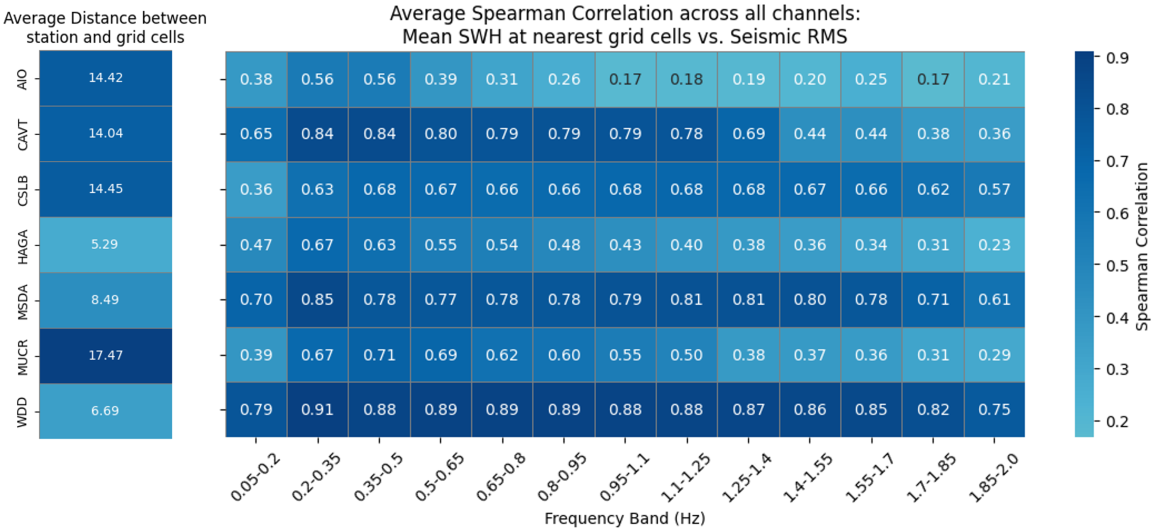


Figure 4.6 The distance between each seismic station and its respective grid cells, alongside the average distance.

are shown in Figure 4.7. The correlation coefficients were calculated for each channel, frequency band and station and subsequently averaged to produce one correlation per station and frequency band. It is imminently clear that the correlation is stronger at certain stations than others over several frequency bands. The highest correlation tends to occur at 0.2-035 Hz and 0.35-0.5 Hz, concurring with the conclusions drawn from the similar analysis conducted by Minio et al. [7], Borzi et al. [19], and Ferretti et al. [18].

Figure 4.7 also shows the average distance between each station and its respective five nearest grid cells. It appears that the distance between the station and the grid cells has a minimal effect on the correlation. This can be seen from stations AIO, CAVT, and CSLB, which are all positioned at relatively similar distances from their respective grid cells (around 14 km), and yet, the correlation ranges from 0.17 to 0.84. However, the stations located in Malta – MSDA and WDD – both have among the highest correlations across the stations and are at relatively close distances to the coast. Lastly, the station closest to its grid cells, HAGA, does not show the highest correlations, and the station furthest from its grid cells, MUCR, tends to show a higher correlation than HAGA at certain frequency bands.

Minio et al. [7] noted that the input features, specifically the seismic RMS, form a time series with low autocorrelation decay. Autocorrelation decay reflects the persistence of temporal relationships within a variable. In this case, the objective was to investigate how seismic RMS values related to their past values over time. It is well understood that sea states often remain elevated after high winds have passed for



	Average Distance	0.05-0.2	0.2-0.35	0.35-0.5	0.5-0.65	0.65-0.8	0.8-0.95	0.95-1.1	1.1-1.25	1.25-1.4	1.4-1.55	1.55-1.7	1.7-1.85	1.85-2.0
AIO	14.42	0.38	0.56	0.56	0.39	0.31	0.26	0.17	0.18	0.19	0.20	0.25	0.17	0.21
CAVT	14.04	0.65	0.84	0.84	0.80	0.79	0.79	0.79	0.78	0.69	0.44	0.44	0.38	0.36
CSLB	14.45	0.36	0.63	0.68	0.67	0.66	0.66	0.68	0.68	0.68	0.67	0.66	0.62	0.57
HAGA	5.29	0.47	0.67	0.63	0.55	0.54	0.48	0.43	0.40	0.38	0.36	0.34	0.31	0.23
MSDA	8.49	0.70	0.85	0.78	0.77	0.78	0.78	0.79	0.81	0.81	0.80	0.78	0.71	0.61
MUCR	17.47	0.39	0.67	0.71	0.69	0.62	0.60	0.55	0.50	0.38	0.37	0.36	0.31	0.29
WDD	6.69	0.79	0.91	0.88	0.89	0.89	0.89	0.88	0.88	0.87	0.86	0.85	0.82	0.75

Figure 4.7 The distance between each seismic station and its respective grid cells, alongside the average distance.

some hours, indicating a dependence on current and previous weather conditions. By extension, if seismic RMS is influenced by sea conditions, a similar temporal dependency would be expected in the seismic signal itself.

The autocorrelation of the seismic hourly RMS decays more rapidly at the lower end of the frequency bands than at the higher end. This is shown in Figure 4.8 for each station. For clarity, autocorrelation values have been normalised between 0 and 1.

The most rapid decay is observed within the 0.05-0.2 Hz frequency band, where autocorrelation drops sharply – from 1 to approximately 0.15-0.45 – within just 20 hours. This indicates that, at these lower frequencies, the correlation between seismic RMS at hour 0 and hour 20 falls to only 15-45% of its initial value. Frequency bands up to 0.5 Hz generally display similar behaviour, with fast autocorrelation decay suggesting a more dynamic, transient signal influenced by changing environmental conditions.

In contrast, higher frequency bands (greater than 0.5 Hz) exhibit a more gradual decay, with autocorrelation values typically remaining between 0.5 and 0.9 after 20 hours. This suggests greater persistence in the signal, likely attributed to locally sourced and stable noise, such as anthropogenic (human) activity. These patterns are particularly pronounced at station CAVT, where high-frequency autocorrelation shows a clear 24-hour cyclical pattern. Station CAVT also happens to be just around 3 km away from the town of Castelvetrano, which contributes to this fluctuation. This observation aligns with existing research, which suggests that daily human-induced noise tends to diminish in frequencies below 0.6 Hz, consistent with the decay behaviour observed in this study [73].

The different decay rates across frequency bands reflect the inherent timescales of the dominant seismic noise sources. The rapid decay in lower frequency bands suggests that the underlying processes, such as wave impact on the coastline, are more short-lived and vary quicker. In contrast, the slower decay at higher frequencies implies the presence of more temporally stable sources, such as continuous anthropogenic activity.

Data skewness impacted the model selection of this research, and further to the critical evaluation of Minio et al.'s approach [7] in Chapter 3. The distributions of the seismic RMS for each station and frequency band were obtained. The charts in Figure 4.9 show that the hourly seismic RMS exhibits quite a strong skewness towards zero.

The exploratory data analysis revealed insights into patterns in the data, and provided a basis for the choice of stations, corresponding training data and selection of training models. Subsequent ML models were trained upon the seismic hourly RMS and sea SWH on stations AIO, CAVT, CSLB, HAGA, MSDA, MUCR and WDD.

4.4 Model Selection

The selection of models for this study was guided by a combination of research-driven and practical considerations. Key factors included performance benchmarks reported in existing literature, the low autocorrelation decay observed in the exploratory analysis, the high skewness present in the data, the interference of human noise within the signals, and the need for a computationally efficient solution that can be trained and deployed on standard consumer-grade hardware. These considerations led to an informed decision on a model that balances predictive performance with real-world feasibility.

The results presented Table 3.4 were the only available benchmark results for this study. Recall that Minio et al. [7] make use of a RF classifier, using the hourly seismic RMS data from all the stations as the input, and the SWH at all the grid cells as the target variables. The RF classifier was trained with 200 trees, 40 maximum features per split, and a maximum tree depth of 15 nodes. An R^2 of 0.89 was achieved alongside a mean prediction error of 0.21±0.23 m using this set of hyperparameters.

As understood earlier, anthropogenic activity (caused by human activity such as heavy vehicles and construction) was a recurrent factor in the signal, which contributed to a lower decay, especially at the higher end of the frequency range considered. This was key because the persistent patterns in the data over time did not arise due to intrinsic memory in the seismic signal itself but were caused by the repeated occurrence of similar external conditions, such as human activity. This created regular, structured patterns in the higher frequency components of the signal. Although this

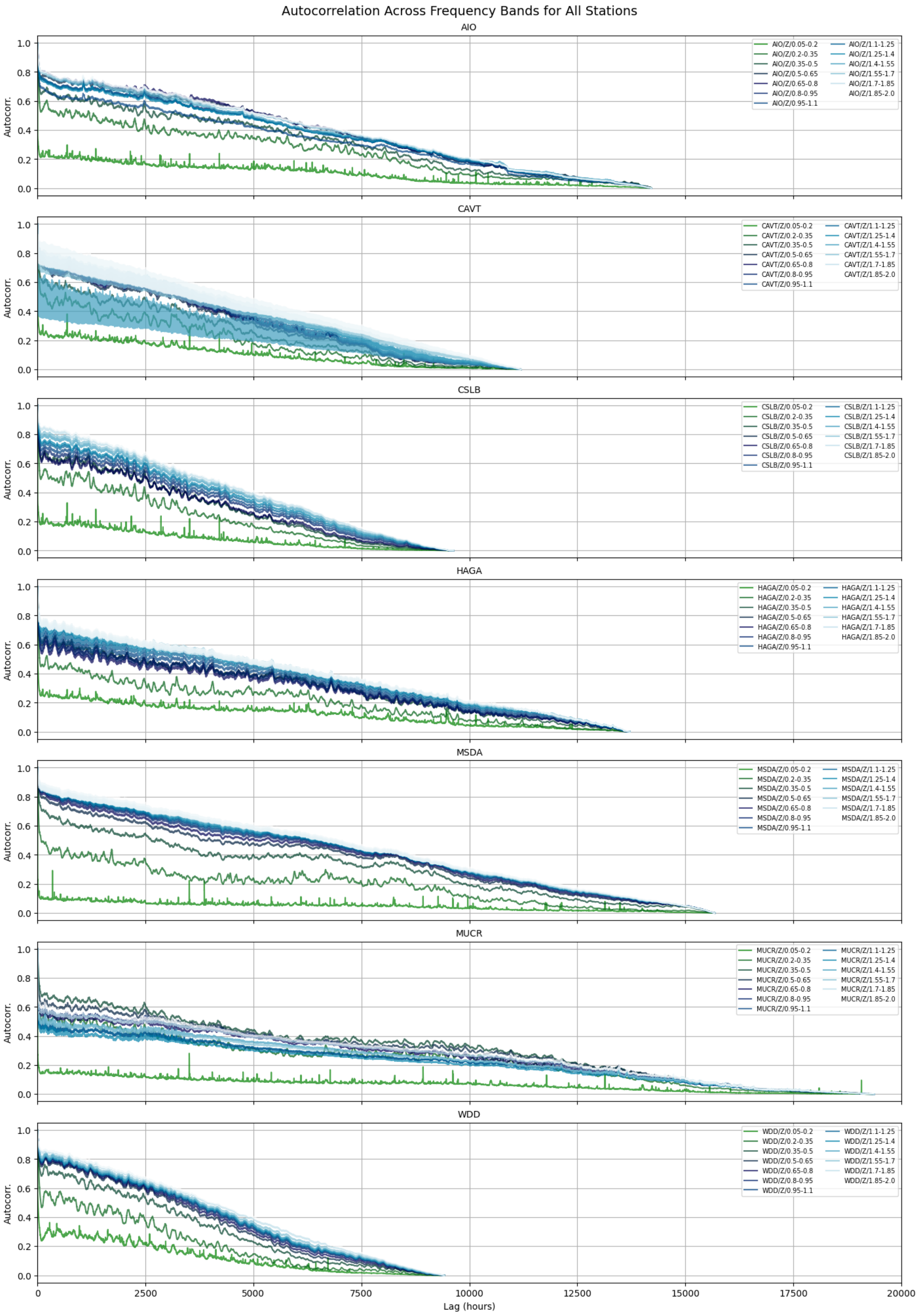


Figure 4.8 The autocorrelation decay for all stations and frequency bands, normalised between 0 and 1 for the Z-channel.

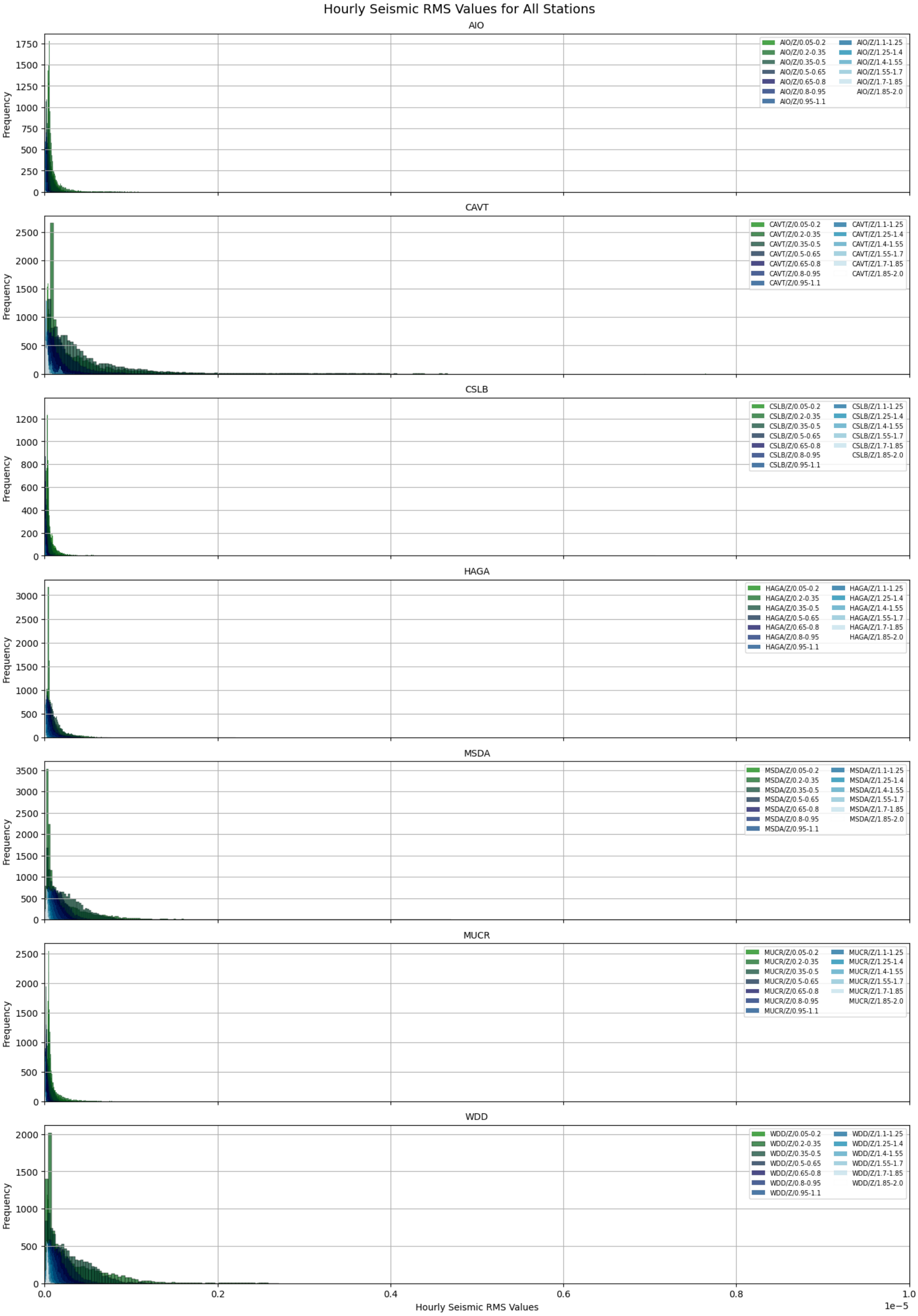


Figure 4.9 Histograms showing the hourly seismic RMS data for all stations and frequency bands for the Z-channel.

mimicked temporal continuity, the underlying seismic movements were better characterised by periodic external inputs rather than an actual long-term memory.

Moreover, a high skewness towards zero can be problematic with the wrong model selection. For instance, linear models such as linear regression perform better when features are symmetrically distributed (close to normal). Distance-based models such as KNN and support vector machine (SVM) are also impacted by skewness since a few large values in a zero-skewed dataset can lead to poor fitness. Transformations, such as the Box-Cox transformation, exist that can be applied to bring the distribution closer to the normal distribution [56].

On the other hand, tree-based models such as decision trees, RF, and extreme gradient boosting (XGBoost) are not influenced by skewness. Since they are based on thresholds defined to split data, rather than by making distributional assumptions, they become a preferred choice for the scenario described within this research, supported by similar problems in other industries and areas [74].

Tree-based models such as RF models are computationally efficient and can be trained and deployed on standard personal computing hardware, making them well-suited for applications with limited access to high-performance computing resources. Additionally, their relatively low memory and processing requirements compared to deep learning models support broader deployment flexibility [26].

This approach attempts to ‘outsmart’ the need for such extensive transformations, to reduce the computational load to the minimum required. Further to all the above considerations, the RF regressor was selected as the ML model for this research, using `scikit-learn` (version 1.5.2).

4.5 Creation of a Baseline

To assess the effectiveness of the proposed modelling approach, a baseline was established for comparison. Minio et al. [7] provide a baseline described in Table 3.4. Although the method used by Ferretti et al. [18] was not AI-based, their best MAE is of 0.19m, which can be used as a baseline to measure changes in performance between AI and numerical based methods. To reproduce and incrementally measure improvement, a separate baseline test was conducted in this study.

The developed baseline involved fundamental signal pre-processing steps, including the extraction of hourly seismic RMS from the raw seismic waveforms and the organisation of the corresponding sea SWH data. The RMS values were then checked against a predefined threshold, with values below 1×10^{-9} replaced by NaN to indicate noise. Subsequently, each feature, corresponding to a specific frequency band at a station and channel, was examined to determine if the number of NaN values exceeded 5,000; set in alignment with Minio et al.'s approach [7]. Features with NaNs above this threshold were discarded from the analysis. Linear interpolation was applied to the remaining features to handle any NaN values that persisted.

For features with a skewness greater than 0.7, Box-Cox transformation was used to correct the data. Additionally, data from the Earthquake Catalogue was retrieved to account for potential seismic disturbances. Earthquake events in the Mediterranean region and globally, with magnitudes exceeding 5.5 and 7.0 on the Richter scale, respectively, were considered to impact the seismic RMS signal. Corresponding data points were removed if they fell within the defined earthquake criteria.

The target variables were identified by applying the nearest grid cells algorithm for each station to obtain five grid cells per station, and the mean and median SWH of these grid cells were computed as additional target variables. This resulted in a total of 39 input features (13 frequency bands $\times 3$ channels) and seven target variables per station. Seven data sets were created for the seven stations included in this research, as listed in Table 4.1.

For each station, 70% of the data were used for training, with the remaining 30% were reserved for testing. To create the training and test sets, the data were divided into 40 non-consecutive chunks. From these, 70% of the chunks (28 chunks) were randomly selected to form the training set, while the remaining 30% made up the test set. This approach was implemented using code directly adapted from Minio et al.'s methodology, and accounted for the temporal variation in the data. An RF regressor was then trained on the training set, using 200 estimators, a maximum depth of 15, and a maximum of 40 features – hyperparameters that Minio et al. identified as yielding the best performance in their study [7].

4.6 Experimental Setup and Hyperparameters

All experiments were conducted on a personal laptop running Microsoft Windows 10 Home (Version 10.0.19045). The system is equipped with an Intel Core i5-8250U CPU (1.6 GHz), 8 GB of RAM, and integrated Intel UHD Graphics 620. The device is an ASUS UX410UAR model. No discrete graphics processing unit (GPU) was used for training. The simplicity of the computational resources extends from one of the objectives: to create models that can work on consumer-grade hardware.

The overarching objective of any ML application is to strike a balance between overfitting and accuracy. This ensures that the model has healthy generalisability qualities to perform well on unseen data. To this end, further to the baseline provided by Minio et al. [7], and that outlined above, a hyperparameter grid search experiment was carried out. The following parameters formed part of the grid search, while the rest were left as the default values.

The max_features hyperparameter, which determines the number of features to consider when searching for the best split at each node in a decision tree, requires a balance between model stability and accuracy. While using a large number of features may seem advantageous, it can reduce diversity among the trees in the ensemble, thereby limiting the benefits of randomisation. Conversely, selecting too few features can result in weaker individual trees due to limited information at each split [75]. To explore this trade-off, the grid search included three values for max_features – 0.5, log2, and sqrt, corresponding to 50% of the total input features, the logarithm base 2 and the square root of the number of input features, respectively.

An important hyperparameter for tuning is n_estimators, which determines the number of trees in the forest. Since the approach aimed to increase diversity by adjusting the sample size of the features at each node, it was crucial to ensure that the number of trees was sufficiently large, potentially larger than the default of 100, to provide accurate predictions [76]. Consequently, the grid search considered n_estimators values of 100, 200, and 300.

The maximum depth of the tree (max_depth) controls the maximum number of decision nodes a tree can have, from the root node to the leaf. Tuning this hyperparameter contributes to the trade-off between accuracy and overfitting. Shallow trees may underfit the data, leading to poor performance on the training set but better generalisation to unseen data. In contrast, deeper trees reduce bias by capturing more complex patterns but tend to overfit, resulting in poor generalisation [77]. To this end, a maximum depth of 10, 20 or 30 nodes per tree was considered.

A search on the optimum minimum samples required to split a leaf node – min_samples_leaf – was carried out. A lower minimum allows more complex trees since there are just a few samples per node, potentially leading to higher variance and overfitting. On the other hand, a higher minimum forces a more conservative approach to splitting nodes, risking underfitting [78]. The default for this hyperparameter is 1 in scikit-learn, and grid search was performed on a minimum samples to split of 1, 3 or 5.

In addition to min_samples_leaf, another related hyperparameter, min_samples_split, crucial in determining how a tree grows was included in the tuning. This parameter controls the minimum number of samples required to split an internal node. A lower value allows the model to create deeper trees with potentially more detailed decision boundaries, increasing the risk of overfitting, while a higher value can lead to simpler models that might underfit by not capturing enough complexity in the data. Grid search was performed on values of min_samples_split of 2, 5, and 10 to identify the optimal configuration for the dataset.

In addition to the method used by Minio et al. to handle the temporal features was retained, bootstrapping was also applied to all experiments. Bootstrapping involved generating random samples of data from the original dataset, with replacement. During training, this process effectively introduced increased shuffling of the data points. Through bootstrapping, the dissolution of temporal features was further reinforced, enhancing the model's ability to generalise.

Research suggests that using a smaller feature sample size (`max_features`), bootstrapping samples for each tree, and increasing the tree depth and other hyperparameter constraints can result in less correlated trees. However, this requires a larger number of trees to achieve convergence [79]. This insight justifies the inclusion of a higher number of trees in the hyperparameter search. In summary, the grid search considered the following hyperparameters, with bootstrapping enabled throughout:

- Number of features to consider when looking for the best split (`max_features`): 0.5, $\log_2$, sqrt.
- Number of trees in the forest (`n_estimators`): 100, 200, 300.
- Maximum depth of the tree (`max_depth`): 10, 20, 30.
- Minimum number of samples to be at a leaf node (`min_samples_leaf`): 1, 3, 5.
- Minimum number of samples to split an internal node (`min_samples_split`): 2, 5, 10.

The core libraries used were scikit-learn version 1.5.2, and obspy version 1.4.1. All libraries and versions used are provided in the requirements.txt file within the source code, which is available at:
`https://github.com/erikasbailey/SeismoWave-MScAI-ErikaSB`.

4.7 Evaluation Metrics and Performance Analysis

Evaluation metrics offer various insights into the performance of ML models, as a means to assess their accuracy and robustness. By looking at evaluation metrics through a test set, the effectiveness was gauged on scenarios that the models were not exposed to during training – a key indicator of the model's real-world applicability and reliability – where hypotheses meet facts. By thoroughly evaluating the models using appropriate metrics, confidence in their predictive power and generalisation ability can be established [80].

One of the primary objectives of this research was to improve upon the evaluation framework used by Minio et al. [7], ensuring a more rigorous and comprehensive assessment of model performance, especially from an AI perspective. To achieve this, a set of standard evaluation metrics, the mean absolute error (MAE), MSE, RMSE, and coefficient of determination (R²), were employed. These metrics provide a holistic view of the model's predictive accuracy, error magnitude, and goodness of fit, respectively [81].

The mean absolute error, given by Equation 4.1 , is a widely used metric for evaluating regression models. It was employed to measure the average absolute difference between observed and predicted values, providing a straightforward interpretation of model error in the same units as the target variable. The root mean square error, given by Equation 4.3, as its name implies, was calculated as the square root of the mean squared error, given by Equation 4.2. Taking the square root of the MSE did not affect how models compare in terms of ranking, but it did convert the error back into the original units of the target variable. This makes the RMSE easier to interpret and more meaningful in a real-world context. In this study, it translated to an error measured in metres, corresponding to the SWH, which is far more intuitive than the square metre unit used in the MSE. As such, both MAE and RMSE offered complementary insights since the MAE reflected the average magnitude of error, and RMSE placed more emphasis on larger errors due to the squaring component [64]. The MAE is also one of the metrics used in the baseline for Minio et al. (refer to Table 3.4), and to this end, a concrete comparison between this research and the baseline was facilitated.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.1}$$

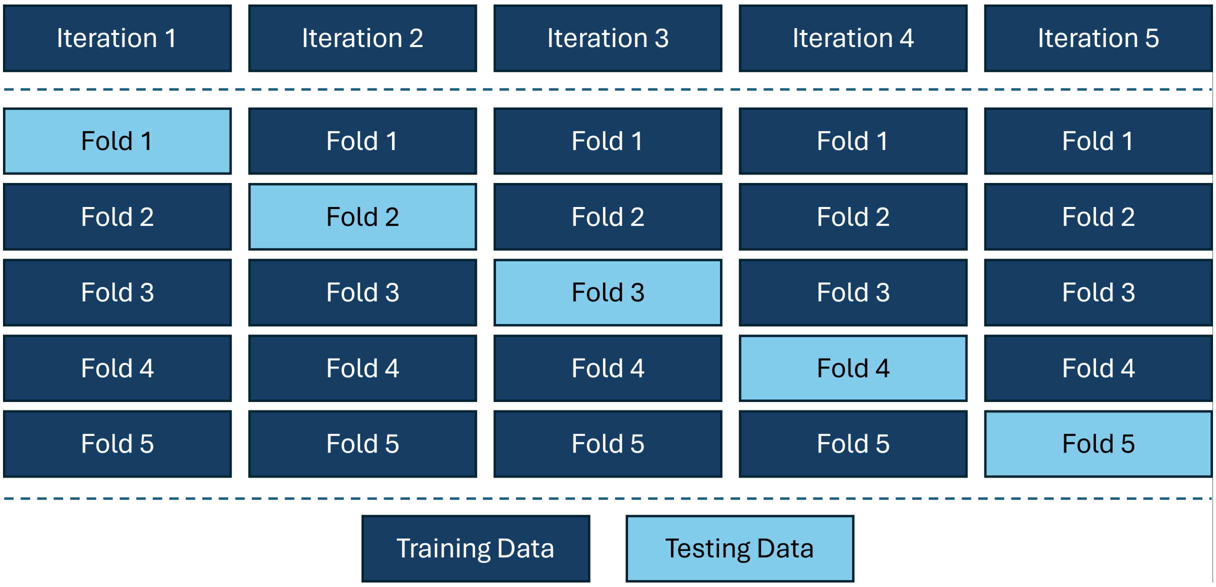
$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}} \tag{4.3}$$

where n is the sample of observations $y_i, i = 1, 2, ..., n$ and corresponding model predictions $\hat{y}_i$.

In addition to the error-based metrics, the coefficient of determination (R²) was also computed. This metric was used in Minio et al.'s baseline evaluation (Table 3.4) and provides a complementary perspective to the MAE, MSE, and RMSE. While the error metrics quantify the magnitude of prediction error, they do not explicitly indicate how well the model captures the variance in the observed data.

The R² metric addresses this by measuring the proportion of variance in the target variable that is explained by the model's predictions. The formula for R², shown in Equation 4.4 is derived from the MSE of the model (numerator), and the MSE between the mean of the ground truth observations and the observations themselves.



Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Fold 1	Fold 1	Fold 1	Fold 1	Fold 1
Fold 2	Fold 2	Fold 2	Fold 2	Fold 2
Fold 3	Fold 3	Fold 3	Fold 3	Fold 3
Fold 4	Fold 4	Fold 4	Fold 4	Fold 4
Fold 5	Fold 5	Fold 5	Fold 5	Fold 5

Figure 4.10 Data split for k-fold cross-validation, when $k = 5$.

It ranges from 0 to 1, unless the model predictions are worse than simply predicting the mean of the target variable, in which case the value of R^2 is negative, with higher values indicating a better fit. Unlike absolute error metrics, R^2 offered a more intuitive sense of model performance relative to the variability in the ground truth, reducing ambiguity in interpretation. This made it especially useful for assessing how well the model generalised on unseen data [63].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(\overline{y} - y_i)^2} \tag{4.4}$$

where $\overline{y}$ represents the mean of the ground truth observations.

Furthermore, K-fold cross-validation was applied to the best-performing stations to ensure that performance estimates were reliable and not biased by any specific subset of the data [82]. This approach enabled a more robust analysis of how well the model generalises to different sets of data. For this set of tests, five folds were used, along with the best hyperparameters resulting from the grid search exercise for that station. A value of $k = 5$ is commonly used in the machine learning literature as it provides a reliable estimate of model generalisation performance without incurring the high computational cost associated with larger values such as $k = 10$. While lower values $k = 3$ reduce computational overhead further, they provide fewer splits and therefore less reliable estimates of model generalisation [83]. Iteratively, a fold was retained for testing, while the rest were used for training, as shown in Figure 4.10. This guaranteed robustness by proving consistent results across all folds, whereby every data point was allowed to form part of the training set and test set.

Mirroring the methodology of Minio et al. [7], the consecutive data were once again divided into 40 chunks, which were then randomly shuffled with a seed of 27.

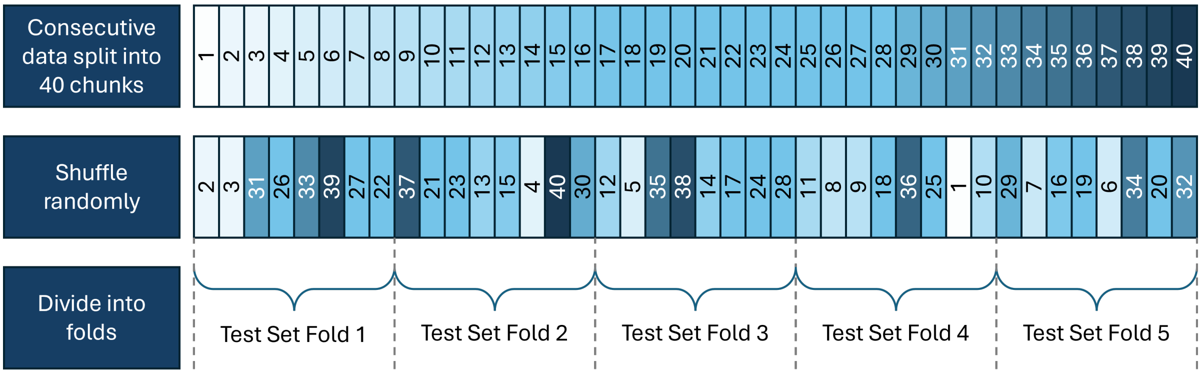


Figure 4.11 Data shuffling for k-fold cross-validation, when $k = 5$.

Out of these, eight chunks were allocated to the test set, while the remaining 32 were used for training. This process is shown in Figure 4.11. This approach was designed to ensure that both training and testing sets included a diverse distribution of data across different times of the year. Without this chunking and shuffling strategy, each fold in a standard five-fold cross-validation would retain a consecutive set of 20% of the data for training, potentially omitting nearly an entire season from the training set – an issue that could limit the model's ability to generalise across seasonal variations. This would also limit the analysis due to the omission of different seasons for each model.

Moreover, an error analysis was conducted for high-error cases, identifying potential causes of poor performance in specific models and stations. This analysis also highlighted opportunities for future research and model improvement.

In summary, the RF regressor was evaluated using a comprehensive set of metrics: MAE, MSE, RMSE, and the coefficient of determination (R^2). These metrics offer complementary insights into model accuracy, sensitivity to outliers, and overall predictive performance. K-fold cross-validation was employed to ensure robust and generalisable evaluation across different data subsets, helping to mitigate overfitting and bias from a single train-test split. Together, these evaluation techniques balanced statistical rigour with practical interpretability, providing a sound foundation for assessing the model's effectiveness.

4.8 Reproducibility and Implementation Details

While the approach presented strives to maintain scientific rigour, certain limitations are acknowledged. A notable shortcoming that may affect reproducibility is that a random seed was not set during the training of the Random Forest model or when splitting the data for the baseline models or hyperparameter tuning. Notwithstanding this, a random seed was set at the K-fold cross-validation stage. As a result, the inherent stochasticity of these processes may affect the reproducibility of results.

The source code is available at `https://github.com/erikasbailey/SeismoWave-MScAI-ErikaSB`, and can be executed by setting the working directory to the main folder. For ease of reference, a flowchart summarising the methodology is provided in Figure 4.12.

In summary, the methodology outlined in this chapter presents a systematic approach to improving upon the work of Minio et al. by extensively refining data and hyperparameter tuning. The key innovations, such as reducing data interpolation and producing a model for each station, contributed to more efficient and robust models. Additionally, the use of error metrics and the coefficient of determination (R^2) ensured a comprehensive evaluation of model performance. Despite certain limitations, such as the lack of a fixed random seed, the approach provides a solid foundation for the subsequent evaluation and performance analysis of the models. The next chapter will present the results of these evaluations, providing insight into the effectiveness of the proposed methodology.

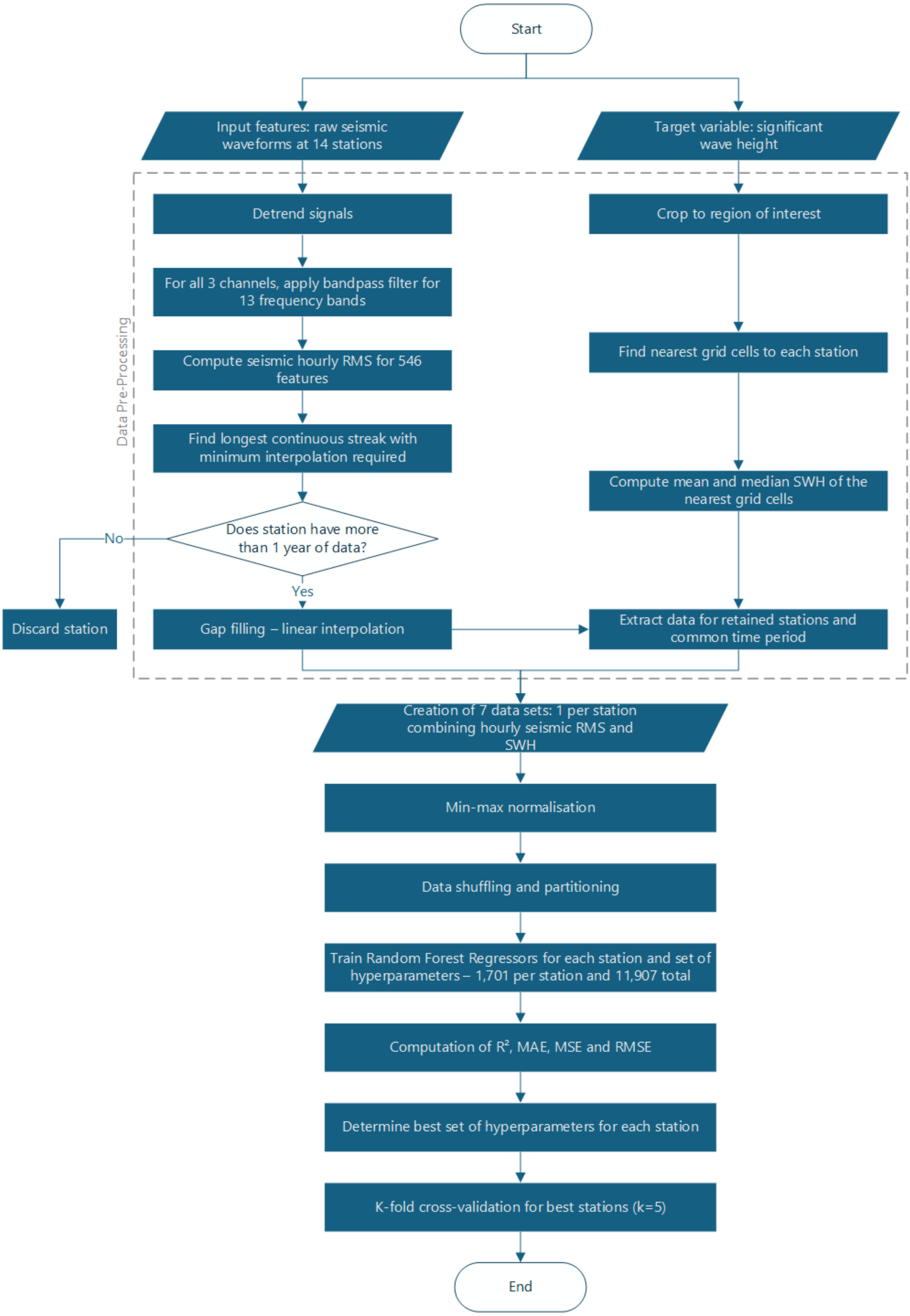

Start
Input features: raw seismic waveforms at 14 stations
Target variable: significant wave height
Data Pre-Processing
Detrend signals
For all 3 channels, apply bandpass filter for 13 frequency bands
Compute seismic hourly RMS for 546 features
Find longest continuous streak with minimum interpolation required
Does station have more than 1 year of data?
No
Discard station
Yes
Gap filling – linear interpolation
Crop to region of interest
Find nearest grid cells to each station
Compute mean and median SWH of the nearest grid cells
Extract data for retained stations and common time period
Creation of 7 data sets: 1 per station combining hourly seismic RMS and SWH
Min-max normalisation
Data shuffling and partitioning
Train Random Forest Regressors for each station and set of hyperparameters – 1,701 per station and 11,907 total
Computation of R², MAE, MSE and RMSE
Determine best set of hyperparameters for each station
K-fold cross-validation for best stations (k=5)
End

Figure 4.12 Flowchart summary of the methodology.

5 Results and Evaluation

To ensure that the evaluation of this work remains grounded in its core aims, it is helpful to revisit the primary objectives established at the outset of the research in Chapter 1:

1. Establish a scientifically robust baseline for the relationship between hourly seismic RMS amplitude and SWH.
2. Develop a cost-effective solution that can run on consumer-grade hardware.
3. Design an efficient pipeline with minimal preprocessing to explore the extent of what AI models can learn from near-raw data.
4. Identify optimal hyperparameters for each station through a grid search.
5. Maximise the use of real-world data while applying minimal interpolation, favouring a *less-is-more* and *quality-over-quantity* philosophy.

To achieve these goals, seismic waveform data were collected from seven stations, each providing three component channels – the Z-channel representing vertical movement, the E-channel representing horizontal, east-west movement, and the N-channel representing horizontal, north-south movement, as depicted in Figure 2.1. The data were processed across thirteen frequency bands to capture relevant spectral features. The target variable, SWH, was estimated at the five closest grid cells to each station, as well as the mean and median values across those grid cells. This pairing of seismic and sea data was performed on an hourly basis. Seismic data were obtained from the EIDA, while SWH measurements were sourced from CMEMS.

The results in this chapter are organised to reflect the progression of the study – from baseline analysis through to model performance, hyperparameter tuning, and K-fold cross-validation. Each section presents the outcomes in relation to the objectives above, followed by a critical evaluation of the findings, limitations, and implications for future work.

5.1 Baseline Model

As an initial step beyond the approach proposed by Minio et al. [7], RF regressors were trained independently across various stations. The methodology, described in detail in Chapter 4.5, closely replicated the original pipeline in order to establish a fair and meaningful baseline for comparison.

To ensure consistency with the previous work, the same preprocessing steps were applied. These included clipping RMS values below 1×10^{-9} to NaN, excluding

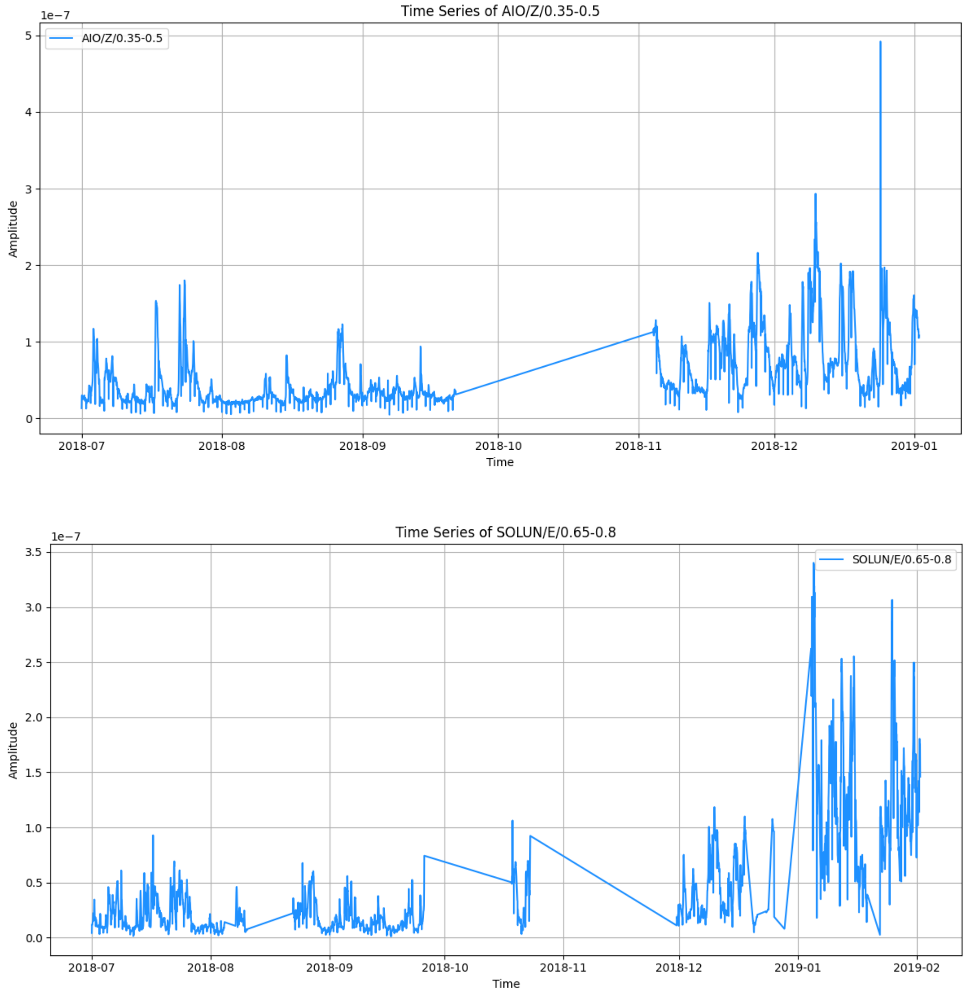


Figure 5.1 Depiction of linear interpolation of hourly seismic RMS over an extended period.

features with more than 5,000 missing values, and applying linear interpolation to the remaining gaps. Through this processing, stations CAVT, PZIN, CLTA, HPAC and MSRU were completely excluded from the analysis, since they had no features with less than 5,000 missing values. The Box-Cox transformation was used to reduce skewness, and periods influenced by regional or distant seismic events were filtered out using an earthquake catalogue. Of the 1,369 earthquake entries retrieved, 30 global earthquakes met the threshold, while no regional Mediterranean earthquakes were identified.

The major change between this baseline and that of Minio et al., is in the development of a model per station, using the data from the nearest grid cells to each station, which were identified using the nearest grid cells algorithm.

The suitability of linear interpolation, as employed by Minio et al. [7], has been extensively evaluated in Chapter 3. To illustrate the practical implications of such a technique, Figure 5.1 presents two graphical examples highlighting its effect on individual features at two representative stations. While the chart displays just two cases, similar patterns were observed across all stations and features. The figure clearly demonstrates that linear interpolation introduces synthetic values that do not accurately reflect real-world occurrences, ultimately providing unreliable input data to the model over extended periods of time.

The RF models were configured with a fixed set of hyperparameters: 200 estimators, a maximum tree depth of 15, and a maximum of 40 features considered per split. These were established by Minio et al. as the ideal hyperparameters [7]. Bootstrapping was enabled, while all other hyperparameters were kept at their default values to mirror the original baseline as closely as possible.

The resulting models served as a reference point for evaluating the effectiveness of the proposed pipeline and its associated improvements, including those related to computational efficiency, improved preprocessing, and hyperparameter optimisation. Performance metrics for each station are presented in Table 5.1. To facilitate analysis and comparison, the results for the five individual grid cells are presented as the mean and standard deviation thereof.

The best R^2 reported by Minio et al. was 0.89, accompanied by a MAE of approximately 0.21 ± 0.23 m [7]. In comparison, the strongest performance achieved in reproduced baseline was an R^2 of 0.869 at station CSLB when predicting the median SWH from the five nearest grid cells, which is comparable to Minio et al.'s baseline. At this location, the model also achieved a notably low MAE of 0.138m, compared with other stations. Averaged across all stations for the same mean SWH prediction task, the baseline yielded an average R^2 of 0.70 and an average MAE of 0.161 m. These results show a slightly lower maximum R^2, but offering improved average predictive accuracy as reflected in the lower MAE, positioning them at a comparable level to Minio et al [7].

Table 5.1 demonstrates that the performance across the grid cells was consistent, with relatively small values for the standard deviation at each station and for each metric. Station MUCR typically had the highest standard deviation in the error metrics, up to 49 cm, owing to the dispersion in the nearest grid cells, as shown in Figure 4.2. In most cases, the best performance was obtained when predicting the mean SWH across the five grid cells nearest to each respective station. A consistent relationship across the R^2, MSE, MAE, and RMSE metrics for each station was also noted. This is especially prominent in Figure 5.2, which depicts how at station MMGO, the lower R^2 is paired with the highest MSE, MAE and RMSE. This consistency supports the credibility of the baseline model evaluations, as stations with higher R^2 values generally exhibited lower error scores and vice versa. This indicates that the models achieved a reasonable balance between predictive accuracy and overall fit for the given set of hyperparameters.

Table 5.1 Table of results for the reproduced baseline.

Station	Metric	Individual Grid Cells	Mean SWH	Median SWH
AIO	R^2	0.348 ± 0.007	0.350	0.344
CSLB	R^2	0.867 ± 0.006	0.868	0.869
HAGA	R^2	0.635 ± 0.016	0.639	0.639
MMGO	R^2	0.329 ± 0.009	0.330	0.331
MPNC	R^2	0.849 ± 0.012	0.861	0.855
MSDA	R^2	0.828 ± 0.036	0.843	0.852
MUCR	R^2	0.833 ± 0.006	0.840	0.837
SOLUN	R^2	0.693 ± 0.012	0.698	0.700
WDD	R^2	0.796 ± 0.062	0.841	0.847
AIO	MSE	0.090 ± 0.009	0.089	0.095
CSLB	MSE	0.045 ± 0.002	0.044	0.045
HAGA	MSE	0.066 ± 0.005	0.065	0.065
MMGO	MSE	0.253 ± 0.019	0.252	0.251
MPNC	MSE	0.034 ± 0.009	0.030	0.031
MSDA	MSE	0.063 ± 0.009	0.056	0.054
MUCR	MSE	0.056 ± 0.020	0.052	0.064
SOLUN	MSE	0.056 ± 0.014	0.054	0.05
WDD	MSE	0.095 ± 0.013	0.067	0.082
AIO	MAE	0.210 ± 0.009	0.209	0.215
CSLB	MAE	0.137 ± 0.002	0.137	0.138
HAGA	MAE	0.158 ± 0.005	0.156	0.157
MMGO	MAE	0.244 ± 0.006	0.243	0.243
MPNC	MAE	0.113 ± 0.018	0.109	0.110
MSDA	MAE	0.159 ± 0.008	0.147	0.150
MUCR	MAE	0.159 ± 0.034	0.156	0.173
SOLUN	MAE	0.137 ± 0.015	0.135	0.130
WDD	MAE	0.183 ± 0.018	0.157	0.175
AIO	RMSE	0.300 ± 0.016	0.298	0.308
CSLB	RMSE	0.211 ± 0.005	0.210	0.213
HAGA	RMSE	0.257 ± 0.009	0.255	0.256
MMGO	RMSE	0.503 ± 0.019	0.502	0.501
MPNC	RMSE	0.182 ± 0.026	0.174	0.177
MSDA	RMSE	0.250 ± 0.018	0.237	0.233
MUCR	RMSE	0.232 ± 0.049	0.228	0.253
SOLUN	RMSE	0.236 ± 0.028	0.233	0.224
WDD	RMSE	0.308 ± 0.021	0.258	0.286

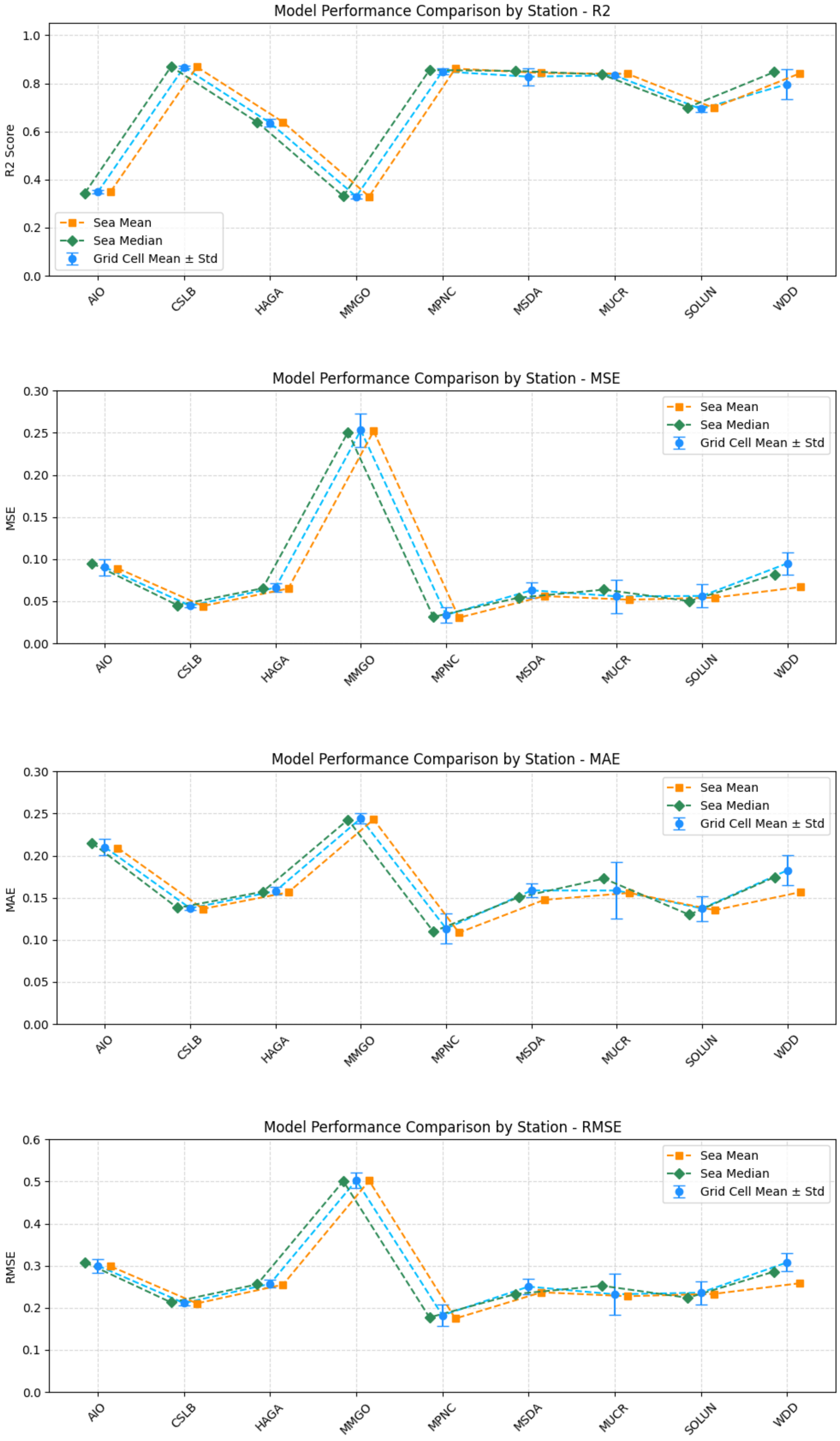


Figure 5.2 Performance by station and metric for baseline model.

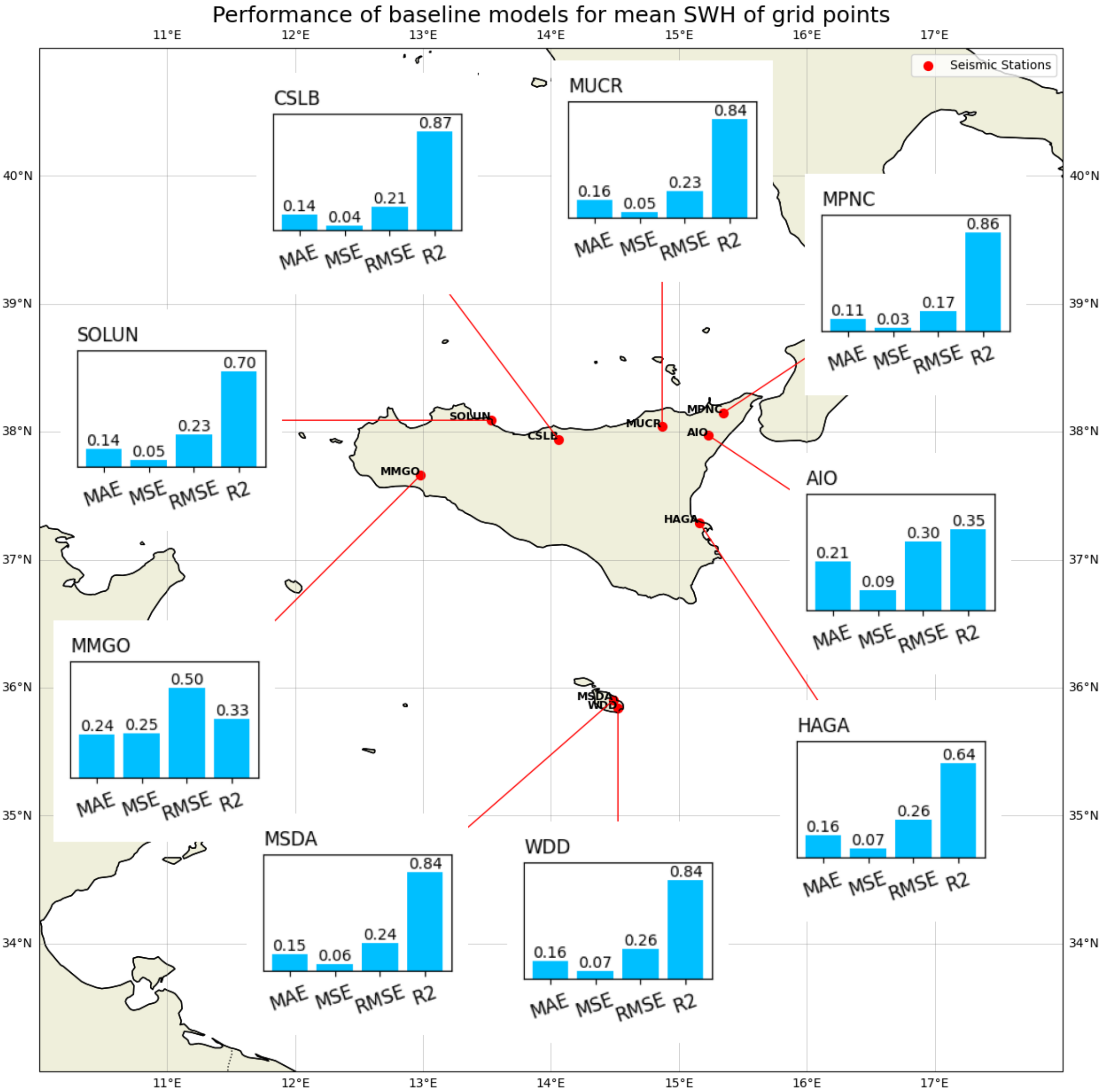


Figure 5.3 The results achieved for the baseline models trained on the mean SWH of the five nearest grid cells to each station.

The value of R² varied significantly across the stations, ranging from as low as 0.33 at station MMGO to as high as 0.869 at station CSLB. A spatial overview of these results is presented in Figure 5.3, which illustrates the geographical distribution of model performance. These findings highlighted substantial variability in the strength of the relationship between seismic hourly RMS and SWH across the study area. Notably, five stations – three along the northern coast of Sicily and the two located in Malta – achieved excellent performance, with R² values exceeding 0.84. In contrast, several other locations yielded weaker fits, with R² scores between 0.33 and 0.70. These variations suggested that the sensitivity of seismic RMS to sea state fluctuations is highly location-dependent and may be influenced by the local environment, geographical factors or the quality of the respective station's instrumentation.

The RMSE values were consistently higher than the corresponding MAE values across all stations. This discrepancy is expected, as RMSE penalizes larger errors more heavily due to the squaring term in its calculation (Equation 4.3). The presence of this gap suggests that while the models generally performed well, occasional large prediction errors had a disproportionate impact on the RMSE, indicating the existence of outliers or instances where the model struggled.

These baseline results supported the underlying hypothesis that seismic RMS can serve as a useful predictor for sea SWH. However, the variability in model performance across stations suggested that a uniform approach to model configuration was insufficient. The implication of linear interpolation suggested that model performance may benefit from an improved preprocessing pipeline with a different approach to tackling missing data.

Moreover, while the fixed hyperparameter settings provided a valuable point of reference, they likely failed to capture the full complexity of the underlying relationships at each site. This highlighted the need for station-specific model tuning and motivated the proposed adaptive methodology that can incorporate local conditions. Factors such as proximity to open sea, coastal topography, and anthropogenic noise may all influence the strength of the relationship, reinforcing the conclusion that a one-size-fits-all model configuration is unlikely to generalise well across geographically diverse locations.

5.2 Model Performance

Recall that, as detailed in Chapter 4, several key modifications were introduced to the methodology described in the reproduced baseline model, with the aim of aligning more closely with the objectives of this study:

- Noise filtering, skewness correction, and the use of the earthquake catalogue were deliberately omitted in accordance with Objective 3, which sought to minimise preprocessing and evaluate the extent to which models could learn from raw or near-raw data.

- For each station, the longest continuous segment of data with minimal interpolation was identified and used for training and evaluation, in line with Objectives 3 and 5, which emphasised data integrity and real-world applicability.

- Station-specific models were developed using only locally relevant data, thereby reducing the computational burden and enabling deployment on consumer-grade hardware, in support of Objective 2.

- A hyperparameter tuning exercise was performed for each station to determine the optimal hyperparameters, addressing Objective 4 and improving model performance through targeted tuning.

Collectively, these adjustments to the methodology contributed to a more robust exploration of the relationship between seismic RMS and SWH, fulfilling Objective 1, and targeting the primary research question.

The following analysis presents a detailed evaluation of the proposed modelling pipeline. First, the outcomes of the hyperparameter tuning exercise are explored to identify the best-performing models and the corresponding optimal parameter configurations for each station. This is followed by a cross-validation study, aimed at assessing the stability and generalisability of the models across different subsets of the data. The conducted analyses provide a detailed overview of the models' performance, their reliability, and their limitations within the constraints of this study.

5.2.1 Hyperparameter Tuning

The principal advantage of shifting from a single, unified model to individual models per station was the increased relevance and specificity of both the input data and the resulting predictions. This station-specific approach enabled independent hyperparameter tuning, allowing each model to be optimised according to the characteristics of its local dataset. As a result, model accuracy was improved without compromising computational feasibility. Optimal configurations were identified through a grid search over five hyperparameters. The results presented below include a comparative overview of the best-performing hyperparameters across stations, along with a discussion of potential factors contributing to the observed variation.

From Figure 5.4, it is evident that the greatest fluctuations in performance occurred due to changes in the maximum tree depth (`RF_max_depth`) and the number of estimators used in the RF regressor (`RF_n_estimators`). These two hyperparameters had the most significant impact on the model's R^2 scores across the stations considered. This suggested that they have more importance when considering model complexity and learning capacity.

In contrast, modifying the number of features considered when looking for the best split (`max_features`), the minimum number of samples required to split an internal node (`min_samples_split`), and the minimum number of samples required to be at a leaf node (`min_samples_leaf`) led to comparatively smaller changes in performance. This indicated that the model was relatively stable and immune to variations in these hyperparameters.

Figure 5.5 offers similar insights, highlighting that when evaluating the model from both perspectives – its ability to generalise, and minimise error – the most

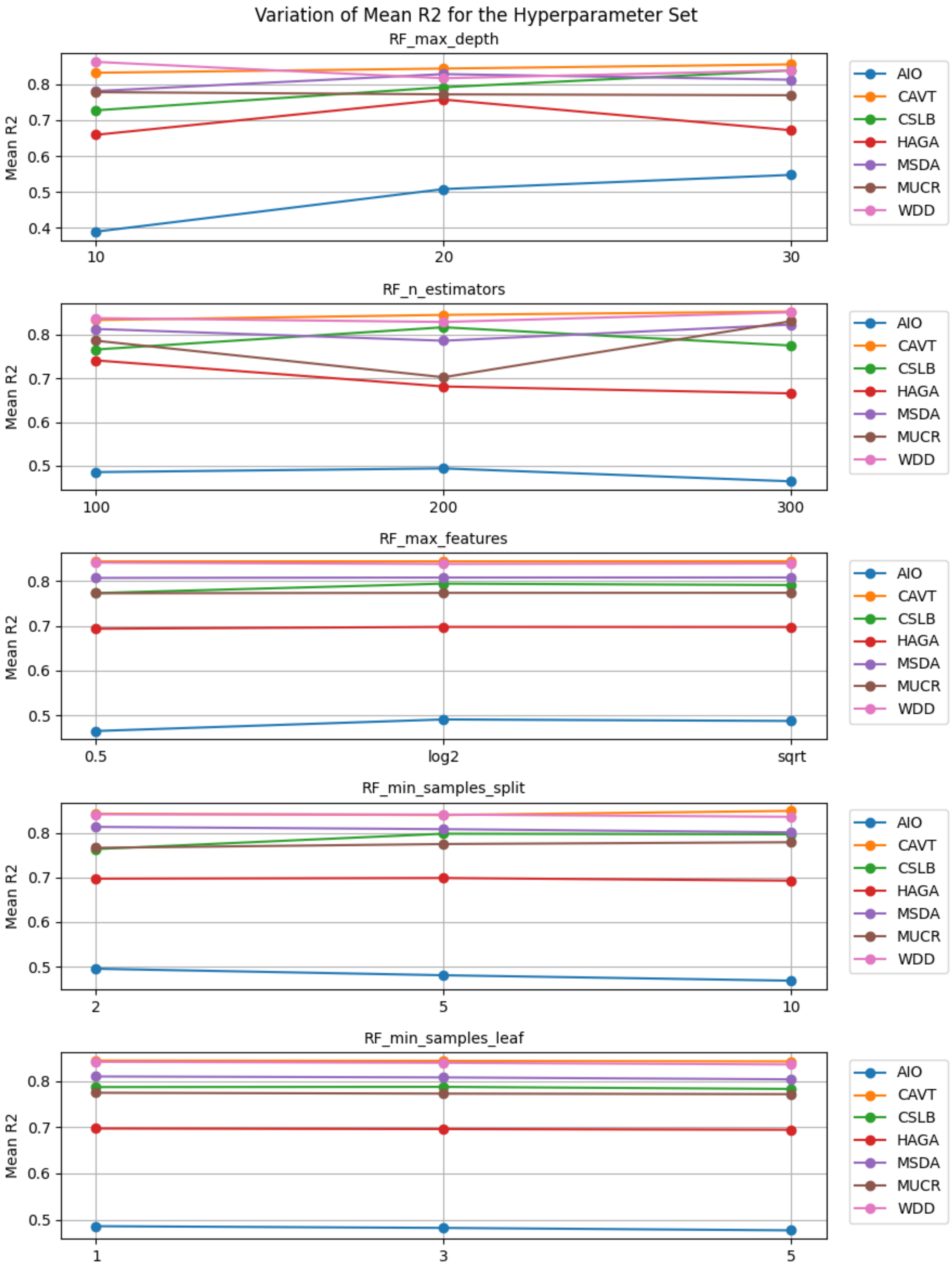


Figure 5.4 The variation of the mean R^2 across all stations for the different hyperparameters.

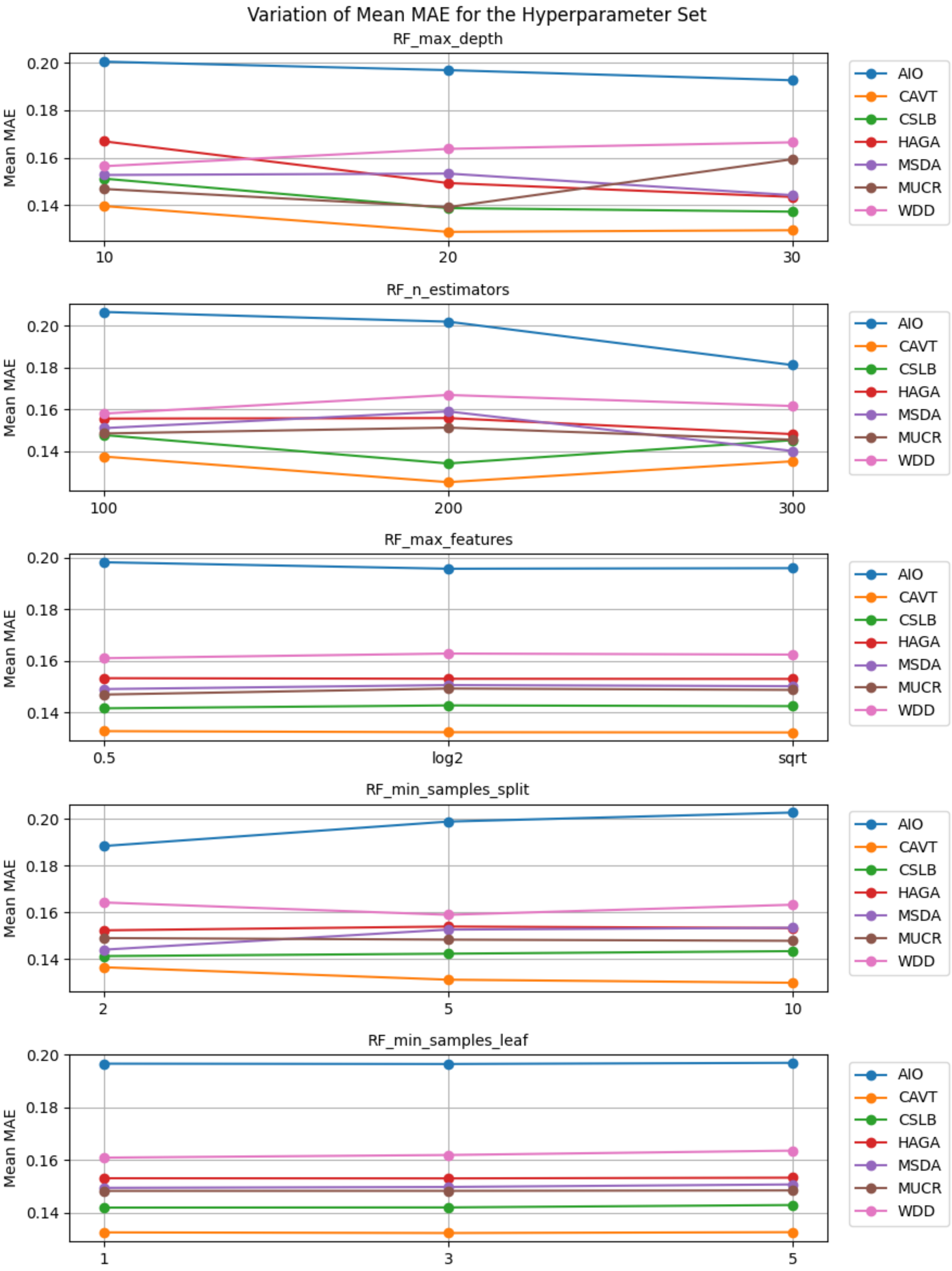


Figure 5.5 The variation of the mean MAE across all stations for the different hyperparameters.

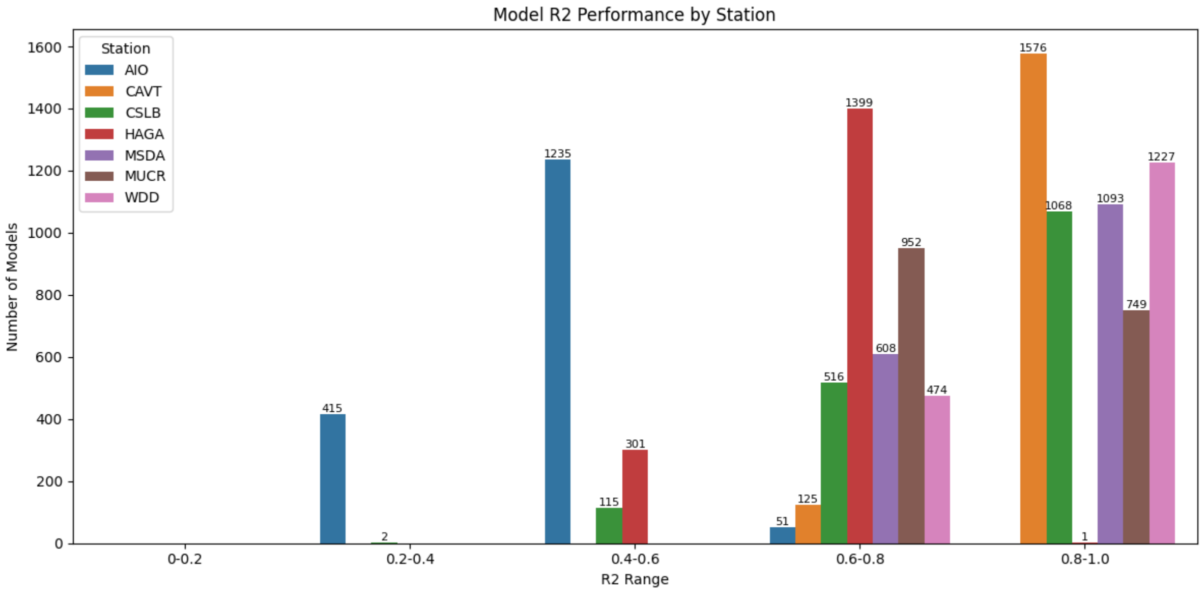


Figure 5.6 Histogram showing the distribution of R² for the various stations and hyperparameters.

influential hyperparameters were the maximum tree depth and the number of estimators. These two parameters continued to demonstrate the greatest impact on the model's performance, emphasising their importance in balancing model complexity and predictive power.

The grid search exercise resulted in 243 models being trained for each station and each target variable: 243 hyperparameter combinations ×7 stations ×7 target variables = 11,907 RF regressors. The distribution of the performance based on R^2 for of these models on the test set is shown in Figure 5.6. The histogram clearly depicts the consistent poor performance of station AIO, whereby just 3.0% of all the trained models result in R^2 exceeding 0.6, and the greatest R^2 is just 0.611. On the other hand, a relatively consistently strong performance at stations CAVT, WDD, MSDA and CSLB is observed, with 92.7%, 72.1%, 64.3% and 62.7% of the models trained for the respective stations achieving R^2 greater than 0.8. On the other hand, stations HAGA and MUCR achieve a consistent mid-range performance. The majority of the regressors trained at those locations achieved an R^2 between 0.6 and 0.8, with 82.2% and 56.0% of the models trained at stations HAGA and MUCR obtaining an R^2 within this category.

Interestingly, Figures 5.4, 5.5 and 5.6 all suggest that station AIO consistently performs worse than all other stations across all hyperparameter configurations. This implies that station AIO may have presented unique challenges or peculiarities that the model struggled to address. The station may have instrument-related issues such as incorrect calibration or synchronisation, which was not revealed through exploratory data analysis. It is also possible that the ideal hyperparameters for station AIO lie beyond those tested in this study, warranting a more focused investigation into hyperparameter values that were not included in the initial search.

Tables 5.2–5.8 show the optimal set of hyperparameters achieved for each station and the various evaluation metrics. These provided an insight to the variation, if any, caused by adjustments made to the hyperparameters, along with the hyperparameters that cause these changes.

Station AIO

At station AIO, the highest R^2 achieved was 0.607. As shown in Table 5.2, slight variations in hyperparameters yielded marginal improvements in error metrics. Specifically, the MAE was reduced by 0.0093 m, the MSE by 0.0062 m^2, and the RMSE by 0.0118 m, relative to the model with the lowest R^2. However, these small gains in predictive accuracy came at the expense of model explanatory power, with R^2 decreasing to approximately 0.577. Given that R^2 values below 0.6 are typically considered to indicate a weaker fit between predicted and observed values, such a trade-off is considered unjustified. Consequently, the optimal set of hyperparameters for station AIO was taken to be the configuration that maximised R^2, thereby preserving the strongest overall model fit.

Station CAVT

The selection of optimal hyperparameters for station CAVT was comparatively straightforward. A single configuration yielded the highest R^2, as well as the lowest MSE and RMSE, as shown in Table 5.3, providing a clear basis for selection. Although a marginally lower MAE was achieved with an alternative set – improving MAE by just 0.00025 m – this came at the cost of slight reductions in the other performance metrics. Given the negligible benefit in MAE and the stronger overall performance, the hyperparameters resulting in the highest R^2, lowest MSE, and lowest RMSE were selected as optimal for station CAVT.

Station CSLB

Station CSLB demonstrated a pattern similar to that observed at station AIO, with some variation in performance across different hyperparameter sets. However, the overall performance at CSLB was substantially stronger than station AIO, with a maximum R^2 of 0.881, presented in Table 5.4. While alternative configurations offered slight reductions in error – most notably an improvement of up to 0.061 m in RMSE – these gains came at the cost of a significant decrease in R^2, by as much as 0.134. Given the importance of maintaining model generalisability and explanatory power, this trade-off was deemed unjustified. As such, the hyperparameters yielding the highest R^2 were selected as optimal for station CSLB.

Station HAGA

Station HAGA produced the lowest values across all three error metrics: MAE, MSE, and RMSE, for a single set of hyperparameters, as shown in Table 5.5. These improvements amounted to 0.030 m, 0.035 m^2, and 0.084 m respectively, over the configuration with the best R^2. However, these relatively minor reductions in error were accompanied by a notable decrease in model fit, with R^2 dropping by 0.175. Given that the maximum gain in RMSE represented only an 8.4 cm improvement in SWH estimation, this trade-off was considered disproportionate. Accordingly, the hyperparameters associated with the highest R^2 were selected as optimal for station HAGA.

Station MSDA

Station MSDA exhibited consistent performance across different hyperparameter configurations, despite only two of the four top-performing models in Table 5.6 sharing identical settings. This outcome supports the earlier analysis presented in Figures 5.4, 5.5, which identified the number of estimators and the maximum tree depth as the most influential hyperparameters. For station MSDA, these parameters remained largely consistent across the best-performing models, resulting in comparable values for both goodness of fit, and error metrics. Given the marginal differences between the top configurations, and in line with the selection criteria applied to other stations, the hyperparameters yielding the highest R^2 were chosen as optimal.

Station MUCR

At station MUCR, the hyperparameter configuration that produced the lowest MSE and RMSE, as shown in Table 5.7, led to a reduction in R^2 of more than 0.124 compared to the best-performing model in terms of explanatory power. Given the marginal improvements in error and the significant drop in R^2, this configuration was excluded from further consideration. An alternative set of hyperparameters resulted in the lowest MAE, accompanied by a slight reduction in R^2 relative to the optimal model. However, when considering the trade-off between predictive performance and the risk of overfitting, the configuration yielding the highest R^2 was selected as the optimal choice for station MUCR.

Table 5.2 Optimal hyperparameters and results for station AIO.

	Station AIO			
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	30	30	30	30
`RF_n_estimators`	200	300	300	300
`RF_max_features`	$\log_2$	$\log_2$	$\log_2$	$\log_2$
`RF_min_samples_split`	2	5	5	5
`RF_min_samples_leaf`	1	3	5	5
MAE	0.18243	**0.17308**	0.17309	0.17309
MSE	0.07107	0.06497	**0.06490**	**0.06490**
RMSE	0.26659	0.25487	**0.25476**	**0.25476**
R^2	**0.60686**	0.57728	0.57767	0.57767

Station WDD

Lastly, station WDD presented the most straightforward case among all evaluated locations. A single set of hyperparameters, shown in Table 5.8, satisfied both the error and goodness-of-fit criteria, making the selection unambiguous. This configuration produced the lowest MSE and RMSE, as well as the highest R^2 across all stations.

Summary of Hyperparameter Tuning Results

In summary, Table 5.9 presents the selected optimal hyperparameters for each station, alongside the evaluation metrics. The variation in these configurations across locations supports the underlying hypothesis that station-specific conditions necessitate tailored modelling approaches. This outcome aligns with and fulfils Objective 4 of this study, demonstrating the value of performing targeted hyperparameter tuning to improve model performance at each site. Additionally, Table 5.10 shows the mean and standard deviation of the performance metrics at each station. The high standard deviation justifies the need for a station by station approach to both training and evaluation.

Table 5.3 Optimal hyperparameters and results for station CAVT.

Station CAVT				
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	30	30	30	30
`RF_n_estimators`	200	200	200	200
`RF_max_features`	sqrt	0.5	sqrt	sqrt
`RF_min_samples_split`	2	10	2	2
`RF_min_samples_leaf`	1	3	1	1
MAE	0.10066	**0.10042**	0.10066	0.10066
MSE	**0.02291**	0.02317	**0.02291**	**0.02291**
RMSE	**0.15137**	0.15223	**0.15137**	**0.15137**
R^2	**0.89238**	0.89117	**0.89238**	**0.89238**

Table 5.4 Optimal hyperparameters and results for station CSLB.

Station CSLB				
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	10	30	20	20
`RF_n_estimators`	200	100	200	200
`RF_max_features`	$\log_2$	0.5	$\log_2$	$\log_2$
`RF_min_samples_split`	5	2	10	10
`RF_min_samples_leaf`	3	5	1	1
MAE	0.14298	**0.11822**	0.12327	0.12327
MSE	0.05519	0.03268	**0.03025**	**0.03025**
RMSE	0.23492	0.18077	**0.17394**	**0.17934**
R^2	**0.88108**	0.76862	0.74718	0.74718

Table 5.5 Optimal hyperparameters and results for station HAGA.

Station HAGA				
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	20	30	30	30
`RF_n_estimators`	100	200	200	200
`RF_max_features`	sqrt	$\log_2$	$\log_2$	$\log_2$
`RF_min_samples_split`	2	10	10	10
`RF_min_samples_leaf`	1	1	1	1
MAE	0.15251	**0.12278**	**0.12278**	**0.12278**
MSE	0.06361	**0.02817**	**0.02817**	**0.02817**
RMSE	0.25221	**0.16785**	**0.16785**	**0.16785**
R^2	**0.78357**	0.60880	0.60880	0.60880

Table 5.6 Optimal hyperparameters and results for station MSDA.

Station MSDA				
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	30	20	20	20
`RF_n_estimators`	100	100	100	100
`RF_max_features`	$\log_2$	sqrt	0.5	0.5
`RF_min_samples_split`	10	2	2	2
`RF_min_samples_leaf`	1	3	1	1
MAE	0.12207	**0.11417**	0.11430	0.11430
MSE	0.03282	0.02528	**0.02514**	**0.02514**
RMSE	0.18116	0.15901	**0.15854**	**0.15854**
R^2	**0.86198**	0.85512	0.85597	0.85597

Table 5.7 Optimal hyperparameters and results for station MUCR.

Station MUCR				
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	30	20	20	20
`RF_n_estimators`	100	200	200	200
`RF_max_features`	0.5	0.5	0.5	0.5
`RF_min_samples_split`	10	2	5	5
`RF_min_samples_leaf`	1	5	1	1
MAE	0.14089	**0.12563**	0.12637	0.12637
MSE	0.04067	0.03030	**0.03007**	**0.03007**
RMSE	0.20166	0.17406	**0.17340**	**0.17340**
R^2	**0.86200**	0.85512	0.73789	0.73789

Table 5.8 Optimal hyperparameters and results for station WDD.

Station WDD				
	Highest R^2	Lowest MAE	Lowest MSE	Lowest RMSE
`RF_max_depth`	10	10	10	10
`RF_n_estimators`	100	100	100	100
`RF_max_features`	$\log_2$	$\log_2$	$\log_2$	$\log_2$
`RF_min_samples_split`	5	5	5	5
`RF_min_samples_leaf`	3	3	3	3
MAE	**0.10175**	**0.10175**	**0.10175**	**0.10175**
MSE	**0.02073**	**0.02073**	**0.02073**	**0.02073**
RMSE	**0.14398**	**0.14398**	**0.14398**	**0.14398**
R^2	**0.92060**	**0.92060**	**0.92060**	**0.92060**

Table 5.9 Optimal hyperparameters selected for each station and corresponding evaluation metrics.

	AIO	CAVT	CSLB	HAGA	MSDA	MUCR	WDD
`RF_max_depth`	30	30	10	20	30	30	10
`RF_n_estimators`	200	200	200	100	100	100	100
`RF_max_features`	$\log_2$	sqrt	$\log_2$	sqrt	$\log_2$	0.5	$\log_2$
`RF_min_samples_split`	2	2	5	2	10	10	5
`RF_min_samples_leaf`	1	1	3	1	1	1	3
MAE	0.18243	0.10066	0.14298	0.15251	0.12207	0.14089	0.10175
MSE	0.07107	0.02291	0.05519	0.06361	0.03282	0.04067	0.02073
RMSE	0.26659	0.15137	0.23492	0.25221	0.18116	0.20166	0.14398
R^2	0.60686	0.89238	0.88108	0.78357	0.86198	0.86200	0.92060

Table 5.10 Summary of performance across all stations for the optimal hyperparameters.

	MAE	MSE	RMSE	R^2
Mean	0.13476	0.04386	0.20456	0.82978
Standard Deviation	0.02701	0.01838	0.04487	0.09907
Minimum	0.10066	0.02073	0.14398	0.60686
Maximum	0.18243	0.07107	0.26659	0.92060

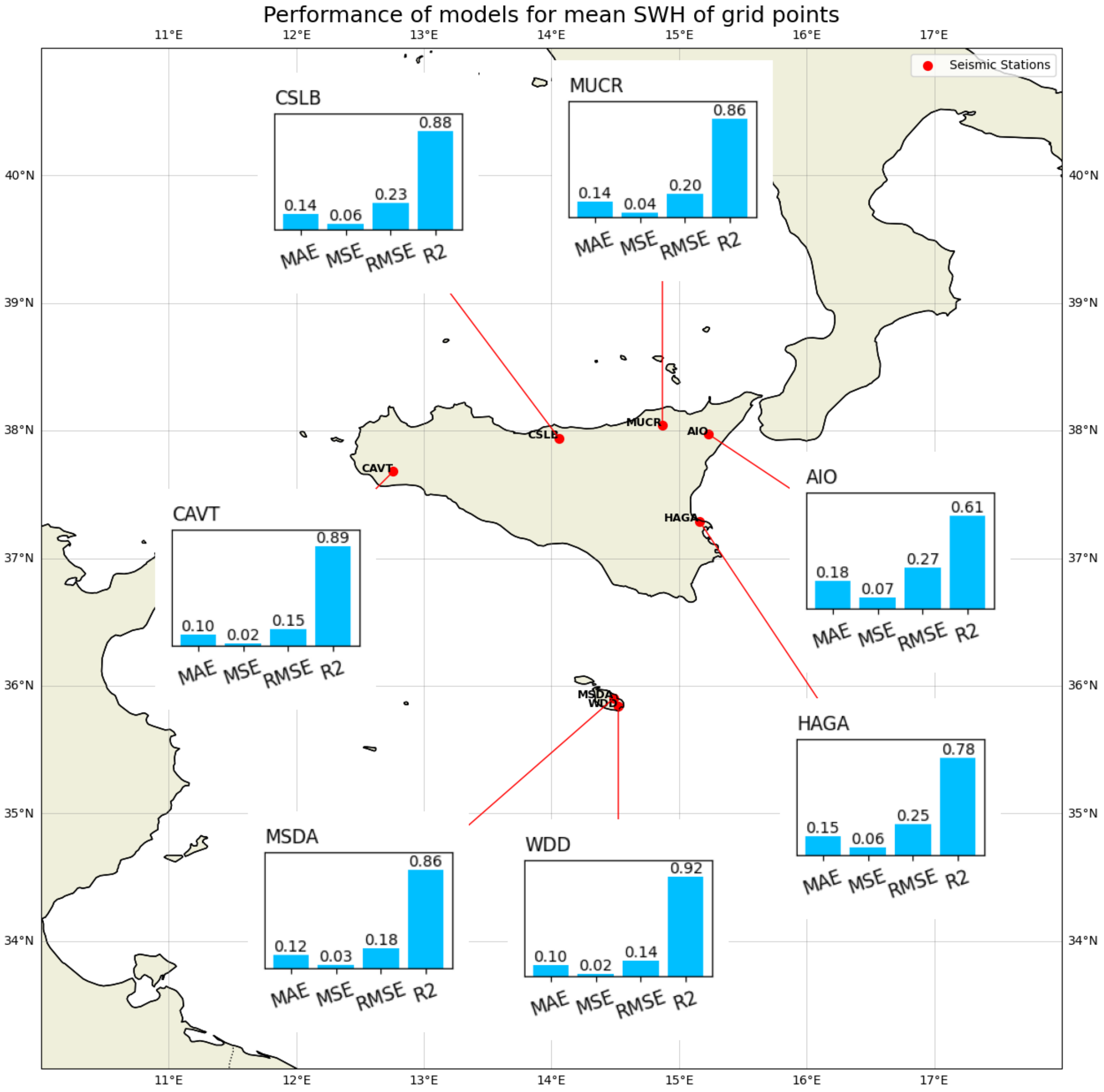


Figure 5.7 The results achieved for the models trained with the optimal hyperparameters per station on the mean SWH of the five nearest grid cells to each station.

Figure 5.7 shows the geographical distribution of model performance. A ranging performance was observed, once again denoting a variability in the strength of the relationship between seismic hourly RMS and SWH over the region of interest. The three stations along the northern and north-western coast of Sicily, and the two stations located in Malta achieved the best overall performance, with an R^2 exceeding 0.86. The other two stations achieved R^2 values of 0.61 and 0.78. The reasons for such variations will be discussed in the subsequent analysis of the K-fold cross validation results, followed by a comparison to the baseline models.

5.2.2 K-Fold Cross Validation

Out of the 11,907 RF regressors trained during the hyperparameter tuning process, one optimal model was selected per station, resulting in a total of seven final models. To evaluate the robustness and generalisability of these models, a five-fold cross-validation procedure was implemented, using the optimal set of hyperparameters determined through the earlier experiment. As illustrated in Figure 4.10 and Figure 4.11, this approach ensured that each data point was used once for testing and four times for training across the folds. The data underwent random stratification and importantly, no data were shared between the training and test sets within any single fold, preserving the integrity of the validation process and providing a reliable assessment of model performance.

The following analysis represents the performance of the K-fold cross validation, for each station.

Station AIO

Earlier analysis revealed that station AIO had the poorest performance across all stations, at its optimal set of hyperparameters, obtaining an R^2 of 0.60686; 0.17671 less than the next worst performing station, HAGA. From the K-fold cross-validation exercise, this discrepancy was further contextualised. The models trained for AIO again achieved a low mean performance and a relatively significant standard deviation, as shown in Table 5.11. The higher standard deviation indicated some instability in generalisation. The repeated concern with this station's data may have caused overfitting on certain folds and underfitting on others.

Figure 5.8a illustrates the predicted versus actual SWH across the full test set for the best-performing fold at station AIO. Additionally, Figure 5.8b presents two examples comparing the predicted and actual SWH at the locations of the five grid cells nearest to station AIO. The test period spans a variety of sea states and seasonal conditions, allowing for analysis of the model's temporal generalisation capacity. Notably, the plot reveals clear prediction errors such as consistent underestimation of SWH from 7 June to 22 June 2019, and overestimation from 22 July to 5 August 2019. These suggest that the model was able to capture general trends, but struggled to adapt to sudden shifts in SWH.

Station CAVT

Station CAVT emerged as one of the best-performing locations during the hyperparameter tuning, achieving the second-highest R^2 overall. This strong initial performance was further substantiated by the K-fold cross-validation, which revealed consistently high predictive accuracy with minimal variability across folds. The R^2 values ranged from 0.823 to 0.895, reflecting the model's robustness and reliability when exposed to different data. These results are shown in Table 5.12.

Figure 5.9a illustrates the model's ability to closely follow the mean SWH, capturing the overall shape and fluctuations with notable precision. The majority of deviations between predicted and actual values appeared during periods of extreme weather, likely attributable to the scarcity of such events in the training dataset. This underrepresentation may have hindered the model's capacity to generalise well to rare but high-impact wave conditions.

As shown in Figure 5.9b, the predictions at individual grid cells are highly accurate, with errors as low as 1 cm. In practical terms, such discrepancies are negligible for many applications, indicating that the model's predictions can be both precise and dependable. Collectively, these results indicate that CAVT is a reliable station for wave height estimation using seismic signals.

Station CSLB

The hyperparameter search conducted earlier revealed that station CSLB achieved a similar generalisability to station CAVT, at an R^2 of 0.88108. The MAE, MSE and RMSE, on the other hand, were slightly higher at station CSLB.

The K-fold cross validation exercise for station CSLB revealed a low mean R^2 of 0.64204 and a comparatively high standard deviation of 0.29353 in performance across the five folds, shown in Table 5.13. This is owing to one of the folds performing very poorly, achieving an R^2 of -0.05. This, alongside the lower correlation between the mean SWH and seismic RMS, shown in Figure 4.7, suggested that influences external to the parameters of this research may play a part in the relationship between seismic RMS and the SWH. Nonetheless, at the best data split, an R^2 of 0.88938 was achieved, consistent with that achieved during the search for the best hyperparameters.

Figure 5.10a demonstrates that the model at station CSLB was capable of tracking the overall trend of the mean SWH, albeit with more variability than that observed at other high-performing stations. The time series revealed a combination of accurate predictions with intermittent overestimations and underestimations. Notably, the pronounced peak between 14 August and 24 August 2020 was captured reasonably well, although a few sharp anomalies were underpredicted relative to the observed values. One timestamp from this period is captured by Figure 5.10b, showing a reasonable performance at most grid cells during a period of the highest SWH within the test set.

When compared with station CAVT (Figure 5.9a), the CSLB predictions exhibit a slightly noisier profile, with less precise alignment to the actual signal, particularly during rapid fluctuations. This suggests that while both models have established a reasonably strong relationship between SWH and seismic RMS, the CSLB model is more prone to error, potentially due to the factors emanating beyond the parameters of this research. Nonetheless, the overall performance remains strong, with temporal patterns generally well-replicated and only minor divergences from actual values.

Station HAGA

Building upon the hyperparameter tuning stage, station HAGA was identified as a mid-range performer in terms of R², while achieving MAE, MSE, and RMSE values comparable to those observed at station CSLB, the second-best performer overall. This suggests that while the model may not have captured the full variance in the data, it still produced relatively low prediction errors.

However, the cross-validation results revealed some instability, with one of the folds producing a notably weak performance. The minimum R² across the five folds was 0.45504, potentially due to factors such as instrument noise or calibration issues. The mean R² of 0.66332 ± 0.12320 further reflects this variability, indicating moderate predictive reliability overall. The best-performing fold reached an R² of 0.75471, accompanied by mid-range error metrics, compared to other stations, suggesting a reasonable trade-off between accuracy and generalisation. These results are shown in Table 5.14.

The test set associated with the best fold featured less extreme sea states compared to other stations, with a maximum SWH of approximately 2.4 m, as shown in Figure 5.11a. The model generally followed the temporal pattern of observed SWH, with a significant overestimation between 4 March and 1 April 2019. Outside of this period, the predictions remained relatively consistent, albeit with persistent low-amplitude error throughout the signal. It is worth noting that the perceived magnitude of error in the plot is visually exaggerated due to the reduced vertical axis scale which was roughly half that of station CSLB in Figure 5.10a.

Spatially, the model performed well across the five nearest grid cells to station HAGA, as illustrated in Figure 5.11b. In several instances, prediction errors were below 1 cm, demonstrating the model's capacity even at grid-cell level.

Station MSDA

Station MSDA exhibited the most stable performance across all five folds, with a mean R² of 0.83860 and a remarkably low standard deviation of just 0.02846. The maximum R² achieved was 0.86563, closely aligned with results from the hyperparameter tuning phase, while the minimum was 0.79543 – still well above the lowest-performing models at other stations. Interestingly, the highest R² value was associated with the highest observed error metrics across folds (Table 5.15), suggesting a potential trade-off between explained variance and absolute prediction accuracy.

The test set corresponding to the best-performing fold provided a balanced representation of all four seasons, as shown in Figure 5.12a. The model demonstrated strong performance in tracking periods of elevated SWH, particularly between 25 December 2020 and 27 January 2021. Nonetheless, consistent with patterns observed at other stations, there were instances of both over- and underestimation, with underpredictions more prominent during episodes of extreme wave activity.

At grid-cell level, model performance remained robust, as evidenced in Figure 5.12b. Several of the selected grid cell predictions showed errors lower than 1 cm, highlighting the model's effectiveness in capturing fine-scale variations in wave height near station MSDA.

Station MUCR

Station MUCR demonstrated a reasonably stable performance across the five-fold cross-validation, achieving a mean R^2 of 0.79882 with a standard deviation of 0.04678. The peak R^2 value of 0.86660 closely matched that obtained during the hyperparameter tuning phase, indicating consistency across different evaluation methods. The error metrics MAE, MSE, and RMSE were relatively low and aligned with those of other stable-performing stations, notably CAVT and MSDA.

The test set for the best-performing fold revealed strong model responsiveness to rapid changes in SWH, particularly evident between 2 April and 22 April 2022. The model also performed well under more extreme sea conditions, such as those observed between 23 August and 2 October 2020. As with other stations, the upper extremes of SWH were sometimes underestimated, though such underestimations were less pronounced outside of high-severity events.

Grid-cell performance, illustrated in Figure 5.13b, was generally strong. Across two selected timestamps, errors at the grid cell level ranged from negligible to a maximum of 8 cm, demonstrating acceptable variance and reinforcing the station's overall robustness in wave height estimation.

Station WDD

Prior to the K-fold cross-validation, station WDD had already emerged as the top-performing model during the hyperparameter tuning stage, setting high expectations for subsequent validation.

As shown in Table 5.17, WDD maintained a high mean R^2 and low standard deviation across the five folds, reinforcing its reputation for consistent and accurate performance. It outperformed its Maltese counterpart MSDA and other strong performers from Sicily. The highest R^2 observed was 0.91263, closely matching the result from the earlier tuning phase and further validating the model's reliability.

The time series in Figure 5.14a demonstrates the model's ability to closely track actual SWH, with minimal divergence between predicted and observed values. While some instances of over- and underestimation remained, the associated errors appeared considerably smaller compared to other stations. The test set encompassed a diverse range of sea states across all four seasons, providing a robust basis for evaluating the model's generalisation capabilities.

Figure 5.14b further supports these findings. Even under more extreme sea conditions, the model produced errors no greater than 9 cm, while maintaining very small errors under typical conditions. Overall, WDD stood out as the most reliable and high-performing station in terms of wave height prediction based on seismic data.

Summary of K-Fold Cross Validation Results

The K-fold cross-validation results suggested a generally strong and consistent predictive relationship between seismic RMS values and corresponding SWH measurements across all stations. In most cases, performance during cross-validation closely mirrored the results obtained during the hyperparameter tuning stage, validating the model stability and effectiveness.

The performance consistency, across all stations and folds per metric, is summarised in Figure 5.15. Stations WDD and MSDA stood out from a stability and performance perspective, closely followed by station CAVT and MUCR, indicating strong generalisability across different subsets of the data. Stations AIO, CSLB and HAGA demonstrated wider performance ranges across folds, suggesting sensitivity to data partitioning or underlying quality issues in the training data.

A recurring pattern observed across multiple stations was the underestimation of peak SWH during periods of more extreme sea conditions. These discrepancies often extended beyond the peaks themselves, reflecting the models' limited exposure to such conditions during training, and potentially indicating a data imbalance or a nonlinear relationship not fully captured by the current model design.

Despite these limitations, all models displayed a good ability to generalise across both temporal and spatial domains, as evidenced by low errors at individual grid cell predictions and stable performance across various seasonal conditions. These results were obtained through data exposed to minimal data processing, in alignment with Objectives 2, 3 and 5 of this research. This further supports the potential of using seismic data as a reliable resource for estimating ocean wave activity, indicated by Objective 1 of this study, while also highlighting areas for refinement, particularly in handling edge cases associated with extreme environmental variability.

Table 5.11 K-fold cross validation results for station AIO.

	MAE	MSE	RMSE	R^2
Fold 1	0.169307	0.062620	0.250240	0.577653
Fold 2	0.205318	0.081181	0.284923	0.356452
Fold 3	0.251851	0.158015	0.397511	0.512832
Fold 4	0.164635	0.059645	0.244223	0.296591
Fold 5	0.198994	0.084443	0.290591	0.271027
Mean	0.198021	0.089181	0.293498	0.402911
Standard Deviation	0.034968	0.040008	0.061644	0.135529
Minimum	0.164635	0.059645	0.244223	0.271027
Maximum	0.251851	0.158015	0.397511	0.577653
Best R^2	**0.169307**	**0.062620**	**0.250240**	**0.577653**

Table 5.12 K-fold cross validation results for station CAVT.

	MAE	MSE	RMSE	R^2
Fold 1	0.119997	0.034024	0.184456	0.824425
Fold 2	0.152407	0.055495	0.235574	0.836915
Fold 3	0.133355	0.042184	0.205388	0.895242
Fold 4	0.116028	0.031395	0.177187	0.871979
Fold 5	0.169018	0.066633	0.258133	0.823155
Mean	0.138161	0.045946	0.212148	0.850343
Standard Deviation	0.022350	0.014892	0.034272	0.031919
Minimum	0.116028	0.031395	0.177187	0.823155
Maximum	0.169018	0.066633	0.258133	0.895242
Best R^2	**0.133355**	**0.042184**	**0.205388**	**0.895242**

Table 5.13 K-fold cross validation results for station CSLB.

	MAE	MSE	RMSE	R^2
Fold 1	0.163621	0.109277	0.330571	-0.05001
Fold 2	0.150255	0.051887	0.227788	0.731844
Fold 3	0.141376	0.052180	0.228430	0.889382
Fold 4	0.112758	0.024956	0.157974	0.755026
Fold 5	0.15789	0.057534	0.239862	0.883940
Mean	0.145180	0.059167	0.236925	0.642036
Standard Deviation	0.019956	0.030765	0.061577	0.393528
Minimum	0.112758	0.024956	0.157974	-0.050010
Maximum	0.163621	0.109277	0.330571	0.889382
Best R^2	**0.141376**	**0.052180**	**0.228430**	**0.889382**

Table 5.14 K-fold cross validation results for station HAGA.

	MAE	MSE	RMSE	R^2
Fold 1	0.123114	0.030660	0.175099	0.660083
Fold 2	0.184792	0.121465	0.348518	0.693702
Fold 3	0.167826	0.058617	0.242110	0.754709
Fold 4	0.121683	0.027827	0.166813	0.455042
Fold 5	0.156743	0.048462	0.220140	0.753045
Mean	0.150831	0.057406	0.230536	0.663316
Standard Deviation	0.027816	0.037998	0.072965	0.123203
Minimum	0.121683	0.027827	0.166813	0.455042
Maximum	0.184792	0.121465	0.348518	0.754709
Best R^2	**0.167826**	**0.058617**	**0.242110**	**0.754709**

Table 5.15 K-fold cross validation results for station MSDA.

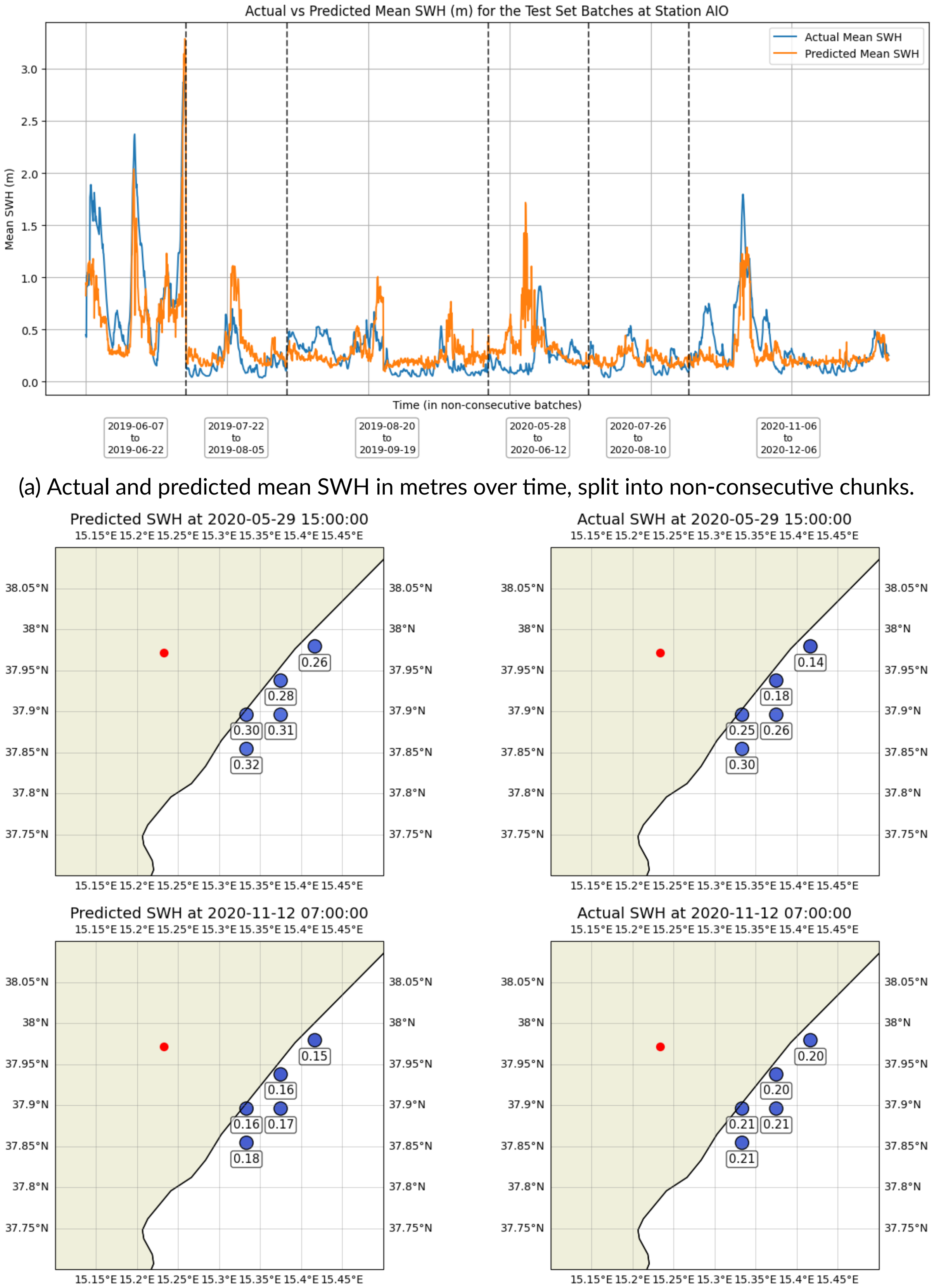
	MAE	MSE	RMSE	R^2
Fold 1	0.145279	0.048429	0.220067	0.826091
Fold 2	0.149912	0.055408	0.235390	0.865631
Fold 3	0.119357	0.025263	0.158943	0.846620
Fold 4	0.139735	0.040937	0.202329	0.859208
Fold 5	0.146098	0.043086	0.207572	0.795433
Mean	0.140076	0.042625	0.204860	0.838596
Standard Deviation	0.012140	0.011198	0.028659	0.028458
Minimum	0.119357	0.025263	0.158943	0.795433
Maximum	0.149912	0.055408	0.235390	0.865631
Best R^2	**0.149912**	**0.055408**	**0.235390**	**0.865631**

Table 5.16 K-fold cross validation results for station MUCR.

	MAE	MSE	RMSE	R^2
Fold 1	0.146368	0.046904	0.216573	0.754810
Fold 2	0.169425	0.060320	0.245601	0.866596
Fold 3	0.116771	0.031602	0.177770	0.821905
Fold 4	0.137248	0.037855	0.194565	0.792589
Fold 5	0.142982	0.041905	0.204708	0.758209
Mean	0.142559	0.043717	0.207843	0.798822
Standard Deviation	0.018906	0.010842	0.025458	0.046775
Minimum	0.116771	0.031602	0.177770	0.754810
Maximum	0.169425	0.060320	0.245601	0.866596
Best R^2	**0.169425**	**0.060320**	**0.245601**	**0.866596**

Table 5.17 K-fold cross validation results for station WDD.

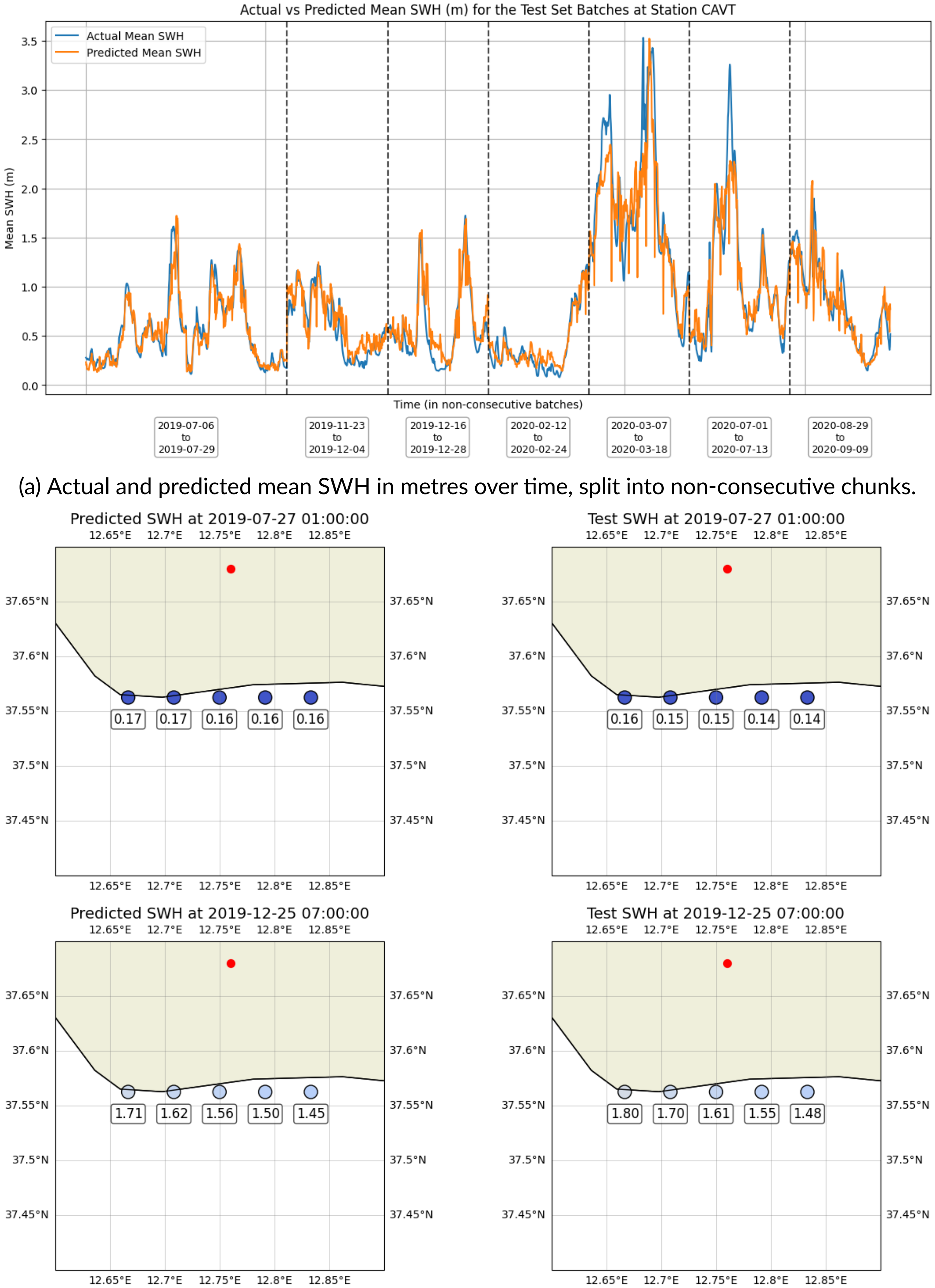
	MAE	MSE	RMSE	R^2
Fold 1	0.101956	0.023402	0.152978	0.885760
Fold 2	0.187367	0.124087	0.352260	0.823331
Fold 3	0.141248	0.041203	0.202985	0.877316
Fold 4	0.110847	0.030799	0.175497	0.879675
Fold 5	0.127221	0.036189	0.190235	0.912634
Mean	0.133728	0.051136	0.214791	0.875743
Standard Deviation	0.033575	0.041312	0.079065	0.032505
Minimum	0.101956	0.023402	0.152978	0.823331
Maximum	0.187367	0.124087	0.352260	0.912634
Best R^2	**0.127221**	**0.036189**	**0.190235**	**0.912634**



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

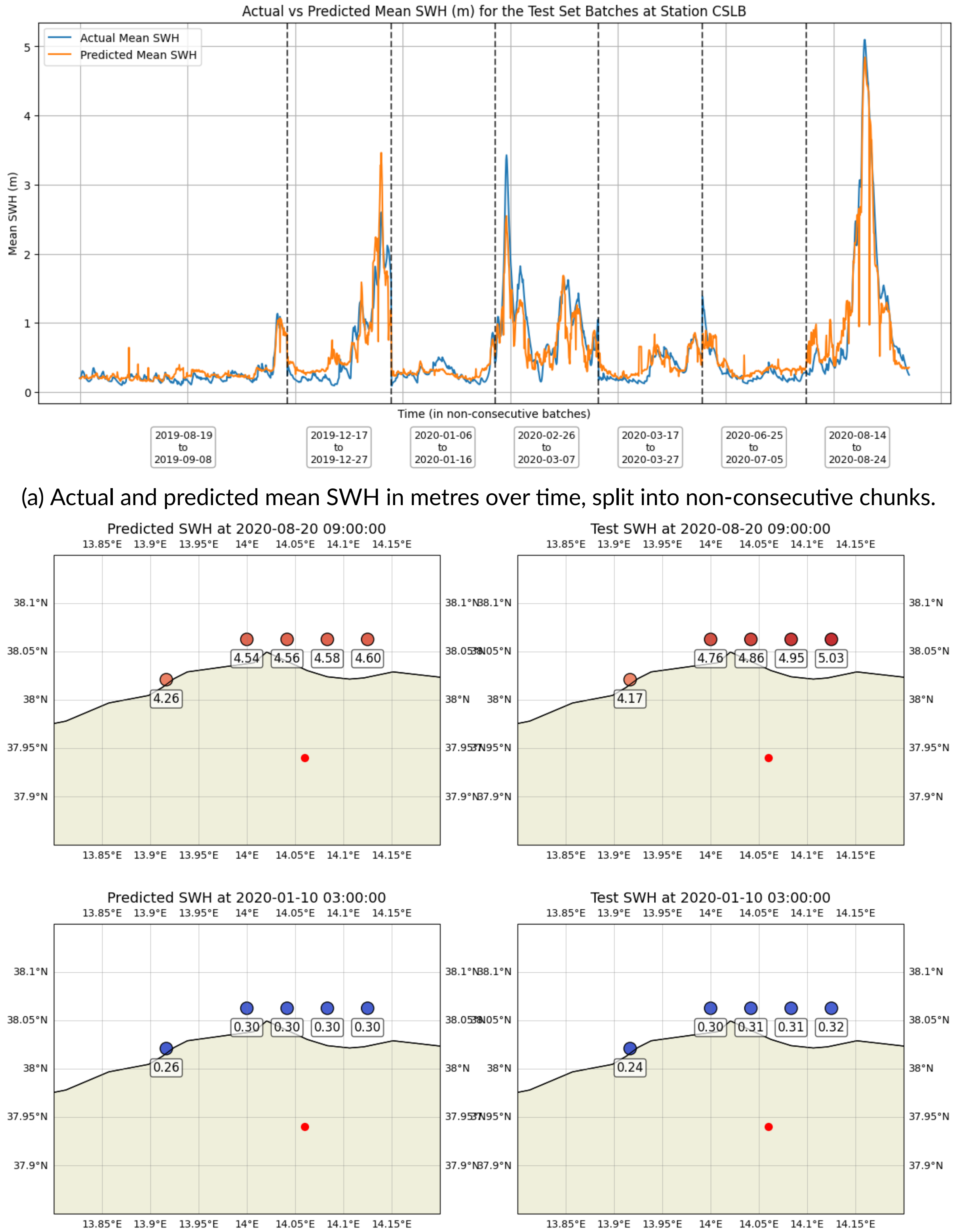
Figure 5.8 Data visualisation for best fold results at station AIO.



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

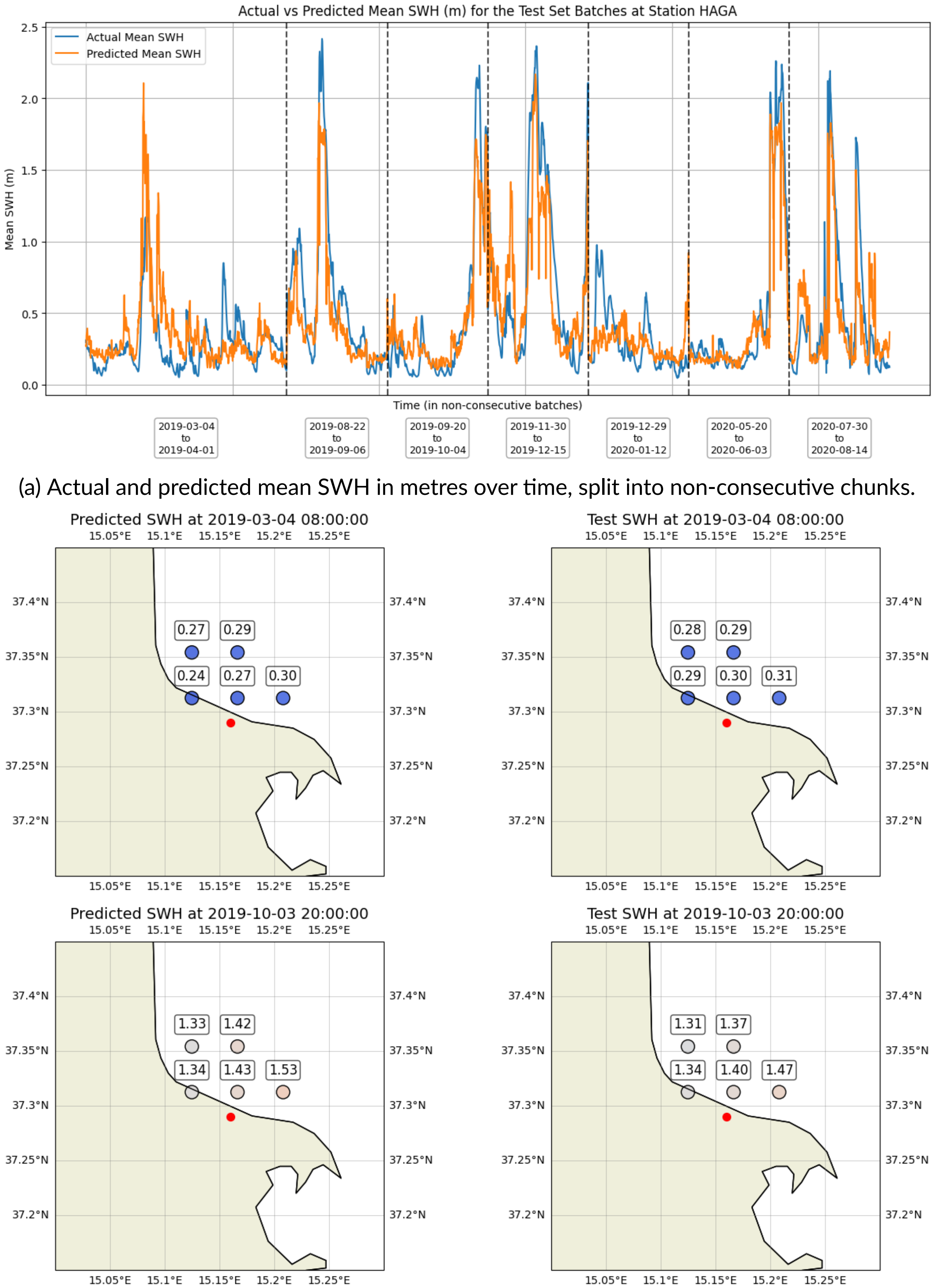
Figure 5.9 Data visualisation for best fold results at station CAVT.



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

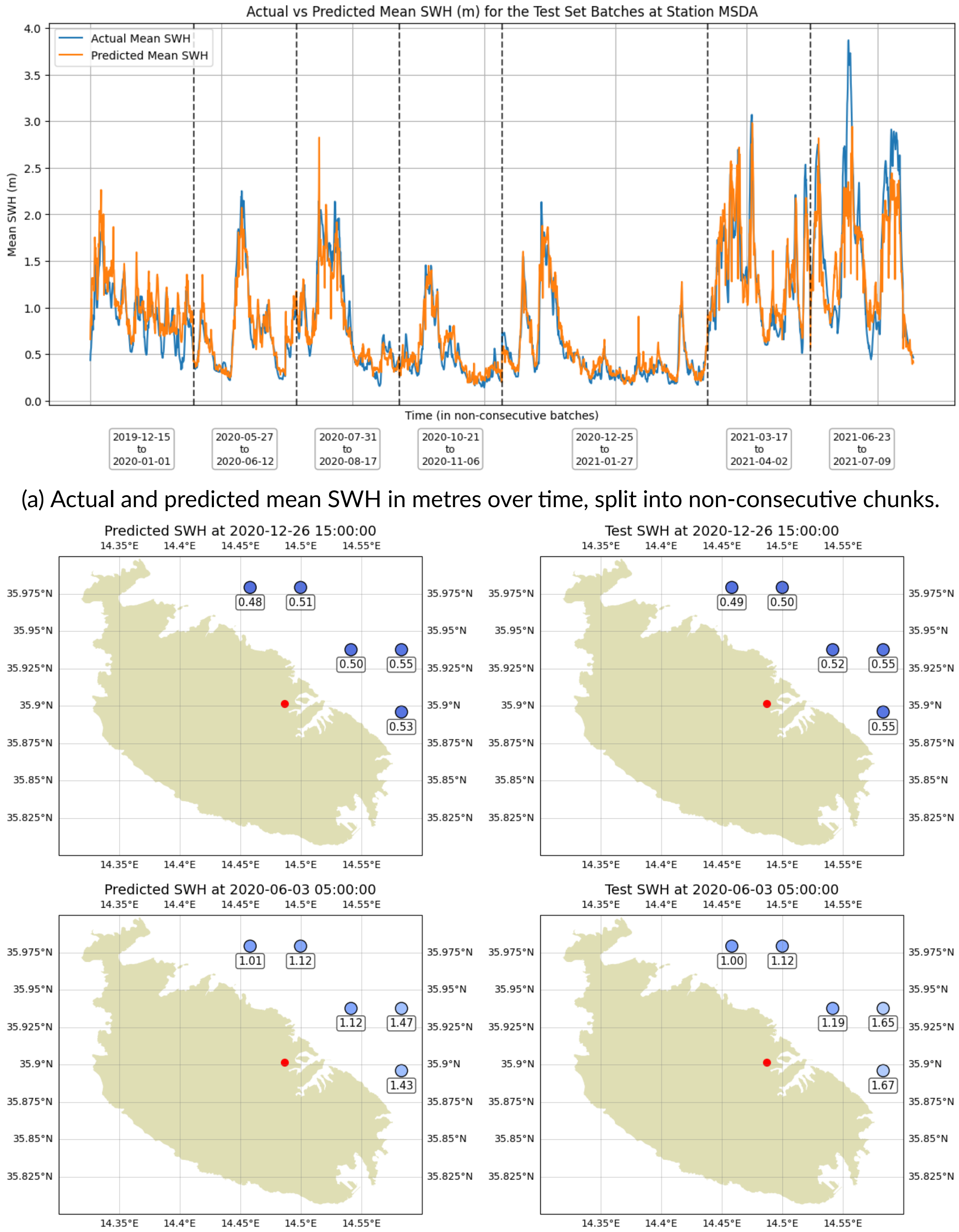
Figure 5.10 Data visualisation for best fold results at station CSLB.



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

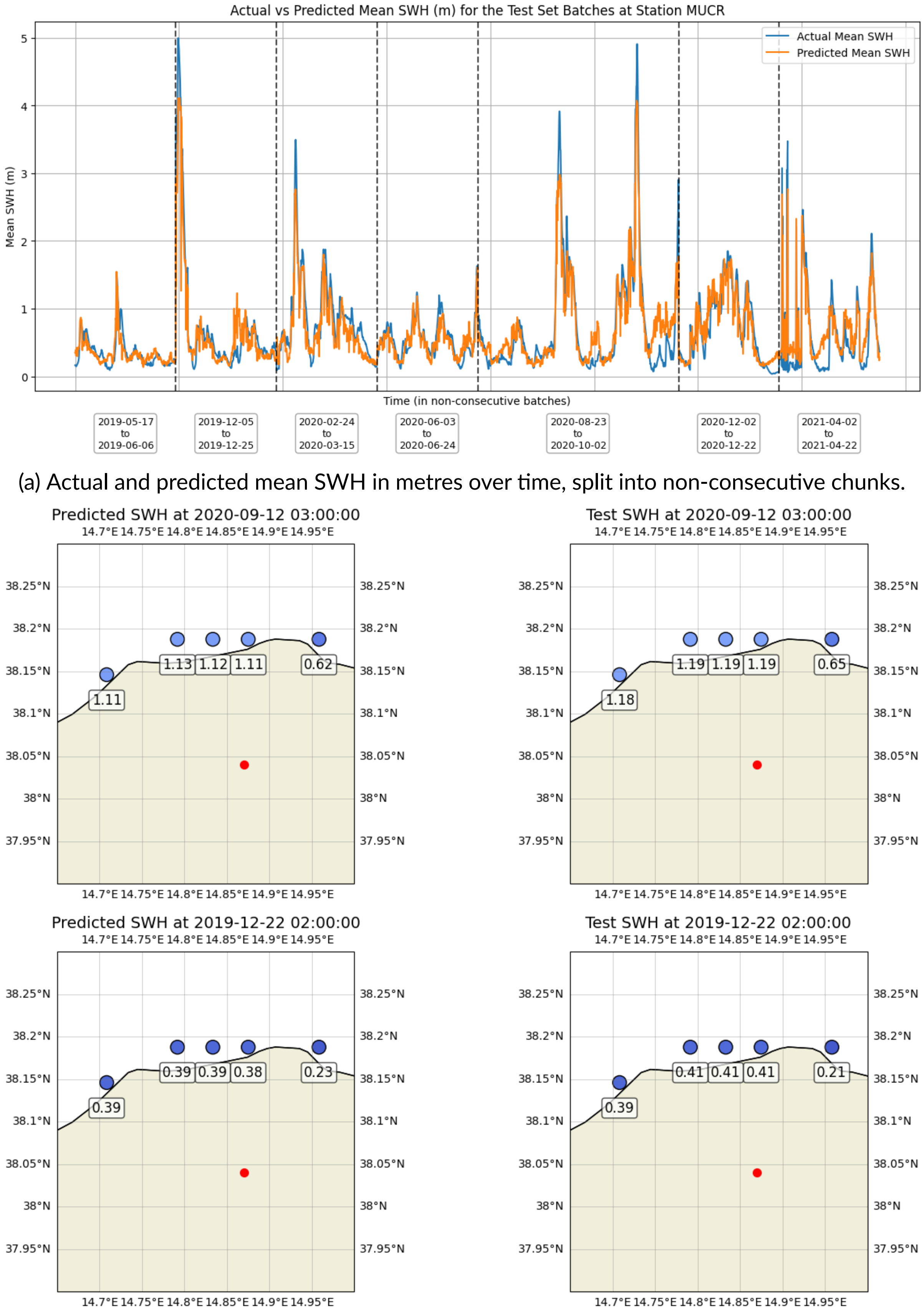
Figure 5.11 Data visualisation for best fold results at station HAGA.



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

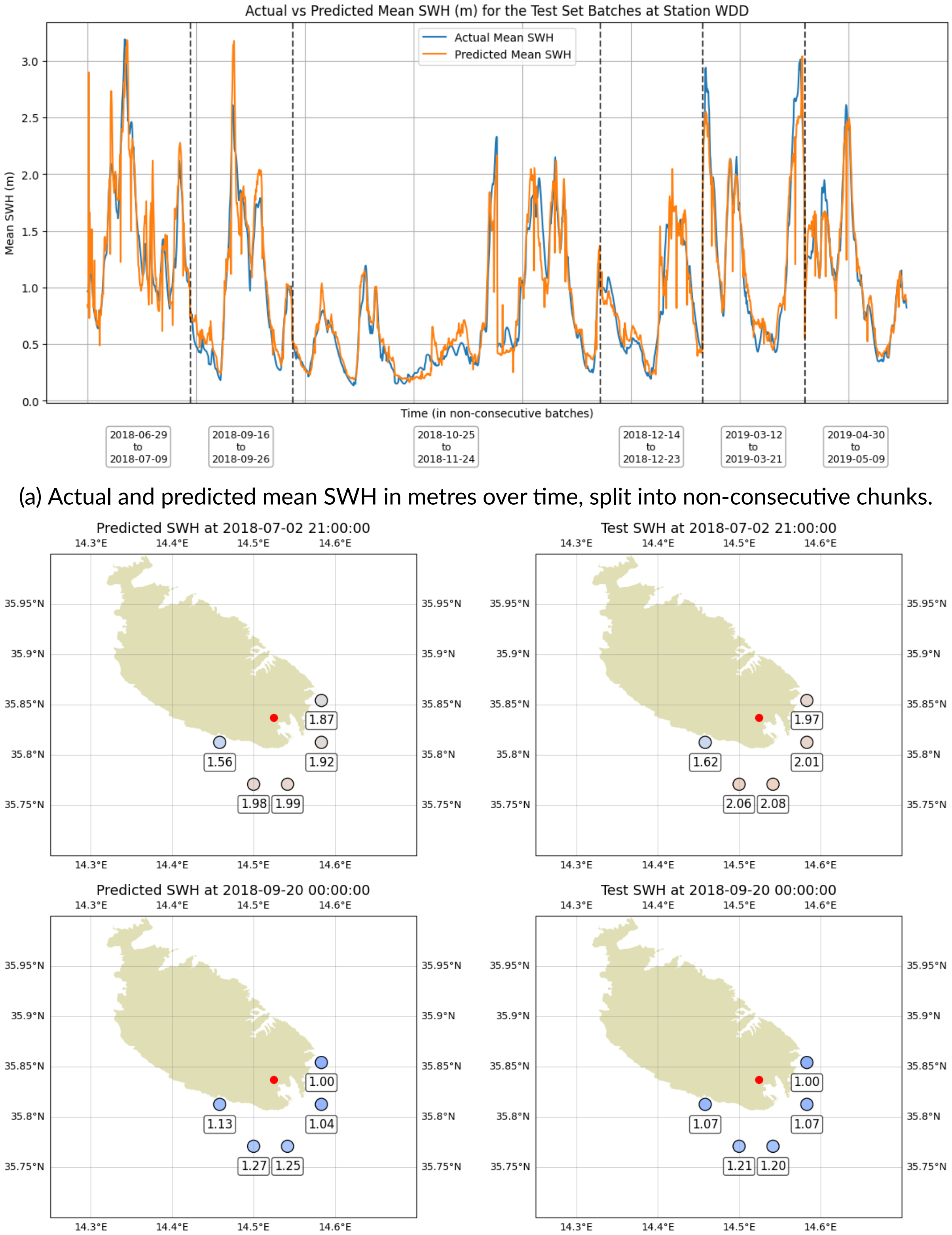
Figure 5.12 Data visualisation for best fold results at station MSDA.



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

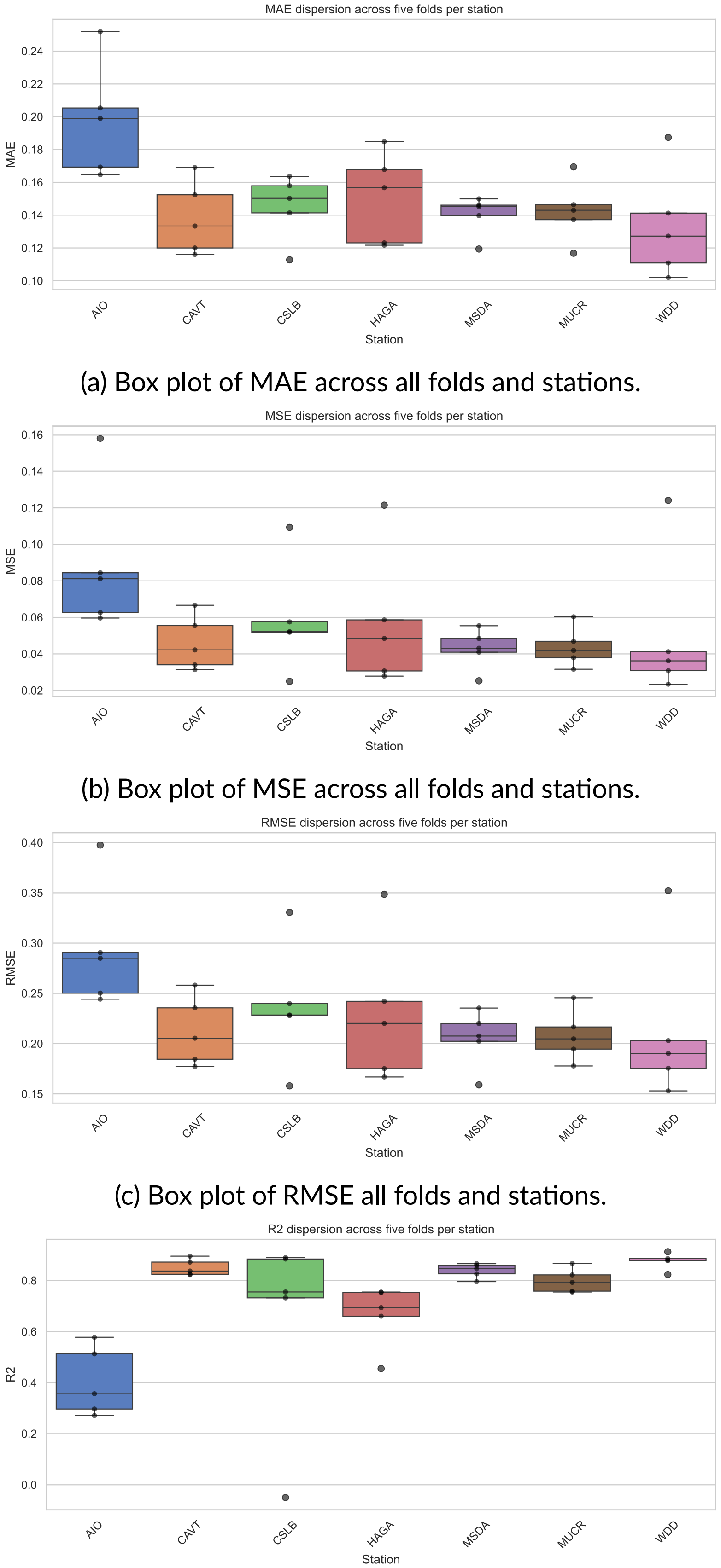
Figure 5.13 Data visualisation for best fold results at station MUCR.



(a) Actual and predicted mean SWH in metres over time, split into non-consecutive chunks.

(b) Two examples of actual and predicted SWH at the five grid cells nearest to the station.

Figure 5.14 Data visualisation for best fold results at station WDD.



(a) Box plot of MAE across all folds and stations.

(b) Box plot of MSE across all folds and stations.

(c) Box plot of RMSE all folds and stations.

(d) Box plot of R^2 across all folds and stations.

Figure 5.15 Dispersion of MAE, MSE, RMSE and R^2 across folds for all stations.

Table 5.18 Comparison to baseline results

	MAE (m)	MSE (m^2)	RMSE (m)	R^2
Numerical Methods				
Ferretti et al.	0.19	–	–	–
AI-Based Solutions				
Cannata et al.	~0.1	–	–	–
Minio et al.	0.21±0.23	–	–	0.89
Baranbooei et al. Scenario 1	0.6132	–	0.8780	0.8363
Baranbooei et al. Scenario 2	0.6816	–	0.9505	0.8059
Recreated Baseline				
Mean across stations	0.16101	0.07887	0.26627	0.69662
Standard Deviation	0.03844	0.06308	0.08931	0.20469
Minimum	0.10850	0.03045	0.17449	0.32966
Maximum	0.24297	0.25200	0.50199	0.86804
Final Models				
Mean across stations	0.13476	0.04386	0.20456	0.82978
Standard Deviation	0.02701	0.01838	0.04487	0.09907
Minimum	0.10066	0.02073	0.14398	0.60686
Maximum	0.18243	0.07107	0.26659	0.92060

5.3 Comparison with Baseline Models

The performance of the proposed models was placed in the context of established baselines from previous research and recreated models. The recreated baselines represented an approach as similar as possible to that of Minio et al. [7], but with a model trained for each station. They served as a benchmark for evaluating improvements offered by seismic-based regression techniques. The evaluation was conducted using the same metrics as the baselines (MAE, MSE, RMSE and R^2) to ensure fair and consistent comparisons.

The summary of baseline results in Table 3.4 led to Objective 1 of this research, which was to establish a scientifically robust baseline for the relationship between hourly seismic RMS amplitude and SWH. Table 5.18 shows that the baseline models have a consistent performance, however, an MAE of approximately 0.7m in sea wave height could be improved, since it can be considered quite significant in the practical context of this research. Overall, with the exception of Baranbooei et al. [67], the baseline results lack detail, precision and variety in the evaluation metrics.

In terms of MAE, the final models that were developed in this research achieve reductions of just over 50% compared to the results reported by Minio et al. [7], and up to 85% compared with the results reported by Baranbooei et al. Scenario 2 [67]. The RMSE also improved by around 85% between Baranbooei et al. Scenario 2 [67] and the final models. This substantial improvement supports the hypothesis articulated in Objective 4, which suggested that models tuned to local environmental characteristics would yield better predictive accuracy.

Furthermore, the highest MAE observed across the seven stations was 0.18243 m, closely aligned with the results obtained by Ferretti et al. [18] in a method employing MCMC. Crucially, this performance surpassed that of the recreated baseline implemented as part of this study, reinforcing the claim that carefully designed AI models can outperform traditional techniques even when relying on minimally processed input data, and using low computation cost models, in line with Objectives 2 and 3 of this research.

This finding was also in line with Objective 3, which anticipated that seismic data, in near-raw form, could be directly mapped to meaningful sea parameters without the need for extensive feature engineering. The results indicated that the models not only learned patterns effectively but also generalised well across different conditions, offering a robust alternative to the heavily preprocessed approaches reviewed.

The relatively lower performance of the developed models during periods of elevated SWH aligns with the observations reported by Baranbooei et al. [67], who similarly noted a decline in model accuracy under extreme sea state conditions.

To further benchmark performance, a recreated baseline model yielded a mean MAE of 0.16101 m and an R^2 of 0.69662. Notably, the spread in performance across stations was considerable, with MAE ranging from 0.10850 m to 0.24297 m, and R^2 values ranging from 0.32966 to 0.86804. In contrast, the final models demonstrated a consistent improvement across all evaluation metrics. The mean MAE decreased to 0.13476 m, a reduction of approximately 16.3% compared to the recreated baseline. Similarly, the mean RMSE dropped to 0.20456 m, and the average R^2 improved to 0.82978, indicating better fit and reduced residual variance. The performance gains were also more stable across stations, evidenced by lower standard deviations in all metrics. The minimum and maximum MAE for the final models were 0.10066 m and 0.18243 m respectively, while the R^2 ranged from 0.60686 to a peak of 0.92060.

To assess whether the improvements of the final models over the baseline were statistically significant across stations, a Wilcoxon signed-rank test was applied to paired metrics for stations where both baseline and final model results were available. The Wilcoxon signed-rank test is a non-parametric paired test that evaluates whether the median difference between paired observations is significantly different from zero, without assuming a specific distribution of the differences [84]. This approach is suitable in this context, as only a limited number of paired observations (stations) were available for comparison. The test was conducted for MAE, MSE, RMSE, and R^2, comparing the baseline and final model metrics across stations. The results indicate that the increase in R^2 was statistically significant ($p = 0.0312$), suggesting that the final models capture more variance than the baseline. In contrast, reductions in MAE ($p = 0.0938$), RMSE ($p = 0.0938$), and MSE ($p = 0.6875$) were not statistically significant at the 5% level, indicating that the observed decreases in error metrics could plausibly occur by chance across stations. These results provide evidence that, while the final models improve predictive accuracy in terms of R^2, improvements in absolute error metrics are less consistent across all stations.

Although the methodologies used for the recreated baseline and final models differed, particularly in terms of data preprocessing and station selection, six common stations allowed for a comparison. A key limitation of the baseline models was the use of linear interpolation to address missing values in the seismic time series, replicating the approach by Minio et al. [7]. As illustrated in Figure 5.1, this method led to the inclusion of unrealistic, synthetic data, which degraded model performance. In contrast, the final models employed a more conservative interpolation strategy, restricting linear gap-filling to a maximum of eight consecutive hours, at the expense of fewer data points, in alignment with Objective 5. This constraint significantly improved data quality and thereby model accuracy, which improved on stations for which there was at least one year of data.

Another critical distinction between the two approaches was in the selection and tuning of model hyperparameters. The baseline models employed a uniform set of hyperparameters across all stations, directly adopted from the configuration used by Minio et al. [7]. While this approach offered simplicity and reproducibility, it did not take into account local variability in the relationship between seismic energy and wave height. In the final models, however, as noted in Objective 4, station-specific hyperparameter tuning was conducted, enabling the models to better capture regional dynamics. This station-level optimisation revealed the adaptability of RF regressors, whereby seismic stations were in close physical proximity to their corresponding sea data sources. This entailed expanding the region of interest over Minio et al. [7], to include data as close to the stations towards the north of Sicily. The performance gains observed validate the importance of individualised model calibration in physically distributed systems.

Lastly, it is good to reiterate that while Minio et al. [7] made use of professional hardware, with the associated costs, the experiments within this dissertation were carried out on a personal computer, without a GPU, in line with Objective 2 of this research.

Overall, these findings validated the efficacy of the proposed models in capturing the complex relationship between RMS seismic amplitude and sea SWH, surpassing both traditional approaches, and recreated baselines. The improvement in accuracy and consistency across geographically dispersed stations demonstrated the potential of AI to enhance marine environmental monitoring from onshore seismic data.

5.4 Error Analysis

The evaluation of model performance across all seven stations (AIO, CAVT, CSLB, HAGA, MSDA, MUCR, and WDD) revealed consistent strengths as well as some limitations in the predictive ability of the trained models. While most stations demonstrated relatively low error margins and strong R^2 values, some patterns of underperformance were identified.

The K-fold cross validation results, particularly the visual comparison of actual versus predicted SWH in Figures 5.8a–5.14a, revealed that the models frequently underestimated peak wave heights, especially during stormier periods. The RF regressors tended to smooth out these extremes. The practical implications of such performance would be underestimating wave heights during rough sea states, which can pose safety concerns for marine forecasting applications.

Station AIO stood out as the weakest performer, exhibiting significantly lower R^2 values and higher error metrics even after model tuning. This disparity persisted despite improvements made between the recreated baseline and final models. The causes were not apparent from initial exploratory data analysis, prompting a post-evaluation investigation to assess possible contributing factors.

To further examine this issue, the SWH data was segmented using key percentiles (25%, 50%, 75%, 90%, and 95%) for each station. As shown in Figure 5.16, stations such as CAVT, MSDA, and WDD exhibited broader SWH distributions, with their 95th percentile exceeding 1.75m – and WDD even surpassing 2m. In contrast, AIO, CSLB, HAGA, and MUCR showed compressed distributions, with 95th percentiles under 1.5m. This implied that high wave events were relatively rare in these datasets.

A key takeaway was that the RF models struggled most with estimations beyond the typical data range, which was not unexpected given their reliance on exposure to such occurrences through data frequency. Since high wave heights exceeding 2.5m occurred in fewer than 5% of samples across all stations, the models had limited exposure to these events during training.

Although station AIO shared similar wave height distribution characteristics with station HAGA, its performance was substantially worse. This suggests that factors outside the scope of this study, such as instrument calibration, seismic noise contamination, or local bathymetric or topographic effects may be affecting the seismic signal quality at this station.

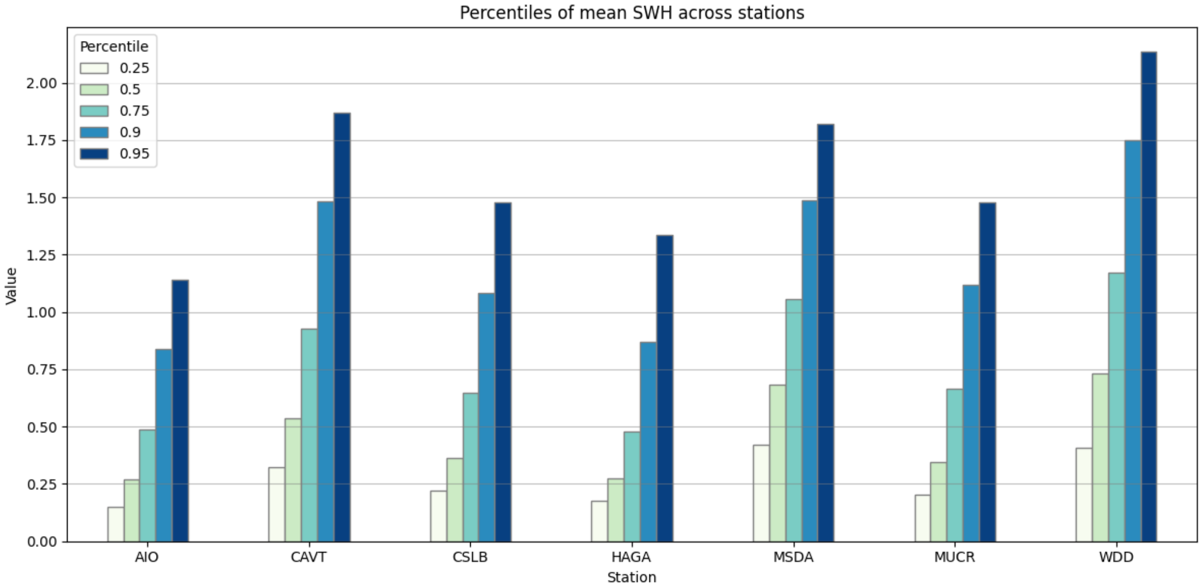


Figure 5.16 The percentiles of the mean SWH of the five nearest grid cells to each station.

To evaluate the effect of the underrepresentation of high sea states in the training data, the test set values of the mean SWH from the five nearest grid cells to each station were binned into 1-meter intervals. The corresponding MAE was then calculated for each interval. As illustrated in Figure 5.17, there is a clear trend of increasing error with increasing target SWH. This trend is particularly steep beyond 3 m, where the error begins to rise exponentially.

Notably, not all stations exhibited SWH values exceeding 3 m, reflecting the rarity of such events in the dataset. For those that did, the lack of sufficient training data in this range appears to have significantly compromised the model's generalisation ability. The steep increase in MAE at higher intervals indicated that the models struggled to extrapolate effectively beyond the conditions they were most frequently exposed to during training.

This highlighted a limitation in the datasets' distribution: the model's predictive capacity was tuned to lower sea states, which dominate the data. The influence of this imbalance was evident in the magnitude and variability of the error, particularly at the higher end of the SWH spectrum.

5.5 Summary of Key Findings

To conclude this chapter, a summarised version of the key findings is provided hereunder. These findings serve a two-fold purpose: first, to highlight the research objectives that have been successfully met; and second, to draw attention to any additional insights uncovered throughout the experimentation and analysis phases.

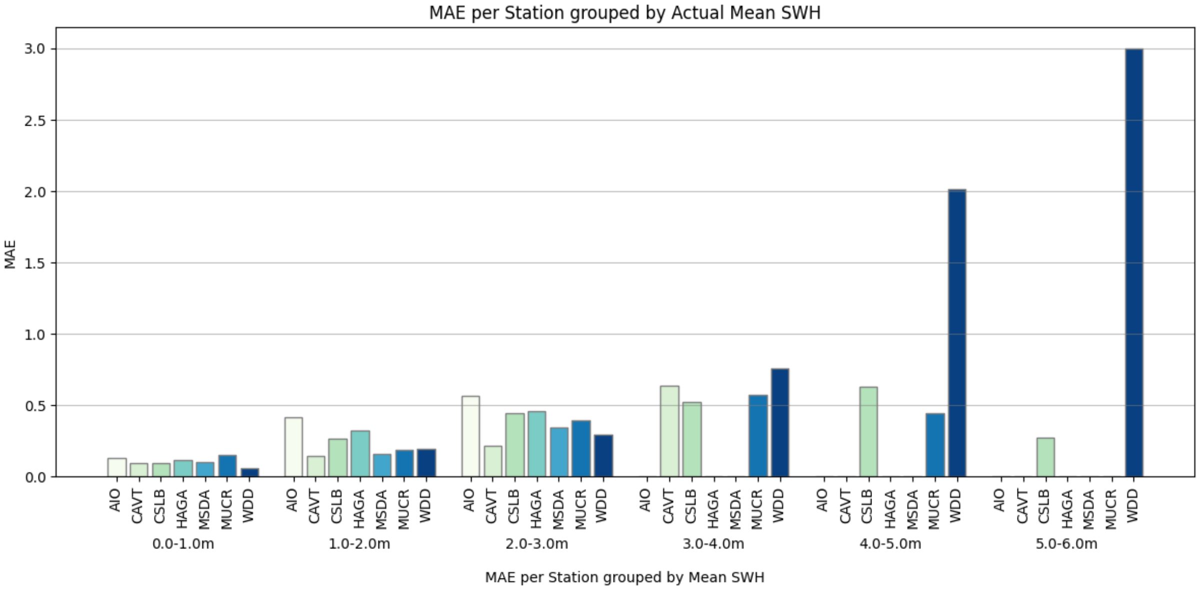


Figure 5.17 Average MAE per station grouped by actual mean SWH of the test sets.

5.5.1 Objectives

The key findings directly tied to the core objectives of this research are outlined below.

Objective 1 – Baseline Relationship

The recreated baseline model, built using comparable methods to those in prior literature, particularly Minio et al. [7], provided a realistic benchmark for evaluation. The baseline demonstrated a mean R^2 of 0.69662, and a MAE of 0.16101 m across all stations. However, less robust interpolation methods and lack of station-level hyperparameter tuning impacted the models' ability to generalise. Further to this research, the 'Final Models' referred to within this dissertation are proposed to be the baseline models for further research in this area, since they are built upon a solid methodology, as detailed in Chapter 4. Moreover, these models provided a mean R^2 of 0.82978, and a MAE of 0.13746 m across stations AIO, CAVT, CSLB, HAGA, MSDA, MUCR and WDD. The minimum MAE observed is 0.10066 at station CAVT, while the highest R^2 was 0.92060 at station WDD.

Objective 2 – Cost-Effective Solution

As detailed earlier, all experiments were conducted on a personal laptop running Microsoft Windows 10 Home (Version 10.0.19045). The system is equipped with an Intel Core i5-8250U CPU (1.6 GHz), 8 GB of RAM, and integrated Intel UHD Graphics 620. The device is an ASUS UX410UAR model. No discrete GPU was used for training. To this end, models cannot only be run, but can also be trained on consumer-grade hardware, making it a cost-effective solution. From a computational cost perspective, the methodology adopted is inherently more efficient than its predecessors, since a model is developed for each station, using data in physical proximity to it. Previous work saw one model being trained for all stations, using all station data at the input, and all sea data within the region of interest as the target variable, requiring high-end hardware.

Objective 3 – Efficient and Deployable Pipeline

The comparison to the recreated baseline saw that less preprocessing returned an improvement of 0.13316 and 0.02625m in the mean R^2 and the mean MAE across all stations, respectively. The maximum R^2 at station WDD increased from 0.841 at the recreated baseline to 0.921 at the final models. This verified that the RF regressors are not only able to learn from near-raw seismic data, with minimal pre-processing, but actually can perform even better under these conditions.

Objective 4 – Location-Specific Hyperparameter Tuning

One of the fundamental changes in methodology over the only known detailed AI-based solution was the segregation of models – rather than one model covering the whole region of interest, separate models for each station were designed. This provided two main benefits: the RF regressors were able to learn local characteristics to the respective stations, and the hyperparameters of the RF regressors were individually tuned and optimised. The variance in hyperparameters for each station, as shown in Table 5.9, confirmed that such a strategy was beneficial. Most stations did benefit from a higher number of trees and an increased depth. Varying other hyperparameters caused less fluctuations in performance, as shown in Figure 5.4 and 5.5.

Objective 5 – Prioritising Real-World Data

Another fundamental improvement in methodology over Minio et al.'s approach was the use of the 'longest stretch' algorithm, to determine the longest continuous stretch of data available for each station, with minimal interpolation. The original linear interpolation of up to 5,000 data points was used in the recreated baseline (impact shown in Figure 5.1), while the 'longest stretch' algorithm was employed in the final models. The improvement in all performance metrics is a reflection of the improved methodology, which encompassed this algorithm.

5.5.2 Pertinent Findings

Through the research that was carried out, other pertinent findings were made that are not tied directly to a core objective, but could aid subsequent research.

Finding 1 – Station AIO

From the recreated baseline until the final models, it was evident that station AIO was underperforming when compared with other stations. The exact reasons for this are unknown; however, exploratory data analysis did not reveal any peculiarities tied to this station alone. The issue may lie in sensor calibration resulting in inaccurate measurements, or a geographical parameter outside the scope of this research.

Finding 2 – Station CSLB

K-fold cross validation revealed that at the best performing fold, station CSLB had a high R^2 when compared with other stations. However, one of the folds led to a very poor fitness of R^2. This suggested that certain data points within station CSLB contribute to a poor overall performance. Such a data quality issue may arise from sensor calibration leading to inaccurate measurements, but it is unlikely that a geographical parameter caused this performance, since it occurred at only one fold.

Finding 3 – Class imbalance

Although the work carried out in this research focuses on regression, the consistent underestimation during periods of higher SWH suggested that a class imbalance existed in the dataset, whereby there is not enough exposure to such conditions in the training set. The subsequent error analysis that was carried out verified this – less than 5% of the available data referred to SWH exceeding 2.5 m.

6 Conclusion

This research has successfully established a robust baseline for the relationship between seismic hourly RMS amplitude and SWH. Beginning with the reproduction of an existing methodology, several opportunities for enhancement were identified. Raw data were queried, extracted, and used to reconstruct a baseline, upon which a series of innovations and refinements were implemented. The final models were then rigorously compared against both the reproduced baselines and those established in the literature.

By adhering closely to the methodologies reviewed in prior studies, notably Minio et al. [7], the performance indicators presented in the literature were independently verified. Notably, this research advanced upon Minio et al.'s approach by training models individually for each seismic station and its corresponding five nearest grid cells, rather than employing a single model across all stations. A comprehensive suite of evaluation metrics was generated for both the recreated baseline and subsequent models, thereby establishing a robust framework for model assessment and fulfilling Objective 1 of this study.

Moreover, Objective 2 concerned the feasibility of training and deploying models on consumer-grade hardware, which was successfully achieved. Both the recreated baselines and final models were developed and tested using readily accessible computing resources. This underscores the practical applicability of the proposed methodology, particularly in settings with limited financial or infrastructural capacity, thereby contributing to greater accessibility of artificial intelligence technologies.

In pursuit of Objective 3, the final models were deliberately designed to minimise data processing overhead. This led to the development of computationally lightweight models, capable of efficient inference even in resource-constrained environments. Reducing the need for extensive preprocessing steps not only shortened inference times but also enhanced the operational efficiency of the models once deployed.

In addressing Objective 4, hyperparameter tuning was conducted individually for each station, thereby tailoring the models to the local characteristics of the seismic and oceanographic conditions. This station-specific optimisation strategy resulted in notable improvements, with the mean coefficient of determination (R^2) across all stations increasing by up to 0.13316 compared to the recreated baseline. Similarly, the mean absolute error (MAE) was reduced by 0.02355 m, reflecting a significant enhancement in predictive performance.

Finally, Objective 5 emphasised the utilisation of real-world data to ensure practical relevance. While initial baselines relied on linear interpolation to fill data gaps of up to 200 days, potentially introducing significant error, the final models employed a longest-stretch algorithm. This strategy limited training to data segments with at least one continuous year of records, permitting interpolation over gaps of no more than eight hours. This methodological refinement contributed to a substantial improvement in model accuracy and reliability.

6.1 Strengths and Limitations

Several methodological innovations contributed significantly to the improvements observed over the recreated baseline models. Central changes included the application of the longest-stretch algorithm and the strategy of hyperparameter tuning tailored to individual stations. The deliberate emphasis on a *less-is-more* and *quality-over-quantity* philosophy led to models trained predominantly on high-quality, real-world data. This approach fundamentally diverged from the baseline methodology, which relied heavily on linear interpolation to address missing data, thereby introducing synthetic information that did not reliably represent plausible real-world scenarios.

The station-specific hyperparameter optimisation strategy yielded substantial performance gains. For instance, at station AIO, the R^2 improved markedly from 0.350 in the recreated baseline to 0.606 in the final models. Similarly, at station WDD, R^2 increased from 0.841 to 0.921, highlighting the effectiveness of localised model tuning in capturing station-specific variability.

Nonetheless, certain limitations were identified. Analysis of the data from station AIO suggested the presence of underlying quality issues, likely stemming from sensor inaccuracies, potentially introduced through calibration or synchronisation errors. Furthermore, while initial evaluation of station CSLB indicated strong performance, K-fold cross-validation revealed inconsistencies across the five folds, with one fold exhibiting a negative R^2. This suggests the presence of heterogeneous or poor-quality data segments, underlining the need for more robust data quality assessment prior to model training.

Another recurring limitation observed across the final models was the underestimation of SWH during periods of extreme weather. Detailed error analysis attributed this to a class imbalance within the dataset, with less than 5% of observations corresponding to SWH values exceeding 2.5 m. This imbalance hindered the models' ability to accurately capture and predict extreme events, a common challenge in real-world environmental modelling.

6.2 Impact and Significance

This research contributes meaningfully to the subdomain of AI focused on real-world environmental modelling, particularly in contexts characterised by noisy, incomplete, or imperfect datasets. By demonstrating that lightweight, station-specific models can be trained and deployed effectively on consumer-grade hardware without the need for extensive data preprocessing, this study advances the development of accessible and scalable AI solutions. In doing so, it addresses a critical gap within AI research in general, which often focuses on idealised or heavily curated datasets, and rarely considers deployment constraints in resource-limited environments.

The methodological innovations introduced, particularly the use of a longest-stretch algorithm for data selection and localised hyperparameter tuning, offer a replicable framework for other applications where data scarcity, quality variability, and operational constraints are prominent. These techniques collectively enhance model generalisability and robustness, setting a precedent for the development of AI systems that are not only accurate but also practical for real-world environmental monitoring and forecasting tasks.

The broader implications of this research align directly with the United Nations Sustainable Development Goals 12 (Responsible Consumption and Production) and 14 (Life Below Water). Given that the livelihoods of over three billion people depend on the oceans and coastal regions, improving the ability to monitor and predict marine conditions, with minimum environmental impact, is of profound societal importance. Accurate and accessible wave height estimations can support coastal communities, policymakers, and disaster management agencies by enhancing early warning systems through real-time information, improving maritime safety, and informing sustainable coastal development practices.

Additionally, by reducing computational requirements and infrastructure dependency, this research democratises access to advanced predictive models, making them feasible for use by non-governmental organisations, local authorities, and research institutions working on constrained budgets. In this way, the work not only advances technical knowledge within AI but also contributes to more equitable and sustainable research and innovation.

6.2.1 Comparison with Recent Work by Baranbooei et al.

Baranbooei et al. [67] published a study closely aligned with the thematic scope of this dissertation very recently (April 2025), aiming to model the relationship between seismic signals and SWH using AI techniques. This contribution was discussed in detail in Section 3. While their work contributes to the field, it also offers a valuable point of comparison for contextualising the originality and methodological innovations introduced in this research.

This study shares a foundational objective and exhibits methodological parallels with the earlier work of Minio et al. [7], upon which this dissertation builds and improves. However, there are several critical distinctions that make the approach proposed within this dissertation unique.

Firstly, the geographical domains under investigation differ. Baranbooei et al. focus on the northeast Atlantic Ocean surrounding Ireland – a region characterised by complex and open-ocean dynamics. In contrast, this dissertation concentrates on the enclosed Mediterranean Sea. While Baranbooei et al. report a good model performance across SWH values up to 10 m, the SWH within the region of interest of this dissertation rarely exceeded 5 m. These differing environmental and data characteristics necessitate tailored analytical strategies, highlighting the importance of context-specific model design.

A further key difference lies in the spatial configuration of the data. Baranbooei et al. utilise seismic and oceanographic data collected from widely dispersed sources, with the buoy providing SWH measurements situated at a considerable distance from the coastline. Conversely, this dissertation employs data sources located within a geographically compact region. The approach developed in this dissertation explicitly accounts for the decrease of correlation between seismic and ocean wave data over distance, by training distinct models for each location. This modelling strategy enhances predictive precision and reflects a deliberate design choice rooted in data behaviour.

Methodologically, Baranbooei et al. employ ANNs for regression, whereas this research explores the performance of RF regressors. The evaluation of RF models provides renewed insight into the modelling of this complex relationship and contributes to the broader understanding of algorithm suitability in this context.

Finally, in terms of preprocessing, Baranbooei et al. [67], like Minio et al. [7], exclude data from periods of significant seismic activity, under the assumption that such events introduce noise. In contrast, this dissertation challenges that assumption, demonstrating that excluding these data may be an unnecessary complication. The omission of this step streamlines the data pipeline and prioritises minimal preprocessing.

Taken together, these distinctions underscore the originality of the methodology proposed in this dissertation. By addressing regional data characteristics and critically re-examining preprocessing conventions established in literature, this work makes a contribution to the research on seismic-oceanic data integration.

6.3 Future Work

Building on the findings and limitations identified in this study, several promising avenues for future research have been identified.

By applying advanced gap-filling techniques for seismic data such as DL-based techniques, as outlined in Chapter 2, model robustness could be significantly enhanced. While the longest-stretch algorithm used in this research restricted gaps to a maximum of eight hours, extending this threshold by applying more sophisticated interpolation methods could preserve greater volumes of real-world data without introducing synthetic bias. Crucially, any proposed method would need to maintain the overall computational efficiency of the system to ensure continued deployability on consumer-grade hardware, in line with the accessibility objectives of this study.

Additionally, targeted data augmentation strategies specifically aimed at addressing the class imbalance for extreme weather events represent a critical opportunity for improvement. Techniques such as synthetic minority over-sampling technique (SMOTE) can be used to synthetically generate data and address class imbalance through observing data in the neighbourhood. New data points are synthesised based on feature similarity to the periods of extreme wave height, which are underrepresented [85]. Such methods could be employed to enrich the training datasets with plausible extreme wave height scenarios. This would help the models better capture and predict rare but high-impact events, which are of particular importance for coastal safety and disaster risk management.

These future research directions aim to refine model performance, enhance reliability under extreme conditions, and broaden the practical applicability of seismic-based ocean wave height prediction systems, all while maintaining the core principle of accessibility that underpinned this study.

Ultimately, this research not only advanced the technical capabilities of AI-driven environmental models, but also contributed to the broader goal of fostering accessible, sustainable solutions that can support the resilience and well-being of coastal communities worldwide.

References

[1] R. Yang, F. Zhang, and M. Hou, "Oceanplan: Hierarchical planning and replanning for natural language auv piloting in large-scale unexplored ocean environments," in *Proceedings of the 18th International Conference on Underwater Networks & Systems*, 2024, pp. 1–5.

[2] U. N. D. of Economic and S. Affairs. "United nations sustainable development goals," Accessed: Apr. 25, 2025. [Online]. Available: `https://sdgs.un.org/goals`.

[3] IOC-UNESCO, *Global Ocean Science Report 2020—Charting Capacity for Ocean Sustainability*, K. Isensee, Ed. Paris: UNESCO Publishing, 2020.

[4] H. Islam, R. M. Campos, T. R. Ferreira, and C. Guedes Soares, "Hydrodynamic assessment of a biofouled wave buoy in coastal zone," in *International Conference on Offshore Mechanics and Arctic Engineering*, American Society of Mechanical Engineers, vol. 84324, 2020.

[5] T. Pedersen and E. Siegel, "Wave measurements from subsurface buoys," in *2008 IEEE/OES 9th Working Conference on Current Measurement Technology*, IEEE, 2008, pp. 224–233.

[6] J. Lampkin and R. White, "Space junk," in *Space Criminology: Analysing Human Relationships with Outer Space*, Springer, 2023, pp. 71–92.

[7] V. Minio, A. M. Borzì, S. Saitta, S. Alparone, A. Cannata, G. Ciraolo, D. Contrafatto, S. D'Amico, G. Di Grazia, G. Larocca, and F. Cannavò, "Towards a monitoring system of the sea state based on microseism and machine learning," *Environmental Modelling & Software*, vol. 167, p. 105 781, 2023, ISSN: 1364-8152. DOI: `https://doi.org/10.1016/j.envsoft.2023.105781`.

[8] A. Toffoli and E. M. Bitner-Gregersen, "Types of ocean surface waves, wave classification," *Encyclopedia of Maritime and Offshore Engineering*, pp. 1–8, 2017.

[9] T. Toomey, A. Amores, M. Marcos, A. Orfila, and R. Romero, "Coastal hazards of tropical-like cyclones over the mediterranean sea," *Journal of Geophysical Research: Oceans*, vol. 127, no. 2, e2021JC017964, 2022.

[10] B. McDonagh, E. Clementi, A. C. Goglio, and N. Pinardi, "The characteristics of tides and their effects on the general circulation of the mediterranean sea," *Ocean Science*, vol. 20, no. 4, pp. 1051–1066, 2024. DOI: `10.5194/os-20-1051-2024`.

[11] B. Gutenberg, "Microseisms," in *Advances in Geophysics*, vol. 5, Elsevier, 1958, pp. 53–92.

[12] A. Besedina and T. A. Tubanov, "Microseisms as a tool for geophysical research. a review," *Journal of Volcanology and Seismology*, vol. 17, no. 2, pp. 83–101, 2023.

[13] F. Ardhuin, L. Gualtieri, and E. Stutzmann, "How ocean waves rock the earth: Two mechanisms explain microseisms with periods 3 to 300 s," *Geophysical Research Letters*, vol. 42, no. 3, pp. 765–772, 2015.

[14] S. Bonnefoy-Claudet, F. Cotton, and P.-Y. Bard, "The nature of noise wavefield and its applications for site effects studies: A literature review," *Earth-Science Reviews*, vol. 79, no. 3, pp. 205–227, 2006, ISSN: 0012-8252. DOI: `10.1016/j.earscirev.2006.07.004`.

[15] P. D. Bromirski and F. K. Duennebier, "The near-coastal microseism spectrum: Spatial and temporal wave climate relationships," *Journal of Geophysical Research: Solid Earth*, vol. 107, no. B8, 2002. DOI: `https://doi.org/10.1029/2001JB000265`.

[16] K. Hasselmann, "A statistical analysis of the generation of microseisms," *Reviews of Geophysics*, vol. 1, no. 2, pp. 177–210, 1963.

[17] L. Li, "Understanding seismic body waves retrieved from noise correlations: Toward a passive deep earth imaging," Ph.D. dissertation, Université Grenoble Alpes, 2018.

[18] G. Ferretti, A. Zunino, D. Scafidi, S. Barani, and D. Spallarossa, "On microseisms recorded near the ligurian coast (italy) and their relationship with sea wave height," *Geophysical Journal International*, vol. 194, pp. 524–533, Jul. 2013. DOI: `10.1093/gji/ggt114`.

[19] A. M. Borzì, V. Minio, R. De Plaen, T. Lecocq, S. Alparone, S. Aronica, F. Cannavò, F. Capodici, G. Ciraolo, S. D'Amico, D. Contrafatto, G. Di Grazia, I. Fontana, G. Giacalone, G. Larocca, C. Lo Re, G. Manno, G. Nardone, A. Orasi, M. Picone, G. Scicchitano, and A. Cannata, "Integration of microseism, wavemeter buoy, hf radar and hindcast data to analyze the mediterranean cyclone helios," *Ocean Science*, vol. 20, no. 1, pp. 1–20, 2024. DOI: `10.5194/os-20-1-2024`.

[20] M. S. Longuet-Higgins, "A theory of the origin of microseisms," *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 243, no. 857, pp. 1–35, 1950.

[21] P. D. Bromirski, F. K. Duennebier, and R. A. Stephen, "Mid-ocean microseisms," *Geochemistry, Geophysics, Geosystems*, vol. 6, no. 4, 2005.

[22] J. Ferreira and C. G. Soares, "Modelling distributions of significant wave height," *Coastal Engineering*, vol. 40, no. 4, pp. 361–374, 2000.

[23] H. Sverdrup, W. Munk, S. I. of Oceanography, and U. S. H. Office, *Wind, Sea and Swell: Theory of Relations for Forecasting* (H.O. pub). Hydrographic Office, 1947. [Online]. Available: `https://books.google.com.mt/books?id=DvPyLfd1xdAC`.

[24] Z.-H. Zhou, *Machine learning*. Springer nature, 2021.

[25] E. Alpaydin, *Machine learning*. MIT press, 2021.

[26] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[27] P. L. Fernández-Cabán, F. J. Masters, and B. M. Phillips, "Predicting roof pressures on a low-rise structure from freestream turbulence using artificial neural networks," *Frontiers in Built Environment*, vol. 4, p. 68, 2018.

[28] J. Ghosh and Y. Shin, "Efficient higher-order neural networks for classification and function approximation," *International Journal of Neural Systems*, vol. 3, no. 04, pp. 323–350, 1992.

[29] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[30] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[31] A. Dziewonski, S. Bloch, and M. Landisman, "A technique for the analysis of transient seismic signals," *Bulletin of the seismological Society of America*, vol. 59, no. 1, pp. 427–444, 1969.

[32] J. Ronen, "Wave-equation trace interpolation," *Geophysics*, vol. 52, no. 7, pp. 973–984, 1987.

[33] S. Fomel, "Seismic reflection data interpolation with differential offset and shot continuation," *Geophysics*, vol. 68, no. 2, pp. 733–744, 2003.

[34] X. Chai, G. Tang, S. Wang, K. Lin, and R. Peng, "Deep learning for irregularly and regularly missing 3-d data reconstruction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 7, pp. 6244–6265, 2020.

[35] D. Yoon, Z. Yeeh, and J. Byun, "Seismic data reconstruction using deep bidirectional long short-term memory with skip connections," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 7, pp. 1298–1302, 2020.

[36] T. Samson and F. Aweda, "Forecasting rainfall in selected cities of southwest nigeria: A comparative study of random forest and long short-term memory models," *Acadlore transactions on geosciences*, vol. 3, no. 2, pp. 79–88, 2024.

[37] X. Gao and X. Bian, "Autonomous driving of vehicles based on artificial intelligence," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 4, pp. 4955–4964, 2021.

[38] A. Singla and T. Malhotra, “Challenges and opportunities in scaling ai/ml pipelines,” *Journal of Science & Technology*, vol. 5, no. 1, pp. 1–21, 2024.

[39] K. Gallagher, K. Charvin, S. Nielsen, M. Sambridge, and J. Stephenson, “Markov chain monte carlo (mcmc) sampling methods to determine optimal models, model resolution and model choice for earth science problems,” *Marine and Petroleum Geology*, vol. 26, no. 4, pp. 525–535, 2009.

[40] J. Dunkley, M. Bucher, P. G. Ferreira, K. Moodley, and C. Skordis, “Fast and reliable markov chain monte carlo technique for cosmological parameter estimation,” *Monthly Notices of the Royal Astronomical Society*, vol. 356, no. 3, pp. 925–936, 2005.

[41] M. Johannes and N. Polson, “Mcmc methods for continuous-time financial econometrics,” in *Handbook of financial econometrics: Applications*, Elsevier, 2010, pp. 1–72.

[42] H.-C. Chiu, “Stable baseline correction of digital strong-motion data,” *Bulletin of the Seismological Society of America*, vol. 87, no. 4, pp. 932–944, 1997.

[43] M. Müller, *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer International Publishing Switzerland, 2015. [Online]. Available: `https://www.researchgate.net/publication/290440858_The_Fourier_Transform_in_a_Nutshell`.

[44] S. Carabott, *In pictures: After storm helios - flowing valleys, angry waves and fallen walls*, Accessed: 23 November 2024, Feb. 2023. [Online]. Available: `https://timesofmalta.com/article/storm-aftermath-pictures-flowing-valleys-aggressive-waves.1012965`.

[45] B. Båth, *Spectral analysis in geophysics*. Elsevier, 2012, pp. 12, 13.

[46] SLB, *Signature*, Accessed: 23 November 2024, 2023. [Online]. Available: `https://glossary.slb.com/en/terms/s/signature`.

[47] A. Z. G Korres M Ravdas, *Copernicus monitoring environment marine service (cmems).* Data set. Accessed: 31 March 2025, 2019. [Online]. Available: `https://doi.org/10.25423/cmcc/medsea_multiyear_wav_006_012`.

[48] F. Ardhuin, J. E. Stopa, B. Chapron, F. Collard, R. Husson, R. E. Jensen, J. Johannessen, A. Mouche, M. Passaro, G. D. Quartly, et al., “Observing sea states,” *Frontiers in Marine Science*, vol. 6, p. 124, 2019.

[49] A. Cannata, F. Cannavò, S. Moschella, G. Di Grazia, G. Nardone, A. Orasi, M. Picone, M. Ferla, and S. Gresta, “Unravelling the relationship between microseisms and spatial distribution of sea wave height by statistical and machine learning approaches,” *Remote Sensing*, vol. 12, no. 5, p. 761, 2020.

[50] J. Hancock and T. M. Khoshgoftaar, “Leveraging lightgbm for categorical big data,” in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, IEEE, 2021, pp. 149–154.

[51] E. Fix, *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine, 1985, vol. 1.

[52] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, “Learning k for knn classification,” *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, Jan. 2017, ISSN: 2157-6904. DOI: `10.1145/2990508`.

[53] U. .-. S. for a Changing World, *Api documentation - earthquake catalog*, Accessed: 23 November 2024, 2024. [Online]. Available: `https://earthquake.usgs.gov/fdsnws/event/1/`.

[54] F. Ardhuin, E. Stutzmann, M. Schimmel, and A. Mangeney, “Ocean wave sources of seismic noise,” *Journal of Geophysical Research: Oceans*, vol. 116, no. C9, 2011.

[55] R. E. Anthony, A. T. Ringler, D. C. Wilson, M. Bahavar, and K. D. Koper, “How processing methodologies can distort and bias power spectral density estimates of seismic background noise,” *Seismological Research Letters*, vol. 91, no. 3, pp. 1694–1706, Apr. 2020, ISSN: 0895-0695. DOI: `10.1785/0220190212`.

[56] L. Blum, M. Elgendi, and C. Menon, “Impact of box-cox transformation on machine-learning algorithms,” *Frontiers in Artificial Intelligence*, vol. 5, p. 877 569, 2022.

[57] S. Patro and K. K. Sahu, “Normalization: A preprocessing stage,” *arXiv preprint arXiv:1503.06462*, 2015.

[58] Y. Guo, L. Fu, and H. Li, “Seismic data interpolation based on multi-scale transformer,” *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[59] H. Kaur, N. Pham, and S. Fomel, “Seismic data interpolation using cyclegan,” in *SEG technical program expanded abstracts 2019*, Society of Exploration Geophysicists, 2019, pp. 2202–2206.

[60] A. A. Khan, O. Chaudhari, and R. Chandra, “A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation,” *Expert Systems with Applications*, vol. 244, p. 122 778, 2024.

[61] S. Peiris and R. Hunt, “Revisiting the autocorrelation of long memory time series models,” *Mathematics*, vol. 11, no. 4, 2023, ISSN: 2227-7390. DOI: `10.3390/math11040817`.

[62] J. A. Cook and J. Ranstam, “Overfitting,” *British Journal of Surgery*, vol. 103, no. 13, pp. 1814–1814, Nov. 2016, ISSN: 0007-1323. DOI: `10.1002/bjs.10244`.

[63] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation,” *Peerj computer science*, vol. 7, e623, 2021.

[64] T. O. Hodson, “Root mean square error (rmse) or mean absolute error (mae): When to use them or not,” *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–10, 2022.

[65] J. Li, A. D. Heap, A. Potter, and J. J. Daniell, “Application of machine learning methods to spatial interpolation of environmental variables,” *Environmental Modelling & Software*, vol. 26, no. 12, pp. 1647–1659, 2011. DOI: `10.1016/j.envsoft.2011.07.004`.

[66] D. Craig, C. J. Bean, I. Lokmer, and M. Möllhoff, “Correlation of wavefield-separated ocean-generated microseisms with north atlantic source regions,” *Bulletin of the Seismological Society of America*, vol. 106, no. 3, pp. 1002–1010, 2016. DOI: `10.1785/0120150181`.

[67] S. Baranbooei, C. J. Bean, M. Rezaeifar, and S. E. Donne, “Determining offshore ocean significant wave height (swh) using continuous land-recorded seismic data: An example from the northeast atlantic,” *Journal of Marine Science and Engineering*, vol. 13, no. 4, p. 807, 2025.

[68] A. Moni, D. Craig, and C. J. Bean, “Separation and location of microseism sources,” *Geophysical Research Letters*, vol. 40, no. 12, pp. 3118–3122, 2013.

[69] M. Rajurkar, U. Kothyari, and U. Chaube, “Artificial neural networks for daily rainfall—runoff modelling,” *Hydrological Sciences Journal*, vol. 47, no. 6, pp. 865–877, 2002.

[70] I. N. di Geofisica e Vulcanologia (INGV), *Rete sismica nazionale (rsn)*, Data set. Accessed: 23 November 2024, 2005. [Online]. Available: `https://doi.org/10.13127/sd/x0fxnh7qfy`.

[71] J. F. Tan, R. R. Stewart, and J. Wong, “Classification of microseismic events via principal component analysis of trace statistics,” *CSEG RECORDER*, vol. 1, pp. 34–38, 2010.

[72] A. Lyubushin, “Microseismic noise in the low frequency range (periods of 1–300 min): Properties and possible prognostic features,” *Izvestiya, Physics of the Solid Earth*, vol. 44, pp. 275–290, 2008.

[73] J. Groos and J. Ritter, “Time domain classification and quantification of seismic noise in an urban environment,” *Geophysical Journal International*, vol. 179, no. 2, pp. 1213–1231, 2009.

[74] S. Chowdhury, Y. Lin, B. Liaw, and L. Kerby, “Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance,” in *2022 International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*, IEEE, 2022, pp. 17–25.

[75] P. Probst, M. N. Wright, and A.-L. Boulesteix, “Hyperparameters and tuning strategies for random forest,” *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 9, no. 3, e1301, 2019.

[76] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, “How many trees in a random forest?” In *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8*, Springer, 2012, pp. 154–168.

[77] A. Nadi and H. Moradi, “Increasing the views and reducing the depth in random forest,” *Expert Systems with Applications*, vol. 138, p. 112 801, 2019.

[78] M. Daviran, A. Maghsoudi, R. Ghezelbash, and B. Pradhan, “A new strategy for spatial predictive mapping of mineral prospectivity: Automated hyperparameter tuning of random forest approach,” *Computers & Geosciences*, vol. 148, p. 104 688, 2021.

[79] P. Probst, “Hyperparameters, tuning and meta-learning for random forest and other machine learning algorithms,” Ph.D. dissertation, lmu, 2019.

[80] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, “Investigation of performance metrics in regression analysis and machine learning-based prediction models,” 2022.

[81] A. V. Tatachar, “Comparative assessment of regression models based on model evaluation metrics,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, no. 09, pp. 2395–0056, 2021.

[82] S. Geisser, “The predictive sample reuse method with applications,” *Journal of the American statistical Association*, vol. 70, no. 350, pp. 320–328, 1975.

[83] X. Zhang and C.-A. Liu, “Model averaging prediction by k-fold cross-validation,” *Journal of Econometrics*, vol. 235, no. 1, pp. 280–301, 2023.

[84] I. C. A. Oyeka, G. U. Ebuh, et al., “Modified wilcoxon signed-rank test,” *Open Journal of Statistics*, vol. 2, no. 2, pp. 172–176, 2012.

[85] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.