POLITECNICO DI TORINO
Repository ISTITUZIONALE

Adaptive Conformalized RNN Ensembles for Uncertainty-Aware Motion Prediction of Floating Offshore Wind Turbines in Stochastic Marine Environments

Original
Adaptive Conformalized RNN Ensembles for Uncertainty-Aware Motion Prediction of Floating Offshore Wind Turbines in Stochastic Marine Environments / Tamuly, P., Del Ser, J., Fathnejat, H., Nava, V.. - In: RESULTS IN ENGINEERING. - ISSN 2590-1230. - 32:(2026). [10.1016/j.rineng.2026.114034]

Availability:
This version is available at: 11583/3018006 since: 2026-11-17T04:28:42Z

Publisher:
Elsevier

Published
DOI:10.1016/j.rineng.2026.114034

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

22 November 2026

Results in Engineering 32 (2026) 114034

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering

Adaptive conformalized RNN ensembles for uncertainty-aware motion prediction of floating offshore wind turbines in stochastic marine environments

Pranjal Tamuly [a,*], Javier Del Ser [b,c], Hamed Fathnejat [d], Vincenzo Nava [e]

[a] Department of Economics, Engineering, Society and Business Organization, University of Tuscia, Viterbo, 01100, Italy
[b] TECNALIA, Basque Research & Technology Alliance (BRTA), 48160, Derio, Spain
[c] Department of Mathematics, University of the Basque Country (UPV/EHU), 48940, Leioa, Spain
[d] Construction Research Centre, National Research Council Canada, Building M-20, 1200 Montréal Road, Ottawa, K1A 0R6, ON, Canada
[e] Politecnico di Torino, Corso Duca degli Abruzzi, 24, Torino, 10129, Italy

ARTICLE INFO

Keywords:
Offshore structures
Floating wind turbines
Mooring systems
Recurrent neural network
Ensemble learning
Conformal prediction

ABSTRACT

Reliable motion prediction of Floating Offshore Wind Turbine (FOWT) platforms under stochastic metocean conditions is critical for structural safety, fatigue assessment, and operational control. Although Recurrent Neural Networks (RNNs) effectively capture temporal dynamics, their deterministic nature lacks uncertainty quantification, potentially leading to unforeseen failures, over-engineered designs, and increased maintenance costs. This study addresses this gap by introducing an Adaptive Conformalized RNN Ensembles (ACRNNE) framework that builds on conformal prediction and bootstrap-based RNN ensembles. The framework integrates three adaptive mechanisms: (1) an inverse-error weighting strategy that dynamically adjusts the contribution of each ensemble member at every prediction step, (2) a residual pool with weighted conformal inference and (3) an online update of the coverage level driven by the observed outcomes of earlier forecasts. Unlike conventional RNN models, the proposed method provides prediction intervals that adapt to the current metocean condition. The ACRNNE framework is thoroughly benchmarked against four established uncertainty-quantification techniques. ACRNNE achieves more consistent empirical coverage across the diverse metocean conditions than the evaluated baselines, with the largest gains for high, long-period waves. Furthermore, evaluation across multi-degree-of-freedom responses demonstrates robust performance with both LSTM and GRU architectures. Overall, ACRNNE enables uncertainty-aware motion prediction of FOWT platforms through empirically calibrated conformal intervals, supporting risk-informed decision-making in their operation and management.

1. Introduction

The offshore wind sector presents a significant opportunity for advancing renewable energy solutions. In 2024, wind energy met 19% of electricity demand in the European Union [1]. The substantial increase in offshore wind capacity, from 8503 MW in 2014 to 73,185 MW in 2023, underscores its vital role in meeting energy demands and reducing carbon emissions [2].

To leverage this expanding offshore potential, attention has increasingly shifted towards technologies capable of operating in deeper waters, most notably Floating Offshore Wind Turbines (FOWTs). FOWTs enable access to deeper and more resource-rich offshore areas that are inaccessible to traditional fixed-bottom turbines. One of the primary components of an FOWT is the supporting floating platform, which, during operation, is subject to the coupled interactions of wind, waves, and currents. This coupling results in time-varying and autocorrelated platform responses, which have a direct impact on turbine performance, component fatigue, and the reliability of the overall system. For this reason, accurately predicting platform motions is essential for the safe and efficient operation of FOWTs, facilitating optimal control strategies, maintenance planning, operational limit setting, and mooring system design [3]. Short-term motion forecasts allow advanced control systems to proactively adjust blade pitch or mooring line tension, thereby mitigating structural loads and preventing failure. Conversely, inaccurate predictions may result in unsafe conditions or excessively cautious responses, diminishing energy production. These challenges highlight the need for comprehensive forecasting frameworks to deliver precise motion predictions and adequately assess related uncertainties [4].

* Corresponding author.
E-mail address: pranjal.tamuly@unitus.it (P. Tamuly).

https://doi.org/10.1016/j.rineng.2026.114034
Received 22 June 2026; Received in revised form 23 October 2026; Accepted 5 November 2026
Available online 6 November 2026
2590-1230/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Problem Statement. Accurately modelling these coupled responses of FOWTs, however, remains a significant challenge. While small-scale physical FOWT model testing can be used for this purpose, the concurrent physical phenomena require adherence to different conflicting scaling laws [5], with hydrodynamics and aerodynamics governed by Froude and Reynolds scaling laws, respectively. Consequently, numerical modelling, possibly validated by experimental hybrid approaches, exhibits a balance between fidelity and practical constraints while enabling comprehensive analysis of platform dynamics under diverse operating conditions. Hence, effective numerical prediction requires time-series models capable of capturing the complexity of FOWT dynamics.

Physical origins of predictive performance degradation. FOWT motions arise from coupled wave and current loading, aerodynamic forces, and interactions with the rotor, controller and mooring system. For example, resonant amplification can occur when the wave peak period approaches a natural period of the platform. In higher and steeper waves, quadratic viscous drag on columns and bracings becomes a relevant source of nonlinear damping and excitation. Aerodynamic loading also changes across the operating regimes: the aerodynamic damping of platform pitch and surge depends on the operating condition and on the controller tuning, and above rated wind speed the blade-pitch controller can reduce or even reverse it. In addition, the catenary mooring system provides a nonlinear restoring stiffness, so that currents and mean thrust shift the operating point of the system. These mechanisms can change the amplitude and temporal structure of platform motions across metocean conditions, making forecast errors condition-dependent. Intervals calibrated across pooled conditions may therefore achieve adequate average coverage while under-covering particular conditions.

Existing Approaches. Early numerical modelling methods for time series predominantly utilised Autoregressive (AR) [6], Moving Average (MA), and Autoregressive Integrated Moving Average (ARIMA) approaches, alongside their seasonal variants (SARIMA). These linear models have demonstrated effectiveness in applications like wind speed forecasting and energy demand modelling [7–9]. However, they typically model each time series independently, limiting their ability to capture the complex, non-linear, and multivariate temporal interactions and dependencies, which are characteristic of FOWT dynamics. Moreover, these models struggle to represent long-term dependencies and exhibit inferior performance under sudden environmental perturbations, which are predominant in offshore marine environments. Furthermore, linear time-invariant models have limited ability to represent the amplitude-dependent viscous damping, the controller-dependent aerodynamic damping and the non-linear mooring restoring forces discussed above.

These limitations of linear time-series models have motivated a transition towards machine learning approaches. Such modelling alternatives address both non-linearity and long-term temporal dependencies by learning and adapting their parameters directly from data, making them well-suited for predicting non-linear and dynamically evolving systems. Among such data-driven models, deep neural networks, especially Recurrent Neural Networks (RNNs), have recently become practical tools for modelling complex temporal patterns [10]. RNN-based models can learn non-linear relationships and long-term dependencies from historical dynamic motions of FOWTs. Several recent works have resorted to RNN-based models for FOWT motion modelling. Shi et al. [11] developed a multi-input Long Short-Term Memory (LSTM) network that incorporates historical platform motions and environmental inputs, such as wave height and mooring tension, enhancing accuracy for short-term motion predictions. Zhang et al. [12] proposed a hybrid DWT-SARIMA-LSTM model for short-term offshore wind power forecasting, demonstrating that LSTM can efficiently capture complex patterns in offshore wind time series. Likewise, Deng et al. [3] developed a self-attention-augmented LSTM model to enhance the capture of main features in motion response across different metocean conditions. Hanifi et al. [13] further advanced offshore wind power forecasting by combining wavelet packet decomposition with optimised LSTM and CNN architectures, improving prediction accuracy through hyperparameter optimization. In a more recent hybrid approach, it has been demonstrated that combining a time-frequency feature fusion algorithm with LSTM networks significantly improves prediction performance under both normal and extreme sea conditions [14]. Contemporaneously, Yin et al. [15] proposed an optimised Bi-LSTM model for short-term prediction of FOWT platform motions using wave–motion correlations under noisy environments.

Research Gap. Although RNN-based models have demonstrated strong predictive capabilities, their effectiveness depends heavily on the availability of large volumes of high-quality training data. While some recent studies [16–19] have addressed data scarcity through data-driven augmentation techniques based on deep generative models for offshore structures, these approaches primarily focus on generating complete synthetic time-series responses. Such methodologies produce entirely new response sequences for various metocean conditions, but do not forecast future platform motions from historical observations. Moreover, most of these RNN-based studies produce only point forecasts and do not quantify forecast uncertainty. A few probabilistic approaches provide prediction intervals, but whether these intervals remain reliable across different metocean conditions is seldom examined.

In floating wind turbine applications, where platform motions influence both structural integrity and power generation efficiency, understanding prediction uncertainty is therefore as crucial as the forecast itself. Despite this, robust uncertainty quantification for FOWT motion prediction remains largely unexplored in the literature, even though it is essential for reliable engineering decisions under diverse real-world metocean scenarios. Shao and Liu [20] addressed uncertainty quantification for dynamic responses of offshore wind turbines using manifold learning to construct probability box models, though their approach focuses on fixed-bottom turbines and does not address floating platforms. A recent step towards addressing this gap is the Bayesian framework proposed by Li et al. [4], which provides probabilistic motion forecasts by combining real-time monitoring data with numerical simulations. Similarly, Fernandez-Navamuel et al. [21] developed an uncertainty-aware inverse model for damage identification in FOWTs using autoencoders, where multimodal uncertainties stemming from the ill-posed nature of the inverse problem are represented through a multivariate Gaussian mixture model. However, these approaches rely on accurate prior probabilistic models and often entail significant computational expense, which limits their practicality for continuous online forecasting and real-time operational decision-making. From a structural safety perspective, Guo et al. [22] recently proposed an efficient random finite element method for reliability analysis and design of multi-line dynamically installed anchors for FOWTs, highlighting the critical need for probabilistic structural assessment methods capable of addressing the compound uncertainties inherent in floating offshore wind systems.

Despite the growing interest noted in these recent works, two research gaps still persist in uncertainty quantification for FOWT motion dynamics: (i) the majority of FOWT studies emphasise deterministic accuracy or generative realism over well-calibrated prediction intervals with strong statistical properties; and (ii) even when calibrated prediction is examined in time-series contexts, adaptive interval calibration under varying metocean conditions and across multiple degrees of freedom is seldom addressed. Moreover, the limitations discussed above regarding recent uncertainty treatment approaches in offshore renewable structures highlight the need for a distribution-free uncertainty quantification technique that imposes minimal assumptions on data distribution while adapting to the dynamic characteristics of time series.

This need becomes even more evident when considering the wider range of uncertainty quantification methods applied in regression and time-series modelling. Classical probabilistic frameworks (e.g., Bayesian linear models and Gaussian processes) offer calibrated uncertainties but struggle when confronted with the high dimensionality, strong

temporal variability, and rapidly varying data streams typical of operational engineering systems [23]. In the context of deep learning, commonly used uncertainty proxies include Monte Carlo dropout, deep ensembles, and quantile regression. Although flexible, these techniques often rely on heuristic assumptions regarding coverage guarantees and thus cannot reliably ensure formal statistical validity [24–26]. In contrast, Conformal Prediction (CP) offers distribution-free, finite-sample coverage guarantees by calibrating a conformity score on held-out data, making it applicable across a wide range of predictors [27]. Conformalized Quantile Regression (CQR) leverages learned quantiles to produce adaptive, heteroscedastic intervals that maintain valid coverage under exchangeability assumptions [27]. Extensions of CP to sequential and online settings (via sliding-window calibration and online recalibration) further demonstrate its potential to maintain valid coverage under distribution shifts, making it particularly relevant for time-varying forecasting scenarios in offshore environments [28,29].

Research Hypothesis. Among the various CP extensions designed for sequential data, the Ensemble Batch Prediction Interval (EnbPI) algorithm stands out as particularly relevant for forecasting FOWT motions. EnbPI constructs distribution-free prediction intervals for dynamic time series by conformalizing residuals obtained from a bootstrap ensemble [30]. EnbPI is model-agnostic, straightforward to implement and has been validated in multiple forecasting domains. Standard EnbPI weights all ensemble members equally and calibrates a single interval width from a sliding window of past residuals. With a long residual window, EnbPI may respond slowly to changes in forecast difficulty. Several studies have investigated *adaptive* or *online* conformalization to tackle this limitation. Among them, Adaptive Conformal Inference (ACI) updates the miscoverage level online from the observed coverage errors [31], and its variants, such as the aggregation of several step sizes through online expert aggregation, further enhance coverage and interval efficiency in the presence of shifts [28]. Parallel lines integrate ensembles with conformalized quantile regression (EnCQR) to derive distribution-free prediction intervals specifically designed for evolving, heteroscedastic series [32]. Recent online control policy methods enhance stability and efficiency through strongly adaptive regret guarantees and systematic learning-rate schedules [29]. These advances highlight the significance of *time-aware residual updates* for calibrated uncertainty quantification in dynamic systems.

Novelty and Contribution. In the context of structural safety, reliable prediction intervals for FOWT platform motions carry direct engineering consequences: prediction intervals that systematically under-cover the true structural response yield inadequate safety margins, potentially compromising mooring integrity, fatigue life estimates, and operational control thresholds. The EnbPI method represents a notable advancement in conformal prediction for time series, grounding probabilistic interval construction in formal coverage theory. However, two critical limitations emerge when applying standard EnbPI to FOWT structural response time series under realistic metocean variability. First, a single prediction interval width, calibrated from a long sliding window of residuals, may not adequately represent the forecast difficulty of the current metocean condition, so that prediction intervals reaching the nominal coverage on average can under-cover the energetic conditions on which structural integrity decisions rely. Second, equal ensemble weighting can fail to account for time-varying model reliability.

To address these limitations and their structural safety implications, this paper presents an Adaptive Conformalized RNN Ensembles (ACRNNE) framework for probabilistic structural response estimation in floating offshore wind turbines. First, an *inverse error weighting mechanism* is introduced, whereby ensemble member contributions are dynamically adjusted at each prediction step based on their recent performance, so that structurally critical uncertainty estimates follow the most recently observed model reliability instead of historical average performance. Second, a *residual pool with weighted conformal inference* is implemented, wherein each residual is stored together with the identities of its contributing members, so that the conformal quantile is obtained by weighting all accumulated residuals, enabling continuous recalibration as new forecast outcomes become available. Third, the coverage level is updated online from the observed outcomes of earlier forecasts, adapting to the varying metocean conditions. The main contribution is the integration and systematic evaluation of these three adaptive mechanisms for FOWT motion forecasting with delayed feedback. The specific contributions of this work can be summarised as follows:

- The ACRNNE framework is developed for probabilistic FOWT structural response estimation, combining inverse error weighting, a weighted residual pool and an online coverage-level update to achieve more consistent empirical coverage across the evaluated metocean conditions.
- ACRNNE is benchmarked against four uncertainty quantification approaches (ensemble, bootstrap, split conformal prediction and EnbPI) to evaluate empirical coverage and interval efficiency. An ablation study examines the three adaptive mechanisms individually and compares them with the complete framework.
- An evaluation across surge, heave and pitch shows improved empirical coverage in energetic metocean conditions relative to the competing conformal baselines.
- The evaluation with LSTM and GRU ensembles examines the consistency of the adaptive calibration across the two recurrent architectures, highlighting the flexibility of the proposed framework to accommodate different RNN-based time-series modelling choices.
- A regime-transition stress test examines adaptation to abrupt changes in forecast difficulty and identifies delayed recovery of coverage.

The remainder of this paper is organised as follows. Section 2 introduces the foundations of conformal prediction and the ACRNNE framework. Section 3 details the experimental setup, including the generation of the dataset, the baselines, the evaluation protocol and the performance metrics. Section 4 presents and discusses the results, covering the overall performance, the coverage across metocean conditions, the contribution of the individual mechanisms, the physical interpretation of the differences among degrees of freedom, the behaviour in boundary metocean conditions and after abrupt regime changes, and the influence of the recurrent architecture. Finally, Section 5 draws the conclusions and outlines future research directions.

2. Proposed framework

This section introduces the theoretical foundations and algorithmic design of the proposed ACRNNE framework for uncertainty-aware motion prediction in FOWTs. First, we review the principles of conformal prediction and outline the standard EnbPI method, establishing the distribution-free basis for prediction interval construction through bootstrap ensembles and out-of-bag residual calibration (Subsection 2.1). Second, we formalise the complete algorithmic workflow and detail the adaptive mechanisms that distinguish the proposed approach, namely the inverse-error weighting strategy (used to dynamically modulate ensemble member influence), the weighted residual pool and the online update of the coverage level, operating under the delayed feedback of multi-step forecasting (Subsection 2.2).

2.1. Fundamentals of conformal prediction

The proposed ACRNNE framework integrates deep learning with distribution-free statistical inference to construct prediction intervals in place of deterministic point estimates. In this context, CP provides a robust framework for uncertainty quantification in regression models by delivering intervals with formal, finite-sample coverage guarantees

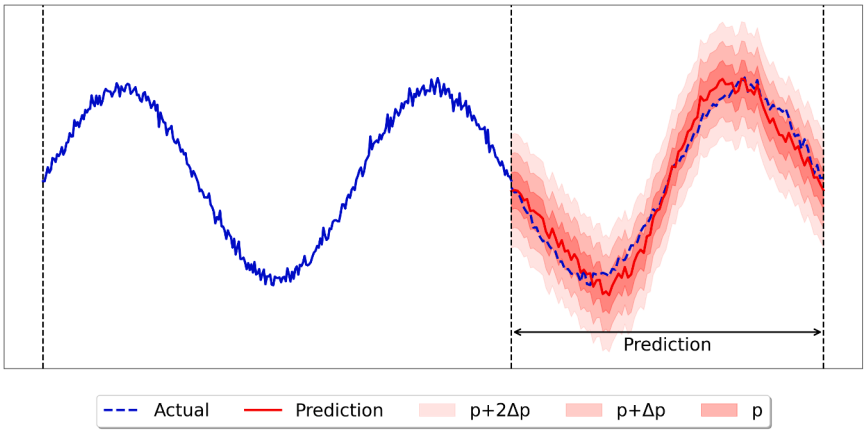


Fig. 1. Schematic illustration of FOWT motion forecasting with nested prediction intervals at nominal coverage levels p, $p + \Delta p$ and $p + 2\Delta p$, where p is the nominal coverage probability of the innermost interval and Δp is the increment between successive levels.

[33,34]. Conformal methods operate by evaluating how well a candidate's response conforms to previously observed data through a nonconformity score, which measures the discrepancy between model predictions and true outcomes. For a new pair $(X^{\text{new}}, Y^{\text{new}})$, the resulting prediction interval $\Gamma(X^{\text{new}})$ satisfies:

$$\mathbb{P}\big(Y^{\text{new}} \in \Gamma(X^{\text{new}})\big) \geq 1 - \alpha, \tag{1}$$

where Y^{new} is the unknown continuous response and $\alpha \in (0, 1)$ is a user-specified miscoverage level. This guarantee ensures that the true response falls within the interval with probability at least $1 - \alpha$, under the standard assumption of exchangeability; the probability is taken jointly over the calibration data and the new pair, so the coverage is marginal and not conditional on a particular input. To construct such an interval, conformal prediction relies on a nonconformity function $\mathcal{A}(X, Y)$ that quantifies how atypical a response Y is for a given input X relative to the training data. Importantly, the nonconformity scores used for interval formation are computed on a *calibration dataset*, i.e., a held-out subset of data that is not used for fitting the predictive model. Given a calibration dataset $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, nonconformity scores $S_i = \mathcal{A}(X_i, Y_i)$ are computed and their finite-sample corrected $(1 - \alpha)$-quantile is used to derive the prediction interval for a new input X^{new} as:

$$\Gamma(X^{\text{new}}) = \Big\{ y \in \mathbb{R} \ : \ \mathcal{A}(X^{\text{new}}, y) \leq q_{1-\alpha} \Big\}, \tag{2}$$

where $q_{1-\alpha} = S_{(k)}$ is the k-th smallest of the n scores, with $k = \lceil (n + 1)(1 - \alpha) \rceil$. If $k > n$, the theoretical formulation sets $q_{1-\alpha} = \infty$; the implementation used in this study instead bounds the interval by the largest score (Section 2.2). In regression problems, this set reduces to an interval (or a union of intervals, depending on $\mathcal{A}(\cdot, \cdot)$), providing an estimate of predictive uncertainty.

Time-series forecasting can be viewed as a specific form of regression in which predictions must be generated for an ordered sequence of future points, requiring prediction intervals to be estimated in advance over the forecast horizon. Motivated by this setting, the present study adopts a sequential prediction scheme based on a sliding-window evaluation, in which a forecast of the motion K steps ahead is issued at every time step from a window of past observations, and the outcome of each forecast becomes available only after this lead time has elapsed. This setup reflects operational forecasting conditions, where only the most recent observations are available for model updates and interval construction. Fig. 1 provides a conceptual illustration of the observed history, point forecasts and nested prediction intervals. The shaded bands represent nominal coverage levels p, $p + \Delta p$ and $p + 2\Delta p$, where p is the coverage probability of the innermost interval and Δp is the increment between successive levels. Higher nominal coverage levels correspond to wider intervals.

Ensemble Prediction Intervals. In this context, EnbPI extends naive conformal predictors to address some limitations in time series prediction. While classical conformal prediction typically relies on a single model to generate nonconformity scores, EnbPI enhances this approach by combining predictions from multiple models trained through bootstrap sampling. This ensemble-based strategy combines the diversity of the members with the conformal calibration of their residuals. EnbPI begins by creating an ensemble of models through bootstrap sampling, where each model is trained on a randomly sampled subset of the training data. This bootstrap procedure ensures diversity in the ensemble by exposing each model to slightly different training data distributions. During inference, EnbPI averages the predictions of all ensemble members and calibrates the interval width using the residual pool initialised from OOB predictions. The width of these intervals is determined by the empirical distribution of the out-of-bag residuals (nonconformity scores), which EnbPI updates in a sliding window as new observations become available [30]. This approach constructs prediction intervals without making strong assumptions about the underlying data distribution, providing practically reliable uncertainty estimates even for autocorrelated and dynamically evolving processes.

To describe the EnbPI approach in detail, let us consider a training dataset $\mathcal{D}_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^{N_{tr}}$, where each $X_i \in \mathbb{R}^{L \times D}$ represents a windowed time-series dataset comprising L past observations with D features, and $Y_i \in \mathbb{R}$ represents the motion K steps after the last sample of the window (K-step-ahead forecasting). The training phase involves M independent models $\{f_m(x; \boldsymbol{\theta}_m)\}_{m=1}^{M}$ using bootstrap sampling data. For each model m, a bootstrap sample $\mathcal{D}_m = [(X_i, Y_i)]_{i \in \mathcal{I}_m}$ is created by randomly sampling with replacement time series samples from the training dataset. Here, $\mathcal{I}_m$ is a sequence (multiset) of indices drawn with replacement from $\{1, 2, \ldots, N_{tr}\}$, so that repeated draws are preserved. It is assumed that $|\mathcal{I}_m| = |\mathcal{D}_{\text{train}}| = N_{tr}$, i.e. bootstraps have the same size as the original training dataset. Therefore, for a given $m \in \{1, \ldots, M\}$, some original training sequences may appear multiple times in $\mathcal{D}_m$, while others may not appear at all. The ensemble prediction for a new test point X^{new} is computed as the average of the predictions of every member in the ensemble:

$$\hat{Y}^{\text{EnbPI}} = \frac{1}{M} \sum_{m=1}^{M} f_m(X^{\text{new}}; \boldsymbol{\theta}_m), \tag{3}$$

where $\hat{Y}_m = f_m(X^{\text{new}}; \boldsymbol{\theta}_m)$ represents the prediction issued by model m. Note that the naive EnbPI approach relies on equal weights to combine the predictions of the members in the ensemble. However, in our proposed adaptive variant, these weights are dynamically adjusted based on recent model performance, as we will later explain in detail.

To construct the prediction intervals, EnbPI employs the concept of "out-of-bag" (OOB) predictions on the training data, so that no

separate calibration subset is required. Given a training sample (X_i, Y_i), $i = 1, \ldots, N_{tr}$, the OOB prediction $\widehat{Y}_i^{\mathrm{OOB}}$ is computed as the average prediction from all models that were not trained on that sample. The nonconformity score is defined as the absolute residual between the true value and this OOB prediction:

$$\mathcal{A}^{\mathrm{EnbPI}}(X_i, Y_i) = \left| Y_i - \widehat{Y}_i^{\mathrm{OOB}} \right|, \qquad \widehat{Y}_i^{\mathrm{OOB}} = \frac{1}{|\mathcal{M}_i^{\mathrm{OOB}}|} \sum_{m \in \mathcal{M}_i^{\mathrm{OOB}}} f_m(X_i; \theta_m), \tag{4}$$

where $\mathcal{M}_i^{\mathrm{OOB}} \subseteq \{1, \ldots, M\}$ denotes the set of models for which the i-th training sample was *out of its bag* (out-of-bag); samples with $\mathcal{M}_i^{\mathrm{OOB}} = \emptyset$ are skipped. Once initial non-conformity scores are computed over the training samples, the prediction interval for a new test sequence X^{new} is estimated as:

$$\Gamma^{\mathrm{EnbPI}}(X^{\mathrm{new}}; \alpha) = \left[\widehat{Y}^{\mathrm{EnbPI}} - q_{1-\alpha},\ \widehat{Y}^{\mathrm{EnbPI}} + q_{1-\alpha} \right], \tag{5}$$

where $q_{1-\alpha}$ represents the finite-sample corrected $(1-\alpha)$-quantile of the OOB nonconformity scores, defined as above. ACRNNE modifies how the ensemble forecasts and these scores are weighted and updated, and how the quantile level is chosen, as described in the next subsection.

2.2. Proposed adaptive mechanisms

As can be inferred from Eqs. (3)–(5), the standard version of the EnbPI approach employs out-of-bag nonconformity scores updated in a sliding window, a fixed coverage level, and equal weighting of all ensemble members. This approach can be suboptimal for complex dynamic processes, as the initial error distribution may not reflect the model's performance on more recent data, and equal weighting does not account for varying model performance over time. To overcome these limitations, in this work we propose three adaptive mechanisms: dynamic model weighting, continuous residual pool refinement and an online update of the coverage level.

The adaptive mechanism operates through initialization and dynamic updating phases. First, the system establishes an initial pool of residuals using out-of-bag (OOB) predictions on the training data. For each training sample, the method identifies which bootstrap models did not use that sample during training and employs only those models to make a prediction. The prediction error (residual) is computed as the difference between the true value and this OOB prediction; training samples without any OOB member are skipped. Each residual is stored together with the identities of its contributing members, providing a weighted residual pool for conformal quantile calculation. This approach reduces the optimism of in-sample residuals, since each residual is computed by models that did not see the sample during training. All forecasts, targets and residuals are expressed in physical units, after the normalisation described in Section 3 has been reversed.

During inference, the adaptive mechanism updates both model weights and the residual pool at each time step. Throughout, X_h denotes the observed input window at forecast origin h, and Y_h its K-step-ahead target, which is observed at origin $h + K$. After making a prediction $\widehat{Y}_h$ at origin h of the motion K steps ahead and observing its true value Y_h once these K steps have elapsed, the system: (1) computes individual prediction errors for each ensemble member, (2) updates model weights using inverse error weighting:

$$w_m^{(h+K)} = \frac{1/(e_m^{(h)} + \epsilon)}{\sum_{m'=1}^{M} 1/(e_{m'}^{(h)} + \epsilon)}, \tag{6}$$

where $e_m^{(h)} = |f_m(X_h) - Y_h|$ is the prediction error of model m for the forecast issued at origin h, measured once the true value Y_h has been observed, and ϵ is a small constant to prevent division by zero. The resulting weights $w_m^{(h+K)}$ are applied from origin $h + K$ onwards; equivalently, the weights used at origin h are computed from the error $e_m^{(h-K)} = |f_m(X_{h-K}) - Y_{h-K}|$ of the most recent forecast whose outcome has been observed, as in Algorithm 1, so that no unobserved value enters the forecast issued at origin h. The system then (3) appends the new residual $Y_h - \widehat{Y}_h$ to the pool along with metadata indicating which models contributed. This adaptive weighting ensures that well-performing models receive higher influence in subsequent predictions, while the growing residual pool maintains a comprehensive history of prediction errors. Critically, the conformal quantile is computed using a dynamic weighted quantile procedure over all accumulated residuals. At each prediction step, every historical residual (including those from training) is re-weighted by summing the current model weights for the models that originally generated it. This means residuals from models currently performing poorly are down-weighted, even if those models performed well when the residual was generated. This dynamic re-weighting mechanism allows the residual distribution to follow the most recently observed model performance. Formally, the pool at origin h is the indexed list $S_h = [(r_j, \mathcal{M}_j)]_{j=1}^{n_h}$ of the n_h stored residuals r_j and the sets $\mathcal{M}_j$ of members that generated them, so that repeated residual values are retained. Each entry receives the raw and normalised weights

$$a_{j,h} = \sum_{m \in \mathcal{M}_j} w_m^{(h)}, \qquad \widetilde{w}_{j,h} = \frac{a_{j,h}}{\sum_{\ell=1}^{n_h} a_{\ell,h}}. \tag{7}$$

A residual appended during prediction has $\mathcal{M}_j = \{1, \ldots, M\}$ and therefore a raw weight of one, since all members contributed to the forecast. Even with equal member weights, the OOB residuals do not receive equal weights, because their raw weight $|\mathcal{M}_j|/M$ depends on the number of members that excluded the sample.

Finally, (4) the coverage level is updated online. Instead of a fixed miscoverage level α, ACRNNE uses a level α_h that is updated from the observed outcomes of earlier forecasts, following Adaptive Conformal Inference [31]:

$$\alpha_h = \begin{cases} \alpha_{h-1}, & h \le K, \\ \alpha_{h-1} + \gamma \left[\alpha - \mathbb{I}\left(Y_{h-K} \notin \widehat{C}_{h-K} \right) \right], & h > K, \end{cases} \qquad \alpha_0 = \alpha, \tag{8}$$

where $\widehat{C}_{h-K}$ is the interval issued at origin $h - K$, the update is applied only once its outcome has been observed, and $\mathbb{I}(\cdot)$ is a binary indicator function taking value 1 if its argument is true (0 otherwise). A miss decreases α_h and widens the subsequent intervals, whereas a sequence of hits gradually narrows them. The step size is $\gamma = 0.005$ and the target level is $\alpha = 0.1$. The interval at origin h is $\widehat{C}_h = [\widehat{Y}_h - q_h,\ \widehat{Y}_h + q_h]$, where q_h is obtained from the n_h absolute residuals in the pool as:

$$q_h = \begin{cases} \max_{1 \le j \le n_h} |r_j|, & \alpha_h \le 0, \\ Q_h\left(\min\left\{ \lceil (n_h + 1)(1 - \alpha_h) \rceil / n_h,\ 1 \right\} \right), & 0 < \alpha_h < 1, \\ 0, & \alpha_h \ge 1, \end{cases} \tag{9}$$

with the weighted quantile:

$$Q_h(\tau) = \inf \left\{ q \ge 0 : \sum_{j=1}^{n_h} \widetilde{w}_{j,h}\, \mathbb{I}\left(|r_j| \le q \right) \ge \tau \right\}. \tag{10}$$

If $\lceil (n_h + 1)(1 - \alpha_h) \rceil > n_h$, which occurs when $\alpha_h < 1/(n_h + 1)$, standard conformal prediction would return an unbounded interval ($q_h = \infty$); the minimum in Eq. (9) caps the quantile level at one, so that the half-width is instead set to the largest residual in the pool. The level α_h itself is not clipped. These boundary rules define our implementation of an ACI-inspired update and differ from the endpoint conventions of the original formulation of Gibbs and Candès [31], in which a non-positive level yields an unbounded interval; here, the half-width is instead bounded by the largest residual in the pool. The boundary cases do occur in practice: in energetic metocean conditions with repeated misses, α_h can become non-positive for a substantial fraction of the forecast origins, during which the interval half-width equals the largest residual in the pool. Moreover, the adjusted level is applied to the weighted quantile of the residual pool (Eq. (10)), not to the unweighted empirical quantile of the original formulation, so that the coverage-level update and the member weights act jointly on the interval width. All

Algorithm 1 ACRNNE: Adaptive Conformalized RNN Ensemble Framework (K-step-ahead forecasting).

1: **Input:** Training data $\mathcal{D}_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^{N_{tr}}$, # models M, significance α, lead K, step size γ, constant $\epsilon > 0$
2: **Output:** Prediction intervals $\widehat{C}_h$ for forecast origins $h = 1, \ldots, H$
3: **Training Phase:**
4: **for** $m = 1, \ldots, M$ **do**
5: Bootstrap sample: index sequence $\mathcal{I}_m$ of N_{tr} draws from $\{1, \ldots, N_{tr}\}$ with replacement; $\mathcal{D}_m = [(X_i, Y_i)]_{i \in \mathcal{I}_m}$
6: Train model f_m on $\mathcal{D}_m$
7: **end for**
8: **Initialise Residual Pool from Training Data:**
9: $\mathcal{S} \leftarrow [\,]$ ▷ Empty indexed list
10: **for** $i = 1, \ldots, N_{tr}$ **do**
11: $\mathcal{M}_i^{\text{OOB}} \leftarrow \{m : i \notin \mathcal{I}_m\}$ ▷ Out-of-bag models for sample i
12: **if** $\mathcal{M}_i^{\text{OOB}} = \emptyset$ **then** skip sample i
13: $\widehat{Y}_i^{\text{OOB}} \leftarrow \frac{1}{|\mathcal{M}_i^{\text{OOB}}|} \sum_{m \in \mathcal{M}_i^{\text{OOB}}} f_m(X_i)$ ▷ OOB prediction
14: Append $(Y_i - \widehat{Y}_i^{\text{OOB}},\ \mathcal{M}_i^{\text{OOB}})$ to $\mathcal{S}$ ▷ Store residual + contributing model IDs
15: **end for**
16: **Sequential Prediction Phase:**
17: **for** each test sample **do**
18: $w_m \leftarrow 1/M$ for all m; $\alpha_0 \leftarrow \alpha$; $\mathcal{S}$ reset to the initial pool ▷ Reset
19: **for** $h = 1, \ldots, H$ **do**
20: $\alpha_h \leftarrow \alpha_{h-1}$
21: **if** $h > K$ **then** ▷ Outcome of origin $h - K$ is observed
22: $e_m \leftarrow |f_m(X_{h-K}) - Y_{h-K}|$ for all m
23: $w_m \leftarrow \frac{1/(e_m+\epsilon)}{\sum_j 1/(e_j+\epsilon)}$ for all m ▷ Inverse error weighting
24: Append $(Y_{h-K} - \widehat{Y}_{h-K},\ \{1, \ldots, M\})$ to $\mathcal{S}$ ▷ Raw weight one
25: $\alpha_h \leftarrow \alpha_{h-1} + \gamma\,(\alpha - \mathbb{I}(Y_{h-K} \notin \widehat{C}_{h-K}))$ ▷ Coverage-level update
26: **end if**
27: $\widehat{Y}_h \leftarrow \sum_{m=1}^{M} w_m f_m(X_h)$ ▷ Weighted ensemble prediction
28: **Construct interval via weighted quantile:**
29: $n_h \leftarrow |\mathcal{S}|$ ▷ Includes newly appended residuals
30: Re-weight all residuals in $\mathcal{S}$ using current weights $\{w_m\}$:
31: **for** $j = 1, \ldots, n_h$ **do**
32: $a_j \leftarrow \sum_{m \in \mathcal{M}_j} w_m$
33: **end for**
34: Normalise: $\widetilde{w}_j \leftarrow a_j / \sum_{\ell=1}^{n_h} a_\ell$ ▷ Weights sum to unity
35: $q_h \leftarrow$ Eq. (9): weighted quantile Q_h of $\{|r_j|\}$ with weights $\{\widetilde{w}_j\}$ at level $\min\{\lceil (n_h+1)(1-\alpha_h)\rceil / n_h, 1\}$; $\max_j |r_j|$ if $\alpha_h \le 0$; 0 if $\alpha_h \ge 1$
36: $\widehat{C}_h \leftarrow [\widehat{Y}_h - q_h, \widehat{Y}_h + q_h]$ ▷ Y_h is observed at origin $h + K$
37: **end for**
38: **end for**

adaptive quantities (weights, residual pool and coverage level) are re-initialised at the start of each test record.

Algorithm 1 presents the ACRNNE framework, and Fig. 2 provides a visual overview of the complete pipeline. The method trains M bootstrap ensemble models (lines 4–7), then initialises a residual pool using out-of-bag (OOB) predictions on training data (lines 8–15). For each training sample i, only out-of-bag models $\mathcal{M}_i^{\text{OOB}}$ that excluded sample i from their bootstrap sets are used to compute the OOB prediction, so that each residual is computed by models that did not see the sample during training. The resulting residuals are stored with their contributing model IDs for subsequent adaptive re-weighting. During sequential prediction (lines 16–38), the framework operates in an *online K-step-ahead forecasting* mode: at origin h, the forecast uses only data observed up to origin h, the model weights are updated from the error $|f_m(X_{h-K}) - Y_{h-K}|$ of the most recent forecast whose outcome has been *observed*, and the residual of each forecast, together with the corresponding coverage-level update, is applied only *after* its target has been observed, K steps later. No future, unobserved value enters the prediction or interval construction at origin h. A central element is the weighted conformal quantile computation (lines 28–35): each stored residual in pool $\mathcal{S}$ carries metadata indicating which models generated it, and at prediction step h, each residual receives a raw weight equal to the sum of the current weights of its generating models (Eq. (7)). Since individual residual weights can vary, the weights are normalised before computing the quantile to ensure they sum to unity, enabling proper weighted quantile computation. This dynamic re-weighting ensures that residuals from the members that were most accurate on the latest observed forecasts influence the quantile more strongly than those from less accurate members, allowing the interval width to reflect recent, not past, model reliability. Combined with the online coverage-level update, the resulting quantile q_h widens the intervals after repeated misses and narrows them after repeated hits, so that the interval width adapts to the current record.

The finite-sample guarantee of standard conformal prediction is derived under exchangeability between calibration and test observations. Platform-motion records are temporally autocorrelated, and ACRNNE further incorporates data-dependent ensemble weights, a growing residual pool, and an adaptive miscoverage level. Consequently, the present evaluation focuses on the empirical coverage achieved on held-out test records. Accordingly, the reported intervals should be interpreted as empirically calibrated uncertainty estimates, not as distribution-free coverage guarantees.

3. Research questions and experimental setup

A comprehensive assessment of our proposed method's performance, advantages, and limitations in the prediction of FOWT motions is

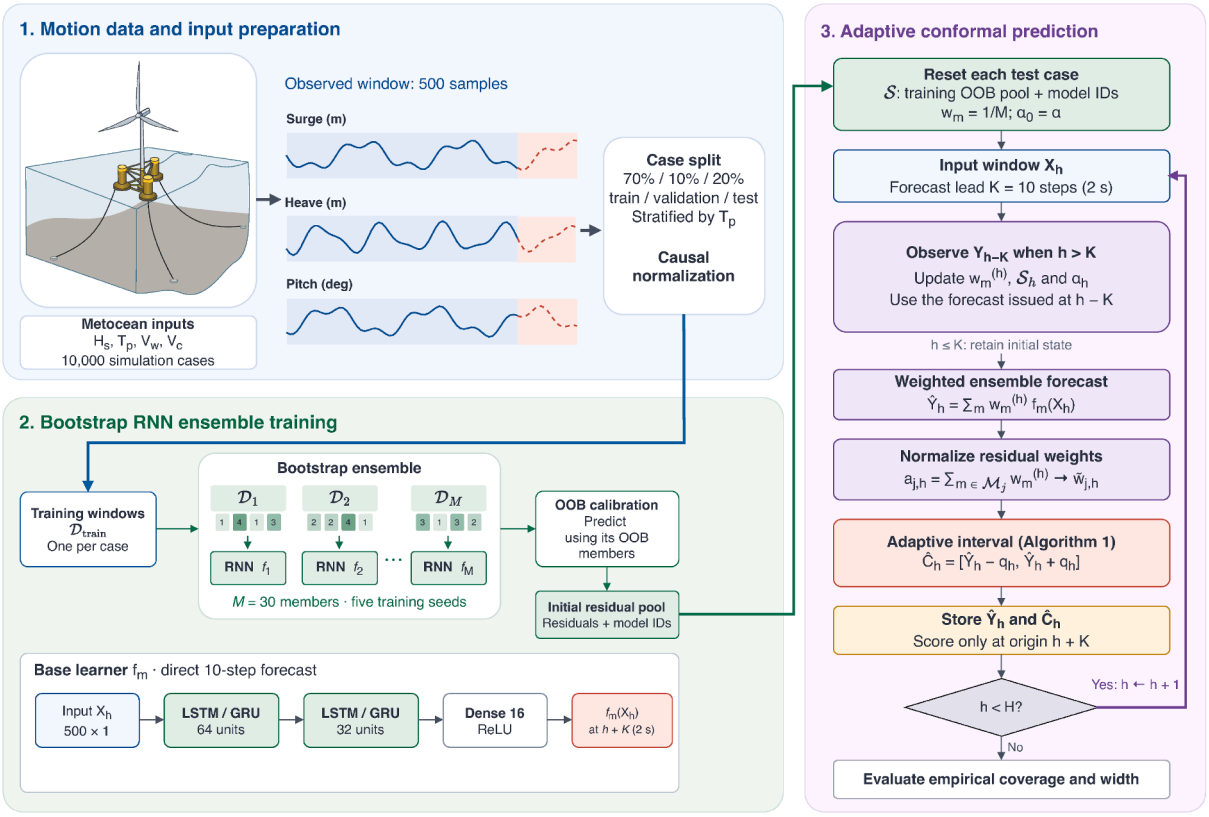


Fig. 2. ACRNNE workflow for conformal prediction, from data preparation, in which the blue segments show the observed 500-sample (100 s) input history and the red dashed segments the subsequent period over which 2 s-ahead forecasts are issued, through bootstrap ensemble training to the online construction of the adaptive intervals. Specific settings shown, such as the base learners, data split, window length, ensemble size and K, refer to the present experiments; the framework can be applied with other choices. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

provided by organising a structured experimental numerical campaign. Specifically, our experimental setup is designed to address three Research Questions (RQs):

- **RQ1.** *Are the prediction intervals produced by the proposed ACRNNE approach consistent in coverage and interval efficiency over time, across different degrees of freedom (DoF) of the FOWT platform and across metocean conditions?* Understanding how coverage and interval width vary over time and across DoFs is requisite for assessing the method's reliability with respect to uncertainty quantification in the estimation of FOWT motions, and more generally when predicting autocorrelated responses. Coverage that is adequate only on average may conceal failures in the metocean conditions that govern structural loads.
- **RQ2.** *How do the coverage and interval efficiency of the ACRNNE compare to those generated by alternative uncertainty quantification approaches?* The ACRNNE approach is benchmarked against existing methods (bootstrap, ensemble, split conformal prediction and EnbPI) to quantify its relative performance, and an ablation study examines the adaptive mechanisms individually and compares them with the complete framework. The ultimate goal of this comparison is to highlight the performance gains of the proposed ACRNNE framework with respect to traditional uncertainty quantification techniques.
- **RQ3.** *How does the choice of the RNN variant within the ensemble influence the resulting prediction intervals?* The effect of different RNN architectures on the ensemble's output is examined to determine whether certain variants provide more reliable or accurate prediction intervals. This analysis informs model selection and provides insight into the flexibility of the approach across different neural network configurations.

In the rest of this section, we detail the experimental setup devised to provide empirical evidence for our responses to the above RQs:

Data generation. The dataset used for developing, validating, and testing the novel ACRNNE framework is generated synthetically. Indeed, due to both the infancy of the FOWT sector and confidentiality constraints, the availability of real-world measurements or data obtained under controlled experimental conditions remains limited. For this reason, datasets produced using validated numerical simulators [11] are widely adopted. Simulating the coupled response of a floating offshore wind turbine requires modelling the complex interactions between ocean waves, aerodynamics, and structural dynamics, which entails solving multiple interdependent governing equations. In this context, the National Renewable Energy Laboratory (NREL) has developed *OpenFAST*, an open-source, modular simulation framework designed to predict the aero-hydro-servo-elastic behaviour of wind turbines under diverse operating conditions. To ensure its reliability, OpenFAST has undergone extensive verification and validation efforts carried out by NREL and other independent researchers [35–37]. Owing to its modular structure and proven fidelity, OpenFAST is therefore commonly used to simulate the response of floating platforms under prescribed metocean conditions for dataset generation. Despite the extensive validation of OpenFAST, the configuration used in this study involves several simplifying assumptions. Hydrodynamic loads are computed by combining linear potential-flow theory with Morison-type quadratic drag terms with empirically determined, constant drag coefficients. The environmental inputs are also idealised: wave loads are simulated from a Pierson–Moskowitz spectrum, and wind, waves and current are assumed aligned and stationary over each simulation. Real-world artefacts such as sensor noise, marine growth and mooring degradation are also not

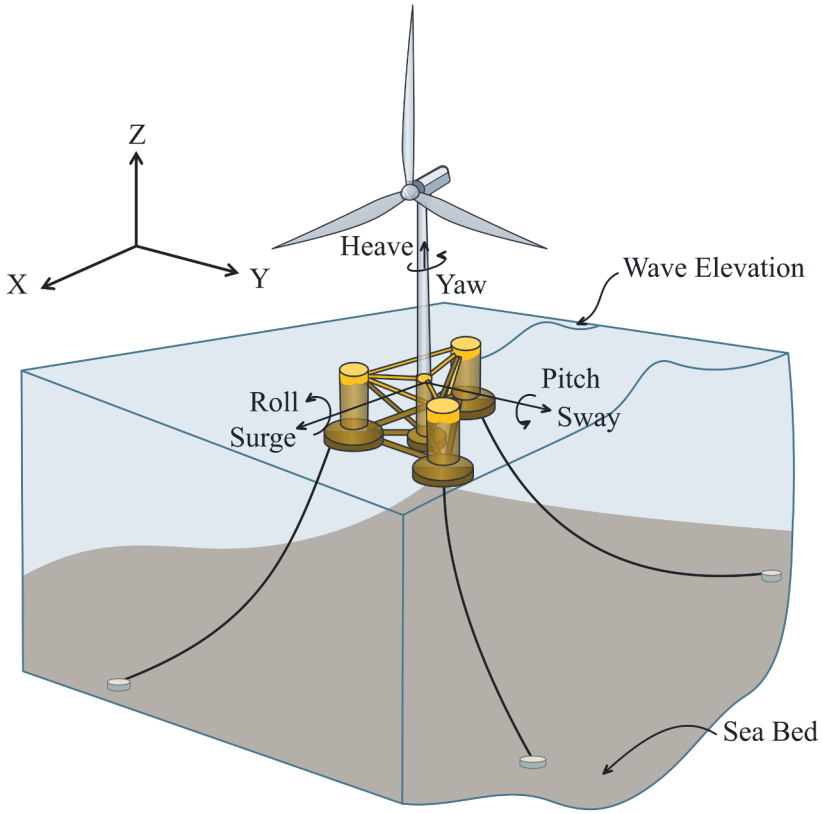


Fig. 3. Schematic diagram of the DeepCwind semi-submersible platform with three-bladed rotor configuration.

represented. These simplifications are deliberate within the scope of this work. OpenFAST simulations are used here as a validated baseline that is representative of the platform dynamics. This baseline allows the proposed framework to be assessed systematically over 10,000 metocean conditions with known simulated response histories, which is not feasible with the limited field data currently available for FOWTs. Although the ACRNNE procedure is not specific to OpenFAST, its present training and evaluation inherit the assumptions of the simulated dataset. Its coverage on measured responses, which are affected by additional sources of uncertainty such as sensor noise and drift, turbulent and misaligned wind and wave conditions with directional spreading, marine growth and mooring degradation, therefore requires separate validation.

In this paper, the case studies rely on the NREL 5 MW baseline floating offshore wind turbine [38], mounted on the DeepCwind OC4 semi-submersible platform [39] and modelled in OpenFAST. An overview of this platform configuration is presented in Fig. 3. The floater consists of a central column connected to three peripheral offset columns, with an array of slender bracings providing additional structural rigidity. The platform is station-kept by three catenary mooring lines attached 14.0 m below the mean sea level at the fairleads, as depicted in Fig. 4. The mooring system is modelled using MoorDyn, the dedicated mooring dynamics module within OpenFAST, and its key parameters are summarised in Table 1. The simulations are performed with OpenFAST v3.5.5, coupling the structural (ElastoDyn), aerodynamic (AeroDyn), control (ServoDyn), hydrodynamic (HydroDyn) and mooring (MoorDyn) modules; the main settings, including the wave spectrum, the random wave seeds, the wind and current models and the controller, are summarised in Table 2.

We trained our models across a broad spectrum of metocean conditions. These conditions include variations in wave height, wave period, wind speed, and current velocity. The specific metocean parameters considered in this study are listed in Table 3 [40], and are representative of the normal operating range for the semi-submersible FOWT under investigation. Environmental loads, i.e., due to waves, wind, and current,

Table 1
Properties of the mooring line used in our experiments.

Line Properties	Value
Mooring Line Diameter	0.0766 m
Mooring Line Length	835.35 m
Equivalent Line Mass Density in Air	113.35 kg/m
Equivalent Mooring Line Stiffness	753.6×10^6 N

Table 2
Main settings of the OpenFAST simulations.

Item	Setting
Solver	OpenFAST v3.5.5; time step 0.0125 s
Structure	ElastoDyn (platform 6 DoF, tower and blade flexibility)
Aerodynamics	AeroDyn v15, blade-element momentum with unsteady airfoil aerodynamics
Control	ServoDyn, NREL 5 MW baseline controller (OC3-Hywind variant)
Hydrodynamics	HydroDyn: WAMIT potential flow, Morison drag, mean-drift forces
Mooring	MoorDyn v2, lumped-mass model
Waves	Pierson–Moskowitz spectrum, 0.31–4.9 rad/s, independent random seed per run
Wind and current	Steady, uniform hub-height wind; sub-surface current profile
Directions	Wind, waves and current aligned (0 deg)
Duration	700 s, first 500 s discarded; 200 s recorded at 0.2 s

Table 3
Simulation setup, environmental conditions, and total combinations.

Property	Value	Units	Number of Combinations
Sampling time step	0.20	s	-
Transient time	500	s	-
Significant height, H_s	2.1:0.43:6.02	m	10
Peak period, T_p	2:1.44:15	s	10
Wind speed, V_w	3:2.44:25	m/s	10
Current velocity, V_c	0.55:0.06:1.1	m/s	10
Total unique combinations			$\mathbf{10 \times 10 \times 10 \times 10 = 10,000}$

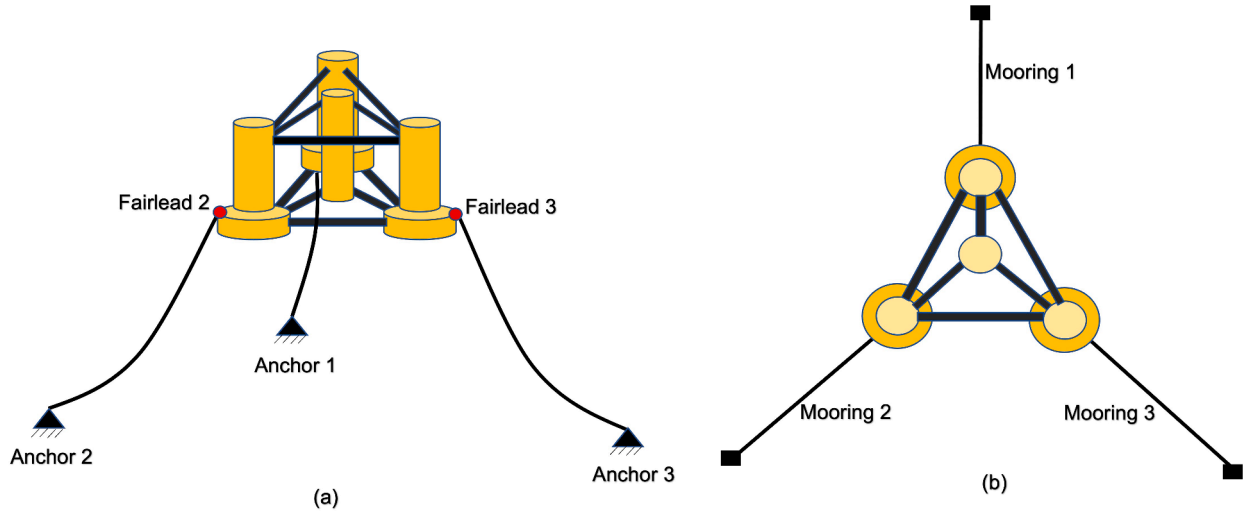


Fig. 4. Multipoint mooring line configuration of floating platform: (a) front view and (b) top view.

are assumed to be aligned throughout all simulations. Each environmental parameter (significant wave height H_s, peak period T_p, current velocity V_c and wind speed V_w) is discretized into 10 distinct values within its respective range, and the independent combination of these values yields a total of 10,000 unique simulation cases. For each case, the numerical model records six time-history responses associated with the six DoFs of the floating platform: surge, sway, heave, roll, pitch, and yaw. However, a recent study by Sharma and Nava [41] using the same platform configuration observed that sway, roll, and yaw motions exhibit negligible variation due to the platform's geometric symmetry and the unidirectional alignment of environmental loads. Motivated by these findings, the present work focuses on the surge, heave, and pitch DoFs, which capture the dominant dynamic behaviour of the system. Each OpenFAST simulation spans a total duration of 700 s. The initial 500 s correspond to the transient response as the system evolves towards steady-state behaviour, while the remaining portion of the time series (200 s, recorded at 0.2 s) forms the dataset used for model development and evaluation.

The only stochastic input of the simulations is the random phase of the wave components, drawn from an independent seed for every simulation; wind and current are steady. The prediction intervals therefore quantify the uncertainty of the 2 s-ahead forecast of a given realisation conditional on its observed history, with coverage measured over the sampled metocean conditions and realisations. Since each metocean combination is represented by a single realisation, the variability of the response across realisations of the same metocean condition is not characterised separately.

RNN Variants and Baselines. The ACRNNE framework is implemented using two prominent recurrent neural network (RNN) architectures as base learners: Long Short-Term Memory (LSTM) [42] and Gated Recurrent Unit (GRU) [43]. Employing both architectures enables the framework to capture a broad range of temporal dynamics and demonstrates the flexibility of the proposed approach in accommodating heterogeneous model behaviours within the ensemble.

To thoroughly assess the advantages brought by ACRNNE, the performance of the proposed framework is benchmarked against four established uncertainty quantification techniques. These baseline methodologies encompass distinct paradigms (namely, conformal prediction, ensemble modelling, and bootstrap aggregation, together with their combination in EnbPI), each contributing fundamental principles that are extended and unified within ACRNNE. Specifically:

- **EnbPI:** EnbPI uses a bootstrap ensemble to obtain out-of-bag residuals from the training data [30]. At each forecast origin, it forms an equally weighted ensemble prediction and constructs a symmetric interval from the finite-sample corrected quantile of a sliding window of residuals, whose length equals the number of out-of-bag residuals (7,000) and which is updated as new forecast outcomes become available. In the present study, EnbPI uses a fixed miscoverage level α and equal member weights; ACRNNE modifies these two elements through adaptive weighting and an online miscoverage update.
- **Split Conformal Predictor (split CP):** This method provides the theoretical foundation for distribution-free uncertainty quantification [27,33,34]. It relies on a single neural network trained on 80% of the training simulations, with the remaining 20% held out as calibration set to derive prediction intervals from the empirical distribution of residuals. These static nonconformity scores form the baseline conformal backbone upon which ACRNNE is built.
- **Ensemble Method:** This approach captures model uncertainty through diversity across multiple networks trained on identical data but initialised differently [24,25]. Prediction intervals are constructed from the spread of ensemble outputs, spanning the 5th to 95th percentiles of the forecasts of 30 networks. This method provides the foundation for model diversity, which is further enhanced in our adaptive weighting strategy. Ensemble learning has also demonstrated strong potential for structural performance prediction in marine environments; Guo et al. [44] developed an ensemble learning framework for life-cycle performance prediction of

Table 4
RNN architectures and hyperparameter settings for ACRNNE models.

Parameter	Value
Architecture (per motion model)	
Base architectures	LSTM, GRU
Input sequence length	500 timesteps
Output features	1 (motion 2 s ahead)
Recurrent units (Layer 1)	64
Recurrent units (Layer 2)	32
Dense layer units	16
Activation (Dense)	ReLU
Dropout rate	0.2
Training	
Optimizer	Adam
Learning rate	0.001
Batch size	32
Maximum epochs	50
Early stopping patience	10 (on validation loss)
Ensemble size	30
Implementation	PyTorch 2.9.1 (CUDA 13.0), NVIDIA GB10 GPU

coastal and marine structures, confirming the applicability of such methods in offshore structural engineering.

- **Bootstrap Method:** Extending the ensemble paradigm, the bootstrap approach introduces data uncertainty by training models on different bootstrap resamples of the original dataset [45]. The distribution of predictions reflects both model and data variability; the interval spans the 5th to 95th percentiles of the forecasts of the 30 bootstrap members, which are shared with EnbPI and ACRNNE. This forms the basis of the bootstrap ensemble used in standard EnbPI, from which ACRNNE is evolved. The utility of bootstrap resampling for structural reliability assessment has been further demonstrated by Modak and Chakraborty [46], who integrated bootstrap-based polynomial chaos expansion with an active learning framework for structural safety analysis, confirming the suitability of bootstrap-based methods for uncertainty-critical engineering problems.

The proposed ACRNNE method advances beyond these baselines by combining the distribution-free approach of conformal prediction with the model diversity provided by ensemble and bootstrap techniques. Crucially, it introduces adaptive residual updating and an online update of the coverage level during the prediction phase, allowing the interval estimation to adjust continuously to the observed response and to the current accuracy of the ensemble members. This adaptation mitigates the limitations of a fixed interval width, particularly for autocorrelated time series.

To ensure fairness and methodological consistency, all baseline methods and the proposed ACRNNE framework employ identical RNN architectures, hyperparameters, and training configurations, as summarised in Table 4. The only structural difference is the requirement of split CP to reserve a separately held-out calibration set, which is intrinsic to its formulation and not a tuning choice. This choice ensures that differences in performance arise from the uncertainty quantification methodology, not from difference in model complexity or training strategy.

Data Preprocessing and Partitioning. The dataset preparation follows a multi-stage preprocessing pipeline designed to enhance model convergence and prediction accuracy. First, mean removal is performed on each input window. For a given input window $\mathbf{x}_h = [x_{h-L+1}, \dots, x_h]$ of length $L = 500$, the window mean is computed as:

$$\mu_h = \frac{1}{L} \sum_{i=0}^{L-1} x_{h-i}, \tag{11}$$

which is then subtracted from each time step of the window and from the forecast target to obtain the centred sample:

$$\widetilde{x}_{h-i} = x_{h-i} - \mu_h, \quad \widetilde{Y}_h = Y_h - \mu_h. \tag{12}$$

Only the observed history enters the centring, so no future value is used in the normalisation. This mean-centering operation centres the data around zero while preserving the amplitude and frequency characteristics of FOWT motions. This preprocessing step addresses the varying operating points across different metocean conditions, allowing the models to focus on learning the dynamic variations instead of static offsets.

Following mean removal, the dataset is partitioned into training, validation and testing subsets using a 70–10–20 split across the independent simulation cases, stratified by peak period, so that no simulation contributes to more than one subset. For each training seed, one 500-sample input window is selected at a random valid position from each training case and shared across methods. Within the training phase, the centred windows are scaled by a single factor, equal to the standard deviation of all centred training-window values, with the scaler fitted exclusively on the training data. The same scale and the corresponding input-window mean are used to transform the targets and to restore the predictions to physical units. The same scaling transformation is subsequently applied to the test data to prevent data leakage, ensuring that no information from the test set influences model training.

Within each test case, the framework is evaluated in an *online K-step-ahead* forecasting mode: at every origin h, the model uses only the most recent 500 observed timesteps to predict the motion $K = 10$ steps (2 s) ahead; once this value becomes available, K steps later, it is used to update the ensemble weights, the residual pool and the coverage level. The predictor never accesses values beyond the current sampling instant. This procedure is repeated for 250 consecutive forecast origins per case, simulating a streaming sensor-driven operation in which the framework continuously re-calibrates as new measurements arrive; each forecast looks 2 s ahead, the evaluation covers 50 s per test simulation, and the state of every method is re-initialised at the start of each test simulation.

Model Configuration and Training. All methods employ identical RNN architectures as detailed in Table 4, with hyperparameter settings consistent across implementations. For ensemble-based methods (Bootstrap, Ensemble, EnbPI and ACRNNE), 30 models are trained to balance diversity and computational cost. Training employs the Adam optimizer with a learning rate of 0.001, batch size of 32, and early stopping with patience of 10 epochs based on validation loss. The validation simulations are used to monitor convergence and prevent overfitting. Each experiment is repeated with five independent random seeds, which control the network initialisation, the bootstrap samples, the positions of the training windows and the calibration subset of split CP, while the data partition is fixed so that all methods are compared on identical test simulations. This architectural and training standardisation ensures that performance differences across methods can be attributed exclusively to their uncertainty quantification strategies and not to variations in model capacity or optimization protocols.

Performance Metrics. The performance of ACRNNE is evaluated using a comprehensive set of metrics that capture different aspects of prediction quality. All metrics are computed on the window-centred data (after mean removal as described in the preprocessing section), ensuring that the evaluation focuses on the dynamic motion components and excludes static offsets. Targets, forecasts and interval bounds are expressed in physical units (m, deg) relative to the mean of the corresponding input window; residuals, coverage and interval widths are identical in centred and uncentred coordinates. The metrics are:

- The Mean Absolute Error (MAE) is computed to provide a linear measure of the average magnitude of errors between predicted and actual values. For this analysis, all actual and predicted values from every timestep across all test samples are first collected. The MAE is then calculated as the average of the absolute differences between these corresponding values. Formally:

$$\mathrm{MAE} = \frac{1}{H \cdot N} \sum_{n=1}^{N} \sum_{h=1}^{H} |y_{n,h} - \hat{y}_{n,h}|, \tag{13}$$

  where $y_{n,h}$ is the observed platform motion targeted by the forecast issued at origin h of test sample n, i.e. the motion K steps after the origin, and $\hat{y}_{n,h}$ is the corresponding prediction. The calculation averages the absolute errors across all forecast origins (H) and test samples (N).
- The Prediction Interval Coverage Probability (PICP) is used to quantify the reliability of the prediction intervals. It is defined as the proportion of observed values that fall within their predicted lower and upper bounds. For each timestep in the test set, an indicator function determines if the true value is captured within the interval. The PICP is then calculated as the mean of these indicator values across all timesteps for all test samples, as shown in the equation:

$$\mathrm{PICP} = \frac{1}{H \cdot N} \sum_{n=1}^{N} \sum_{h=1}^{H} \mathbb{I}\big(y_{n,h} \in [L_{n,h}, U_{n,h}]\big), \tag{14}$$

  where $U_{n,h}$ and $L_{n,h}$ are the upper and lower bounds of the prediction interval issued at origin h for test sample n. Here, $\mathbb{I}(\cdot)$ is the

indicator function defined in Section 2.2. For a model calibrated to a confidence level of $1-\alpha$, the PICP is expected to be close to this value.

- The Prediction Interval Normalised Average Width (PINAW) is employed to measure the precision of the uncertainty estimates. A smaller PINAW is desirable, as it indicates sharper, more informative prediction intervals, provided the PICP is maintained at the nominal level. In this study, the average width of all prediction intervals across the entire test set is first calculated. This average width is then normalised by the range of the true observed values, where the range (R) is defined as the difference between the global maximum and minimum of the window-centred true values across all test samples and time steps. This is expressed as:

$$\text{PINAW} = \frac{\overline{W}}{R} \quad \text{where} \quad \overline{W} = \frac{1}{H \cdot N} \sum_{n=1}^{N} \sum_{h=1}^{H} (U_{n,h} - L_{n,h}), \tag{15}$$

and $R = \max_{n,h}(y_{n,h}) - \min_{n,h}(y_{n,h})$ represents the global range across all test samples and time steps. This normalisation allows for a fair comparison of interval widths across datasets with different scales.

- The Coverage Width-Based Criterion (CWC) synthesises both reliability (PICP) and precision (PINAW) into a single score. This metric rewards intervals that are both narrow and achieve the target coverage level, while penalising intervals that are wide or fail to meet the desired coverage. The CWC value is computed as:

$$\text{CWC} = (1 - \text{PINAW}) \cdot \exp\left(-\eta \cdot (\text{PICP} - (1-\alpha))^2\right), \tag{16}$$

where α is the significance level (equal to 0.1 for 90% prediction intervals). The penalty parameter η controls the severity of the penalty for coverage deviations, and is set equal to 30 in our study. The $(1 - \text{PINAW})$ term rewards narrower intervals, while the exponential term penalises models whose PICP deviates from the target coverage of $(1-\alpha)$. A higher CWC score is indicative of a better balance between coverage and width. However, this variant of the criterion of [47] penalises deviations of the average coverage in both directions and does not reflect how coverage is distributed across metocean conditions.

Since average coverage may hide failures in particular conditions, the analysis additionally reports the *lowest group coverage*, defined as the lowest PICP among the ten peak-period groups of the grid. Results are averaged over the five seeds. The 95% confidence intervals reported in Appendix A (Table A.8) are t-based across the five training seeds on the fixed test split, and paired Wilcoxon signed-rank tests compare the methods using per-simulation metrics averaged across the seeds. The results presented and discussed in the following section are derived from the complete test set of 2000 simulations, each evaluated under the online 2 s-ahead protocol at 250 consecutive forecast origins (50 s), and are averaged over five random seeds.

4. Results and discussion

This section presents the evaluation of the ACRNNE method, addressing the three research questions outlined in Section 3. The analysis focuses on the temporal, cross-DoF and cross-condition consistency of the prediction intervals, a comparative assessment against established baseline methods including an ablation of the ACRNNE mechanisms, and the impact of the chosen RNN architecture.

RQ1. Are the prediction intervals produced by the proposed ACRNNE approach consistent in coverage and interval efficiency over time and across different DoFs and metocean conditions of the FOWT platform?

We begin our analysis of the results with Fig. 5, which illustrates the coverage and interval width across surge, heave, and pitch as a function of the peak period. The results are presented as a function of T_p because the data are split by peak period: each of the ten T_p values of the grid contributes the same number of simulations to the test set (200), so the coverage of every T_p class is estimated from an equally large sample.

Figs. 5(a)– 5(c) show empirical coverage as a function of peak period for surge, heave and pitch, respectively. All conformal methods overcover heave and pitch in metocean conditions with shorter peak periods, where the semi-submersible response is comparatively small. With increasing T_p, the coverage of EnbPI and split CP declines, falling below the 90% target in the longest-period cases, particularly for heave. In contrast, ACRNNE remains closer to the nominal level across the tested

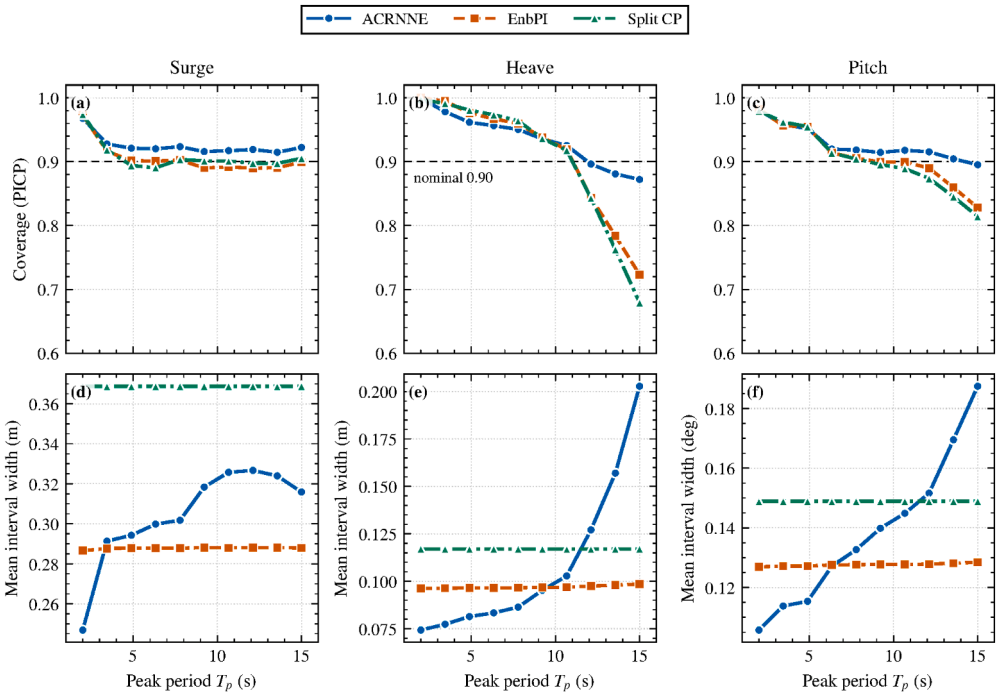


Fig. 5. Comparison of empirical coverage and mean interval width of the conformal methods as functions of peak period T_p.

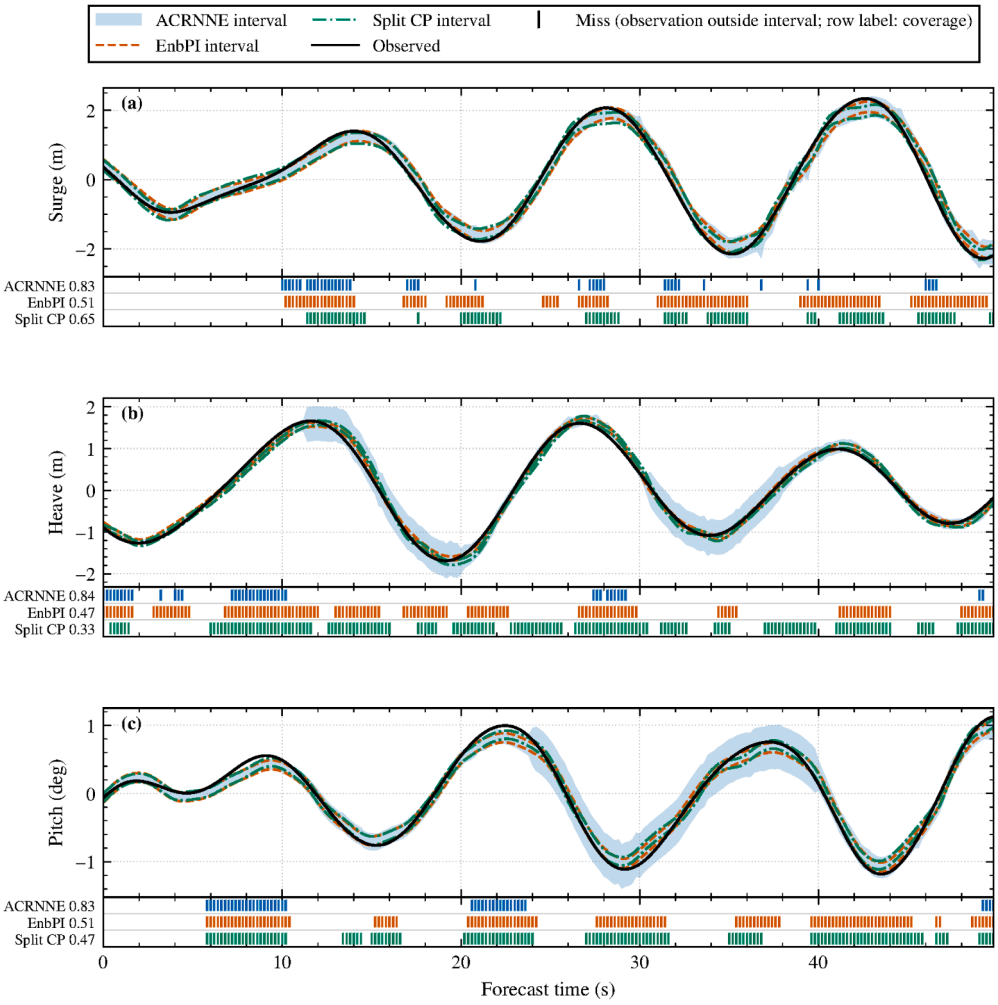


Fig. 6. Representative energetic test case (H_s = 6.02 m, T_p = 15 s): 2 s-ahead forecasts with the prediction intervals of ACRNNE, EnbPI and split CP.

peak-period range. Surge coverage changes little with T_p for all methods. Figs. 5(d)– 5(f) show the corresponding mean interval widths. Split CP and EnbPI produce widths that vary little with peak period. Consequently, their intervals are relatively wide in metocean conditions with shorter peak periods and become insufficiently wide in those with longer peak periods. For split CP, this pattern follows from its fixed calibration quantile. For EnbPI, the long residual window is only marginally refreshed during each test record, so its quantile and interval width change little.

The differences among the DoFs can be related to the platform dynamics. Fig. 5 shows that coverage is more sensitive to peak period for heave and pitch than for surge, particularly for EnbPI and split CP. A possible explanation is that heave and pitch have shorter natural periods [39], but closer to resonance with the upper bound of wave periods range considered. Surge, in contrast, exhibits a much longer natural period and its motion is dominated by wave drift. These differences in frequency response may contribute to variations in the forecast-error distributions across metocean conditions, which nearly constant interval widths do not adequately accommodate. ACRNNE adapts its intervals using observed forecast errors and achieves broadly comparable average coverage across the three DoFs (Table 6), although heave and pitch retain slightly lower coverage in the boundary metocean conditions (Table 5). This interpretation is qualitative; the individual contributions of hydrodynamic excitation, aerodynamic loading and controller coupling are not isolated in this study.

ACRNNE changes its interval width according to the recently observed forecast errors. Fig. 6 illustrates the forecasts and intervals for one representative energetic test case (H_s = 6.02 m, T_p = 15 s). EnbPI and split CP keep a nearly constant interval width throughout the record and therefore miss a large share of the observations in all three DoFs. These misses are not randomly scattered in time but occur in clusters, around the turning points of the oscillations and along the steep slopes between them, where the motion changes fastest. In contrast, ACRNNE responds to these changes after updating the residuals: its miscoverage level decreases and the interval widens so most subsequent observations fall inside it, whereas the interval contracts again once the forecasts succeed. As a result, ACRNNE covers most observations missed by both baselines in every DoF, widening its intervals mainly after observed misses. This behaviour illustrates the main advantage of ACRNNE over conformal intervals with a nearly constant width: the interval reacts to the forecast errors observed in the current record instead of reflecting only the average difficulty of the calibration data.

Two additional tests assess interval behaviour in severe boundary sea states and after abrupt changes in forecast difficulty. The first considers the most severe metocean conditions within the test grid, combining high significant wave heights with long peak periods. These sea states

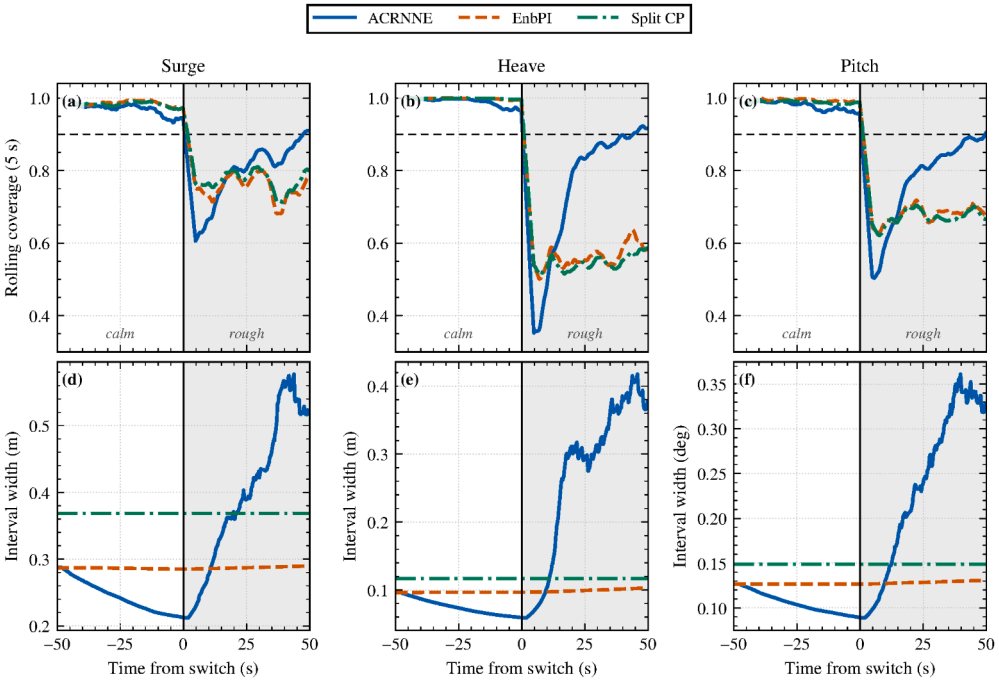


Fig. 7. Rolling coverage (top) and interval width (bottom) around an abrupt switch from a low-wave-height, short-period to a high-wave-height, long-period test case.

Table 5
Coverage in boundary metocean conditions, combining the two highest H_s and T_p levels of the grid.

	LSTM			GRU		
Method	Surge	Heave	Pitch	Surge	Heave	Pitch
ACRNNE	**0.861**	**0.841**	**0.847**	**0.861**	**0.842**	**0.850**
EnbPI	0.754	0.566	0.679	0.759	0.597	0.695
Split CP	0.769	0.548	0.666	0.766	0.578	0.680
Ensemble	0.552	0.634	0.474	0.607	0.661	0.495

lie at the boundary of the simulated parameter grid. As shown in Table 5, all reported methods remain below the nominal coverage in these conditions. Nevertheless, ACRNNE retains a substantially higher coverage than the alternatives, because it widens its intervals accordingly, whereas EnbPI and split CP fall well behind.

The second test is a chained error-regime stress test. A low-wave-height, short-period test case ($H_s \leq 3.41$ m, $4.89 \leq T_p \leq 6.33$ s) is concatenated with a high-wave-height, long-period test case ($H_s \geq 5.15$ m, $T_p \geq 13.56$ s). The relevant online state of each method, including ensemble weights, residual pools and the coverage level where applicable, is carried across the switch. Results are averaged over 100 such pairs per seed and five seeds. This is a hypothetical stress test of an abrupt change in forecast difficulty, not a physically simulated metocean transition. Each forecast continues to use the observed motion history of its own stationary case; consequently, the test does not represent a continuous environmental transition.

Fig. 7 shows that ACRNNE contracts its interval during the first phase and therefore initially has lower coverage than EnbPI and split CP after the switch. As delayed forecast outcomes become available, ACRNNE widens its interval and exceeds the evaluated baselines, which remain under-covered during the high-wave-height, long-period phase. It approaches the nominal level only towards the end of the 50 s window. The adaptive update progressively improves coverage after the switch and ultimately outperforms the other conformal methods, although its response is necessarily governed by the availability of the 2 s-ahead forecast outcomes.

RQ2. How do the coverage and interval efficiency of the ACRNNE compare to those generated by alternative uncertainty quantification approaches?

We now compare ACRNNE against established baseline methods: EnbPI, Ensemble, split Conformal Prediction, and Bootstrap. All metrics in Table 6 are averaged over the complete test set and five seeds, separately for each DoF and architecture. This comparison highlights fundamental differences in performance.

Table 6 distinguishes intervals based only on ensemble spread from those calibrated using observed forecast errors. Ensemble and Bootstrap intervals substantially under-cover: their empirical coverage ranges from 0.51 to 0.65 across surge, heave, and pitch. Thus, disagreement among the ensemble members alone does not adequately represent the total 2 s-ahead prediction error. The conformal methods instead calibrate the interval width on observed forecast residuals and consequently achieve empirical coverage close to the nominal level on average. These residuals reflect the combined effects of stochastic wave excitation, incomplete information in the observed motion history, and model training and approximation errors. Their individual contributions are not separately quantified in this study.

Since average coverage can conceal under-coverage in individual metocean conditions, Table 6 also reports the lowest group coverage, defined as the lowest PICP across the ten peak-period groups of the test set. On this measure, ACRNNE achieves the highest coverage among the benchmarked methods for every DoF and both architectures. Its average coverage is slightly conservative and its intervals are moderately wider than those of EnbPI. However, this additional width improves coverage in the peak-period groups where the other methods may under-cover. Compared with split CP, ACRNNE achieves higher lowest group coverage while maintaining lower point-forecast errors and narrower intervals in most DoF–architecture combinations. Its point-forecast accuracy is comparable to that of EnbPI, whereas its lowest group coverage is higher; the improvement therefore arises primarily from interval calibration and not from more accurate point predictions.

Although EnbPI obtains the highest CWC score, this does not contradict the coverage results. CWC rewards narrow intervals and penalises deviations of *average* coverage from the nominal level, including

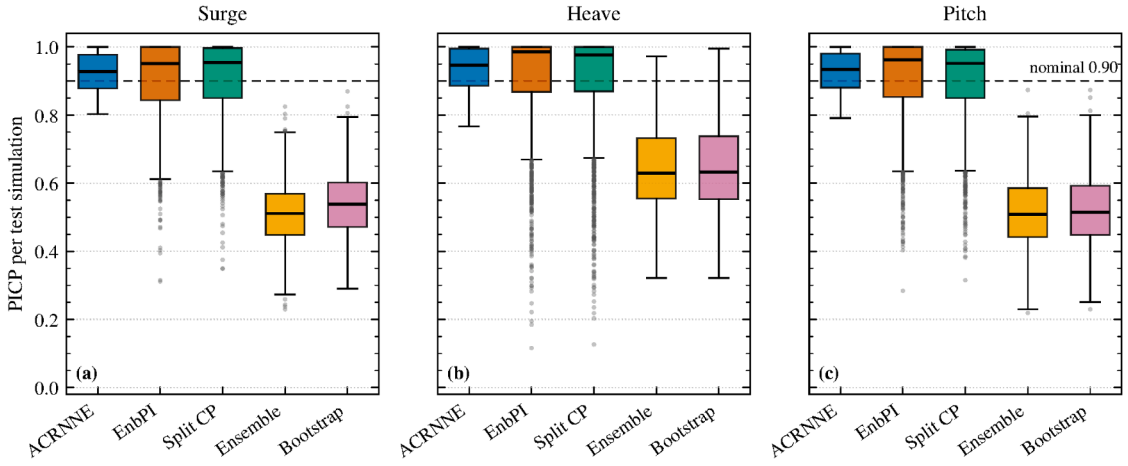


Fig. 8. Distribution of the coverage (PICP) across the 2000 test simulations for LSTM ensembles.

Table 6
Performance of the uncertainty quantification methods for 2 s-ahead forecasts (MAE in m for surge and heave and in deg for pitch).

		Surge					Heave					Pitch				
	Method	PICP	Lowest	PINAW	MAE	CWC	PICP	Lowest	PINAW	MAE	CWC	PICP	Lowest	PINAW	MAE	CWC
LSTM	ACRNNE	0.925	**0.913**	0.0498	0.0636	0.933	0.936	**0.872**	0.0211	0.0205	0.942	0.928	**0.895**	0.0387	**0.0279**	0.940
	EnbPI	0.906	0.886	**0.0471**	0.0648	**0.952**	0.911	0.723	**0.0188**	0.0203	**0.978**	0.909	0.828	**0.0356**	0.0281	**0.962**
	Split CP	0.908	0.885	0.0604	0.0817	0.936	0.905	0.679	0.0227	0.0257	0.974	0.903	0.814	0.0415	0.0332	0.955
	Ensemble	0.508	0.390	0.0191	**0.0634**	0.011	0.644	0.570	0.0093	**0.0200**	0.152	0.515	0.397	0.0146	0.0279	0.012
	Bootstrap	0.539	0.432	0.0211	0.0648	0.023	0.650	0.578	0.0096	0.0203	0.153	0.521	0.410	0.0149	0.0281	0.014
GRU	ACRNNE	0.928	**0.909**	0.0412	**0.0515**	0.937	0.933	**0.876**	0.0180	0.0179	0.951	0.928	**0.902**	0.0345	**0.0245**	0.943
	EnbPI	0.909	0.878	**0.0392**	0.0528	**0.958**	0.909	0.750	**0.0161**	**0.0178**	**0.982**	0.908	0.840	**0.0318**	0.0250	**0.966**
	Split CP	0.906	0.870	0.0410	0.0560	0.957	0.904	0.705	0.0202	0.0223	0.978	0.900	0.827	0.0353	0.0284	0.962
	Ensemble	0.603	0.515	0.0202	0.0534	0.077	0.644	0.569	0.0083	0.0178	0.149	0.535	0.454	0.0134	0.0247	0.020
	Bootstrap	0.617	0.554	0.0209	0.0528	0.095	0.646	0.570	0.0083	**0.0178**	0.149	0.544	0.448	0.0138	0.0250	0.023

conservative coverage. It does not account for how coverage varies across metocean conditions. Consequently, a method can achieve a high CWC score while under-covering in the peak-period groups where forecasting is more difficult. The lowest group coverage reported in Table 6 makes this behaviour visible. Paired Wilcoxon signed-rank tests compare the per-case absolute coverage errors, $|\text{PICP} - 0.90|$, of ACRNNE with those of EnbPI and split CP. ACRNNE has significantly lower coverage error in every DoF–architecture combination ($p < 10^{-6}$; Table A.9). This indicates that the improvement of ACRNNE is systematic across the test simulations (see also Fig. 8).

Beyond the average metrics, Fig. 8 shows the distribution of simulation-level coverage across the held-out test cases. The Ensemble and Bootstrap intervals under-cover in almost every simulation, confirming that model spread alone is insufficient to represent the total forecast error. EnbPI and split CP achieve high median coverage because their nearly constant interval widths fully cover many simulations with comparatively low forecast difficulty, which mainly correspond to metocean conditions with shorter peak periods. However, both distributions have pronounced lower tails. ACRNNE substantially reduces this lower tail and keeps the coverage of most test simulations closer to the nominal level. This behaviour follows its online update, which adjusts interval width after observed misses within each test record. Thus, ACRNNE provides more uniform empirical coverage across the held-out simulations. GRU ensembles show the same pattern.

Table 7 examines the contribution of the three adaptive mechanisms. The rows labelled W, P and L apply inverse-error weighting, residual-pool updating, and the online coverage-level update separately; the L row corresponds to an ACI-inspired update of the coverage level (Section 2.2). The final row combines all three mechanisms, which corresponds to the proposed ACRNNE. The results show that the mechanisms play different roles. Weighting alone (W) and residual-pool updating alone (P) leave the lowest group coverage well below the nominal level for heave and pitch. The online coverage-level update (L), however, substantially improves lowest group coverage for every DoF and both architectures. This improvement is accompanied by wider intervals, particularly for heave and pitch, because the method increases the interval width after observed misses.

The online coverage-level update accounts for most of the improvement in lowest group coverage. Compared with L alone, the complete ACRNNE framework reduces PINAW by approximately 1–3% across the evaluated degrees of freedom and architectures, while maintaining broadly similar average coverage. The lowest group coverage improves modestly for heave and pitch, and remains similar for surge. These results suggest that the weighting and residual-pool mechanisms provide additional refinements to interval efficiency. However, because W and P are added together in this comparison, the observed refinement cannot be attributed to either mechanism separately.

RQ3. How does the choice of RNN variant within the ensemble influence the resulting prediction intervals of ACRNNE?

To address this third RQ, we investigate the influence of the underlying model architecture on uncertainty quantification. The performance depends not only on the weighting mechanism but also on the base models. All metrics reported here are averaged over the complete test set and five seeds. Table 6 shows that GRU ensembles provide more accurate 2 s forecasts than LSTM ensembles for all three DoFs, with a lower MAE and the largest gain for surge, and correspondingly narrower intervals, with a lower PINAW. The smaller parameter count of the GRU may favour generalisation given the limited number of training windows, although the present comparison does not isolate this effect. GRU members also achieve lower mean validation losses across all three DoFs. The

Table 7
Ablation of the ACRNNE mechanisms for 2 s-ahead forecasts on the complete test set (2,000 simulations; means over five seeds).

		Surge				Heave				Pitch			
	Mechanisms	MAE	PICP	PINAW	Lowest	MAE	PICP	PINAW	Lowest	MAE	PICP	PINAW	Lowest
LSTM	W	**0.0636**	0.908	**0.0470**	0.887	0.0205	**0.909**	0.0187	0.726	**0.0279**	**0.907**	0.0355	0.830
	P	0.0648	**0.907**	0.0470	0.888	**0.0203**	0.912	**0.0187**	0.731	0.0281	0.910	**0.0355**	0.832
	L	0.0648	0.924	0.0509	0.910	**0.0203**	0.937	0.0217	0.862	0.0281	0.928	0.0391	0.891
	W + P + L (ACRNNE)	**0.0636**	0.925	0.0498	**0.913**	0.0205	0.936	0.0211	**0.872**	**0.0279**	0.928	0.0387	**0.895**
GRU	W	**0.0515**	**0.910**	**0.0390**	0.874	0.0179	**0.906**	**0.0160**	0.748	**0.0245**	**0.907**	**0.0317**	0.850
	P	0.0528	0.910	0.0391	0.880	**0.0178**	0.910	0.0160	0.756	0.0250	0.909	0.0318	0.843
	L	0.0528	0.927	0.0422	**0.911**	**0.0178**	0.934	0.0185	0.868	0.0250	0.927	0.0351	0.897
	W + P + L (ACRNNE)	**0.0515**	0.928	0.0412	0.909	0.0179	0.933	0.0180	**0.876**	**0.0245**	0.928	0.0345	**0.902**

W: inverse-error weighting; P: residual-pool update; L: online coverage-level update; W + P + L: complete ACRNNE.

coverage behaviour, however, is similar for the two architectures: ACRNNE achieves similar average and lowest group coverage with both architectures, and Table 7 shows the same pattern across the mechanisms for both. Hence, for the two evaluated architectures, the base learner mainly affects the accuracy and the interval width, and the similar coverage obtained with both suggests that the adaptive calibration is effective across the two evaluated base learners.

Taken together, the results provide a holistic picture, where ACRNNE improves coverage across groups of metocean conditions, particularly for high, long-period waves, for which split CP and EnbPI, with nearly constant interval widths, under-cover. This improvement originates mainly from the online coverage-level update and is paid for with wider intervals in those conditions.

5. Conclusions and future research directions

This study has addressed the challenge of quantifying uncertainty in the motion prediction of Floating Offshore Wind Turbines operating under diverse metocean conditions, in support of their structural integrity and safe, reliable operation. To this end, we have introduced an Adaptive Conformalized RNN Ensembles framework that integrates the distribution-free principles of conformal prediction with the modelling diversity offered by bootstrap ensembles. The framework integrates three adaptive mechanisms: (i) an inverse-error weighting mechanism that dynamically emphasises high-performing models; (ii) a weighted residual pool that continuously updates uncertainty estimates; and (iii) an online update of the coverage level, all operating under the delayed feedback of 2 s-ahead forecasting. Together, these mechanisms improve the uniformity of empirical coverage across the tested metocean conditions, including the conditions in which split conformal prediction and EnbPI, whose widths change little in the present setting, under-cover. Nevertheless, empirical coverage in the boundary metocean conditions remains below the nominal 90% level, ranging from 84.1% to 86.1% across the evaluated degrees of freedom and architectures. The improvement therefore represents a reduction in under-coverage, not its elimination.

The experimental campaign conducted in this study has provided clear evidence of the effectiveness and robustness of the proposed ACRNNE framework. The framework's evaluation has yielded several key conclusions:

- *Coverage and interval efficiency:* The conformal methods achieve average coverage at or above the nominal level, whereas intervals based solely on ensemble or bootstrap member spread substantially under-cover. ACRNNE achieves higher lowest group coverage than EnbPI and split conformal prediction across all evaluated degrees of freedom and architectures. Its intervals are narrower than those of split conformal prediction in most configurations, but wider than those of EnbPI. EnbPI achieves the highest CWC scores, illustrating the trade-off between interval width, average calibration and coverage across metocean groups.
- *Robustness across diverse conditions:* Unlike the conformal baselines with nearly constant interval widths, which under-cover the target on a large fraction of cases, particularly in high-wave-height, long-period conditions, ACRNNE reduces under-coverage across metocean conditions and test simulations by increasing interval width where recent forecasts miss, at the cost of moderately wider intervals. The ablation study attributes most of this gain to the online coverage-level update.
- *Effect of the base learner:* GRU ensembles are more accurate than LSTM ensembles (lower MAE) for all three degrees of freedom and produce narrower intervals (lower PINAW), whereas the coverage behaviour is similar for both, suggesting that the adaptive calibration is effective across the two evaluated base learners.
- *Multi-DoF validation:* The framework improves empirical coverage relative to the evaluated baselines across multiple degrees of freedom (surge, heave, pitch). Heave and pitch are the most demanding for high, long-period waves, consistent with the hydrodynamic response of the platform, and the chained regime-transition stress test shows that coverage improves progressively after an abrupt increase in forecast difficulty.

This work carries broader implications beyond structural safety of FOWTs. Improving empirical coverage while continually updating the intervals from observed forecast outcomes addresses a fundamental challenge in time-series uncertainty quantification. As machine learning expands into safety-critical domains with shifting environments (from autonomous vehicles to power grid management), the proposed ACRNNE offers a practical approach for improving empirical interval calibration in autocorrelated and time-varying settings.

Future Research. Several limitations of this study motivate future work. First, the evaluation has relied on validated mid-fidelity numerical simulations, which do not reproduce all the physical phenomena and uncertainties under real conditions (including complex hydrodynamic and mooring effects, directional spreading and misaligned environmental loading). Measured responses are further affected by additional sources of uncertainty, such as sensor noise and drift, turbulent and misaligned wind and wave conditions with directional spreading, marine growth and mooring degradation. Assessing whether ACRNNE maintains near-nominal coverage on such measured responses is a necessary next step before operational deployment. Moreover, the wind is steady and the metocean conditions are stationary within each simulation, and conditions outside the simulated range are not tested; the train/test split is made at the level of simulation cases, so neighbouring metocean combinations of the full-factorial grid may appear in both subsets. The evaluation therefore assesses performance within the simulated design space, not extrapolation beyond the simulated metocean range. The regime-transition analysis utilises two separate simulations and therefore may not represent actual change in metocean conditions during operation. The comparisons establish performance relative to the selected baseline configurations. Future research will extend the evaluation to several longer prediction horizons and compare ACRNNE with other adaptive conformal methods, such as ACI. As a natural extension of the present work, several wave realisations per metocean condition will be considered to characterise the variability of the data and how it propa-

gates to the uncertainty at the model output, and to separate aleatory from epistemic contributions. In addition, uncertainty attribution methods could identify which model inputs contribute most to the width of the quantified intervals. Further promising avenues for extending this work include: i) integrating adaptive prediction intervals directly into Model Predictive Control (MPC) frameworks to enable more robust load mitigation; ii) exploring mixed-architecture ensembles that automatically select LSTM or GRU members based on the signal dynamics of the motion variable; iii) adapting the framework to other marine platforms with different dynamics, such as wave energy converters or multipurpose platforms; iv) exploring faster adaptation mechanisms, such as condition-dependent coverage levels; and v) expanding the approach to handle other complex processes with changing dynamical conditions.

CRediT authorship contribution statement

Pranjal Tamuly: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization; **Javier Del Ser:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology; **Hamed Fathnejat:** Writing – review & editing, Writing – original draft, Visualization, Data curation, Investigation, Software; **Vincenzo Nava:** Writing – review & editing, Writing – original draft, Supervision, Project administration.

Declaration of competing interest

There are no relevant financial or non-financial competing interests to report.

Acknowledgments

This work is supported by the Italian Ministry of University and Research (MUR) under the National Recovery and Resilience Plan (PNRR), through the funding scheme for projects presented by young researchers (project code SOE2024_0000165, CUP J87G25000140007). Javier Del Ser acknowledges funding support from the Basque Government through ELKARTEK funds (ref. KK-2026/00050) and the consolidated research group MATHMODE (ref. 1866-26). Javier Del Ser and Vincenzo Nava have received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101342459 (MOST-PINN).

Data availability

The simulated dataset used in this study is openly available in Zenodo at https://doi.org/10.5281/zenodo.23529260.

Appendix A. Confidence intervals and paired statistical tests

This appendix examines the sensitivity of the results to training randomness. Table A.8 reports the mean performance metrics and their 95% confidence intervals across the five training seeds. The seeds differ in network initialisation, bootstrap resampling, training-window positions, training-data shuffling, dropout realisations, and the split-conformal calibration subset, while the case-level training, validation and test split remains fixed.

The confidence intervals therefore quantify variation across training runs on the same held-out test set. They do not represent sampling variability across new test sets or repeated wave realisations. Table A.9 complements these results by reporting the paired differences between ACRNNE and the two conformal baselines for each test case.

Table A.8 shows that the empirical coverage of ACRNNE remains stable across the five training seeds. Its lowest group coverage is higher than that of EnbPI and split CP for all three degrees of freedom and for both LSTM and GRU ensembles. Split CP shows wider confidence intervals, particularly for lowest group coverage, indicating greater variation across training seeds.

The paired comparisons in Table A.9 show lower absolute coverage error for ACRNNE than for EnbPI and split CP in every DoF–architecture combination. In contrast, the paired MAE differences between ACRNNE and EnbPI are small. Together, these results indicate that the principal benefit of ACRNNE is more reliable interval coverage across peak-period groups.

Table A.8
Mean performance metrics and 95% confidence intervals ($mean_{int}$) across the five training seeds on the fixed test split. Values are reported as the mean with the confidence-interval half-width as a subscript. Lowest denotes the minimum PICP across the ten peak-period (T_p) groups. MAE is reported in m for surge and heave, and in deg for pitch.

		Surge				Heave				Pitch			
	Method	PICP	Lowest	PINAW	MAE	PICP	Lowest	PINAW	MAE	PICP	Lowest	PINAW	MAE
LSTM	ACRNNE	$0.925_{0.001}$	$0.913_{0.002}$	$0.0498_{0.0040}$	$0.0636_{0.0027}$	$0.936_{0.001}$	$0.872_{0.003}$	$0.0211_{0.0008}$	$0.0205_{0.0003}$	$0.928_{0.002}$	$0.895_{0.002}$	$0.0387_{0.0013}$	$0.0279_{0.0005}$
	EnbPI	$0.906_{0.005}$	$0.886_{0.005}$	$0.0471_{0.0022}$	$0.0648_{0.0035}$	$0.911_{0.004}$	$0.723_{0.017}$	$0.0188_{0.0004}$	$0.0203_{0.0005}$	$0.909_{0.006}$	$0.828_{0.006}$	$0.0356_{0.0015}$	$0.0281_{0.0006}$
	Split CP	$0.908_{0.011}$	$0.885_{0.011}$	$0.0604_{0.0093}$	$0.0817_{0.0136}$	$0.905_{0.012}$	$0.679_{0.066}$	$0.0227_{0.0040}$	$0.0257_{0.0048}$	$0.903_{0.014}$	$0.814_{0.019}$	$0.0415_{0.0079}$	$0.0332_{0.0050}$
	Ensemble	$0.508_{0.031}$	$0.390_{0.032}$	$0.0191_{0.0010}$	$0.0634_{0.0007}$	$0.644_{0.044}$	$0.570_{0.040}$	$0.0093_{0.0007}$	$0.0200_{0.0009}$	$0.515_{0.017}$	$0.397_{0.025}$	$0.0146_{0.0003}$	$0.0279_{0.0007}$
	Bootstrap	$0.539_{0.037}$	$0.432_{0.027}$	$0.0211_{0.0004}$	$0.0648_{0.0035}$	$0.650_{0.010}$	$0.578_{0.015}$	$0.0096_{0.0003}$	$0.0203_{0.0005}$	$0.521_{0.015}$	$0.410_{0.018}$	$0.0149_{0.0006}$	$0.0281_{0.0006}$
GRU	ACRNNE	$0.928_{0.002}$	$0.909_{0.004}$	$0.0412_{0.0025}$	$0.0515_{0.0027}$	$0.933_{0.001}$	$0.876_{0.002}$	$0.0180_{0.0007}$	$0.0179_{0.0005}$	$0.928_{0.003}$	$0.902_{0.003}$	$0.0345_{0.0014}$	$0.0245_{0.0006}$
	EnbPI	$0.909_{0.005}$	$0.878_{0.006}$	$0.0392_{0.0023}$	$0.0528_{0.0036}$	$0.909_{0.003}$	$0.750_{0.010}$	$0.0161_{0.0007}$	$0.0178_{0.0006}$	$0.908_{0.007}$	$0.840_{0.006}$	$0.0318_{0.0017}$	$0.0250_{0.0009}$
	Split CP	$0.906_{0.010}$	$0.870_{0.018}$	$0.0410_{0.0083}$	$0.0560_{0.0125}$	$0.904_{0.007}$	$0.705_{0.085}$	$0.0202_{0.0035}$	$0.0223_{0.0025}$	$0.900_{0.013}$	$0.827_{0.047}$	$0.0353_{0.0049}$	$0.0284_{0.0052}$
	Ensemble	$0.603_{0.043}$	$0.515_{0.052}$	$0.0202_{0.0015}$	$0.0534_{0.0030}$	$0.644_{0.040}$	$0.569_{0.058}$	$0.0083_{0.0004}$	$0.0178_{0.0005}$	$0.535_{0.028}$	$0.454_{0.045}$	$0.0134_{0.0007}$	$0.0247_{0.0010}$
	Bootstrap	$0.617_{0.035}$	$0.554_{0.048}$	$0.0209_{0.0013}$	$0.0528_{0.0036}$	$0.646_{0.031}$	$0.570_{0.041}$	$0.0083_{0.0004}$	$0.0178_{0.0006}$	$0.544_{0.016}$	$0.448_{0.039}$	$0.0138_{0.0004}$	$0.0250_{0.0009}$

Table A.9
Mean paired differences between ACRNNE and each baseline (ACRNNE minus baseline) in absolute coverage error, interval width, and MAE. Coverage error is defined as |PICP – 0.90|. Paired Wilcoxon signed-rank tests confirm lower coverage error for ACRNNE in every comparison ($p < 10^{-6}$).

		Surge			Heave			Pitch		
	vs.	Cov. error	Width	MAE	Cov. error	Width	MAE	Cov. error	Width	MAE
LSTM	EnbPI	-0.040	+0.0167	-0.0012	-0.050	+0.0118	+0.0002	-0.041	+0.0112	-0.0002
	Split CP	-0.038	-0.0643	-0.0182	-0.051	-0.0082	-0.0052	-0.039	-0.0101	-0.0053
GRU	EnbPI	-0.040	+0.0121	-0.0014	-0.048	+0.0099	+0.0001	-0.040	+0.0094	-0.0004
	Split CP	-0.038	+0.0017	-0.0046	-0.049	-0.0112	-0.0045	-0.040	-0.0030	-0.0038

References

[1] WindEurope, Wind Energy in Europe: 2024 Statistics and the Outlook for 2025–2030, Technical Report, WindEurope, 2025. Published in February 2025, https://windeurope.org/data/products/wind-energy-in-europe-2024-statistics-and-the-outlook-for-2025-2030/.

[2] International Renewable Energy Agency (IRENA), Renewable Energy Statistics 2024, International Renewable Energy Agency, Abu Dhabi, 2024.

[3] S. Deng, D. Ning, R. Mayon, The motion forecasting study of floating offshore wind turbine using self-attention long short-term memory method, Ocean Eng. 310 (2024) 118709.

[4] N. Li, G. Zou, Y. Feng, L. Ali, Probabilistic prediction of floating offshore wind turbine platform motions via uncertainty quantification and information integration, J. Mar. Sci. Eng. 12 (6) (2024) 886.

[5] C. Chen, Y. Ma, T. Fan, Review of model experimental methods focusing on aerodynamic simulation of floating offshore wind turbines, Renew. Sustain. Energy Rev. 157 (2022) 112036.

[6] L.M. Huang, W.Y. Duan, Y. Han, D.H. Yu, A. Elhandad, Extending the scope of AR model in forecasting non-stationary ship motion by using AR-EMD technique, J. Sh. Mech. 19 (9) (2015) 1033–1049.

[7] B.J. Jin, B.H. Zhang, C.H. Deng, J.L. Wu, Wavelet-ARMA model revised by neural network to predict wind power, Adv. Mater. Res. 724 (2013) 669–674.

[8] R.G. Kavasseri, K. Seetharaman, Day-ahead wind speed forecasting using f-ARIMA models, Renew. Energy 34 (5) (2009) 1388–1393.

[9] M. Rezaee, R. Fathi, V. Jahangiri, M.M. Ettefagh, A. Jamalkia, M.H. Sadeghi, Detection of damages in mooring lines of spar type floating offshore wind turbines using fuzzy classification and ARMA parametric modeling, Int. J. Struct. Stab. Dyn. 21 (08) (2021) 2150111.

[10] H. Fathnejat, B. Ahmadi-Nedushan, S. Hosseininejad, M. Noori, W.A. Altabey, A data-driven structural damage identification approach using deep convolutional-attention-recurrent neural architecture under temperature variations, Eng. Struct. 276 (2023) 115311.

[11] W. Shi, L. Hu, Z. Lin, L. Zhang, J. Wu, W. Chai, Short-term motion prediction of floating offshore wind turbine based on muti-input LSTM neural network, Ocean Eng. 280 (2023) 114558.

[12] W. Zhang, Z. Lin, X. Liu, Short-term offshore wind power forecasting–a hybrid model based on discrete wavelet transform (DWT), seasonal autoregressive integrated moving average (SARIMA), and deep-learning-based long short-term memory (LSTM), Renew. Energy 185 (2022) 611–628.

[13] S. Hanifi, H. Zare-Behtash, A. Cammarano, S. Lotfian, Offshore wind power forecasting based on WPD and optimised deep learning methods, Renew. Energy 218 (2023) 119241.

[14] B. Song, Q. Zhou, R. Chang, Short-term motion prediction of FOWT based on time-frequency feature fusion LSTM combined with signal decomposition methods, Ocean Eng. 317 (2025) 120046.

[15] J. Yin, J. Ding, Y. Yang, J. Yu, L. Ma, W. Xie, D. Nie, M. Bashir, Q. Liu, C. Li, et al., Wave-induced motion prediction of a deepwater floating offshore wind turbine platform based on Bi-LSTM, Ocean Eng. 315 (2025) 119836.

[16] H. Fathnejat, V. Nava, Data augmentation for damaged scenarios in floating offshore wind turbines: an approach based on diffusion architecture, hierarchical variational approximation and healthy data distribution, Eng. Comput. (2025) 1–21.

[17] P. Tamuly, S. Sharma, V. Nava, Integrated damage detection and time-series data augmentation for floating offshore mooring systems via variational semi-supervised learning, Ocean Eng. 330 (2025) 121199.

[18] H. Fathnejat, V. Nava, From augmentation to translation: data generation by conditional hierarchical variational autoencoder, enhancing monitoring mooring systems in floating offshore wind turbines, Eng. Appl. Artif. Intell. 163 (2026) 112951.

[19] P. Tamuly, V. Nava, Damage detection in mooring systems of floating offshore wind turbines using semi-supervised GAN with image-transformed limited labelled data, Adv. Struct. Eng. (2025) 13694332251391469.

[20] Y. Shao, J. Liu, Uncertainty quantification for dynamic responses of offshore wind turbine based on manifold learning, Renew. Energy 222 (2024) 119875.

[21] A. Fernandez-Navamuel, N. Gorostidi, D. Pardo, V. Nava, E. Chatzi, Gaussian mixture autoencoder for uncertainty-aware damage identification in a floating offshore wind turbine, Wind Energy Sci. 10 (5) (2025) 857–885.

[22] X. Guo, P. Yi, J. Liu, A novel efficient RFEM for reliability analysis and design of multi-line dynamically installed anchor for floating offshore wind turbines, Struct. Saf. 111 (2024) 102520. https://doi.org/10.1016/j.strusafe.2024.102520

[23] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, MIT Press, 2006.

[24] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016.

[25] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Adv. Neural Inf. Process. Syst. 30 (2017).

[26] N. Tagasovska, D. Lopez-Paz, Single-model uncertainties for deep learning, Adv. Neural Inf. Process. Syst. 32 (2019).

[27] Y. Romano, E. Patterson, E. Candes, Conformalized quantile regression, Adv. Neural Inf. Process. Syst. 32 (2019).

[28] M. Zaffran, O. Feron, Y. Goude, J. Josse, A. Dieuleveut, Adaptive conformal predictions for time series, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 25834–25866.

[29] A. Bhatnagar, H. Wang, C. Xiong, Y. Bai, Improved online conformal prediction via strongly adaptive online learning, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 2337–2363.

[30] C. Xu, Y. Xie, Conformal prediction interval for dynamic time-series, in: International Conference on Machine Learning, PMLR, 2021, pp. 11559–11569.

[31] I. Gibbs, E. Candès, Adaptive conformal inference under distribution shift, in: Advances in Neural Information Processing Systems, 34, 2021, pp. 1660–1672.

[32] V. Jensen, F.M. Bianchi, S.N. Anfinsen, Ensemble conformalized quantile regression for probabilistic time series forecasting, IEEE Trans. Neural Netw. Learn. Syst. 35 (7) (2024) 9014–9025.

[33] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, 29, Springer, 2005.

[34] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (3) (2008).

[35] J. Chen, C. Jin, M.-H. Kim, Systematic comparisons among OpenFAST, Charm3D-FAST simulations and deepCWind model test for 5 MW OC4 semisubmersible offshore wind turbine, Ocean Syst. Eng. 13 (2) (2023) 173–193. https://doi.org/10.12989/ose.2023.13.2.173

[36] A.S. Hsieh, K.A. Brown, N.B. DeVelder, T.G. Herges, R.C. Knaus, P.J. Sakievich, L.C. Cheung, B.C. Houchens, M.L. Blaylock, D.C. Maniaci, High-fidelity wind farm simulation methodology with experimental validation, J. Wind Eng. Ind. Aerodyn. 218 (2021) 104754.

[37] N. Mendoza, et al., Verification and validation of model-scale turbine performance and control strategies for the IEA wind 15 MW reference wind turbine, Energies 15 (20) (2022) 7649.

[38] J. Jonkman, S. Butterfield, W. Musial, G. Scott, Definition of a 5-MW Reference Wind Turbine for Offshore System Development, Technical Report, National Renewable Energy Laboratory, United States, 2009. https://doi.org/10.2172/947422

[39] A. Robertson, J. Jonkman, M. Masciola, H. Song, A. Goupee, A. Coulling, C. Luan, Definition of the semisubmersible floating system for phase II of OC4, Technical Report, National Renewable Energy Lab (NREL), United States, 2014.

[40] H. Jiang, H. Wang, M.A. Vaz, X. Bai, Research on dynamic response prediction of semi-submersible wind turbine platform in real sea test model based on machine learning, Appl. Ocean Res. 142 (2024) 103808.

[41] S. Sharma, V. Nava, Condition monitoring of mooring systems for floating offshore wind turbines using convolutional neural network framework coupled with autoregressive coefficients, Ocean Eng. 302 (2024) 117650.

[42] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[43] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, (2014). arXiv:1409.1259

[44] H. Guo, Y. Dong, E. Bastidas-Arteaga, X. Lei, Life-cycle performance prediction and interpretation for coastal and marine RC structures: an ensemble learning framework, Struct. Saf. 110 (2024) 102496. https://doi.org/10.1016/j.strusafe.2024.102496

[45] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap.(1st ed.), Chapman and Hall/CRC, New York, 1994. https://doi.org/10.1201/9780429246593

[46] A. Modak, S. Chakraborty, An enhanced learning function for bootstrap polynomial chaos expansion-based enhanced active learning algorithm for reliability analysis of structure, Struct. Saf. 109 (2024) 102467. https://doi.org/10.1016/j.strusafe.2024.102467

[47] A. Khosravi, S. Nahavandi, D. Creighton, A.F. Atiya, Comprehensive review of neural network-Based prediction intervals and new advances, IEEE Trans. Neural Netw. 22 (9) (2011) 1341–1356. https://doi.org/10.1109/TNN.2011.2162110