Applied Ocean Research 148 (2024) 104024

Contents lists available at ScienceDirect

Applied Ocean Research

journal homepage: www.elsevier.com/locate/apor

Research paper

Identifying causes of errors between two wave-related data using performance metrics

Takahito Iida

Department of Naval Architecture and Ocean Engineering, Osaka University, Suita, Osaka 5650871, Japan

ARTICLE INFO

Keywords:
Performance metrics
Quantitative comparison
Time-series wave data
Frequency domain's representation

ABSTRACT

Recently, numerous prediction methods of time-series data of wave-related phenomena have been developed. Quantitative evaluation of an error of a predicted result against a reference result is important to improve prediction accuracy. Many performance metrics are engaged to evaluate their accuracy. However, it is difficult for them to identify the error causes because all errors, no matter what the cause, are combined together. This paper presents new representations of performance metrics to separate errors by causes. Performances are evaluated in a frequency domain instead of a time domain. A frequency domain's amplitude and phase are respectively evaluated using performance metrics. In addition, to detect errors due to instantaneous phenomena and changes in an error trend over time, mean errors are defined by three-time intervals. The original metric uses all time-series data. On the other hand, a finite interval mean error at the present time is calculated by the mean of the preceding data. In addition, cumulative mean error at the present time is calculated by the mean of the data up to the present time. These new representations of performance metrics could help to identify error causes. Benchmark tests are carried out to demonstrate the validity of the proposed representations.

1. Introduction

The importance of a comparison of two time-series data (e.g. measured and predicted data) is increasing more and more in response to the recent developments of real-time wave prediction methods (e.g. Al-Ani et al., 2020; Law et al., 2020; Iida and Minoura, 2022) and digital twin technologies (e.g. Lee et al., 2022; Liong and Chua, 2022; Isnaini et al., 2024) in ocean engineering communities. Integration of real-time prediction methods of wave-related data into a real operation of vessels enables us to immediately make a decision on the navigation to avoid any sudden risk or failure (Lee et al., 2022). Such a decision depends on the reliability of the predictions. Therefore, it is essential to reveal the prediction accuracy and its applicability limit.

A primitive method of comparison is to display the two data on a graph and qualitatively evaluate how much they overlap. Since visual perception differs for each person, the judgment of whether the result is acceptable or not also depends on the person. Therefore, quantitative performance metrics are necessary to ensure a unified evaluation. Many metrics have been established so far, such as R-squared, Pearson correction coefficient, mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), and normalized root mean squared error (NRMSE). It is reported that the RMSE is optimal for normal distribution errors, and the MAE is optimal for Laplacian distribution errors (Hodson, 2022). These metrics are widely used in many communities, and the ocean engineering communities also commonly use these metrics (e.g. Fan et al., 2020; Jörges et al., 2021; Wang et al., 2021; Lee et al., 2022). As our target data is related to sea waves, the data often crosses zero and its mean is zero. Therefore, percentage-type errors become infinite or undefined values at zero-cross points. To overcome this disadvantage, a mean arctangent absolute percentage error (MAAPE) was proposed (Kim and Kim, 2016). As for metrics to wave-related data, Perlin and Bustamante (2016) proposed a surface similarity parameter (SSP) based on the Sobolev norm. This compares two data in the frequency domain, and thus wave amplitude and phase information can be considered.

These performance metrics objectively provide the degree of concordance/discrepancy in data. These are helpful to rank the accuracy of results by different prediction methods. However, each performance metric has inherent features, and inadequate choice of the metric could hide shortcomings. In addition, although these performance metrics are often used to only justify a proposed result, it is more important to identify the causes of errors. Otherwise, these prediction methods could not be improved and might be used even for inadequate situations.

In this paper, new representations of the performance metrics are proposed to determine the causes of errors. Existing performance metrics usually calculate a mean error, that is, all errors, regardless of

E-mail address: iida@naoe.eng.osaka-u.ac.jp.

https://doi.org/10.1016/j.apor.2024.104024
Received 13 January 2024; Received in revised form 18 March 2024; Accepted 18 April 2024
Available online 28 April 2024

their causes, are displayed together. However, it is desirable to separate errors by their respective causes. Since wave-related data are expressed by Fourier series expansion, amplitude and phase information in a frequency domain are essential. Therefore, we propose to evaluate these frequency domain's amplitude and phase respectively instead of direct evaluation of time-series data. Two prefixes, frequency domain's amplitude (FA-) and frequency domain's phase (FP-), are applied to existing metrics, such as FA-MAE (frequency domain's amplitude mean absolute error), FA-MAPE (frequency domain's amplitude mean absolute percentage error), and FP-MAE (frequency domain's phase mean absolute error). These metrics enable us to discuss the errors in the classical frequency domain. In addition, since existing metrics are calculated using all time-series data, it is not possible to detect an instantaneous error and change in a trend of errors over time. Therefore, we additionally introduce two mean representations. In the finite interval representation, the error at the current time is defined as the mean of the preceding m data points. On the other hand, in the cumulative representation, the error at the current time is defined as the mean of the data from the beginning up to the current time. By expanding existing performance metrics using these new representations, many causes of errors could be identified. Of course, whether new metrics are proposed or not, it is necessary to carefully examine the causes of errors by seeing time histories and spectra one by one. However, when the amount of data is huge, careful checking of all data is impractical. The proposed metrics could be used for the first screening of the data, and it allows users to focus on the likely source of errors before deep investigation of the error.

To demonstrate the proposed metrics, some typical wave-related time-series data are prepared. These are compared by a set of performance metrics, and it is shown that each cause can be identified by seeing these performance evaluations.

2. Brief review of existing performance metrics

Firstly, some performance metrics commonly used in ocean engineering communities are briefly reviewed. Here, two time-series data $\boldsymbol{x} = (x_1, x_2, \dots, x_N)$, $\boldsymbol{y} = (y_1, y_2, \dots, y_N)$ are considered where N is the total number of the data. The data $\boldsymbol{y}$ is assigned as a reference value, whereas the data $\boldsymbol{x}$ is test data whose error from the reference needs to be investigated. For example, comparisons of true value $\boldsymbol{y}$ vs. observed value $\boldsymbol{x}$, or measured value $\boldsymbol{y}$ vs. predicted value $\boldsymbol{x}$ are expected. Therefore, when percentage-type errors are considered, the ratio (denominator) is defined by $d_i = y_i$ as data x_i has a larger uncertainty. Note that the median $d_i = (x_i + y_i)/2$ may be used for the denominator if the uncertainties of the two data are comparable.

In the field of ocean science, most of the ocean data have been analyzed using statistical evaluation methods. However, this paper focuses more on the evaluation of time-series data, and we do not deeply explain those statistical methods.

1. R-squared (coefficient of determination)
   R-squared (a.k.a. coefficient of determination) is a metric to evaluate the fitness of two data (e.g. William et al., 2003). This could be given as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{1}$$

   where $\overline{y}$ is the mean of data $\boldsymbol{y}$. Note that the R-squared has some variations (Kvålseth, 1985), and the first definition of Kvålseth (1985) is shown here since this may be the most popular definition. Eq. (1) indicates a ratio of the residual sum of squared to the total sum of squared. R^2 is commonly used in statistics, but this is also used for comparing two time-series data. The range of R^2 is $[-\infty, 1]$ where $R^2 = 1$ means a perfect fit, and a high value of R^2 indicates a strong fit. R^2 could be negative when $\sum_i^N (x_i - y_i)^2 > \sum_i^N (y_i - \overline{y})^2$. The minimum value of R^2 is 0 when a linear regression model is compared with a reference. The scatter plot is often shown in addition to the R-squared and the linear regression model.
2. Pearson correlation coefficient
   Pearson correlation coefficient is a metric to evaluate a linear correlation between two data (e.g. William et al., 2003). This is defined as

$$C_r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{2}$$

   where $\overline{x}$ is the mean of data $\boldsymbol{x}$. Eq. (2) indicates a ratio of covariance of two data to the product of their standard deviations. The range of C_r is $[-1, 1]$. $C_r = 1$ corresponds to a perfect positive linear relation whereas $C_r = -1$ is a negative linear relation. In addition, $C_r = 0$ means no linear relation between the two data. Pearson correlation coefficient is simple, easy to interpret, and scale-independent. Therefore, this is used in many fields. In case two wave-related data are compared, the Pearson correction coefficient provides the degree of concordance of wave phases. On the other hand, the concordance rate of wave amplitudes cannot be evaluated.
3. Mean absolute error (MAE)
   Mean absolute error (MAE) is a metric to evaluate an average absolute error between two data, which is given as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{3}$$

   The range of the MAE is $[0, \infty]$ where MAE $= 0$ is a perfect agreement of two data and a small value is a small error. The MAE is simple and easy to interpret. The magnitude of the MAE depends on the scale of the data.
4. Mean absolute percentage error (MAPE)
   Mean absolute percentage error (MAPE) is a metric to evaluate a percentage of absolute error. The MAPE is a percentage representation of the MAE, which is given as

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x_i - y_i}{y_i}\right| \tag{4}$$

   The range of the MAPE is $[0, \infty]$ where MAPE $= 0$ corresponds to 0% error, and a small value indicates a low percentage of error. As the MAPE is percentage-type, this metric is scale-independent. The MAPE is sensitive to small reference values as this can lead to large percentage errors even if the absolute errors are small. If the data has zero or near zero, the MAPE can become undefined or extremely large. Therefore, this metric is not suitable for wave-related data since such data often cross zero.
5. Mean squared error (MSE)
   Mean square error (MSE) is a metric to evaluate an average squared error between two data (e.g. Willmott et al., 1985), which is given as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{5}$$

   The range of the MSE is $[0, \infty]$ where MSE $= 0$ is a perfect agreement of two data, and a small value indicates a small error. The value of the MAE depends on the scale of the data. Since values are squared, extreme values are penalized more than small values, unlike the MAE. Therefore, the MSE is sensitive to outliers.

6. Root mean squared error (RMSE)
Root mean squared error (RMSE) is a metric to evaluate an error between two data (e.g. Willmott et al., 1985). The RMSE is calculated by taking a squared root of the MSE as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{6}$$

The range of the RMSE is $[0, \infty]$ where RMSE = 0 shows a perfect agreement of two data, and a small value is a small error. Similar to the MSE, the RMSE is sensitive to extreme values. In addition, the value of the RMSE depends on the scale of the data. The RMSE can be rewritten as

$$\mathrm{RMSE} = \sqrt{\bar{e}^2 + s^2} \tag{7}$$

where

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i),\ s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i - \bar{e})^2 \tag{8}$$

Here, $\bar{e}$ is a mean error (ME) of two data (i.e. bias) and s is a standard deviation. This indicates that the RMSE is the hypotenuse length of a right-angled triangle, with the side lengths represented by the mean error and the standard deviation (Pythagorean theorem). As the RMSE can be statistically interpreted, this is one of the most popular performance metrics and is widely used. When the mean error (bias) is zero, the RMSE becomes the standard deviation of the error. Such a situation is popular for free surface wave problems as the mean is generally zero. However, when a vessel or offshore structure is considered, we should carefully check mean phenomena, such as steady sinkage, steady trim, added wave resistance, and wave drift force. Either way, it is recommended to show not only the value of the RMSE but also the values of the mean error and standard deviation to evaluate the performance.
There are many arguments over whether RMSE or MAE should be used (Willmott and Matsuura, 2005; Chai and Draxler, 2014; Hodson, 2022). Hodson (2022) stated that the RMSE is optimal for normal (Gaussian) errors, and the MAE is optimal for Laplacian errors where these errors follow the normal distribution and Laplacian distribution, respectively. In any case, it is valuable to plot a histogram of the residuals and check the error distribution. Once the likely distribution is determined, a Quantile–Quantile (Q–Q) plot can also visualize the similarity of the distributions (generally, the normal distribution is used for the reference). Note that the MAE and RMSE are generalized in Willmott et al. (1985).
7. Normalized root mean squared error (NRMSE)
Normalized root mean squared error (NRMSE) is a metric to evaluate an error between two data in normalized form. The NRMSE is calculated by normalizing the RMSE as

$$\mathrm{NRMSE} = \frac{1}{\bar{d}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \tag{9}$$

Here, the RMSE is normalized by $\bar{d}$, and there are some variations for $\bar{d}$, such as $\bar{d} = \bar{y}$, $\bar{e}$, and $y_{\max} - y_{\min}$ where $y_{\max}$ and $y_{\min}$ are maximum and minimum values of $\boldsymbol{y}$. The range of NRMSE is $[0, \infty]$. The NRMSE is sensitive to extreme values and is scale-independent. It is necessary to check the value of $\bar{d}$ to avoid a division by zero.
8. Mean arctangent absolute percentage error (MAAPE)
Mean arctangent absolute percentage error (MAAPE) is a metric to evaluate a percentage error and is defined on a finite range with avoiding zero divisions (Kim and Kim, 2016). The MAAPE is given as

$$\mathrm{MAAPE} = \frac{1}{N}\sum_{i=1}^{N}\arctan\left(\left|\frac{x_i - y_i}{y_i}\right|\right) \tag{10}$$

The MAAPE is represented by the arctangent of the MAPE; the MAAPE indicates the angle whereas the MAPE is the ratio. Therefore, the MAAPE can avoid undefined values by zero divisions. In addition, the range of the MAAPE is finite as $[0, \pi/2]$ where MAAPE = 0 corresponds to perfect agreement. Since the MAAPE is represented by arctangent, this weighs more for small errors. Therefore, this metric is insensitive to extreme errors but sensitive to small errors. From the MAAPE, the mean arctangent absolute error (MAAE) could be also defined as

$$\mathrm{MAAE} = \frac{1}{N}\sum_{i=1}^{N}\arctan(|x_i - y_i|) \tag{11}$$

Since the MAAE is scale-dependent, this expression should not be used for actual evaluations. However, the introduction of the MAAE enables the discussion of differences among the MAE, RMSE, and MAAE as these are all scale-dependent metrics. The MAE, RMSE, and MAAE are different in terms of what kind of error is being weighted. The RMSE has a larger slope for larger errors; the MAAPE has a larger slope for smaller errors. The MAE has a flat slop for any errors. In the case the accuracy is emphasized, the MAAE could be suitable. On the other hand, the RMSE could be suitable in the case the error (especially outliers) is emphasized. Compared to other metrics, the number of users of the MAAPE is small, however, this metric has unique and valuable characteristics.
9. Surface similarity parameter (SSP)
Surface similarity parameter (SSP) is a metric to evaluate a normalized error using the Sobolev norm (frequency domain values) (Perlin and Bustamante, 2016), which is given as

$$\mathrm{SSP} = \sqrt{\frac{\int |X(\omega) - Y(\omega)|^2 d\omega}{\int |Y(\omega)|^2 d\omega}} \tag{12}$$

$$\approx \sqrt{\frac{\sum_{i=1}^{N}|X_i - Y_i|^2}{\sum_{i=1}^{N}|Y_i|^2}} \tag{13}$$

where complex functions $X(\omega)$ and $Y(\omega)$ are Fourier transform of $x(t)$ and $y(t)$. Eq. (13) is a discrete representation of Eq. (12) where $X_i = X(\omega_i)$, $Y_i = Y(\omega_i)$, and ω_i $(i = 1, 2, \ldots, N)$ is a discrete frequency. Note that the original SSP is normalized by the median (Perlin and Bustamante, 2016), but Eq. (12) is normalized by $Y(\omega)$ to uniform the expression with other metrics. The SSP compares two data in the frequency domain, and information of both amplitude and phase are considered. Similar to the MAAPE, the SPP is not a popular metric. Nevertheless, the SSP is a good candidate for the performance metric for wave-related data since the frequency domain analysis is popular and easy to understand.
10. Other metrics
We reviewed some typical performance metrics above, however, these have further variations, such as relative mean error (RME), symmetric mean absolute percentage error (SMAPE), mean absolute scaled error (MASE), and so on. These are given as the following equations:

$$\mathrm{RME} = \frac{\mathrm{MAE}}{\mathrm{MAD}} = \frac{\sum_{i=1}^{N}|x_i - y_i|}{\sum_{i=1}^{N}|y_i - \bar{y}|} \tag{14}$$

$$\mathrm{SMAPE} = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i - y_i|}{(|x_i| + |y_i|)/2} \tag{15}$$

$$\mathrm{MASE} = \frac{\mathrm{MAE}}{\mathrm{MAE}_{\mathrm{Naïve}}} = \frac{\frac{1}{N}\sum_{i=1}^{N}|x_i - y_i|}{\frac{1}{N-1}\sum_{i=2}^{N}|y_{i-1} - y_i|} \tag{16}$$

where MAD is the mean absolute deviation, and $\mathrm{MAE}_{\mathrm{Naïve}}$ is the mean absolute error using the Naïve forecasting, i.e. $x_i = y_{i-1}$. In this paper, these variations are not considered.

Table 1
Summary of the existing performance metrics.

No.	Metric	Characteristics
1	R-squared R^2	R^2 evaluates the fitness of two data, especially for a regression model.
2	Pearson correlation coefficient C_r	C_r evaluates a linear relation of two data. The concordance of wave phase is estimated, and the amplitude is out of the estimation.
3	Mean absolute error (MAE)	The MAE is scale-dependent, weighting-free, and optimal for Laplacian distribution errors.
4	Mean absolute percentage error (MAPE)	The MAPE is percentage type of the MAE and scale-independent. This has a problem for zero division.
5	Mean squared error (MSE)	The MSE is scale-dependent and outlier-sensitive. Generally, the RMSE is used instead of the MSE.
6	Root mean squared error (RMSE)	The RMSE is a hypotenuse length of a triangle with legs of a mean error and standard deviation. This is scale-dependent, outlier-sensitive, and optimal for normal distribution errors.
7	Normalized root mean squared error (NRMSE)	The NRMSE is scale-independent. The interpretation depends on the form of normalization.
8	Mean arctangent absolute percentage error (MAAPE)	The MAAPE represents errors by the angle. This is scale-independent and zero-division-free. This is outlier-insensitive and sensitive to small error.
9	Surface similarity parameter (SPP)	The SPP evaluates errors of complex amplitude in the frequency domain. This is scale-independent.

It is also mentioned that dynamic time warping (DTW) evaluates the similarity of two time-series data (see Senin, 2008). The MAE is based on the Manhattan distance, and thus the distance is calculated by reference data and test data at the same time. On the other hand, the DTW measures the distance between two data whose time scales or speeds are different. Note that methods of distance measures are reviewed by Ding et al. (2008). Edit distance (a.k.a. Levenshtein distance) approaches might be also applicable to evaluate the similarity of two data by transforming real values of time-series data to strings (e.g. a symbolic aggregate approximation, SAX (Lin et al., 2003)). These metrics are generally used for more complicated time-series data, such as for data mining of stock prices or human behavior patterns. Time series of ocean waves and their related, i.e. our targets, are commonly represented by the summation of trigonometric functions. Therefore, these distance measures need not necessarily be applied. Common time-series representation methods are reviewed by Lin et al. (2003).

Characteristics of these metrics are summarized in Table 1.

3. New representations of performance metrics

As seen in Section 2, there are various metrics to compare the two time-series data. However, it is difficult for these metrics to identify the cause of errors because all errors, no matter what the cause, are included together in the evaluation. Therefore, this paper presents new representations of performance metrics to separate errors by causes. Perlin and Bustamante (2016) proposed the SSP which evaluates data in the frequency domain. Since waves can be represented by Fourier series expansion, such an evaluation is essential and easy to understand. However, the SSP includes both amplitude and phase information, and these cannot be decoupled. In order to identify the cause, it is desirable to separately consider the errors of amplitude and phase. Therefore, we introduce two prefixes, a frequency domain's amplitude (FA-) and frequency domain's phase (FP-), which are added in front of the name of the performance metric. When these representations are applied to MAE, MAPE, and ME, they are defined as

$$\text{FA-MAE} = \frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\text{FA-AE}_i \tag{17}$$

$$\text{FA-MAPE} = \frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\text{FA-APE}_i \tag{18}$$

$$\text{FP-MAE} = \frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\text{FP-AE}_i \tag{19}$$

$$\text{FP-ME} = \frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\text{FP-E}_i \tag{20}$$

where

$$\text{FA-AE}_i = \frac{2}{N}\left|(|X_i| - |Y_i|)\right| \tag{21}$$

$$\text{FA-APE}_i = \left|\frac{|X_i| - |Y_i|}{|Y_i|}\right| \tag{22}$$

$$\text{FP-AE}_i = \begin{cases} |\text{PE}_i| & (|\text{PE}_i| \le \pi) \\ 2\pi - |\text{PE}_i| & (|\text{PE}_i| > \pi) \end{cases} \tag{23}$$

$$\text{FP-E}_i = \begin{cases} \text{PE}_i - 2\pi & (\text{PE}_i > \pi) \\ \text{PE}_i & (-\pi < \text{PE}_i \le \pi) \\ \text{PE}_i + 2\pi & (\text{PE}_i \le -\pi) \end{cases} \tag{24}$$

$$\text{PE}_i = \arg(X_i) - \arg(Y_i) \tag{25}$$

Here, $X(\omega)$ and $Y(\omega)$ are complex values that are given by the Fourier transform of $x(t)$ and $y(t)$. The discrete form of $X(\omega)$ is expressed by $X_i = X(\omega_i)$. In this paper, the discrete Fourier transforms are defined by the following equations:

$$\begin{cases} X(\omega_k) = \sum_{n=0}^{N-1} x(t_n)\left(e^{-\frac{2\pi i}{N}}\right)^{kn} \\ x(t_n) = \frac{1}{N}\sum_{k=0}^{N-1} X(\omega_k)\left(e^{-\frac{2\pi i}{N}}\right)^{-kn} \end{cases} \tag{26}$$

Therefore, $|X_i|$ denotes amplitude, and $\arg(X_i)$ represents phase of which range is $(-\pi, \pi]$. Note that N_ω is the number of data within the range of $[\omega_{\min}, \omega_{\max}]$, and thus $N_\omega \le N$. FA-AE$_i$ is an absolute error of amplitude at each frequency, and these means (i.e. MAE) are represented by FA-MAE. As shown in Eq. (21), the value is normalized by $N/2$ because a single-sided spectrum is used. This operation scales it to a component of the original amplitude. Here, we could define another prefix using a frequency domain's spectrum (FS-), such as

$$\text{FS-MAE}_i = \frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\text{FS-AE}_i \tag{27}$$

$$\text{FS-AE}_i = \left|\boldsymbol{\Phi}_X(\omega_i) - \boldsymbol{\Phi}_Y(\omega_i)\right| \tag{28}$$

where

$$\boldsymbol{\Phi}_X(\omega_i) = \frac{1}{2\Delta\omega}\left(\frac{2|X(\omega_i)|}{N}\right)^2 \tag{29}$$

is a spectrum of $X(\omega_i)$. This prefix directly compares the spectra. Since amplitude and spectrum satisfy a relation $2|X|/N = \sqrt{2\boldsymbol{\Phi}_X\Delta\omega}$, FA- and FS- are similar metrics. However, FA-AE$_i \neq \sqrt{2\Delta\omega}$FS-AE$_i$ for the errors. Which is more suitable depends on the problem. The ranges of FA-AE$_i$,

Table 2
Benchmark sets of two-wave data. $\boldsymbol{y}$ is reference data, and $\boldsymbol{x}$ is test data.

Case	$\boldsymbol{y}$	$\boldsymbol{x}$
1	Irregular waves based on Pierson–Moskowitz spectrum	Waves with $+\pi/2$ phase shift from $\boldsymbol{y}$
2		Waves with $+\pi$ phase shift from $\boldsymbol{y}$
3		Waves with $-\pi/2$ phase shift from $\boldsymbol{y}$
4		Waves with different random phase from $\boldsymbol{y}$
5		Waves with y_i multiplied by 0.8
6		Waves with y_i multiplied by 5/3
7		Waves y plus 0.1 (mean value shift)
8	Phase shifted waves using finite-depth dispersion relation	Phase shifted waves using deep water dispersion relation
9	Irregular waves based on Pierson–Moskowitz spectrum	Waves $\boldsymbol{y}$ plus high-frequency noise
10		Waves $\boldsymbol{y}$ plus low-frequency noise
11		Waves $\boldsymbol{y}$ with $x_i > 1.5 \rightarrow x_i = 0$ for modeling data missing
12	Experimentally measured JONSWAP waves (including reflected waves)	Predicted waves using an uni-directional prediction method
13		Predicted waves using a bi-directional prediction method

FA-MAE, FS-AE$_i$ and FS-MAE are $[0, \infty]$, and 0 indicates a perfect agreement. FA-APE$_i$ is an absolute percentage error of amplitude at each frequency, and these means are represented by FA-MAPE. Here, the ranges of FA-APE$_i$ and FA-MAPE are $[0, \infty]$ where 0 is a perfect agreement. Note that representations of FA-RMSE, FA-NRMSE, or FA-MAAPE are possible, but only MAE and MAPE are shown as these two metrics are used in a later section. PE$_i$ is a raw error of phase, and a phase error whose range is deformed as $(-\pi, \pi]$ is defined by FP-E$_i$. On the other hand, the absolute error of phase is given by FP-AE$_i$ where the range is $[0, \pi]$. The mean values of these metrics are named FP-ME and FP-MAE. It should be noted that the percentage representation is not applied to phase functions because the phase does not have magnitude.

Existing metrics evaluate data as one value (mean value) using all data. However, this cannot capture instantaneous errors and changes in a trend of errors over time. Therefore, we further introduce two variations of performance metrics. Here, it is defined that an original performance metric is expressed as follows:

$$\text{Original performance metric} = f\left((x_1, x_2, \dots, x_N), (y_1, y_2, \dots, y_N)\right) \tag{30}$$

where $f(\cdot)$ is any of a performance metric (e.g. RMSE). Eq. (30) indicates that the performance metric is calculated using all time-series data ($i = 1, 2, \dots, N$) of $\boldsymbol{x}$ and $\boldsymbol{y}$. On the other hand, in order to capture the instantaneous error, the performance metric at time i is calculated using the finite number m of data. This metric is named a finite interval performance metric and is given as

$$\text{Finite interval performance metric}(i) = f((x_{i-m+1}, \dots, x_i), (y_{i-m+1}, \dots, y_i)) \tag{31}$$

Here, m can be either a constant or a variable: e.g. m is a constant when a fixed time interval is used, and m is a variable when a time interval is calculated by the number of zero-up cross waves. This metric is defined when $i \geq m$. Here, we define the interval by the past data because we expect real-time prediction. However, if a posteriori estimation is considered, the ith data could be the center of the interval.

Moreover, a cumulative performance metric is defined as

$$\text{Cumulative performance metric}(i) = f((x_1, \dots, x_i), (y_1, \dots, y_i)) \tag{32}$$

where the performance metric at the time i is calculated using the data until i. This representation is useful to investigate the changes in a trend of errors over time. Combinations of these three representations facilitate the identification of causes of errors.

It is worth noting that the finite interval representation is similar to the short-time Fourier transform (Allen and Rabiner, 1977) where the time-series data is divided into segments. This plots the changes in the spectrum in time. On the other hand, Welch's method (Welch, 1967) and the averaged periodogram method (a.k.a. Bartlett's method (Bartlett, 1948)) average segmented spectra to reduce noises. These methods apply frequency domain analysis to segmented time-series data. Our ideas are the extension of their concepts to not only frequency domain analysis but also time domain analysis.

4. Benchmark tests

To investigate how performance metrics behave for comparisons of wave-related data, benchmark tests are carried out. We consider typical causes of errors, such as phase differences, amplitude differences, mean drift, different dispersion, frequency-dependent noises, missing data, and trend change. Corresponding 13 wave data are prepared as the benchmarks, and these are listed in Table 2. The following subsections explain the details of the data and discuss the comparisons.

4.1. Errors due to phase, amplitude, and mean drift (cases 1 to 7)

For cases 1 to 7, irregular waves based on the Pierson–Moskowitz spectrum (Pierson Jr. and Moskowitz, 1964) are used as the reference data $\boldsymbol{y}$. These are given as

$$y(t) = \sum_{n=1}^{N} \sqrt{2\Phi(\omega_n)\Delta\omega_n} \cos(\omega_n t + \varepsilon_n) \tag{33}$$

where

$$\Phi(\omega) = \frac{A}{\omega^5} e^{-\frac{B}{\omega^4}}, \; A = 172.8 \frac{H_{1/3}^2}{T_1^4}, \; B = \frac{691.2}{T_1^4} \tag{34}$$

Here, Φ is the Pierson–Moskowitz type frequency spectrum, T_1 is a mean wave period, and $H_{1/3}$ is a mean wave height. In this paper, mean wave period $T_1 = 8.0$ s and mean wave height $H_{1/3} = 3.0$ m are used. Irregular waves consist of 100 frequency components, and each component has a random phase. The frequency step size is decided based on $\Phi(\omega_n)\Delta\omega = \text{const.}$, i.e. the step size is unequally spaced so as to avoid periodicity with $2\pi/\Delta\omega$. In addition, the time step size is $\Delta t = 0.1$ s. The data are evaluated in the time range of 100 waves of the reference $\boldsymbol{y}$ where the number of waves is counted by zero-up cross method. Cutting-off frequencies $\omega_{\min} = 0.45$ rad/s and $\omega_{\max} = 1.3$ rad/s are used to calculate frequency domain values (i.e. FS-MAE, FA-MAE, FA-MAPE, FP-MAE, and FP-ME).

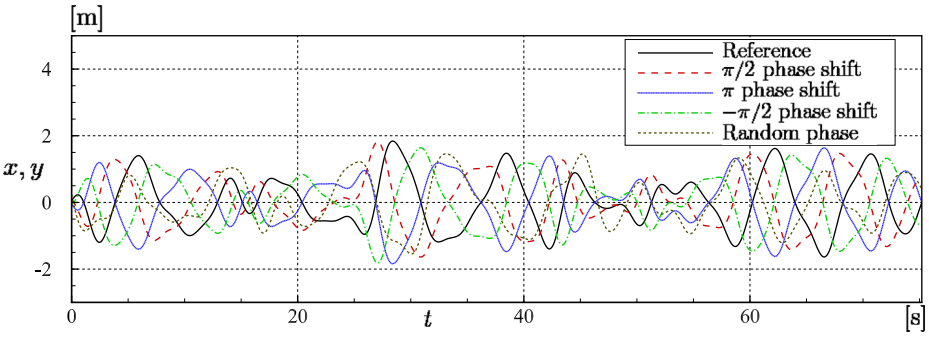
As the first 4 benchmarks (cases 1 to 4), the errors due to the phase difference are considered. Test data $\boldsymbol{x}$ are generated by shifting phases of the reference $\boldsymbol{y}$: $+\pi/2$ phase shift for case 1, $+\pi$ for case 2, $-\pi/2$ for case 3, and random phase for case 4 (different random seeds are used from $\boldsymbol{y}$). On the other hand, cases 5 and 6 change their amplitudes from the reference: $\boldsymbol{y}$ is multiplied by 0.8 for case 5, and $\boldsymbol{y}$ is multiplied by 5/3 for case 6. For case 7, 0.1 m is added to the reference $\boldsymbol{y}$; the mean value is shifted from 0 to 0.1. The first 10 waves of these data are shown in Fig. 1 where cases 1 to 4 are shown in Fig. 1(a) and cases 5 to 7 are in Fig. 1(b).

4.1.1. Discussion on phase differences (cases 1 to 4)

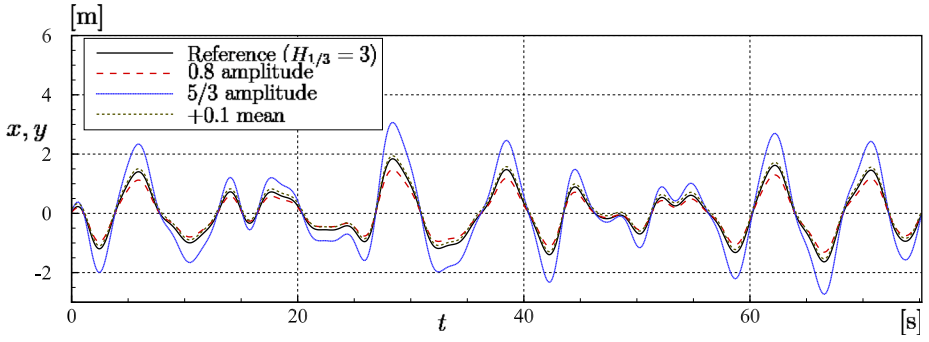
Table 3 shows the evaluated performance values. These are evaluated using Pearson correlation coefficient C_r, MAE, RMSE, ME, SD, MAAPE, FS-MAE, FA-MAE, FA-MAPE, FP-MAE, and FP-ME. The bold values represent values that can identify the cause of the difference between reference and test data.

Table 3
Performance metric for benchmark sets. The values needed when identifying the causes of errors are shown in bold.

Case	y	x	C_r $[-1, 1]$	MAE $[0, \infty]$	RMSE $[0, \infty]$	ME $[0, \infty]$	SD $[0, \infty]$	MAAPE $[0, \pi/2]$	FS-MAE $[0, \infty]$	FA-MAE $[0, \infty]$	FA-MAPE $[0, \infty]$	FP-MAE $[0, \pi]$	FP-ME $[-\pi, \pi]$
1	PM irregular waves	$+\pi/2$ pha.	**0.00**	0.94	1.16	0.00	1.16	0.91	0.01	0.00	0.03	**1.57**	**1.57**
2		$+\pi$ pha.	**−1.00**	1.33	1.65	0.00	1.65	1.11	0.00	0.00	0.00	**3.14**	**3.14**
3		$-\pi/2$ pha.	**0.00**	0.94	1.16	0.00	1.16	0.90	0.01	0.00	0.03	**1.57**	**−1.57**
4		Random pha.	**0.01**	0.88	1.10	0.00	1.10	0.88	0.41	0.03	0.35	**1.51**	**−0.13**
5		×0.8 amp.	1.00	0.13	0.16	0.00	0.16	0.20	0.27	0.02	**0.20**	0.00	0.00
6		×5/3 amp.	1.00	0.44	0.55	0.00	0.55	0.59	1.31	0.06	**0.67**	0.00	0.00
7		+0.1 mean	1.00	0.10	0.10	**0.10**	0.00	0.30	0.00	0.00	0.00	0.00	0.00
8	Finite depth disp.	Deep water disp.	0.97	0.15	0.20	0.00	0.20	0.36	0.00	0.00	0.01	0.14	0.13
9	PM irregular waves	High-freq. noise	0.98	0.13	0.16	0.00	0.16	0.32	0.10	0.01	0.38	0.41	−0.03
10		Low-freq. noise	0.98	0.13	0.16	0.00	0.16	0.33	0.15	0.01	0.17	0.22	0.03
11		$x_i > 1.5$ missing	0.92	0.06	0.33	−0.06	0.32	0.03	0.23	0.02	0.34	0.41	0.01
12	Experiment	Uni prediction	0.73	0.23	0.36	0.00	0.36	0.60	0.02	0.02	0.52	0.59	0.07
13		Bi prediction	0.96	0.11	0.16	0.00	0.16	0.49	0.01	0.01	0.28	0.16	0.00



(a) Cases 1 to 4 (phase difference)



(b) Cases 5 to 7 (amplitude difference and mean drift)

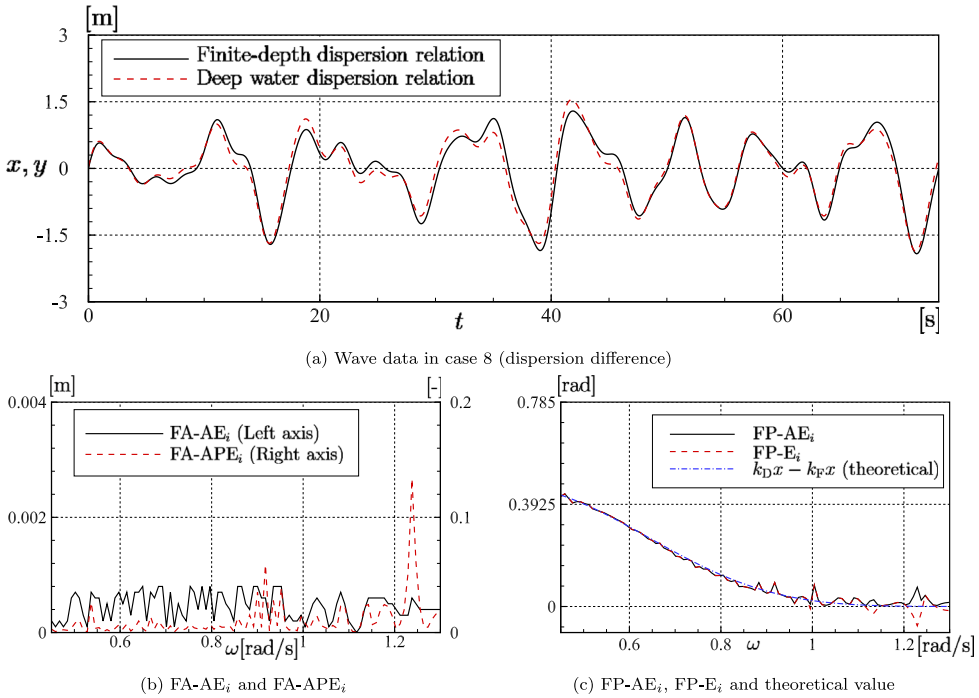
Fig. 1. Reference and test wave data in cases 1 to 7. Waves are based on Pierson–Moskowitz spectrum with mean wave period $T_1 = 8.0$ s and mean wave height $H_{1/3} = 3.0$ m. The first 10 waves of reference data are shown although the full time length is 100 waves.

The differences in cases 1 to 4 are due to phases. Values of FS-MAE, FA-MAE, and FA-MAPE are almost 0 except for case 4; wave amplitudes of test data are the same as the reference. On the other hand, C_r, FP-MAE, and FP-ME show errors for these cases. The correlation coefficient indicates that 0.00 for $\pm\pi/2$ gap, and $C_r = -1.00$ for π gap. Similarly, FP-MAE shows $1.57(\approx \pi/2)$ for $\pm\pi/2$ gap, and $3.14(\approx \pi)$ for π gap. These values indicate phase-originated errors. Nevertheless, these metrics evaluate only phase gap magnitude, and direction (i.e. pulse or minus) is not a concern. FP-ME shows a direct phase difference value as seen in cases 1 to 3. As for case 4, test data have different random phases from the reference data, but their amplitudes are the same. $C_r = 0.01(\approx 0)$ and FP-MAE$= 1.51(\approx \pi/2)$ indicate the possibility of the random phase, but this cannot be distinguished from the case of $\pm\pi/2$ phase shift. To identify whether random or $\pm\pi/2$, FP-ME needs to be checked. FP-ME is zero when the phase is exactly the same or random. However, FP-ME$= -0.13$ implies that the phases are neither $\pm\pi/2$ difference nor perfectly random. In addition, looking at FA-MAE and FA-MAPE, these have non-zero values. These are because the raw Fourier transformed data are used and any smoothing technique is not applied. Therefore, the obtained spectra are spiky and these result in apparent errors. In addition, leakage error occurs for the test data since the same time range as $\boldsymbol{y}$ is used. It could be an option to preprocess the data using a window function and smoothing technique before comparisons. Discussion on smoothing will be shown in Section 4.2.2

4.1.2. Discussion on amplitude differences and mean drift (cases 5 to 7)

The differences in cases 5 and 6 are due to the scales of amplitudes. Therefore, no phase differences are indicated as $C_r = 1$, FP-MAE$= 0$, and FP-ME$= 0$. FA-MAPE identifies these scale-originated amplitude difference: theoretical values are FA-MAPE$= |(0.8 - 1.0)/1.0| = 0.2$ for case 5, and FA-MAPE$= |(5/3 - 1.0)/1.0| = 0.67$ for case 6. The results in Table 3 show identical values. The feature of amplitude differences is whether scale or different relations could be determined by seeing FA-AE$_i$ and FA-APE$_i$ in the frequency domain although this graph is not shown here. Note that $C_r = 1$ when the phase difference is zero, and thus C_r cannot identify the difference in amplitude.

As for case 7, the difference is due to adding a constant value of 0.1 to $\boldsymbol{x}$. This is nothing related to wave components, and thus $C_r = 1$, FA-MAE=0, FA-MAPE=0, FP-MAE=0, and FP-ME=0. Note that the mean



(a) Wave data in case 8 (dispersion difference)

(b) FA-AE$_i$ and FA-APE$_i$

(c) FP-AE$_i$, FP-E$_i$ and theoretical value

Fig. 2. Reference and test wave data in case 8. The reference data is phase-shifted by finite-depth dispersion relation whereas the test data is calculated by deep water dispersion relation. Distance is $x = 50$ m, and water depth $h = 30$ m is used for the finite-depth dispersion waves. (a) The first 10 waves of reference data are shown although the full time length is 100 waves. (b) Distributions of FA-AE$_i$ and FA-APE$_i$ in frequency domain. (c) Distributions of FP-AE$_i$ and FP-APE$_i$ in frequency domain. Theoretical phase difference is also plotted. Cutting-off frequencies are $\omega_{\min} = 0.45$ rad/s and $\omega_{\max} = 1.3$ rad/s.

drift could be appeared at $\omega = 0$ within the frequency spectrum, but this frequency is excluded by applying cutting-off frequency $\omega_{\min}$. ME=0.1 identifies this mean drift. As other conditions are the same, MAE and RMSE also become 0.1.

Summarizing above, C_r, FP-MAE, and FP-ME identify the difference of phase (i.e. horizontal/time translation). To distinguish phase gaps of $\pm\pi/2$, 0, and random, both FP-MAE and FP-ME need to be checked. FA-MAPE identifies the difference of scale of amplitude (i.e. vertical scale) as this metric is percentage type. ME identifies the difference of mean value (i.e. vertical translation) as this metric is magnitude type. MAE, RMSE, and MAAPE are metrics that evaluate the overall concordance between two data. Compared to cases 1 to 7, MAE and RMSE rank these data in the order of cases 7, 5, 6, 4, 3, 1, and 2 (cases 1 and 3 are almost equal). On the other hand, MAAPE ranks them in the order of cases 5, 7, 6, 4, 3, 1, and 2. Among these cases, the first and second changes depending on the metrics.

4.2. Frequency-dependent errors due to noises and dispersion (cases 8 to 10)

Causes of differences between cases 1 to 7 can be identified by using mean performance metrics as shown in Table 3. However, these are not always sufficient for identifying the causes of errors since some errors depend on frequency. Cases 8 to 10 correspond to these frequency-dependent errors.

For case 8, two waves related to different dispersion relations are considered. Waves are generated by applying phase shift $(\omega t - kx)$ to irregular waves used for the reference of cases 1 to 7. Here, k is a wave number, and x is the distance where $x = 50$ m is used. For the reference data $\boldsymbol{y}$, we consider the finite-depth dispersion relation $\omega^2/g = k_F \tanh k_F h$ where g is gravitational acceleration, $h = 30$ m is water depth, and k_F is a wave number for finite-depth dispersion. On the other hand, the compared data $\boldsymbol{x}$ is calculated by deep water dispersion relation $\omega^2/g = k_D$ where k_D is a wave number for deep water dispersion. These waves are shown in Fig. 2(a) (10 waves are plotted against 100 waves).

Moreover, we consider the frequency-dependent noises for cases 9 and 10. The reference data $\boldsymbol{y}$ is the same as those of cases 1 to 7. For the test data $\boldsymbol{x}$, high and low-frequency noises are added to the reference $\boldsymbol{y}$. The noise spectra are given as

$$\boldsymbol{\Phi}(\omega) = ae^{b(\omega - c)} \tag{35}$$

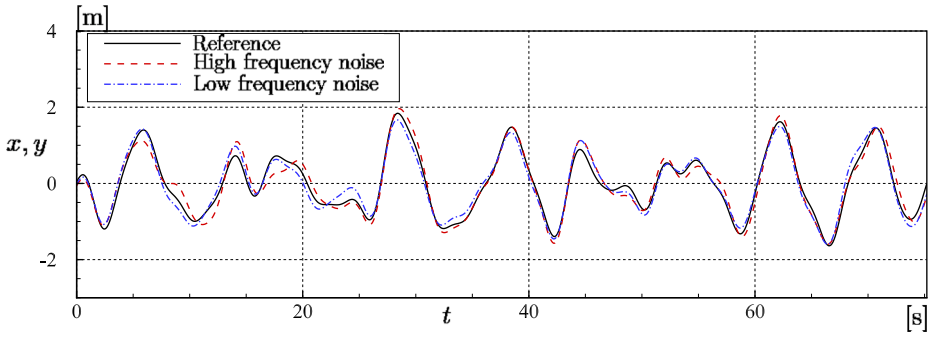
Here, $a = 0.1$, $b = 4.0$, and $c = \omega_{\max} = 1.3$ are used for the high-frequency noise (case 9), and $a = 0.1$, $b = -4.0$, and $c = \omega_{\min} = 0.45$ are used for the low-frequency noise (case 10). The first 10 waves are shown in Fig. 3(a) against 100 waves.

4.2.1. Discussion on different dispersion (case 8)

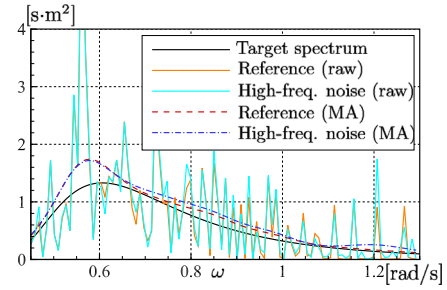
Values of performance metrics for case 8 are shown in Table 3. As FA-MAE= 0.00 and FA-MAPE=0.01, the difference is not by amplitude. On the other hand, FP-MAE=0.14 and FP-ME=0.13 imply the difference in phase. However, the cause could not be identified by these values. Therefore, FA-AE$_i$, FA-APE$_i$, FP-AE$_i$, and FP-E$_i$ are investigated to visualize the tendency for frequencies. Results of FA-AE$_i$ and FA-APE$_i$ are shown in Fig. 2(b), and the results of FP-AE$_i$ and FP-E$_i$ are shown in Fig. 2(c). Here, theoretical phase difference $k_D x - k_F x$ is also plotted in Fig. 2(c). The deep water assumption is valid when the water depth satisfies $h \geq \lambda/2$; frequency satisfies $\omega \geq 1.01$ rad/s. FP-AE$_i$ and FP-E$_i$ indicate that waves by deep water dispersion agree with that by finite depth dispersion for these frequencies. On the other hand, the phase difference increases as the frequency becomes small. As shown in Fig. 2(c), such the phase difference coincides with the theoretical phase difference. As a result, it is identified that the cause of the difference is due to the dispersion relation.

4.2.2. Discussion on frequency-dependent noises (cases 9 and 10)

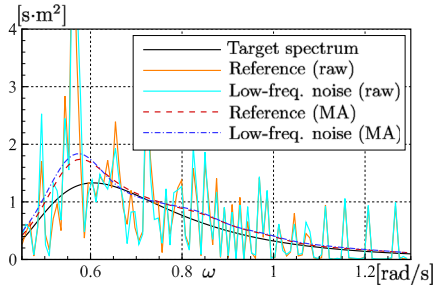
Here, we consider the spectrum errors instead of amplitude errors. To investigate the difference in the spectrum, the obtained spectra are shown in Figs. 3(b) and 3(c), respectively. In addition, results of the FS-AE$_i$ are also plotted in Figs. 3(d) and 3(e). Figs. 3(b) and 3(d) correspond to the results of the high frequency noise, and Figs. 3(c) and



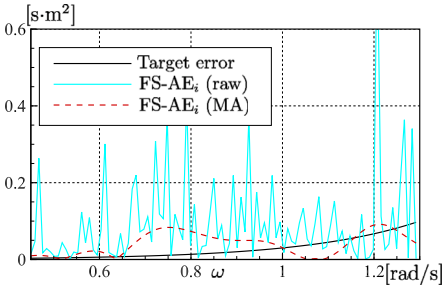
(a) Wave data in cases 9 and 10 (frequency-dependant noises)



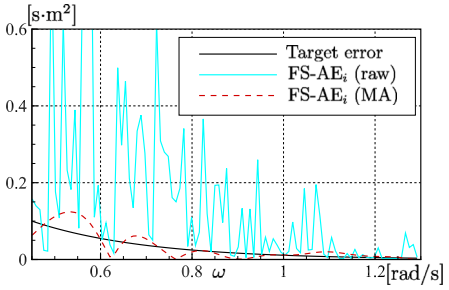
(b) Spectra for case 9 (high-frequency noise)



(c) spectra for case 10 (low-frequency noise)



(d) FS-AE$_i$ for case 9 (high-frequency noise)



(e) FS-AE$_i$ for case 10 (low-frequency noise)

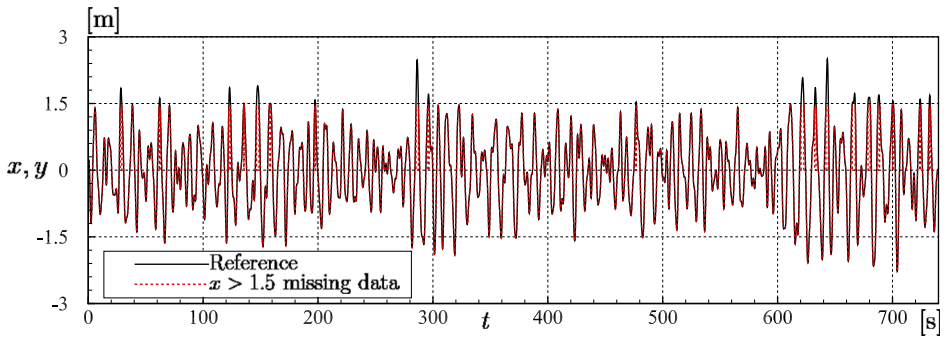
Fig. 3. Reference and test wave data in cases 9 to 10. (a) The first 10 waves of reference data are shown against 100 waves. (b) and (c) Spectra of time-series data. (d) and (e) Distributions of FP-AE$_i$. Cutting-off frequencies are $\omega_{\min}$ = 0.45 rad/s and $\omega_{\max}$ = 1.3 rad/s. The raw results and smoothed results are shown where 3 points centered moving average is applied 50 times to obtain the smoothed spectra.

3(e) are for the high-frequency noise. Target wave spectrum (i.e. (34)) and noises (i.e. (35)) are also put in the figures. Moreover, the results of raw data and smoothed data are compared where the smoothed data are labeled MA (moving average). Smoothed spectra are calculated by iterating 50 times of 3 points centered moving average. The FS-AE$_i$ is calculated by the smoothed spectra. Note that the results of Table 3 are estimated using raw data. Both cases show good agreement with target errors. Table 3 indicates that the comparisons of raw data show FS-MAE=0.10 and 0.15 for high and low noise cases, respectively. On the other hand, the comparisons of smoothed data show FS-MAE=0.04 and 0.03 for these cases, that is, the smoothing levels the errors.

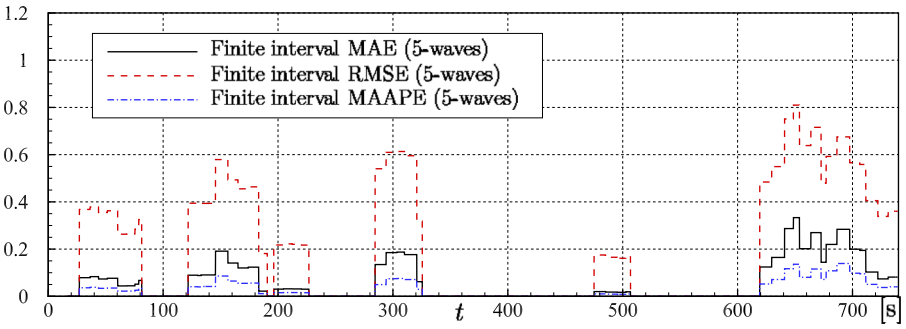
4.3. Errors in time series data (cases 11 to 13)

In order to consider the cases whose causes of differences are related to time-series, cases 11 to 13 are considered. Case 11 uses the same irregular wave data as cases 1 to 7 for the reference $\boldsymbol{y}$. Test data $\boldsymbol{x}$ is almost the same as $\boldsymbol{y}$, but the data are set to 0 when $x_i > 1.5$. This imitates the missing data where such phenomenon sometimes occurs when values exceed the sensor's measurable range or sensors are attached in the vicinity of the water line of an offshore structure. These data are shown in Fig. 4(a). Here, 100 waves are considered and all data are shown.

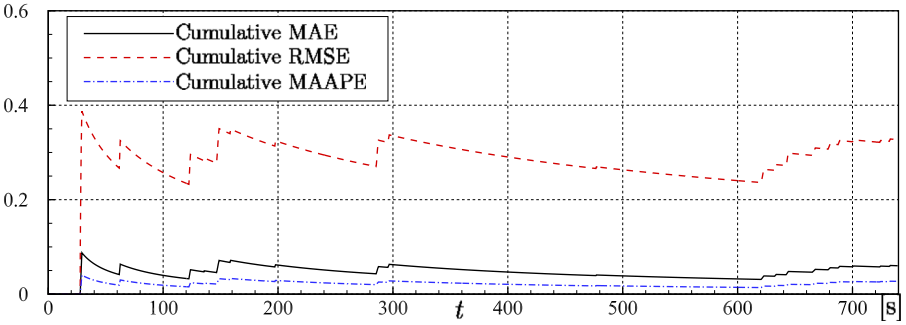
Furthermore, we consider a case where the error trend changes over time. Cases 12 and 13 are related to such a situation. The reference data of cases 12 and 13 are experimentally measured time-series data of water surface elevation. The experiment was carried out in a two-dimensional wave tank at Osaka University, Japan (Iida, 2023). The tank length is 14 m, and irregular waves based on the JONSWAP spectrum (Stansberg et al., 2002) (the significant wave period $T_{1/3}$ = 1.2 s and wave height $H_{1/3}$ = 1.5 cm) were generated by a wave maker at the one end of the tank (this is 1/100 length and 1/10 time scale experiment against real ocean problems). Since the length of the tank is finite-length, waves are reflected by another end. Therefore, measured data contains not only progressive waves but also regressive waves after reflected waves reach the measurement position. Here, such waves are predicted by two methods. The one is based on a uni-directional prediction method (Iida and Minoura, 2022) which predicts waves using the information of up-wave position. The accuracy of this prediction method should become worse after reflected waves reach the measured position. Another method is based on a bi-directional prediction method (Iida, 2023) which predicts waves using information on both up-wave and down-wave positions. Because this method can consider both progressive and regressive wave components, this prediction method should keep the accuracy level. Both methods are based on the same linear finite-depth water wave impulse response function, and thus the output waves are calculated by the convolution integral of the impulse response function and input wave time-series data at the measured positions. The difference is whether or not the data at the



(a) Wave data in case 11 (missing data)



(b) Finite interval performance metrics (5-wave mean is used)



(c) Cumulative performance metrics

Fig. 4. Reference and test data in case 11. The missing data is set to 0 when the value of x is bigger than 1.5. (a) Time history of 100 zero-up cross waves. (b) Finite interval performance metrics (5-waves). (c) Cumulative performance metrics.

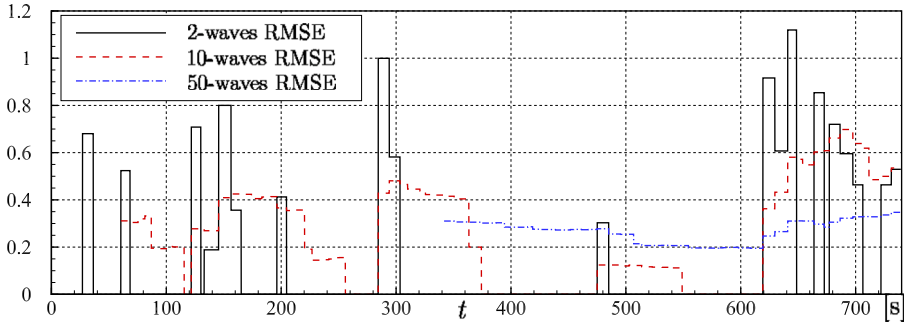


Fig. 5. Comparison of different intervals of finite interval RMSE for case 11.

down-wave position was used in combination with the up-wave position's data. Case 12 denotes the comparison between the experimental data and estimated data using the uni-directional prediction method. Case 13 is the comparison between the experimental data and estimated data using the bi-directional prediction method. Data were measured 60 s, and measured and predicted results are shown in Fig. 6(a). Note that wave data are smoothed for this paper, and thus the results are not exactly the same as (Iida, 2023). To calculate performance metrics, cutting-off frequencies $\omega_{\min} = 3$ rad/s and $\omega_{\max} = 12$ rad/s are used.

4.3.1. Discussion on missing data (case 11)

Table 3 indicated that RMSE=0.33 shows an error of non-negligible magnitude While MAE=0.06 and MAAPE=0.03 indicate good concordance. To investigate the difference with respect to time, finite-time

interval and cumulative performances are calculated. Here, 5 waves are used for the finite-time interval. 5-wave MAE, RMSE, and MAAPE are plotted in Fig. 4(b), and cumulative MAE, RMSE, and MAAPE are plotted in Fig. 4(c). Looking at Fig. 4(b), the performances deteriorate after data missing occurs. These magnitudes are particularly high for the 5-wave RMSE. On the other hand, the 5-wave MAAPE is insensitive to data missing. This is because weighting to outlier is different for these metrics. As discussed in Section 2. 8, the MAAPE is represented by the arctangent, and thus larger values converge to $\pi/2$. On the other hand, the RMSE evaluates larger values to a greater extent as this is represented as a square. Due to this difference, results of cumulative performances show different tendencies. As the MAAPE is insensitive to the outlier, the cumulative MAAPE is almost 0, and the final value (i.e. value in Table 3) is MAAPE=0.03. On the other hand, the error of the cumulative RMSE dramatically increases after the first outlier even though data $\boldsymbol{x}$ is the same as data $\boldsymbol{y}$ except for data missing zones. The MAE does not weigh for any data, and results are in moderation. These results indicate that the RMSE is not a good metric to evaluate overall performance when outliers are included. The MAAPE is insensitive to outliers, and thus the MAAPE is preferable if users want to avoid the influence of outliers. On the other hand, if users want to detect outliers, the finite-interval RMSE is a good metric as this is sensitive to outliers.

To investigate the influence of the range of the finite interval metrics on the performance evaluation, different ranges of intervals are plotted in Fig. 5. 2-waves, 10-waves, and 50-waves RMSE are shown. The shorter the interval, the sharper the peak. On the other hand, the longer interval results in a broader base around the peak. An appropriate interval range depends on the phenomenon to be detected. When the longer interval is used, this becomes a similar result of cumulative metric (see the 50-wave RMSE in Fig. 5 and the cumulative RMSE in Fig. 4(c)). Since cumulative representation needs computational costs for long time-series data, sufficiently long finite interval representation can be an alternative to the cumulative representation.

4.3.2. *Discussion on trend change due to wave reflection (cases 12 and 13)*

To investigate the error trend in time, cumulative performance metrics are shown in Figs. 6(b) and 6(c) where Fig. 6(b) is for the uni-directional prediction method and Fig. 6(c) is for the bi-directional prediction method. Cumulative C_r and RMSE in Fig. 6(b) indicate that the accuracy of the prediction becomes worse after $t = 26$ s. On the other hand, the results of Fig. 6(c) do not show such a tendency. Because the difference between the two prediction methods is whether regressive waves are considered or not, it can be identified that the reflected waves reach the measured position at this time. The cumulative RMSE emphasizes such differences as this weighs more for excessive errors.

Looking at Table 3, the amplitude is moderate as FA-MAPE= 0.28 for the bi-directional prediction. To investigate this error, the frequency domain's amplitudes are plotted in Fig. 7. Fig. 7(a) shows amplitudes $2|X|/N$ and $2|Y|/N$, and these FA-AE$_i$ and FA-APE$_i$ are displayed in Fig. 7(b). These indicate that errors are around $\omega = 5.34$ rad/s and $\omega > 10$ rad/s. The difference in $\omega > 10$ rad/s could be a result of the almost calm surface of the time-series data until $t < 10$ s where waves had not reached the wave probes and noise is dominant. The frequency $\omega = 5.34$ rad/s corresponds to non-dimensional frequency $\omega' = \omega\sqrt{h/g} = 5.34\sqrt{0.45/9.81} = 1.14$ where finite-depth dispersion relation is dominant. Since these prediction methods are based on the analytical solution of the impulse response function, the finite-depth dispersion is approximated. As a result, it is demonstrated that the prediction accuracy of amplitude deteriorates around such a frequency (see Fig. 3 of Iida, 2023). Therefore, the results are reasonable under the limitation of the theory.

5. Example of the procedure to identify error causes

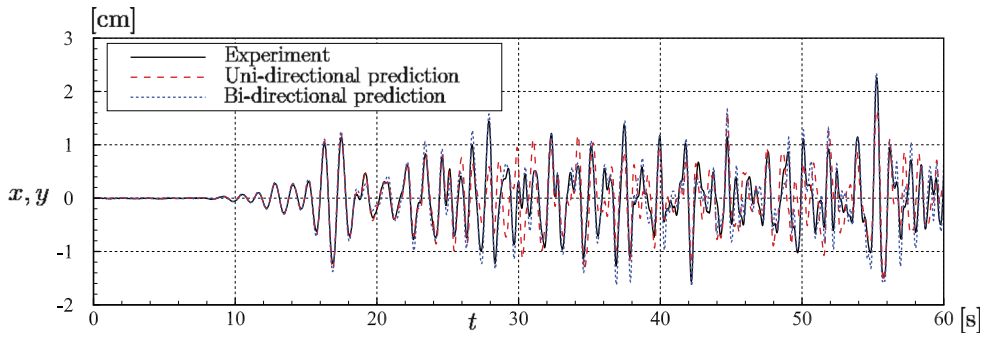
We investigated some typical cases of comparisons of time-series data related to wave phenomena. Since we already know the causes of these errors, we showed the minimum results necessary to identify the causes. However, the causes of errors are unknown for the real problems. Therefore, an example of an estimation procedure is summarized as follows:

1. Check error characteristics in time
   Firstly, it is recommended to check the data's stationarity in time. Inputs outside valid measurement range or instrument malfunctions can result in missing or deformed data. Stationarity could be also broken by changes in physical conditions, such as ship speed change, wind trend change, and wave reflection (in a tank experiment). These errors deteriorate overall performances. To detect such time transition, finite interval and cumulative metrics are useful, e.g. finite interval MAE and cumulative MAE.
2. Check performance metrics over all time-series data
   If data is stationary, overall performances are evaluated using all time-series data. The correlation coefficient C_r can evaluate the phase concordance. Calculating the ME is important to check the drift of the data and estimation. When the ME is zero, the RMSE becomes a standard deviation. To evaluate time-series concordance, a combination of the MAE, RMSE, and the MAAPE is preferable because these have different features; the MAE is scale-dependent and weight-free for errors, the RMSE is scale-dependent and outlier-sensitive, and the MAAPE is scale-independent and small-error-sensitive. When comparing more than two data, the check of the ranking order of these metrics could be helpful to discuss the difference in error trends.
3. Check frequency domain's amplitude and phase
   If an error exists in the overall performance metrics, it is necessary to check the errors in both amplitude and phase, respectively. Firstly, mean performances, FA-MAE, FA-MAPE, FP-MAE, and FP-ME, are calculated, and the likely source of errors is determined. After that, such an error (such as FA-AE$_i$) is displayed in the frequency domain. It would be better to first calculate the values in the frequency domain from the raw data to check for error trends. Then, how smoothing and cut-off frequencies will be applied are decided, and performance metrics are calculated again.
4. Connect the error causes to the physical/theoretical reasons
   Once a source of errors is determined, we need to consider the physical/theoretical reason for the errors. It is necessary to check whether the data is measured/predicted within the applicable range of the measurement apparatus/theory. In some cases, the target phenomena occur beyond the measurable sampling time, range, and theoretical assumptions (such as linearity/weak non-linearity, harmonics, incompressibility, inviscidity, and so on).

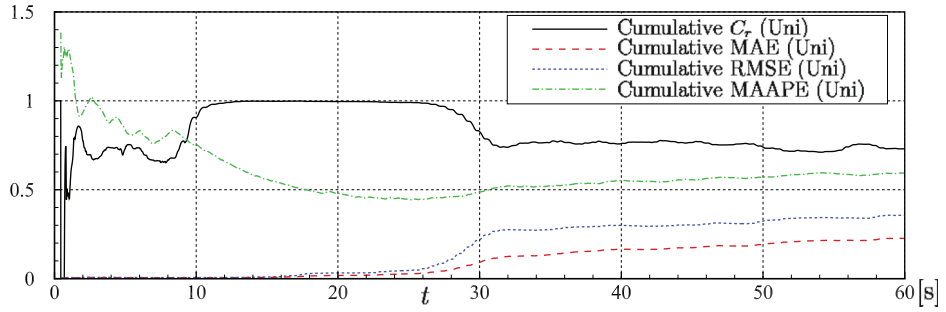
6. Discussion on applicability and limitations

As we are interested in real-time wave predictions, our primary targets are wave-related time-series data ranging from a few seconds to minutes. Therefore, the proposed performance metrics could be applicable to real-time wave predictions (e.g. Al-Ani et al., 2020; Law et al., 2020; Iida and Minoura, 2022; Law et al., 2022; He et al., 2023), predictions of ship and platform's motions (e.g. Naaijen et al., 2018; Liu et al., 2020; Lee et al., 2022; Liong and Chua, 2022; Lee et al., 2023; Kinugasa et al., 2023; Isnaini et al., 2024), wave load (e.g. Ren et al., 2023), energy harvesting of WEC (e.g. Halliday et al., 2005; Korde, 2017), and so on. The proposed metrics could facilitate improving these predictions.

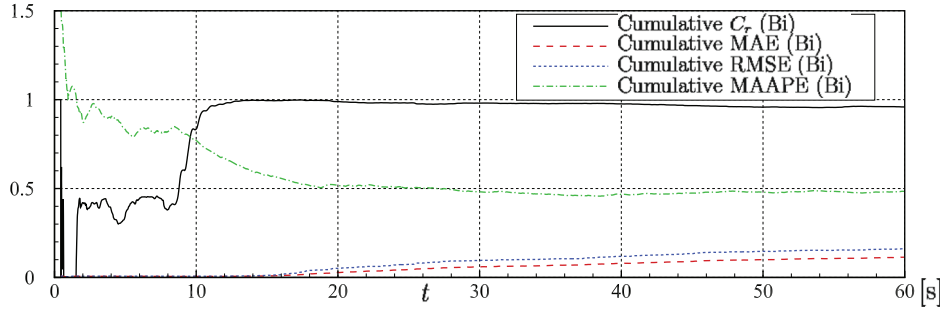
On the other hand, the frequency domain's metrics (i.e. FA- and FP-) might be inadequate for predictions of wave peak (e.g. Kagemoto,



(a) Wave data in cases 11 and 12

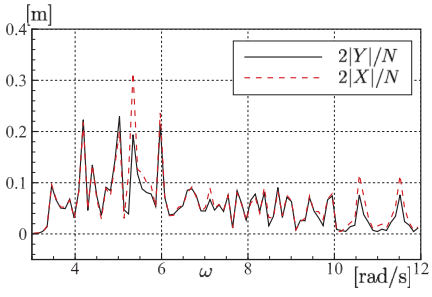


(b) Cumulative performance metrics for uni-directional prediction

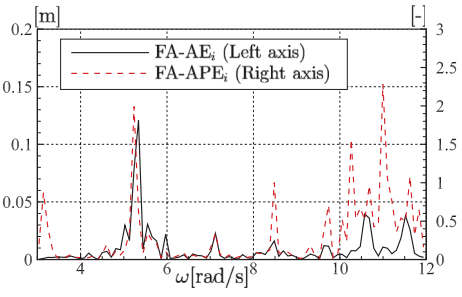


(c) Cumulative performance metrics for bi-directional prediction

Fig. 6. Reference and test data in cases 12 and 13. The reference data is experimentally measured water surface whereas test data are predicted surfaces based on both uni-directional and bi-directional predictions. The reference data includes progressive and regressive wave components. (a) Time history of the surface elevation. (b) Cumulative performance metrics in case 12 (uni-directional prediction). (c) Cumulative performance metrics in case 13 (bi-directional prediction).



(a) Frequency domain's amplitude



(b) FA-AE$_i$ and FA-APE$_i$

Fig. 7. Frequency domain's amplitude and errors in case 13 (results between experiment and bi-directional prediction method).

2020, 2022) and significant wave height (e.g. Fan et al., 2020; Jörges et al., 2021) since these do not contain frequency information. Finite interval and cumulative metrics could be applied to these data to check stationarity in time. Note that the time-series of significant wave height does not cross zero, and the mean value is also important. In addition, the frequency domain's metrics are also inadequate for predictions of strongly nonlinear phenomena, such as slamming (e.g. Terada et al., 2017).

In this paper, we only consider the time-series data and its frequency component, i.e. t and ω. However, the proposed method can be used

for spatial data, such as snapshots of the free surface, and this is transformed into the wave number k domain. In addition, space–time data could be also evaluated by extending the proposed method. This means that this method is applicable to directional spectrum evaluation, i.e. multi-directional wave predictions (e.g. Belmont et al., 2006, 2014; Wang et al., 2021; Kim et al., 2024).

Direct time series metrics are applicable to any time-series data regardless of the applicability of the frequency domain's metrics. However, cumulative representation could require a huge amount of data as time passes. This tendency is especially excessive when the frequency domain's metrics are combined with cumulative ones, e.g. Cumulative FA-MAE. In such a case, finite-interval representation is recommended.

When the data are based on physics or mathematics, identification of a cause of errors could be possible as the behind mechanics of the data may be obvious. On the other hand, it is not easy to interpret the error of the data predicted by AI technologies where these attempts are rapidly increasing recent years, (e.g. Fan et al., 2020; Kagemoto, 2020; Jörges et al., 2021; Lee et al., 2023; Kinugasa et al., 2023). It is thought that insufficient training data in the short and long-wavelength regions, where these corresponding spectral densities are relatively small, could reduce the accuracy of prediction in these regions. It seems also difficult to predict wave phases whose behavior is likely to be random. These errors could be visualized by evaluating spectrum and phase in the frequency domain, i.e. FS- and FP-. It is worth noting that the proposed metrics could be used not only for the evaluation of errors but also for loss functions in the machine learning process, such as Wedler et al. (2022).

In this paper, we are not able to demonstrate the comparison using real ocean wave time-series data. Since real-time wave prediction methods are still developing, most of the comparisons are made using laboratory experiments. We expect these will be applied to real ocean wave problems soon, and then the proposed performance metrics will really be needed.

7. Conclusion

As the quantitative evaluation is important to compare and discuss errors of two time-series data related to waves, we studied how to use performance metrics. Firstly, existing performance metrics, R-squared, Pearson correlation coefficient, mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared error (MSE), root mean squared error (RMSE), normalized RMSE, mean arctangent absolute percentage error (MAAPE), and surface similarity parameter (SSP), are reviewed. Each metric has inherent features, and none is the universally best. It is important to understand these features and choose adequate metrics for a target problem.

These metrics are used to calculate the mean value of all time-series data. As a result, we could not identify causes of errors using such a mean value as all errors are all together regardless of causes. However, it is essential to identify the causes of errors to understand the limitations of the performance and improve accuracy. Therefore, we proposed new representations of performance metrics to be able to separate errors by causes. Since wave-related data can be transformed to the frequency domain by the Fourier transform, errors in the frequency domain's amplitude and phase are independently evaluated. Here, two prefixes, the frequency domain's amplitude (FA-) and frequency domain's phase (FP-), are applied to existing performance metrics, e.g. FA-MAE, FA-MAPE, FP-MAE, and FP-ME. In addition, to deal with instantaneous errors and changes in a trend of errors over time, means of two different intervals are defined whereas original performance metrics are calculated by all data. Finite interval representation evaluates the mean error at time i using preceding m data. On the other hand, cumulative representation evaluates the mean error at time i using all data until i. By combining these new representations of metrics with existing metrics, we can examine the error causes more deeply.

For the demonstrations, some examples of comparisons were shown. Example benchmarks cover typical causes of errors, such as phase difference, amplitude difference, mean drift, different dispersion, frequency noises, data missing, and time trend change due to wave reflection. In all cases, the causes of errors can be identified using proposed performance metrics. Since we already know the cause of these errors, the effort of the investigations could be minimal. However, sources of errors are generally unknown for the real problems, and thus we summarized the estimation procedure. Firstly, we need to check the error characteristics in time. Secondly, performances are estimated using a set of metrics for overall time-series data. After that, performances in the frequency domain need to be checked. When a source of errors is determined, we need to connect the error sources to the physical/theoretical reasons.

We believe our study facilitates the appropriate comparisons of wave-related data and improves the accuracy of time-series prediction methods in the field of ocean engineering.

CRediT authorship contribution statement

Takahito Iida: Writing – review & editing, Writing – original draft, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Takahito Iida reports financial support was provided by JSPS KAKENHI Grant Number JP23K13504.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by JSPS KAKENHI, Japan Grant Number JP23K13504.

References

Al-Ani, M., Belmont, M., Christmas, J., 2020. Sea trial on deterministic sea waves prediction using wave-profiling radar. Ocean Eng. 207, 107297.
Allen, J.B., Rabiner, L.R., 1977. A unified approach to short-time Fourier analysis and synthesis. Proc. IEEE 65 (11), 1558–1564.
Bartlett, M.S., 1948. Smoothing periodograms from time-series with continuous spectra. Nature 161 (4096), 686–687.
Belmont, M., Christmas, J., Dannenberg, J., Hilmer, T., Duncan, J., Duncan, J., Ferrier, B., 2014. An examination of the feasibility of linear deterministic sea wave prediction in multidirectional seas using wave profiling radar: Theory, simulation, and sea trials. J. Atmos. Ocean. Technol. 31 (7), 1601–1614.
Belmont, M., Horwood, J., Thurley, R., Baker, J., 2006. Filters for linear sea-wave prediction. Ocean Eng. 33 (17–18), 2332–2351.
Chai, T., Draxler, R.R., 2014. Root Mean Square Error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geosci. Model Develop. 7 (3), 1247–1250.
Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E., 2008. Querying and mining of time series data: Experimental comparison of representations and distance measures. Proc. VLDB Endow. 1 (2), 1542–1552.
Fan, S., Xiao, N., Dong, S., 2020. A novel model to predict significant wave height based on long short-term memory network. Ocean Eng. 205, 107298.
Halliday, J., Dorrell, D., Wood, A., 2005. The application of short-term deterministic wave prediction to offshore electricity generation. In: International Conference on Renewable Energies and Power Quality.
He, J., Chen, Z., Zhao, C., Chen, X., Wei, Y., Zhang, C., 2023. A robust scheme for deterministic sea wave reconstruction and prediction using coherent microwave radar. IEEE Trans. Geosci. Remote Sens..
Hodson, T.O., 2022. Root-Mean-Square Error (RMSE) or Mean Absolute Error (MAE): When to use them or not. Geosci. Model Dev. 15 (14), 5481–5487.

Iida, T., 2023. Decomposition and prediction of initial uniform bi-directional water waves using an array of wave-rider buoys. Renew. Energy 119137.
Iida, T., Minoura, M., 2022. Analytical solution of impulse response function of finite-depth water waves. Ocean Eng. 249, 110862.
Isnaini, R., Tatsumi, A., Iijima, K., 2024. Future predictions of wave and response of multiple floating bodies based on the Kalman filter algorithm. J. Ocean Eng. Mar. Energy 10 (1), 137–154.
Jörges, C., Berkenbrink, C., Stumpe, B., 2021. Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks. Ocean Eng. 232, 109046.
Kagemoto, H., 2020. Forecasting a water-surface wave train with artificial intelligence-A case study. Ocean Eng. 207, 107380.
Kagemoto, H., 2022. Forecasting a water-surface wave train with artificial intelligence (Part 2)–Can the occurrence of freak waves be predicted with AI? Ocean Eng. 252, 111205.
Kim, I.-C., Ducrozet, G., Leroy, V., Bonnefoy, F., Perignon, Y., Bourguignon, S., 2024. A real-time wave prediction in directional wave fields: Strategies for accurate continuous prediction in time. Ocean Eng. 291, 116445.
Kim, S., Kim, H., 2016. A new metric of absolute percentage error for intermittent demand forecasts. Int. J. Forecast. 32 (3), 669–679.
Kinugasa, M., Yamamoto, Y., Terada, D., Katayama, T., 2023. LSTM-based model predictive control for motion stabilized platform of Doppler LiDAR for offshore wind observation. In: International Conference on Offshore Mechanics and Arctic Engineering, vol. 86861, American Society of Mechanical Engineers, V004T05A020.
Korde, U.A., 2017. Wave-by-wave control of a wave energy converter with deterministic wave prediction. In: European Wave and Tidal Energy Conference, vol. 2017.
Kvålseth, T.O., 1985. Cautionary note about R 2. Amer. Statist. 39 (4), 279–285.
Law, Y.Z., Santo, H., Chua, K.H., 2022. Wave-field prediction based on radar snapshots taken on a moving vessel. J. Phys.: Conf. Ser. 2311, 012022.
Law, Y., Santo, H., Lim, K., Chan, E., 2020. Deterministic wave prediction for unidirectional sea-states in real-time using artificial neural network. Ocean Eng. 195, 106722.
Lee, J.-H., Lee, J., Kim, Y., Ahn, Y., 2023. Prediction of wave-induced ship motions based on integrated neural network system and spatiotemporal wave-field data. Phys. Fluids 35 (9).
Lee, J.-H., Nam, Y.-S., Kim, Y., Liu, Y., Lee, J., Yang, H., 2022. Real-time digital twin for ship operation in waves. Ocean Eng. 266, 112867.
Lin, J., Keogh, E., Lonardi, S., Chiu, B., 2003. A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. pp. 2–11.
Liong, C., Chua, K., 2022. Data assimilation for deterministic prediction of vessel motion in real-time. Ocean Eng. 244, 110356.
Liu, Y., Duan, W., Huang, L., Duan, S., Ma, X., 2020. The input vector space optimization for LSTM deep learning model in real-time prediction of ship motions. Ocean Eng. 213, 107681.
Naaijen, P., Van Oosten, K., Roozen, K., van't Veer, R., 2018. Validation of a deterministic wave and ship motion prediction system. In: International Conference on Offshore Mechanics and Arctic Engineering, vol. 51272, American Society of Mechanical Engineers, V07BT06A032.
Perlin, M., Bustamante, M.D., 2016. A robust quantitative comparison criterion of two signals based on the Sobolev norm of their difference. J. Engrg. Math. 101 (1), 115–124.
Pierson Jr., W.J., Moskowitz, L., 1964. A proposed spectral form for fully developed wind seas based on the similarity theory of SA Kitaigorodskii. J. Geophys. Res. 69 (24), 5181–5190.
Ren, Z., Han, X., Yu, X., Skjetne, R., Leira, B.J., Sævik, S., Zhu, M., 2023. Data-driven simultaneous identification of the 6DOF dynamic model and wave load for a ship in waves. Mech. Syst. Signal Process. 184, 109422.
Senin, P., 2008. Dynamic Time Warping Algorithm Review, vol. 855, (no. 1–23), Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, p. 40.
Stansberg, C., Contento, G., Hong, S.W., Irani, M., Ishida, S., Mercier, R., Wang, Y., Wolfram, J., Chaplin, J., Kriebel, D., 2002. The specialist committee on waves final report and recommendations to the 23rd ITTC. In: Proceedings of the 23rd ITTC. vol. 2, pp. 505–551.
Terada, D., Amano, R., Katayama, T., 2017. A motion estimation method of high speed craft in irregular sea by using onboard monitoring motion time series data for motion control. In: Proceedings of the 16th International Ship Stability Workshop. ISSW2017, Belgrade, Serbia, pp. 5–7.
Wang, Y., Wang, H., Zhou, B., Fu, H., 2021. Multi-dimensional prediction method based on bi-LSTMC for ship roll. Ocean Eng. 242, 110106.
Wedler, M., Stender, M., Klein, M., Ehlers, S., Hoffmann, N., 2022. Surface similarity parameter: A new machine learning loss metric for oscillatory spatio-temporal data. Neural Netw. 156, 123–134.
Welch, P., 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Tran. Audio Electroacoust. 15 (2), 70–73.
William, H., et al., 2003. Econometric Analysis, fifth ed. New York University.
Willmott, C., Ackleson, S., Davis, R., Feddema, J., Klink, K., Legates, D., O'donnell, J., Rowe, C., 1985. Statistics for the evaluation of model performance. J. Geophys. Res. 90 (C5), 8995–9005.
Willmott, C.J., Matsuura, K., 2005. Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82.