Proceedings of the Thirty-sixth (2026) International Ocean and Polar Engineering Conference
Orlando, Florida, USA, May 31-June 5, 2026

ISBN 978-1-880653-72-2; ISSN 1098-6189
www.isope.org

Real-time Interval Forecasting of Ship Motion Responses Using Conformalized Quantile Regression

*Baoyu Qin, Maokun Ye, Decheng Wan**

Computational Marine Hydrodynamics Lab (CMHL), School of Ocean and Civil Engineering
Shanghai Jiao Tong University, Shanghai, China
*Corresponding Author

ABSTRACT

Accurate and reliable short-term forecasting of ship motion responses is crucial for various marine operations. This study proposes a probabilistic forecasting framework that integrates Conformalized Quantile Regression (CQR) with a Long Short-Term Memory (LSTM) network for real-time, uncertainty-aware prediction of ship motion. Using a real-world dataset of ship motion, CQR-LSTM is compared against standard Quantile Regression LSTM (QR-LSTM) across multiple prediction horizons (1-6 seconds) and confidence levels (95%, 90%, 85%). Results indicate that CQR-LSTM generally provides more reliable coverage probabilities approaching theoretical guarantees while maintaining efficient prediction intervals.

KEY WORDS: Conformalized Quantile Regression; ship motion prediction; uncertainty quantification; probabilistic forecasting; long short-term memory networks.

INTRODUCTION

In complex marine environments, ships are subjected to the combined effects of wind, waves, and currents, resulting in six-degree-of-freedom (6-DOF) motions characterized by significant nonlinearity. These motions severely impact high-precision offshore operations that demand high stability. Particularly under harsh sea conditions, violent 6-DOF motions, such as rolling and pitching, can lead to catastrophic accidents, endangering lives and property. Implementing motion compensation to maintain vessel movement within safe limits can drastically enhance operational efficiency and safety. The prerequisite for such compensation is the ability to achieve accurate, real-time prediction of the ship motion response.

The real-time prediction of motion responses for ships and offshore floating structures has undergone extensive development, resulting in a rich theoretical framework. Wiener (1949) proposed a prediction method for stationary time series. Utilizing integral equations as the analytical tool, this method achieved optimal linear prediction by minimizing the mean square error (MSE). Kaplan et al. (1969) employed the convolution method for ship motion prediction. Kalman et al. (1961) introduced Kalman filtering, an efficient optimal recursive estimation algorithm. Its core principle lies in the fusion of prediction and observation, providing the best estimate of the system state through probabilistic and statistical methods amidst uncertain information. Triantafyllou et al. (1983) applied Kalman filtering to study the state estimation and very short-term prediction of the motion of the U.S. Navy's DD-963 destroyer. While these classical methods were indeed pioneering, their limitations have become increasingly prominent when faced with the demands of real-time prediction for highly complex systems like ship motion responses, thereby restricting their practical application.

Yumori (1981) adopted the ARMA model to fit the relationship between waves and ship motions through autocovariance and cross-covariance functions, proving the feasibility of using ARMA models for short-term motion prediction in real marine environments. Yu et al. (2016) employed the AR model to study the very short-term prediction of ship heave and pitch, and conducted motion platform tests. Ma et al. (2012) investigated very short-term prediction technology for Unmanned Surface Vehicles (USVs) using a Seasonal Integrated Autoregressive Moving Average (SARIMA) model based on time series and regular wave test data from USV models. Methods such as AR and ARMA possess advantages in computational efficiency, enabling real-time prediction; however, they remain grounded in linearity assumptions and are unable to effectively capture and model the inherent nonlinear characteristics of ship motions.

With the development of computer science and technology, machine learning methods have been increasingly applied to the real-time prediction of ship motion responses. Among these, Artificial Neural Networks (ANN) are widely utilized due to their strong ability to approximate nonlinear systems and their self-adaptive capabilities. Since real-time ship motion prediction is essentially a time-series forecasting problem, Recurrent Neural Networks (RNN), which are inherently suited for processing sequential data, are naturally applicable. Based on the causal relationship between waves and ship motion responses, Yi et al. (2021) employed an LSTM neural network model to establish a mapping between random wave height history and ship roll motion history, achieving very short-term roll prediction based on wave height data. Besides, Xiao et al. (2022) introduced a method for the very short-term prediction of semi-submersible platform motions based on the Temporal Convolutional Network (TCN) model. Zhao et al. (2024) applied Deep Operator Networks (DeepONet) to ship motion prediction, providing a new tool for functional forecasting. Beyond single-model methods, hybrid models integrating signal processing techniques or multiple neural networks are also extensively used. Wang et al. (2025) proposed a VMD-CNN-LSTM combined prediction model, proving the effectiveness of Variational Mode Decomposition (VMD) for data pre-processing and CNN for feature extraction in enhancing LSTM performance for complex nonlinear motion prediction. Sun et al. (2022) proposed a hybrid model based on LSTM and Gaussian Process Regression (GPR) for short-term ship attitude prediction, achieving reliable interval prediction results while maintaining the high accuracy of LSTM.

The aforementioned studies primarily focus on the accuracy of predicting the motion and attitude of marine structures, with less consideration given to the uncertainty of the prediction results. In offshore operations with stringent safety requirements, a "black-box" prediction model that cannot provide confidence levels is difficult to trust. Providing reliable prediction intervals (PIs) within a deep learning framework remains a significant challenge. Representative methods for interval prediction can be categorized into two types: parametric methods and non-parametric methods. In practical applications, non-parametric methods are widely used for probabilistic forecasting problems as they do not impose restrictions on the distribution of the probability intervals. Wei et al. (2024) combined Quantile Regression and Gated Recurrent Unit (GRU) networks, proposing a QR-GRU model for accurate short-term heave motion prediction of semi-submersible platforms, which demonstrates superior performance in both point and interval forecasts. Gui et al. (2025) proposed a hybrid method combining time series decomposition with CQR to achieve accurate and reliable ultra-short-term photovoltaic power interval forecasting without relying on prior distribution assumptions. Hu et al. (2025) proposed a novel wind power interval prediction method that integrates Neural Ensemble Search with dynamic CQR to generate significantly narrower and more reliable prediction intervals while ensuring theoretical coverage guarantees.

In this study, a probabilistic forecasting framework that integrates CQR with a LSTM network for real-time, uncertainty-aware prediction of ship motion responses is proposed. Experimental validation was conducted using a real-world dataset of ship motion responses. Furthermore, CQR-LSTM is compared against standard Quantile Regression with LSTM across multiple forecast horizons and confidence levels.

METHODOLOGY

Conformalized Quantile Regression

Conformalized Quantile Regression is a method for constructing prediction intervals that combines Conformal Prediction with Quantile Regression (Romano et al., 2019). It provides distribution-free coverage guarantees under finite samples and can adapt to heteroscedasticity. The procedures of CQR are as follows.

First, the available dataset is split into a proper training set, indexed by I_1, and a calibration set, indexed by I_2. Any quantile regression algorithm A can be employed to fit two conditional quantile functions on the proper training set. Specifically, the lower and upper quantile functions are estimated at levels $\alpha_{\text{lo}} = \alpha / 2$ and $\alpha_{\text{hi}} = 1 - \alpha / 2$, respectively:

$$\{\hat{q}_{\alpha_{\text{lo}}}, \hat{q}_{\alpha_{\text{hi}}}\} \leftarrow \mathcal{A}\left(\{(X_i, Y_i) : i \in I_1\}\right) \tag{1}$$

Next, conformity scores are computed on the calibration set to quantify the error of the initial plug-in interval $\hat{C}(x) = [\hat{q}_{\alpha_{\text{lo}}}(x), \hat{q}_{\alpha_{\text{hi}}}(x)]$. For each calibration point $i \in I_2$, the score is defined as

$$E_i := \max\left\{\hat{q}_{\alpha_{\text{lo}}}(X_i) - Y_i, Y_i - \hat{q}_{\alpha_{\text{hi}}}(X_i)\right\} \tag{2}$$

The score E_i is non-negative when Y_i falls outside $\hat{C}(X_i)$ and non-positive otherwise; its magnitude reflects the magnitude of the coverage error. The conformity score thus accounts for both undercoverage and overcoverage.

Let $Q_{1-\alpha}(E, I_2)$ denote the $(1-\alpha)(1+1/|I_2|)$-th empirical quantile of the set $\{E_i : i \in I_2\}$. For a new test point X_{n+1}, the final conformalized prediction interval is constructed by symmetrically adjusting the initial quantile estimates:

$$\begin{aligned} C(X_{n+1}) = \big[&\hat{q}_{\alpha_{\text{lo}}}(X_{n+1}) - Q_{1-\alpha}(E, I_2), \\ &\hat{q}_{\alpha_{\text{hi}}}(X_{n+1}) + Q_{1-\alpha}(E, I_2)\big] \end{aligned} \tag{3}$$

This adjustment guarantees finite-sample marginal coverage under the exchangeability assumption, while preserving the adaptivity to heteroscedasticity inherited from the underlying quantile regression.

LSTM

The recurrent neural network is a deep learning network specifically designed for time series problems. Hochreiterand and Schmidhuber (1997) proposed a Long-Short-Term Memory (LSTM) network structure model, which is a new type of network model based on the RNN. The gradient disappearance or explosion caused by derivative multiplication is alleviated, and the memory ability of the deep learning network is effectively improved. The structure of an LSTM unit is shown on Fig. 1.

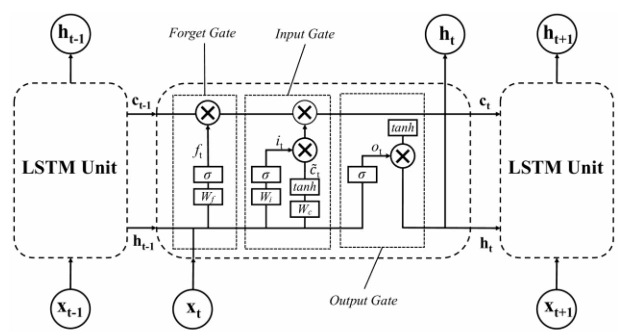


Fig. 1 The structure of LSTM

It consists of four parts: the forget gate, the input gate, the output gate, and the cell state. When the LSTM neural network receives an input, there are three input parameters: the input value at the current time step x_t, the output value from the previous time step h_{t-1}, and the cell state from the previous time step c_{t-1}. When producing an output, there are two output parameters: the output value at the current time step h_t and the cell state c_t at the current time step.

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{4}$$

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{5}$$

$$\tilde{c}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{6}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{7}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{8}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{9}$$

where W represents the weight matrix, b represents the bias vector, σ and tanh represent the activation function, and $\tilde{c}_t$ represents the temporary state at the current moment.

CQR-LSTM

In complex ocean environments, ships are affected by multiple factors, such as wind, waves, and currents, leading to highly nonlinear motion responses, which is appropriate for LSTM to deal with. Simultaneously, CQR can construct efficient prediction intervals. Therefore, the combination of CQR and LSTM provides an effective method for interval prediction. The process of using the CQR-LSTM method for ship motion interval forecasting includes three steps as presented in Fig. 2. First, a quantile regression LSTM model is trained on the training dataset to estimate multiple conditional quantiles of the ship motion responses. Second, using an independent calibration set, a conformal calibration procedure is performed to compute non-conformity scores and determine a calibration factor that adjusts the initial quantile estimates to achieve the desired coverage probability. Finally, the calibrated model is applied to the test set, where it produces prediction intervals that are both empirically reliable and theoretically guaranteed under the assumed exchangeability condition.

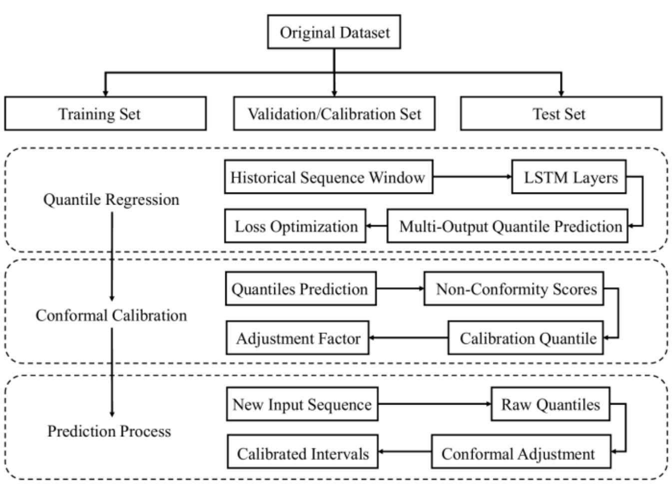


Fig. 2 The framework of CQR-LSTM

Method Evaluation Metrics

Four evaluation metrics are used to assess the performance of predicted interval. Prediction interval coverage probability (PICP) represents the probability that the target value falls within the prediction interval and is a key indicator for evaluating the reliability of interval predictions. A high coverage probability indicates that more target values will be within the constructed prediction interval.

$$P_{\mathrm{PICP}} = \frac{1}{N}\sum_{n=1}^{N}\theta_n \tag{10}$$

$$\theta_n = \begin{cases} 1, & y_n \in [\hat{f}_{\mathrm{low}}, \hat{f}_{\mathrm{up}}] \\ 0, & y_n \notin [\hat{f}_{\mathrm{low}}, \hat{f}_{\mathrm{up}}] \end{cases} \tag{11}$$

Where θ_n is a binary variable, indicating whether the target value y_n is between the upper boundaries $\hat{f}_{\mathrm{up}}$ and lower boundaries $\hat{f}_{\mathrm{low}}$.

Prediction interval average width (PINAW) is used to measure the average width of prediction intervals.

$$W_{\mathrm{PINAW}} = \frac{1}{y_{\max} - y_{\min}}\frac{1}{N}\sum_{n=1}^{N}\left(\hat{f}_{\mathrm{up}} - \hat{f}_{\mathrm{low}}\right) \tag{12}$$

where $y_{\max}$ and $y_{\min}$ are the maximum and minimum of the true values, respectively.

Normalized interval width (NIW) takes both the prediction interval coverage probability and the prediction interval average width into account, standardizing the average width by the interval coverage. When the interval coverage probability is fixed, prediction methods with smaller NIW values have narrower prediction intervals.

$$W_{\mathrm{NIW}} = \frac{W_{\mathrm{PINAW}}}{P_{\mathrm{PICP}}} \tag{13}$$

Coverage width-based criterion (CWC) provides a comprehensive evaluation of the prediction interval width of the model by considering coverage probability, imposing penalties for insufficient prediction interval coverage.

$$W_{\mathrm{CWC}} = W_{\mathrm{PINAW}}\left(1 + I(P_{\mathrm{PICP}} < \mu)e^{-\eta(P_{\mathrm{PICP}} - \mu)}\right) \tag{14}$$

$$I(P_{\mathrm{PICP}} < \mu) = \begin{cases} 1, & P_{\mathrm{PICP}} < \mu \\ 0, & P_{\mathrm{PICP}} \geq \mu \end{cases} \tag{15}$$

where η is the penalty parameter, μ is the confidence level. When the interval coverage of the prediction method falls short of the preset confidence level, the penalty term $e^{-\eta(P_{\mathrm{PICP}} - \mu)}$ causes the interval width to increase significantly, thereby amplifying the shortcomings of the prediction model in terms of interval coverage.

EXPERIMENTAL SETUP

Dataset

The raw data consists of time-series vertical acceleration measurements obtained from the polar research vessel S.A. Agulhas II during its 2019 spring cruise, traveling between Cape Town, South Africa, and the marginal ice zone. The original data was recorded at a sampling frequency of 10 Hz. From this dataset, representative segments were selected and processed via integration and downsampling to obtain the ship's heave time-series curve. The final dataset spans a total duration of 4,800 seconds with a sampling frequency of 1 Hz, which means a sampling interval of 1 s, resulting in a total of 4,800 sample points.

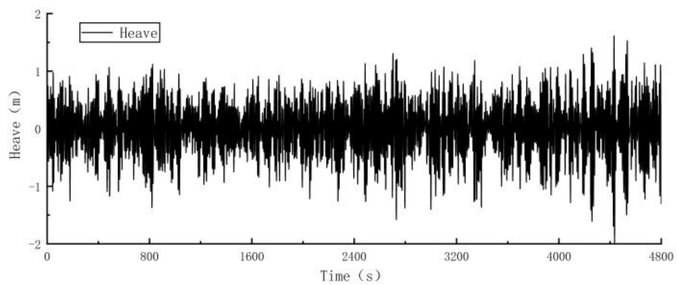


Fig. 3 The time series of heave motion

Upon performing the Fourier Transform, the heave spectrum is obtained as shown below. It can be observed that the dominant frequency is primarily located around 0.1 Hz, while a significant low-frequency component is also present.

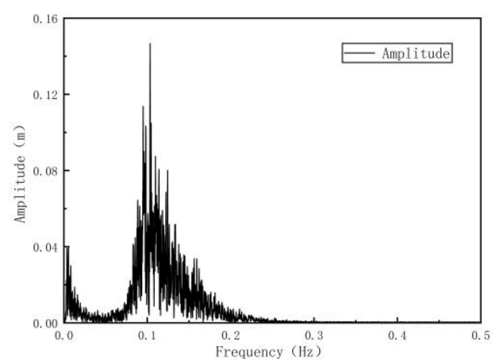


Fig. 4 The spectrum of heave motion

To ensure statistical rigor and the validity of the conformal coverage guarantee, a strictly non-overlapping data partitioning strategy with a 7:2:1 ratio was adopted in this study. For the baseline QR-LSTM model, the dataset was divided into a training set (70%), a validation set (20%) for performance optimization and hyperparameter tuning, and a testing set (10%) for performance evaluation. For the proposed CQR-LSTM framework, the 20% validation portion was further partitioned into two independent subsets: a 10% validation set for model selection and a 10% calibration set dedicated exclusively to calculating the conformal scores, which ensures that the calibration process is entirely independent of the model training and tuning phases, thereby providing a more reliable assessment of the interval forecasting performance.

Convergence Analysis

To determine the optimal length of the historical input sequence, a convergence analysis was conducted prior to the main experiments. Multiple CQR-LSTM models with identical architectures were trained using varying input window sizes (e.g., 30, 40, and 50 time steps). The prediction performance was evaluated for each configuration. The results indicated that for window sizes above 30 steps, model performance does not change much. As a result, a window size of 30 steps was selected as the optimal trade-off, capturing sufficient temporal dependencies for accurate quantile estimation without introducing unnecessary computational complexity.

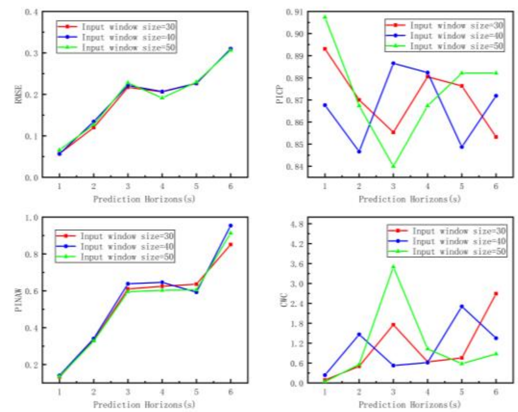


Fig. 5 The performance of CQR-LSTM models with different input window sizes

Model configuration and training procedure

To compare the performance of the CQR-LSTM model and the traditional QR-LSTM model, their hyperparameters are set to identical values. Furthermore, an early stopping mechanism is implemented to prevent overfitting and enhance training efficiency. During the training process, the model's performance on the validation set is continuously monitored. If the prediction performance fails to improve after a specified number of epochs, the training is terminated, and the model parameters that yielded the best performance on the validation set are retained. The specific configurations of the hyperparameters are summarized in Table 1.

Table 1. Hyperparameter Configurations of CQR-LSTM model

Hyperparameters	Values
Number of hidden layers	3
Number of hidden units	256
Batch size	16
Maximum epochs	200
Early stopping patience	20
Optimizer	Adam
Learning rate	0.001

The model is designed with an input window of 30 steps and outputs three quantiles. For example, corresponding to a 95% confidence interval, the output quantiles are 0.025, 0.5 and 0.975. The conformal calibration step utilizes the calibration dataset to compute the calibration factor Q, targeting a preset confidence level.

RESULTS AND DISCUSSION

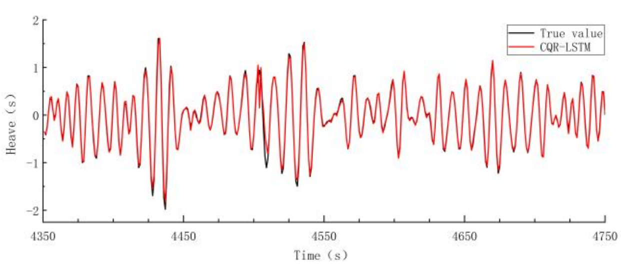
Point prediction

Before presenting the comparative results, it is important to clarify the selection of baseline models. Although hybrid architectures such as VMD-CNN-LSTM or QR-GRU have shown promising results in point prediction, this study specifically employs QR-LSTM as the primary benchmark. The objective is to conduct a controlled comparative study that isolates the impact of the CQR procedure. By using a consistent LSTM backbone, we can directly quantify the improvements in interval reliability and coverage provided by conformalization, without interference from complex network architectural variations.

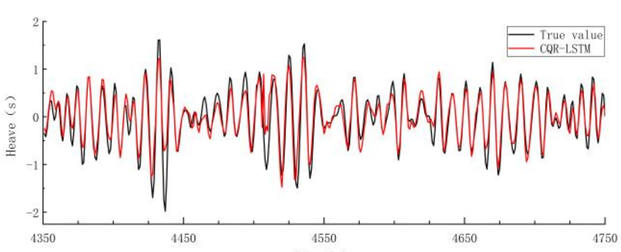
From the perspective of probability theory, neural networks are designed to learn a conditional distribution. In the context of time-series forecasting, by inputting historical data to predict a future time point, the model is essentially trained to learn the distribution of values at a future moment given the historical observations.

By implementing different loss functions, the model yields different characteristics of this distribution: the Mean Square Error (MSE) corresponds to the conditional mean, the Mean Absolute Error (MAE) corresponds to the conditional median, and the Quantile Loss corresponds to the conditional quantiles. Based on these quantile results, prediction intervals can be constructed. Consequently, the point prediction performance of CQR-LSTM and QR-LSTM remains similar across different confidence levels even though there is a slight difference between the validation datasets. Therefore, only the prediction results of CQR-LSTM at a single confidence level are presented here.

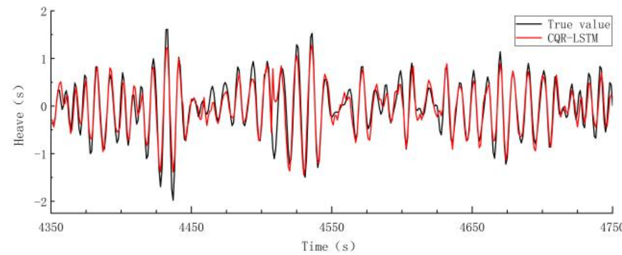
Fig. 6 and Table 2 indicate that CQR-LSTM achieves impressive accuracy in point prediction. When predicting the immediate next step, the coefficient of determination (R^2) reaches as high as 0.987. As the lead time increases, the prediction accuracy gradually declines. However, at a prediction horizon of 5 s, the R^2 remains at 0.849, indicating that the model captures the underlying trends of the ship's heave time-series data, which lays a solid foundation for the follow-up interval prediction.



(a) Point prediction results at a prediction horizon of 1 s



(b) Point prediction results at a prediction horizon of 3 s



(c) Point prediction results at a prediction horizon of 5 s

Fig. 6 Point prediction results of CQR-LSTM model

Table 2. The error evaluation of CQR-LSTM point prediction

Prediction horizons	R^2	RMSE	MAE
1 s	0.987	0.066	0.035
3 s	0.833	0.234	0.168
5 s	0.849	0.222	0.163

Interval prediction

In this section, CQR-LSTM model is compared with QR-LSTM model, using a real-world dataset of heave motion of the polar research vessel across multiple prediction horizons (1-6 seconds) and confidence levels. Fig. 7 illustrates the prediction results of the two models at a prediction horizon of 3 s with a pre-specified confidence level of 90%. This figure is presented as a schematic representation, and a detailed analysis will be conducted based on specific interval prediction metrics. It can be observed that in regions with high prediction difficulty and significant uncertainty, such as peaks and valleys, the prediction intervals are wider to ensure that the ground truth is successfully covered. Conversely, in areas characterized by lower prediction difficulty and uncertainty, the prediction intervals are narrower, thereby enhancing interval efficiency while maintaining coverage. While both models provide prediction intervals that generally meet the requirements, there are discernible differences in terms of coverage probability and interval width.

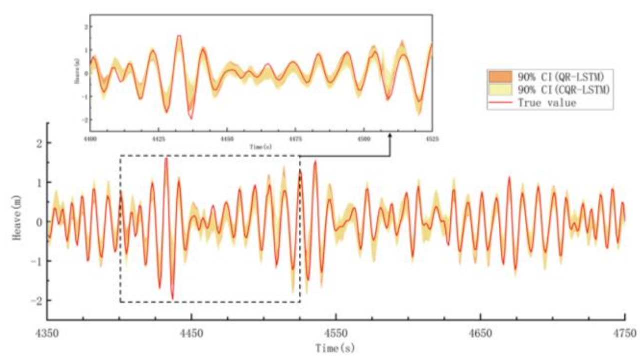


Fig. 7 The comparison of two models under a prediction horizon of 3 s

As observed from Table 3, Table 4, and Table 5, the model performance across different confidence levels exhibits several trends. As the prediction horizon increases, the Prediction Interval Coverage Probability (PICP) generally decreases, indicating that the uncertainty of long-term forecasting grows and the coverage capability of both methods weakens. Simultaneously, the Prediction Interval Normalized Average Width (PINAW) and Normalized Interval Width (NIW) generally increase, suggesting that the prediction intervals widen to account for the heightened uncertainty. The Coverage Width Criterion (CWC) overall shows an upward trend, implying that the comprehensive predictive performance deteriorates as the time steps increase. Furthermore, as the pre-specified confidence level decreases, it is expected that the PICP and PINAW/NIW will decline for the same method and prediction step. Besides, the variation in CWC depends on the trade-off between coverage error and interval width.

Comparing the two methods, in most cases, the CQR-LSTM outperforms or performs comparably to the QR-LSTM in terms of PICP. Through conformal calibration, the CQR-LSTM can more adaptively adjust its intervals, bringing the PICP closer to the target confidence level while simultaneously controlling the interval width. Regarding PINAW, the values for CQR-LSTM are typically slightly larger than or similar to those of QR-LSTM. This can be attributed to the strong inherent nonlinearity and significant predictive uncertainty of the ship heave data; consequently, the conformal calibration mechanism must increase the interval width to satisfy the specified confidence level requirements. Therefore, in this study, a significant reduction in interval width was not observed. Besides, as illustrated in Fig. 8, the CQR-LSTM generally demonstrates superior performance in terms of the comprehensive metric CWC.

It is worth mentioning that neither model is satisfactory at a prediction horizon of 1 s except the CQR-LSTM model in a confidence level of 95%. The QR-LSTM exhibits severe over-coverage with excessively wide intervals, whereas the CQR-LSTM shows significant under-coverage with intervals that are too narrow. This phenomenon may be attributed to the fact that quantile regression is more susceptible to outliers in the training data when estimating extreme quantiles, such as 2.5% and 97.5%, leading to interval inflation. Furthermore, the quantile adjustment in conformal prediction is highly sensitive to the distributional tails. If the residual distribution of the calibration set does not match that of the test set, the adjustments at the tails may become inaccurate.

Table 3. The evaluating metrics of CQR-LSTM and QR-LSTM in confidence level of 95%

Prediction horizons	Method	PICP	PINAW	NIW	CWC
1 s	QR-LSTM	0.971	0.195	0.054	0.054
	CQR-LSTM	0.956	0.201	0.056	0.056
2 s	QR-LSTM	0.916	0.392	0.109	0.702
	CQR-LSTM	0.922	0.403	0.112	0.558
3 s	QR-LSTM	0.927	0.724	0.202	0.850
	CQR-LSTM	0.927	0.742	0.207	0.871
4 s	QR-LSTM	0.922	0.706	0.197	0.977
	CQR-LSTM	0.916	0.773	0.215	1.386
5 s	QR-LSTM	0.889	0.638	0.178	3.950
	CQR-LSTM	0.927	0.775	0.216	0.911
6 s	QR-LSTM	0.866	0.845	0.235	16.07
	CQR-LSTM	0.906	1.012	0.282	2.871

Table 4. The evaluating metrics of CQR-LSTM and QR-LSTM in confidence level of 90%

Prediction horizons	Method	PICP	PINAW	NIW	CWC
1 s	QR-LSTM	0.966	0.180	0.050	0.050
	CQR-LSTM	0.878	0.142	0.040	0.156
2 s	QR-LSTM	0.881	0.380	0.106	0.386
	CQR-LSTM	0.866	0.346	0.096	0.629
3 s	QR-LSTM	0.834	0.536	0.149	4.119
	CQR-LSTM	0.857	0.595	0.166	1.558
4 s	QR-LSTM	0.853	0.577	0.161	1.824
	CQR-LSTM	0.870	0.609	0.170	0.930
5 s	QR-LSTM	0.828	0.565	0.157	5.891
	CQR-LSTM	0.847	0.595	0.166	2.515
6 s	QR-LSTM	0.849	0.804	0.224	3.085
	CQR-LSTM	0.878	0.881	0.245	0.967

Table 5. The evaluating metrics of CQR-LSTM and QR-LSTM in confidence level of 85%

Prediction horizons	Method	PICP	PINAW	NIW	CWC
1 s	QR-LSTM	0.933	0.146	0.041	0.041
	CQR-LSTM	0.818	0.141	0.039	0.238
2 s	QR-LSTM	0.795	0.289	0.080	1.367
	CQR-LSTM	0.809	0.289	0.080	0.698
3 s	QR-LSTM	0.809	0.538	0.150	1.301
	CQR-LSTM	0.826	0.547	0.152	0.659
4 s	QR-LSTM	0.738	0.439	0.122	33.26
	CQR-LSTM	0.824	0.519	0.144	0.677
5 s	QR-LSTM	0.788	0.488	0.136	3.116
	CQR-LSTM	0.816	0.575	0.160	1.059
6 s	QR-LSTM	0.771	0.673	0.188	9.691
	CQR-LSTM	0.857	0.769	0.214	0.214

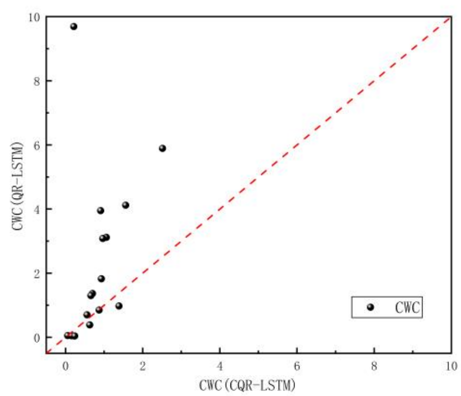


Fig. 8 Comparison of CWC between two models

CONCLUSIONS

This study proposes a probabilistic forecasting framework that integrates Conformalized Quantile Regression (CQR) with Long Short-Term Memory (LSTM) networks for the real-time, uncertainty-aware prediction of ship motion responses. The proposed CQR-LSTM model was systematically evaluated against the standard QR-LSTM using a real-world heave motion dataset from the polar research vessel S.A. Agulhas II across multiple prediction horizons (1-6 s) and confidence levels (95%, 90%, and 85%).

Experimental results demonstrate that the CQR-LSTM framework generally provides more reliable prediction intervals, with coverage probabilities closely aligning with theoretical confidence levels. Through conformal calibration, CQR-LSTM effectively adjusts initial quantile estimates to achieve the desired coverage, a strength that is particularly pronounced at longer prediction horizons where uncertainty is heightened. Although the interval widths of CQR-LSTM are occasionally slightly larger than those of QR-LSTM, reflecting the model's adaptive adjustment to ensure coverage within highly nonlinear and heteroscedastic motion environments, the method consistently exhibits superior or comparable performance in comprehensive metrics such as the Coverage Width Criterion (CWC).

However, certain limitations were observed, particularly at very short prediction horizons where both models exhibited coverage bias. QR-LSTM showed interval inflation leading to over-coverage, while the narrow intervals of CQR-LSTM resulted in under-coverage. This suggests that quantile regression remains sensitive to extreme quantile estimation, and that conformal calibration relies heavily on the exchangeability between the residuals of the calibration and test sets.

In conclusion, the CQR-LSTM framework provides a robust approach for the interval prediction of ship motions, achieving a favorable balance between reliability and efficiency.

Despite the promising results, this study has certain limitations. The current experiments are based on a dataset from a single cruise and focus primarily on heave motion, which serves as a proof of concept for the proposed CQR-LSTM framework. Although the model demonstrates high reliability for heave motion, its generalizability to different ship types, varying sea states, and other degrees of freedom remains to be further validated. Future work will focus on testing the proposed method across more diverse maritime datasets and exploring its adaptive performance in complex multi-DOF ship motion scenarios. What's more, future research will also explore the integration of CQR with more sophisticated deep learning models.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (2025YFF0519800), to which the authors are most grateful.

REFERENCES

Gui, Q, Xu, W, Li, X, Luo, L, Ye, H, and Wang, Z (2025). "Ultra-Short-Term Photovoltaic Power Interval Forecasting Based on Time-Series Decomposition and Conformal Quantile Regression," *Electric Power*, 58(5), 21–32.

Hu, J, Deng, Y, and Che, J (2025). "A novel wind power interval prediction method based on neural ensemble search and dynamic conformalized quantile regression," *Applied Soft Computing*, 180, 113476.

Kalman, RE, and Bucy, RS (1961). "New results in linear filtering and prediction theory," *Journal of Basic Engineering*, 83(1), 95–108.

Kaplan, P (1969). "A study of prediction techniques for aircraft carrier motions at sea," *Journal of Hydronautics*, 3(3), 121–131.

Ma, W, Shi, S, Zou, J, Pang, Y, and Zhu, K (2012). "Extreme short-time prediction technology of high speed unmanned surface vehicle motion," *Journal of Traffic and Transportation Engineering*, 12(1), 38–44.

Romano, Y, Patterson, E, and Candès, EJ (2019). "Conformalized Quantile Regression," *Proc 33rd Conf Neural Inf Process Syst (NeurIPS)*, Vancouver, Canada, NeurIPS, 1, 1–11.

Sun, Q, Tang, Z, Gao, J, and Zhang, G (2022). "Short-term ship motion attitude prediction based on LSTM and GPR," *Applied Ocean Research*, 118, 102927.

Triantafyllou, MS, Bodson, M, and Athans, M (1983). "Real time estimation of ship motions using Kalman filtering techniques," *IEEE Journal of Oceanic Engineering*, OE-8(1), 9–19.

Wang, Y, Li, J, Wang, S, Zhang, H, Yang, L, and Wu, W (2025). "Extreme short-term prediction of unmanned surface vessel nonlinear motion under waves," *Journal of Marine Science and Engineering*, 13(3), 610.

Wei, Q, Wu, B, Li, X, and Zhang, F (2024). "Interval Prediction of the Heave Motions of a Semi-submersible Platform Based on QR-GRU," *Proc 34th Int Offshore and Polar Eng Conf*, Rhodes, Greece, ISOPE, 1, 2343–2348.

Wiener, N (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, Wiley, 163.

Xiao, F, Li, X, Yang, J, Guo, X, and Li, Y (2022). "Extreme Short-Time Prediction of Semi-Submersible Platform Motion Based on TCN," *Ocean Engineering Equipment and Technology*, 9(4), 1–9.

Yi, W, and Gao, Z (2021). "Very Short-term Prediction of Ship Rolling Motion in Random Transverse Waves Based on LSTM Neural Network," *Journal of Wuhan University of Technology (Transportation Science & Engineering)*, 45(6), 1113–1117.

Yu, Y, Ma, X, and Ji, S (2016). "Short Time Prediction of Ship Motion Based on AR Model and Stewart Platform Experiment," *Journal of Shanghai Ship and Shipping Research Institute*, 39(4), 4–7.

Yumori, I (1981). "Real time prediction of ship response to ocean waves using time series analysis," *Proc OCEANS '81*, Boston, MA, USA, IEEE, 1082–1089.

Zhao, Y, Zhao, J, Wang, Z, Lu, S, and Zou, L (2024). "A comprehensive comparison study between Deep Operator networks neural network and long short-term memory for very short-term prediction of ship motion," *Journal of Hydrodynamics*, 36(6), 1167–1180.