HPC-Driven Oceanographic Predictions with Graph Neural Networks (GNNs) and Gated Recurrent Units (GRUs)

Paraskevi Vourlioti, Theano Mamouka, Maria Banti, Charalampos Paraskevas, Stylianos Kotsopoulos, Vasileios Alexandridis and Georgia Kalantzi

September 18, 2024

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2023) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

Proceedings of the First EuroHPC user day

HPC-Driven oceanographic predictions with Graph Neural Networks (GNNs) and Gated Recurrent Units (GRUs)

Paraskevi Vourlioti[a,*],Theano Mamouka[a], Maria Banti[a], Charalampos Paraskevas[a], Stylianos Kotsopoulos[a], Vasileios Alexandridis[b], Georgia Kalantzi[b]

[a]*Neuralio A.I. P.C., 12th km Thessalonikis—N. Moudanion, 57001 Thermi, Greece*
[b]*AlongRoute, 15 km N.R. Thessaloniki-Moudania, 57001 Thermi, Greece*

Abstract

In this work, we utilized the high-performance computing (HPC) capabilities of the Vienna Scientific Cluster (VSC5) to develop and validate advanced AI models for oceanographic forecasting, with a focus on predicting Significant Wave Height (SWH). Using the computational power of VSC5, particularly its NVIDIA A100 GPUs, allowed us to process and analyze over 500 million data points. The most promising model, a Graph Neural Networks - Gated Recurrent Unit GNN-GRU hybrid, was trained to generate six hourly forecasts for SWH and achieved a Mean Absolute Error (MAE) of 0.0071 and an R-squared (R^2) value of 0.98 against test data, demonstrating high accuracy and efficiency. The first validation results outside the training period are promising and efforts to refine the model are on-going. This work highlights the importance of HPC to the advancement of oceanographic forecasting, by enabling the processing of extremely large datasets and the creation of models that are up to the demanding standards of the marine sector. Subsequent efforts will center around enhancing these models, mitigating detected biases, and creating an operational deployment framework that can facilitate prompt decision-making in maritime operations.

Peer-review under responsibility of the scientific committee of the Proceedings of the First EuroHPC user day
Keywords: Significant Wave Height; HPC; GNN-GRU; CMEMS; CDS; EuroCC

1. Introduction

Almost 3% of greenhouse gas emissions worldwide are attributable to maritime transportation, which is expected to quadruple in less than 30 years. By 2050, the International Maritime Organization (IMO) [1] wants to cut these emissions by 70%. Due to inaccurate weather forecasts, weather routing —a method of reducing fuel use and

* Corresponding author.
E-mail address: vvourlioti@neuralio.ai

1877-0509 © 2024 The Authors. Published by ELSEVIER B.V.

Peer-review under responsibility of the scientific committee of the Proceedings of the First EuroHPC user day

emissions— has only succeeded in achieving 3–10% savings. The chaotic nature of physical processes in weather and climate systems, combined with the immense computational demands of traditional numerical forecast models, has made it challenging to improve the accuracy and resolution of these models. Despite significant research efforts over the past 50 years, these models still struggle to provide highly precise forecasts due to the inherent unpredictability of the systems they are trying to model and the limitations of current computing power.

With the advancements in Artificial Intelligence (AI) methods and the services of Copernicus Marine Service [2] and Copernicus Data Store (CDS) [3], the time has come to develop new research and development processes, close to the private sector, which requires reliable products. The Copernicus Marine Environment Monitoring Service (CMEMS) is a crucial component of the EU Copernicus program, providing regular, systematic information on the physical, biogeochemical, and sea-ice conditions of the global ocean and European regional seas. CMEMS serves over 15,000 users and supports various applications, ranging from maritime safety to climate forecasting. The service's mission includes delivering short-term marine forecasts, analyzing past and present marine conditions, and providing data for climate change monitoring [4].

It is apparent that there are on-going efforts of the research community to develop models to predict oceanographic parameters and an effort on how to combine different architectures and the wealth of data. The target variables, vital for ship routing, include the Significant Wave Height (SWH). Ocean waves with a high SWH may submerge ships and destroy ocean or coastal infrastructure [5]. It endangers human life, agriculture output, and the viability of aquaculture goods. As a result, precise forecasting of SWH is critical since it can assist in avoiding social and economic losses. Furthermore, SWH prediction can provide various benefits. For example, improving ship routes based on SWH predictions might help avoid rough seas, decreasing sailing time and fuel costs. Furthermore, SWH prediction can help organize military and amphibious operations [5]. As climate change is expected to influence oceanographic conditions, leading to more frequent and severe storms, the ability to accurately predict SWH will become increasingly important, and further emphasizing the need for ongoing research and innovation in the field.

Due to the challenges in data collection and the limitations of computing power, Significant Wave Height (SWH) predictions were primarily based on empirical or numerical models ([5],[6],[7]). These methodologies do not have leaning ability that leads to a low prediction accuracy [5]. With the advent of machine learning, Bayesian Networks, XGBoost, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) have been utilized for SWH prediction [8]. A common setback of these models was the negligence of the temporal dependencies of the data that made the models sensitive to noise that consequently lowered their reliability [5]. To address this issue, many researchers have investigated with promising results the Recurrent Neural Networks (RNN) and Long Short Term Memory (LSTM) methodologies ([9],[10]).

Following these efforts, a trend to use deep leaning approaches in SWH prediction emerged, taking advantage of the stronger feature extraction capacity compared to the machine learning methodologies. Convolutional Neural Networks (CNN) were found to outperform other deep-leaning approaches with works like [11] that extracted information from Synthetic Aperture Radar (SAR) images to predict SWH. Although CNNs are designed for visual data, promising results with CNNs resulted from a CNN-LSTM framework that used raw ocean images but with setbacks in night and foggy conditions [12]. Another promising and emerging application of CNNs is by correcting the predictions of numerical models [13].

Despite the ongoing efforts, two key challenges are identified that remain unresolved for accurate SWH prediction: effectively capturing the relationships between various input types while learning the complex non-linear mapping and temporal dependencies with SWH data and distinguishing between occasional extreme sea conditions and seasonal variations in SWH, which involves understanding both short-term and long-term patterns [4]. To this end in the works of [4], that utilized a Wavelet Graph Neural Network (WGNN) showed in their experimental results better performance than other models, including some numerical models and other deep-learning methods.

Understanding the computational limitations and the promising results of CNNs our first model was a CNN-based multi-stage model that predicts deltas instead of absolute values with noise in the input data to enhance the robustness. The validation of this model was done by comparing the model’s forecasts with satellite measurements of Significant Wave Height (SWH), against the CMEMS operational forecasts compared to the same satellite measurements. A sample size of 1.914 collocated points within the Mediterranean was collected and the comparison demonstrated a better performance than the operational CMEMS forecasts, especially in the case of larger waves (>1 m) and a better fitting of the regression line at about 7% for AlongRoute's model.

After the promising results of the first model, the team has sought ways to improve the in-house modelling framework to evaluate the product (i.e., operational oceanographic forecasts) with shipping companies. For the business model to work, it was crucial to develop more sophisticated models with:

1. Extensive training on larger amount of input data and crucial variables, and
2. Spatio-temporal characteristics that allow to extend the forecast horizon.

It became apparent that the first requirement could be satisfied with computational power. To this end, starts-up that need access to GPUs and computational resources are benefited by EuroCC (European Competence Center), that is the European network of 33 NCCs (National Competence Centers) national HPC competence centers with the aim of bridging the existing HPC skills gaps and infrastructure while promoting cooperation across Europe [14]. To gain access to the resources, a short project proposal was submitted outlining the amount of computational power needed and the scope of the project to EuroCC Austria. This granted the team with 100.000 CPU-hours or ~4.500 GPU-hours in the Vienna Scientific Cluster (VSC5) system. Utilizing supercomputer capabilities and offering technical support to its users, the Vienna Scientific Cluster (VSC) is an alliance of multiple Austrian universities. VSC-4 and VSC-5, the fastest supercomputers in Austria, are the flagship systems of the VSC family at the moment. Both systems are powering science and research at the leading academic institutions in Austria and fulfill the demand for high computing power in the areas such as physics, chemistry, meteorology, life sciences, and many others [15].

The second requirement was fulfilled by utilizing the most recent successful methodologies utilizing the computational power to process large data sets in HPC and perform training on extended data. To do so, we opted for a deep leaning methodology that is based in GNNs that have shown to be outperforming other deep learning methodologies while the temporal scale being resolved with Recurrent Gated Units, (GNU) a framework utilized in SWH predictions but in combination with a multivariate multi step time series forecasting [16]. This paper presents the usage of the VSC5, to train and validate models that can predict oceanographic parameters in extended temporal resolutions, the challenges faced in deploying the models, first validation results and future work both research-wise and business-wise.

2. Materials and Methods

To prepare a data driven model, our work was based on finding data that are significant and relate to the physics behind wave generation. This was achieved by utilizing reanalysis oceanographic data from the Copernicus Marine store (wave and physics related) as well as reanalysis atmospheric data from the Copernicus Data Store (CDS) [3]. In climate reanalysis, consistent time series of numerous climate variables are produced through the integration of past observations and models. Reanalysis are among the most frequently employed datasets in the geophysical sciences. They offer an extensive record of the observed climate as it has changed over the past few decades, with data collected on 3D grids at sub-daily intervals. This makes them suitable in our objective of training an Artificial Intelligence (AI) model that learns the relationship between the variables of interest and is capable to produce operationally forecasts of the SWH and wave direction given the previous state of the sea. The plethora of data availability and the access to VSC5 allowed as to investigate before the development of our model the most significant variables that did affect SWH. Feature importance algorithms (regression, Random Forest) combined with guidance from the experts, concluded into the set of the variables to train our model. The access to HPC facilitated the processing of all the variables included in the different data sets as well as the selection of the most important ones. This was extremely crucial for our model, considering that in an operational framework, we opt to retrain our model with up-to-date data that matter and enhancing the model accuracy.

2.1. Variable Selection

The data collection mechanisms to train the model were based on the offered mechanisms of CMEMS and CDS. Simple bash scripts were created to download the data for a full year from all three data sources and all their variables. The user must be registered in both CMEMS and CDS to obtain credentials. The data collection begun for the Mediterranean Sea (Fig.1) for two reasons. Firstly, when compared to available data sets, higher spatial resolution is available for the Mediterranean, compared to the global data sets. Secondly, the Mediterranean Sea offers a closed boundary condition system that is easier to simulate compared to other seas. The year of 2014 was selected as was it was marked by several intense storms that generated high waves, especially in the western Mediterranean, complicating naval and commercial navigation.

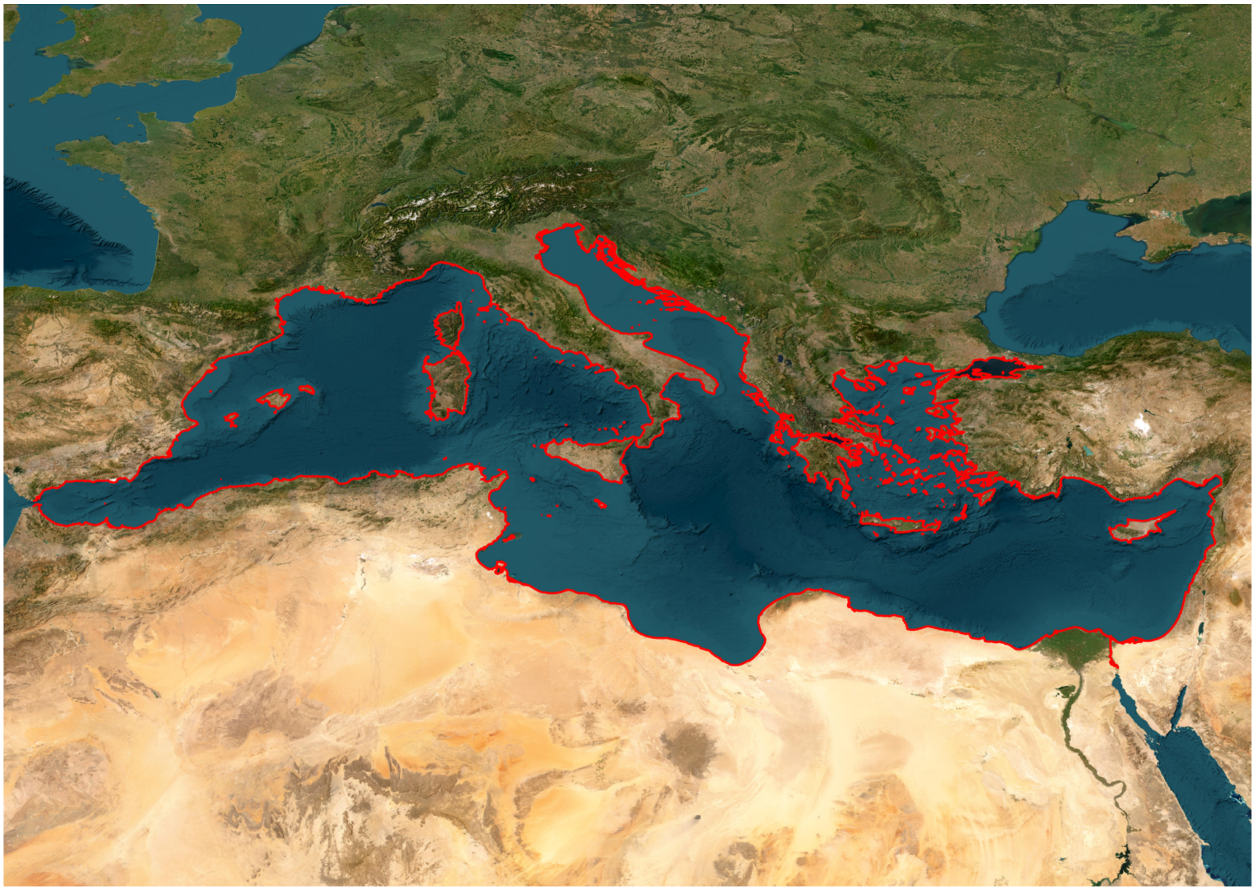

Fig. 1. Area of the Mediterranean Sea, used for training.

In Table 1, the three data sources considered to train our model are presented. The datasets were first carefully selected in terms of providing operational forecasts counterparts, so as to later set-up the model in operational mode, but also to have equivalent information at global scale. Nonetheless, it must be noted that moving outside the Mediterranean Sea comes with losing the higher spatial resolution data sets. As can be seen in Table 1, the data from CMEMS have the same spatial grid and temporal resolution, meaning that the UERRA data had to be re-mapped to the MEDSEA data grid spatially. On the other hand, the MEDSEA data had to also match the 6-hourly temporal resolution of UERRA.

One year of data occupied almost 80G storage in VSC5 and the regression/RF model to finalize the selection of data (see Table 1) was sent in the CPU nodes as it was not computationally intensive to run. This is because temporal and spatial aggregations were applied to the data to check temporal and spatial feature importance. These aggregations reduced the memory burden to handle this amount of data. Finally, 10 variables were selected to train our model.

Table 1. Reanalysis data, their characteristics, and the number of selected variables to train the model.

Reanalysis Data	Spatial Resolution	Temporal Resolution	Number of total variables	Number of selected variables
MEDSEA_MULTIYEAR_WAV_006_012 (CMEMS)	0.042° × 0.042°	Hourly	18	5
MEDSEA_MULTIYEAR_PHY_006_004 (CMEMS)	0.042° × 0.042°	Hourly	10	2
UERRA (CDS)	5.5km x 5.5km	6-hourly	19	3

2.2. Model Development

In this work, we aimed to introduce an extended temporal forecast framework that meets both our scientific interests and the business perspective that demands forecasts for more than a few hours ahead in the future. Graph Neural Networks (GNNs) combined with Gated Recurrent Units (GRUs) offer a powerful approach for significant wave height forecasting. GNNs excel in capturing spatial relationships within complex oceanographic data, while GRUs effectively model temporal dependencies. By integrating these two models, the GNN-GRU framework can simultaneously analyze spatial and temporal patterns in wave data, leading to more accurate and reliable wave height predictions. This hybrid approach is particularly well-suited for handling the dynamic and interconnected nature of ocean environments.

PyTorch [17] was used for building and training the model, with the Adam optimizer managing the learning process and Mean Squared Error (MSE) as the loss function. The data was preprocessed by normalizing it, using MinMaxScaler, and we created sequences of input data to feed into the model. The dataset was split into 60% for training, 20% for validation, and 20% for testing. To fine-tune the model's hyperparameters, including the hidden dimension size, learning rate and batch size, we utilized Optuna [18], an optimization framework that allowed the detection of the best settings for these parameters after 20 trials. Following the training process, we evaluated the model's performance on the test set, by calculating metrics such as mean absolute error, root mean squared error, and R-squared to assess accuracy.

The model was trained on an NVIDIA A100 GPU at the VSC, showing strong performance and efficiency (Table 2). The model was trained on a large dataset of about 559 million data points, completing the process in approximately 13h50min, which reflects the capability of the hardware used. It must be noted that in order to process this amount of data, the A100 GPU with 512GB RAM was necessary, otherwise memory issues rose during run-time. Moreover, the usage of data frames for each data source (three different reanalysis datasets) in python and their merging into one data frame, led to large amounts of memory allocation that in turn crushed our runs. This was overcome not only by using the GPUs with the largest amount of RAM memory but also cleaning from memory all data frames the minute they were no longer necessary in the processing chain. The model achieved a Mean Absolute Error (MAE) of 0.0071, suggesting that its predictions are quite accurate with only a small average error. The R-squared (R^2) value of 0.98 shows that the model effectively captures most of the variability in the data. Additionally, the Root Mean Squared Error (RMSE) of 0.015 further confirms the accuracy of the model's predictions. Overall, the model demonstrates good predictive accuracy and efficiency in training. We trained three separate models, the first one being able to predict 6 hours ahead, the second one 12 hours ahead and the last one 24 hours ahead. The most stable and promising was the model built to forecast 6 hours ahead and our first validation results for this model are shown in the next section.

Table 2. Reanalysis data, their characteristics, and the number of selected variables to train the model.

VSC Partition	Total data points	Training Time	MAE	R^2	RMSE
zen3_0512_a100x2	559.238.400	13h50min	0.0071	0.98	0.015

3. Results

The validation strategy we followed is running our models and making predictions outside of the training period focusing on challenging dates with high waves. The models are initialized with forecasts and analysis datasets to mimic an operational set-up. Then we opted to understand if the seasonality is captured by our models by running our model for a month's period and comparing them to both reanalysis data and satellite SWH data from CMEMS. The reanalysis data, being gridded fields, can help us understand if the model is capturing the general circulation and the satellite data, as external validation, require also an extended period to collect a big enough size of overpassing satellite measurements.

A selection of days with SWH above 1.5 m was first located through filtering the data from the CMEMS with a focus on dates outside the training period. The first validation results are shown here for our model predictions, outside the training period and for 14/01/2019, 00 UTC and compared to reanalysis data. As can be seen in Fig. 2(a), our models can forecast the general wave circulation quite well when compared to the reanalysis data, with an average MAE of 0.34 m. The high waves seen near the coasts of France and cascading southwards are well captured in general by the GNN-GRU (Fig. 2(a)), although some local misalignments are seen west and east of the main event. Although this is a good result, we do have some overestimations by the model that we are seeking to understand. In Fig. 3(b), a scatter plot of model prediction versus reanalysis is given for the examined date, in which a lump is seen in low observed SWH that is overestimated by the model in the area of the Aegean Sea. Our first impression was that our model was not understanding well the coastlines and that it lets the SWH built up as if no land was there. This came apparent when creating a buffer zone around coastlines of around 4km (one grid box with regards to our spatial resolution), and the same plot was created as can be seen in Fig. 3(a) for all the Mediterranean. Although the bulk is gone, some points persist, and it is under investigation in our next steps. Here an extended validation that is undergoing will help us understand the model biases and its shortcomings.

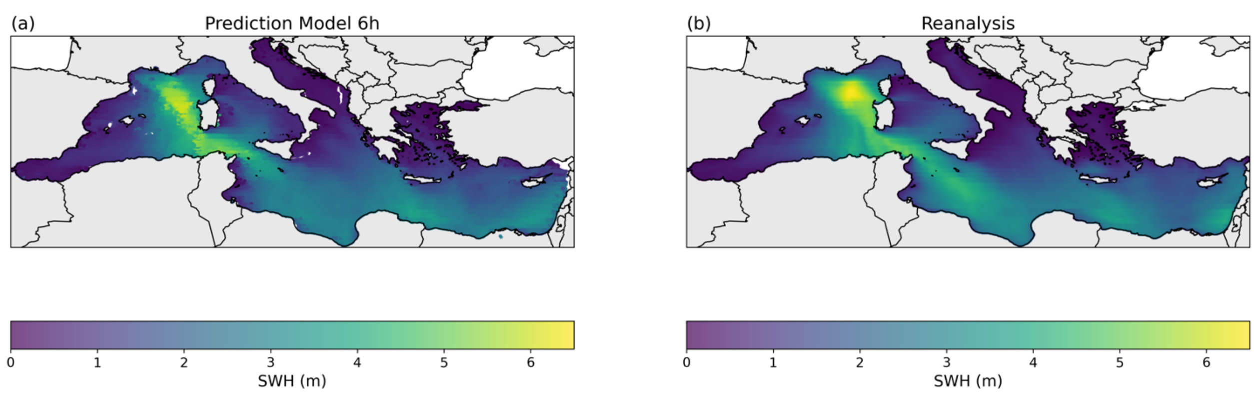

Fig. 2. (a) Model prediction on 14/01/2014; (b) Reanalysis data of SWH for the same time.

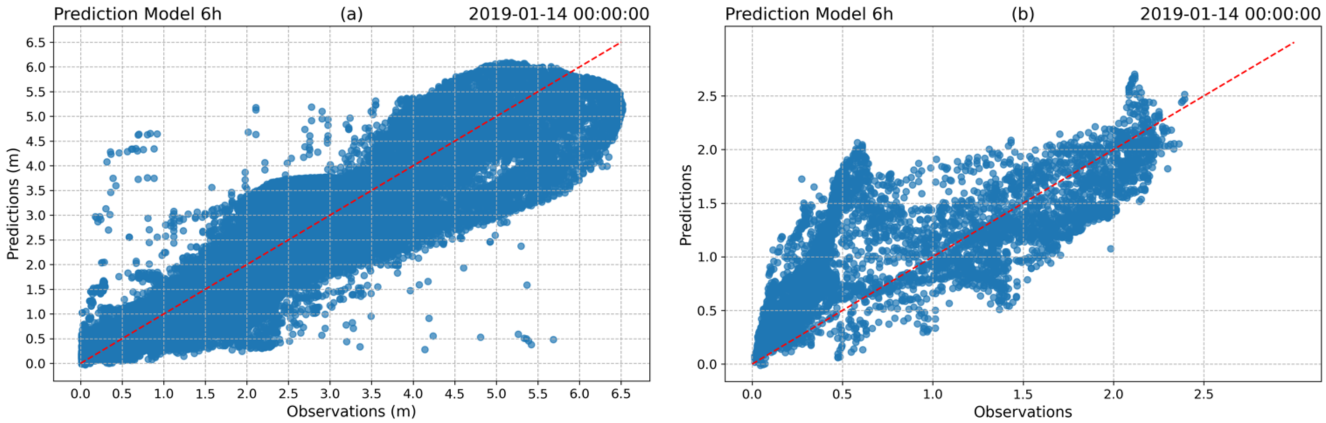

Fig. 3. (a) Scatter plot of predicted versus reanalysis data for Mediterranean (with buffer zone round coastlines); (b) Scatter plot of predicted versus reanalysis data for Aegean (without buffer zone around coastlines).

4. Discussion

Utilizing HCP resources granted though EuroCC, we were able to train and work on validating AI models for oceanographic parameters prediction and built a set up for future operational workflows that will benefit the shipping operations value chain. We have worked on the VSC5 cluster and it's NVIDIA A100 GPUs in which we developed and deployed a GNN-GRU that can handle spatiotemporal data and showed good performance metrics against the test dataset. Given the enormous amount of data processed, the most challenging part was not the usage of the HPC itself but rather handling the pre-processing step of the data in terms of memory. This was achieved by utilizing the GPU with the 512 GB RAM but also by cleaning up the cache and freeing the memory from redundant data frames. The training of each created model took around 14 hours on two GPUs, covering the entirety of the Mediterranean Sea and for one year of data (2014) accumulating to more than 500 million data points.

A significant advantage of having access to these HPC resources was the ability to perform hyperparameter tuning for the network with 20 trials, which consumed the majority of the total training time (about 90%). Without access to such computational power, we would have been forced to skip this critical step, likely resulting in the use of estimated configurations rather than the optimal ones. This would have negatively impacted the accuracy and reliability of our model's predictions. Thus, the HPC resources were instrumental in ensuring that our model was as well-tuned as possible, ultimately leading to better performance in predicting oceanographic parameters.

The validation of the model begun with checking the ability of the model to understand the general circulation in days with challenging sea conditions, with our model scoring good average MAE but showing large overestimations of small observed SWH as well as underestimations of observed SWH between 2.5m and 4.5m. This was attributed to the challenges of small islands and coastlines that our model seemed to miss and was overcome by creating a buffer zone around the coastlines. Some points did persist even after the buffer zone and are under investigation. Future work includes validating our model against satellite and reanalysis data for extended period to understand the strength and weakliness and calibrate the model towards optimal performance.

The access to HPC allows us to train the model with even more data, given the increase of data volumes increases the accuracy of our model or as an alternative to use simple bias correction in the output of the model. Once we secure the best model configuration, the next step is to provide these data in inference mode, in our premises, for the interested shipping companies.

Acknowledgements

This work got support from the Austrian National Competence Centre for High-Performance Computing, High-Performance Data Analytics and Artificial Intelligence (EuroCC Austria). It is legally represented by Advanced Computing Austria ACA GmbH and funded by the EuroCC 2 project that has received funding from the European High-Performance Computing Joint Undertaking (EuroHPC JU) and participating countries under grant agreement No. 101101903. The computational results presented have been achieved [in part] using the Vienna Scientific Cluster (VSC).

References

[1] International Maritime Organization (2023) Strategy on reduction of GHG emissions from ships. Report

[2] Von Schuckmann, K., Le Traon, P. Y., Smith, N., Pascual, A., Brasseur, P., Fennel, K. S., et al. (2018). Copernicus marine service ocean state report. J. Oper. Oceanograp. 11, 1–142. doi: 10.1080/1755876X.2018.1489208

[3] *CDS*. Climate Copernicus. Retrieved August 28, 2024, from https://cds.climate.copernicus.eu/cdsapp#!/home.

[4] Le Traon, P. Y., Reppucci, A., Alvarez Fanjul, E., Aouf, L., Behrens, A., Belmonte, M., ... & Zacharioudaki, A. (2019). From observation to information and users: The Copernicus Marine Service perspective. Frontiers in Marine Science, 6, 234.

[5] Chen, D., Liu, F., Zhang, Z., Lu, X., & Li, Z. (2021, July). Significant wave height prediction based on wavelet graph neural network. In 2021 IEEE 4th international conference on big data and artificial intelligence (BDAI) (pp. 80-85). IEEE.

[6] T. W. Group, "The wam model—a third generation ocean wave pre- diction model," Journal of Physical Oceanography, vol. 18, no.12, pp. 1775 – 1810, 1988.

[7] L. Mentaschi, G. Besio, F. Cassola, and A. Mazzino, "Performance evaluation of wavewatch iii in the mediterranean sea," Ocean Modelling, 06 2015.

[8] J. Mahjoobi and E. Adeli Mosabbeb, "Prediction of significant wave height using regressive support vector machines," Ocean Engineering, vol. 36, no. 5, pp. 339–347, 2009.

[9] S. Mandal and N. Prabaharan, "Ocean wave forecasting using recurrent neural networks," Ocean Engineering, vol. 33, no. 10, pp. 1401–1410, 2006.

[10] K. Osawa, H. Yamaguchi, M. Umair, M. A. Hashmani, and K. Horio, "Wave height and peak wave period prediction using recurrent neural networks," in 2020 International Conference on Computational Intelli- gence (ICCI), 2020, pp. 1–4.

[11] B. Quach, Y. Glaser, J. E. Stopa, A. A. Mouche, and P. Sadowski, "Deep learning for predicting significant wave height from synthetic aperture radar," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 3, pp. 1859–1867, 2021.

[12] H. Choi, M. Park, G. Son, J. Jeong, J. Park, K. Mo, and P. Kang, "Real- time significant wave height estimation from raw oceanimages based on 2d and 3d deep neural networks," Ocean Engineering, vol. 201, p. 107129, 2020.

[13] J. Mooneyham, S. C. Crosby, N. Kumar, and B. Hutchinson, "Swrl net: A spectral, residual deep learning model for improving short-term wave forecasts," Weather and Forecasting, vol. 35, no. 6, pp. 2445 – 2460, 2020.

[14] Marchant, D. G., & Jensen, T. L. EuroCC WP 8.2/8.3/8.5 DK Survey.

[15] Austrian initiative on high performance computing*. Vienna Scientific Center. Retrieved August 25, 2024, from https://www.vsc.ac.at/home/*

[16] Lawal, Z. K., Yassin, H., Teck Ching Lai, D., & Che Idris, A. (2024). Understanding the Dynamics of Ocean Wave-Current Interactions Through Multivariate Multi-Step Time Series Forecasting. Applied Artificial Intelligence, 38(1), 2393978.

[17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32

[18] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019, July). Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2623-2631).