Journal of *Marine Science and Engineering*

MDPI

Article

Deep-Learning Error Correction of Regional Significant Wave Height Forecasts

Wenbo Li [1,2], Song Gao [1], Zhifeng Wang [3,*], Wei Zhang [1], Wenqian Chen [3], Rui Li [1], Che Yue [1] and Shengjian Chen [1]

[1] North China Sea Marine Forecasting and Hazard Mitigation Center, Ministry of Natural Resources, Qingdao 266100, China
[2] College of Engineering, Ocean University of China, Qingdao 266100, China
[3] The State Key Laboratory of Coastal and Offshore Engineering, Ocean University of China, Qingdao 266100, China
* Correspondence: wzf1984@ouc.edu.cn

Abstract

Deep-learning post-processing can correct systematic errors in numerical wave forecasts, but its reliability under high-wave conditions remains unclear. We develop an error-correction framework in which deep networks predict the full future error trajectory of operational forecasts, and compare three architectures, a pure LSTM, a pure Transformer and a hybrid LSTM–Transformer, tuned identically with Bayesian optimization, using hourly observations from six buoys in the Bohai Sea and Yellow Sea over 2023–2025. On the full test sample all models improve the 72 h forecasts by 15.37% to 23.93% in RMSE, with the hybrid model attaining the lowest pooled RMSE; on the high-wave subset with SWH above 1.5 m, however, all three degrade the forecast, increasing RMSE by 7.58% to 9.05%. An exact decomposition of the squared-error change shows that fluctuation-capturing skill remains intact while the degradation is driven by the mean term: trained on samples dominated by low and moderate waves, the correctors learn a globally averaged adjustment whose sign opposes the reversed raw bias at high waves. Neither loss reweighting nor filtered training resolves this degradation at the 72 h horizon. Shortening the horizon to 24 h restores positive skill, with RMSE reductions of 4.43% to 7.88%, and reverses the ranking in favor of the pure LSTM. These findings indicate that architecture rankings are horizon-dependent, and they are consistent with the interpretation that extreme-condition skill is limited by the information content of the predictors rather than by model capacity.

Keywords: significant wave height; error correction; deep learning; high-wave conditions

Check for updates

Academic Editor: Eugen Rusu

Received: 21 September 2026
Revised: 10 October 2026
Accepted: 13 October 2026
Published: 16 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Ocean waves are fundamental to maritime safety, offshore engineering, coastal protection and disaster mitigation, and significant wave height (SWH) is the primary quantity in operational wave forecasting and early-warning services. In semi-enclosed shallow seas such as the Bohai Sea and the northern Yellow Sea of China, where winter storms and cold-air outbreaks frequently generate severe sea states, even moderate SWH forecast errors can delay warnings and cause economic losses. Third-generation spectral wave models such as WAM [1], WAVEWATCH III [2,3] and SWAN [4], now form the backbone of operational forecasting, and wave modeling has reached considerable maturity [5]. In such systems the wave model is typically forced by surface winds from a mesoscale atmospheric model such as WRF, which downscales global analyses and forecasts to the kilometer scale needed to

resolve coastal wind jets and rapidly developing cyclones over semi-enclosed seas [6,7]. Nevertheless, numerical forecasts inevitably contain systematic, state-dependent errors arising from imperfect wind forcing, parametrized physical processes such as whitecapping, depth-induced breaking and wave–current interaction, finite resolution and inaccurate bathymetry [5], which vary with lead time, sea state and local environment.

To reduce such errors without rerunning expensive numerical models, statistical post-processing has long been essential to operational forecasting, dating back to the model output statistics (MOS) of Glahn and Lowry [8] and later formalized through ensemble model output statistics (EMOS) [9] and quantile mapping [10]. Machine learning (ML) has substantially expanded this toolbox: neural networks can learn flexible nonlinear mappings from ensemble forecasts to predictive distributions [11], although no single method dominates under all conditions [12]. More broadly, deep learning has become a transformative tool for Earth system science [13], including medium-range weather forecasting models that rival numerical prediction [14].

In the ocean-wave community, ML has been applied along two main lines. The first replaces or emulates the numerical model and forecasts SWH directly, from early feed-forward networks [15,16] to recurrent and sequence-to-sequence networks [17], LSTM networks [18,19], hybrid gradient-boosting–LSTM systems [20] and attention-based architectures [21,22]. The second line, which is closer to the present study, keeps the numerical model in the loop and learns to correct its errors. Neural networks have long been used to improve numerical wave forecasts through assimilation-type corrections [23,24] and post-processing layers [25]. Zhang et al. [26] developed a deep-learning real-time bias correction for ECMWF SWH forecasts in the western North Pacific, achieving substantial error reductions up to 240 h lead time. Notably, Wang et al. [27] reported that correction skill deteriorates markedly under high-wave conditions, with RMSE improvement dropping to a few percent for SWH exceeding 4 m. Comparable correction strategies exist for other marine hazards, such as deep-learning offset correction of storm-surge forecasts [28].

The choice of network architecture is central to sequence-based error correction. LSTM networks [29] maintain gated memory well suited to the temporal persistence of forecast errors and are widely adopted in wave-height prediction [18,19]; convolutional–recurrent hybrids such as ConvLSTM [30] combine local memory with structured spatiotemporal patterns. Transformers use self-attention to model long-range dependencies without recurrence, and have performed strongly in multi-horizon time-series benchmarks and, recently, spatiotemporal wave forecasting [21,31,32]. Hybrid LSTM–Transformer designs attempt to combine local memory with global attention, but to our knowledge they have not been systematically compared against each other and their single-architecture counterparts for SWH error correction. Moreover, since corrector skill is sensitive to hyperparameters, we tune all models under an identical protocol using the Optuna Bayesian optimization framework [33].

Despite this progress, three gaps remain. First, virtually all existing studies evaluate corrections with all-sample metrics, whereas skill under high-wave conditions, precisely where accurate forecasts matter most, has rarely been assessed; the limited evidence suggests skill degrades with SWH, but the causes and remedies are poorly understood. Second, whether the relative advantage of different architectures (LSTM, Transformer, or hybrid) depends on forecast lead time has not been investigated for wave-error correction. Third, although class-imbalance remedies such as sample weighting and event-focused filtered training are well known in the broader ML literature, their effectiveness in rare high-wave regimes has not been quantified.

The present study addresses these gaps using hourly observations from six buoys in the Bohai Sea and Yellow Sea and the corresponding operational wave forecasts over

2023–2025. Within a unified error-correction framework, deep networks predict the full series of future forecast errors from t + 7 to t + 78 h in a single multi-step forecast, and the corrected forecast is obtained by subtracting the predicted error from the raw forecast. A pure LSTM, a pure Transformer and a hybrid LSTM–Transformer are tuned with Optuna under identical protocols, and their performance is evaluated on the full test sample, on a high-wave subset, and across two forecast horizons. Our main findings are as follows. (i) All three architectures significantly reduce the all-sample RMSE of the 72 h forecasts, with the hybrid model attaining the lowest pooled RMSE. (ii) Under high-wave conditions, however, all three models degrade the forecast: they apply a globally averaged downward adjustment that points the wrong way where waves are highest, and neither loss weighting nor filtered training at the 72 h horizon resolves this failure, which an error-decomposition criterion traces to the systematic component of the correction rather than to any loss of fluctuation-capturing skill. (iii) When the horizon is shortened to 24 h, big-wave-filtered training succeeds and the ranking reverses: the pure LSTM attains the lowest pooled RMSE, reducing the error by 7.88%. These results show that the best architecture is horizon-dependent rather than universal, and that extreme-condition skill is constrained by the information content of the predictors rather than by model capacity.

The remainder of this paper is organized as follows. Section 2 describes the buoy observations, numerical forecasts and quality control. Section 3 presents the correction framework, architectures, tuning protocol, big-wave training strategies and evaluation metrics. Section 4 reports the results, and Section 5 concludes.

2. Data

2.1. Buoy Observations

Hourly in situ observations of SWH were obtained from six offshore buoys moored in the Bohai Sea and the Yellow Sea (Figure 1), operated by the North China Sea Marine Forecasting and Disaster Mitigation Center, Ministry of Natural Resources. This north–south distribution allows the correction models to be evaluated under contrasting wave regimes: the Bohai Sea is a semi-enclosed shallow basin dominated by fetch-limited wind seas, whereas the Yellow Sea stations are additionally exposed to swell propagating from the open northwestern Pacific. The study period spans 00:00 1 January 2023 to 23:00 31 December 2025 (local time). As shown in Figure 2, the missing-data ratios remain below 4% at all stations throughout the period, except at B2 in 2024 (slightly above 5%), indicating generally good data completeness.

2.2. Numerical Wave Forecasts

The numerical forecasts were produced by a regional wave forecasting system based on the third-generation spectral wave model WAVEWATCH III® v6.07. The computational domain covers 117° E–127° E, 32° N–41° N, encompassing the entire Bohai Sea and Yellow Sea at a horizontal resolution of $1/36^\circ \times 1/36^\circ$. The model was forced by 3 km surface wind fields from the Weather Research and Forecasting (WRF) model, version 4.6.1, run in forecast mode with initial and boundary conditions from the NCEP Global Forecast System at $0.5^\circ \times 0.5^\circ$ resolution, with forecasts issued twice daily (00:00 and 12:00 UTC) up to a 168 h lead time. The physics configuration comprises the YSU planetary boundary layer scheme [34], the WSM six-class graupel microphysics scheme, the Kain–Fritsch (new Eta) cumulus parameterization, the thermal diffusion land surface scheme, and the Dudhia shortwave [35] and RRTM longwave [36] radiation schemes. The YSU scheme was selected because it represents boundary layer processes under unstable and rapidly developing conditions, which are typical of the strong wind events that generate high waves in the Bohai and Yellow Seas, and it is widely used in regional wind and wave studies over these seas.

Wave spectra at the open boundaries were taken from the operational 1/12° northwestern Pacific wave forecasting system of the NMFC (North China Sea Marine Forecasting Center). Bathymetry was synthesized from high-precision nautical charts of the China Maritime Safety Administration in coastal waters and the global ETOPO1 dataset [37] offshore. The ST4 source-term package was used for wind input and whitecapping dissipation [38]; detailed model settings are listed in Table 1. Hourly SWH fields were output, and values at the six buoy locations were extracted by bilinear interpolation for the correction experiments.

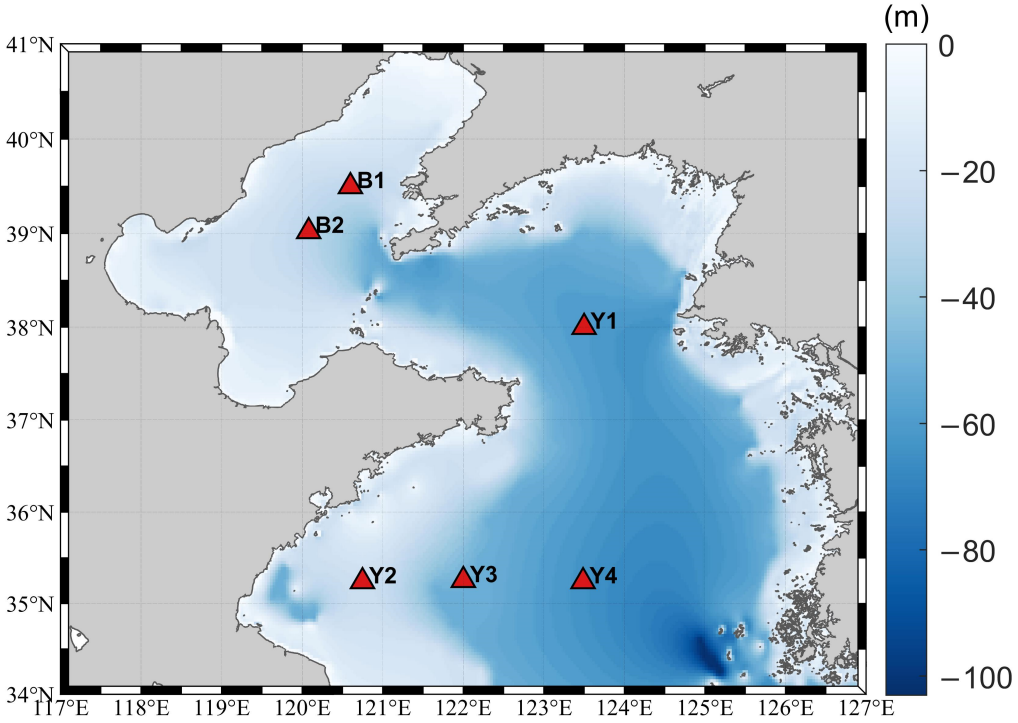


Figure 1. Locations of the six buoy stations in the Bohai Sea and the Yellow Sea.

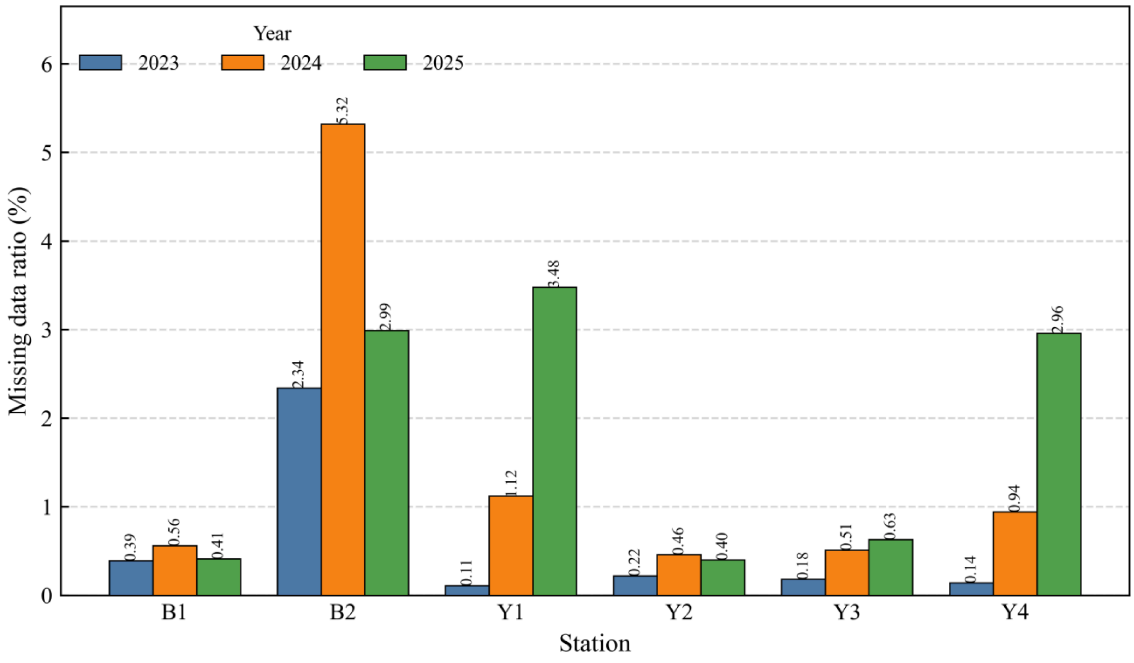


Figure 2. Annual missing-data ratios of SWH observations at the six buoy stations during 2023–2025.

Table 1. Models and parameters.

Models and Parameters	Attribute
PR3	Higher-order programs using Tolman averaging techniques
UQ	Third-order (UQ) dissemination programs
ST4	Input-dissipation packages: Ardhuin et al. source-term package
LN1	Selection of linear input: Cavaleri and Malanotte-Rizzoli with filter
NL1	Discrete interaction approximation (DIA)
BT1	JONSWAP base friction formula
REF0	No reflection
DB1	Selective depth caused Bartels-Janssen fracture
TR0	No triad interactions utilized
BS0	No bottom scattering utilized
WNT1	Selection of wind interpolation approach (time): approximately linear speed interpolation
WNX1	Selection of wind speed interpolation approach (spatial): approximate linear velocity insertion
CRT1	Choosing the currently interpolated approach (time): linear interpretation
CRX1	Select current insertion approach (spatial): approximate linear velocity insertion
NOGRB	User provided switches for GRIB packages: packages not covered

3. Methodology

3.1. Error-Correction Framework

Let $H_m(t{+}\mathrm{h})$ and $H_o(t{+}\mathrm{h})$ denote the numerical-model forecast and the buoy-observed SWH, respectively, at lead time h for a forecast issued at time t. The raw forecast error is defined as

$$e(t+h) = H_o(t+h) - H_m(t+h) \tag{1}$$

Instead of predicting SWH directly, the deep-learning models developed in this study predict the forecast error itself. For each issue time t, the model receives two groups of inputs. The first group describes the recent error behavior of the numerical model over the most recent 24 h (t − 17 to t + 6 h, aligned with the target window): the historical forecast errors $e(t-k)$ and the corresponding historical model forecasts $H_m(t-k)$. The second group describes the target period: the raw model forecasts $H_m(t{+}\mathrm{h})$ for the 72 lead times $\mathrm{h} = 7, \ldots, 78$, together with a normalized lead-time coordinate that informs the network of the position of each output step within the forecast horizon. Note that the observed SWH is not used as a direct input; observational information enters the model exclusively through the historical error channel. The overall error-correction framework is illustrated in Figure 3. From these inputs the network produces the complete future error trajectory in a single multi-step (one-shot) forecast:

$$\hat{e}(t+h) = f_\theta(\mathbf{X}(t)), \quad h = 7, 8, \ldots, 78, \tag{2}$$

where $\mathbf{X}(t)$ denotes the above input feature set at issue time t, f_θ is a deep network with trainable parameters θ, and the output vector $\hat{e}(t{+}7), \ldots, \hat{e}(t{+}78)$ contains the predicted errors for all 72 lead times simultaneously. The corrected forecast is then obtained by subtracting the predicted error from the raw model forecast:

$$H_c(t+h) = H_m(t+h) - \hat{e}(t+h) \tag{3}$$

Three features of this design deserve emphasis. First, error prediction keeps the corrected forecast tied to the numerical model: when the predicted error is small or unskillful, the correction degrades gracefully toward the raw forecast rather than drifting from the physical solution. Second, the one-shot strategy generates all 72 lead times in a single

forward pass, avoiding the error accumulation of recursive forecasting. Third, a single unified model is trained on samples pooled from all six stations, so that the network can exploit the error structure shared among stations rather than learning each station in isolation.

The target window starts at t + 7 h because the first hours of the operational forecast are effectively spin-up with small errors of limited practical value. The most recent 24 h of errors serve as predictors because forecast errors in this region exhibit strong short-range persistence. All predictors and target errors are standardized station-by-station using training-period statistics only, preventing information leakage; samples with missing values in any required column are excluded, with the impact on data availability reported in Section 2. The chronological split assigns the first 60% of the study period to training, the next 20% to validation (early stopping and hyperparameter optimization), and the final 20% to independent testing, with no shuffling, so that the test period represents a genuine operational situation on unseen data.

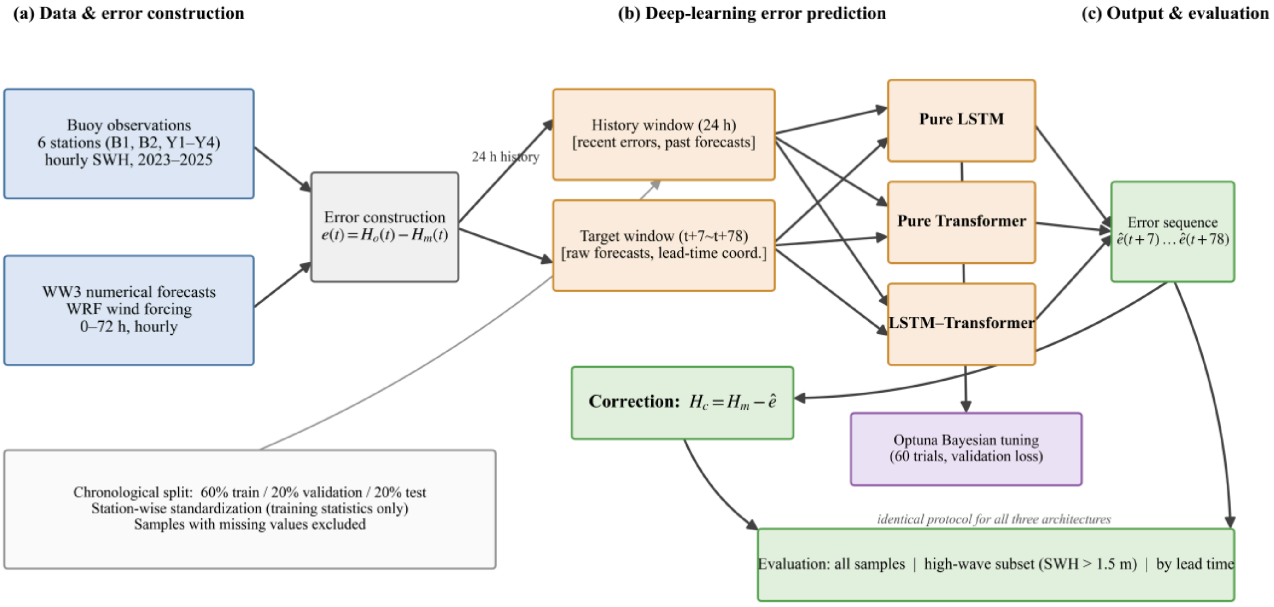


Figure 3. Overall framework of the deep-learning error-correction system.

3.2. Model Architectures

All three models receive the same input information: the recent forecast errors, the past model forecasts, the raw forecasts over the target window and the lead-time coordinate. This predictor set contains all the information available to a corrector at issue time, following the established practice of neural-network error correction for wave forecasts [16]. They differ only in how this information is organized and how temporal dependencies are extracted, allowing the effect of the architecture itself to be isolated. The three architectures are illustrated in Figure 4. Colors group the functional components: blue boxes are inputs, yellow boxes are projection and positional-encoding layers, pink boxes are LSTM layers, green boxes are attention-based blocks, purple boxes are the output head and prediction, and the orange box is the final error-correction step. Red arrows denote the key and value flow from the encoder memory to the cross-attention layers.

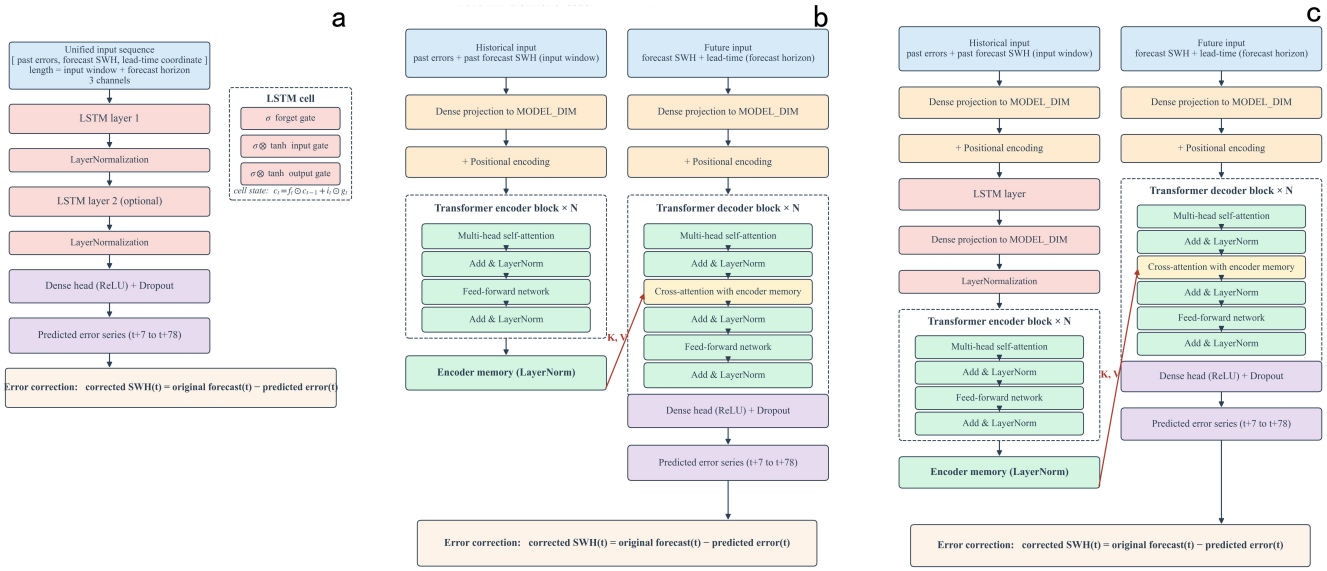


Figure 4. Schematic architectures of (**a**) the pure LSTM, (**b**) the pure Transformer, and (**c**) the hybrid LSTM–Transformer error-correction models.Colors group the functional components (see text for details).

The pure LSTM model reads a single unified sequence that concatenates the history and target windows, with three channels per time step (past errors, model forecasts, lead-time coordinate). The sequence is processed by one or two stacked LSTM layers [29], each followed by layer normalization; the terminal hidden state passes through a fully connected layer with ReLU activation and dropout, and a final linear layer produces the complete 72-dimensional error vector at once. Owing to its gated recurrent memory, this architecture is expected to be particularly effective at exploiting the strong short-range persistence of forecast errors.

The pure Transformer model organizes the inputs in encoder–decoder form: the encoder reads the 24 h history sequence, and the decoder the target-window sequence. All recurrent connections are replaced by multi-head self-attention [31]; each block additionally contains a position-wise feed-forward network, residual connections and layer normalization, and sinusoidal positional encodings are added to the input embeddings to preserve time ordering. In the decoder, cross-attention allows every output step to attend to the encoded history. Since attention relates any two time steps through a single operation regardless of distance, this architecture should favor the extraction of long-range dependencies in the error series.

The hybrid model combines the two mechanisms asymmetrically. In the encoder, the embedded history sequence is first processed by an LSTM layer returning the full output sequence; after linear projection and layer normalization, the outputs are refined by stacked Transformer encoder blocks, forming the memory. The decoder is entirely attention-based, with cross-attention reading the encoded memory, followed by a fully connected prediction head. The rationale is that the LSTM pathway provides local memory and an inductive bias toward temporally adjacent information when compressing the history, while the attention pathway aggregates information across the whole window without distance decay, potentially outperforming both single-architecture models, at the cost of a larger parameter count.

For all three models, the 72-dimensional error prediction is produced by a time-distributed linear output layer. Training uses the Adam optimizer implemented in TensorFlow 2.15 (Google LLC, Mountain View, CA, USA) with mean squared error loss , gradient clipping, early stopping and learning-rate reduction on the validation loss. The main

hyperparameters, including hidden dimensions, number of heads and blocks, dropout rate, L2 coefficient, batch size and learning rate, are tuned separately for each architecture with Optuna [33] under an identical protocol described in Section 3.3, ensuring a fair comparison.

3.3. Bayesian Hyperparameter Optimization

The skill of deep-learning models is highly sensitive to hyperparameters, and a comparison of architectures is only meaningful if each model reaches its own best configuration. All three models were therefore tuned with the Optuna framework, which implements Bayesian optimization via the Tree-structured Parzen Estimator (TPE); by modeling the objective from the history of completed trials, TPE samples preferentially in promising regions and is substantially more efficient than grid or random search under a limited budget.

For each trial, a candidate model was trained on the training set for at most 60 epochs with early stopping (patience of 6 on the validation loss), and evaluated by the validation mean squared error of the predicted errors; the test set was never touched during optimization. After 60 trials, the best configuration was retrained for up to 100 epochs, with the validation set used only for early stopping (patience of 10) and learning-rate scheduling, and the final model was evaluated once on the independent test set.

The protocol was identical for the three architectures: 60 trials, and the same data splits, objective function, per-trial training budget and random seed. The search spaces were designed to be comparable in width: shared hyperparameters such as hidden dimensions, dropout rates, L2 coefficients, learning rates and batch sizes have identical ranges, while architecture-specific ones, namely the number of LSTM layers and the numbers of heads and Transformer blocks, have their own ranges. The complete search spaces and optimal configurations are listed in Table 2, where the best validation loss denotes the minimum MSE of the standardized prediction errors during tuning.

Table 2. Hyperparameter search spaces and the optimal configurations selected by Optuna for the three models (60 TPE trials each, minimizing validation MSE).

Hyperparameter	Pure LSTM	Pure Transformer	LSTM–Transformer
MODEL_DIM	—	(64,96,128) Best: 128	(64,96,128) Best: 64
NUM_LSTM_LAYERS	1,2 Best: 1	—	—
LSTM_UNITS	(32,64,96,128,160) Best: 160	—	(32,64,96,128) Best: 96
NUM_HEADS	—	(2,4,8) Best: 2	(2,4,8) Best:2
NUM_TRANSFORMER_BLOCKS	—	1,2 Best: 1	1,2 Best: 2
FF_DIM	—	(64,96,128,192,256) Best: 192	(64,96,128,192,256) Best: 256
DENSE_UNITS	(32,64,96,128,160) Best: 96	(32,64,96,128) Best: 96	(32,64,96,128) Best: 128
DROPOUT_RATE	(0,0.05,0.10,0.15,0.20) Best:0.2	(0,0.05,0.10,0.15,0.20) Best: 0.05	(0,0.05,0.10,0.15,0.20) Best: 0.0
L2_REG	10^{-7}–10^{-4} Best: 4.2×10^{-6}	10^{-7}–10^{-4} Best: 7.3×10^{-7}	10^{-7}–10^{-4} Best: 1.2×10^{-7}
LEARNING_RATE	10^{-5}–5×10^{-4} Best: 10^{-4}	10^{-5}–5×10^{-4} Best: 4×10^{-4}	10^{-5}–5×10^{-4} Best: 10^{-4}
BATCH_SIZE	(32,64,96,128) Best: 96	(32,64,96,128) Best: 96	(32,64,96,128) Best: 64
Best validation loss	0.9795	0.9012	0.8951

3.4. Big-Wave-Oriented Training Strategies

High-wave events are relatively infrequent in the training data: at the six buoy stations, observations with SWH exceeding 1.5 m account for about 12.5% of all hourly records.

Models trained by minimizing the mean squared error over all samples are therefore dominated by calm and moderate sea states, and their skill under high-wave conditions, where accurate forecasts are most needed, is not guaranteed. To examine whether this class imbalance can be alleviated by training-side interventions, two big-wave-oriented strategies were applied to all three architectures. Both modify the training procedure only; validation and test evaluations remain identical to the baseline experiments, so any change in skill is attributable solely to the training strategy.

The first strategy retains all training samples but assigns larger weights to the lead times at which the observed wave is high. For the target window of each sample, the weight at lead time h is defined as

$$w(h) = 1 + \alpha \cdot max(H_o(t+h) - H_{th},\, 0) \tag{4}$$

where $H_o(t+\mathrm{h})$ is the observed SWH, $H_{\mathrm{th}} = 1.5$ m is the high-wave threshold, and $\alpha = 1.0$ controls the strength of the weighting. The weight equals 1 for calm and moderate conditions and grows linearly with the exceedance above the threshold. The weighted mean squared error is used as the training loss. Because the weighting depends on the observed SWH, which is available at training time for every target lead time, no additional information is required at inference.

The second strategy modifies the training set itself: only the samples whose target window contains at least one observed SWH exceeding the threshold are retained for training, i.e.,

$$\max_h H_o(t+h) > H_{th}, \quad H_{th} = 1.5\,m \tag{5}$$

All samples failing this criterion are discarded, so that the model is trained exclusively on high-wave events. This strategy maximizes the exposure of the network to large-wave error patterns, at the cost of a substantially reduced training set. The filtered models are therefore expected to suffer from reduced statistical robustness, which is examined in Section 4.

For both strategies, all hyperparameters were re-tuned with Optuna under the identical protocol described in Section 3.3, because the optimal configuration under a modified training objective or training set is not necessarily the same as that of the baseline. The evaluation protocol, including the all-sample metrics and the high-wave subset metrics, was kept unchanged in all cases. The same filtering strategy was additionally applied to a 24 h forecast-horizon configuration.

To assess the statistical significance of the RMSE differences while accounting for temporal dependence among the strongly overlapping forecast samples, we applied a block bootstrap with 1000 resamples, using blocks defined as all target points of one station within one calendar day, and report 95% percentile confidence intervals.

3.5. Evaluation Metrics and the High-Wave Subset

3.5.1. Metrics

The raw and corrected forecasts are evaluated against the buoy observations using four complementary metrics: the root mean squared error (RMSE), the mean absolute error (MAE), the mean error (bias), and the coefficient of determination (R^2):

$$\begin{aligned} \mathrm{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(H_{p,i} - H_{o,i}\right)^2}, \quad \mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|H_{p,i} - H_{o,i}\right|, \\ \mathrm{Bias} &= \frac{1}{N}\sum_{i=1}^{N}\left(H_{p,i} - H_{o,i}\right), \quad R^2 = 1 - \frac{\sum_{i=1}^{N}\left(H_{p,i} - H_{o,i}\right)^2}{\sum_{i=1}^{N}\left(H_{o,i} - \overline{H}_o\right)^2} \end{aligned} \tag{6}$$

where $H_{p,i}$ and $H_{o,i}$ denote the predicted (raw or corrected) and observed SWH of the i-th record, and N is the number of records. RMSE and MAE measure the overall error magnitude; bias measures the systematic component of the error, with positive values indicating overestimation and negative values indicating underestimation. It is particularly important in this study because the correction acts through a shift in the forecast; R^2 measures the fraction of observed variance explained by the forecast. For a corrected forecast to be considered an improvement, it must reduce RMSE relative to the raw forecast.

3.5.2. High-Wave Subset

In addition to the all-sample evaluation, the skill is assessed separately under high-wave conditions. The high-wave subset is defined point-wise: an individual record enters the subset if the observed SWH at that record exceeds 1.5 m,

$$H_o(t + \mathrm{h}) > 1.5\,\mathrm{m} \tag{7}$$

Note that the selection is based on the observation, not on the forecast, so that the subset is identical for the raw and all corrected forecasts and the comparison remains fair. All metrics in Section 4 are reported both over all samples and over this subset. The 1.5 m threshold follows the operational alerting practice for the study area. Its robustness is examined below using alternative absolute thresholds of 1.0 and 2.0 m and station-specific 90th and 95th percentile definitions.

3.5.3. A Covariance Criterion for Correction Skill

The conditions under which an error-correction scheme can improve a forecast can be made explicit. With the correction defined in Equation (3), the corrected error in SWH is e_c(t + h) = e(t + h) − ê(t + h). Expanding the squared error and taking expectations gives the exact decomposition

$$\Delta\mathrm{RMSE}^2 \equiv \mathrm{MSE_c} - \mathrm{MSE_{raw}} = [\mathrm{Var}(\hat{e}) - 2\mathrm{Cov}(\hat{e}, e)] + \left[\mathrm{E}[\hat{e}]^2 - 2{\cdot}\mathrm{E}[e]{\cdot}\mathrm{E}[\hat{e}]\right] \tag{8}$$

which separates the contribution of the error fluctuations (the fluctuation term, first bracket) from that of the systematic offsets. When the true and predicted errors are approximately centered, i.e., their means are much smaller than their standard deviations, the mean term is negligible, and Equation (8) reduces to $\Delta\mathrm{RMSE}^2 \approx \mathrm{Var}(\hat{e}) - 2\mathrm{Cov}(\hat{e}, e)$. The correction then reduces the RMSE if and only if

$$\mathrm{Cov}(\hat{e}, e) > (1/2){\cdot}\mathrm{Var}(\hat{e}) \tag{9}$$

which, in other words, means the predicted error must covary sufficiently strongly with the true error. This criterion is architecture-independent: it depends only on the information about the future error contained in the predictors.

Two implications of the full decomposition in Equation (8) deserve emphasis. First, on the full sample, where forecast biases are modest, the centered approximation holds and the fluctuation term governs the outcome. Second, the mean term becomes decisive whenever the mean predicted error $\mathrm{E}[\hat{e}]$ is large and of the opposite sign to the mean true error $\mathrm{E}[e]$: the cross term $-2{\cdot}\mathrm{E}[e]{\cdot}\mathrm{E}[\hat{e}]$ is then positive and can overwhelm a nominally skillful fluctuation term. This second regime is not hypothetical; it is precisely what the high-wave subset realizes, where the sign reversal of the raw bias turns the mean term into the dominant source of degradation.

4. Results

4.1. Hyperparameter Convergence and Optimal Architectures

Figure 5 shows the optimization histories of the three models over the 60 TPE trials; scatter points denote the validation loss of individual trials and solid lines the best validation loss achieved up to that trial. All three models converge well within the budget: the pure LSTM reaches its optimum of 0.9795 already at trial 24, whereas the pure Transformer and the hybrid model continue improving until trials 52 and 37, reaching 0.9012 and 0.8951. The attention-based models thus benefit more from the search budget because their richer spaces of heads, blocks, and feed-forward dimensions take more trials to explore, while the smaller LSTM space is exhausted quickly. The validation ranking of hybrid < Transformer < LSTM matches the ranking on the independent test set, indicating that tuning identified genuine architectural differences rather than overfitting the validation set.

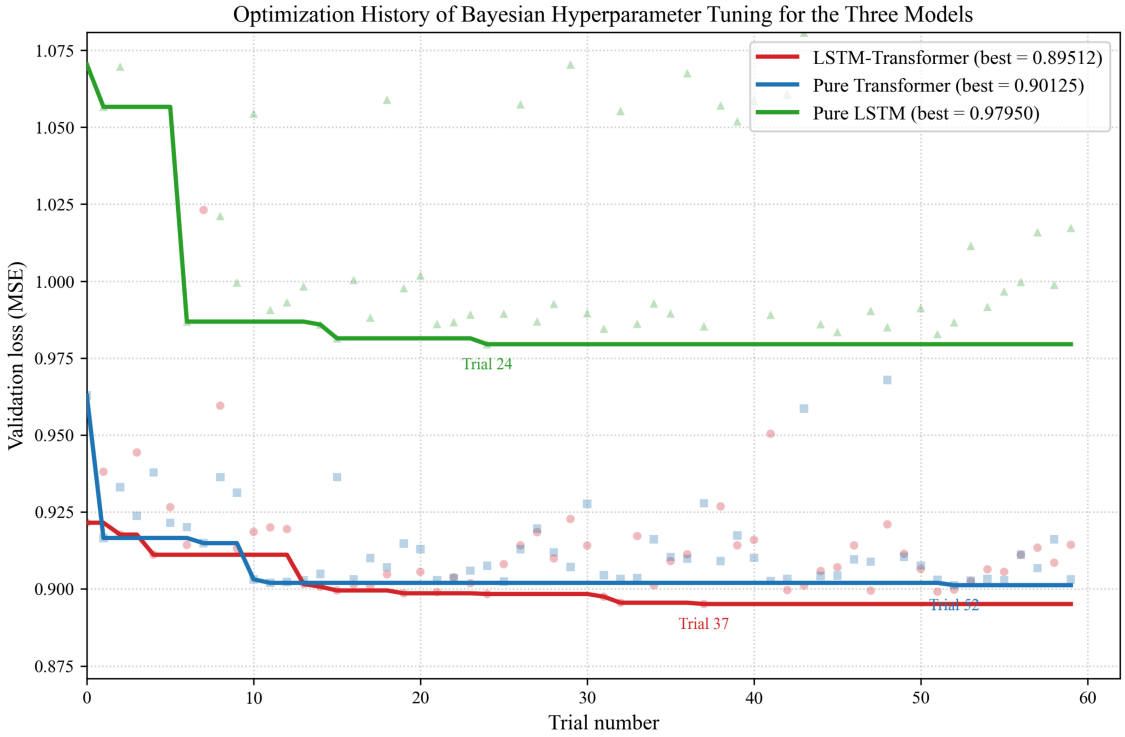


Figure 5. Optimization histories of the Optuna TPE tuning for the three models. Scatter points show the validation loss of individual trials; solid lines show the best validation loss achieved up to each trial. Numbers indicate the trial at which each model reached its optimum.

The optimal configurations differ systematically between the architectures. The pure LSTM selects a single LSTM layer with 160 units, which is the widest value in its search space, indicating that it compensates for the absence of attention by maximizing recurrent capacity rather than by stacking layers. The pure Transformer prefers a wide but shallow configuration: an embedding dimension of 128, two heads, one encoder/decoder block, and the largest feed-forward dimension of 192. The hybrid model trades width for depth, combining a narrower embedding of 64 with a 96-unit LSTM encoder and two Transformer blocks. Regularization differs markedly: the pure LSTM requires strong dropout (0.20) whereas the attention-based models need almost none (0.05 and 0.0), and the optimal L2 coefficients of the attention-based models (7.3×10^{-7} and 1.2×10^{-7}) are an order of magnitude smaller than that of the LSTM (4.2×10^{-6}). The optimal learning rates are of the order of 10^{-4} for all three models (1×10^{-4}, 4×10^{-4} and 1×10^{-4}).

The three optimized architectures contain 87,208, 313,665 and 445,633 trainable parameters for the pure LSTM, pure Transformer and hybrid models, respectively. Tuning each architecture with 60 Optuna trials required approximately 30 min on a single NVIDIA L20 GPU (NVIDIA Corp., Santa Clara, CA, USA). Although the hybrid model is the largest, it exceeds the pure Transformer by only 42%, and the block-bootstrap analysis reported in Section 4.2 shows that its advantage over the pure Transformer is not statistically significant. On the high-wave subset, the three models degrade by statistically indistinguishable amounts despite a five-fold difference in parameter count. Model capacity therefore does not drive the performance differences reported in this study.

4.2. Correction Performance Evaluated on All Samples

Table 3 summarizes the all-sample performance of the raw and corrected SWH forecasts on the test set, evaluated over lead times t + 7 h to t + 78 h and pooled across all six stations. Three features stand out.

Table 3. All-sample performance of the raw and corrected SWH forecasts on the test set.

Station	Model	RMSE (m)	MSE (m^2)	MAE (m)	R^2	I_{RMSE} (%)
B1	LSTM	0.362/0.331	0.132/0.109	0.258/0.237	0.599/0.666	8.56%
	Transformer	0.362/0.306	0.132/0.094	0.258/0.219	0.599/0.715	15.47%
	L-Transformer	0.362/0.304	0.132/0.092	0.258/0.215	0.599/0.719	**16.02%**
B2	LSTM	0.359/0.341	0.129/0.116	0.252/0.242	0.609/0.647	5.01%
	Transformer	0.359/0.316	0.129/0.100	0.252/0.222	0.609/0.698	11.98%
	L-Transformer	0.359/0.315	0.129/0.099	0.252/0.219	0.609/0.700	**12.26%**
Y1	LSTM	0.399/0.357	0.160/0.128	0.285/0.249	0.633/0.706	10.53%
	Transformer	0.399/0.352	0.160/0.124	0.285/0.245	0.633/0.716	11.78%
	L-Transformer	0.399/0.349	0.160/0.122	0.285/0.242	0.633/0.720	**12.53%**
Y2	LSTM	0.439/0.337	0.193/0.114	0.314/0.210	0.098/0.354	23.23%
	Transformer	0.439/0.286	0.193/0.082	0.314/0.185	0.098/0.536	34.85%
	L-Transformer	0.439/0.282	0.193/0.080	0.314/0.179	0.098/0.547	**35.76%**
Y3	LSTM	0.423/0.334	0.179/0.112	0.299/0.216	0.382/0.614	21.04%
	Transformer	0.423/0.299	0.179/0.089	0.299/0.203	0.382/0.691	**29.31%**
	L-Transformer	0.423/0.299	0.179/0.089	0.299/0.199	0.382/0.691	**29.31%**
Y4	LSTM	0.400/0.333	0.161/0.111	0.275/0.221	0.581/0.709	16.75%
	Transformer	0.400/0.303	0.161/0.092	0.275/0.207	0.581/0.760	24.25%
	L-Transformer	0.400/0.301	0.161/0.091	0.275/0.203	0.581/0.763	**24.75%**
ALL	LSTM	0.397/0.336	0.158/0.113	0.279/0.227	0.514/0.652	15.37%
	Transformer	0.397/0.304	0.158/0.092	0.279/0.209	0.514/0.716	23.43%
	L-Transformer	0.397/0.302	0.158/0.091	0.279/0.204	0.514/0.719	**23.93%**

Note: In this and the following tables, I_{RMSE} is computed from the rounded RMSE values shown, so that all values can be reproduced directly from the table entries.Bold values indicate the largest I_{RMSE} in each group, i.e., the most effective correction.

First, the correction is effective at every station and for every architecture. The pooled RMSE decreases from 0.397 m for the raw forecast to 0.336, 0.304 and 0.302 m for the pure LSTM, pure Transformer and hybrid models, corresponding to reductions of 15.37%, 23.43% and 23.93%, while R^2 increases from 0.514 to at most 0.719. Station-level improvements are consistent, with I_RMSE ranging from 5.01% for the LSTM at B2 to 35.76% for the hybrid model at Y2, and no case of degradation. Block-bootstrap confidence intervals confirm that the all-sample improvements are statistically significant for all three architectures, with raw-minus-corrected RMSE differences of +0.068 [0.060, 0.075], +0.094 [0.083, 0.104] and +0.097 [0.086, 0.109] m for the LSTM, Transformer and hybrid models. The hybrid model significantly outperforms the pure Transformer, although by a small margin (+0.003 [0.001, 0.006] m).

Second, the architecture ranking agrees with the validation ranking during tuning: the hybrid model best, the pure Transformer second, and the pure LSTM last. This confirms that the protocol generalizes to unseen data. The advantage of the attention-based models over the pure LSTM is largest at the shallow, low-energy stations (e.g., Y2: 35.76% vs. 23.23%; Y3: ~29% vs. 21.04%), where the raw forecast is least skillful (R^2 = 0.098 at Y2) and the error carries the most exploitable structure. This ranking is also evident in the Taylor diagram (Figure 6): at every station the corrected forecasts shift toward higher correlation than the raw forecast, with the largest displacement at the least skillful stations, most visibly Y2, whose correlation rises from about 0.82 to 0.88 while its normalized standard deviation relaxes from above 1 toward unity; at the already skillful Bohai stations the four symbols nearly overlap, consistent with the small I_{RMSE} there.

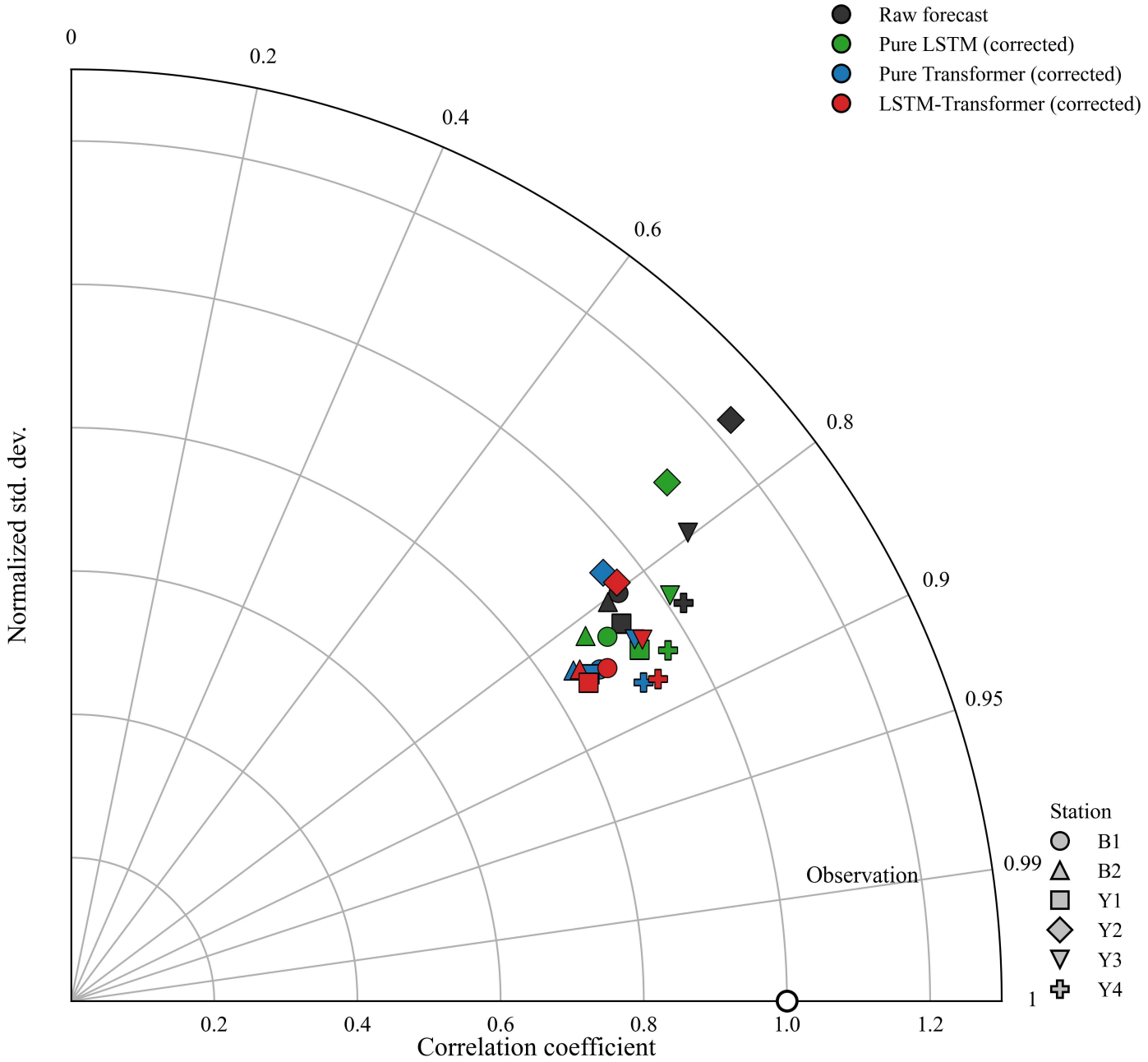


Figure 6. Taylor diagram of the raw and corrected SWH forecasts. One symbol per model per station; colors denote models and shapes denote stations.

Third, the spatial pattern of I_{RMSE} is inversely related to raw-forecast skill: the largest gains occur at Y2 and Y3, where raw RMSE exceeds 0.42 m and R^2 falls below 0.39, whereas the smallest gains occur at B1 and B2, where the raw forecast is already comparatively good ($R^2 \approx 0.60$). This indicates that the correction primarily removes the systematic, structured part of the forecast error, which is largest where the raw model is least skillful, rather than irreducible random noise.

Figure 7a shows the RMSE as a function of lead time. The raw forecast error grows from about 0.25 m at t + 7 h to more than 0.50 m beyond t + 72 h. All three corrected curves lie clearly below the raw curve at every lead time, and the absolute benefit increases with lead time: at short leads (t + 7–15 h) the corrected RMSE is 0.13–0.22 m, while at t + 78 h it remains at 0.40–0.44 m against 0.52 m for the raw forecast. The two Transformer-based models track each other closely throughout, whereas the pure LSTM is consistently worse from the shortest leads onward (about 0.20 m versus 0.13–0.15 m at t + 7–10 h), the gap widening further beyond approximately t + 28 h. This indicates that attention better exploits both the short- and long-range error structure in the inputs. Figure 7b compares the error distributions. The raw forecast error is positively skewed with a pronounced positive tail, reflecting the systematic SWH overestimation by the numerical model. After correction, all three distributions become sharper, approximately symmetric and centered near zero, with the positive bias and long right tail largely removed. This confirms that the models correct not only the bulk error but also its systematic high-bias component.

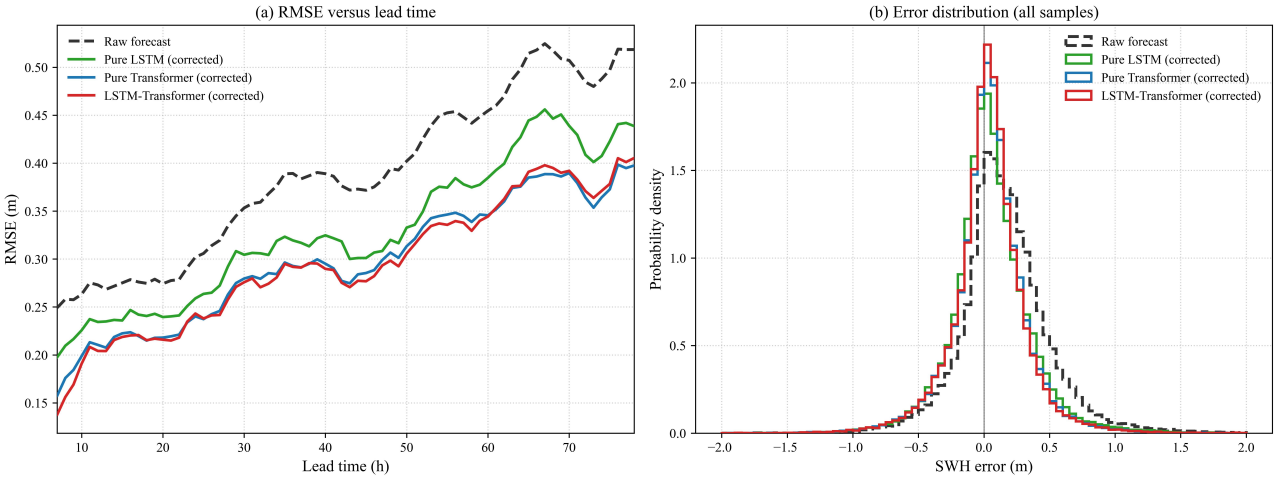


Figure 7. (**a**) RMSE versus lead time and (**b**) error distributions of the raw and corrected SWH forecasts.

4.3. Degradation on the High-Wave Subset and the Failure of Targeted Remedies

4.3.1. A Consistent Failure Across Architectures

The all-sample gains reported in Section 4.2 conceal a systematic failure under high-wave conditions. Following the point-level definition of Section 3.5.2, we extracted all test records with observed SWH exceeding 1.5 m, which account for 8.5% of the target points, and re-evaluated the raw and corrected forecasts on this subset.

All three architectures degrade the forecast instead of improving it. The raw RMSE on the subset is 0.475 m, whereas the corrected RMSE increases to 0.512, 0.518 and 0.511 m for the pure LSTM, pure Transformer and hybrid models, corresponding to I_{RMSE} of −7.79%, −9.05% and −7.58%. On the subset, the raw R^2 is already as low as 0.060, and correction drives it negative, down to −0.087 or even −0.118. The degradation occurs at nearly all lead times rather than only at long leads (Figure 8b), which points to a cause at the training level rather than at any particular horizon. The three architectures also degrade by almost the same amount, with differences below 0.01 m, in clear contrast to their well-separated

ranking on the full sample. The high-wave degradation is statistically significant for the LSTM and Transformer models, with raw-minus-corrected RMSE differences of −0.037 [−0.056, −0.018] and −0.043 [−0.076, −0.005] m, and marginally non-significant for the hybrid model (−0.034 [−0.069, +0.005] m). Pairwise differences between architectures on this subset are small: the LSTM does not differ significantly from the other two models, while the hybrid model degrades significantly less than the pure Transformer (+0.009 [0.001, 0.016] m). The failure therefore appears to be inherited from the training signal rather than from the models themselves. Figure 9 further shows that this failure is not an artifact of the chosen 1.5 m threshold. Across five contiguous sea-state bins, the benefit of correction decreases monotonically with observed SWH: the error boxes narrow around zero in the two lowest bins but widen and shift to negative beyond 1.5 m. This behavior is robust to the definition of high waves. With a 1.0 m threshold the correction remains beneficial (I_{RMSE} of +6.56% to +10.56%), while with a 2.0 m threshold the degradation deepens to −12.03% to −17.07%. Station-specific percentile thresholds, defined individually at each station by the 90th and 95th percentiles of its observed SWH, account for the different wave regimes of the Bohai Sea and Yellow Sea: the resulting P90 thresholds range from 1.2 m at Y2 to 1.9 m at Y1, confirming that a fixed 1.5 m threshold corresponds to very different exceedance frequencies across stations. Under these relative definitions the pooled degradation persists, with I_{RMSE} between −2.84% and −10.44%, and the station-level pattern in Figure 10 reinforces the mechanism: correction degrades at the four stations whose relative thresholds lie in the genuinely high sea-state range, whereas it remains skill-positive at Y2, whose P90 threshold of 1.2 m samples the moderate range where correction is beneficial, and it remains neutral-to-positive at Y3, where the raw forecast is nearly unbiased. The transition from skill-positive to skill-negative correction is thus continuous and tied to the sea-state regime rather than to the choice of a fixed or relative threshold. The bias columns of Table 4 provide direct evidence for the mechanism. On the full sample the raw forecast is positively biased (+0.165 m), and the correctors rightly learn to subtract a positive offset. On the high-wave subset, however, the raw bias reverses to −0.132 m: the numerical model underestimates SWH during high-wave events, as is typical under rapidly growing wind seas. The correctors, optimized on the full training set, cannot represent this sign reversal: the learned correction remains dominated by the majority pattern and subtracts an offset from forecasts that are already too low. The corrected bias is therefore amplified to −0.237, −0.299 and −0.278 m, and the MAE increases from 0.350 m to 0.387–0.404 m.

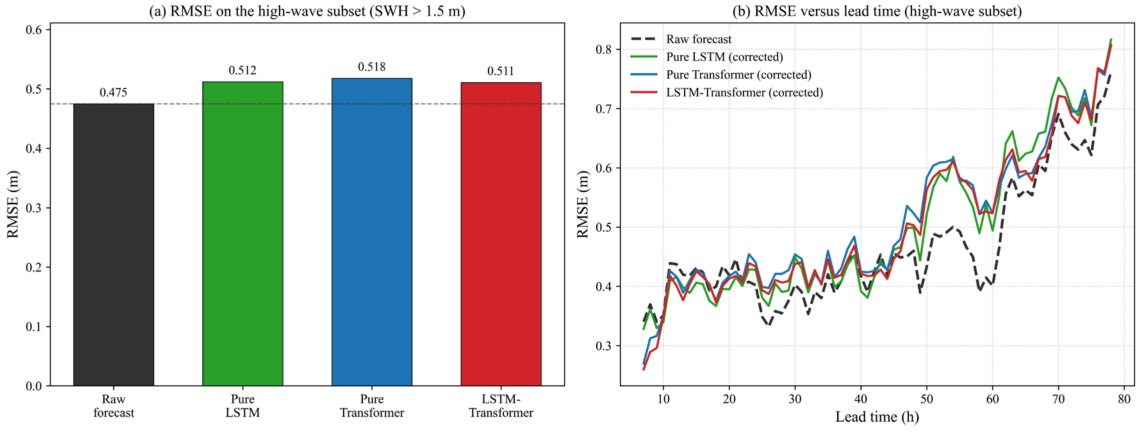


Figure 8. (**a**) RMSE of the raw and corrected forecasts on the high-wave subset (observed SWH > 1.5 m), and (**b**) RMSE versus lead time on the same subset. Test set, all six stations pooled.

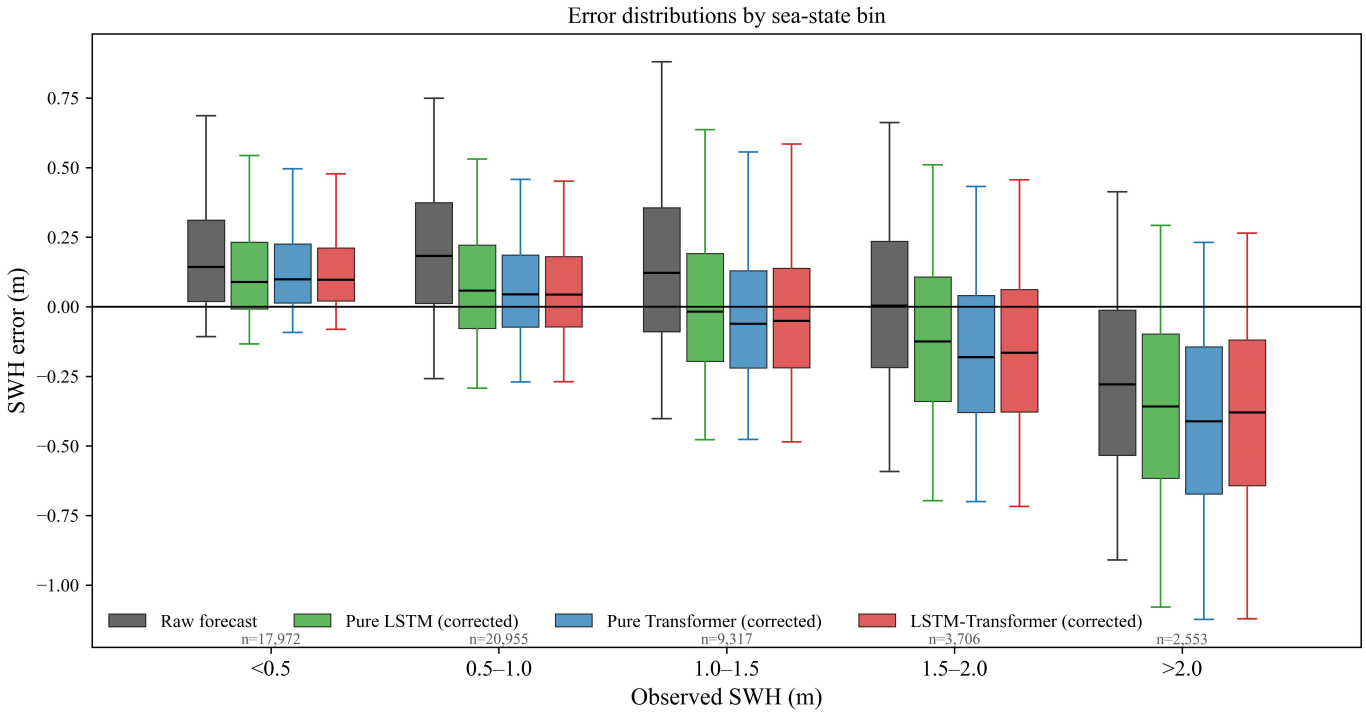


Figure 9. Distributions of the SWH error (forecast/corrected minus observed) within five sea-state bins defined by the observed SWH. Boxes span the interquartile range with the median marked; whiskers extend to the 5th and 95th percentiles. Sample sizes of each bin are annotated at the bottom.

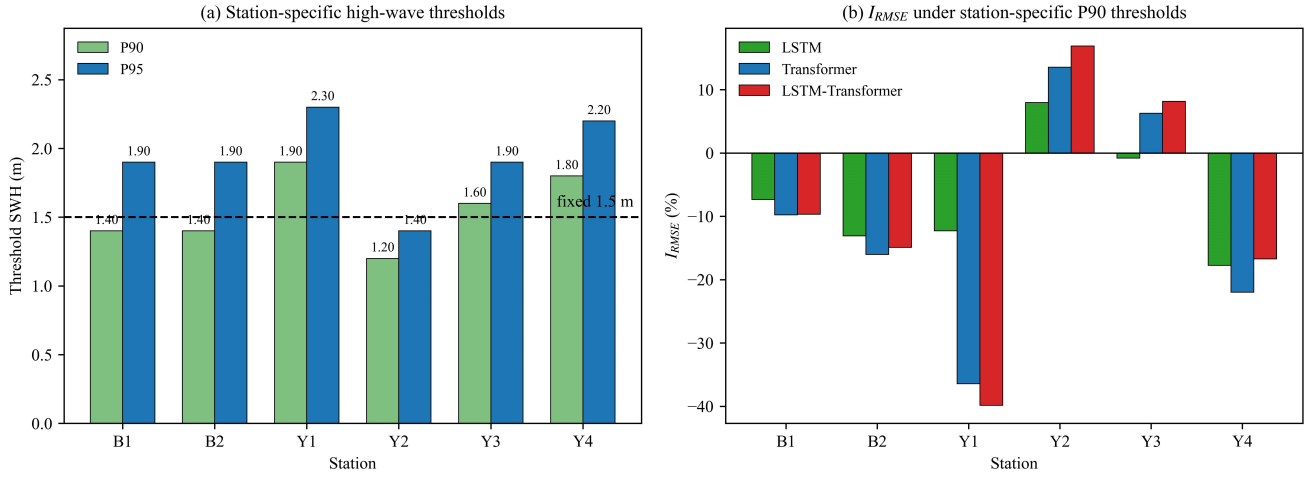


Figure 10. Station-specific high-wave thresholds and correction skill. (**a**) The 90th and 95th percentiles of observed SWH at each station, with the fixed 1.5 m threshold shown as a dashed line. (**b**) I_{RMSE} of the three models above each station's own 90th percentile threshold.

Table 4. Performance of the raw and corrected SWH forecasts on the high-wave subset.

Model	RMSE_raw	RMSE_cor	MAE_raw	MAE_cor	Bias_raw	Bias_cor	I_{RMSE} (%)
LSTM	0.475	0.512	0.350	0.387	−0.132	−0.237	−7.79
Transformer	0.475	0.518	0.350	0.404	−0.132	−0.299	−9.05
L-Transformer	0.475	0.511	0.350	0.395	−0.132	−0.278	−7.58

The high-wave degradation is not confined to individual stations, but its magnitude varies systematically with the local bias of the raw forecast. Station-level I_{RMSE} ranges from −0.04% to −31.86% (Figure 11), and the bias sign reversal occurs at five of the six stations. The exception is instructive: at Y3, where the raw forecast is nearly unbiased under high waves (bias of −0.023 m), the correction does not degrade and the two attention-based models even improve (I_{RMSE} of +7.04% and +8.75%), consistent with the decomposition of Section 3.5.3, in which the mean-term penalty vanishes when E[e] is close to zero. The station-level pattern therefore supports the proposed mechanism rather than contradicting it: the pooled degradation is shared by all stations, but its strength is modulated by the local systematic bias of the raw forecast.

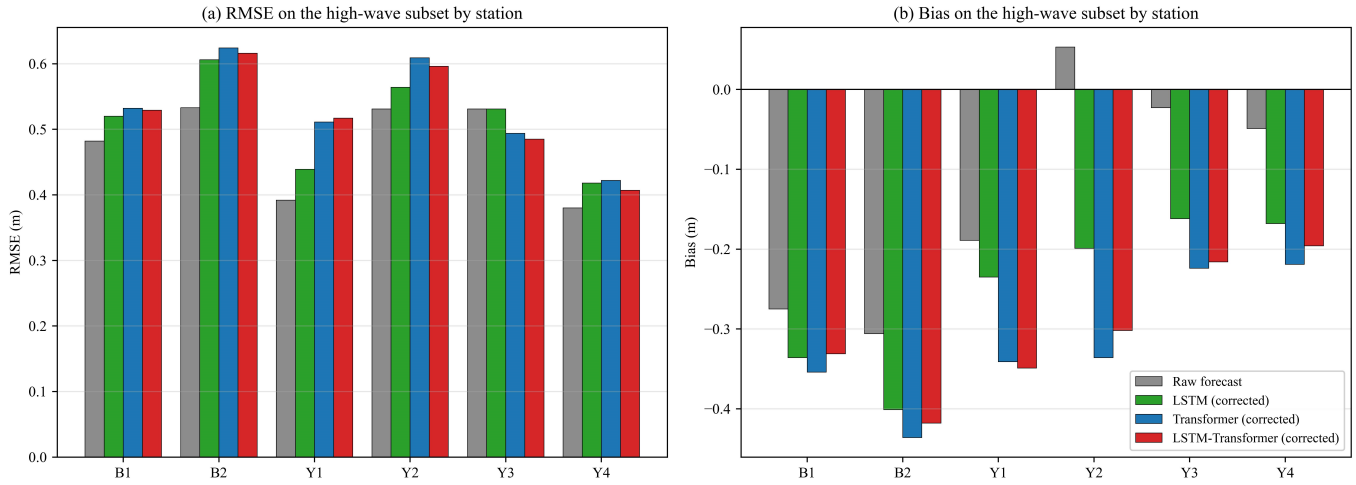


Figure 11. Station-level performance on the high-wave subset: (**a**) RMSE of the raw forecast and the three corrected forecasts; (**b**) corresponding bias. The degradation occurs at five of the six stations, and its magnitude tracks the local bias of the raw forecast.

This behavior is exactly what the decomposition of Section 3.5.3 predicts, although through a different channel than one might first expect. Figure 12 verifies the criterion on the full sample: Cov($\hat{e}$, e) closely tracks Var($\hat{e}$) but remains comfortably above the Var($\hat{e}$)/2 threshold for all three models at every lead time, and the predicted $\Delta\mathrm{RMSE}^2$ from the fluctuation term of Equation (8) reproduces the observed squared-error reduction almost exactly. The high-wave subset, however, exposes the failing channel. The fluctuation-side statistics remain healthy: on the subset, Cov($\hat{e}$, e) attains 0.0141, 0.0399 and 0.0371 m^2 for the LSTM, Transformer and hybrid model, still above the corresponding Var($\hat{e}$)/2 values of 0.0130, 0.0253 and 0.0247 m^2. What collapses is the zero-mean assumption. With E[e] = 0.132 m on the subset and the correction-induced shift in mean error E[$\hat{e}$] = Bias_cor − Bias_raw ranging from −0.105 to −0.167 m, ê opposes the sign of e, and the mean term of Equation (8) contributes $-2\mathrm{E}[e]\mathrm{E}[\hat{e}] \approx +0.028$ m^2 per model, quantitatively accounting for the observed degradation of +0.0353 to +0.0427 m^2 once the (still negative) fluctuation term is added. The criterion and the bias sign reversal are thus two views of the same failure: because training gradients are dominated by the 91.5% of low- and moderate-wave samples, the models learn a globally averaged correction whose systematic component points the wrong way precisely where the waves are highest. That all three architectures degrade by nearly the same amount confirms that the bottleneck lies in the training signal rather than in model capacity.

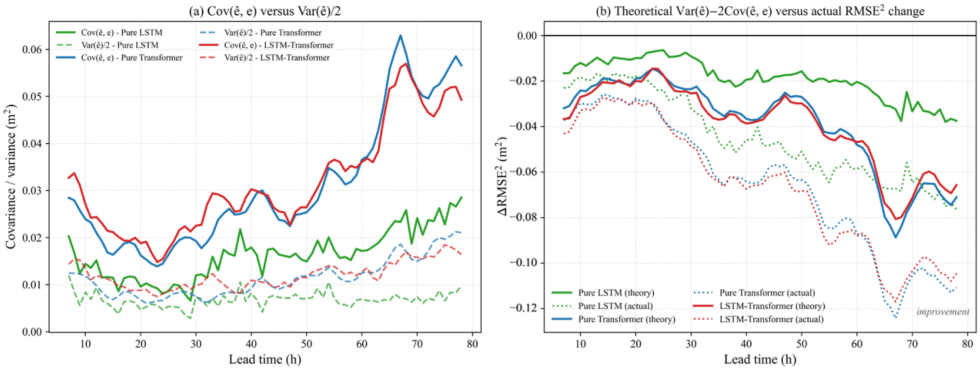


Figure 12. Verification of the covariance skill criterion on the full test sample. (**a**) Cov($\hat{e}$, e), Var($\hat{e}$) and the Var($\hat{e}$)/2 threshold versus lead time; (**b**) predicted versus observed ΔRMSE^2. Negative ΔRMSE^2 denotes improvement. Curves pooled over all stations.

4.3.2. Targeted Remedies Do Not Help at the 72 h Horizon

A natural response to the imbalance diagnosed above is to strengthen the representation of high-wave samples during training. We tested the two strategies of Section 3.4, namely point-weighted loss and high-wave-filtered training. Each was re-tuned with the same 60-trial Optuna budget and evaluated under the same protocol as the baseline models, so that any difference is attributable to the training strategy alone. Results on the high-wave subset are summarized in Table 5 and Figure 13.

Table 5. Performance of the two high-wave-oriented training strategies on the high-wave subset and on the full test set, for all three architectures.

Strategy	Model	RMSE_raw_HW (m)	RMSE_cor_HW (m)	I_{RMSE} (%)
Point-weighted loss	LSTM	0.475	0.480	−1.05%
	Transformer	0.475	0.509	−7.16%
	L-Transformer	0.475	0.487	−2.53%
High-wave-filtered	LSTM	0.475	0.478	−0.63%
	Transformer	0.475	0.499	−5.05%
	L-Transformer	0.475	0.511	−7.58%

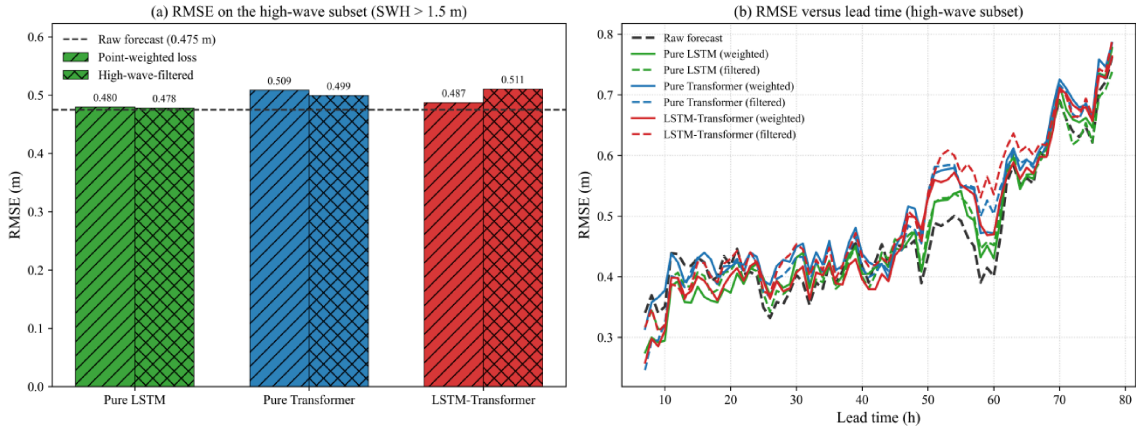


Figure 13. Performance of the two high-wave-oriented training strategies on the high-wave subset: (**a**) RMSE pooled over lead times and (**b**) RMSE versus lead time. The dashed line marks the raw forecast. Test set, all six stations pooled; baseline values are given in Table 4.

Neither strategy eliminates the degradation, although both mitigate it to different degrees. Under the point-weighted loss, the corrected subset RMSE is 0.480, 0.509 and

0.487 m for the pure LSTM, pure Transformer and hybrid models (I_{RMSE} = −1.05%, −7.16%, −2.53%). All three values remain above the raw 0.475 m. Under filtered training the corresponding values are 0.478, 0.499 and 0.511 m (−0.63%, −5.05%, −7.58%). Relative to the baseline degradation, point weighting substantially reduces the penalty for the pure LSTM and the hybrid model, yet no strategy–architecture combination turns the high-wave correction skill-positive. Meanwhile both remedies erode the all-sample skill that motivated the correction: the pooled RMSE reduction shrinks from 15.37–23.93% (baseline) to 15.11–21.74% under point weighting and 10.30–20.25% under filtered training. The trade-off is thus unfavorable in both directions: high-wave performance is not rescued, while bulk performance deteriorates.

The covariance criterion again explains why. Reweighting or filtering changes how often high-wave samples contribute to the gradient, but not the information available about them: the inputs at a high-wave lead time contain the same future forecast that is itself biased low, and the historical error window contains mostly the low-wave conditions preceding the event. The fluctuation-side statistics remain intact: Cov($\hat{e}$, e) stays above Var($\hat{e}$)/2 for all three models. The degradation is therefore driven by the mean term of Equation (8), which neither strategy can reverse: the learned correction retains the negative mean adjustment inherited from the positive full-sample bias, so that $-2E[e]E[\hat{e}]$ remains positive on the subset. In other words, the high-wave failure at the 72 h horizon is not a sample-scarcity problem that stronger training can fix, but an information-deficiency problem: at long lead times the future forecast is nearly the only predictor of its own error, and where that forecast is systematically biased, the error field carries too little predictable structure to exploit. This interpretation is corroborated in Section 4.4: shortening the horizon to 24 h enriches the predictable error information, which not only removes the degradation but also reverses the architecture ranking.

4.4. The 24 h Scheme: The Problem Dissolves as the Horizon Shortens

Sections 4.2 and 4.3 diagnose the high-wave failure as stemming from insufficient predictable error information at long lead times. This diagnosis makes a testable prediction: shortening the forecast horizon should make the error field more learnable and make the degradation diminish or vanish. We therefore retrained and re-tuned all three architectures with the horizon restricted to 24 h, keeping inputs, data partition, tuning budget and evaluation protocol identical to the 72 h scheme, and re-evaluated them on the high-wave subset (Table 6, Figure 14).

Table 6. Performance of the raw and corrected SWH forecasts on the high-wave subset under the 24 h scheme. Positive I_{RMSE} denotes improvement over the raw forecast.

Model	RMSE_raw	RMSE_cor	MAE_raw	MAE_cor	Bias_raw	Bias_cor	I_{RMSE} (%)
LSTM	0.406	0.374	0.320	0.299	−0.172	−0.145	7.88
Transformer	0.406	0.378	0.320	0.302	−0.172	−0.159	6.90
L-Transformer	0.406	0.388	0.320	0.312	−0.172	−0.188	4.43

The prediction is confirmed. On the high-wave subset, all three architectures now improve the raw forecast: the corrected RMSE decreases from 0.406 m to 0.374, 0.378 and 0.388 m for the pure LSTM, pure Transformer and hybrid models (I_{RMSE} = +7.88%, +6.90%, +4.43%), in stark contrast to the uniform degradation of −7.58% to −9.05% under the 72 h scheme (Table 4), while R^2 rises from 0.353 to 0.453, 0.440 and 0.411. Figure 15 shows that the contrast is systematic: the 24 h curves remain above zero at essentially all lead times, whereas the 72 h curves fall below zero and deteriorate with lead time. This locates the crossover between learnable and unlearnable error regimes within the 24–72 h window.

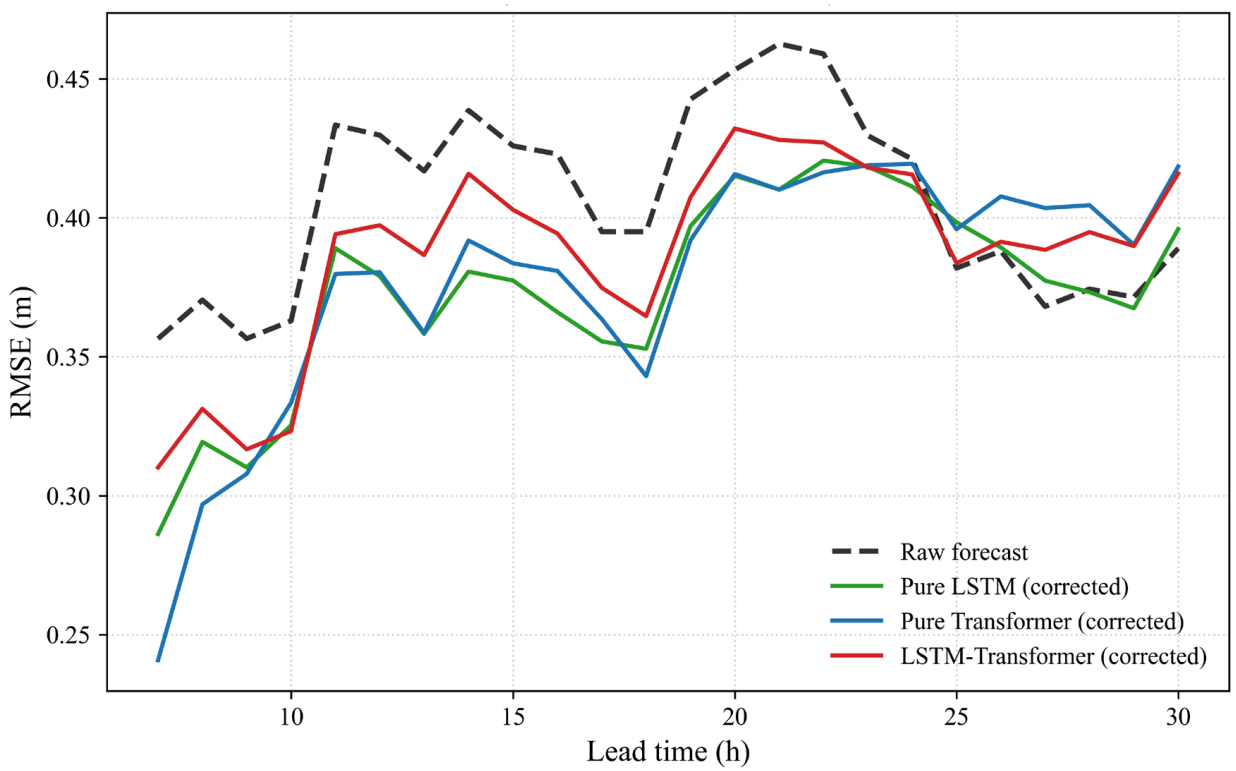


Figure 14. RMSE of the raw and corrected forecasts versus lead time on the high-wave subset under the 24 h scheme. Test set, all six stations pooled.

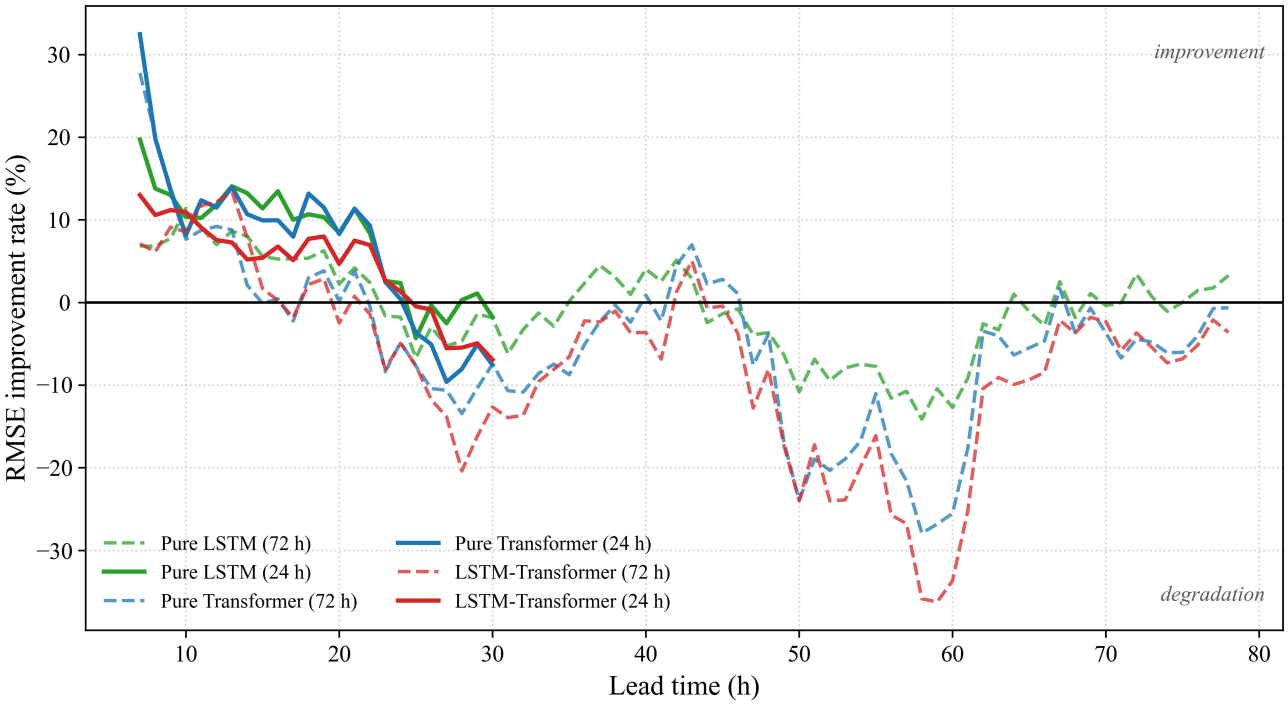


Figure 15. Lead-time-resolved RMSE improvement rate of the corrected forecasts on the high-wave subset under the 24 h (solid) and 72 h (dashed) schemes. Positive values denote improvement over the raw forecast; test set, all six stations pooled.

Two further observations are noteworthy. First, the architecture ranking reverses: the pure LSTM, the weakest in the 72 h scheme, becomes the best performer, followed by the pure Transformer and the hybrid model. The reversal can be traced to the temporal

persistence of the forecast-error field. Model errors, especially their systematic biases, are strongly autocorrelated, so the recent errors in the 24 h history window remain a strong predictor of error sign and magnitude over the next 24 h; at short leads the correction task is essentially extrapolation of a persistent, locally structured signal, ideally suited to the LSTM's local sequential memory. The pure Transformer's global receptive field, by contrast, is of little use when the exploitable signal is concentrated in the most recent hours, and the hybrid's divided capacity is matched to neither extreme. At longer leads, persistence decays, the history window is exhausted as an information source, and performance is governed by the ability to extract weak long-range error structure from the future forecast itself, the regime in which attention excels. This single mechanism accounts for both ranking reversals observed in this study: architecture rankings in error correction are not absolute but horizon-dependent. Second, the bias evidence corroborates the sign-reversal mechanism of Section 4.3.1: the raw bias on the subset is -0.172 m at short leads, and the pure LSTM and pure Transformer corrections now reduce its magnitude (to -0.145 and -0.159 m) instead of amplifying it. That is, the models correct in the subset-consistent direction once the training signal permits.

Together, Tables 4 and 6 delimit the practical scope of deep-learning error correction in this region: reliably beneficial for bulk conditions at all lead times and for high waves at short leads, but counterproductive for high waves beyond about 24 h, where the future forecast, itself biased, is nearly the sole carrier of error information. Remedies for this regime should therefore target the information side rather than the training side, e.g., by incorporating atmospheric predictors (wind speed, pressure tendencies) or wave-model diagnostic fields that carry advance information about wave growth.

To illustrate how the two schemes behave in a real high-wave event, Figure 16 presents the forecasts issued at 20:00 on 1 and 2 October 2025 for stations Y2, Y3 and Y4, during which observed SWH rose from below 0.5 m to a peak of about 4 m. Under the 72 h scheme, all three correctors degrade the raw forecast for observations above 1.5 m: RMSE increases from 0.40–0.46 m (raw) to 0.49–0.73 m, and the negative bias is amplified, which is consistent with the systematic underestimation of high waves diagnosed in Section 4.3. Under the 24 h scheme the outcome depends on the issue time. For the 1 October issue, when the event was still developing, corrections remain mixed and the two attention-based models still degrade; the pure LSTM already reduces RMSE at all three stations including a decrease from 0.446 m to 0.420 m at Y4. For the 2 October issue, issued as the event matured, all correctors improve markedly: RMSE decreases by 18% to 23%, for example from 0.381 m to 0.294 m at Y4 and from 0.244 m to 0.182 m at Y2, and the systematic negative bias of the raw forecast (-0.15 to -0.21 m) is reduced to near zero or reversed to a small positive value. Across both issues the pure LSTM is the most robust corrector under high-wave conditions, in agreement with the architecture ranking found for the 24 h scheme in Section 4.4.

The results above also clarify the strengths and weaknesses of the Transformer architecture for wave forecast correction. Its strength is evident on the full sample: the attention mechanism captures error structure that persists over tens of hours, so the attention-based models achieve the largest gains at the 72 h horizon. Its weaknesses are practical rather than fundamental: the pure Transformer carries 313,665 trainable parameters, about 3.6 times the 87,208 of the pure LSTM, and its larger search space makes tuning more expensive for a modest additional gain. More importantly, the decomposition shows that no architecture can compensate for predictors that are themselves biased under high-wave conditions. The attention mechanism is therefore not ineffective in general: it performs best where predictable error structure exists, and its degradation at long leads under high waves has the same origin as for the other models, namely that the future forecast is nearly the only carrier of error information there. The horizon-dependent ranking should thus be read

as guidance for matching architecture to regime rather than as a verdict on any single architecture.

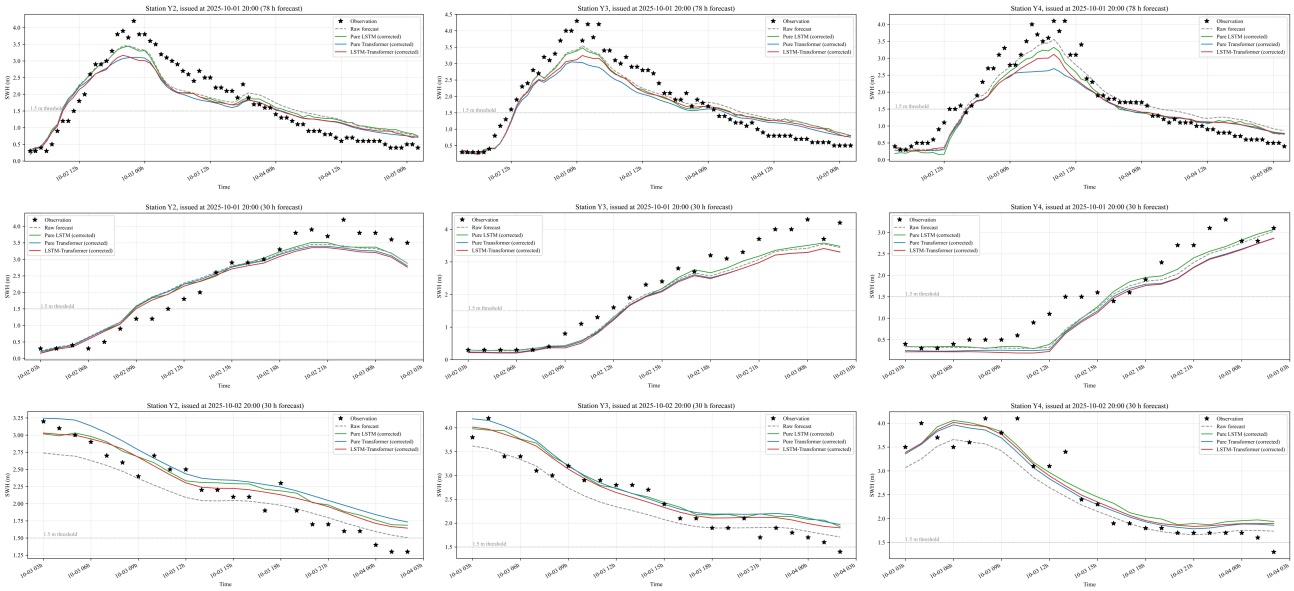


Figure 16. Observed, raw and corrected SWH at stations Y2, Y3 and Y4 during the high-wave event of 2–3 October 2025, for the 72 h scheme (**top**) and the 24 h scheme (**middle** and **bottom**).

5. Conclusions

This study developed and evaluated deep-learning error-correction models for 72 h significant wave height forecasts at six buoy stations in the Bohai Sea and Yellow Sea, comparing three architectures under an identical input protocol, tuning budget (60 Optuna TPE trials per model) and evaluation framework: a pure LSTM, a pure Transformer, and an asymmetric LSTM–Transformer hybrid. An exact decomposition of the squared-error change, $\Delta\mathrm{RMSE}^2 = [\mathrm{Var}(\hat{e}) - 2\mathrm{Cov}(\hat{e}, e)] + [\mathrm{E}[\hat{e}]^2 - 2\mathrm{E}[e]\mathrm{E}[\hat{e}]]$, was introduced as an architecture-independent framework for interpreting when and why correction succeeds or fails. The main findings are as follows.

(1) Under the 72 h scheme, all three architectures significantly improve the raw forecast on the full test sample, with the hybrid model performing best: the pooled RMSE decreases from 0.397 m to 0.302 m (a 23.93% reduction), and R^2 rises from 0.514 to 0.719. The architecture ranking (hybrid > pure Transformer > pure LSTM) is consistent between the validation and test sets, and the spatial pattern of improvement, which is largest where the raw forecast is least skillful, indicates that the models primarily remove systematic, structured error components rather than irreducible noise.

(2) On the high-wave subset (observed SWH > 1.5 m), however, all three architectures consistently degrade the forecast ($I_{\mathrm{RMSE}} = -7.58\%$ to -9.05%; R^2 turning negative), with differences among architectures below 0.01 m. The failure originates in a sign reversal of the raw bias between regimes (from +0.165 m on the full sample to −0.132 m at high waves): because such conditions are rare in the training record, the correctors learn a globally averaged adjustment that subtracts in the wrong direction where waves are highest. The decomposition formalizes this: on the subset the fluctuation term remains skillful, with $\mathrm{Cov}(\hat{e}, e)$ staying above $\mathrm{Var}(\hat{e})/2$, while the degradation is driven by the mean term, whose cross component $-2\mathrm{E}[e]\mathrm{E}[\hat{e}] \approx +0.028$ m^2 accounts for the major part of the observed squared-error increase of +0.035 to +0.043 m^2. The bottleneck thus appears to lie in the training signal rather than in model capacity.

(3) Neither targeted remedy reverses the degradation at the 72 h: point-weighted loss and high-wave-filtered training both leave the subset I_{RMSE} negative (-0.63% to -7.58%),

while both erode the all-sample skill, with the RMSE reduction shrinking from 15.37–23.93% to between 10.30% and 21.74% depending on strategy and architecture. Reweighting or filtering changes how often high-wave samples contribute to the gradient, but not the information available about them: at long leads the future forecast, itself biased, is nearly the sole carrier of error information. Our experiments support the interpretation that the long-lead limitation reflects the information content of the predictors rather than sample scarcity or model capacity, although a direct ablation with additional physical predictors, such as forecast wind speed and direction, pressure tendency and wave-model diagnostics, remains necessary to confirm this mechanism. In addition, exploring alternative sampling schemes and building larger high-wave sample archives are identified as priorities for future work.

(4) When the horizon is restricted to 24 h, the degradation disappears: on the high-wave subset all three architectures turn skill-positive with I_{RMSE} between +4.43% and +7.88% and R^2 up to 0.453, and the architecture ranking reverses, with the pure LSTM attaining the lowest RMSE. The reversal is consistent with the temporal persistence of the error field: recent observed errors remain a strong, locally structured predictor of near-future errors within 24 h, a regime well matched to recurrent memory, whereas the global advantage of attention appears to pay off only at longer leads. Architecture rankings in error-correction studies are thus not absolute but horizon-dependent.

These findings lead to a practical recommendation: the correction strategy should be switched by lead time. At short leads within about 24 h, a pure LSTM correction can be applied confidently at all sea states, including high waves; at longer leads, the hybrid or pure Transformer correction should be retained for bulk conditions but withheld or flagged for predicted high-wave events. For the long-lead high-wave regime, remedies should target the information side, such as incorporating atmospheric predictors or wave-model diagnostic fields that carry advance signals of wave growth, rather than reweighting the training distribution. The decomposition proposed here offers a transferable diagnostic for deciding, in other regions and forecasting systems, whether a learned correction can be trusted before it is applied. Finally, the present conclusions are drawn from six buoys in the Bohai Sea and Yellow Sea; their extension to other basins and to open-ocean conditions remains to be verified.

Author Contributions: Conceptualization, Z.W. and S.G.; Data curation, S.C. and C.Y.; methodology, W.L. and W.C.; software, W.L. and W.C.; validation, W.Z. and R.L.; writing—original draft preparation, W.L.; writing—review and editing, Z.W. and R.L.; funding acquisition, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was firstly funded by the Key R&D Program of Shandong Province, China (No. 2024SFGC0201), and secondly by the Hainan Provincial Natural Science Foundation of China (No. 423QN259).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. The WAMDI Group. The WAM model—A third generation ocean wave prediction model. *J. Phys. Oceanogr.* **1988**, *18*, 1775–1810. [CrossRef]
2. Tolman, H.L. A third-generation model for wind waves on slowly varying, unsteady, and inhomogeneous depths and currents. *J. Phys. Oceanogr.* **1991**, *21*, 782–797. [CrossRef]
3. The WAVEWATCH III Development Group. *User Manual and System Documentation of WAVEWATCH III® Version 6.07*; Tech. Note 333, NOAA/NWS/NCEP/MMAB; U. S. Department of Commerce: College Park, MD, USA, 2019; p. 465.

4. Booij, N.; Ris, R.C.; Holthuijsen, L.H. A third-generation wave model for coastal regions: 1. Model description and validation. *J. Geophys. Res. Ocean.* **1999**, *104*, 7649–7666. [CrossRef]
5. Cavaleri, L.; Alves, J.-H.; Ardhuin, F.; Babanin, A.; Banner, M.; Belibassakis, K.; Benoit, M.; Donelan, M.; Groeneweg, J.; Herbers, T.; et al. Wave modelling—The state of the art. *Prog. Oceanogr.* **2007**, *75*, 603–674. [CrossRef]
6. Zhang, H.; Wang, Z.; Ma, X. Wind and wave climatic characteristics and extreme parameters in the Bohai Sea. *J. Mar. Sci. Eng.* **2025**, *13*, 826. [CrossRef]
7. Liang, B.; Fan, F.; Liu, F.; Gao, S.; Zuo, H. 22-year wave energy hindcast for the China East Adjacent Seas. *Renew. Energy* **2014**, *71*, 200–207. [CrossRef]
8. Glahn, H.R.; Lowry, D.A. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **1972**, *11*, 1203–1211. [CrossRef]
9. Gneiting, T.; Raftery, A.E.; Westveld, A.H.; Goldman, T. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **2005**, *133*, 1098–1118. [CrossRef]
10. Vannitsem, S.; Bremnes, J.B.; Demaeyer, J.; Evans, G.R.; Flowerdew, J.; Hemri, S.; Lerch, S.; Roberts, N.; Theis, S.; Atencia, A.; et al. Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Am. Meteorol. Soc.* **2021**, *102*, E681–E699. [CrossRef]
11. Rasp, S.; Lerch, S. Neural networks for postprocessing ensemble weather forecasts. *Mon. Weather Rev.* **2018**, *146*, 3885–3900. [CrossRef]
12. Schulz, B.; Lerch, S. Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Weather Rev.* **2022**, *150*, 235–257. [CrossRef]
13. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [CrossRef]
14. Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; Tian, Q. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **2023**, *619*, 533–538. [CrossRef]
15. Deo, M.C.; Naidu, C.S. Real time wave forecasting using neural networks. *Ocean Eng.* **1998**, *26*, 191–203. [CrossRef]
16. Makarynskyy, O. Improving wave predictions with artificial neural networks. *Ocean Eng.* **2004**, *31*, 709–724. [CrossRef]
17. Pirhooshyaran, M.; Snyder, L.V. Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence-to-sequence networks. *Ocean Eng.* **2020**, *207*, 107424. [CrossRef]
18. Fan, S.; Xiao, N.; Dong, S. A novel model to predict significant wave height based on long short-term memory network. *Ocean Eng.* **2020**, *205*, 107298. [CrossRef]
19. Minuzzi, F.C.; Farina, L. A deep learning approach to predict significant wave height using long short-term memory. *Ocean Model.* **2023**, *181*, 102151. [CrossRef]
20. Hu, H.; van der Westhuysen, A.J.; Chu, P.; Fujisaki-Manome, A. Predicting Lake Erie wave heights and periods using XGBoost and LSTM. *Ocean Model.* **2021**, *164*, 101832. [CrossRef]
21. Liu, Y.; Lu, W.; Wang, D.; Lai, Z.; Ying, C.; Li, X.; Han, Y.; Wang, Z.; Dong, C. Spatiotemporal wave forecast with transformer-based network: A case study for the northwestern Pacific Ocean. *Ocean Model.* **2024**, *188*, 102323. [CrossRef]
22. Lim, B.; Zohren, S. Time-series forecasting with deep learning: A survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. [CrossRef]
23. Krasnopolsky, V.M.; Chalikov, D.V.; Tolman, H.L. A neural network technique to improve computational efficiency of numerical oceanic models. *Ocean Model.* **2002**, *4*, 363–383. [CrossRef]
24. Deshmukh, A.N.; Deo, M.C.; Bhaskaran, P.K.; Nair, T.M.B.; Sandhya, K.G. Neural-network-based data assimilation to improve numerical ocean wave forecast. *IEEE J. Ocean. Eng.* **2016**, *41*, 944–953. [CrossRef]
25. James, S.C.; Zhang, Y.; O'Donncha, F. A machine learning framework to forecast wave conditions. *Coast. Eng.* **2018**, *137*, 1–10. [CrossRef]
26. Zhang, W.; Sun, Y.; Wu, Y.; Dong, J.; Song, X.; Gao, Z.; Pang, R.; Guoan, B. A deep-learning real-time bias correction method for significant wave height forecasts in the Western North Pacific. *Ocean Model.* **2024**, *187*, 102289. [CrossRef]
27. Wang, B.; Xiao, J.; Feng, C.; Pan, X.; Wang, B. Deep learning-based bias correction model for spatiotemporal significant wave height prediction using multi-channel VMRNN. *Oceans* **2026**, *7*, 54. [CrossRef]
28. Nader, N.; Giaremis, S.; Dawson, C.; Kaiser, C.; Mohammadiporshokooh, K.; Kaiser, H. StormNet: Improving storm surge predictions with a GNN-based spatio-temporal offset forecasting model. *arXiv* **2026**, arXiv:2604.20688.
29. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
30. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; Woo, W.-C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008. [CrossRef]
32. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]

33. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19), Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631. [CrossRef]
34. Hong, S.-Y.; Noh, Y.; Dudhia, J. A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Weather Rev.* **2006**, *134*, 2318–2341. [CrossRef]
35. Dudhia, J. Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.* **1989**, *46*, 3077–3107. [CrossRef]
36. Mlawer, E.J.; Taubman, S.J.; Brown, P.D.; Iacono, M.J.; Clough, S.A. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res. Atmos.* **1997**, *102*, 16663–16682. [CrossRef]
37. Amante, C.; Eakins, B.W. ETOPO1: 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis. In *NOAA Technical Memorandum NESDIS NGDC-24*; National Oceanic and Atmospheric Administration: Boulder, CO, USA, 2009.
38. Ardhuin, F.; Rogers, E.; Babanin, A.V.; Filipot, J.; Magne, R.; Roland, A.; van der Westhuysen, A.; Queffeulou, P.; Lefevre, J.; Aouf, L.; et al. Semiempirical dissipation source functions for ocean waves. Part I: Definition, calibration, and validation. *J. Phys. Oceanogr.* **2010**, *40*, 1917–1941. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.