https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

A scale-invariant evaluation framework for geoscientific scalar time-series models: demonstration with wave model validation

Fayun Wang[a,*]

[a] *Hainan Academy of Ocean and Fisheries Sciences, Haikou, 571126, China*

[*] Corresponding author.

E-mail address: wangfy@hnhky.cn (F.Y. Wang)

Abstract: In the validation of geoscientific scalar time-series models, the selection of evaluation metrics is often based on experience and conventions, lacking a systematic design framework. Moreover, commonly used metrics are predominantly dimensional error measures whose values are inherently tied to the magnitude of the variable being assessed, making fair comparisons across different stations or datasets difficult. To address these issues, this study proposes a scale-invariant evaluation framework termed the "Golden Triangle", consisting of the range-normalized root mean square error (NRMSE), the range-normalized mean absolute error (NMAE), and the Pearson correlation coefficient (PCC). It is mathematically proven that all three metrics possess scale invariance, meaning their evaluation results are unaffected by the absolute magnitude of the variable, thereby enabling direct cross-station and cross-dataset comparisons. The three metrics capture extreme error sensitivity, average error level, and trend consistency, respectively, complementing one another to form a multi-dimensional diagnostic capability and providing a systematic design framework for metric selection. An empirical case study using data from four global deep-sea buoys and two wave reanalysis models demonstrates that the "Golden Triangle" framework effectively eliminates the influence of wave height magnitude differences among stations, overcomes the cross-station ranking distortion caused by dimensional dependence, and yields reasonable performance rankings. Furthermore, in the comparison between the full dataset and the high-energy-state dataset, the framework exhibits excellent consistency, confirming its reliability and sensitivity to extreme events. The "Golden Triangle" framework offers a logically coherent and generalizable evaluation tool for the validation of geoscientific scalar time-series models.

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

1. Introduction

Numerical models and data-driven models have become core tools for both research and engineering applications across various geoscience disciplines. From meteorological forecasting and ocean dynamic processes to hydrological routing and environmental quality simulation, the reliability of model results directly determines the scientific validity of research conclusions and the soundness of decision-making. Model validation, as the critical step in assessing model performance, requires a unified and standardized metric system to enable fair comparisons across studies, stations, and models. However, current validation practices in different geoscience branches lack unified specifications regarding the selection of evaluation metrics and normalization strategies. Consequently, the cross-station and cross-study comparability of results remains generally insufficient, and a widely recognized, standardized scalar time-series model validation system has yet to be established.

This issue is equally prominent in the field of ocean wave simulation and prediction. With the mature application of third-generation spectral wave models—such as WAM (WAMDI Group, 1988), SWAN (Booij et al., 1999), and WAVEWATCH III (Tolman, 1991)—and the widespread adoption of machine learning methods in wave height prediction (Deo and Naidu, 1999; Pourzangbar et al., 2023), model validation has become an indispensable part of research. A review of the existing literature reveals that evaluation metrics such as root mean square error (RMSE), mean absolute error (MAE), Pearson correlation coefficient (PCC), coefficient of determination (R^2), scatter index (SI), and bias are widely used. In the field of numerical modeling, for instance, Rusu et al. (2008) selected RMSE, bias, SI, and PCC to validate a WAM-SWAN nested system off the west Iberian coast; Saket and Etemad-Shahidi (2012) employed mean absolute percentage error (MAPE) and PCC to quantify the wave energy simulation performance of the SWAN model in the northern Gulf of Oman; Sun et al. (2021) used RMSE, MAE, bias, SI, and PCC to verify the accuracy of a WW3 model in the Xisha Islands waters; Aydoğan and Ayat (2018) also adopted RMSE and PCC as evaluation metrics in their long-term wave trend study in the Black Sea; Ponce de León and Guedes Soares (2005) relied on RMSE, bias, and PCC to characterize model performance in their numerical study on island sheltering effects; Stopa et al. (2016) employed normalized bias, RMSE, SI,

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

and PCC to compare multiple source term parameterization schemes; Gonçalves et al. (2014) and Vieira et al. (2020) uniformly adopted RMSE, bias, SI, and PCC for wave energy assessment and climate analysis in the Canary Islands and the Persian Gulf, respectively; Qiu et al. (2025) investigated the spatial distribution characteristics of waves around South China Sea island reefs using the SWAN model and completed model validation with RMSE, bias, SI, and PCC. In the field of machine learning, Fan et al. (2020) adopted RMSE, MAE, MAPE, R^2, and Nash-Sutcliffe efficiency (NSE) to evaluate the wave height prediction performance of a LSTM model at ten global stations; Feng and Xu (2026) compared different neural networks using RMSE, MAE, mean square error (MSE), and PCC in storm surge prediction; Xie et al. (2025) combined RMSE, MAE, MSE, MAPE, and R^2 to assess model prediction results; Li et al. (2024) primarily relied on R^2 to evaluate model goodness-of-fit in wave run-up prediction; Wang et al. (2026) selected RMSE, MSE, R^2, and other evaluation metrics in their recent ensemble learning study; Li et al. (2025) employed RMSE and MSE to evaluate the performance of a convolutional network for reconstructing the South China Sea wave field; Xie et al. (2026) applied a causality-mining approach combined with convolutional neural networks to predict significant wave height (Hs) in the South China Sea, using RMSE, MAE, MAPE, and R^2 to quantify the forecast accuracy of various causality-based deep learning models at multiple lead times. These studies collectively demonstrate that such metrics are effective for comparing model performance within a single dataset, as they adequately reflect how well the model reproduces the bulk wave height.

However, the above metric system has notable limitations. On the one hand, RMSE and MAE are dimensional absolute errors whose numerical values inherently depend on the wave height magnitude at the station being evaluated: a nearshore station with a mean Hs of only 0.3 m may yield an RMSE of merely 0.1 m, whereas an open-ocean station with a mean wave height of 2 m may produce an RMSE as high as 0.6 m. This does not directly indicate that the model performs better at the former station, because the relative error levels could be entirely reversed. To eliminate the influence of dimension, SI—the RMSE normalized by the observed mean—has been widely adopted for relative assessment in the community (Ris et al., 1999; Liu et al., 2024). However, the mean only reflects the central tendency of the data and cannot

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

cover the full range of variation or extreme values. When the wave height ranges differ considerably across stations, mean normalization overestimates the SI value, failing to achieve truly fair cross-station comparisons. On the other hand, although PCC and R^2 are dimensionless, they only reflect the degree of linear correlation and are insensitive to systematic bias, thus failing to capture the absolute level of error and making it difficult to fully characterize model performance when used alone.

To address this issue, the range-normalized RMSE (NRMSE) and the range-normalized MAE (NMAE) were adopted in prior studies to conduct fair model comparisons across different stations, validating the applicability of this normalization approach in wave contexts (Wang et al., 2025, 2026). However, that prior work merely employed these metrics as evaluation tools for specific studies, without systematically establishing the mathematical nature of their scale invariance, nor integrating complementary metrics to form a standardized, universal evaluation framework.

In practice, model performance often varies with environmental factors such as water depth and topography, exhibiting substantial differences across different regions. When researchers seek to answer questions such as "at which station does the same model perform better, given different wave height magnitudes?" or "how can modeling results from different studies be benchmarked against one another?", the existing metric system struggles to provide reliable answers. This issue is not unique to the wave domain. In the validation of all types of geoscientific scalar time-series models—including those for temperature, precipitation, water quality, and hydrology—problems such as dimensional dependence, inconsistent normalization practices, and over-reliance on single-type metrics are pervasive, hindering cross-study and cross-regional comparison of results as well as methodological improvement. The forecast verification framework established by Murphy and Winkler (1987) emphasizes the reliability of forecast systems, while Gneiting and Raftery (2005) further pointed out that verification requires not only numerical accuracy but also the capability to effectively predict extreme events—a perspective that is generally lacking in existing normalization approaches.

Based on the above analysis, this paper proposes a universal evaluation framework for geoscientific scalar time-series models, grounded in the scale-invariant

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

mathematical property of range normalization. This framework, termed the "Golden Triangle," consists of NRMSE, NMAE, and PCC. NRMSE captures extreme error sensitivity, NMAE reflects the average error level, and PCC assesses trend consistency; the three metrics complement one another from distinct dimensions, providing a comprehensive characterization of model performance. By eliminating dimensional effects through range normalization, this framework enables standardized and fair comparisons across stations, models, and datasets within the context of the same physical variable. To verify the effectiveness of the framework, this study takes the observed data from four global deep-sea buoys with significantly different wave height magnitudes, along with two independent wave reanalysis models, as a case study, and compares the evaluation results of dimensional metrics with those of the "Golden Triangle" metrics. The case results demonstrate that the framework can reasonably distinguish the relative performance of models across different stations, validating the practicality and robustness of the method. The mathematical foundation of the framework does not depend on any specific physical quantity, making it applicable to the validation needs of various types of geoscientific scalar time-series models.

2. Review of existing model validation metrics

2.1 Classification and development of commonly used evaluation metrics

Model validation metrics in the geosciences have evolved over several decades into three relatively mature categories, corresponding to three independent evaluation perspectives—absolute error, relative error, and correlation—applicable to both traditional numerical models and emerging data-driven models.

The first category consists of absolute error metrics, represented by RMSE and MAE. These metrics are physically intuitive, directly quantify the absolute magnitude of errors, and have been widely applied across various geoscience branches, including oceanography, meteorology, and hydrology. They are primarily suited for comparing different model configurations within the same station and the same dataset, where they can clearly reflect the absolute level of error. However, constrained by their dimensional dependence, they inherently lack the capacity for cross-station and cross-dataset comparisons.

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

The second category consists of relative error metrics, whose core idea is to eliminate dimensional effects through normalization, thereby enabling relative accuracy assessment. Among these, the most widely used is SI, which normalizes the RMSE by the observed mean. SI has a long history of application in wave model validation and has served as a classical standard for nearshore wave verification since the early development of third-generation wave models (Ris et al., 1999). SI offers good interpretability when the data distribution is stationary and lacks prominent extreme values. In addition to mean normalization, existing studies have also adopted other normalization bases, such as extreme values and the observed range; however, the selection criteria vary, and a unified consensus has yet to be established across the field. Among these, range normalization, owing to its ability to cover the full dynamic range of the data, offers unique advantages in conditions involving extreme events, but still lacks systematic mathematical justification and a standardized framework.

The third category consists of correlation metrics, represented by PCC and R^2. These metrics focus on the degree to which the simulated time series matches the observed time series in terms of temporal evolution patterns and are inherently dimensionless. However, they only reflect the degree of linear correlation and are completely insensitive to systematic bias, thus failing to capture the true magnitude of errors. Using them alone can easily lead to evaluation bias.

2.2 Common issues in current methodological frameworks

Taken together, the existing metric system still suffers from three common issues, which make it difficult to support standardized cross-scenario model evaluation.

In this paper, "cross-scenario" broadly refers to cross-station, cross-dataset, and cross-model comparisons of relative performance. It should be noted that absolute error metrics such as RMSE can compare different models when the station and dataset are identical; however, once the station or dataset changes, these metrics can no longer provide a fair comparison. The framework proposed in this paper is designed specifically to address cross-scenario evaluation.

The first issue is the lack of a unified consensus on normalization bases. Different studies adopt different normalization schemes—such as the mean or

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

extreme values—according to the context of their own research, leading to inconsistent physical interpretations of relative error results. This makes it difficult to establish a universal benchmark for model performance and hinders the accumulation of cross-study findings as well as model improvement and benchmarking.

The second issue is the insufficient capacity for fair cross-scenario comparison. Absolute error metrics, constrained by their dimensional dependence, are incapable of cross-scenario comparison. SI, as a representative mean-normalized RMSE metric, eliminates dimensional effects, but its applicable conditions are limited to daily operating conditions with stationary data and no prominent extreme values; it cannot reflect the simulation accuracy of extreme events. However, in practical engineering applications such as wave modeling, meteorology, and hydrology, the simulation accuracy of extreme events is often the central concern in model performance evaluation. Under this requirement, SI is not an ideal evaluation metric.

The third issue is the absence of a standardized composite evaluation framework. Individual metrics inherently have evaluation blind spots, yet existing studies tend to combine metrics arbitrarily based on empirical preferences. A standardized evaluation framework that simultaneously accounts for scale invariance, multi-dimensional complementarity, and broad applicability has yet to be established, making it difficult to diagnose model deficiencies comprehensively and systematically.

The above issues are not unique to any single sub-discipline; they are pervasive in the validation of various types of geoscientific scalar time-series models, including those for waves, temperature, precipitation, hydrology, and water quality. Establishing a logically coherent and generalizable standardized evaluation framework is thus of considerable methodological significance for advancing the standardization of geoscientific model validation.

3. A scale-invariant evaluation framework for scalar time-series models

3.1 Definition of metrics

To construct a universal evaluation framework applicable to geoscientific scalar time-series models, this study systematically integrates three existing metrics: RMSE and MAE are each normalized by the observed range to yield NRMSE and NMAE,

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

and these are then combined with PCC. The three metrics evaluate model performance from three distinct dimensions—extreme error sensitivity, average error level, and trend consistency—together forming a complementary "Golden Triangle" evaluation system.

Let the observed series at a given station be $x = (x_1, x_2, \dots, x_n)$, and the corresponding model-simulated series be $y = (y_1, y_2, \dots, y_n)$. The formulas for each metric are defined as follows:

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \quad (2)$$

$$NRMSE = \frac{RMSE}{max(x) - min\,(x)} \quad (3)$$

$$NMAE = \frac{MAE}{max(x) - min\,(x)} \quad (4)$$

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (5)$$

where $\bar{x}$ and $\bar{y}$ are the means of the observed series and the simulated series, respectively; $max(x)$ and $min\,(x)$ are the maximum and minimum values of the observed series. Smaller RMSE and MAE values indicate smaller absolute model errors; smaller NRMSE and NMAE values indicate smaller relative model errors. PCC ranges from −1 to 1, with values closer to 1 indicating a stronger linear correlation between the simulated and observed values.

3.2 Scale invariance

In model validation across different stations or datasets, the numerical magnitude of the observed variable often differs considerably. When conducting cross-scenario comparison, an evaluation metric should not be influenced by the absolute magnitude of the variable, but should only reflect the relative error level. This property is termed scale invariance. Only evaluation metrics possessing scale invariance can be directly compared across scenarios, free from interference by the absolute size of the numerical values.

Definition: Let the observed series at Station A be $x_A = (x_1, x_2, \dots, x_n)$, and the

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

corresponding model-simulated series be $y_A = (y_1, y_2, \dots, y_n)$. Consider Station B, whose observed and simulated values are both k times those of Station A, i.e., $x_B = k \cdot x_A$, $y_B = k \cdot y_A$. Under this setting, Stations A and B differ in variable magnitude, but the relative simulation performance of the model is identical. If a metric M satisfies $M(x_A, y_A) = M(x_B, y_B)$, then M is said to possess scale invariance.

RMSE test:

$$RMSE(x_B, y_B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(ky_i - kx_i)^2} = \sqrt{k^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} = k \cdot RMSE(x_A, y_A) \quad (6)$$

MAE test:

$$MAE(x_B, y_B) = \frac{1}{n}\sum_{i=1}^{n}|ky_i - kx_i| = k \cdot \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| = k \cdot MAE(x_A, y_A) \quad (7)$$

NRMSE test:

$$NRMSE(x_B, y_B) = \frac{RMSE(x_B, y_B)}{max(x_B) - min(x_B)} = \frac{k \cdot RMSE(x_A, y_A)}{k \cdot (max(x_A) - min(x_A))} = NRMSE(x_A, y_A) \quad (8)$$

NMAE test:

$$NMAE(x_B, y_B) = \frac{MAE(x_B, y_B)}{max(x_B) - min(x_B)} = \frac{k \cdot MAE(x_A, y_A)}{k \cdot (max(x_A) - min(x_A))} = NMAE(x_A, y_A) \quad (9)$$

PCC test:

$$PCC(x_B, y_B) = \frac{\sum_{i=1}^{n}(k \cdot x_i - k \cdot \bar{x})(k \cdot y_i - k \cdot \bar{y})}{\sqrt{\sum_{i=1}^{n}(k \cdot x_i - k \cdot \bar{x})^2 (k \cdot y_i - k \cdot \bar{y})^2}} = \frac{k^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{k^2 \cdot \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} = PCC(x_A, y_A) \quad (10)$$

From the above tests, it can be seen that RMSE and MAE do not possess scale invariance: when the variable is uniformly scaled by a factor of k, they are also scaled by k. This means that even if the relative model performance is identical at Stations A and B, RMSE and MAE will still differ due to the difference in variable magnitude, making it impossible to distinguish whether the difference stems from genuine model performance or data dimensionality. NRMSE, NMAE, and PCC all possess scale invariance: when the variable is uniformly scaled, their values remain unchanged. This indicates that these metrics eliminate the influence of dimensionality, enabling direct and fair comparisons across stations with different magnitudes. Notably, NRMSE and NMAE can be used for cross-station ranking: stations with

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

smaller values indicate that the model achieves smaller relative errors at those stations.

3.3 Comparison between range normalization and mean normalization

This section discusses the mathematical properties of mean normalization and compares its applicable conditions with those of range normalization.

Mean normalization corresponds to SI, which is defined as the RMSE normalized by the observed mean:

$$SI = \frac{RMSE}{\bar{x}} \tag{11}$$

SI test:

$$SI(x_B, y_B) = \frac{RMSE(x_B, y_B)}{\bar{x}_B} = \frac{k \cdot RMSE(x_A, y_A)}{k \cdot \bar{x}_A} = SI(x_A, y_A) \tag{12}$$

From the above test, it can be seen that mean normalization likewise possesses scale invariance.

Although both range normalization and mean normalization possess scale invariance, the information captured by their denominators differs. Range normalization uses the observed range as the denominator, measuring the proportion of model error relative to the full dynamic range of the data. The range is determined by the extreme values at both ends of the data and is highly sensitive to extreme events; thus, range normalization is capable of reflecting the relative performance of the model over the entire variation interval. Mean normalization uses the observed mean as the denominator, measuring the proportion of model error relative to the average level of the data. The mean reflects the central tendency of the data and is relatively less sensitive to extreme values; thus, mean normalization is more focused on evaluating model performance under daily operating conditions.

The two normalization strategies are not inherently superior or inferior to one another; rather, they differ in their applicable conditions. When the data series is relatively stationary with no prominent extreme values, or when the research objective focuses on daily performance, mean normalization is an appropriate choice. When the data series contains significant extreme events and the engineering application is particularly concerned with the simulation accuracy of such events, range

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

normalization is more suitable. Geoscientific scalar time-series models generally require the assessment of model reliability under extreme conditions; therefore, range normalization better meets the practical needs of geoscientific model validation.

3.4 Analysis of metric complementarity

NRMSE, NMAE, and PCC all satisfy scale invariance, and the three metrics are mutually independent and complementary in their evaluation dimensions, together forming a complete diagnostic system.

NRMSE inherits the squared weighting characteristic of the root mean square error, making it more sensitive to large-magnitude errors. It can highlight simulation deviations during extreme events, focusing on reflecting the extreme error level of the model over the full dynamic range. However, because it is dominated by large errors, the contribution of small errors under daily operating conditions can be diluted in the final result. NMAE treats all errors with equal weight, providing a stable reflection of the average relative error level of the model, and aligning more closely with the overall accuracy performance under daily operating conditions. However, it is insensitive to deviations of a few extreme values and cannot independently capture extreme-value simulation capability. PCC only measures the temporal trend consistency between the simulated and observed series. Completely independent of the absolute magnitude of errors, it can reflect the model's ability to capture the evolution of physical processes. However, it is entirely insensitive to systematic bias—even if the simulated values are systematically higher or lower, a high correlation coefficient can still be obtained as long as the temporal fluctuations are synchronized.

The three metrics approach model evaluation from three independent dimensions—extreme error sensitivity, average error level, and trend consistency. The evaluation blind spots that a single metric cannot cover can be compensated by the other metrics, forming a multi-dimensional diagnostic capability. For example, a high PCC combined with a high NRMSE suggests that the model captures the correct trend but incurs large extreme errors, indicating that improving numerical accuracy during extreme events should be prioritized. A low PCC coupled with a low NRMSE suggests that the model poorly captures the trend but that the relative error is

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

controllable, indicating that optimizing the temporal evolution mechanisms and timing consistency of physical processes should be prioritized. This kind of multi-dimensional evaluation provides more refined feedback for model developers and users, helping to identify model deficiencies in a targeted manner and optimize parameterization schemes.

In summary, the "Golden Triangle" evaluation framework, jointly constituted by NRMSE, NMAE, and PCC, possesses two core strengths—scale invariance and dimensional complementarity. It provides a unified evaluation benchmark for cross-scenario model performance comparison, and this constitutes the most fundamental design value of the framework.

4. Case study: cross-basin wave model validation

The preceding sections systematically constructed the "Golden Triangle" evaluation framework and proved its scale invariance. To verify the framework's effectiveness in practical applications, this study selects observed wave data from four stations across different global ocean basins and data from two reanalysis models as an empirical case study. This case was chosen based on two considerations. First, the wave field possesses a mature and well-developed validation metric system, with RMSE, MAE, SI, and PCC all widely applied, making it convenient for comparing the differences between evaluation systems. Second, the stations across different ocean basins exhibit significant differences in wave height magnitude and distribution characteristics, allowing for the simultaneous testing of the framework's cross-scenario comparison capability and its effectiveness in evaluating extreme events.

4.1 Study area and data

The observed Hs data used in this case study were obtained from four deep-sea wave buoys of the National Data Buoy Center (NDBC) (National Data Buoy Center, 2026). As shown in Fig. 1, these buoys are located in the North Atlantic Ocean (41001), the Gulf of Mexico (42001), the Gulf of Alaska (46001), and the Central Pacific Ocean (51001), respectively. All stations have water depths exceeding 3000 m, representing open-ocean environments across different ocean basins. The recording intervals of the raw NDBC wave data are not entirely uniform. To match the temporal resolution of the model data, all valid Hs observations falling within the same hour

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

were arithmetically averaged and uniformly resampled into a 1-hourly time series in this study. Time steps with no valid observations after resampling were retained as blank missing values without any interpolation or gap-filling. During model validation and error assessment, the observed wave heights and the corresponding model data were directly matched in time based on the observation timestamps. Detailed information for each station is provided in Table 1.

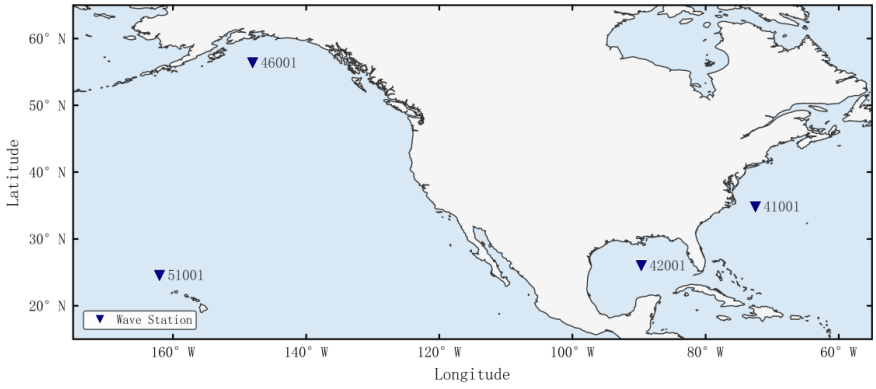


Fig. 1. Geographical locations of the NDBC buoy stations used in this study

Table 1

Basic information of the wave measurement stations

Station	Latitude	Longitude	Water Depth (m)	Observation Period	Resampled Sample Size
41001	34.738 N	72.503 W	4462	2022/01–2025/12	23379
42001	25.922 N	89.638 W	3195	2022/01–2025/12	24585
46001	56.296 N	148.027 W	4123	2022/01–2025/12	25377
51001	24.475 N	162.030 W	4895	2022/01–2025/12	34809

Figure 2 presents violin plots of the observed Hs at each station, overlaid with box plots indicating the median and interquartile range, with extreme values marked by triangle symbols. The figure clearly reveals substantial structural differences in the wave height data across the four stations. The figure clearly reveals substantial structural differences in the wave height data across the four stations. Station 42001 exhibits a concentrated distribution of low values, with a median of only 0.86 m, 95% of wave heights below 2.39 m, and an overall range of 0.11–5.44 m, making it the

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

station with the most concentrated data distribution and the lowest wave height magnitude among the four. Station 46001 displays typical high-sea-state characteristics, with a median of 2.18 m, 95% of wave heights below 5.07 m, and an overall range of 0.48–11.11 m, showing an extremely pronounced disparity between low and high values. Station 41001 has the largest dynamic range of wave heights; its minimum is 0.50 m and median is 1.58 m, but extreme wave heights can reach 11.52 m. Station 51001 shows a generally elevated and relatively concentrated distribution, with a median of 2.07 m, 95% of wave heights below 3.92 m, and an overall range of 0.76–8.37 m.

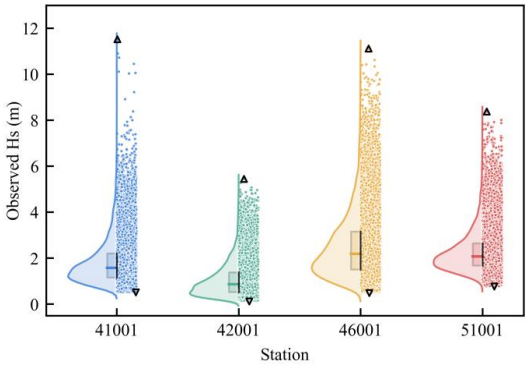


Fig. 2. Comparison of observed Hs distributions among four stations

The above data exhibit fundamental differences in wave height magnitude, distribution shape, concentration, and extreme value characteristics. The wave height ranges across the stations vary dramatically, from 5.33 m to 11.01 m (more than a twofold difference), covering typical deep-water wave characteristics from mild sea states to extreme storms. This significant cross-magnitude variation provides an ideal verification setting for testing the cross-scenario fair comparison capability of the "Golden Triangle" framework.

4.2 Model data

To validate the "Golden Triangle" evaluation framework, two widely used global wave reanalysis datasets were selected as model outputs in this study.

The first dataset is the ERA5 global reanalysis wave data, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF; Hersbach et al.,

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

2018). The ERA5 wave fields are generated by the third-generation wave model WAM. It should be noted that while this dataset assimilates satellite altimeter observations, it does not assimilate the NDBC buoy observations used in this study, thereby ensuring the independence and objectivity of the validation. The spatial resolution of the ERA5 wave data is relatively coarse at 0.5° × 0.5°; therefore, the model grid points are far from the exact buoy locations. In this study, bilinear interpolation was applied to interpolate the values of the four nearest grid points to the exact buoy coordinates to obtain the model values at the buoy locations.

The second dataset is the WAVERYS global high-resolution wave reanalysis product (Product ID: GLOBAL_MULTIYEAR_WAV_001_032; Copernicus Marine Service, 2024). WAVERYS is generated by the third-generation wave model MFWAM, featuring a spatial resolution of 0.2° × 0.2° and a temporal resolution of 3 hours. Because the WAVERYS grid is finer with grid points closer to the buoy locations, and all buoys are situated in deep open waters where the wave field is highly spatially homogeneous, the nearest grid point data were directly extracted without spatial interpolation. To maintain consistency with the temporal resolution of the observations, linear interpolation was applied in the time dimension to obtain the model values at a 1-hour temporal resolution. Due to the boundary effect of linear interpolation, the interpolated results for the last two hours of December 31, 2025, were missing, and these two hours were excluded as missing values during subsequent pairing.

4.3 Results

4.3.1 Time series comparison

Figure 3 presents the time series of observed and simulated significant wave height from the ERA5 and WAVERYS models at the four stations (data from the year 2023 are shown as a representative example). Overall, the simulated curves exhibit high consistency with the observed wave height series in terms of temporal evolution trends, indicating high reliability of the simulation results. However, the degree of waveform agreement differs across stations. Direct visual inspection can neither accurately quantify the performance differences of the models across stations nor eliminate the potential interference of wave height magnitude on the evaluation.

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

Therefore, further quantitative analysis using statistical metrics is necessary.

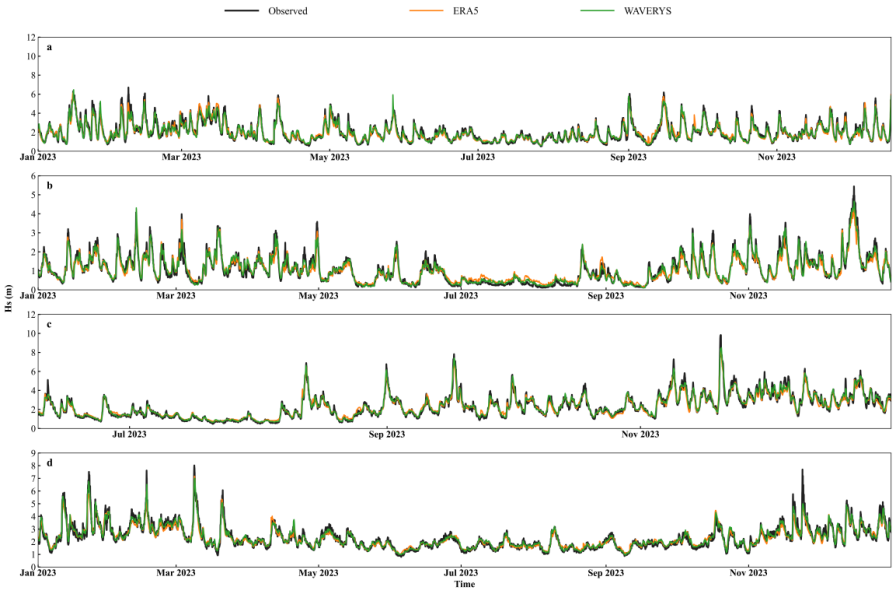


Fig.3. Time series of observed and simulated Hs at (a) Station 41001, (b) Station 42001, (c) Station 46001, and (d) Station 51001. (Data from the year 2023 is shown as a representative example).

4.3.2 Cross-station error comparison

Table 2 lists the mean Hs, range, and the values of RMSE, MAE, NRMSE, NMAE, and PCC for the two models at the four stations. The dimensional error metrics are first examined. For the ERA5 model, Station 42001 has a mean Hs of 1.04 m, with RMSE and MAE of 0.206 m and 0.147 m, respectively, which are the lowest among the four stations; Station 46001 has the highest mean Hs of 2.49 m, with RMSE and MAE of 0.313 m and 0.220 m, which are the highest among the four stations. A similar pattern is observed for the WAVERYS model, exhibiting the lowest error at Station 42001 and the highest at Station 46001. However, because RMSE and MAE lack scale invariance, their values are influenced by both the wave height magnitude and the actual model performance. Consequently, although these metrics can be used to compare different models at the same station, they cannot be directly applied to fair cross-scenario comparisons. For instance, the RMSE at Station 42001 is smaller than that at Station 46001, but this does not indicate that the model

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

performs better at Station 42001, because the difference in wave height magnitude confounds the interpretation.

Table 2

Statistical metrics for ERA5 and WAVERYS models at four stations

Station	Mean Hs (m)	Median Hs (m)	Range (m)	Model	*RMSE* (m)	*MAE* (m)	*NRMSE*	*NMAE*	*PCC*
41001	1.82	1.58	11.01	ERA5	0.283	0.196	0.026	0.018	0.957
				WAVERYS	0.242	0.164	0.022	0.015	0.969
42001	1.04	0.86	5.33	ERA5	0.206	0.147	0.039	0.028	0.96
				WAVERYS	0.162	0.113	0.030	0.021	0.974
46001	2.49	2.18	10.64	ERA5	0.313	0.220	0.029	0.021	0.979
				WAVERYS	0.252	0.175	0.024	0.016	0.983
51001	2.26	2.07	7.61	ERA5	0.256	0.183	0.034	0.024	0.958
				WAVERYS	0.192	0.138	0.025	0.018	0.976

In contrast, NRMSE and NMAE possess scale invariance, thereby eliminating the influence of wave height magnitude. For the ERA5 model, although Station 46001 has a larger mean Hs than Station 42001, its NRMSE is 0.029, which is smaller than the value of 0.039 at Station 42001. The NRMSE values are 0.026 at Station 41001, 0.039 at Station 42001, 0.029 at Station 46001, and 0.034 at Station 51001, yielding a ranking of 41001 < 46001 < 51001 < 42001. The NMAE values follow a similar ranking. A consistent pattern is also found for the WAVERYS model.

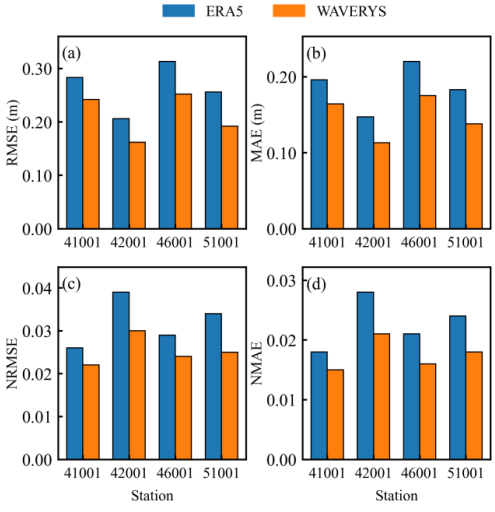


Fig.4. RMSE, MAE, NRMSE and NMAE of ERA5 and WAVERYS models at four stations.

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 4 presents bar charts of RMSE, MAE, NRMSE, and NMAE for the two models at the four stations. The figure intuitively demonstrates that ranking the model performance across stations using RMSE and MAE yields results that differ from those obtained using NRMSE and NMAE. This fully demonstrates that metrics lacking scale invariance can produce misleading evaluation results in cross-scenario comparisons, while NRMSE and NMAE, owing to their scale invariance, produce rankings that are independent of wave height magnitude, reflecting the true relative performance of the model and achieving fair cross-station comparisons.

4.3.3 Performance under extreme events

To examine whether the range-normalized metrics can reflect model performance during extreme events, the 90th percentile of the full-dataset observed Hs series at each station is taken as the threshold, and all time steps with Hs values not lower than this threshold are extracted to form the high-energy-state dataset. The range-normalized root mean square error computed exclusively for the high-energy states, denoted as $NRMSE_{high}$, is calculated for each station and compared in ranking with the NRMSE based on the full dataset.

Table 3

High-energy states statistics

Station	Thresh(m)	N_{high}	$Range_{high}$(m)	$RMSE_{high}$(ERA5)	$NRMSE_{high}$(ERA5)	$RMSE_{high}$(WAVERYS)	$NRMSE_{high}$(WAVERYS)
41001	3.11	2348	8.41	0.594	0.071	0.533	0.063
42001	2.01	2461	3.43	0.404	0.118	0.315	0.092
46001	4.30	2542	6.81	0.663	0.097	0.521	0.076
51001	3.38	3488	4.98	0.520	0.104	0.357	0.072

Note: High-energy states defined by Hs $\geqslant$ threshold, where threshold = 90th percentile of observed Hs; N_{high} = number of measured points in high-energy states. $Range_{high}$ = max − min of observed Hs; $NRMSE_{high}$ computed only for high-energy states.

Table 3 lists the basic statistics of the high-energy-state data and the $RMSE_{high}$ and $NRMSE_{high}$ values for the two models at each station. The data in Table 3 once again confirm the limitation of RMSE. For both the ERA5 and WAVERYS models, the minimum $RMSE_{high}$ occurs at Station 42001, but its $NRMSE_{high}$ is the highest. This contradictory ranking within the same dataset indicates that RMSE is severely constrained by wave height magnitude, making it difficult to distinguish the relative model performance during extreme events.

A comparison of Table 2 and Table 3 shows that the $NRMSE_{high}$ values at all

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

stations exceed the corresponding full-dataset NRMSE values. This is because the high-energy-state periods contain only large-wave subsets, where the range is narrower, and the absolute model error in the large-wave region tends to be relatively larger, thereby elevating the $NRMSE_{high}$. Nevertheless, this change does not affect the consistency of the model performance ranking across stations.

Figure 5 directly compares the NRMSE rankings derived from the full dataset and the high-energy states: For the ERA5 model, the full-dataset and high-energy-state rankings are completely identical, both being 41001 < 46001 < 51001 < 42001. For the WAVERYS model, the overall trend remains stable, with Station 41001 always the lowest and Station 42001 always the highest, with only a local swap in rankings between Stations 51001 and 46001. In-depth analysis reveals that this local swap stems from the weak differences in the error metrics between the two stations, making the ranking sensitive to variations in the sample subset. Under the full dataset, the NRMSE values for Stations 46001 and 51001 are 0.024 and 0.025, with a difference of only 0.001; under the high-energy subset, their NRMSE values are 0.076 and 0.072, with a difference of 0.004. Regardless of the dataset condition, the differences between the two stations account for a very small proportion of the overall distribution range of NRMSE across stations: the full-dataset range is 0.022–0.030, and the high-energy range is 0.063–0.092. Compared to station pairs with stable rankings under both conditions, such as 41001 and 42001, the deviation between 46001 and 51001 is insufficient to maintain a stable performance ranking; thus, the WAVERYS model can be considered to exhibit comparable performance at these two stations.

The objective identification of such weak differences precisely reflects the advantage of the range-normalized metrics. If RMSE were used for comparison, the RMSE values at Stations 46001 and 51001 under high-energy states would be 0.521 m and 0.357 m, respectively, yielding a relative difference based on the mean of approximately 37%. This apparent difference, extremely amplified by wave height magnitude, masks the essence that the models' simulation performance is similar at the two stations. In contrast, NRMSE removes the interference of wave height magnitude, accurately identifying the similar performance levels at the two stations while maintaining the consistency of relative trends between the full dataset and

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

extreme events across most stations.

Thus, the range-normalized metrics can be effectively used to evaluate the models' simulation capability under extreme conditions.

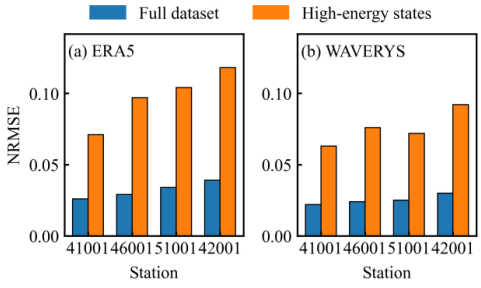


Fig.5. Comparison of NRMSE between full dataset and high-energy states for (a) ERA5 and (b) WAVERYS.

4.3.4 Validation of metric complementarity

The complementarity among the three metrics within the "Golden Triangle" framework has been theoretically established in the preceding sections. This section provides an empirical validation of this complementary relationship using the observed and model data. Taking the ERA5 model at Station 42001 as an example, its PCC reaches 0.960, which is at a relatively high level among the four stations, indicating that the model well captures the temporal variation patterns of wave height. However, its NRMSE is 0.039, the highest among the four stations, indicating that the relative error magnitude is relatively large. Taken together, these results suggest that although the model captures the trend well at Station 42001, the overall error is relatively large, and improving the numerical simulation accuracy should be prioritized in subsequent work. Conversely, the PCC of the ERA5 model at Station 41001 is 0.957, the lowest among the four stations. However, its NRMSE is merely 0.026, remaining at a relatively low level, indicating that the relative error of the model at Station 41001 is controllable but the trend-fitting capability is slightly insufficient, and optimizing the temporal evolution mechanisms of physical processes should be prioritized.

The above cases clearly demonstrate that no single metric can fully capture

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

model performance: a high PCC may mask relatively large errors, and a low NRMSE may be accompanied by a poor ability to simulate trends. Only by combining all three metrics can a complete multi-dimensional diagnosis of model performance be achieved. This complementarity, together with the scale invariance of the "Golden Triangle" framework, constitutes its core strengths: it enables fair cross-scenario comparisons while systematically diagnosing model deficiencies and their directions for improvement from multiple dimensions.

5 Discussion

5.1 Advantages of range-normalized metrics for cross-scenario comparison

The core advantage of range normalization lies in its scale invariance, which does not depend on any particular case or data type. For the same physical variable, as long as the observational data differ, the evaluation results given by range-normalized metrics can be directly compared across stations and datasets—something that dimensional metrics such as RMSE and MAE cannot achieve.

Compared with mean normalization, the denominator selection strategy of range normalization makes it more robust under conditions of complex data distribution. Mean normalization uses the mean as the denominator and performs well when the data distribution is relatively symmetric. However, when the data exhibit a strongly right-skewed distribution, although extreme values will elevate the mean, due to the presence of numerous low-value samples, the mean still remains far below the extreme upper limit of the data, failing to objectively reflect the error level of the model over the full dynamic range. In the present case, the wave height data at Station 42001 are highly concentrated at low values, yet the range can reach 5.33 m, while the mean is only 1.04 m—far below the middle-to-high value range of the data. Using the mean as the normalization base underestimates the model's relative performance near extreme values. Range normalization uses the observed range as the denominator, anchoring the full dynamic range of the data. Its physical meaning is "the proportion of model error relative to the total variation amplitude of the variable". It is free from interference by the skewness and concentration of the data distribution, and its evaluation results are more stable and more comparable.

5.2 Diagnostic value of multi-metric complementarity

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

The ultimate goal of geoscientific model validation should not be to obtain a vague conclusion of "good" or "poor," but rather to precisely identify the directions of model deficiencies, thereby providing a clear basis for subsequent improvement. A single metric can only reflect one aspect of model performance and cannot independently undertake the task of diagnosis.

The three metrics within the "Golden Triangle" framework correspond to three core issues in model debugging: NRMSE corresponds to extreme error sensitivity, evaluating the simulation accuracy of extreme events; NMAE corresponds to the average error level, evaluating the overall performance over the full period; PCC corresponds to trend consistency, evaluating the temporal variation patterns. Used in combination, the three metrics can decompose model error into two dimensions—magnitude deviation and trend deviation—providing a clear diagnostic pathway for locating model deficiencies. In the present case, the contrast between the ERA5 model data at Station 42001 and Station 41001 clearly demonstrates this diagnostic value: Station 42001 has a PCC of 0.960, indicating that the model's trend-fitting capability is excellent, but its NRMSE is 0.039, the highest among the four stations, indicating that the magnitude deviation is relatively large; Station 41001 has an NRMSE of 0.026, indicating that the overall relative error is controllable, but its PCC is 0.957, the lowest among the four stations, indicating that its ability to capture temporal variation patterns is slightly insufficient.

This multi-dimensional diagnostic capability provides direct guidance for practical work such as the optimization of parameterization schemes, the adjustment of grid resolution, and the selection of data assimilation strategies in geoscientific models. Model developers can use the feedback from the "Golden Triangle" to select improvement directions in a targeted manner according to practical needs.

5.3 Scope of applicability and limitations

The "Golden Triangle" evaluation framework was originally designed to provide a standardized cross-scenario evaluation tool for the validation of continuous scalar time-series data. Any geoscientific model that takes time as the index and produces continuous numerical outputs—such as temperature and precipitation forecasting in meteorology, runoff and soil moisture simulation in hydrology, wave height and water

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

level prediction in oceanography, and dissolved oxygen and pollutant concentration simulation in environmental science—can directly adopt this framework for multi-dimensional diagnosis of model performance. Users only need to provide the observed series and model-simulated series of the physical variable to complete the calculation and comparative analysis of all metrics.

Two aspects require attention during application. First, the denominator of range normalization is determined by the maximum and minimum values of the observed series, making it relatively sensitive to anomalous extremes in the data. Before calculating the normalized metrics, necessary quality control should be performed on the observed data to remove clearly identified outliers. Conversely, the sensitivity of range normalization to outliers also has diagnostic value: if the NRMSE value is far lower than expected, it suggests that unidentified anomalous extremes may exist in the observed series, warranting a retrospective check of the original data quality. Second, the current framework focuses on the amplitude evaluation and trend diagnosis of scalar variables and cannot be directly applied to vector physical quantities that contain both magnitude and direction dimensions; in such cases, additional directional error metrics need to be introduced as a supplement.

5.4 Positioning of this study

In current geoscientific model validation practice, researchers typically employ multiple metrics simultaneously for evaluation. However, the selection of metric combinations is often based on individual experience or disciplinary conventions, and normalization approaches also vary across studies. As a result, findings from different studies are difficult to benchmark against one another, hindering the effective accumulation of domain knowledge. Few studies have systematically constructed a universal evaluation framework from the two dimensions of cross-scenario comparability and multi-dimensional complementarity. The "Golden Triangle" framework proposed in this paper directly responds to this gap.

This framework takes scale invariance as its mathematical foundation, resolving the fundamental problem that dimensional metrics cannot enable fair cross-station and cross-dataset comparisons. It provides a unified theoretical basis for benchmarking model performance across different studies. At the same time, through its three-

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

dimensional combination of NRMSE, NMAE, and PCC, the framework elevates model diagnosis from merely "providing an error value" to "identifying the specific dimensions of deficiencies," offering clear directional guidance for subsequent work such as the refinement of parameterization schemes and the optimization of data assimilation strategies.

Any model validation involving continuous scalar time-series data can directly adopt this framework without the need for domain-specific adjustments. This generality endows the framework with the potential to evolve toward a universal disciplinary standard.

6 Conclusions and outlook

6.1 Main conclusions

This study addresses the challenges of cross-scenario fair comparison and the lack of systematic design in metric combinations in geoscientific model validation. The "Golden Triangle" evaluation framework was constructed and empirically validated using observations from four globally distributed deep-sea NDBC buoys, along with the ERA5 and WAVERYS reanalysis datasets, as a case study. The main conclusions are as follows:

From the theoretical perspective, the NRMSE, NMAE, and PCC metrics possess rigorous scale invariance, meaning their evaluation results are unaffected by the absolute magnitude of the variable. They can thus enable fair cross-station and cross-dataset comparisons of model performance. This property provides a unified mathematical foundation for cross-scenario evaluation and resolves the fundamental problem that dimensional metrics cannot achieve fair comparisons across different magnitudes.

From the diagnostic capability perspective, the three metrics—NRMSE, NMAE, and PCC—approach model evaluation from three independent dimensions: extreme error sensitivity, average error level, and trend consistency, complementing one another. Used in combination, they can decompose model error into two dimensions—magnitude deviation and trend deviation—thereby precisely identifying the directions of model deficiencies and providing clear diagnostic guidance for

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

subsequent work such as the optimization of parameterization schemes.

From the empirical perspective, the cross-basin wave model validation case demonstrates that the framework can effectively eliminate the interference caused by differences in wave height magnitude among stations, overcome the cross-station ranking distortion caused by dimensional dependence, and yield reasonable cross-station performance rankings. Furthermore, the framework exhibits excellent consistency when comparing the full dataset with the high-energy-state dataset. This fully validates the practicality and reliability of the method.

6.2 Outlook

The "Golden Triangle" framework is designed to serve the validation needs of various types of geoscientific scalar time-series models. Future work toward this goal can proceed along the following directions.

In terms of application expansion, further empirical validation can be conducted in additional geoscience branches such as meteorology, hydrology, and environmental science, accumulating application cases of the framework under different data characteristics and model types, and testing its applicability and robustness in a wider range of settings.

In terms of framework extension, the current framework focuses on the amplitude evaluation and trend diagnosis of scalar variables and cannot be directly applied to vector physical quantities such as ocean currents and wind speed, which contain both magnitude and direction dimensions. Future work can explore extending the idea of scale invariance to vector settings, thereby constructing a multi-dimensional evaluation system that accounts for both amplitude and directional errors.

Code availability

No novel algorithms or standalone software packages were developed in this study. The Jupyter Notebook files, including scripts for raw data preprocessing, dataset merging, metric calculation, table generation, and figure plotting, are provided as Supplementary Information to support the reproducibility of the results.

Data availability

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.

All wave observation and reanalysis model data used in this study are publicly available. The buoy wave observations were obtained from the National Data Buoy Center. ERA5 wave reanalysis data were retrieved from the Copernicus Climate Change Service, and WAVERYS wave reanalysis data were obtained from the Copernicus Marine Service. The processed datasets and raw data subsets for case validation are provided as Supplementary Information. Statistical metrics supporting the conclusions are presented in the figures and tables of the main text. Readers can independently reproduce the core findings of this study using these publicly available datasets and the corresponding supplementary files.

Author contribution

Fayun Wang: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Software, Visualization, Writing – original draft, Writing – review & editing.

Competing interests

The author declares that he has no conflict of interest.

Acknowledgements

The author would like to thank the Hainan Academy of Ocean and Fisheries Sciences for providing the research environment, institutional support.

References

Aydoğan, B. and Ayat, B.: Spatial variability of long-term trends of significant wave heights in the Black Sea, Appl. Ocean Res., 79, 20 – 35, doi:10.1016/j.apor.2018.07.001, 2018.

Booij, N., Ris, R. C., and Holthuijsen, L. H.: A third-generation wave model for coastal regions 1. Model description and validation, J. Geophys. Res., 104, 7649–7666, doi:10.1029/98JC02622, 1999.

Copernicus Marine Service (CMEMS): Global Ocean Waves Reanalysis (WAVERYS), Product ID: GLOBAL_MULTIYEAR_WAV_001_032, doi:10.48670/moi-00022, 2024.

Deo, M. C. and Naidu, C. S.: Real time wave forecasting using neural networks, Ocean Eng., 26, 191–203, doi:10.1016/S0029-8018(97)10025-7, 1999.

Fan, S. T., Xiao, N. H., and Dong, S.: A novel model to predict significant wave height based on long short-term memory network, Ocean Eng., 205, 107298, doi:10.1016/j.oceaneng.2020.107298, 2020.

Feng, X. C. and Xu, H.: Accurate storm surge prediction in hurricane area of the Atlantic Ocean

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



using a new multi-recursive neural network based on gate recursive unit, J. Ocean Eng. Sci., 11, 726–739, doi:10.1016/j.joes.2024.01.001, 2026.

Gneiting, T. and Raftery, A. E.: Weather Forecasting with Ensemble Methods, Science, 310, 248–249, doi:10.1126/science.1115255, 2005.

Gonçalves, M., Martinho, P., and Guedes Soares, C.: Assessment of wave energy in the Canary Islands, Renew. Energy, 68, 774–784, doi:10.1016/j.renene.2014.03.017, 2014.

Hersbach, H., et al.: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS), doi:10.24381/cds.adbb2d47, 2018.

Li, C. L., Xu, H., Feng, X. C., and Sun, Q.: Enhanced intelligent reconstruction study on wind wave height field in the South China Sea, J. Ocean Eng. Sci., 10, 897–915, doi:10.1016/j.joes.2025.02.001, 2025.

Li, Y., Peng, T., Xiao, L. F., Wei, H. D., and Li, X.: Wave runup prediction for a semi-submersible based on temporal convolutional neural network, J. Ocean Eng. Sci., 9, 528–540, doi:10.1016/j.joes.2022.08.005, 2024.

Liu, Y., Zhang, X. T., Chen, G., Dong, Q., Guo, X. X., Tian, X. L., Lu, W. Y., and Peng, T.: Deterministic wave prediction model for irregular long-crested waves with Recurrent Neural Network, J. Ocean Eng. Sci., 9, 251–263, doi:10.1016/j.joes.2022.08.002, 2024.

National Data Buoy Center: National Data Buoy Center (NDBC) data, available at: https://www.ndbc.noaa.gov/, last access: 20 October 2026.

Murphy, A. H. and Winkler, R. L.: A General Framework for Forecast Verification, Mon. Weather Rev., 115, 1330–1338, doi:10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2, 1987.

Ponce de León, S. and Guedes Soares, C.: On the sheltering effect of islands in ocean wave models, J. Geophys. Res., 110, C09020, doi:10.1029/2004JC002682, 2005.

Pourzangbar, A., Jalali, M., and Brocchini, M.: Machine learning application in modelling marine and coastal phenomena: a critical review, Front. Environ. Eng., 2, 1235557, doi:10.3389/fenve.2023.1235557, 2023.

Qiu, L. G., Liang, X. L., Chen, B., Wang, S. J., and Wang, F. Y.: Wave distribution characteristics of Xuande Atoll, Xisha Islands, J. Trop. Oceanogr., 44, 50–64, doi:10.11978/2024240, 2025.

Ris, R. C., Holthuijsen, L. H., and Booij, N.: A third-generation wave model for coastal regions 2. Verification, J. Geophys. Res., 104, 7667–7681, doi:10.1029/1998JC900123, 1999.

Rusu, L., Pilar, P., and Guedes Soares, C.: Hindcast of the wave conditions along the west Iberian coast, Coastal Eng., 55, 906–919, doi:10.1016/j.coastaleng.2008.02.029, 2008.

Saket, A. and Etemad-Shahidi, A.: Wave energy potential along the northern coasts of the Gulf of Oman, Iran, Renew. Energy, 40, 90–97, doi:10.1016/j.renene.2011.09.024, 2012.

Stopa, J. E., Ardhuin, F., Babanin, A., and Zieger, S.: Comparison and validation of physical wave parameterizations in spectral wave models, Ocean Modell., 103, 2–17, doi:10.1016/j.ocemod.2015.09.003, 2016.

Sun, Z., Zhang, H. C., Liu, X. L., Ding, J., Xu, D. L., and Cai, Z. W.: Wave energy assessment of the Xisha Group Islands zone for the period 2010–2019, Energy, 220, 119721, doi:10.1016/j.energy.2020.119721, 2021.

Tolman, H. L.: A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents, J. Phys. Oceanogr., 21, 782–797, doi:10.1175/1520-0485(1991)021<0782:ATGMFW>2.0.CO;2, 1991.

https://doi.org/10.5194/egusphere-2026-4605
Preprint. Discussion started: 4 November 2026
© Author(s) 2026. CC BY 4.0 License.



780 Vieira, F., Cavalcante, G., and Campos, E.: Analysis of wave climate and trends in a semi-enclosed
781 basin (Persian Gulf) using a validated SWAN model, Ocean Eng., 196, 106821,
782 doi:10.1016/j.oceaneng.2019.106821, 2020.
783 WAMDI Group.: The WAM Model—A Third Generation Ocean Wave Prediction Model, J. Phys.
784 Oceanogr., 18, 1775–1810, doi:10.1175/1520-0485(1988)018<1775:TWMTGO>2.0.CO;2, 1988.
785 Wang, F. Y., Wang, S. J., Chen, B., Liang, N. A., and Qiu, L. G.: Application of an intelligent wave
786 forecasting method to waters around the islands and reefs in the South China Sea, Haiyang Xuebao,
787 47, 136–149, doi:10.12284/hyxb20250111, 2025.
788 Wang, F. Y., Ye, Q., Chen, B., Wang, S. J., Liang, N., and Qiu, L. G.: Mitigating domain shift for
789 cross-station wave height prediction in South China Sea island and reef waters: proposal and
790 systematic evaluation of three novel methods, J. Ocean Eng. Sci., doi:10.1016/j.joes.2026.05.013,
791 in press.
792 Wang, M., Ying, F. X., and Jia, J.: Ocean wave power flux forecasting using a stacking ensemble of
793 LSTM and LightGBM, Renew. Energy, 256, 124597, doi:10.1016/j.renene.2025.124597, 2026.
794 Xie, M. S., Sun, W. J., Han, Y., Ren, S., Li, C. H., Ji, J. L., Yu, Y., Zhou, S. Y., and Dong, C. M.:
795 Causal Matrix Long Short-Term Memory Network for interpretable significant wave height
796 forecasting, J. Mar. Sci. Eng., 13, 1872, doi:10.3390/jmse13101872, 2025.
797 Xie, M. S., Sun, W. J., Han, Y., and Dong, C. M.: Interpretable significant wave height prediction
798 in the South China Sea using PCMCI-driven fused causal neural networks, J. Ocean Eng. Sci.,
799 doi:10.1016/j.joes.2026.06.002, in press.