https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Probabilistic forecasting of wind turbine remaining useful life using conformalised quantile regression

Ali Eftekhari Milani[1], Donatella Zappalá[1], Shawn Sheng[2], and Simon Watson[1]

[1]Delft University of Technology, Kluyverweg 1, Delft, 2629 HS, Netherlands
[2]National Wind Technology Center, National Laboratory of the Rockies, Golden, USA

Correspondence: Ali Eftekhari Milani (a.eftekharimilani@tudelft.nl)

Abstract. In recent years, numerous machine learning methods have been developed to predict the remaining useful life (RUL) of wind turbine components. However, uncertainties in modelling the future progression of degradation often preclude accurate point forecasts of failure times. Quantifying this uncertainty is therefore crucial to ensuring reliable predictions as it empowers operators to make risk-informed maintenance decisions. This work proposes a probabilistic RUL forecasting framework that leverages a convolutional autoencoder (CAE) to extract health indicators (HIs) from supervisory control and data acquisition (SCADA) signals, accurately capturing component degradation over time. To facilitate HI extraction, a Convolutional Neural Network-based normal behaviour modelling framework is employed as a feature extractor, and residuals of component temperature signals, rather than the raw signals, are supplied to the CAE. These HIs are then fed into a Long Short-Term Memory-based conformalised quantile regression framework to probabilistically predict RUL, calibrating confidence intervals to reliably represent uncertainty. This proposed approach effectively models degradation while alleviating the impact of high noise in field data. Its application to two case studies demonstrates that, while achieving similar performance to existing approaches using the simulated Commercial Modular Aero-Propulsion System Simulation (CMAPSS) dataset, the proposed approach significantly outperforms when using a real SCADA dataset with gearbox failures, reducing point prediction errors by approximately 67%. Furthermore, the generated prediction intervals are better calibrated and, on average, 42% shorter, providing more informative and reliable uncertainty estimates.

1 Introduction

Predicting the remaining useful life (RUL) of wind turbine components is crucial for transitioning from reactive and schedule-based maintenance strategies to predictive maintenance, which contributes to operational efficiency and reduces maintenance costs in wind farms. Accurate RUL predictions allow operators to schedule maintenance proactively, optimise turbine performance, and minimise unexpected failures (Li et al. (2026)).

Accurately predicting RUL over extended periods is challenging due to the complex nature of degradation processes and inherent randomness in the future operational and environmental conditions. This complexity leads to uncertainties in the prediction of RUL, which is classically categorised into two distinct types of uncertainty: aleatoric and epistemic. Aleatoric uncertainty refers to the inherent, irreducible randomness of a phenomenon. In the context of wind turbines, this stems from

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

stochastic external factors, such as variable wind conditions, complex load fluctuations, and natural sensor noise. Epistemic uncertainty, conversely, arises from a lack of knowledge and is theoretically reducible with more information. This uncertainty originates from the model itself, caused by limited understanding of component degradation mechanisms, simplifying model assumptions, or insufficient historical training data (McMorland et al. (2022); Sankararaman and Goebel (2020)). The combination of these factors makes precise point forecasts unreliable. Consequently, effective RUL prediction methods must move beyond simple regression to provide confidence intervals that rigorously quantify these uncertainties. This ensures well-calibrated predictions that neither underestimate risk nor overestimate safety.

Modern wind turbines are equipped with supervisory control and data acquisition (SCADA) systems, which continuously monitor a wide range of operational parameters (Chesterman et al. (2023)). In contrast, other data sources, such as vibration measurements, are less common and are typically available only for shorter durations (Casolo et al. (2024)). As a result, SCADA data often serve as the primary input for RUL prediction models, enabling tracking of component degradation throughout the turbine's lifespan and ensuring broad applicability. However, SCADA datasets typically provide low-resolution summary statistics at 10-minute intervals, and are prone to significant noise (Chatterjee and Dethlefs (2021)). The fluctuating operational and environmental conditions captured in SCADA data further complicate the development of effective RUL prediction methods.

Methods for probabilistic RUL predictions using SCADA data are relatively scarce in the literature. In Peter et al. (2022), an anomaly index is constructed from one-class support vector machine (SVM)-generated alarms, and its future trajectory is forecasted using an autoregressive moving average (ARIMA) model (Harvey (1990)) to predict wind turbine generator failure. However, in this work, RUL is predicted only at a few specific instances near failure, rather than providing a continuous prediction, with performance at longer prediction horizons not assessed. Furthermore, the accuracy of the ARIMA-based confidence intervals is not assessed. In (Zhang et al. (2025)), a Light Gradient Boosting Machine is used to model normal behaviour by generating residuals reflecting gearbox pump degradation, which are then used in a gated recurrent unit (GRU)-based (Chung et al. (2014)) Bayesian neural network for probabilistic RUL predictions. However, the 97.5% confidence intervals predicted by this approach significantly underestimate the actual uncertainty, as their realised coverage, i.e. the ratio of the true RUL values within the predicted RUL intervals, is only around 73% in the test set. In more recent work (De Florio et al. (2025)), a physics-informed machine learning (PIML) framework for probabilistic RUL prediction of bearings using vibration and SCADA data is developed. By embedding fracture mechanics priors, this framework constrains the solution space and produces physically consistent degradation trajectories, with Monte Carlo ensembles used to estimate epistemic uncertainty. However, the realised coverage of the confidence intervals is not assessed.

Conformal prediction has recently gained traction as a framework for generating distribution-free confidence intervals with formal coverage guarantees (Lei et al. (2018)). This characteristic is vital for safety-critical systems, as it eliminates the reliance on rigid parametric assumptions (e.g. Gaussian errors), which often fail in real-world degradation processes, thereby ensuring that the predicted uncertainty bounds remain valid and trustworthy (Angelopoulos and Bates (2023)). Several studies have applied this framework for RUL prediction with uncertainty quantification. For example, (Javanmardi and Hüllermeier (2023)) compare the performance of various conformal prediction approaches on the Commercial Modular Aero-Propulsion System

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Simulation (CMAPSS) dataset (Saxena et al. (2008)), a simulated turbofan engine dataset, showing that a deep convolutional neural network (CNN)-based (O'Shea and Nash (2015a)) conformalised quantile regression (CQR) (Romano et al. (2019)) approach outperforms others in both accuracy and prediction interval coverage. In (Mao et al. (2024)) and (Piao et al. (2025)), CQR-based methods using multi-scale CNNs and Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber (1997)) networks are adopted to probabilistically predict RUL in the CMAPSS and N-CMAPSS (Arias Chao et al. (2021)) datasets. In (Noot et al. (2025)), LSTM- and transformer-based (Vaswani et al. (2017)) models are used alongside split conformal prediction for confidence interval estimation, whereas (Wang et al. (2025)) proposes a split conformal prediction approach for RUL prediction under missing sensor data, combining latent representation learning and diffusion-based data imputation. Despite achieving satisfactory results on simulated datasets, these methods have not been tested on real wind turbine data. The high noise levels and widely varying operational and environmental conditions in SCADA data make it challenging to replicate the same performance. To the best of the authors' knowledge, conformal prediction has not yet been applied to wind turbine RUL prediction.

To address these limitations, this work proposes a comprehensive probabilistic framework for wind turbine RUL prediction. An unsupervised convolutional autoencoder (CAE) (Eftekhari Milani et al. (2024)) is employed to construct health indicators (HIs) that accurately model component degradation by effectively decoupling the underlying degradation trend from operational and environmental factors and sensor noise. The constructed HI is then post-processed to further suppress noise. As an additional noise-reduction measure, a CNN-based normal behaviour modelling framework is employed as a feature extractor to facilitate HI construction. The HIs are then used in an LSTM-based CQR framework to generate accurate probabilistic RUL predictions with well-calibrated confidence intervals.

The rest of this paper is organised as follows: Section 2 describes the methodology for HI construction and RUL prediction. Section 3 applies the method to the widely used CMAPSS dataset, enabling comparison with existing approaches. Section 4 demonstrates the performance of the framework in a field SCADA dataset featuring multiple gearbox bearing failures and compares it with a benchmark approach originally developed for CMAPSS. Finally, conclusions are drawn, and potential directions for future work are discussed in Section 5.

2 Methodology

A run-to-failure sensor dataset can be represented as a discrete multivariate time series $\boldsymbol{S} = \{s_{i,t}\}$ where $i = 1, \ldots, C$, with C being the number of SCADA channels, and $t = 1, \ldots, T$, with t being the time instance and T corresponding to the observed failure time. The RUL, R, is then defined as:

$$R(t) = T - t \tag{1}$$

The objective of this work is to develop a method that, utilising sensor signals observed up to time t_k, generates a probabilistic prediction of $R(t_k)$. Specifically, the model outputs a predicted median and a confidence interval calibrated to a target reliability. This reliability is formally quantified by the expected coverage probability $1 - \alpha$, which means that the true

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

RUL is contained within the predicted interval with a probability of $1-\alpha$, where α represents the target miscoverage rate. The construction of this interval and the median is achieved through predicting three quantiles $\hat{q}_{\alpha/2}[R(t_k)]$, $\hat{q}_{0.5}[R(t_k)]$, and $\hat{q}_{1-\alpha/2}[R(t_k)]$. $\hat{q}_{\alpha/2}[R]$ represents the predicted $\alpha/2$-quantile of R, i.e., a value below which the true value of R is expected to fall with probability $\alpha/2$. Hence, $\hat{q}_{\alpha/2}[R(t_k)]$ and $\hat{q}_{1-\alpha/2}[R(t_k)]$ represent the lower and upper intervals providing $1-\alpha$ confidence, and $\hat{q}_{0.5}[R(t_k)]$ the predicted median.

To ensure reliability, the realised coverage of these predicted intervals, defined as the proportion of true RUL values falling within the predicted intervals, should match the defined nominal confidence level of $1-\alpha$. The condition below must therefore be validated on a dedicated test set:

$$\mathbb{P}\{\hat{q}_{\alpha/2}[R(t)] \leq R(t) \leq \hat{q}_{1-\alpha/2}[R(t)]\} \geq 1-\alpha \tag{2}$$

where $\mathbb{P}$ is the probability over all time instances in the test set. If the realised coverage is substantially higher than $1-\alpha$, the intervals are too wide, and the model is over-conservative, sacrificing precision for unnecessary safety. Conversely, if the realised coverage is lower than $1-\alpha$, the intervals are too narrow, indicating over-confidence, which can compromise decision-making. Therefore, predicted intervals should be well-calibrated, neither underestimating nor overestimating uncertainty. The realised coverage of the intervals should be as close to the nominal $1-\alpha$ as possible.

Conformal prediction provides a framework that guarantees the coverage validity condition in Eq. 2 for finite samples, without assuming any specific distribution (Romano et al. (2019)). This characteristic is vital for safety–critical systems, as such parametric assumptions often fail in real-world degradation processes, compromising the validity of the predicted intervals (Angelopoulos and Bates (2023)). Standard conformal approaches, such as naive conformal prediction (Vovk et al. (2022)), often yield intervals of constant or weakly varying lengths, which can lead to over-conservatism. To address this limitation, CQR is adopted in this work (Romano et al. (2019)), allowing the construction of prediction intervals that are both flexible and well-calibrated.

Directly predicting R from $\boldsymbol{S}$ is challenging due to the high noise level in SCADA signals and the large variability in the operational modes of wind turbines. To address this, the proposed method adopts a two-step approach, shown in Figure 1:

1. An unsupervised model based on a CAE constructs an HI, denoted by H, given $\boldsymbol{S}$. This step models the conditional distribution $\mathbb{P}(H|\boldsymbol{S})$. H is represented as a univariate time series defined in the range 0 to 1, where 0 represents the initial pristine state (0% degradation) and 1 the final failed state (100% degradation).

2. A CQR model, implemented with an LSTM neural network, predicts the three quantiles of R given the HI constructed in step 1. This step models the conditional distribution $\mathbb{P}(R|H)$.

2.1 Unsupervised HI construction using CAE

The component HIs are derived using the unsupervised method proposed in Eftekhari Milani et al. (2024), which has been shown to capture degradation trends more effectively than other approaches in the literature. The method employs a CAE trained via a hybrid optimisation approach combining particle swarm optimisation (PSO) and backpropagation. The goal is to

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

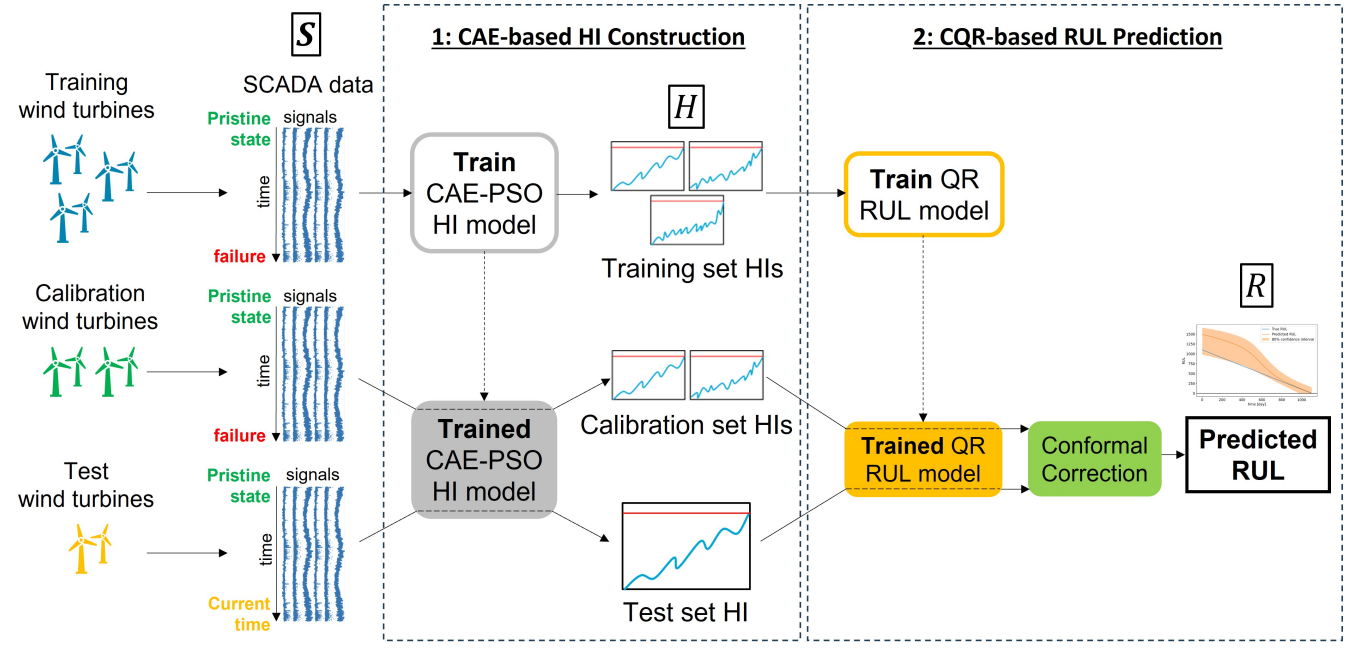

Figure 1. Flowchart of the proposed two-step method for probabilistic RUL prediction.

simultaneously minimise the reconstruction error and maximise the monotonicity of the HI extracted from the CAE's bottleneck layer. Because component degradation is typically irreversible, an HI is expected to follow a monotonic trend. Therefore, monotonicity is a widely adopted criterion for constructing meaningful HIs (She and Jia (2019); Yang et al. (2022)). Maximising monotonicity enables the HI to more accurately represent the component's degradation process, thereby enhancing the accuracy of RUL prediction. The CAE relies on PSO training to maximise the HI monotonicity. This is because monotonicity is defined over the entire run-to-failure trajectory. Such a global property cannot be effectively optimised using standard backpropagation applied sequentially to individual data instances. The training, therefore, maximises the fitness function:

$$f' = f_{\mathrm{M}} - f_{\mathrm{R}} - f_0 - f_1 \tag{3}$$

where f_{M} is the monotonicity of the HI obtained from the bottleneck layer, quantified using the Mann-Kendall (MK) metric (Pohlert (2015)), f_{R} is the CAE reconstruction error, measured as the mean squared error (MSE) between the input and output, $f_0 = |H(0)|$ is the absolute initial HI value, and $f_1 = |1 - H(end)|$, where $H(end)$ denotes the final HI value. By maximising f', the terms f_0 and f_1 are minimised, enforcing HI values of 0 and 1 at the pristine and failed states, respectively.

SCADA measurements exhibit high noise levels and are affected by varying operational and environmental conditions, making signal reconstruction and denoising more difficult for the CAE compared to the cleaner laboratory vibration data used in Eftekhari Milani et al. (2024). In Eftekhari Milani et al. (2025), the method is adapted to SCADA signals by incorporating an additional term, f_{std}, defined as the average weekly rolling standard deviation of the HI. This term, weighted by a factor of 3 to balance its scale with the other loss terms, is minimised during training to produce less noisy HIs. Furthermore, a weight of

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

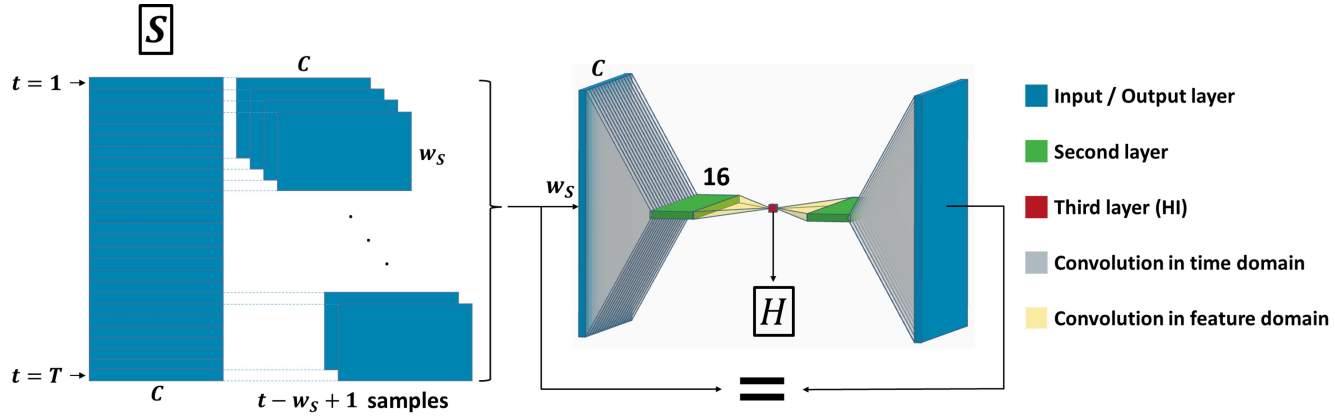

Figure 2. CAE architecture

3 is applied to f_{R} to compensate for the slow minimisation of the reconstruction error, which can slow down the entire training process. The adapted training fitness function used in this work is therefore:

$$f = f_{\mathrm{M}} - 3f_{\mathrm{R}} - f_0 - f_1 - 3f_{\mathrm{std}} \tag{4}$$

The CAE architecture is shown in Figure 2. A rolling time window of length w_S, set to its default value of 10 as in Eftekhari Milani et al. (2024), is applied to the SCADA dataset before inputting it to the CAE. Hence, the HI value at time t_k is computed from the signal values at timestamps $t_k - w_S + 1, t_k - w_S + 2, \ldots, t_k$. In the encoder, a convolution in the time domain using 16 filters of size w_S transforms the input shape $(w_S, C, 1)$ into $(1, C, 16)$. A subsequent convolution in the feature domain with a single filter of size C then produces the HI value at the current timestamp. The decoder mirrors this architecture to reconstruct the input shape $(w_S, C, 1)$. The hyperparameters follow the configuration proposed in Eftekhari Milani et al. (2024), with the sole exception of the number of filters in the first layer, which is increased from 8 to 16 through trial and error. Once the HI is built, the post-processing method proposed in Eftekhari Milani et al. (2025) is applied to reduce local variations arising from rapidly changing operational and environmental conditions. This method fits a non-parametric locally weighted scatterplot smoothing (LOWESS) regression curve (Cleveland and Devlin (1988)) to the HI. The smoothed curve is subtracted from the HI, and its cumulative maximum is re-added, resulting in a more stable and representative degradation profile. This post-processing method is illustrated in Figure 3 reproduced from Eftekhari Milani et al. (2025), which demonstrates its impact on RUL prediction.

2.1.1 Normal behaviour modelling-based feature extraction

SCADA signals can exhibit significant noise, which may hinder the CAE's ability to effectively construct the HI. To address this, a normal-behaviour modelling-based feature extraction approach is proposed to extract degradation-related features from these noisy signals before they are fed into the CAE for HI construction.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

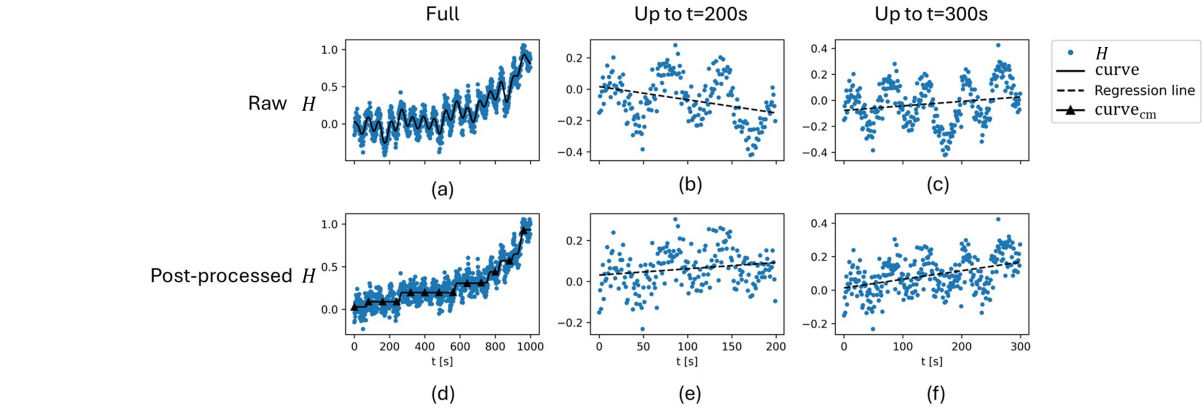

Figure 3. Example of HI post-processing on a hypothetical HI (Eftekhari Milani et al. (2025)): (a–c) raw HI, its first 200s, and its first 300s; (d–f) post-processed HI, its first 200s, and its first 300s

.

For such a noisy signal, its behaviour under healthy component conditions is modelled as a function of the remaining SCADA channels by training a CNN on data from the period during which the wind turbine is assumed to operate in healthy conditions. At least a full year of training data is required to ensure that the CNN effectively captures the signal behaviour across all seasonal variations. The CNN's predicted healthy-state signal values are subtracted from the measured values to obtain residuals, which are then normalised using their mean and standard deviation computed from the training period to produce z-scores. The absolute values of these residual z-scores are taken as degradation features. These features, together with the remaining SCADA channels, are then fed into the CAE to construct the HI.

The CNN architecture is determined through trial and error among various neural network configurations, employing fully connected, recurrent, and convolutional layers, with the objective of minimising prediction error over the training period. A rolling time window of length w_F is applied to the input data, producing input samples of shape $(w_F, C-1)$, where $C-1$ is the number of SCADA channels other than the one being modelled. The model comprises two consecutive 2D convolutional layers with 64 filters of size 3 and ReLU activation function, followed by a max pooling layer (O'Shea and Nash (2015b)), a 2D convolutional layer with 16 filters of size 3, a global average pooling layer (O'Shea and Nash (2015b)), and the final output layer with a single neuron and ReLU activation function. The model's complete architecture, including the output shape of each layer, is shown in Figure 4.

2.2 Probabilistic RUL prediction using LSTM-based CQR

In this step, a model is developed that predicts the three quantiles of the RUL at time t_k, i.e., $\hat{q}_{\alpha/2}[R(t_k)]$, $\hat{q}_{0.5}[R(t_k)]$, and $\hat{q}_{1-\alpha/2}[R(t_k)]$, given the historical health indicator $H(t)$ up to $t = t_k$. First, an LSTM-based quantile regression (QR) model is used to estimate these quantiles. Then, the upper and lower quantiles are calibrated using a conformal prediction approach to guarantee the finite-sample coverage condition defined in Eq. 2.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

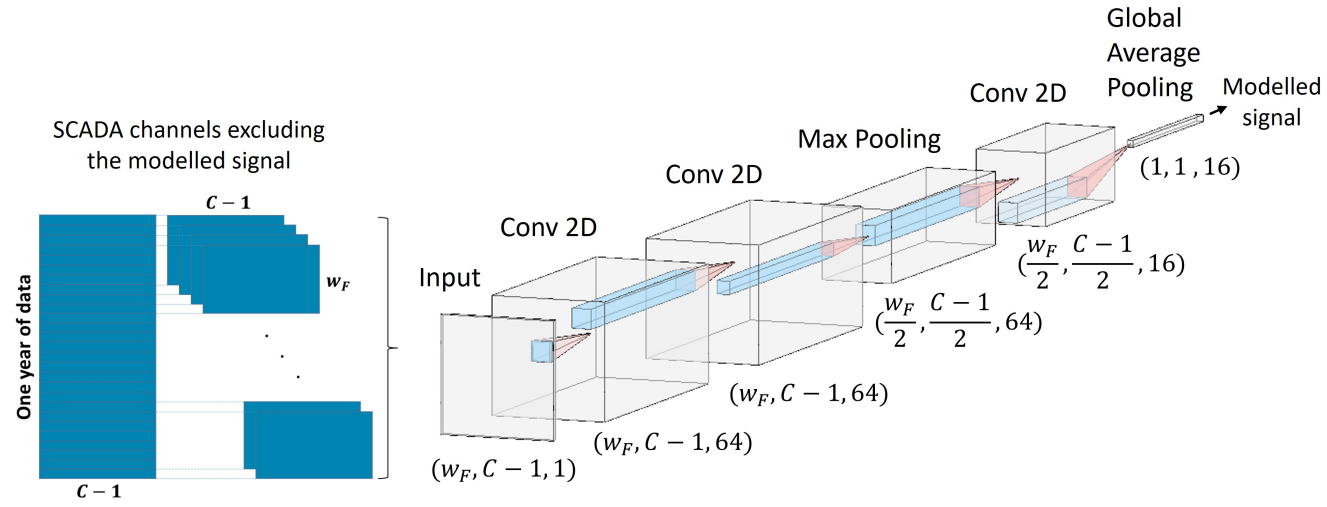

Figure 4. CNN architecture

.

2.2.1 LSTM-based QR

QR is a method for estimating conditional quantiles, such as the median or the 90th percentile, of a response variable, providing a more comprehensive view of the variable's distribution compared to traditional mean regression (Koenker (2005)). Rather than focusing solely on the conditional mean, QR models the relationship between input and response variables at different points of the response distribution. In contrast to classical regression, where the conditional mean of the response variable is estimated by minimising the sum of the squared residuals on the training samples, QR estimates a conditional quantile, q_τ, of the response variable by minimising the "Pinball loss" (Steinwart and Christmann (2011)) defined as:

$$\mathcal{L}_\tau(y,\hat{y}) = \begin{cases} \tau(y-\hat{y}) & \text{if } y-\hat{y} > 0, \\ (1-\tau)(\hat{y}-y) & \text{otherwise} \end{cases} \tag{5}$$

where τ is the target quantile. In the work, the proposed QR model simultaneously estimates $\hat{q}_{\alpha/2}[R(t_k)]$, $\hat{q}_{0.5}[R(t_k)]$, and $\hat{q}_{1-\alpha/2}[R(t_k)]$, given $H(t)$ with $t = t_k - w_H + 1, \ldots, t_k - 1, t_k$. Increasing window length w_H enables the model to incorporate more historical information from earlier timestamps when predicting the RUL quantiles at time, t_k, thereby improving performance. However, excessively large values of w_H increase model complexity and computational burden, and can negatively affect performance due to the risk of overfitting arising from the increased number of parameters, as well as the inclusion of distant historical data that adds noise rather than useful predictive information. The model architecture, shown in Figure 5, is determined through a trial-and-error process using a validation set. It comprises three LSTM layers, each with 64 cells. The output of the final LSTM layer is connected to three parallel output layers, each containing one neuron with an exponential

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

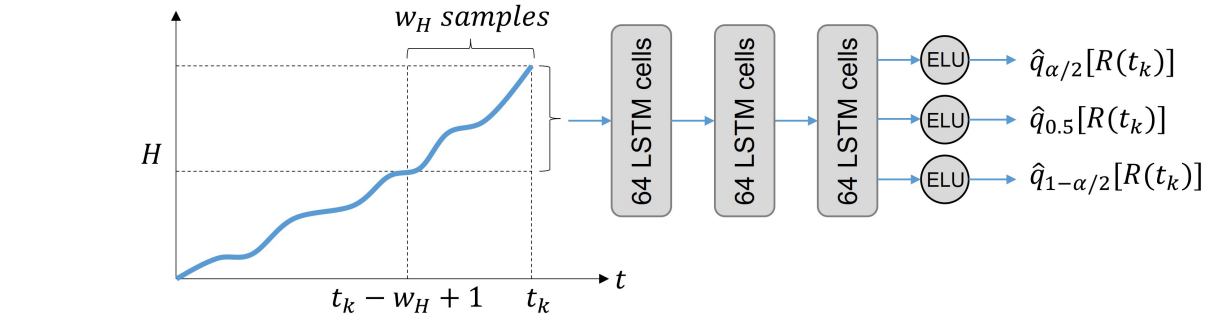

Figure 5. LSTM architecture

linear unit (ELU) activation function, defined as:

$$ELU(y) = \begin{cases} y & \text{if } y > 0, \\ e^y - 1 & \text{otherwise} \end{cases} \tag{6}$$

In this work, the commonly used ReLU activation function was found to be unstable during training, causing the optimisation to stall. The ELU activation avoids this issue by stabilising and speeding up the training through producing smooth negative outputs with non-zero gradients, which helps the network continue learning and results in more stable and faster convergence (Clevert et al. (2015)).

Each output neuron is trained to predict one of the three quantiles of R. This is achieved by assigning to each neuron a Pinball loss (Eq. 5) with the α value corresponding to the quantile it estimates. The overall model training loss function is then defined as the sum of the three individual quantile losses:

$$\mathcal{L} = \mathcal{L}_{\alpha/2} + \mathcal{L}_{0.5} + \mathcal{L}_{1-\alpha/2} \tag{7}$$

2.2.2 Calibration of the quantiles using conformal prediction

While the QR model generates initial estimates for the RUL intervals, it offers no formal guarantee that these intervals will achieve the target coverage on unseen data. To address this limitation and satisfy the finite-sample coverage requirement in Eq. 2, a calibration procedure is required. This is implemented by randomly partitioning the initial training dataset into two disjoint subsets: a training set, $\mathcal{D}_{train}$, and a calibration set, $\mathcal{D}_{calib}$. The QR model is trained exclusively on $\mathcal{D}_{train}$, reserving $\mathcal{D}_{calib}$ strictly for conformal correction of the interval estimates.

After training, the QR model predicts the $\hat{q}_{\alpha/2}[R(t)]$ and $\hat{q}_{1-\alpha/2}[R(t)]$ for each instance in the calibration set. A non-conformity score is then computed for each calibration sample as:

$$E(t) = \max\{\hat{q}_{\alpha/2}[R(t)] - R(t), R(t) - \hat{q}_{1-\alpha/2}[R(t)]\} \tag{8}$$

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

The conformal correction value, q_c, is obtained as the $\lceil(1+\frac{1}{|\mathcal{D}_{calib}|})(1-\alpha)\rceil$-th empirical quantile of the calibration set non-conformity scores (Angelopoulos and Bates (2023)), where $|\mathcal{D}_{calib}|$ is the number of calibration samples and α is the miscoverage rate. The quantile estimates predicted by the QR model are then corrected as:

$$\begin{cases} \hat{q}^c_{\alpha/2}[R(t)] = \hat{q}_{\alpha/2}[R(t)] - q_c \\ \hat{q}^c_{1-\alpha/2}[R(t)] = \hat{q}_{1-\alpha/2}[R(t)] + q_c \end{cases} \tag{9}$$

It is important to note that the correction value, q_c, can be either positive or negative. This is because the non-conformity score in Eq. (8) is negative when $R(t_k)$ lies within the initial interval estimates, and positive when it falls outside. Therefore, this conformal correction simultaneously addresses both over-coverage and under-coverage of the RUL intervals (Romano et al. (2019)).

To evaluate the performance of the predicted RUL quantiles, the three most commonly used metrics for evaluating probabilistic RUL predictions (Javanmardi and Hüllermeier (2023); Mao et al. (2024)) are considered:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{q}_{0.5}[R(t)] - R(t))^2} \tag{10}$$

$$Cov = \frac{1}{T}\sum_{t=1}^{T}\mathbf{I}_{\{\hat{q}^c_{\alpha/2}[R(t)]<R(t)<\hat{q}^c_{1-\alpha/2}[R(t)]\}} \tag{11}$$

$$MIW = \frac{\sum_{t=1}^{T}(\hat{q}^c_{1-\alpha/2}[R(t)] - \hat{q}^c_{\alpha/2}[R(t)])}{T} \tag{12}$$

where $RMSE$ measures the point-prediction error of the median RUL estimate $\hat{q}_{0.5}$, Cov measures the empirical coverage of the predicted interval estimates, where the indicator function $\mathbf{I}_{\{condition\}}$ returns 1 if the $condition$ is satisfied and 0 otherwise, and MIW is the mean width of these intervals.

These metrics complement each other by addressing distinct aspects of predictive quality: accuracy, reliability, and informativeness. While $RMSE$ assesses the accuracy of the model's central tendency, it fails to capture the associated risk. Cov addresses this by validating the reliability of the confidence intervals, ensuring safety, yet it can be trivially maximised by generating infinitely wide intervals. Therefore, MIW acts as a crucial counter-metric to coverage, measuring the sharpness and practical utility of the predictions. The ideal model achieves the target coverage with the minimum possible interval width.

3 Case study 1

In this case study, the proposed method is applied to the CMAPSS dataset, and its performance is compared with existing methods in the literature.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

3.1 Dataset

CMAPSS is a simulation software developed by the National Aeronautics and Space Administration for modelling the dynamics of turbofan engines for research in fault detection, diagnosis, and RUL prediction (Saxena et al. (2008)). The dataset comprises four experiments featuring different operational and failure modes. Each sample includes three operational parameters—flight altitude, Mach number, and throttle position—and 21 sensor measurements capturing pressures, temperatures, and flow rates at key engine stations. These parameters are listed and described in Table 1.

This work utilises the first experiment (FD001), which comprises 100 units for training and 100 units for testing. The degradation mechanism simulated in FD001 corresponds to a fault in the high-pressure compressor. In the training set, all units are run to failure, whereas in the test set, the units are truncated at a given operating cycle. FD001 is widely adopted as the primary benchmark in the prognostic literature due to its controlled experimental design. Unlike the other subsets (e.g. FD002 or FD004), which introduce six operating conditions and multiple failure modes, FD001 operates under a single operating condition (sea level) and a single fault mode. This isolation eliminates the confounding effects of regime-shifting and complex fault interactions, allowing for a focused evaluation of the proposed prognostic algorithm's ability to model the underlying degradation trend.

3.2 Data pre-processing and division

Data channels with constant measurements were removed. These include o_3, s_1, s_5, s_{10}, s_{16}, s_{18}, and s_{19}, resulting in 17 remaining channels. Out of the 100 available initial training units, 90 were randomly selected as the training set while the remaining 10 units were used as the calibration set, as described in Javanmardi and Hüllermeier (2023). All data were scaled to the interval $[0, 1]$ using min-max normalisation computed from the training set.

A piecewise linear definition of RUL is widely adopted in the literature for the CMAPSS dataset (Heimes (2008); Javanmardi and Hüllermeier (2023); Mao et al. (2024)). This labelling strategy reflects the fact that component health state in the CMAPSS dataset starts to deteriorate after a period of stable operation. Consequently, the ground truth RUL is modelled as a constant value during the early life of the engine and only begins to decrease linearly after a fault inception point. This is mathematically defined by capping the RUL at a saturation value, R_{max}:

$$R(t) = \min(R_{\text{max}}, T - t) \tag{13}$$

where R_{max} is set to 125 cycles, T is the failure time, and t is the cycle index. Capping the RUL is crucial for effective model training. If a strictly linear RUL is used from the very first cycle, the model will be forced to predict distinct, high RUL values (e.g. 200 vs. 150) based on sensor data that represents a statistically identical health state. This contradiction creates an unlearnable mapping, which introduces significant noise into the loss function and prevents the model from converging.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Table 1. Description of operational settings and sensor measurements in the CMAPSS dataset (FD001).

Symbol	Description	Units
Operational Settings		
o_1	Flight altitude	kft
o_2	Mach number	-
o_3	Throttle resolver angle (TRA)	%
Sensor Measurements		
s_1	Total temperature at fan inlet (T2)	°R
s_2	Total temperature at low-pressure compressor (LPC) outlet (T24)	°R
s_3	Total temperature at high-pressure compressor (HPC) outlet (T30)	°R
s_4	Total temperature at low-pressure turbine (LPT) outlet (T50)	°R
s_5	Pressure at fan inlet (P2)	psia
s_6	Total pressure in bypass-duct (P15)	psia
s_7	Total pressure at HPC outlet (P30)	psia
s_8	Physical fan speed (Nf)	rpm
s_9	Physical core speed (Nc)	rpm
s_{10}	Engine pressure ratio (EPR)	-
s_{11}	Static pressure at HPC outlet (Ps30)	psia
s_{12}	Ratio of fuel flow to Ps30 (phi)	pps/psi
s_{13}	Corrected fan speed (NRf)	rpm
s_{14}	Corrected core speed (NRc)	rpm
s_{15}	Bypass ratio (BPR)	-
s_{16}	Burner fuel-air ratio (farB)	-
s_{17}	Bleed enthalpy (htBleed)	-
s_{18}	Demanded fan speed (Nf_dmd)	rpm
s_{19}	Demanded corrected fan speed (PCNfR_dmd)	rpm
s_{20}	High-pressure turbine (HPT) coolant bleed (W31)	lbm/s
s_{21}	LPT coolant bleed (W32)	lbm/s

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

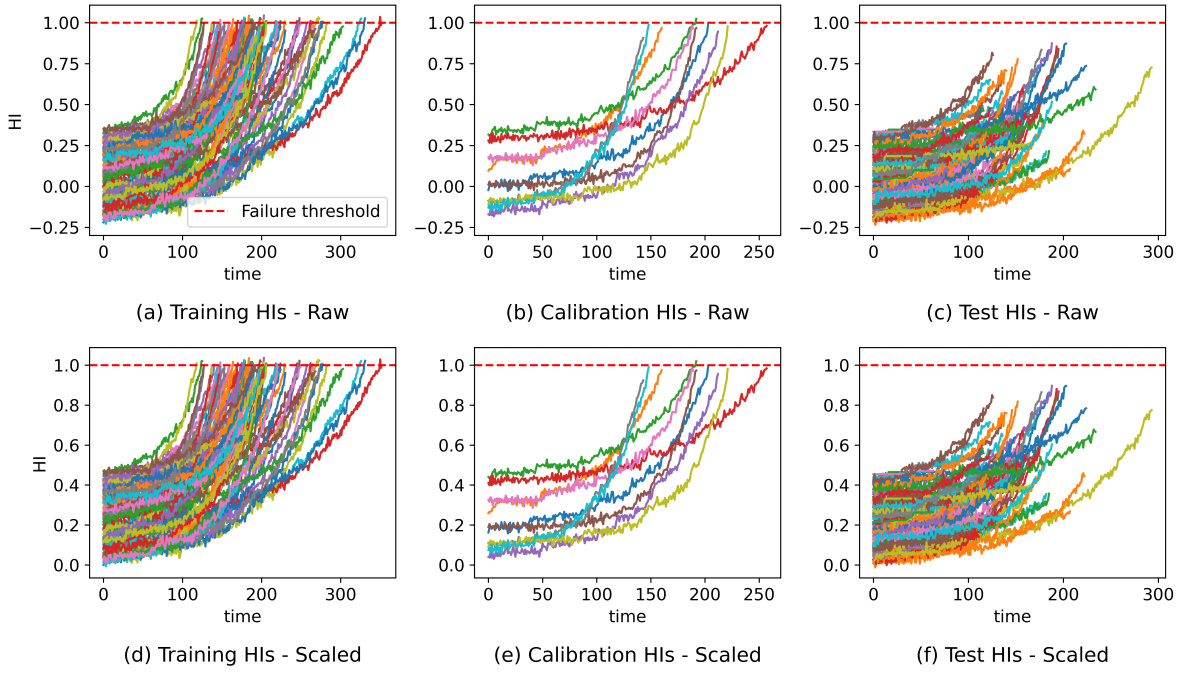

Figure 6. CMAPSS FD001 HIs constructed using the CAE

3.3 HI construction

The CAE was trained using the training signals, constructing the training set HIs. It was then used to predict the calibration and test set HIs, as shown in Figure 6(a–c). The HIs in both the training and calibration sets accurately capture the failure time, reaching a value close to 1 at the point of failure. Since each engine unit in this dataset starts operating with a given initial degradation level, the initial timestamp does not necessarily refer to a pristine condition. This results in a spread of initial HI values, which is consistent across the training, calibration, and test sets and reflects the variability embedded in the CMAPSS simulator.

To align the constructed HIs with the definition used in this work, namely, that an HI represents the percentage of component degradation, and to facilitate the training of the RUL prediction model, the minimum initial HI value in the training set, $H_0 \approx -0.22$, was treated as the pristine condition. Physically, this assumption is grounded in the monotonic and irreversible nature of degradation. In a heterogeneous fleet where units initiate operation with varying degrees of pre-existing wear, the unit exhibiting the lowest initial degradation serves as the best available proxy for a pristine state. Consequently, all HIs were rescaled according to:

$$H_{scaled} = \frac{H - H_0}{1 - H_0} \tag{14}$$

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

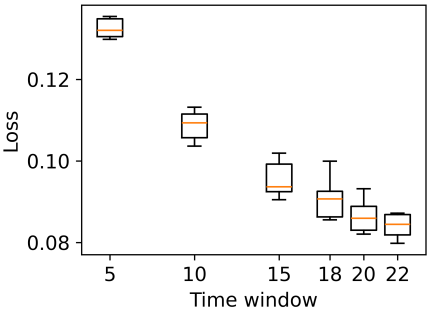

Figure 7. Grid search results for selecting the CQR time window, w_H, on the CMAPSS dataset.

where H is an HI from the training, calibration, or test set. The scaled HIs are shown in Figure 6(d–f).

3.4 RUL prediction

After constructing the HIs, the LSTM-based CQR model was used to predict the RUL for each test case. The QR model was trained on the training set HIs using the Adam optimiser with default parameters (Kingma and Ba (2014)). In each training epoch, 20% of the training instances were randomly selected as a validation set. The training was terminated when the validation loss did not improve for 50 consecutive epochs.

The window size parameter, w_H, was selected via grid search. For each candidate value, the QR model was trained and validated 10 times on the training and calibration sets. The corresponding validation losses are shown in Figure 7. The loss decreases as w_H increases and plateaus around $w_H = 20$. Consequently, larger windows provide marginal performance improvement while increasing model complexity. Because the shortest HI sequence in the test set has a length of 22, the maximum admissible w_H is 22. Hence, a $w_H = 22$ was selected.

After training, the QR model was used to predict the RUL quantiles for both the calibration and test sets. The conformal correction value, q_c, was computed using the calibration predictions, and the test set quantiles were corrected according to Eq. 9. Example RUL predictions for 6 out of 100 test units are shown in Figure 8.

The performance is summarised in Table 2. The RUL prediction was repeated under different confidence intervals to allow comparison with existing methods proposed in the literature. It is important to note that there are slight differences in implementation between these studies. For instance, Javanmardi and Hüllermeier (2023) adopts the same 90/10 training–calibration split as this work, whereas Mao et al. (2024) and Wang et al. (2025) do not report the split ratio. Allocating a larger portion of the data for training improves point prediction performance, but can reduce the reliability of the predicted intervals. Piao et al. (2025) adopts cross-validation rather than a fixed training/validation split, and Wang et al. (2025) uses $R_{\max} = 130$ instead of the widely used 125. This variation in the RUL definition leads to systematic discrepancies in the error calculation, rendering direct $RMSE$ comparisons problematic. Consequently, a strict like-for-like comparison between these approaches is not possible.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

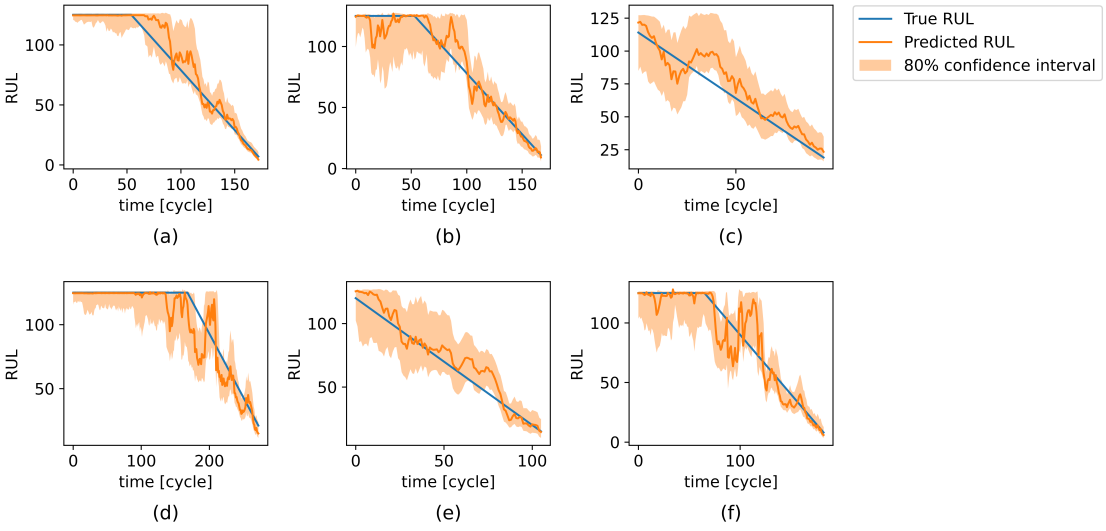

Figure 8. RUL predictions for six test set units using the proposed approach.

Nonetheless, the proposed approach generally produces confidence intervals with more accurate empirical coverage, although it does not outperform existing methods in terms of MIW or $RMSE$. This is expected, as the proposed approach is specifically designed to handle the complexities inherent in real SCADA datasets, namely, high noise and variable operating conditions, rather than simulated datasets like CMAPSS, which are comparatively clean and stationary. When applied to such datasets, this robust design may be less aggressive in curve fitting than models specialised solely for simulated environments, resulting in a marginal penalty in accuracy.

Regardless of these comparative nuances, this case study is essential for benchmarking the proposed method against a standardised community reference, ensuring its generalisability on a widely recognised benchmark. This demonstrates that the method is robust not only for complex real-world data but also in controlled experimental settings.

4 Case study 2

Following the evaluation using simulated data, this case study tests the proposed method on a real-world SCADA dataset comprising multiple field gearbox bearing failures (Desai et al. (2020)). To maintain experimental consistency, the hyperparameter configuration remains aligned with case study 1, with any necessary deviations explicitly noted. The performance of the proposed framework is benchmarked against a state-of-the-art method originally developed for the CMAPSS dataset. This comparison serves a dual purpose: to validate the proposed method's efficacy and robustness in an industrial environment and to highlight the challenges inherent in transferring models from clean, simulated benchmarks to the noisy, non-stationary conditions typical of field operations.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Table 2. Performance comparison of the proposed method with methods in the literature (A–E refer to Javanmardi and Hüllermeier (2023), Mao et al. (2024), Piao et al. (2025), Wang et al. (2025), and Noot et al. (2025), respectively).

Method	$RMSE$	50%		60%		70%		75%		80%		90%	
		Cov	MIW	Cov	MIW	Cov	MIW	Cov	MIW	Cov	MIW	Cov	MIW
Proposed	12.72	**50%**	**18.19**	**63%**	**22.60**	**79%**	29.31	**78%**	31.84	**83%**	32.33	**92%**	42.87
A	13.61	71%	21.00	80%	24.11	87%	**27.86**	–	–	88%	**31.10**	–	–
B	12.62	–	–	–	–	–	–	87%	**27.53**	–	–	94%	37.47
C	13.62	–	–	–	–	–	–	–	–	–	–	95%	42.31
D	12.16	–	–	–	–	–	–	–	–	–	–	**92%**	**31.74**
E	**11.57**	34%	–	–	–	–	–	61%	–	–	–	–	–

Table 3. SCADA channels used in this case study

SCADA channels	**Unit**
Power	W
Rotor speed	rpm
Wind speed	m/s
Gearbox oil temperature	$^{\circ}C$
Gearbox bearing temperature	$^{\circ}C$
Ambient temperature	$^{\circ}C$
Nacelle temperature	$^{\circ}C$

4.1 Dataset

The SCADA dataset used in this case study was collected from a wind farm located in Texas, comprising 100 1.5-MW wind turbines. The signals comprise 10-minute averaged measurements recorded from 2009 to 2019. The SCADA channels used in this study are listed in Table 3. Other channels, such as “status code” and “OK time counter”, were excluded from the analysis, as they exhibited inconsistent recording patterns or provided no meaningful information regarding component health. Six wind turbines, WT1–6, with identical gearbox configurations experienced bearing failures in either the high-speed shaft (HSS) or the intermediate-speed shaft (IMS) of the gearbox, each requiring replacement. These events represent the first gearbox-related failures in each of the six wind turbines, enabling an analysis of the degradation from a pristine state to failure. In this context, the term “pristine” refers to the condition captured at the beginning of the available data history, which corresponds to the period shortly after the turbines were commissioned and entered operational service. The failure dates are reported in Table 4.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

Table 4. Gearbox bearing failure events and component lifetimes.

Wind turbine	Failure type	Failure date	Lifetime (days)
WT1	HSS	2012-02-20	1145
WT2	IMS	2013-05-07	1587
WT3	IMS	2013-05-23	1603
WT4	HSS	2013-05-28	1608
WT5	IMS	2012-02-22	1147
WT6	HSS	2011-01-05	734

4.2 Data preprocessing and division

The data preprocessing step involves removing data points wherein the wind turbine is not operating, characterised by zero rotor speed or negative power. Additionally, erroneous measurements with missing values or those outside the physical bounds, such as very high or low temperatures, were excluded, as they are typically caused by sensor problems or data recording errors. Overall, 13% of the initial data points were dropped. Subsequently, the signals were averaged over daily time frames, reducing the ratio of missing data points from 13% to approximately 1%, thereby improving signal continuity. This temporal aggregation is appropriate given that gearbox degradation typically evolves over timescales of months or years, rather than minutes. This averaging also reduces noise in the signals and lowers the computational burden by decreasing the number of data points. While this downsampling risks masking short-term transient anomalies or rapid shock events, it significantly enhances the signal-to-noise ratio for prognostic modelling. As the objective of this work is to forecast the macroscopic degradation trajectory over an extended horizon, the loss of high-frequency detail is outweighed by the gain in trend stability and robustness. The data sequence length before and after data preprocessing is reported in Table 5.

Table 5. Data availability for each wind turbine, comparing the sequence lengths before and after data preprocessing across different time frames.

	10-minute data sequence lengths		Daily data sequence lengths	
WT	Before pre-processing	After pre-processing	Before pre-processing	After pre-processing
1	167830	145565	1174	1169
2	231560	205129	1617	1610
3	236872	206685	1656	1622
4	233666	206508	1634	1624
5	164062	144011	1147	1141
6	107374	90824	756	744

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

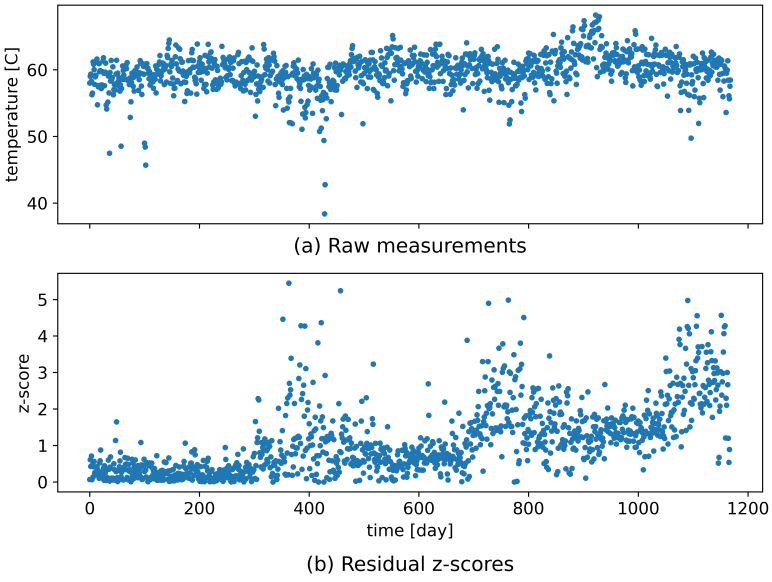

Figure 9. WT1 gearbox bearing temperature: (a) raw measurements and (b) their residual z-scores.

To assess the performance across all six failure cases, a leave-one-out cross-validation approach was adopted. In each fold of the cross-validation, one turbine was set aside for testing, whereas the remaining five turbines were randomly split into a training set of three turbines and a calibration set of two turbines. Each channel was scaled to the $[0, 1]$ range using min-max normalisation based on the training set minimum and maximum values.

4.3 HI construction

Since the gearbox oil is cooled, its temperature provides limited information on degradation, which is typically associated with elevated temperatures. Similarly, the nacelle temperature is highly correlated with the ambient temperature and is therefore difficult to use as a condition monitoring signal. As a result, these two SCADA channels were excluded from CAE training, leaving the gearbox bearing temperature as the only gearbox-specific signal available for HI construction, alongside power, rotor speed, wind speed, and ambient temperature, which represent operation- and environment-specific signals.

Due to the high level of noise, HIs built using the raw gearbox bearing temperature measurements tend to be sub-optimal. Therefore, instead of raw gearbox bearing temperatures, their residual z-score time series, obtained using the proposed normal behaviour modelling-based feature extraction method described in Section 2.1.1, were used as CAE input. This stands in contrast to the first case study, wherein the clean, simulated nature of the CMAPSS data allowed for the direct use of raw signals without such feature extraction. Figure 9 shows these two time series for the case of WT1 up to failure time.

For each cross-validation fold, the CAE was trained on the training turbines and used to build HIs for all turbines in the training, calibration, and test sets. Hence, for each fold, three training HIs, two calibration HIs, and one test HI were built.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

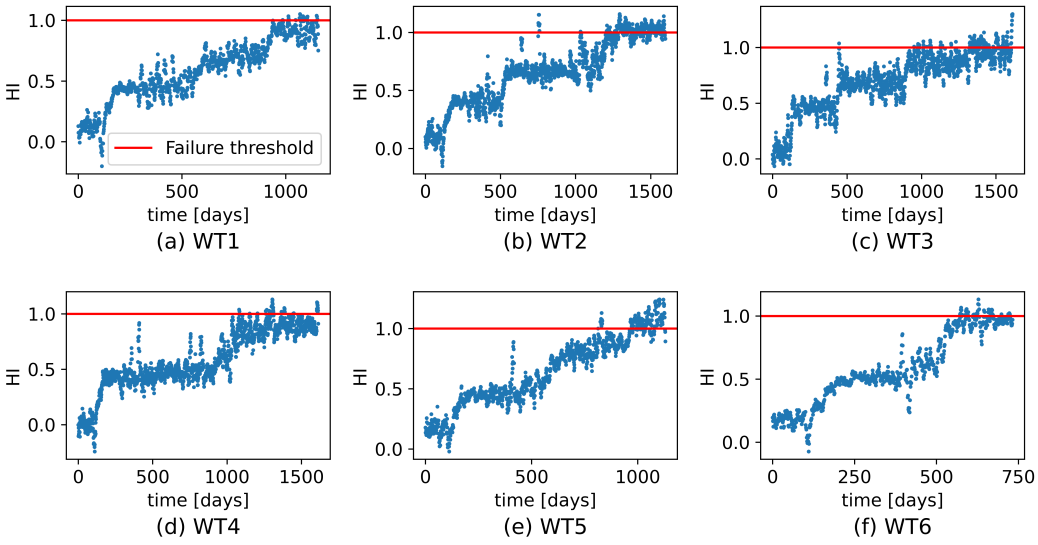

Figure 10. Test HIs obtained for the six failure cases.

The test HIs are shown in Figure 10. They generally exhibit a monotonic trend, with values close to 0 in the pristine state and approaching 1 near failure. However, as expected in a field setting, they are significantly noisier than the HIs built for the CMAPSS dataset, shown in Figure 6. This volatility poses a significant challenge for standard regression models, leading to erratic and unreliable RUL point forecasts. The proposed RUL prediction approach mitigates this issue through two mechanisms: the LSTM architecture leverages long-term temporal dependencies to filter out transient noise while the conformal prediction framework automatically widens the prediction intervals in response to increased variance. This mechanism ensures that the uncertainty is accurately quantified, preventing overconfident predictions in regions of high signal instability.

4.4 RUL prediction

Once the HIs were constructed, the LSTM-based CQR approach, described in Section 2.2, was used to predict the RUL median and 80% confidence intervals for each test case. The QR model was trained in a manner similar to that of case study 1. The w_H parameter was selected using a grid search approach. In the first fold of cross-validation, a w_H value was set, and the QR model was trained and validated 10 times on random train-calibration splits. The results are shown in Figure 11, for which a w_H value of 50 was selected.

After training, the RUL quantiles were predicted for the calibration and test sets. The conformal correction value q_c was obtained from the calibration set predictions, and the test quantiles were corrected according to Eq. 9. Figure 12 shows the test set RUL predictions. A distinct characteristic observed across all units is the progressive narrowing of confidence intervals as the system nears failure. This reflects the model's increasing certainty as degradation patterns become more distinct with the progression of component wear. Furthermore, while the median predictions exhibit local fluctuations driven by the inherent

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

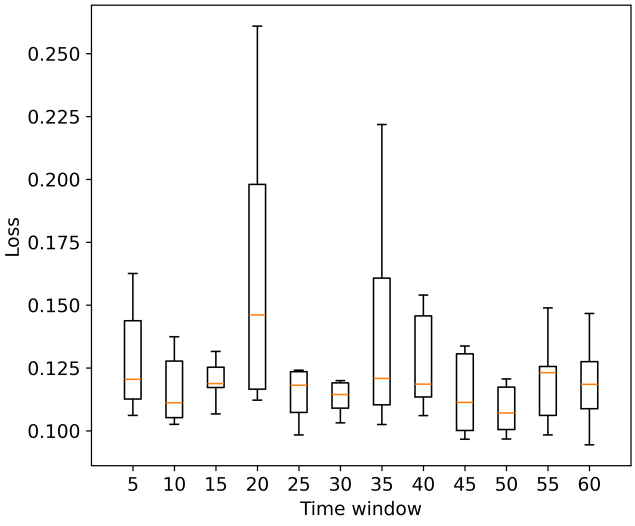

Figure 11. Time window (w_H) selection via grid search for the SCADA dataset.

volatility of the field data, the adaptive width of the confidence intervals dynamically compensates for this variance, ensuring that the true RUL remains reliably covered even during periods of signal instability. However, for some turbines, predictions exhibit relatively large errors and wide confidence intervals even in the vicinity of failure. This error is likely attributable to the high degree of heterogeneity in fault progression and significant sensor noise, which can mask the degradation patterns and force the model to maintain conservative uncertainty bounds to ensure validity.

4.5 Comparison with the literature

The best-performing approach proposed in Javanmardi and Hüllermeier (2023) was adapted and implemented using the SCADA dataset for comparison. This approach was chosen due to its strong performance on CMAPSS and the open-source availability of the code on the first author's GitHub[1], which enables its application to the SCADA dataset. The approach is a CQR model based on a deep convolutional neural network (DCNN), trained using sensor signals to predict the RUL. In contrast to the proposed method, which constructs intermediate HIs from sensor data and post-processes them to reduce noise before feeding them into an LSTM-based CQR model, this approach directly maps raw sensor signals to RUL predictions via a DCNN. A rolling time window, w_D, was applied to the sensor signals, creating 2D input samples of shape (w_D, C). The DCNN comprises four 2D convolutional layers with 10 filters of shape $(10, 1)$ and $tanh$ activation function, followed by a 2D convolutional layer with one filter of shape $(3, 1)$ and $tanh$ activation function. The output is passed through a dropout layer with a dropout rate of 0.5 and a fully connected layer with 100 neurons, using the $tanh$ activation function. The output layer has three neurons, each predicting one RUL quantile. The model was trained using the Adam optimiser with default parame-

[1] https://github.com/alireza-javanmardi/conformal-RUL-intervals

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

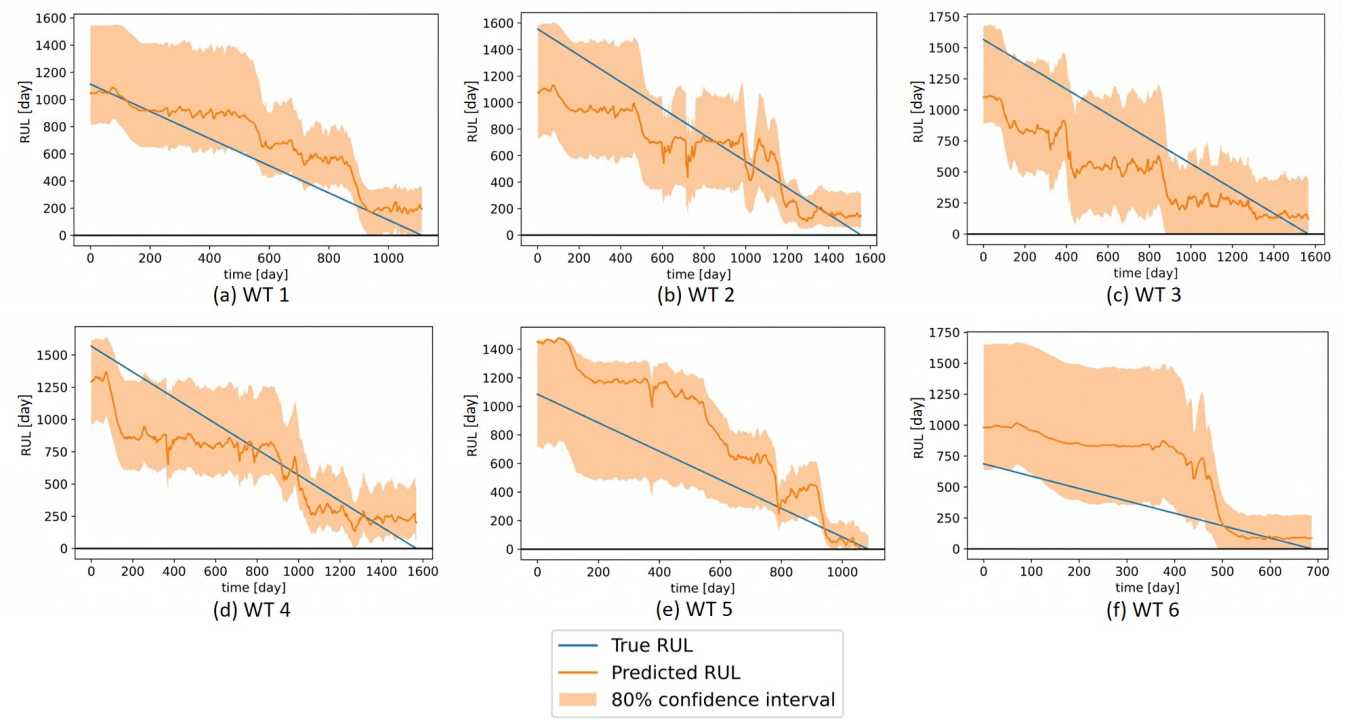

Figure 12. Predicted RULs for the test set using the proposed CQR approach.

ters. An early-stopping criterion was applied to the model's training to ensure proper training and prevent overfitting. In each training epoch, 20% of the training instances were randomly selected as a validation set. The training was terminated when the validation loss did not improve for 50 consecutive epochs.

The w_D parameter was set via grid search, similar to the proposed method. The results are shown in Figure 13, from which $w_D = 45$ is selected. Test RUL predictions using this method are shown in Figure 14, and a detailed comparison of performance is reported in Table 6. The results show that the DCNN-CQR method significantly underperforms the proposed method. The point prediction error of the proposed method, measured using the RMSE metric, is around 67% lower than that of the DCNN-CQR method on average, indicating a better capture of degradation patterns despite noisy and highly variable operational conditions in the SCADA signals. This robustness stems from the CAE's ability to decouple the underlying degradation signal from the noise and operating conditions in sensor signals, a process further enhanced by post-processing the constructed HIs to suppress local fluctuations. Furthermore, the predicted intervals from the proposed method achieve more accurate empirical coverage, being closer to the defined target 80%, and are 42% shorter on average as measured using the MIW metric, indicating higher precision. Both methods satisfy the finite sample coverage requirement in Eq. 2, except for WT6, whose failure occurs considerably earlier than the other five turbines. In this case, the test set is less representative of the training and calibration sets, reducing prediction accuracy. This issue could be alleviated by including additional failure cases, thereby improving representativeness across cross-validation folds.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

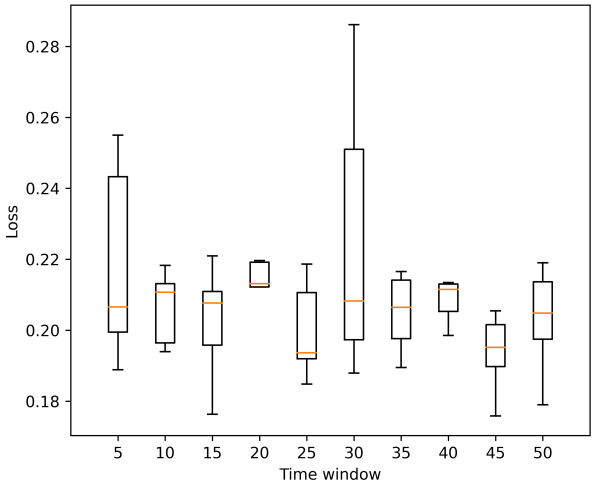

Figure 13. The grid search results to set the time window (w_D) parameter of the DCNN-CQR approach in the SCADA dataset.

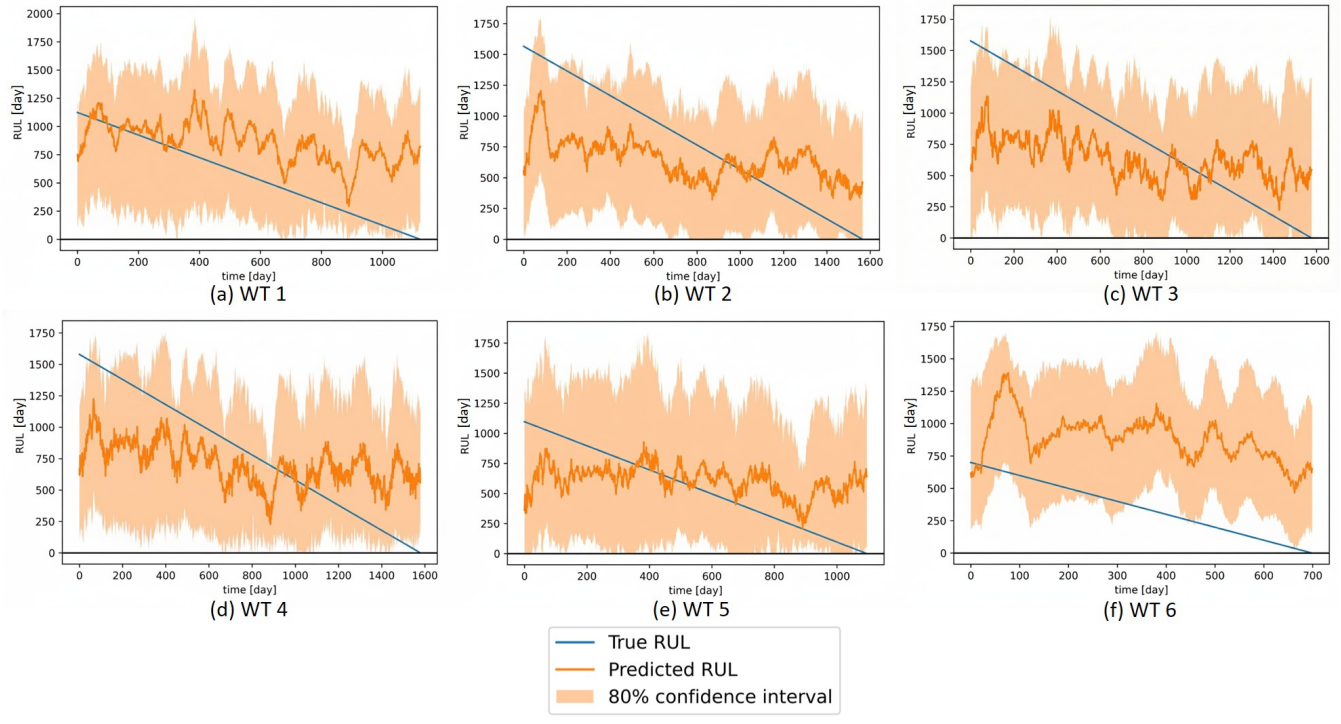

Figure 14. Test set RUL predictions using the DCNN-CQR approach.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Table 6. Performance comparison between the DCNN-CQR model from Javanmardi and Hüllermeier (2023) and the proposed method.

Metrics	WT1		WT2		WT3	
	DCNN-CQR	**Proposed**	DCNN-CQR	**Proposed**	DCNN-CQR	**Proposed**
$RMSE$	548.91	**179.08**	365.01	**93.83**	375.01	**80.68**
Cov	94.04%	**79.60%**	**91.83%**	93.18%	90.81%	**87.62%**
MIW	1209.64	**580.25**	1101.60	**631.92**	1193.26	**751.06**

Metrics	WT4		WT5		WT6	
	DCNN-CQR	**Proposed**	DCNN-CQR	**Proposed**	DCNN-CQR	**Proposed**
$RMSE$	466.70	**120.19**	376.48	**162.53**	641.26	**315.43**
Cov	87.34%	**85.98%**	97.72%	**94.29%**	45.43%	**76.45%**
MIW	1190.19	**586.15**	1282.39	**599.31**	1009.39	**830.38**

5 Conclusions and future work

In recent years, numerous approaches have been developed to predict the RUL of wind turbine components. While many of these methods demonstrate strong performance, they primarily focus on point predictions, whereas an accurate quantification of uncertainty is essential for reliable decision-making. Existing probabilistic RUL prediction approaches remain limited. They either fail to provide accurate uncertainty estimations or have only been validated on simulated datasets. Applying such methods to field data remains considerably more challenging due to high measurement noise and highly variable operational and environmental conditions. In this work, a two-stage method is proposed. Instead of directly modelling the RUL from raw sensor data, the proposed approach first learns a degradation trajectory using an unsupervised HI model based on a CAE. The constructed HIs are then used as input to an LSTM-based CQR model, which probabilistically estimates RUL. This method performs on par with state-of-the-art models on the simulated CMAPSS dataset. When applied to a field SCADA dataset, it substantially outperforms the DCNN-CQR method from Javanmardi and Hüllermeier (2023) in both point-prediction accuracy and the width and empirical coverage of the prediction intervals. This result demonstrates improved robustness to noise and strongly varying operating conditions in real-world data.

Although the proposed method yields superior performance compared to the selected state-of-the-art baseline, the predictions for some turbines still show relatively large errors and wide confidence intervals, even in the vicinity of failure. Future work will investigate these cases in detail and explore alternative modelling choices and neural architectures for constructing HIs and mapping them to RUL, aiming to further strengthen predictive performance and uncertainty quantification in real-world applications.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

Code availability. The codes used for the analyses and experiments presented in this study are available upon request.

Data availability. The dataset used in this research is not publicly accessible due to proprietary restrictions and confidentiality agreements.

Author contributions. AEM: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, original draft preparation; DZ: funding acquisition, project administration, resources, supervision, validation, review, and editing; SS: data curation, validation, review, and editing; SW: funding acquisition, project administration, resources, supervision, review, and editing

Competing interests. One of the co-authors is a member of the editorial board of Wind Energy Science.

Disclaimer. This work was authored in part by the National Laboratory of the Rockies for the U.S. Department of Energy (DOE), operated under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Critical Minerals and Energy Innovation Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Acknowledgements. The authors appreciate the data sharing support in case study 2 by a wind plant owner in the United States.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE DISCUSSIONS

References

Angelopoulos, A. N. and Bates, S.: Conformal Prediction: A Gentle Introduction, Foundations and Trends® in Machine Learning, 16, 494–591, https://doi.org/10.1561/2200000101, 2023.

Arias Chao, M., Kulkarni, C., Goebel, K., and Fink, O.: Aircraft Engine Run-to-Failure Dataset under Real Flight Conditions for Prognostics and Diagnostics, Data, 6, 5, https://doi.org/10.3390/data6010005, 2021.

Casolo, S., Stasik, A., Zhang, Z., and Riemer-Sorensen, S.: Testing Topological Data Analysis for Condition Monitoring of Wind Turbines, https://doi.org/10.48550/ARXIV.2406.16380, 2024.

Chatterjee, J. and Dethlefs, N.: Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future, Renewable and Sustainable Energy Reviews, 144, 111 051, https://doi.org/10.1016/j.rser.2021.111051, 2021.

Chesterman, X., Verstraeten, T., Daems, P.-J., Nowé, A., and Helsen, J.: Overview of normal behavior modeling approaches for SCADA-based wind turbine condition monitoring demonstrated on data from operational wind farms, Wind Energy Science, 8, 893–924, https://doi.org/10.5194/wes-8-893-2023, 2023.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y.: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, https://doi.org/10.48550/ARXIV.1412.3555, 2014.

Cleveland, W. S. and Devlin, S. J.: Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, 83, 596–610, https://doi.org/10.1080/01621459.1988.10478639, 1988.

Clevert, D.-A., Unterthiner, T., and Hochreiter, S.: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), https://doi.org/10.48550/ARXIV.1511.07289, 2015.

De Florio, M., Appleby, G., Keller, J., Eftekhari Milani, A., Zappalá, D., and Sheng, S.: Gearbox Bearing Crack Growth Prognostics with Physics-Informed Machine Learning and Uncertainty Quantification, https://doi.org/10.5194/wes-2025-157, 2025.

Desai, A., Guo, Y., Sheng, S., Phillips, C., and Williams, L.: Prognosis of Wind Turbine Gearbox Bearing Failures using SCADA and Modeled Data, Annual Conference of the PHM Society, 12, 10, https://doi.org/10.36001/phmconf.2020.v12i1.1292, 2020.

Eftekhari Milani, A., Zappalá, D., and Watson, S.: A hybrid Convolutional Autoencoder training algorithm for unsupervised bearing health indicator construction, Engineering Applications of Artificial Intelligence, https://doi.org/10.1016/j.engappai.2024.109477, 2024.

Eftekhari Milani, A., Zappalá, D., Castellani, F., and Watson, S.: Simulating run-to-failure SCADA time series to enhance wind turbine fault detection and prognosis, Wind Energy Science, 10, 2563–2576, https://doi.org/10.5194/wes-10-2563-2025, 2025.

Harvey, A. C.: ARIMA Models, pp. 22–24, Palgrave Macmillan UK, London, ISBN 978-1-349-20865-4, https://doi.org/10.1007/978-1-349-20865-4_2, 1990.

Heimes, F. O.: Recurrent neural networks for remaining useful life estimation, in: 2008 International Conference on Prognostics and Health Management, p. 1–6, IEEE, https://doi.org/10.1109/phm.2008.4711422, 2008.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Javanmardi, A. and Hüllermeier, E.: Conformal Prediction Intervals for Remaining Useful Lifetime Estimation, International Journal of Prognostics and Health Management, 14, https://doi.org/10.36001/ijphm.2023.v14i2.3417, 2023.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://doi.org/10.48550/ARXIV.1412.6980, 2014.

Koenker, R.: Quantile Regression, Cambridge University Press, ISBN 9780511754098, https://doi.org/10.1017/cbo9780511754098, 2005.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L.: Distribution-Free Predictive Inference for Regression, Journal of the American Statistical Association, 113, 1094–1111, https://doi.org/10.1080/01621459.2017.1307116, 2018.

Li, M., Xu, Z., Li, S., Kikuchi, Y., Dong, Y., Gryllias, K. C., Baraldi, P., Zio, E., and Carroll, J.: Health prognostics and maintenance decision-making for wind energy: A comprehensive overview, Renewable and Sustainable Energy Reviews, 226, 116 269, https://doi.org/10.1016/j.rser.2025.116269, 2026.

Mao, S., Li, X., and Zhao, B.: Remaining useful life prediction based on time-series features and conformalized quantile regression, Measurement Science and Technology, 35, 126 113, https://doi.org/10.1088/1361-6501/ad762c, 2024.

McMorland, J., Flannigan, C., Carroll, J., Collu, M., McMillan, D., Leithead, W., and Coraddu, A.: A review of operations and maintenance modelling with considerations for novel wind turbine concepts, Renewable and Sustainable Energy Reviews, 165, 112 581, https://doi.org/10.1016/j.rser.2022.112581, 2022.

Noot, J.-P., Martin, M., and Birmele, E.: LSTM and Transformers based methods for Remaining Useful Life Prediction considering Censored Data, International Journal of Prognostics and Health Management, 16, https://doi.org/10.36001/ijphm.2025.v16i2.4260, 2025.

O'Shea, K. and Nash, R.: An Introduction to Convolutional Neural Networks, https://doi.org/10.48550/ARXIV.1511.08458, 2015a.

O'Shea, K. and Nash, R.: An Introduction to Convolutional Neural Networks, https://arxiv.org/abs/1511.08458, 2015b.

Peter, R., Zappalá, D., Schamboeck, V., and Watson, S. J.: Wind turbine generator prognostics using field SCADA data, Journal of Physics: Conference Series, 2265, 032 111, https://doi.org/10.1088/1742-6596/2265/3/032111, 2022.

Piao, S., Huang, R., and Tsung, F.: CRULP: Reliable RUL Estimation Inspired by Conformal Prediction, IEEE Transactions on Instrumentation and Measurement, 74, 1–11, https://doi.org/10.1109/tim.2024.3522678, 2025.

Pohlert, T.: Non-Parametric Trend Tests and Change-Point Detection, https://doi.org/10.13140/RG.2.1.2633.4243, 2015.

Romano, Y., Patterson, E., and Candès, E. J.: Conformalized Quantile Regression, https://doi.org/10.48550/ARXIV.1905.03222, 2019.

Sankararaman, S. and Goebel, K.: Uncertainty in Prognostics and Systems Health Management, International Journal of Prognostics and Health Management, 6, https://doi.org/10.36001/ijphm.2015.v6i4.2319, 2020.

Saxena, A., Goebel, K., Simon, D., and Eklund, N.: Damage propagation modeling for aircraft engine run-to-failure simulation, in: 2008 International Conference on Prognostics and Health Management, IEEE, https://doi.org/10.1109/phm.2008.4711414, 2008.

She, D. and Jia, M.: Wear indicator construction of rolling bearings based on multi-channel deep convolutional neural network with exponentially decaying learning rate, Measurement, 135, 368–375, https://doi.org/10.1016/j.measurement.2018.11.040, 2019.

Steinwart, I. and Christmann, A.: Estimating conditional quantiles with the help of the pinball loss, Bernoulli, 17, https://doi.org/10.3150/10-bej267, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention Is All You Need, https://doi.org/10.48550/ARXIV.1706.03762, 2017.

Vovk, V., Gammerman, A., and Shafer, G.: Algorithmic Learning in a Random World, Springer International Publishing, ISBN 9783031066498, https://doi.org/10.1007/978-3-031-06649-8, 2022.

Wang, W., Wang, Z., Cai, Z., Hu, C., and Si, S.: Robust uncertainty quantification for online remaining useful life prediction with randomly missing and partially faulty sensor data, Reliability Engineering & System Safety, 262, 111 177, https://doi.org/10.1016/j.ress.2025.111177, 2025.

Yang, Z., Baraldi, P., and Zio, E.: A method for fault detection in multi-component systems based on sparse autoencoder-based deep neural networks, Reliability Engineering and System Safety, 220, 108 278, https://doi.org/10.1016/j.ress.2021.108278, 2022.

https://doi.org/10.5194/wes-2026-36
Preprint. Discussion started: 24 April 2026

Zhang, W., Vatn, J., and Rasheed, A.: Gearbox pump failure prognostics in offshore wind turbine by an integrated data-driven model, Applied Energy, 380, 124 829, https://doi.org/10.1016/j.apenergy.2024.124829, 2025.