Intelligent Systems

Project: Digital Twin for the Azores Free Technological Zone (DATAz) - Wave Properties Forecasting

Bernardo Morais Chagas (103639)

October 17, 2025

Repository link: **GitHub Repository**

Contents

1 Executive Summary

This project develops a data-driven forecasting model for wave properties in the Azores region, integrated in the Digital Twin of the Ocean (DTO) initiative. Because traditional numerical weather prediction methods are computationally intensive, this work explores efficient alternatives using ERA5 reanalysis data. Key wave parameters such as mean wave direction (mwd), mean wave period (mwp), and significant wave height (swh) are forecasted using a multivariate time-series approach that incorporates both temporal and spatial dependencies.

The methodology combines an autoencoder for dimensionality reduction with a Multilayer Perceptron (MLP) for one-hour-ahead predictions. The autoencoder compresses 252 input features into a 32-dimensional latent space, preserving essential information while reducing computational cost. The MLP then predicts wave properties based on these encoded features. Results show low normalized errors across all variables, demonstrating accurate and efficient forecasts.

Future work includes expanding the model to predict all atmospheric and oceanographic variables, enabling multi-step forecasts, and implementing a validation set for hyperparameter optimization to enhance robustness.

2 Introduction

Weather forecasting is considered a complex task that requires solving large-scale numerical models with high spatial and temporal resolution. These solvers demand a large computational power, making the process energy-intensive and dependent on supercomputing infrastructures. Within the framework of the DATAz – A Digital Twin for the Azores Free Technological Zone Project, there is a need to explore more efficient and accessible alternatives to traditional numerical weather prediction methods. The goal is to enable the evaluation of weather and oceanic scenarios in a virtual environment that does not rely solely on high-performance computing systems. This solution would allow the study of what-if scenarios, giving a powerful tool to the decision makers group.

As a first step towards this objective, and in the context of Intelligent Systems Course, a preliminary study was conducted using the ”ERA5 hourly time-series data on single levels from 1940 to present” dataset from the Copernicus Climate Data Store (CDS). This dataset provides high-resolution, hourly climate data derived from a combination of model outputs and historical observations. Focusing on the Azores Free Technological Zone, this initial approach allows for testing and validation of simplified data-driven modelling strategies, setting the foundation for the development of a lightweight, AI-supported digital twin of the ocean.

3 Problem Definition

The objective of this work is to develop a data-driven model capable of forecasting wave variables relevant to oceanic behaviour in the Azores region. The problem is framed as a multivariate time series forecasting task, where the goal is to predict the future state of key ocean parameters—such as wave height, surface wind, and sea surface temperature—based on past temporal and spatial patterns extracted from the ERA5 reanalysis data.

This forecasting capability is essential for the development of the Digital Twin of the Ocean (DTO), as it enables the simulation and anticipation of dynamic ocean conditions.

4 ERA 5 Dataset and Data Preparation

The dataset used in this work, ERA5, provides hourly atmospheric and oceanographic records with a spatial resolution of 0.25° in both latitude and longitude. It is the fifth-generation global atmospheric reanalysis produced by the ECMWF (for the Copernicus Climate Change Service). This data was obtained by assimilating historical observations (from weather stations, ships, buoys, satellites, etc.) [1] with numerical models (model states constrained by real observations) [2]. This process produced a coherent, gridded information system that provides a consistent estimate of the atmospheric state over time.

The dataset includes a total of 17 variables. The first three variables (latitude, longitude, and time) serve as coordinate references, used as indices to map the data spatially and temporally.

The remaining 14 physical variables, listed in Table 1, are the ones used to train, test, and later forecast wave properties.

Variable	**Description**	**Units**
`u10`	U-component of 10 m wind	m/s
`v10`	V-component of 10 m wind	m/s
`d2m`	2 m dew point temperature	K
`t2m`	2 m temperature	K
`msl`	Mean sea level pressure	Pa
`sst`	Sea surface temperature	K
`skt`	Skin temperature	K
`sp`	Surface pressure	Pa
`ssrd`	Surface solar radiation downwards	J/m^2
`strd`	Surface thermal radiation downwards	J/m^2
`tp`	Total precipitation	m
`mwd`	Mean wave direction	degrees
`mwp`	Mean wave period	s
`swh`	Significant wave height	m

Table 1: List of ERA5 variables used for model training and testing.

In order to use this data, the python notebook "0. Download_Dataset.ipynb" was built to allow, after API configuration, the download of the data of a given space and time interval. the files would be downloaded in a Zip file, having, for each point, two different files. They were then extracted and merged into a single file per spatial point.

After defining the regions in observation, the information was then stored in ”dataset_train” and ”dataset_test” folders

The python notebook ”1. Preprocessing.ipynb” was built to compile the datasets in each folder into a main one, denominated ”combined.nc”. Some visualization procedures were developed.

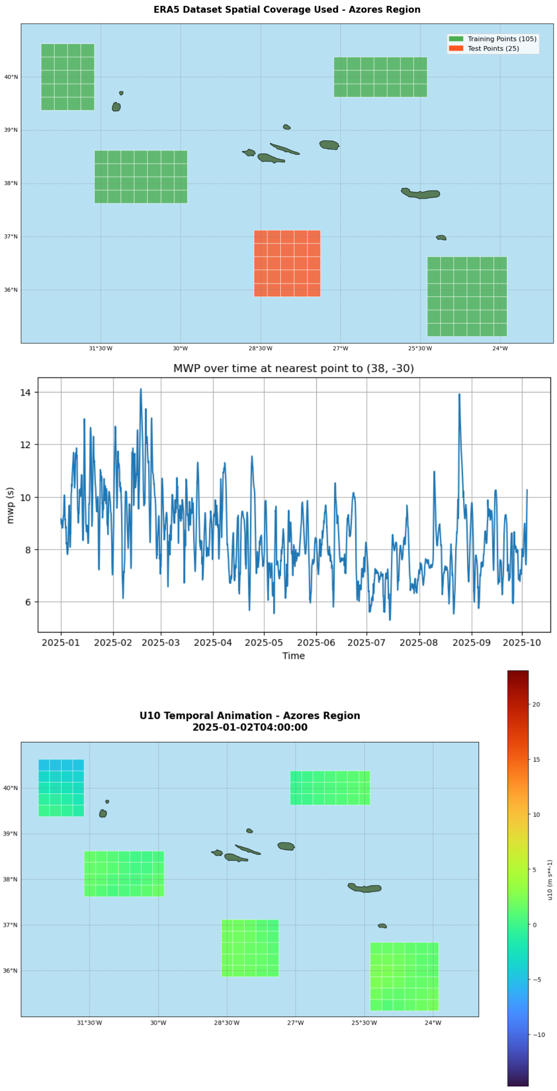

Figure 1: 1) ’ERA5 Dataset Spatial Coverage Use; 2) MPW over time at nearest point (38,-30); 3) U10 temporal animation

5 Methodology

5.1 Definition of the general approach

The first step consisted of defining the study region by selecting a square-shaped area inserted in Azores Archipelago region. The chosen geographical boundaries range from 35° to 41° North in latitude and from 23° to 33° West in longitude. This region adequately covers the islands and surrounding oceanic area of interest, allowing for a representative analysis of local atmospheric and oceanographic dynamics. Was also chosen to use data from January 1st 2025 to October 3rd 2025, having a total of 6624 registers for each spacial point.

The adopted approach to solve this forecasting problem follows a discrete state-space model architecture, in which the output vector depends on both the system state and the external input. For each spatial point, the corresponding wave properties were forecasted one step ahead in time. The output vector was estimated based on the wave properties from the previous one and two time steps (state vector), together with the remaining 11 meteorological and oceanographic variables (input vector).

To enhance prediction accuracy and better capture the spatial dependencies inherent to the problem, the model also incorporated information from the neighbours points. Given that wave behaviour is spatially correlated, the evolution of a target point is influenced by the surrounding environmental conditions. Therefore, a grid-based approach was implemented, where wave properties were predicted one hour ahead using not only the data from the central point but also from its eight adjacent neighbours.

This relationship can be expressed as follows:

$$Y_{1..n}^{t+1} = F\big(Y_{1..n}^{(t,t-1)}, , I_{1..m}^{(t,t-1)}\big) \tag{1}$$

where:

- $Y_{1..n}^{t+1}$ is the forecasted n wave properties for a spatial point at time $t+1$;
- $Y_{1..n}^{(t,t-1)}$ is the previous n wave properties for a spatial point at t and $t-1$;
- $I_{1..m}^{(t,t-1)}$ denotes the input model variables used to describe, combining the central point weather information with all the neighbours information

Focusing on a specific example, for a given central point, in order to forecast the wave properties one time sample (1 hour) ahead, the model input layer size would be (8+1, points) * (14, features) * (2 ,time steps) = 252.

5.2 Selecting the Study Spacial Region

Before downloading the extended space dataset for the all predefined region, was calculated first that, by having 24 rows over latitude, 41 columns over longitude, each point with 6624 time samples, would result in 6,5 millions registers in total.

In order to solve a data leakage problem that could happen by using the same dataset to both train and test subsets. By also considering a possible bias given by the different sea deep in each point, 4 distinguish squares were used as train set and 1 square for test set. This distributions it represented in the figure bellow.

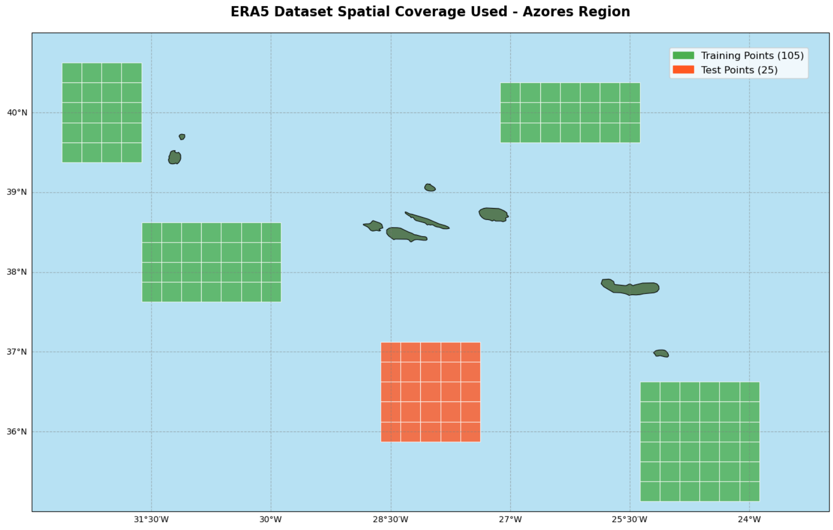

Figure 2: Spacial distribution of the datasets used for both train and test the model

6 Architecture Design and Training

Given the large input dimensionality, an autoencoder was first implemented to reduce the size of the input feature space. By designing the encoder with a latent dimension of 32, the data size was reduced by approximately a factor of eight, enabling more efficient training and better generalization of the forecasting model.

Subsequently, a Multilayer Perceptron (MLP) model was developed to forecast the wave-related properties based on the encoded features. The overall workflow is schematically illustrated in Figure 3.

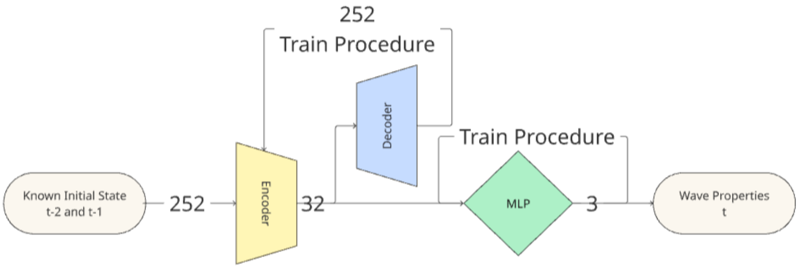

Figure 3: Schematic representation of the designed forecasting model.

To generate the training and testing datasets, a Python program named `GetData.py` was implemented. It contains the function `get_train_data`, which returns a vector list with the specified number of data samples. The function operates by randomly selecting a central spatial point at a given time step, identifying its eight surrounding neighbours, and aggregating the values of all features within the defined time horizon. In this way, for each selected point, the function outputs a single vector containing 252 values representing the spatio-temporal context of that location.

6.1 Autoencoder

Prior to training, the input data were normalized using a Standard Scaler transformation. Since the autoencoder relies on reconstructing its inputs to evaluate reconstruction error, no explicit target vector was required.

The model compresses the original 252-dimensional feature space into a 32-dimensional latent representation and then reconstructs the input, minimizing the reconstruction loss. After training, the autoencoder was evaluated on an independent test set. The mean normalized root mean square error (NRMSE) obtained was 0.112, indicating a satisfactory reconstruction performance.

A visual representation of the autoencoder's different stages is shown in the figure bellow.

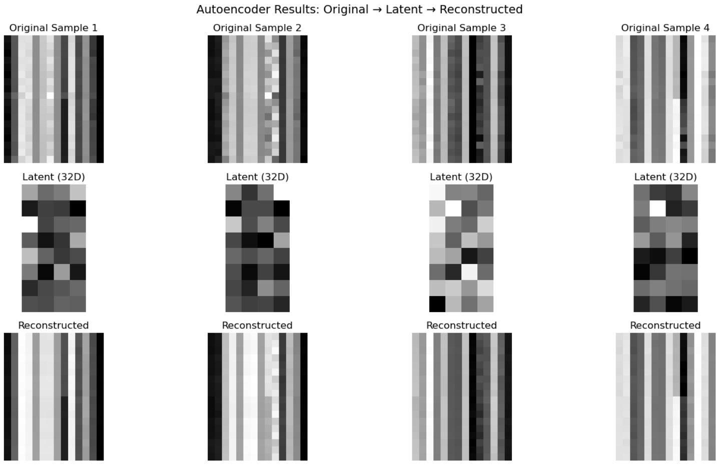

Figure 4: Autoencoder results showing the original, latent, and reconstructed feature spaces.

In addiction, after training the model, the standard scalar and the autoencoder where saved in folder "model" so this this same transformation properties can be used in the next phase

6.2 Multilayer Perceptron (MLP)

To develop the MLP model, both the training and testing datasets were first generated. Each dataset consists of input vectors with 252 features and corresponding target vectors with 3 output variables, representing the wave parameters of interest.

After data scaling and encoding through the trained autoencoder, the dimensionality of the input space was reduced to 32 latent features. These encoded features were then used as inputs to a Multilayer Perceptron composed of two hidden layers. The first hidden layer has 128 neurons (four times the size of the latent dimension),while the second hidden layer has 64 neurons (corresponding to half of the previous layer's size). Rectified Linear Unit (ReLU) activation functions were applied to introduce non-linearity, and a dropout layer was included to mitigate overfitting.

Although the use of a validation set is generally recommended for hyperparameter tuning, it was not implemented in this phase due to the relatively low complexity of the problem and the limited number of tunable parameters.

The MLP was trained using the Adam optimization algorithm with a learning rate of 0.001 and a mean squared error (MSE) loss function. The training was performed over 50 epochs with a batch size of 64, iteratively updating the model weights to minimize the prediction error between the estimated and true target values. During each epoch, the network processed batches of training samples, computed the forward and backward passes, and adjusted its parameters according to the gradient information provided by the optimizer. This iterative process allowed the MLP to progressively learn the nonlinear relationships between the encoded atmospheric–oceanic inputs and the target variables.e

7 Results

The forecasting model was evaluated for a one-hour prediction horizon of wave variables in the Azores region. The trained MLP received encoded features from the autoencoder, in a 32-dimensional latent space, and produced predictions for three target wave parameters: mean wave period (mwp), mean wave direction (mwd), and significant wave height (swh).

Table 2 summarizes the quantitative evaluation of the model on the test dataset for one-hour forecasting usign Mean Squared Error (MSE), Mean Absolute Error (MAE) and Normalized MAE (NMAE) .

Target	**MSE**	**MAE**	**NMAE**
mwd	100.72	7.84	0.0218
mwp	0.0597	0.1929	0.0217
swh	0.0454	0.1853	0.0253

Table 2: Performance metrics of the MLP model for one-hour wave forecasting.

To further validate the model performance, the following figures were generated. The first figure allows the visualization of the temporal evolution of true and predicted sequences over three consecutive time steps. The second figure displays the scatter comparison between predicted and true values for each test sample. The third represents the distribution of prediction errors in a histogram format.

7.1 Mean Wave Direction (mwd)

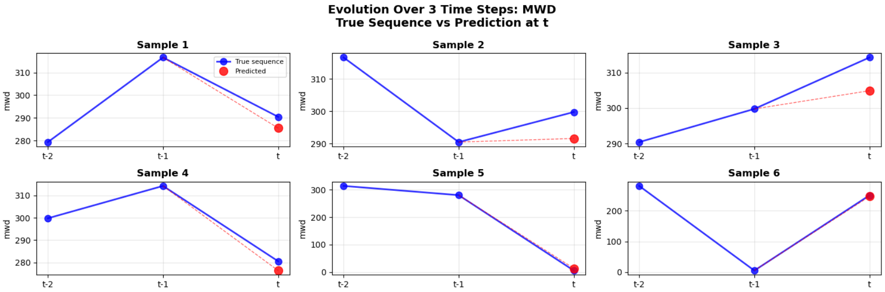

Figure 5: Temporal evolution over three time steps for `mwd` (True vs. Predicted).

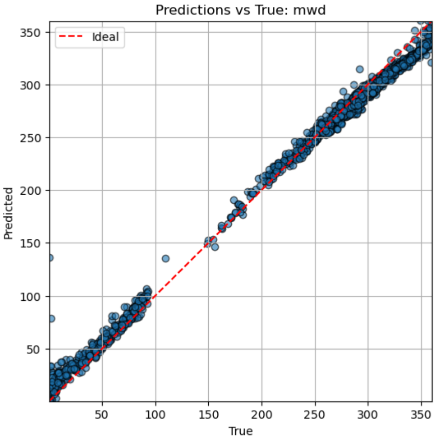

Figure 6: Predicted vs. true values for `mwd`.

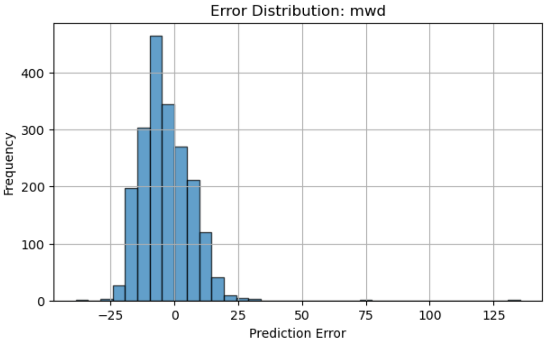

Figure 7: Error distribution for `mwd`.

7.2 Mean Wave Period (mwp)

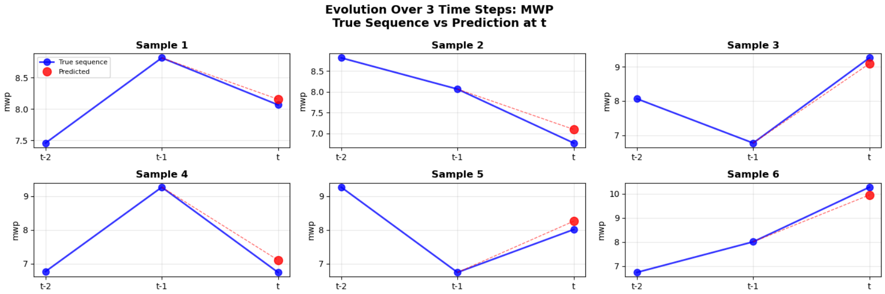

Figure 8: *Temporal evolution over three time steps for* `mwp` *(True vs. Predicted).*

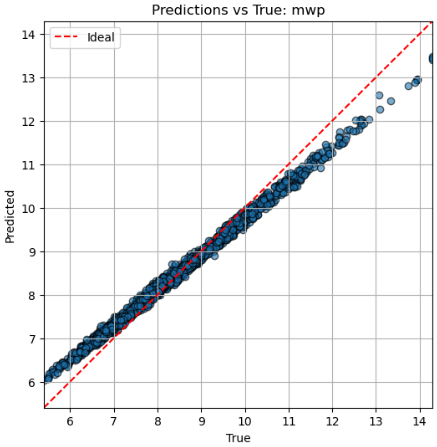

Figure 9: *Predicted vs. true values for* `mwp`.

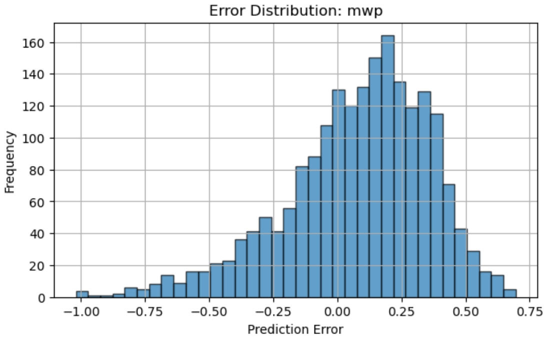

Figure 10: *Error distribution for* `mwp`.

7.3 Significant Wave Height (swh)

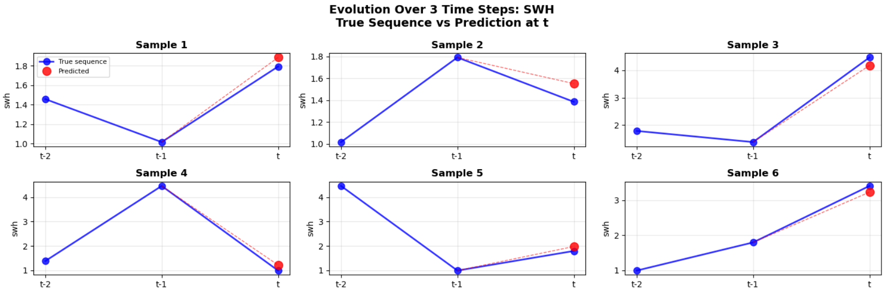

Figure 11: Temporal evolution over three time steps for `swh` (True vs. Predicted).

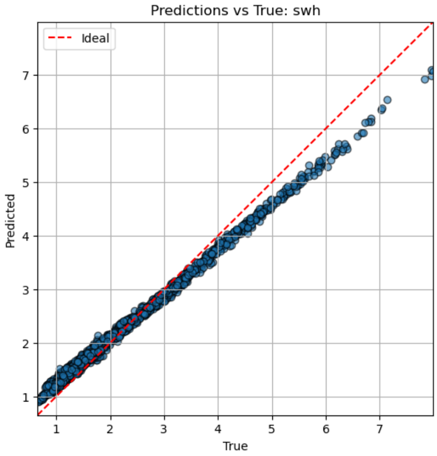

Figure 12: Predicted vs. true values for `swh`.

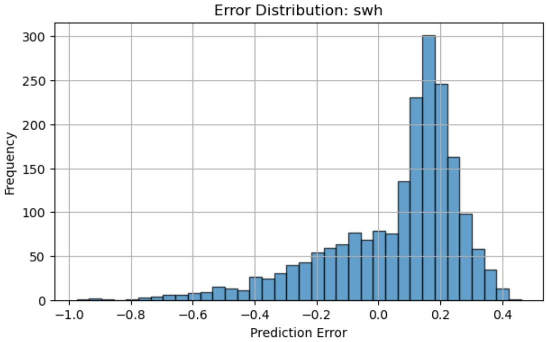

Figure 13: Error distribution for `swh`.

8 Discussion

Based on the results presented in Section 7. Results, the forecasting model demonstrated an overall good performance for one-hour wave prediction. This conclusion was supported by the relatively low MAE and NMAE values across all variables indicate that the model was able to learn and forecast well from the previous two steps in time features information to the one hour ahead wave properties.

Among the three predicted variables, the mean wave direction (mwd) exhibited the highest Mean Squared Error (MSE = 110.72) and Mean Absolute Error (MAE = 7.84), given the higher range of the data (0 to 360 degrees). When computed the normalized error (NMAE = 0.022) was verified that it was the lowest, confirming that the predictions are still consistent and physically meaningful relative to the variable's scale.

In contrast, the mean wave period (mwp) and significant wave height (swh) showed substantially lower MSE and MAE values. The model successfully captured short-term fluctuations in both variables, suggesting that the encoded features extracted by the autoencoder were able to retain essential information about the energy and temporal structure of the wave field. The accurate estimation of swh.

These results indicate that the compact representation produced by the autoencoder effectively preserved spatial and temporal dependencies while significantly reducing input dimensionality. However, the relatively higher uncertainty in mwd forecasts suggests that additional context, such as higher spatial resolution, longer temporal horizons, or more advanced temporal modeling, could further improve prediction stability. Incorporating physical constraints or hybrid physics–ML architectures could also enhance interpretability and robustness.

After training the model, it was possible to forecasts the wave properties within a few seconds on a standard personal computer. This highlights its computational efficiency and suitability for near real-time applications. Such performance demonstrates, and once more validates, the proposed autoencoder-MLP framework.

9 Conclusion

This project demonstrated the feasibility of short-term wave forecasting in the Azores region using a data-driven approach that integrates an autoencoder with a multilayer perceptron. By compressing 252-dimensional spatio-temporal inputs into a 32-dimensional latent representation, the model achieved efficient training and accurate one-hour-ahead predictions of mean wave direction, mean wave period, and significant wave height.

The low normalized errors across all target variables indicate that the proposed architecture successfully captures essential spatial and temporal dependencies while remaining computationally lightweight.

Beyond validating the concept, the study highlights the potential of combining unsupervised feature extraction with supervised learning for environmental forecasting tasks. The approach balances performance and computation power needed offering a scalable foundation for future extensions.

10 Next Steps/ Opportunities

The implemented methodology allows to predict the wave properties one hour ahead, using the actual information.

Increasing the output model dimension from 3 (wave properties) to 14 (all features), would allow us predict all weather properties of the central point as well one hour ahead. Given that the model is based on the information of the itself point, and the neighbours, increasing the output layer to 252, would allow us to predict all the features of all the small grid points. This data could be consequently encoded and used as $t-1$ input, improving significantly the forecasting capability, enabling the model to predict more than one hour ahead.

Given that a output layer of 252 would not be feasible, a more correct approach is offered where the output dimension is set to 32 instead. The output vector would be then the encoded features. To get to know the wave and weather properties, the trained decoded would be used. This procedure is pendent on model validation as well as hyper parameter tuning, being advised the creation of a validation set. The figure bellow represents the new sequence suggested:

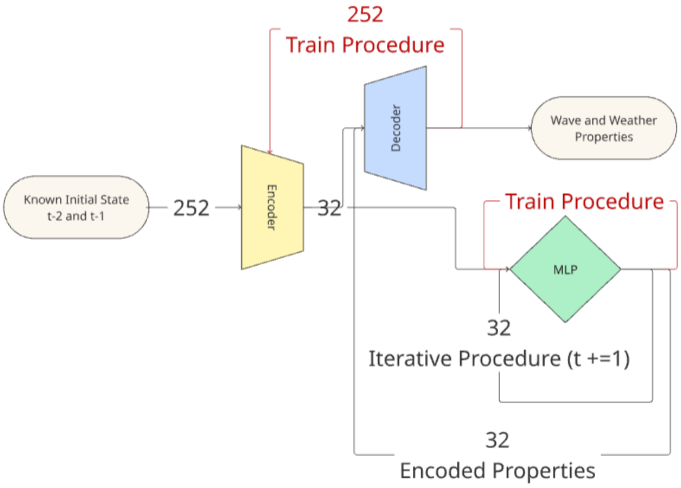
252
Train Procedure
Decoder
Wave and Weather Properties
Known Initial State t-2 and t-1
252
Encoder
32
Train Procedure
MLP
32
Iterative Procedure (t +=1)
32
Encoded Properties

Figure 14: Extended Forecasting Model

In addition, it would be advisable to create a validation set to better tune the hyperparameters without biasing the model. Methodologies such as cross-validation or one hot encoding could also be beneficial.

References

[1] ECMWF, "Observations," available at: https://www.ecmwf.int/en/research/data-assimilation/observations.

[2] ECMWF, "Assimilation methods," available at: https://www.ecmwf.int/en/research/data-assimilation/assimilation-methods.