Journal of Ocean Engineering and Technology 38(6), 426-437, December, 2024 eISSN 2287-6715
https://doi.org/10.26748/KSOE.2024.065

Original Research Article

Study for Filling Missing Wave Data in Geomundo Ocean Buoy Using Artificial Neural Networks

Seongyun Shin[1], Seonghyun Park[1], Kwang Hyo Jung[2] and Sung Boo Park[3]

[1]Graduate student, Department of Naval Architecture and Ocean Engineering, Pusan National University, Busan, Korea
[2]Professor, Department of Naval Architecture and Ocean Engineering, Pusan National University, Busan, Korea
[3]Senior Offshore Floater Engineer, Ship and Offshore Performance Research Center, Samsung Heavy Industries, Co. Ltd, Daejeon, Korea

KEYWORDS: Artificial neural network, Machine learning, Missing, Significant wave height, Peak wave period, Wave direction

ABSTRACT: *This study aimed to propose an Artificial neural network (ANN) model to fill missing wave data using Bayesian optimization of hyperparameters. Ocean environmental data obtained by ocean buoys have been missed due to the malfunction or maintenance of monitoring system or extremely harsh weather condition during a storm. It is important of the continuity of measured data to analyze ocean environmental condition for the engineering purpose such as the design condition for offshore structure and the assessment of wave condition for a long term return period using the extreme analysis. Five ANN models were applied to estimate three wave parameters of significant wave height, peak wave period, and wave direction using of measurement data at Geomundo ocean buoy for eight years (2010-2017). The wind data of European Centre for Medium-Range Weather Forecasts were employed to estimate the wave parameters with ANN models to fill missing wave data at Geomundo ocean buoy. By comparison of each ANN model result, it could be suggested Bidirectional gated recurrent unit network, Gated recurrent unit network, Feed-forward neural network for the best model to fill the significant wave height, peak wave period and wave direction, respectively. These three ANN models could be applied to fill a long-term missing wave data at ocean buoys.*

1. Introduction

For coastal and offshore structures and vessels directly affected by metocean conditions, the effects of these conditions should be incorporated into their design to minimize human and material damage during installation, operation, and decommissioning. Given that ocean waves constitute a major component of the dynamic load that affects structures, it is indispensable to obtain and analyze ocean wave data. For instance, the American Petroleum Institute (API) recommends that ocean wave, wind, and ocean currents corresponding to a 100-year return period should be reflected in design criteria for permanent mooring systems for the station keeping of offshore structures with a service life of ≥20 years (API, 2005).

Ocean wave data collection methods include in situ observations, satellite remote sensing, simulation based on numerical modeling, or reanalysis. Reanalysis data are derived based on a process known as data assimilation, which combines historical observations and the numerical model. One of the advantages of reanalysis data is that it can provide global coverage with consistent spatiotemporal resolution. The accuracy of the data relies on the performance of the numerical model and the availability of observational data; however, its accuracy may decrease if observational data for a particular region and time are not available. Although in situ observations provide highly accurate ocean wave data, data may be missing because of damage to physical components, such as sensors and batteries in observation equipment; network failures; relocation due to extreme ocean environments; and equipment maintenance and replacement. More specifically, the issue of missing data because of extreme ocean environments is regarded as missing not at random (MNAR) as the value of the data is not unrelated to the cause of the missing data, and therefore, the analysis that simply removes missing data may contain bias in its results (Kim et al., 2021). Therefore, it is important to apply the appropriate method depending on the type and nature of the missing data, the availability of data that can be utilized for handling the missing data, and the purpose of the handled missing data.

Kim et al. (2021) investigated the trends in time series techniques to handle the missing data by categorizing the techniques as statistical technique, matrix-based technique, regression-based technique,

Received 26 July 2024, revised 7 October 2024, accepted 8 October 2024
Corresponding author Kwang Hyo Jung: +82-51-510-2343, kjung@pusan.ac.kr

recurrent neural network (RNN)-based technique, and generative adversarial network (GAN)-based technique. The authors identified that the performance of artificial neural network (ANN) models was good in most cases. RNNs and GANs are both machine learning models and ANNs. Machine learning is a method that enables models to learn patterns from given data without explicit programming and perform tasks such as prediction, classification, and clustering. Meanwhile, ANNs are machine learning models developed to emulate the human brain. Furthermore, Kim et al. (2021) noted that the imputation method based on the neural network model is time-consuming for training and processing. ANNs have been proven useful in computer vision systems for self-driving and large language models for Web search and translation; however, the issue of consuming a significant amount of time is expected to be gradually resolved with continued advances in hardware, such as graphic processing unit (GPU) and tensor processing unit (TPU), as well as software constituting the machine learning pipeline.

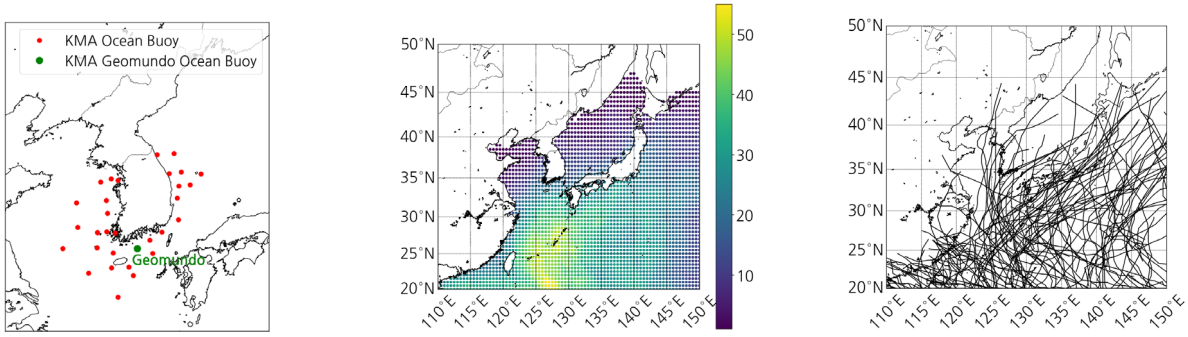
Studies have been conducted on ANN models for forecasting the significant wave height of ocean buoys (Minuzzi and Farina, 2023; Pang and Dong, 2023; Alizadeh and Nourani; 2024) and for predicting significant wave height, average wave period, and wave direction (Wei, 2021). Long-short term memory (LSTM) and gated recurrent unit (GRU) are often used as the basic structure of ANN models. Nonetheless, forecast accuracy decreases with an increase in the forecast lead time, which represents a time gap between the input and output of the forecast model, making it difficult to use the models to fill in long-term missing data. Based on machine learning, Park et al. (2021) studied a model to predict significant wave height and proposed a time series length appropriate for selecting input variables and predicting significant wave height through the Pearson correlation analysis. Bujak et al. (2023) suggested a feed-forward neural network (FNN) model that utilized data from the Integrated Surface Database (ISD) of the National Oceanic and Atmospheric Administration (NOAA) as the input to fill in missing significant wave height in wave buoys, and compared the model to the Mediterranean sea waves hindcast (Korres et al., 2019). Furthermore, Guijo-Robio et al. (2023) proposed a complementary method based on an Evolutionary ANN (EANN), which applied the transfer function and the evolutionary algorithm, to fill in the missing significant wave height data in multiple ocean buoys adjacent to specific sea areas (Gulf of Alaska, Northeast Coast) and compared their model with the bidirectional recurrent imputation for time series (BRITS) model (Cao et al., 2018) and the self-attention-based imputation for time series (SAITS) model (Du et al., 2023). Generally, models based on the ISD and adjacent ocean buoy data are not applicable to missing significant wave height in ocean buoys installed in multiple oceans, because the availability of their input data relies on the location of the missing significant wave height.

This study proposes an ANN model to fill in missing ocean wave data in an ocean buoy. To train, validate, and test the model, this study collected ocean wave data in a Geomundo ocean buoy from the Open MET Data Portal of the Korea Meteorological Administration (KMA) and wind data from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5). With the collected wind data as the input, this study developed the models to estimate the significant wave height, peak wave period, and wave direction. The ANN models applied in this study included FNN, Long-Short Term Memory Network (LSTMN), Bidirectional LSTMN (Bi-LSTMN), Gated Recurrent Unit Network (GRUN), and Bidirectional GRUN (Bi-GRUN). By employing the Bayesian optimization technique, this study identified the combination of hyperparameters with the minimum mean absolute error (MAE) for the validation set for each model. Based on the test set, the models whose training and validation process were completed were compared for performance to fill in the missing data for significant wave height, peak wave period, and wave direction.

2. Data Collection and Analysis

2.1 Ocean Buoy Data from the Korea Meteorological Administration

It is possible to collect ocean buoy data (Fig. 1(a)) in the waters around the Korean Peninsula from the Korea Meteorological Administration's Open MET Data Portal. The Korea Meteorological Administration's ocean buoy data provides 12 types of data on an hourly basis: wind speed, wind direction, gust speed, local air pressure, humidity, air temperature, water temperature, maximum wave height, significant wave height, mean wave height, wave period, and wave direction. This study utilized the Geomundo ocean buoy data, and the



(a) Locations of KMA ocean buoys (b) Frequency of typhoons (c) Tracks of typhoons

Fig. 1 Locations of KMA ocean buoys and statistics of typhoons from 2010–2017

wave period of the buoy provided the peak wave period. The data regarding significant wave height, peak wave period, and wave direction from 2010 to 2017 were collected and utilized as targets for the ANN models. Meanwhile, the data on wind speed and direction were collected and compared to the wind data from ECMWF ERA5 (Fig. 4). The Geomundo ocean buoy data were derived from a disk-shaped buoy located approximately 85 km south of Yeosu (latitude 34.0014° and longitude 127.5014°). The water depth at the location is approximately 80m. Figs. 1 (b) and (c) illustrate the frequency and tracks of typhoons from 2010 to 2017, suggesting the frequent inclusion of the Geomundo ocean buoy in the South Sea within the radius of strong winds. Table 1 presents the percentage of missing data by year. When one of the features among significant wave height, peak wave period, or wave direction was missing, the time point was regarded as missing. In 2012, there were relatively more missing data (Fig. 2(a)), and significant amounts of data were missing from July to September, when typhoons affecting the Korean Peninsula are frequent (Fig. 2(b)). The longest consecutive period of missing data was approximately 39 days (940 hours). Except for 2012, less than 3.1% of the data was missing.

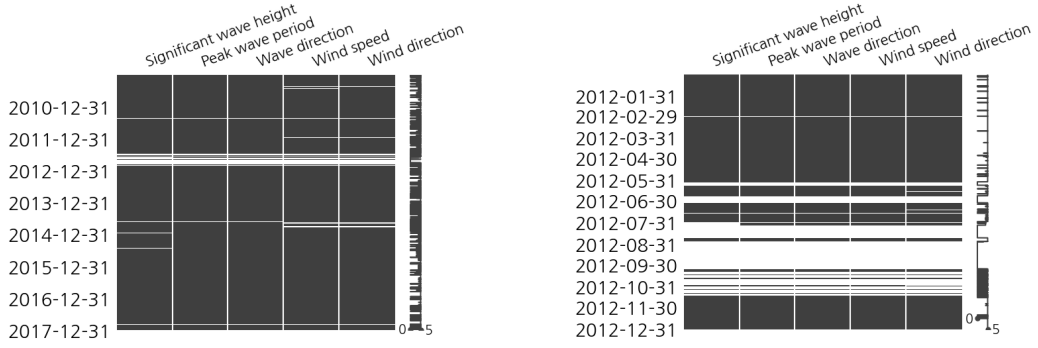
Table 2 lists the sea state codes designated by the World Meteorological Organization (WMO, 2019), categorizes time points by the range of significant wave height per code, and presents the most probable peak wave period by each of the categorized sea state codes

Table 1 Missing rate

Year	No. of raw data	No. of missing data	% Missing
2010	8,760	72	0.82
2011	8,760	166	1.89
2012	8,784	2,526	28.76
2013	8,760	247	2.82
2014	8,760	269	3.07
2015	8,760	211	2.41
2016	8,784	67	0.76
2017	8,760	55	0.63
Total	70,128	3,613	-

(Bales et al., 1981). Based on Eq. (1) (DNV, 2014), the estimated value of the peak shape parameter (γ) of the Joint North Sea Wave Project (JONSWAP) spectrum was calculated with the median significant wave height and the most probable peak wave period.

$$\begin{aligned} &\text{i)}\ \gamma = 5 && \text{for} \quad \frac{T_p}{\sqrt{H_s}} \leqq 3.6 \\ &\text{ii)}\ \gamma = \exp\left(5.75 - 1.15\frac{T_p}{\sqrt{H_s}}\right) && \text{for} \quad 3.6 \leq \frac{T_p}{\sqrt{H_s}} < 5 \\ &\text{iii)}\ \gamma = 1 && \text{for} \quad 5 \leq \frac{T_p}{\sqrt{H_s}} \end{aligned} \tag{1}$$



(a) 2010–2017 (b) 2012

Fig. 2 Missing value visualization of KMA Geomundo ocean buoy data

Table 2 Wave parameter categorization by WMO Sea State

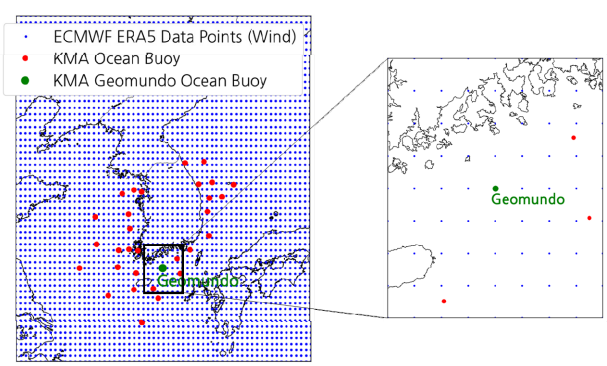
Sea state	Significant wave height (m)		Peak wave period (s)	Estimated Peak shape parameter by Eq. (5)	Characteristics	Sample size	
	Range	Median	Most probable			Number	%
0	0.000	0.000	-	-	Calm (glassy)	0	0.00
1	0.000–0.100	0.050	-	-	Calm (rippled)	0	0.00
2	0.100–0.500	0.300	8.0	1	Smooth (wavelets)	11533	17.34
3	0.500–1.250	0.875	6.4	1	Slight	33945	51.03
4	1.250–2.500	1.875	6.4	1.455	Moderate	19433	29.22
5	2.500–4.000	3.250	7.1	3.390	Rough	1372	2.06
6	4.000–6.000	5.000	10.7	1.280	Very rough	175	0.26
7	6.000–9.000	7.500	12.8	1.455	High	46	0.07
8	9.000–14.000	11.500	12.8	4.093	Very high	11	0.02
9	Over 14.000	-	-	-	Phenomenal	0	0.00

In Eq. (1), T_p represents the peak wave period and H_s represents the significant wave height. The peak shape parameter, which is also referred to as the peak enhancement factor, indicates a higher value when ocean waves have not fully grown and ocean wave energy is clustered around the peak wave period (Cho et al., 2020). The number of the data in the sample was in the order of Sea State Codes 3, 4, and 2, and all other codes represented less than 3% of the total sample.

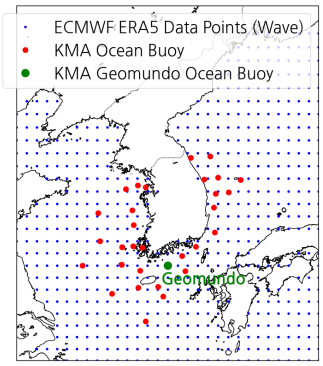
2.2 ECMWF Reanalysis Data

ERA5 is the reanalysis data that assimilate and integrate measurement data into the CY41R2 version of the ECMWF's Integrated forecasting system (IFS) (ECMWF, 2016) (Hersbach et al., 2020). This study collected wind data from ECMWF ERA5 and utilized it as the input for the ANN model. Additionally, it collected and compared the reanalysis ocean wave data from ERA5 with ocean wave data from ocean buoy and ocean wave data estimated from the ANN model in Chapter 4.

For the reanalysis data, ERA5 uses satellite and in situ observations as the input. The wind data measured by in situ observational instruments, such as drifting buoys and moored buoys, are included in the input data of ERA5, whereas the in situ observational ocean wave data are not included. ERA5 is provided in a temporal resolution of 1 hour intervals and a spatial resolution of latitude and longitude intervals of 0.25° and 0.5° for wind and ocean wave data, respectively (Figs. 3 (a), (b)). ERA5 10 m wind (10 m above the surface), significant wave height, peak wave period, and mean wave direction data were collected for eight years (2010–2017) at the grid points closest to the Geomundo ocean buoy. It corresponds to a location

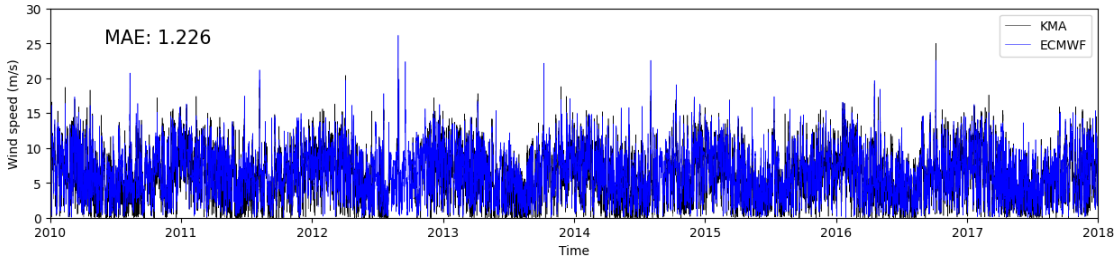


(a) Locations of ECMWF and KMA wind data

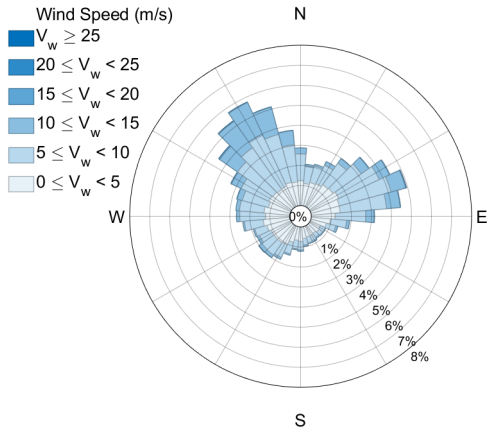


(b) Locations of ECMWF and KMA wave data

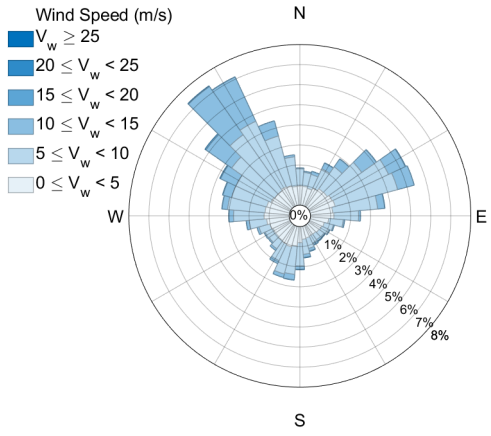
Fig. 3 Locations of ECMWF ERA5 data points and KMA ocean buoys



(a) Time series



(b) Rose diagram from KMA wind



(c) Rose diagram from ECMWF wind

Fig. 4 Comparison of wind data from of the Geomundo ocean buoy and ECMWF ERA5 during 2010–2017

approximately 202 m southwest (219°) (latitude 34.0000°, longitude 127.5000°) from the location of the Geomundo ocean buoy (latitude 34.0014°, longitude 127.5014°) (Fig. 3(a)). Wind data from the Geomundo ocean buoy and ECMWF ERA5 are represented as time series (Fig. 4(a)) and wind rose (Figs. 4 (b), (c)). The wind rose is based on the north (0°) and represents the direction from which the wind blows. The most dominant wind direction was similar with 330° from the Geomundo ocean buoy and 320° from ECMWF ERA5, and when the northwest wind range (270°–360°) was excluded, the most dominant wind direction was the same with 70° from the Geomundo ocean buoy and ECMWF ERA5.

3. Design of ANN Models and Machine Learning

3.1 Data Preprocessing

This study split eight years of data (2010–2017) to train, validate, and test the ANN models. The data for a six-year period (2010–2015) were utilized as the training set. The data for 2016 were employed as the validation set to explore the optimal combination of hyperparameters for each ANN model and to determine the endpoint of training. To compare the performance of the models after the training and validation process, this study utilized the 2017 data as the test set (Fig. 5).

To enable the ANN models to learn the periodic features of the arc degree data (wind direction and wave direction) with the range of 0°–360°, this study applied Eqs. (2) and (3) to convert the direction data in the polar coordinate system into x and y components of the two Cartesian coordinate systems and applied them as the input and output.

$$Direction_x = \cos\left(\frac{Direction_{\deg}}{180}\times\pi\right) \tag{2}$$

$$Direction_y = \sin\left(\frac{Direction_{\deg}}{180}\times\pi\right) \tag{3}$$

In Eqs. (2) and (3), $Direction_{\deg}$ represents the arc degree data, and $Direction_x$ in Eq. (2) and $Direction_y$ in Eq. (3) represent x and y components of the arc degree data, respectively. To prevent the issue of slow learning due to the difference in scale between the input data features of the ANNs, this study standardized the input data of all datasets (training, validation, and test sets) based on Eq. (4).

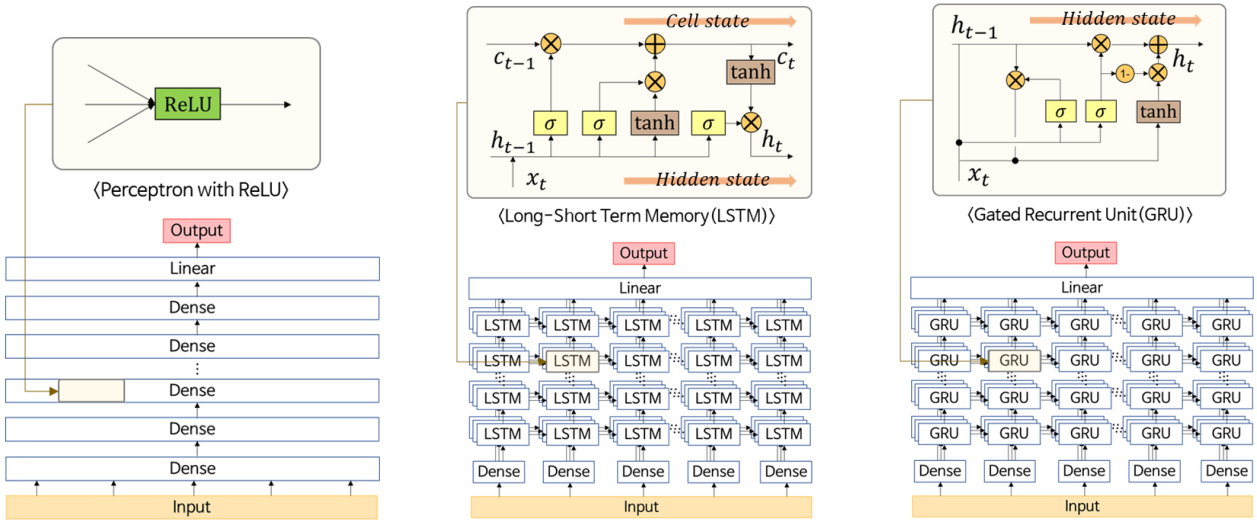
$$x_{i,standardized} = \frac{x_i - \mu_{train}}{\sigma_{train}} \tag{4}$$

In Eq. (4), x_i implies the input data, $x_{i,standardized}$ represents the standardized input data, and μ_{train} and σ_{train} refer to the mean and standard deviation of the training set input data, respectively.

3.2 Artificial Neural Network Models

This study compared five ANN models (FNN, LSTM, GRUN, Bi-LSTM, Bi-GRUN, and Bi-LSTM) to fill in the missing ocean wave data from the ocean buoy. Fig. 6 (a) illustrates the structure of the FNN model. The FNN is the most basic type of ANN and consists of the input layer, hidden layer, and output layer, and a node in each layer is fully connected to the node in the previous layer. Therefore, it is also called the dense layer. The rectified linear unit (ReLU) was applied as

Fig. 5 Data splitting



(a) FNN (b) LSTMN (c) GRUN

Fig. 6 ANN models

the activation function for the dense layer. Time series data input to the FNN is forward propagated to the output layer with values from time points entered in the input layer all at once. Figs. 6 (b) and (c) depict the structure of the LSTMN and GRUN models. The ANN model other than the FNN is the RNN, which receives time series data sequentially as the input and generates output through recurrent connections by considering the input at each time point along with information from the previous time point. In LSTMN and GRUN, the input time series data is sequentially entered and propagated in chronological order and linearly combined in the linear layer to produce the output. In Bi-LSTMN and Bi-GRUN, the time series of the input data is propagated once in chronological order and once in reverse chronological order, and linearly combined in the linear layer. The LSTM was developed to address the issue of long-term dependencies, in which the structure of the Vanilla RNN cannot handle patterns in a long time series. Although it can retain information even in long time series data by controlling the cell state with the input gate, output gate, and forget gate, its disadvantages include a complex structure and high computational cost (Hochreiter and Schmidhuber, 1997). The GRU is a model that integrates the input gate and forget gate of the LSTM and simplifies it into the update gate and reset gate (Chung et al., 2014). Compared to the LSTM, the GRU requires less time for computation. Additionally, if the patterns of the time series data are simple, the GRU may generally outperform the LSTM.

3.3 Model Training and Validation

The initial ANN model, which has not been trained, has a few initialized weights and biases. During training, the ANN model forward propagates the input of the training set, obtains the output value, inserts it into the loss function along with the target value, and thereafter, obtains the loss. The loss is backpropagated to update weights and biases, which train the ANN model to achieve the output that minimizes the loss for the training set. This study evaluated performance by calculating the loss in the validation set at the end of each epoch, in which a single training is run on all the data in the training set. The early termination technique was applied, wherein training was terminated early if the loss did not decrease during 10 validations. For the loss function, MAE was applied (Eq. (5)).

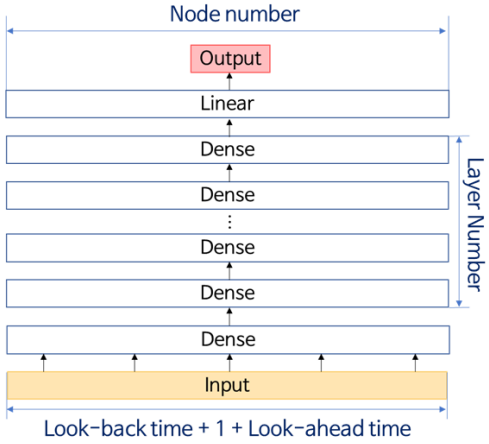
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y_i}| \tag{5}$$

In Eq. (5), y_i is the target value, $\hat{y_i}$ the output value of the model, and n the number of data. Adaptive moment estimation (Adam) was applied as the optimizer, which updates weights and biases during backpropagation. Momentum was applied to reduce the probability of convergence into a local minimum, induce faster convergence in the loss function, and dynamically adjust the learning rate to attain the optimum. This study set Adam's learning rate to 0.001.

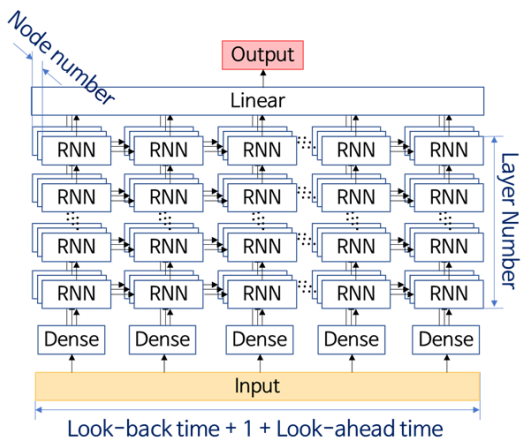
3.4 Hyperparameters and Hyperparameter Optimization

Various hyperparameters were applied for each of the ANN models. Table 3 lists the types and ranges of hyperparameters for each of the ANN models. Look-back time and look-ahead time represent the temporal length of the input time series in the ANN model (Fig. 7). Look-back time represents how far back in the past the first time point in the input time series is from the target point, whereas look-ahead time indicates how far forward in the future the last point in the input time series is from the target point. The range of look-back time and look-ahead time was from a minimum of 0 hours to a maximum of 168 hours (7 days) with the interval set to a constant log scale from 0 hours to 24 hours (1 day) and to 24 hours (1 day) from 24 hours (1 day) to 168 hours (7 days). The batch size was set to a constant logarithmic interval from 2^5 to 2^{11}. Dropout reduces overfitting by randomly removing connections between nodes in every batch. The dropout rate was set to 0.02 intervals from 0 to 0.5. The node number and layer number are the hyperparameters that affect the capacitance of the ANN model (Fig. 7). The layer number implies the number of hidden layers. If it is too high, it may cause gradient loss during backpropagation, and the gradient is not transferred to layers near the input layer, which may slow down convergence in the model.

This study applied Bayesian optimization for hyperparameter



(a) FNN



(b) RNN based model (LSTMN, Bi-LSTMN, GRUN, Bi-GRUN)

Fig. 7 Model hyperparameter diagram

Table 3 Model hyperparameters

Hyperparameter	Range	
	FNN	LSTMN, Bi-LSTMN, GRUN, Bi-GRUN
Node number	2^[0:1:12][1)]	2^[0:1:7]
Layer number	[1:1:25]	[1:1:8]
Look-back time (hour)	[0, 1, 2, 4, 8, 12, 24, 48, 72, 96, 120, 144, 168]	
Look-ahead time (hour)	[0, 1, 2, 4, 8, 12, 24, 48, 72, 96, 120, 144, 168]	
Batch size	2^[5:1:11]	
Dropout rate	[0:0.02:0.5]	

[1)]In [a:b:c], a: Minimum value, b: Interval, c: Maximum value

optimization (HPO) to identify the best performing combination of hyperparameters for each of the ANN models. This technique estimates a set of hyperparameters and a function of losses based on the surrogate model, and applies the acquisition function to select the combination of hyperparameters to be employed in the succeeding trial. It can reduce the number of trials more efficiently than grid search, which trains and validates every combination of hyperparameters. With the Gaussian process and expected improvement for the surrogate model and acquisition function, this study conducted 200 trials for each of the ANN models.

4. Results

Five ANN models (FNN, LSTMN, Bi-LSTMN, GRU, and Bi-GRUN) were applied to three ocean wave parameters (significant wave height, peak wave period, and wave direction) in ocean wave data from ocean buoy, and this study developed a total of 15 optimized ANN models through the training and validation process. During the validation process, Bayesian optimization was applied to identify the combination of six hyperparameters (look-back time, look-ahead time, batch size, dropout size, node number, and layer number), which minimize the MAE for the validation set (Validation MAE). Three wave parameters (significant wave height, peak wave period, and wave direction) were estimated for the dataset period (2010–2017) with 15 ANN models optimized with the combination of hyperparameters (Table 4), which minimize the Validation MAE, to fill in the missing data. To evaluate the performance of the optimized ANN model, this study applied the MAE for the ocean wave data from Geomundo ocean buoy (Test MAE) in the test set period (2017) as an indicator of performance evaluation, and compared the root mean squared error (RMSE) (Test RMSE) (Eq. (6)) if the Test MAE was equal.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad (6)$$

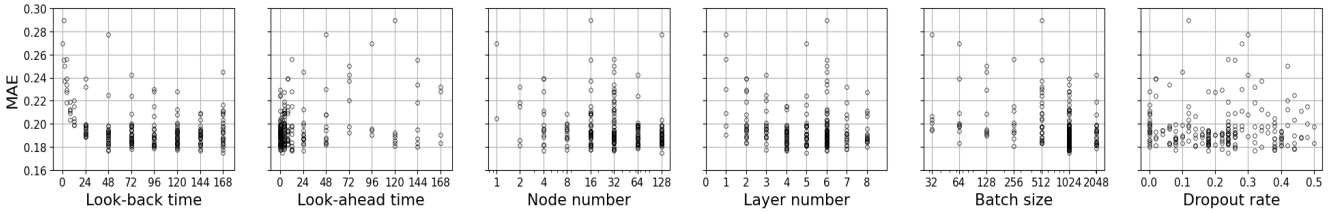
In Eq. (6), y_i is the target value, $\hat{y_i}$ the estimated value of the model, and n the number of data. Table 5 presents the Test MAE and Test RMSE of the optimized ANN model and the reanalysis ocean wave data of ECMWF ERA5. Bi-GRUN, GRUN, and FNN were the ANN models that outperformed in filling in missing data for the significant wave height, peak wave period, and wave direction.

4.1 Hyperparameter Optimization

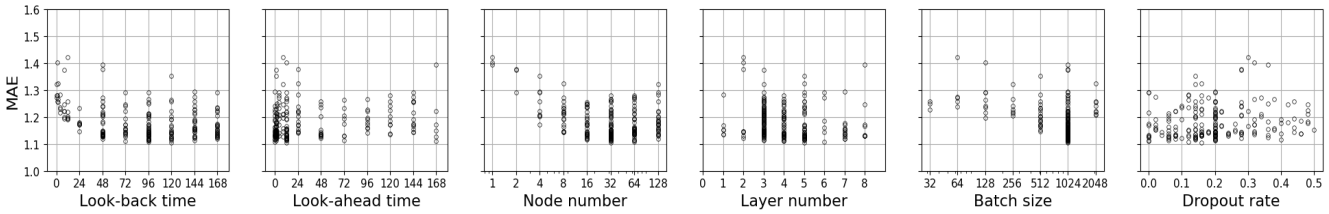
Figs. 8 (a), (b), and (c) illustrate the Validation MAE per hyperparameter for 200 hyperparameter combinations selected during the HPO process for the ANN models with the best Test MAE (significant wave height: Bi-GRUN, peak wave period: GRUN, and wave direction: FNN) (Table 5). The ANN models not depicted in Fig. 8 exhibited a similar trend to the ANN models with the best Test MAE for each type of output in terms of the Validation MAE for look-back time, look-ahead time, and batch size. As look-back time increased from 0 to 48 hours in Figs. 8 (a) and (c), the validation MAE dropped markedly. As look-back time increased from 0 to 96 hours, especially in Fig. 8(b), the validation MAE fell. In terms of look-ahead time, Figs. 8 (a), (b), and (c) all showed the minimum Validation MAE around 0 hours, which confirmed that wind data after the missing time point had a negligible impact on the improved performance of the ocean wave estimation ANN model. In the optimal combination of hyperparameters (Table 4), the dropout rate is distributed in the range of 0.08–0.24 for the significant wave height estimation ANN model, 0.06–0.46 for the peak wave period estimation ANN model, and 0.06–0.50 for the wave direction estimation ANN model. This confirmed a negligible impact on the improved performance of the ANN model, as no explicit trend in the Validation MAE was identified across the overall range of the set hyperparameters.

4.2 Significant Wave Height

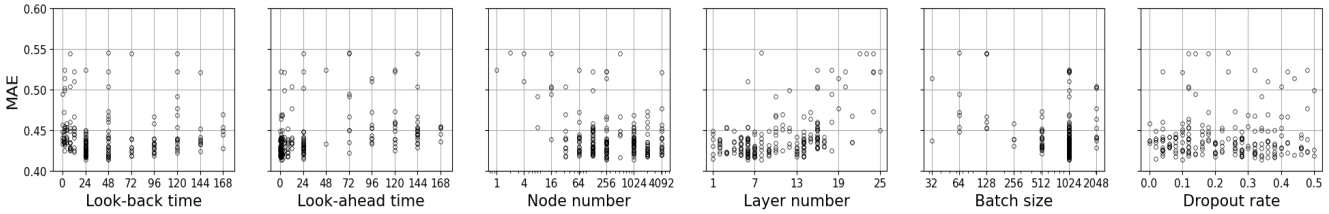
Table 5 presents the Test MAE of the optimized ANN model and the reanalysis ocean wave data of ECMWF ERA5. As the ANN model for estimating significant wave height, Bi-GRUN showed the best performance with 0.169 m—it was approximately 13% lower than the Test MAE of 0.194 m from the significant wave height of ECMWF ERA5. All the five ANN models estimating significant wave height outperformed the significant wave height of ECMWF ERA5. Fig. 9 (a) illustrates the time series of significant wave height from KMA, ECMWF, and Bi-GRUN over the dataset period (2010–2017), while Figs. 9 (b), (c), and (d) depict the MAE for the training set (Training



(a) Bi-GRUN for significant wave height



(b) GRUN for peak wave period



(c) FNN for wave direction

Fig. 8 Validation MAE of selected hyperparameter combinations in HPO process

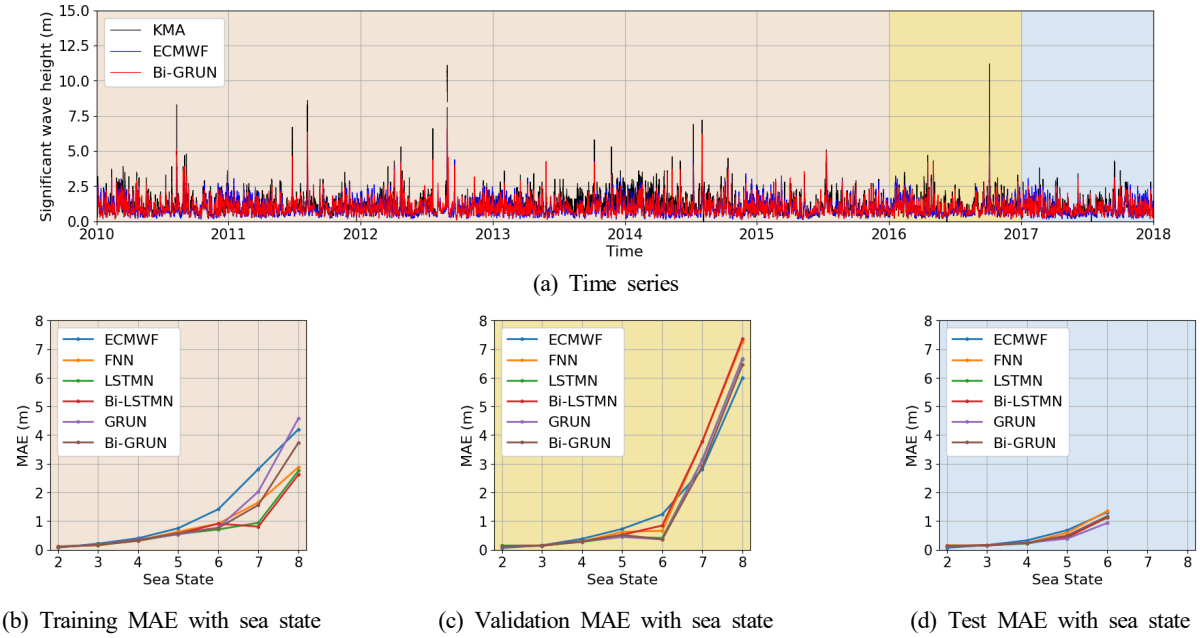
Table 4 Optimum hyperparameter combination of ANN models

Target	Model	Optimum hyperparameter combination						Validation MAE
		Look-back	Look-ahead	Node number	Layer number	batch size	dropout rate	
Significant wave height	FNN	48	4	64	4	1024	0.08	0.194
	LSTMN	96	4	64	5	1024	0.36	0.185
	Bi-LSTMN	144	8	16	8	1024	0.40	0.188
	GRUN	144	24	64	6	1024	0.12	0.180
	Bi-GRUN	168	12	64	5	1024	0.24	0.184
Peak wave period	FNN	96	1	512	9	1024	0.08	1.22
	LSTMN	144	4	64	4	1024	0.46	1.18
	Bi-LSTMN	96	0	64	4	1024	0.06	1.19
	GRUN	120	12	32	5	1024	0.16	1.21
	Bi-GRUN	96	1	8	2	1024	0.18	1.21
Wave direction	FNN	48	1	256	6	1024	0.34	46.1
	LSTMN	96	1	128	8	1024	0.10	45.6
	Bi-LSTMN	48	0	16	8	1024	0.30	46.6
	GRUN	48	24	32	7	1024	0.06	45.9
	Bi-GRUN	48	12	32	4	1024	0.50	45.9

MAE), Validation MAE, and Test MAE by sea state code (Table 2). The MAE of the significant wave height of ECMWF ERA5 was slightly better for Sea State Code 2, while the ANN model was better for Sea State Codes 3 to 6. The MAE of the ANN model increased for Sea State Code 7 and above, and it seems that owing to the extremely low percentage of the samples—less than 0.1% of the total (Table 2)—the ANN model underestimated the significant wave height for higher sea state codes. Furthermore, a higher Validation MAE of the ANN model than the Validation MAE of the significant wave height data of ECMWF ERA5 may be due to a sharp decrease in the input wind speed as Typhoon Chaba (Fig. 11(h)), which occurred during the validation set period, was extremely close to the Gyumundo ocean buoy (Fig. 11(e)), unlike the track of Typhoon Bolaven in 2012 (Fig. 11(d)), which occurred during the training set.

Table 5 Test MAE and RMSE of ANN models

Model	Significant wave height (m)		Peak wave period (s)		Wave direction (°)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
FNN	0.178	0.238	1.20	1.74	**46.2**	63.6
LSTMN	0.176	0.233	1.23	1.80	46.7	64.0
Bi-LSTMN	0.181	0.242	**1.18**	1.77	47.5	65.3
GRUN	0.170	0.229	**1.18**	**1.75**	46.4	64.4
Bi-GRUN	**0.169**	0.229	1.24	1.80	47.4	65.5
ECMWF	0.194	0.264	1.28	1.82	99.3	110.0



(a) Time series

(b) Training MAE with sea state (c) Validation MAE with sea state (d) Test MAE with sea state

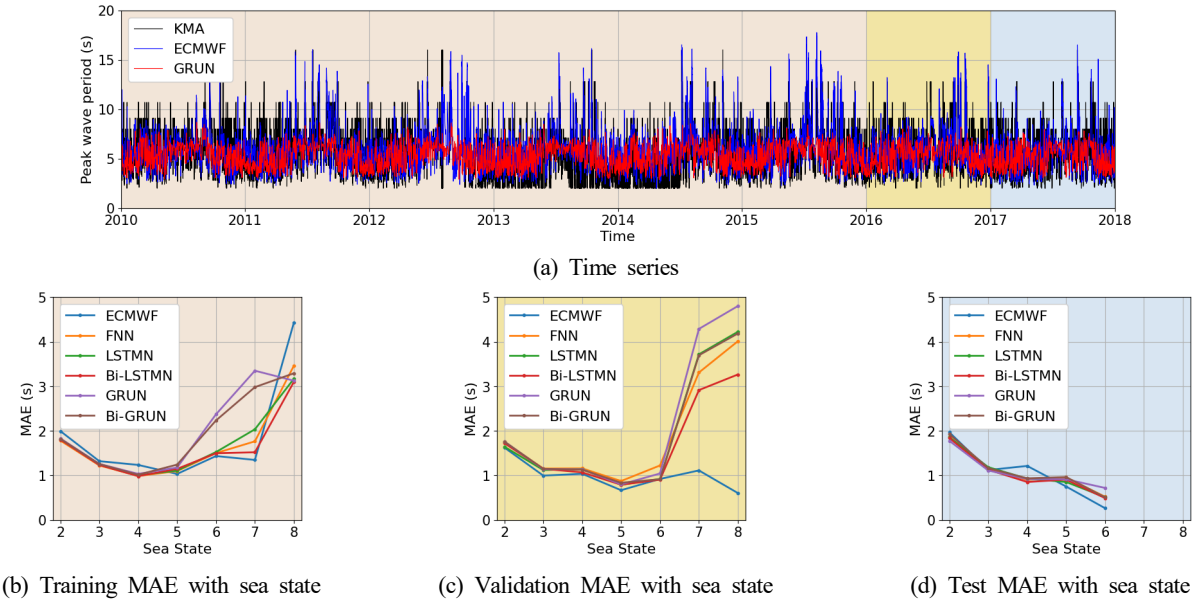
Fig. 9 Time series of significant wave height and MAE for each sea state

4.3 Peak Wave Period

The peak wave period estimated by the optimized ANN model performed better than the peak wave period data from ECMWF ERA5 (Table 5). Bi-LSTMN and GRUN showed the best performance with 1.18 s, which was approximately 8% lower than 1.28 s in the Test MAE for peak wave period data from ECMWF ERA5. Among the ANN models with the lowest Test MAE (Bi-LSTMN and GRUN), GRUN showed a good Test RMSE of 1.75 s. Fig. 10 illustrates the time series of peak wave periods and the MAE by sea state code for the total dataset period. The ANN model tends to underestimate the peak wave period when Typhoon Bolaven in 2012 and Typhoon Chaba in 2016 approached the Geomundo ocean buoy and the peak wave period longer than 10 s was measured (Figs. 11 (c), (g)). It seems to be because of the influence of ocean waves transferred from spatially distant waters. The wind data of ECMWF ERA5 at the single point closest to the Geomundo ocean buoy is considered to have insufficient information to estimate a peak wave period longer than 10 s.

4.4 Wave Direction

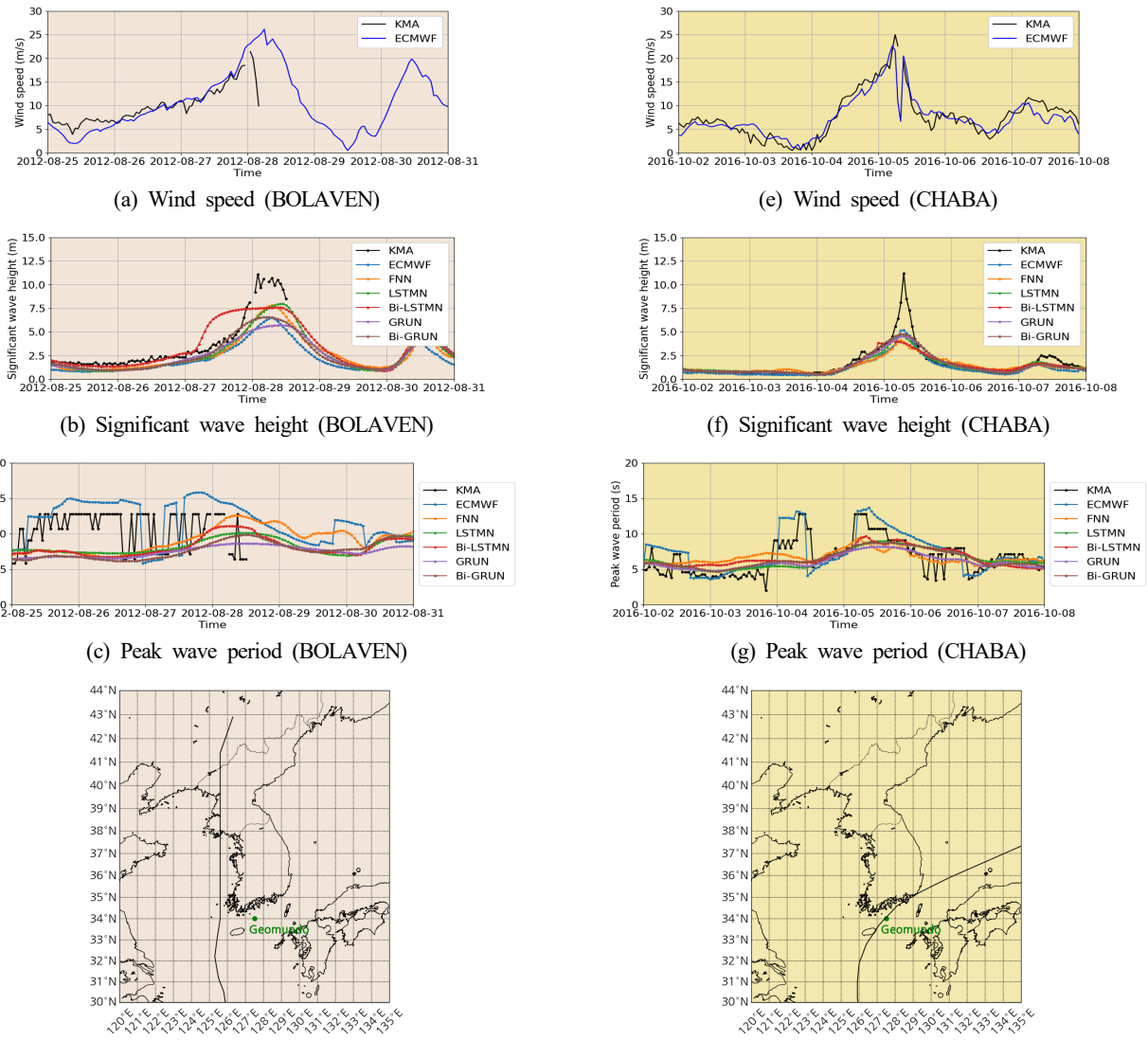
The wave direction estimation of the optimized ANN model also outperformed the wave direction data of ECMWF ERA5. It was confirmed that the optimized ANN model performed more than 50% better than the Test MAE of 99.3° from the wave direction data of ECMWF ERA5 (Table 5). The best performing ANN model was the FNN with the Test MAE of 46.2°. Fig. 12 depicts the time series of wave directions for the entire dataset, while Fig. 13 depicts the wave rose for the ocean buoy, the wave direction data of the ECMWF ERA5, and the FNN corresponding to the time period of the test set (2017). The direction of the wave rose indicates the direction in which ocean wave energy moves. The most dominant wave direction in the FNN (Fig. 13(a)) was 40°, which was the same as that of the ocean buoy (Fig. 13(b)). This confirmed that it produced as good performance as ocean buoy wave direction data. Meanwhile, the wave direction data of ECMWF ERA5 showed the highest frequency of southeast (140°) (Fig. 13(c)), which was the same as the northwest wind (320°)—the direction of the dominant wind in the waters. This may be due to the fact that the ocean wave numerical model of ECMWF ERA5 did not reflect in situ observations in the data assimilation process, and there were errors in coastline and water depth data in the waters where the ocean buoy was installed.



(a) Time series

(b) Training MAE with sea state (c) Validation MAE with sea state (d) Test MAE with sea state

Fig. 10 Time series of peak wave period and MAE for each sea state



(a) Wind speed (BOLAVEN) (e) Wind speed (CHABA)

(b) Significant wave height (BOLAVEN) (f) Significant wave height (CHABA)

(c) Peak wave period (BOLAVEN) (g) Peak wave period (CHABA)

(d) Track of typhoon (BOLAVEN) (h) Track of typhoon (CHABA)

Fig. 11 Comparison of measured data and estimated data results during the passage of typhoons BOLAVEN (2012) and CHABA (2016)

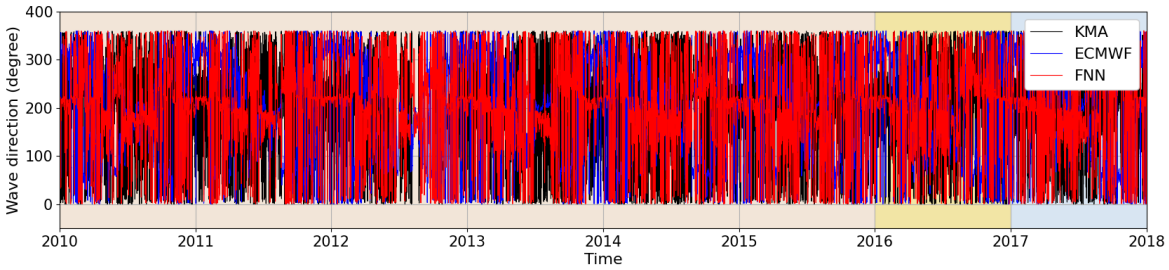
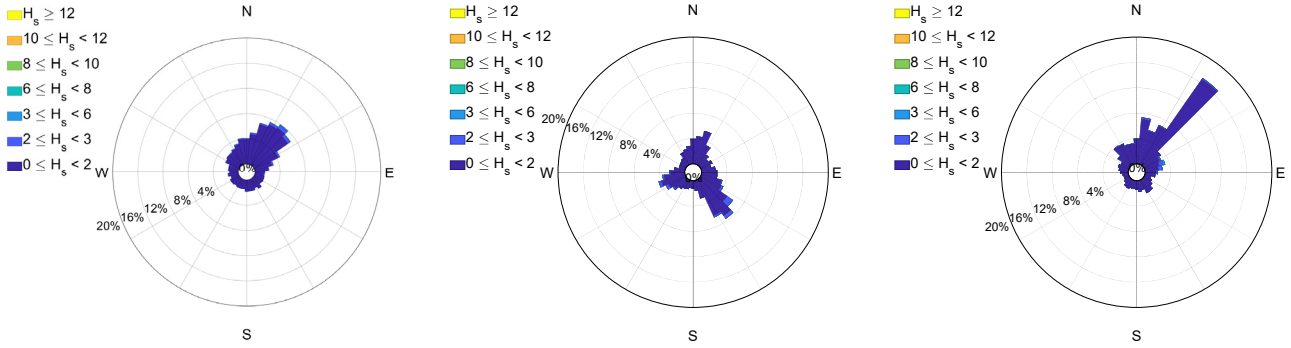


Fig. 12 Time series for wave direction



(a) KMA (b) ECMWF (c) FNN

Fig. 13 Comparison of wave rose diagrams in 2017

5. Conclusions

Ocean wave data acquired from an ocean buoy—an in situ observational instrument—is highly reliable. However, often, there exists the issue of missing data because of equipment failure caused by typhoons or maintenance. Ocean wave data missing over an extended period may lead to inaccurate wave scatter diagrams and underestimated results in extreme value analysis. This study utilized eight years of measured ocean wave information from the Korea Meteorological Administration's Geomundo ocean buoy and the wind data of ECMWF ERA5 to propose the ANN models suitable for each of the ocean wave parameters to fill in missing ocean wave data.

Five ANN models (FNN, LSTMN, Bi-LSTMN, GRUN, and Bi-GRUN) were applied to fill in the missing ocean wave data from the ocean buoy, and wind speed and direction data from ECMWF ERA5 were utilized as the input. MAE and Adam were applied as the loss function and optimizer, respectively. The HPO technique of Bayesian optimization was applied to six hyperparameters (look-back time, look-ahead time, node number, layer number, batch size, and dropout rate), which may affect the performance of the ANN models, to explore the optimal combination of hyperparameters. Through the training and validation process, the five ANN models were applied to three wave parameters (significant wave height, peak wave period, and wave direction) in the ocean buoy wave data. A total of 15 optimized ANN models were developed in this study. The research compared the results of five optimized ANN models for each of the three ocean wave parameters and suggested the most suitable ANN model for each of the ocean wave parameters.

(1) The estimation results of all the optimized ANN models showed a lower Test MAE than the ocean wave data of ECMWF ERA5, which confirmed that the optimized ANN models were better. Next, this study compared the Test MAE of five optimized ANN models for each of the ocean wave parameters and suggested the best ANN model for each parameter. Accordingly, Bi-GRUN, GRUN, and FNN were considered to be the ANN models most suitable for filling in missing data for significant wave height, peak wave period, and wave direction.

(2) The increased Test MAE of the optimized ANN models for a higher sea state code suggests a drop in performance in filling in missing data for significant wave height. This study conjectured that it may be due to the issue of imbalanced data caused by the difference in the number of samples depending on the sea state code of the ocean wave data applied in the training and validation process. Future research may apply data sampling and data augmentation techniques to improve performance. As the eye of Typhoon Chaba in 2016 passed directly over the Geomundo ocean buoy, wind data were missing at the Geomundo ocean buoy, and the significant wave height in the ANN model, estimated with the sharply reduced wind speed in ECMWF ERA5, may have been underestimated. The results of the peak wave period estimated by the ANN model optimized for a peak wave period wave of greater than 10 s showed increased errors, confirming a trend of decreased accuracy. These ocean waves were generated and transferred by high-speed winds from spatially distant waters. Therefore, developing an ANN model based on wind data from waters, considering a typhoon's track, is expected to improve the accuracy of long-term ocean wave estimation.

(3) This study suggested ANN models for filling in missing ocean wave data, and trained, validated, and tested the models based on the

measurement results of the Geomundo ocean buoy installed in the waters around Korea. This study employed the measurement results of a single ocean buoy to evaluate the performance of the ANN models optimized for each of the ocean wave parameters. Nonetheless, the study's finding, that the optimized ANN models were more accurate than the data of ECMWF ERA5, widely applied as reliable ocean wave data worldwide, suggests that AI models are expected to fill in the missing ocean wave data of an ocean buoy.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Funding

This work was supported by a 2-Year Research Grant of Pusan National University and "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2023RIS-007).

References

Alizadeh, M. J., & Nourani, V. (2024). Multivariate GRU and LSTM models for wave forecasting and hindcasting in the southern Caspian Sea. *Ocean Engineering, 298,* 117193. https://doi.org/10.1016/j.oceaneng.2024.117193

American Petroleum Institute (API). (2005). *Design and analysis of stationkeeping system for floating structure* (API RP-2SK).

Bales, S. L., Lee, W. T., Voelker, J. M., & Taylor, D. W. (1981). *Standardized wave and wind environments for NATO operational areas*. David W. Taylor Naval Ship Research and Development Center.

Bujak, D., Bogovac, T., Carević, D., & Miličević, H. (2023). Filling missing and extending significant wave height measurements using neural networks and an integrated surface database. *Wind, 3*(2), 151−169. https://doi.org/10.3390/wind3020010

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. *Advances in neural information processing systems, 31*.

Cho, H. Y., Jeong, W. M., Oh, S. H., & Baek, W. D. (2020). Parameter estimation and fitting error analysis of the representative spectrums using the wave spectrum off the Namhangjin, East Sea. *Journal of Korean Society of Coastal and Ocean Engineers, 32*(5), 363−371. https://doi.org/10.9765/KSCOE.2020.32.5.363

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. https://doi.org/10.48550/arXiv.1412.3555

Det Norske Veritas (DNV). (2014). *Recommend Practice* (DNV-RP-C205). DNV.

Du, W., Côté, D., & Liu, Y. (2023). Saits: Self-attention-based imputation for time series. *Expert Systems with Applications, 219*, 119619. https://doi.org/10.1016/j.eswa.2023.119619

European Centre for Medium-Range Weather Forecasts (ECMWF). (2016). *IFS documentation CY41R2*.

Guijo-Rubio, D., Durán-Rosal, A. M., Gómez-Orellana, A. M., & Fernández, J. C. (2023). An evolutionary artificial neural network approach for spatio-temporal wave height time series reconstruction. *Applied Soft Computing, 146*, 110647. https://doi.org/10.1016/j.asoc.2023.110647

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999−2049. https://doi.org/10.1002/qj.3803

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735−1780.

Kim, E. D., Ko, S. K., Son, S. C., & Lee, B. T. (2021). Technical trends of time-series data imputation. *Electronics and Telecommunications Trends, 36*(4), 145−153. https://doi.org/10.22648/ETRI.2021.J.360414

Korres, G., Ravdas, M., & Zacharioudaki, A. (2019). *Mediterranean Sea Waves Hindcast (CMEMS MED-Waves)* [Data set]. Copernicus Monitoring Environment Marine Service (CMEMS).

Minuzzi, F. C., & Farina, L. (2023). A deep learning approach to predict significant wave height using long short-term memory. *Ocean Modelling, 181*, 102151. https://doi.org/10.1016/j.ocemod.2022.102151

Pang, J., & Dong, S. (2023). A novel multivariable hybrid model to improve short and long-term significant wave height prediction. *Applied Energy, 351*, 121813. https://doi.org/10.1016/j.apenergy.2023.121813

Park, S. B., Shin, S. Y., Jung, K. H., & Lee, B. G. (2021). Prediction of significant wave height in korea strait using machine learning. *Journal of Ocean Engineering and Technology, 35*(5), 336−346. https://doi.org/10.26748/KSOE.2021.021

Wei, Z. (2021). Forecasting wind waves in the US Atlantic Coast using an artificial neural network model: Towards an AI-based storm forecast system. *Ocean Engineering, 237*, 109646. https://doi.org/10.1016/j.oceaneng.2021.109646

World Meteorological Organization (WMO). (2019). Part A—Alphanumeric Codes. *Manual on Codes: International Codes, I. 1.*

Author ORCIDs

Author name	ORCID
Shin, Seongyun	0000-0001-6665-9092
Park, Seonghyun	0009-0005-6886-3368
Jung, Kwang Hyo	0000-0002-8229-6655
Park, Sung Boo	0000-0001-9587-2183