https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

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

Transformer-Based Probabilistic Wind Power Prediction and Power Ramping Verification with Error Tolerance

Ruoke Meng[1,2], Joris Van den Bergh[1,3], Hossein Tabari[1,3], Piet Termonia[1,2], and Geert Smet[1]

[1]Royal Meteorological Institute of Belgium
[2]Ghent University
[3]University of Antwerp

Correspondence: Ruoke Meng (ruoke.meng@meteo.be)

Abstract. Accurate probabilistic wind power forecasts and reliable detection of ramping events are important for the stable operation of wind-integrated power systems. This study implements a Self-Attentive Ensemble Transformer for postprocessing operational wind power ensembles in the Belgian Offshore Zone. This postprocessing delivers corrected ensemble members as output rather than a predictive distribution. The Transformer model achieves near-zero bias in the ensemble mean and a lower Continuous Ranked Probability Score across all lead times than power-curve derived raw ensembles and shows some improvements to Member-by-Member benchmark methods. This model also shows the ability to correct for power overestimation associated with the wake effect. For power ramping, this paper focuses on the predictability of 15% and 30% hourly ramping events. The Transformer produces event frequencies that closely align with the observations over all ramping thresholds. Considering that frequent temporal shifts and intensity underestimations limit a fair evaluation of ramping prediction, we introduce an error-tolerant probabilistic verification framework with buffer concepts and score the model by Buffer Brier Skill Score (BBSS). Incorporating a temporal buffer and magnitude tolerance substantially mitigates penalties for minor prediction errors. This reveals that the models are capable of detecting ramping signals, even though they lack strict precision. While uncertainty is inherent in ramping forecasts, the Transformer demonstrates better-calibrated probabilities than the raw ensembles. Overall, the Transformer method improves probabilistic power prediction and provides an informative probabilistic representation of ramping events.

1 Introduction

Renewable energy has become an increasingly important component of contemporary power systems, with offshore wind playing a growing role in several European countries. Belgium has also strongly engaged in the construction of offshore wind farms in recent years. Since 2020, the Belgian Offshore Zone (BOZ) has a total installed capacity of 2262 MW, and further development is planned for the construction of electrical infrastructure for the Princess Elisabeth Energy Island by 2030 (Borgers et al., 2025). As wind capacity increases, the associated variability in power generation becomes more consequential for system operations. Wind power production is driven by highly variable meteorological conditions, which lead to fluctuations in generated power. Of particular operational relevance are power ramping events, defined as large power changes over short

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

time intervals, which can be a challenge for dispatch and balancing decisions (Wan, 2011). Reliable wind power forecasts are therefore required not only for expected generation but also for the probability of significant ramps.

Wind power prediction relies on wind forecasts by Numerical Weather Prediction (NWP) models, which are then converted to power, e.g., using manufacturer-provided turbine power curves. Ensemble NWP systems represent meteorological uncertainty through multiple members and therefore provide a basis for probabilistic power forecasts. However, raw NWP ensembles commonly exhibit systematic biases and insufficient dispersion (Raftery et al., 2005), and these deficiencies propagate to wind power ensembles derived from operational power curves. Statistical postprocessing is therefore required to improve calibration and dispersion. Classical approaches include Ensemble Model Output Statistics (EMOS), which calibrates parametric predictive distributions (Gneiting et al., 2005). In wind power postprocessing, EMOS has been implemented with truncated normal predictive distributions (Phipps et al., 2022). Another approach is Member-by-Member (MBM) correction, which adjusts each ensemble member individually without relying on distributional assumptions (Van Schaeybroeck and Vannitsem, 2015). However, applying linear parametric corrections may be inadequate in the non-linear context of wind power ensemble spread (Eide et al., 2017). Recently, attention-based Transformer models have been introduced for member-wise postprocessing, offering flexibility for multivariate inputs and the ability to represent dependencies across ensemble members (Finn, 2021), and have been validated with improved temperature and precipitation forecasts (Bouallègue et al., 2024). This development has been further demonstrated to improve both 10m and 100m wind speed results (Van Poecke et al., 2025), which is crucial for wind power but has not been directly adapted for power postprocessing.

Beyond ensemble power predictions, ramping analysis is inherently time-dependent. To explicitly account for the temporal dependence, one approach involves generating probabilistic power scenarios from high-resolution deterministic forecasts using Schaake Shuffle or Gaussian Copulas, which introduce dependency structures based on historical observations or prescribed correlation models (Worsnop et al., 2018). Alternatively, multivariate Gaussian regression models can be applied to ensemble forecasts to characterize the joint distribution of predicted wind speeds (Muschinski et al., 2022). By capturing the underlying temporal correlation structure of forecast errors, these methodologies produce physically consistent scenarios that enable probabilistic analysis of wind power ramping events. However, these approaches rely on generated scenarios and are based on synthetic power derived from wind speed instead of actual power production data.

When evaluating the power and corresponding ramping predictions, strict point-to-point evaluations provide a general overview of model performance, but they may lead to an incomplete understanding of forecast utility, as minor temporal or amplitude shifts are often misinterpreted as severe discrepancies (Vallance et al., 2017; Messner et al., 2020). This limitation is especially pronounced for ramping events, where a forecast with a slight timing shift often suffers from a "double-penalty" effect: being penalized once as a miss and again as a false alarm (Meng et al., 2026; Voyant, 2026). Some probabilistic ramping verification frameworks mitigate this issue by identifying ramps within wide time windows (Worsnop et al., 2018; Muschinski et al., 2022). However, from an operational perspective, Transmission System Operators (TSOs) are primarily concerned with rapid, short-duration power ramps that directly impact grid stability (Sørensen et al., 2020; Demaeyer et al., 2026). Addressing these operational needs involves developing power predictions and establishing verification frameworks that can incorporate a controlled temporal tolerance without eliminating meaningful error discrimination.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

This study focuses on addressing the issues discussed above, including member-based nonlinear postprocessing for ensemble wind power prediction, as well as the analysis of ramping predictability and the development of an appropriate verification framework for probabilistic ramping forecasts. First, we use real operational ensemble predictions from the ECMWF Ensemble Prediction System (EPS) and actual power observations from the Belgian TSO. Second, we apply a Transformer-based postprocessing method for raw wind power calibration, in comparison with the benchmark method of MBM, to eliminate forecast errors and to address potential wind farm wake effects. Similar ideas have improved deterministic forecasts in earlier work (Van den Bleeken et al., 2025; Meng et al., 2026), and here we extend that perspective to multivariate probabilistic postprocessing. Finally, we examine the models' ability to represent power ramping events. To account for intensity and timing errors in ramping verification, we introduce a concept of controlled tolerance and use it to assess probabilistic ramping event predictability. By incorporating error tolerance, the evaluation ensures to reflect the practical flexibility needed for real-world power system operations.

2 Data

The BOZ is located in the North Sea, approximately 22 to 54 kilometers off the Belgian coast (Figure 1). This area comprises nine offshore wind farm projects with a combined total installed capacity of 2262 MW. The Belgian offshore wind farm fleet is among the most densely installed areas in the North Sea, with a power density of approximately 10 MW km^{-2} compared to a regional average of 6.6 MW km^{-2} (Murcia Leon et al., 2020). Furthermore, the BOZ is located closely south of the Dutch Borssele wind farm zone. Consequently, this dense spatial configuration creates a complex flow environment characterized by both intra-farm wake effects and inter-farm wake interactions under certain wind directions (Barthelmie et al., 2009).

Operational power forecasts for the BOZ (Smet et al., 2019, 2023) are produced at the Royal Meteorological Institute of Belgium (RMI) by combining the ECMWF Ensemble Prediction System (EPS) with RMI's limited-area NWP models (LAMs). The extracted meteorological variables include the 100-meter wind speed and the 100-meter wind direction, which are representative of the typical turbine hub heights deployed in the BOZ. Previous work on deterministic wind power prediction and ramping evaluation for BOZ can be found in Meng et al. (2026). In this study, we focus solely on the EPS probabilistic forecasts. The ensemble forecasts utilized in this study are initialized at 00:00 UTC and consist of 51 individual members. These members provide predictions for each hour up to 60 lead hours, the typical forecast range of RMI's local LAM forecasts. While ECMWF updates EPS on 27 June 2023 from 0.2 to 0.1 degree spatial resolution (Haiden et al., 2023), all the EPS forecasts here are extracted at the lower 0.2° spatial resolution on a latitude/longitude grid and then spatially interpolated from the closest grid points to the geographical midpoint of each individual wind farm, as in RMI's operational practice.

Wind power observations are provided by the Belgian transmission system operator, Elia. The historical dataset covers the years 2022 and 2023. The observations for each wind farm are available every 15 minutes, but here we only use integer-hour data to match the temporal resolution of the EPS forecasts. In the operational practice at RMI, the wind power predictions are generated by applying power curves and related technical parameters to the wind forecasts, denoted as EPS-PC. While detailed technical parameters remain undisclosed for confidentiality reasons, the power conversion is described here. Specifically, the

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

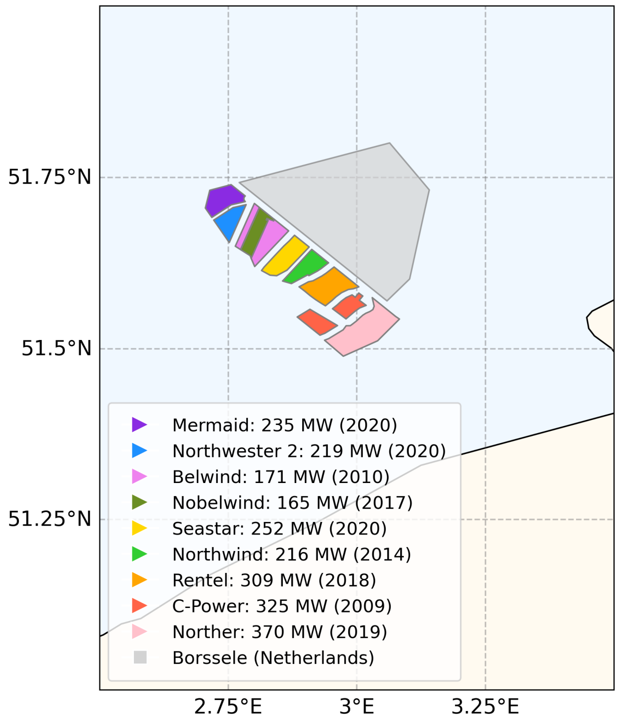

Figure 1. The wind farm projects at the Belgian Offshore Zone (BOZ) and Dutch Borssele wind farm.

power output for each wind farm is first calculated based on the wind speed at the reference location of the park and the total number of installed turbines. Subsequently, a correction factor is applied to each wind farm based on historical production data obtained from Elia. This coefficient compensates for operational losses that might be underestimated by the theoretical curves due to factors such as turbine aging and long-term minor fluctuations in turbine availability.

3 Ensemble power postprocessing methods

3.1 Transformer-based postprocessing

To predict ensemble wind power production while preserving dependencies among members, we apply a Self-Attentive Ensemble Transformer model architecture. The model follows the member-based formulation proposed by Finn (2021), where each member is corrected using self-attentive interactions with the other ensemble members. The framework is subsequently extended by incorporating lead time as an additional dimension within the attention mechanism, enabling the model to capture potential temporal correlations, as introduced by Van Poecke et al. (2025). Our application is tailored to site-specific forecasting by training a separate model for each wind farm. Therefore, our attentive blocks retain the same design for both the attention mechanism and the multilayer perceptron, but omit the spatial dimension.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

In the Transformer module, the latent representations of the ensemble members are transformed into Queries Q, Keys K, and Values V via linear projections. Conceptually, the Value tensor can be interpreted as encoding the physical state of each member, the Query tensor represents the specific information a member requires to correct its trajectory, and the Key tensor dictates what information each member can provide to others. To capture potential correlations, the multi-head self-attention module partitions these projected Q, K, and V tensors across h independent attention heads. This operation divides the total feature dimension $\tilde{c}$ equally, defining the feature dimension per head as $d_k = \tilde{c}/h$. For each attention head, the model computes the attention output A by evaluating the similarity between Q and K and multiplying the result by V:

$$A = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{t \cdot d_k}}\right)V \tag{1}$$

To ensure numerical stability, the dot products are scaled by the square root of the combined temporal and head-specific feature dimensions $\sqrt{t \cdot d_k}$, where t is the number of lead times.

Within each transformer block, the model computes an intermediate representation Z_l using a linear projection with weights W and a residual connection, resulting in the input Z_{l+1} of the next layer:

$$Z_{l+1} = Z_l + WA \tag{2}$$

This interaction mechanism allows each member to effectively borrow information from the collective ensemble to correct systemic biases and adjust the overall spread. After passing through the transformer blocks, the latent features are processed by a member-independent Multi Layer Perceptron (MLP) with the activation function of Gaussian error linear unit (GeLU), which ensures the model is permutation invariant. To constrain the predicted power P within the normalized capacity factor range between 0 and 1, a sigmoid function is applied to the output of this final MLP block:

$$P = \mathrm{Sigmoid}(\mathrm{MLP}(Z_{\mathrm{final}})) = \frac{1}{1 + e^{-\mathrm{MLP}(Z_{\mathrm{final}})}} \tag{3}$$

The input time series of each ensemble member is first mapped into a latent representation with a total feature dimension of $\tilde{c} = 64$ with an embedding layer. The encoder then stacks two ensemble Transformer blocks. Within each block, we employ $h = 4$ attention heads in the self-attentive module, which results in a projected feature dimension of $d_k = 16$ per head. This configuration computes a shared attention matrix by jointly attending over the temporal and feature dimensions, enabling the model to capture full sequence dependencies. The attention output is subsequently processed by the MLP block, where the internal hidden dimension is expanded to 256. To mitigate overfitting, a dropout rate of 0.1 is applied within both the attention and MLP modules. The model parameters are optimized using a learning rate of 0.0001, with the Continuous Ranked Probability Score (CRPS) serving as the training objective in its kernel formulation:

$$\mathrm{CRPS}_{\mathrm{kernel}} = \frac{1}{M}\sum_{m=1}^{M}|P_m - O| - \frac{1}{2M^2}\sum_{m=1}^{M}\sum_{k=1}^{M}|P_m - P_k| \tag{4}$$

where O represents the observed wind power, and P_m and P_k denote the predicted power values from the ensemble size M. The network inputs consist of the 51 ensemble members of the 100-meter wind speed and the 100-meter wind direction, and the power curve converted wind power (EPS-PC). Power values are normalized to the range $[0,1]$ using the installed capacity of each wind farm in the BOZ. Wind direction is represented by its cosine transform. The network is trained separately for each wind farm using the full year of 2022. This training period is randomly split into 85% for weight optimization and 15% for validation. Independent testing is then performed on the 2023 data.

3.2 Benchmark: Member-by-Member postprocessing

To evaluate the performance of the proposed Transformer-based method against a classic approach, we apply the Member-by-Member (MBM) postprocessing proposed by Van Schaeybroeck and Vannitsem (2015). The MBM method applies a linear correction directly to each ensemble member by each lead time separately. For the m-th member of the raw ensemble, denoted as P_m, the calibrated output power $\tilde{P}_m$ is obtained via the following linear transformation:

$$\tilde{P}_m(t) = \alpha(t) + \beta(t)\overline{P}(t) + \tau(t)\left(P_m(t) - \overline{P}(t)\right) \tag{5}$$

where α is the bias correction parameter, and β represents the regression coefficient corresponding to the ensemble mean $\overline{P}$. The term $(P_m - \overline{P})$ represents the deviation of the individual member from the ensemble mean of the target variable, which is scaled by the coefficient τ to adjust the ensemble spread. The regression coefficients α, β, and τ are optimized by minimizing the CRPS over the training dataset. The MBM method in this study is implemented by the `Pythie` Python package (Demaeyer, 2022).

The MBM approach here uses raw EPS-PC ensembles as inputs. The standard MBM approach does not impose explicit bounds on the outputs, while power values are physically constrained between zero and rated capacity. To address this issue, the power values are transformed by $\ln(\frac{P/\text{capacity}}{1-P/\text{capacity}})$, which maps the normalized power to an unbounded range (Bremnes, 2006) for parameter optimization, and is then transformed back to the original scale to evaluate the CRPS loss.

4 Model verification

Power predictions are aggregated from each wind farm to the BOZ level for evaluation. The model predictions are first evaluated deterministically for the ensemble mean. In terms of bias (Figure 2a), the baseline EPS-PC model exhibits a persistent positive bias between approximately 7% to 11%. Both the Transformer and MBM reduce this systematic error and bring the mean bias close to zero. Averaged across all lead times, the Transformer exhibits a slight negative bias, whereas the MBM retains a small positive bias. Regarding Mean Absolute Error (MAE) (Figure 2b), the Transformer consistently reduces the errors by more than 30% compared to EPS-PC across all lead times, while it does not always have a lower MAE than MBM results. Furthermore, the Transformer model effectively mitigates the diurnal variability in the forecast errors that is highly evident in the raw EPS-PC forecasts.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

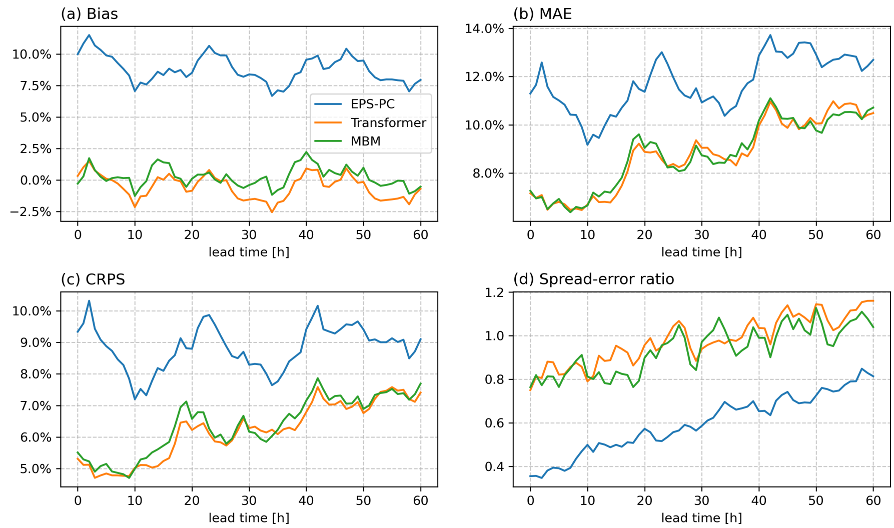

Figure 2. Verification scores of wind power forecasts for EPS-PC, Transformer and MBM methods, expressed as a percentage of BOZ installed capacity. The scores are evaluated on the 2023 test period.

In terms of probabilistic performance, the CRPS provides a joint assessment of calibration and sharpness. As shown in Figure 2c, the CRPS indicates improved overall forecast quality after postprocessing by both MBM and the Transformer. The Transformer method results in lower CRPS values than MBM across most of the lead times, while the improvement regarding MBM is limited. Overall, the major improvement of the Transformer method over the benchmark MBM method lies in intraday forecast periods.

Ensemble calibration is further illustrated by the Spread-Error Ratio (Figure 2d) and rank histograms (Figure 3). Raw EPS-PC suffers from severe under-dispersion, where the ratio starts as low as 0.3 and only reaches roughly 0.7 at later lead times. Both the Transformer model and the MBM demonstrate an improvement in resolving this under-dispersion, but some under-dispersion still remains. In rank histograms, a perfectly calibrated model should produce an approximately uniform distribution across ranks, whereas a U-shape indicates systematic errors and insufficient forecast spread. As illustrated in Figure 3, EPS-PC exhibits a pronounced sloped-shaped distribution with a large peak at the lowest rank. This pattern indicates severe under-dispersion combined with a systematic over-forecasting bias, meaning the observed wind power frequently falls below the lowest member of the raw ensemble. Both Transformer and MBM reduce this extreme peak but retain visible peaks at both ends. Transformer results are closer to the uniform reference line.

Compared with the linear MBM approach, one advantage of the Transformer is its ability to easily incorporate additional predictors and learn nonlinear inter-variant relationships. To further illustrate this capability, we investigate the wind-direction-dependent model performance (Figure 4). Southwesterly flow is dominant in this region, and the BOZ layout has been designed with this prevailing climatology. As a result, wake effects under southwesterly winds are relatively weak, leaving fewer systematic errors for the post-processing models to correct. In contrast, the baseline EPS-PC exhibits higher bias and CRPS when

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

Wind Energy Science Discussions

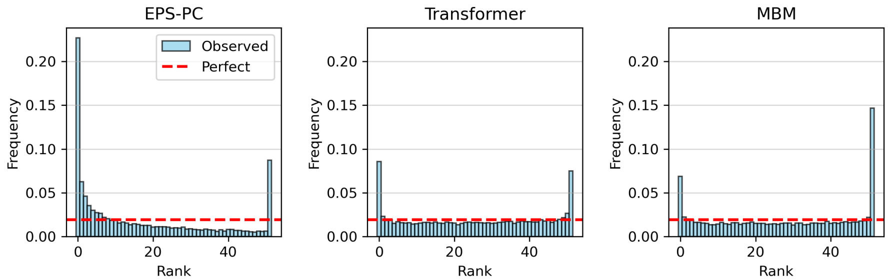

Figure 3. Rank histograms for the postprocessing methods, evaluated on the 2023 test period.

the wind originates from the Northwest, Northeast, and Southeast. A plausible explanation is that EPS-PC does not account for turbine-interaction effects, leading to overestimated wind speeds and, consequently, power production under these stronger wake conditions. While both the MBM and the Transformer reduce these errors in all directions, an additional Transformer test utilizing only a single variable (wind power, denoted as Transformer-1var) is included to indicate the Transformer's skill in addressing directional errors under the same input configuration as the MBM. Averaging the bias over all lead times and partitioning the results by wind direction reveals that the Transformer-1var test yields a directional distribution shape similar to the MBM profile (Figure 4a); both models exhibit more pronounced biases in the NW, NE, and SE directions than in the SW direction. However, the full Transformer model, which integrates directional features, provides a bias distribution that aligns much closer to zero. Although the differences in CRPS among the models are small across wind directions, with the full Transformer still showing a slight advantage (Figure 4b), the directional bias pattern provides clearer evidence of its ability to address wake-related errors.

To further demonstrate the effectiveness of the Transformer method in mitigating wake-related errors, we investigate meteorological conditions in which wake effects are prominent. Wake effects can influence power production across a wind speed range of 5 to 15 m/s, with the most significant power deficits typically occurring around 10 m/s (Barthelmie and Jensen, 2010). As a representative example of these specific conditions, we present a case on 9 May 2023, where the ensemble wind speeds are consistently within this critical interval (Figure 5a). Because the original EPS model lacks wind farm parameterization, the EPS wind prediction overestimates the wind speed, thereby overpredicting wind power in EPS-PC (Figure 5b). In comparison, although the Transformer is not given explicit details on wind farm layouts, it successfully calibrates the prediction distribution to cover the actual values by learning bias patterns associated with these specific weather conditions (Figure 5c). While MBM attempts to improve the result by appropriately increasing the ensemble spread, it still suffers from an overall overestimation, failing to reach the accuracy of the Transformer (Figure 5d).

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

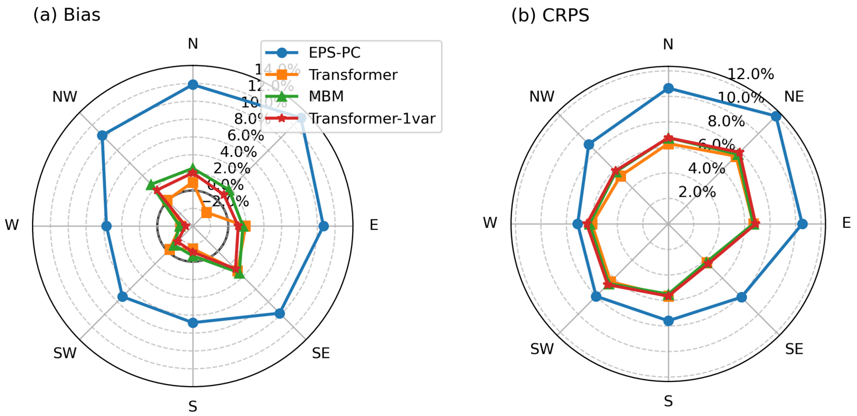

Figure 4. Bias and CRPS of wind power forecasts across different wind direction sectors for EPS-PC, MBM, Transformer, and Transformer-1var test, as a percentage of installed capacity, verified against wind power production data aggregated at BOZ level. The solid circle indicates the zero bias.

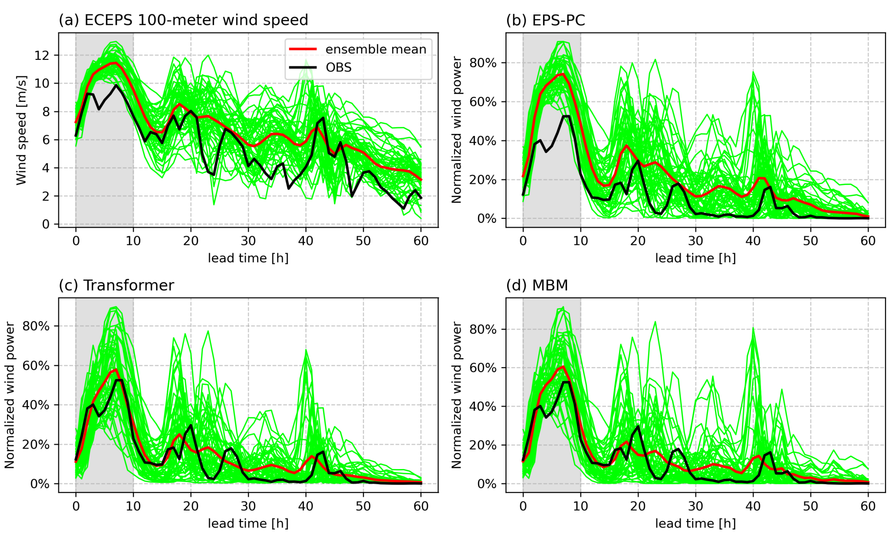

Figure 5. A case on 2023-05-09 for wind speed and power predictions by each model. The gray shadow indicates where the wake effect is prominent. In figure (a), the wind speed observation is the mean of SCADA data at several anonymous wind farms, and the ensemble forecasts are the mean of all wind farms.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

5 Predictability of ramping events

Wind power ramping refers to rapid power changes within a short period. Several characteristics of ramping events are typically of interest: magnitude (the power difference), duration (the time interval over which the change occurs), and direction (up-ramps versus down-ramps) (Gallego-Castillo et al., 2015). A ramping event is identified when the ramping magnitude exceeds a predefined threshold within the selected time interval. In operational practice, power ramping is important for power dispatch and decision-making. Investigating ramp predictability therefore provides an additional perspective on model performance.

The causes of wind power ramping events can be attributed to meteorological and non-meteorological factors. Meteorological drivers include rapid wind speed fluctuations and turbine cut-out controls during extreme high-wind conditions, such as storms (Smet et al., 2019, 2023). Non-meteorological factors concern manual power curtailments for system operations. Grid operators may execute decremental bids to deliberately curtail power output, which can also induce power ramping. We excluded all date times associated with such bids from our dataset.

This study focuses on hourly ramping, i.e., the power change at the minimum temporal resolution of the dataset. For each ensemble member, the hourly ramp is computed as the power difference between consecutive hours, which yields an ensemble of ramping predictions derived from the ensemble power forecasts. According to Elia, a 30% power change within one hour is treated as a ramping event (Elia, 2018; Gonzalez, 2025). In addition to the 30% ramping threshold, we also discuss the 15% threshold to investigate the predictability of small but more frequent ramping events. To appropriately assess model performance in practical applications, we evaluate the ramping predictability separately for the intraday (lead time 0-24 hours) and day-ahead (lead time 24-48 hours) horizons.

5.1 Ramping statistic analysis

The relative frequency of ramping events during the verification period is first examined for both observations and models. For the observations, the relative frequency is calculated as the ratio of cases where the power ramping exceeds the specific ramping threshold to the total number. For the ensemble predictions, the relative frequency is derived by first calculating the occurrence ratio across all valid time steps for each individual ensemble member, and subsequently averaging these frequencies across all members to obtain the ensemble mean frequency. A well-calibrated prediction should have a frequency close to that of observations. Figure 6 illustrates the relative frequency of observed and predicted ramping events for both intraday and day-ahead forecasting horizons. Across all thresholds, the baseline EPS-PC model significantly underestimates the occurrence of both $\pm$15% and $\pm$30% ramping events. In comparison, both the MBM and the Transformer tend to overestimate the ramping frequencies across all analyzed thresholds. Extending the forecasting horizon from intraday to day-ahead causes a comparable reduction in predicted ramping frequencies across all models. This feature generally reflects the increasing uncertainty at longer lead times, where forecasting models tend to produce smoother ensembles with attenuated ramping intensities and less precise timing (Yan et al., 2022). Despite this overall frequency attenuation at the longer forecast lead time, the Transformer maintains a distinct advantage in relative frequency over the raw EPS-PC across all thresholds, though its improvement over the MBM remains limited.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

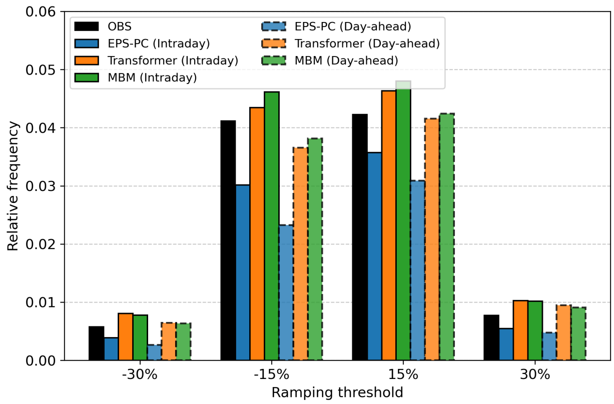

Figure 6. Relative frequency of observed and predicted ramping events for the intraday and day-ahead forecasting horizons in the test year 2023.

The previous analysis evaluates the consistency between predicted and observed ramping in terms of occurrence frequency, but it does not establish an event-to-event correspondence between predictions and observations. To further examine model behavior, we analyze the distribution of predicted ramping magnitudes conditional on observed ramping events. Figure 7 shows the distributions of intraday and day-ahead predictions for cases where ramping events are observed. For all evaluated models, the median predicted magnitudes remain below the corresponding ramping thresholds. In many cases, only a limited fraction of ensemble members exceeds the threshold, indicating a common underestimation of ramping intensity across the ensemble. This limitation is particularly severe for the EPS-PC model, which exhibits the most compressed ensemble spread. While the median of the ramping values for Transformer and MBM remains almost similar to EPS-PC, their expansion spreads to cover more than the thresholds. This indicates that postprocessing is capable of generating sufficient variance to handle large ramping events. A comparison between forecasting horizons shows that intraday predictions tend to be closer to the ramping thresholds than day-ahead forecasts. This indicates that shorter lead times provide some improvement in representing ramping intensity, although underestimation remains present. Overall, these results suggest that even when ramping signals are indicated in the ensemble predictions, many members still fail to reach the corresponding threshold magnitude, which limits the confidence of the ensemble to represent the severity of ramping events. While the postprocessing methods mitigate this issue by maintaining appropriate spreads, capturing the full magnitude of extreme power ramps remains a significant challenge.

5.2 Probabilistic verification with error-tolerance

For probabilistic verification of binary events, the traditional Brier Score (BS) relies on exact point-to-point matching. For ramping events, this requirement can be too restrictive because forecasts that capture the event signal but are slightly displaced in time or magnitude can receive a large penalty. The Fractions Skill Score (FSS) partly addresses this issue in spatial precipitation verification by averaging within a neighborhood window (Necker et al., 2024). However, that formulation is not directly suited to one-dimensional wind power ramping time series, where the timing and intensity of discrete events are operationally

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

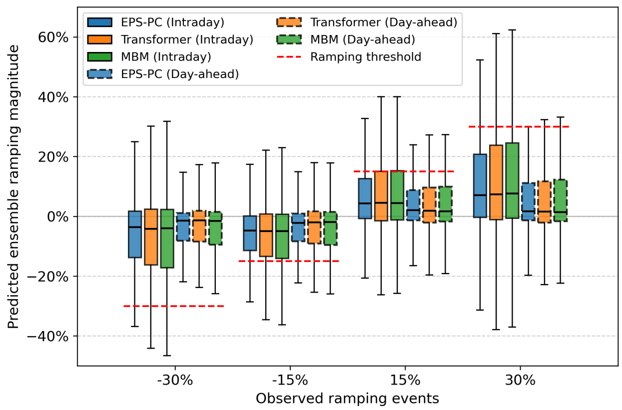

Figure 7. Distributions of predicted ensemble ramping magnitudes for the models under various observed ramping events.

important. Motivated by this difference, we propose a tolerance-based probabilistic verification framework for probabilistic wind power ramping that combines a fractional event definition with a temporal matching rule.

Probabilistic verification of ramping events begins with estimating event probabilities from ensemble forecasts. In ensemble prediction systems, the probability of a ramping event is usually defined as the proportion of ensemble members that exceed a predefined ramping threshold, i.e., the ensemble mean of binary indicators (Worsnop et al., 2018; Muschinski et al., 2022). However, as illustrated by Figure 7, in practical forecasting situations, ramping signals may be present in the forecasts while the predicted magnitude is slightly underestimated. Such rigid binary classification dismisses these useful near-threshold signals as complete misses, unfairly penalizing members that correctly predict the timing and structure but slightly miss the peak magnitude.

To mitigate the magnitude penalty where rigid binary thresholds penalize acceptable forecasts that slightly underestimate the peak intensity, we first convert the binary definition of ramping events into continuous fractional labels for both predictions and observations. Let r denote the absolute magnitude of a power ramp (either predicted or observed), C the target threshold, and ϵ the acceptable magnitude tolerance buffer. We define a fractional mapping:

$$z = \max\left(0, f\left(r, C, \epsilon\right)\right) \quad (6)$$

where the f is a function to allocate tolerance by determining the distance from the predicted value to the target threshold. Here, we use a linear mapping function for fractional processing, so that values closer to the threshold within the tolerant buffer linearly receive a higher fraction, while values farther from the threshold receive a lower fraction:

$$f = \min\left(1, \frac{r - (C - \epsilon)}{\epsilon}\right) \quad (7)$$

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

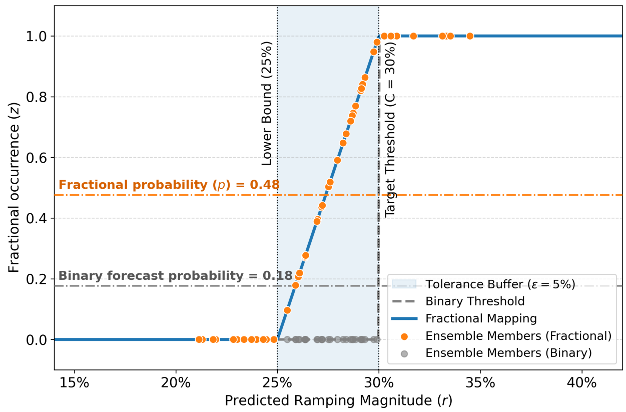

Figure 8. Illustration of the fractional mapping incorporating a magnitude buffer applied to an ensemble forecast. The binary probability (i.e., the probability without a buffer, gray dashed line) strictly penalizes near-miss ensemble members. In contrast, the proposed fractional mapping (blue solid line) assigns fractional occurrences within a defined magnitude tolerance buffer ($\epsilon = 5\%$, shaded area). The magnitude buffer mitigates the magnitude penalty and yields a more representative forecast probability (0.48) compared to the probability without a buffer (0.18).

The forecast fractional probability p_t of a ramping event with threshold C at lead time t is then aggregated by the fractional values of all ensemble members m:

$$p_t = \frac{1}{M} \sum_{m=1}^{M} z_{m,t} \tag{8}$$

For the evaluation target, the observation is also converted to a fractional value of $o_t \in [0,1]$ with the same mapping. To intuitively illustrate this fractional probability, Figure 8 indicates the application of the continuous fractional mapping on a simulated 51-member ensemble forecast. In a scenario where the predicted magnitudes are close but lower than the threshold of 30%, the traditional binary approach strictly assigns a score of zero to the majority of members. Conversely, by assigning proportional fractional mapping z to members falling within the adjustable tolerance buffer (e.g., $\epsilon = 5\%$), the proposed method integrates these near-threshold signals into the overall probability estimation.

In addition to the magnitude penalty, the temporal "double penalty" is another issue: although slight temporal displacements of predicted ramping events relative to actual observations are often operationally treated as useful predictions, they are penalized twice (once as a false alarm and once as a miss) under rigid point-to-point evaluation. In traditional binary verification frameworks, ramping events are identified within time windows and validated to determine whether observed and predicted events occur within that window (Worsnop et al., 2018; Muschinski et al., 2022). However, this approach has two limitations. First, the exact timing of the ramping event is not specified, particularly when the window is wide (Lee et al., 2022). Second, the definition of a ramping event becomes ambiguous, as the event duration can be conflated with the width of the time window.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

To address potential timing discrepancies between observation and prediction, we evaluate the forecast performance by scanning within a temporal tolerance buffer Ω to seek the optimal alignment between the prediction and observation, instead of demanding an exact temporal match. By integrating this temporal buffer mechanism with the fractional occurrences computed above, we formulate a comprehensive error metric. With the fractional probability $p_t \in [0,1]$ and observation $o_t \in [0,1]$, we define the error $\beta(t)$ at time t as the minimum squared difference derived from three complementary evaluation perspectives:

$$\beta(t) = \min\left(\mathcal{L}_{\text{Exact}}(t), \mathcal{L}_{\text{FAM}}(t), \mathcal{L}_{\text{MM}}(t)\right) \tag{9}$$

The three components correspond to the rigorous baseline, false alarm mitigation, and miss mitigation, respectively:

$$\begin{cases} \mathcal{L}_{\text{Exact}}(t) = (p_t - o_t)^2 \\ \mathcal{L}_{\text{FAM}}(t) = (p_t - \max\limits_{\tau\in\Omega_t} o_\tau)^2 \\ \mathcal{L}_{\text{MM}}(t) = (\max\limits_{\tau\in\Omega_t} p_\tau - o_t)^2 \end{cases} \tag{10}$$

where Ω_t represents the temporal window centered at t. The minimization logic interprets the forecast performance through the most favorable valid mechanism:

- **Exact Match ($\mathcal{L}_{\text{Exact}}$):** This term preserves the classical point-to-point Brier Score evaluation. It ensures that when a predicted event perfectly aligns with an observation at the exact time t (a perfect *Hit*), or when neither a forecast nor an event occurs (*True Negative*), the metric accurately rewards this exact correspondence without relying on the temporal buffer.

- **False Alarm Mitigation ($\mathcal{L}_{\text{FAM}}$):** This term addresses the scenario where the model predicts a ramping event ($p_t \approx 1$), but no event occurs at the exact moment t ($o_t = 0$). In a traditional metric, this is a *False Alarm*. However, if an event effectively occurs within the neighborhood ($\max o_\tau \approx 1$), this term reduces the error towards zero, reclassifying the error as a temporally displaced *Hit*.

- **Miss Mitigation ($\mathcal{L}_{\text{MM}}$):** This term addresses the scenario where a ramp actually occurs ($o_t \approx 1$) but is not predicted at t ($p_t = 0$). Traditionally penalized as a *Miss*, this formulation checks if the model anticipates the event within the buffer ($\max p_\tau \approx 1$). If the signal exists nearby, the penalty is waived, acknowledging that the model successfully captured the event with a minor timing offset.

To illustrate the concept of the proposed temporal buffer framework, Figure 9 visualizes the three complementary evaluation mechanisms under slight temporal phase shifts. In traditional point-to-point verification, the misalignments illustrated would lead to a significant penalty at the evaluation time t, which penalizes the model for minor timing errors. By applying the temporal buffer Ω_t, the $\mathcal{L}_{\text{FAM}}$ term effectively searches for actual observations to justify an early or delayed warning (Figure 9b),

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

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

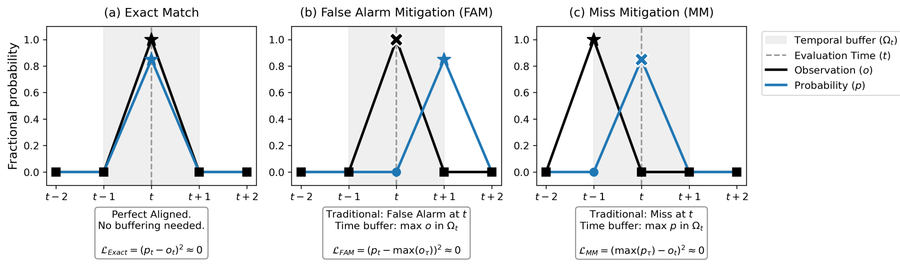

Figure 9. Illustration of verification with the temporal buffer framework. (a) Exact Match evaluation. (b) False Alarm Mitigation (FAM) and (c) Miss Mitigation (MM) mechanisms evaluated within the temporal buffer Ω_t (shaded region). The cross marker indicates that the observation or prediction does not reach the ramping threshold at the evaluation time t. The star marker indicates a signal within Ω_t that should be considered a correct prediction with tolerance.

while the $\mathcal{L}_{\mathrm{MM}}$ term searches the forecast horizon to cover an observed event (Figure 9c). This dynamic minimization mathematically neutralizes the "double-penalty" effect, ensuring the final error metric accurately reflects the model's operational usefulness.

The aggregate Buffer Brier Score (BBS) is the mean over the evaluation period T:

$$\mathrm{BBS} = \frac{1}{T}\sum_{t=1}^{T}\beta(t) \tag{11}$$

It should be noted that BBS is not a proper scoring rule in the formal statistical sense (Gneiting and Raftery, 2007). Because the metric allows temporal and magnitude matching through a minimization step, a forecast may receive partial credit without matching the observation exactly at the evaluated time. Therefore, BBS is not intended to assess probabilistic calibration in the same way as the standard Brier Score, nor is it intended to be used arbitrarily as a loss function. Instead, it should be interpreted as an application-oriented diagnostic metric to understand event-capture skill under explicitly defined tolerances. Only when no buffer is set is the BBS equivalent to the standard BS that satisfies the scoring propriety.

To quantify the relative improvement, the Buffer Brier Skill Score (BBSS) is computed. Similar to the FSS, we use a reference baseline $\mathrm{BBS}_{\mathrm{ref}}$ that represents a scenario where the predictions and observations have strictly zero temporal overlap within the buffer window (i.e., maximal false alarms and misses). For this worst-case reference, the error at time t is directly computed as the sum of their independent squares:

$$\mathrm{BBS}_{\mathrm{ref}} = \frac{1}{T}\sum_{t=1}^{T}(p_t^2 + o_t^2) \tag{12}$$

The final BBSS is calculated as:

$$\mathrm{BBSS} = 1 - \frac{\mathrm{BBS}}{\mathrm{BBS}_{\mathrm{ref}}} \tag{13}$$

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

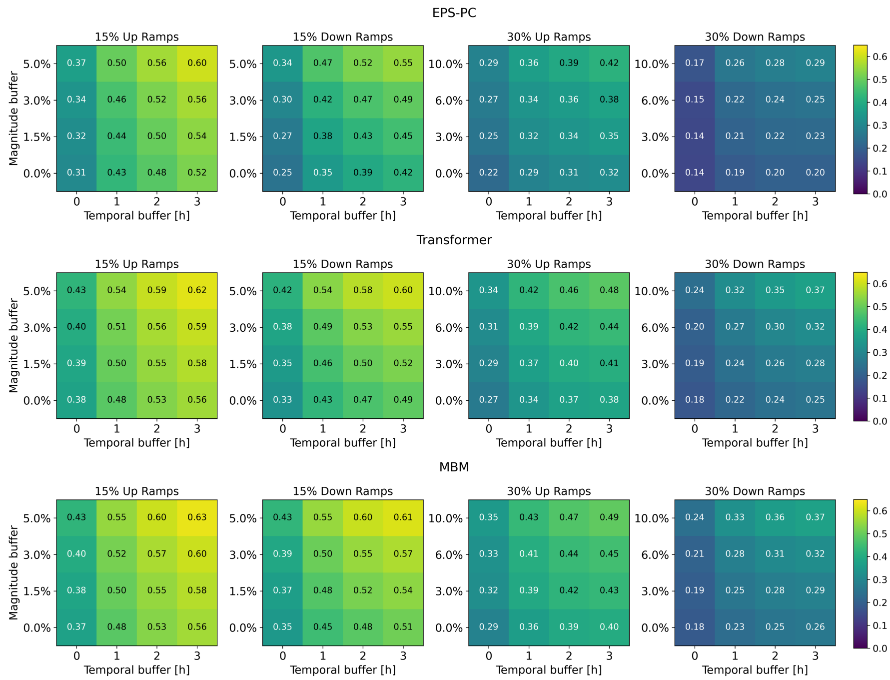

Figure 10. Buffer Brier Skill Score (BBSS) heatmaps for the EPS-PC (top), Transformer (middle), and MBM (bottom) intraday predictions. The scores are evaluated across varying temporal buffers and magnitude buffers (relative to the ramping thresholds) for 15% and 30% up- and down-ramps.

This formulation ensures the skill score in the range $[0, 1]$, where higher values indicate better performance under the chosen validation settings. By combining magnitude and temporal buffers, the metric provides a tolerance-aware view of ramping-event predictability and helps identify buffer settings that may be operationally relevant. Figure 10 illustrates the BBSS for EPS-PC, Transformer, and MBM models across various magnitude and temporal buffer combinations for intraday forecasts. The day-ahead results are included in the appendix (Figure A1). The magnitude buffer is relative to the ramping thresholds. The evaluation is separated into 15% and 30% threshold events for both up-ramps and down-ramps.

Under the rigid evaluation criteria, meaning zero temporal and zero magnitude tolerances, all evaluated models exhibit relatively low skill scores. For instance, the baseline BBSS for 30% down-ramps is 0.14 for the EPS-PC and 0.18 for both MBM and Transformer. Even without applying any buffer, the Transformer and MBM already demonstrate an improvement in ramping prediction skill compared to the raw forecast, while the ramping predictability by the Transformer is not necessarily better than MBM by all ramping thresholds.

Although the Transformer exhibits slightly better skill in intraday power prediction compared to the MBM (Figure 2c), its advantage does not necessarily extend to ramping prediction. A comparison reveals that the Transformer and the MBM perform comparably across most buffer combinations, with the MBM demonstrating a marginally higher predictability for 30% up-ramping events (Figure 10). This behavior can be attributed to their different optimization mechanisms. The MBM performs

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

postprocessing independently for each lead time. Lacking the temporal context between adjacent steps, it is more prone to generating large power gradients, which correspond to stronger ramping signals. Conversely, the Transformer optimizes the wind power prediction across all lead times simultaneously. To minimize the global CRPS under temporal power uncertainty, the Transformer tends to adopt a smoothing strategy. While beneficial for overall power prediction, this temporal smoothing is inherently less favorable for capturing sharp ramping representations.

As the tolerances increase, the performance gap between the postprocessed models and the raw EPS-PC gradually narrows, which is particularly evident for the 15% ramping events. This convergence suggests that although the baseline ensemble members often lack precise temporal synchronization, their ensemble spread still contains meaningful ramping signals that are slightly displaced in time and magnitude. These signals can be partially uncovered by the BBSS framework when tolerance is allowed. Furthermore, up-ramping events demonstrate higher predictability than down-ramping events across all models, which agrees with the verification of the operational deterministic LAMs at RMI (Meng et al., 2026). This difference in predictability between the two directions is more pronounced for the 30% ramping events compared to the 15% cases.

These scores reflect the limits of traditional point-to-point verification, where minor phase shifts or slight intensity underestimation can lead to forecasting failures. As the temporal and magnitude tolerances are applied, a consistent and significant improvement in BBSS is observed across all ramping events for every model. Incorporating a temporal buffer of just 1 to 2 hours recovers a substantial portion of forecasts affected by slight time shifts, suggesting that many operational false alarms and misses are in fact the result of minor timing offsets. Similarly, the models frequently capture the structural trend of the ramping events but slightly underestimate the peak intensity. When both buffers are combined, the models achieve their highest predictability.

From an operational perspective, the proposed error-tolerant verification framework provides more informative guidance for decision-making than rigid point-to-point evaluation metrics. By quantifying forecast skill under controlled temporal and magnitude tolerances, it helps grid operators define warning thresholds that better reflect the practical predictability of ramping events. Identifying where forecast skill begins to level off also provides a basis for balancing system security against economic efficiency, avoiding unnecessarily wide alert windows that may increase false alarms. Furthermore, the observed differences in predictability between up-ramping and down-ramping events suggest that different warning thresholds or operational strategies may be appropriate for the two ramping directions. Overall, the proposed framework can support grid operators in refining scheduling strategies and tailoring warning criteria to the practical limits of forecast predictability.

6 Conclusions

This study investigates probabilistic wind power prediction and ramping verification for offshore wind farms in the Belgian Offshore Zone. The analysis uses operational ensemble forecasts from the ECMWF Ensemble Prediction System and is verified against actual power data from the Belgian TSO. A Transformer-based member-wise postprocessing framework is trained for each wind farm and evaluated against raw power-curve-derived ensembles and a Member-by-Member benchmark.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

For probabilistic power prediction, the postprocessed ensembles show better agreement with observations in terms of calibration and overall distribution quality. Relative to both EPS-PC and MBM, the Transformer achieves a near-zero bias and reduces MAE for the ensemble mean, especially for intraday lead times. Furthermore, it demonstrates superior probabilistic skill by providing the lowest CRPS and enhanced reliability. The model can mitigate errors that potentially relate to wake effects. These results suggest that Transformer-based postprocessing is a promising extension for site-specific probabilistic wind power forecasting.

For ramping verification, the statistical analysis characterizes how event frequencies vary with threshold, direction, and lead time. The baseline EPS-PC model significantly underestimates the relative frequency of ramping events across all thresholds, particularly for ±30% ramping. In contrast, the Transformer and MBM postprocessing provide a substantial improvement, with their predicted frequencies aligning much more closely with observed values. While the predicted frequency decreases from the intraday to the day-ahead forecasting horizon for all models, the Transformer generally maintains better agreement with the observed event frequencies than the baseline methods. This tendency is apparent for the ±30% ramping events at longer lead times.

The study also argues that conventional point-to-point verification can understate forecast utility for ramping events. Small timing offsets and minor magnitude differences can trigger double penalties, causing otherwise informative forecasts to be classified as failures. To address this issue, we implement a tolerance-based verification framework that allows event matching under temporal and magnitude uncertainty. Based on this concept, we introduce the Buffer Brier Skill Score to characterize the probabilistic ramping predictability of the models under temporal and magnitude buffers.

Within this framework, the evaluation across models and event definitions demonstrates that forecast quality depends heavily on event magnitude, ramping direction, and lead time. Prediction skill is generally higher for 15% ramping events than for 30% events, and up-ramps are more predictable than down-ramps in both intraday and day-ahead verification. Under strict settings without temporal or magnitude tolerance, all models show limited absolute skill; however, both postprocessing methods provide clear improvements over the raw EPS-PC. Notably, while the Transformer effectively enhances the baseline, its improvement in ramping predictability over the benchmark MBM remains limited. This comparison further indicates that the Transformer's superior performance in overall wind power prediction does not necessarily translate into a definitively superior ramping prediction. As temporal and magnitude buffers are introduced, the verification skill increases across all approaches, indicating that part of the apparent forecast error is linked to timing displacements and moderate intensity underestimations rather than complete event misses. The performance gap between the postprocessing results and the EPS-PC gradually narrows at larger tolerances.

For future work, a promising direction is to incorporate the concept of verification tolerance into model optimization. Specifically, future research could explore the development of differentiable proper loss functions and their integration with extreme-event optimization strategies. Such an approach may help reduce the tendency to underestimate the intensity of extreme ramping while preserving the probabilistic consistency of the forecasts.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

Appendix A: BBSS for day-ahead forecasts

The evaluation of the day-ahead forecasting horizon reveals a consistent reduction in predictability compared to the intraday results. This general reduction indicates the increased inherent forecasting uncertainty associated with longer lead times. Despite this overall lower performance, the relative ranking among the evaluated models remains consistent with the intraday analysis. The Transformer model performs roughly the same as the MBM and outperforms EPS-PC across all temporal and magnitude buffer combinations. Similar to the intraday discussions, applying temporal and magnitude tolerances partially recovers the forecasting skill for all models in the day-ahead horizon.

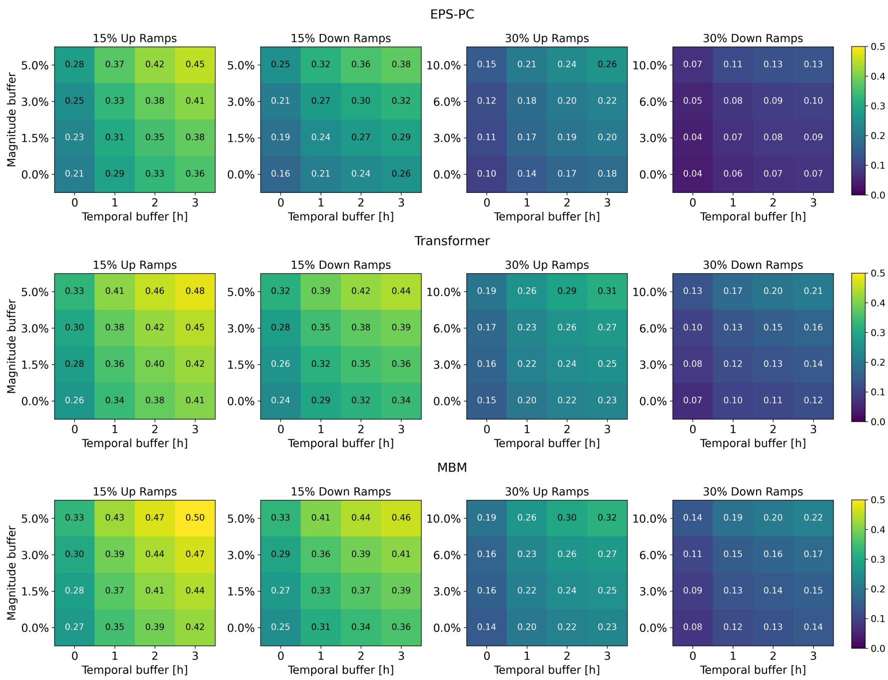

Figure A1. Buffer Brier Skill Score (BBSS) heatmaps for the EPS-PC (top) and Transformer (middle) and MBM (bottom) day-ahead predictions. The scores are evaluated across varying temporal buffers and magnitude buffers (relative to the ramping thresholds) for 15% and 30% up- and down-ramps.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

Code availability. The code for the Transformer model is available upon request by contacting the corresponding author. The code used to perform Member-by-Member postprocessing is available in the Python package `Pythie` (https://github.com/Climdyn/pythie, Demaeyer, 2022).

Data availability. The hourly ECMWF forecast data are not publicly accessible (https://www.ecmwf.int/en/forecasts/accessing-forecasts, last access: 15 July 2026), but 3-hourly ensemble data at 0.25 degree spatial resolution is currently available as open data (https://www.ecmwf.int/en/forecasts/datasets/open-data, last access: 15 July 2026). The power production data of each wind farm at BOZ is provided privately by Elia; however total power production data for the whole BOZ and the power decremental bids information are publicly available in Elia's Open Data Platform (https://opendata.elia.be/explore/dataset/ods031/information/, last access: 15 July 2026).

Author contributions. RM, GS, and JV identified the problem and contributed to the theoretical and conceptual development of the methods. RM, GS, and JV interpreted the results. RM drafted the manuscript, and GS, JV, and HT revised the paper. JV and PT contributed to the project funding and management.

Competing interests. The authors declare that they have no conflict of interest.

Financial support. This research has been supported by the Belgian Science Policy Office (BELSPO) through the E-TREND project (B2/223/P1/E-TREND).

Acknowledgements. This work is part of the Extensible Tools for Renewable ENergy Decision making (E-TREND) project. The authors acknowledge Belgian transmission system operator Elia for providing the 15-minute level power data of individual wind farms to support this work. The authors also acknowledge that Aaron Van Poecke provided demo code for the Transformer implementation; and Bert Van Schaeybroeck and Jonathan Demaeyer helped with `Pythie` package for the MBM solution.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

References

Barthelmie, R. J. and Jensen, L.: Evaluation of wind farm efficiency and wind turbine wakes at the Nysted offshore wind farm, Wind Energy, 13, 573–586, https://doi.org/10.1002/we.408, 2010.

Barthelmie, R. J., Hansen, K., Frandsen, S. T., Rathmann, O., Schepers, J., Schlez, W., Phillips, J., Rados, K., Zervos, A., Politis, E., et al.: Modelling and measuring flow and wind turbine wakes in large wind farms offshore, Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology, 12, 431–444, https://doi.org/10.1002/we.348, 2009.

Borgers, R., Meyers, J., and van Lipzig, N.: Energy production and inter-farm wake losses in future North Sea wind farms, Environmental Research Letters, 20, 074 036, https://doi.org/10.1088/1748-9326/add8a2, 2025.

Bouallègue, Z. B., Weyn, J. A., Clare, M. C., Dramsch, J., Dueben, P., and Chantry, M.: Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers, Artificial Intelligence for the Earth Systems, 3, e230 027, https://doi.org/10.1175/AIES-D-23-0027.1, 2024.

Bremnes, J. B.: A comparison of a few statistical models for making quantile wind power forecasts, Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology, 9, 3–11, https://doi.org/10.1002/we.182, 2006.

Demaeyer, J.: Pythie, Zenodo [code], https://doi.org/10.5281/zenodo.7233538, 2022.

Demaeyer, J., Van den Bergh, J., Schicker, I., Tabari, H., Schmid, F., Schroedter-Homscheidt, M., Vannitsem, S., and De Cruz, L.: Workshop report: EUMETNET-E-TREND joint workshop on postprocessing, forecasting and nowcasting for renewable energy, Journal of the European Meteorological Society, 5, 100 044, https://doi.org/10.1016/j.jemets.2026.100044, 2026.

Eide, S. S., Bremnes, J. B., and Steinsland, I.: Bayesian model averaging for wind speed ensemble forecasts using wind speed and direction, Weather and Forecasting, 32, 2217–2227, https://doi.org/10.1175/WAF-D-17-0091.1, 2017.

Elia: Analysis, benchmark and mitigation of storm and ramping risks from offshore wind power in Belgium, Tech. rep., Elia, https://www.elia.be/-/media/project/elia/elia-site/electricity-market-and-system---document-library/balance-responsible-party-and-system-imbalance/2018/2018-study-report-on-offshore-integration_en.pdf, last access: 15 July 2026, 2018.

Finn, T. S.: Self-Attentive Ensemble Transformer: Representing Ensemble Interactions in Neural Networks for Earth System Models, https://arxiv.org/abs/2106.13924, 2021.

Gallego-Castillo, C., Cuerva-Tejero, A., and Lopez-Garcia, O.: A review on the recent history of wind power ramp forecasting, Renewable and Sustainable Energy Reviews, 52, 1148–1157, https://doi.org/10.1016/j.rser.2015.07.154, 2015.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American statistical Association, 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Monthly weather review, 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Gonzalez, J.: Offshore Ramping Events Detector, https://innovation.eliagroup.eu/en/projects/offshore-ramping-events-detector, last access: 15 July 2026, 2025.

Haiden, T., Janousek, M., Vitart, F., Bouallègue, Z. B., and Prates, F.: Evaluation of ECMWF forecasts, including the 2023 upgrade, https://doi.org/10.21957/d47ba5263c, 2023.

Lee, J. C., Draxl, C., and Berg, L. K.: Evaluating wind speed and power forecasts for wind energy applications using an open-source and systematic validation framework, Renewable Energy, 200, 457–475, https://doi.org/10.1016/j.renene.2022.09.111, 2022.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

Meng, R., Smet, G., Van den Bleeken, D., Van Poecke, A., Tabari, H., Hellinckx, P., Termonia, P., and Van den Bergh, J.: Evaluating the Prediction of Wind Power Ramping Events in the Belgian Offshore Zone, Meteorological Applications, 33, e70 203, https://doi.org/10.1002/met.70203, 2026.

Messner, J. W., Pinson, P., Browell, J., Bjerregård, M. B., and Schicker, I.: Evaluation of wind power forecasts—An up-to-date view, Wind Energy, 23, 1461–1481, https://doi.org/10.1002/we.2497, 2020.

Murcia Leon, J. P., Koivisto, M. J., Sørensen, P., and Magnant, P.: Power fluctuations in high installation density offshore wind fleets, Wind Energy Science Discussions, 2020, 1–23, https://doi.org/10.5194/wes-6-461-2021, 2020.

Muschinski, T., Lang, M. N., Mayr, G. J., Messner, J. W., Zeileis, A., and Simon, T.: Predicting power ramps from joint distributions of future wind speeds, Wind Energy Science, 7, 2393–2405, https://doi.org/10.5194/wes-7-2393-2022, 2022.

Necker, T., Wolfgruber, L., Kugler, L., Weissmann, M., Dorninger, M., and Serafin, S.: The fractions skill score for ensemble forecast verification, Quarterly Journal of the Royal Meteorological Society, 150, 4457–4477, https://doi.org/10.1002/qj.4824, 2024.

Phipps, K., Lerch, S., Andersson, M., Mikut, R., Hagenmeyer, V., and Ludwig, N.: Evaluating ensemble post-processing for wind power forecasts, Wind Energy, 25, 1379–1405, https://doi.org/10.1002/we.2736, 2022.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Monthly weather review, 133, 1155–1174, https://doi.org/10.1175/MWR2906.1, 2005.

Smet, G., Van den Bergh, J., and Termonia, P.: Probabilistic storm forecasts for wind farms in the North Sea, Joint ALADIN-HIRLAM Newsletter, pp. 17–24, https://orfeo.belnet.be/handle/internal/7467, last access: 15 July 2026, 2019.

Smet, G., Van den Bergh, J., and Termonia, P.: Probabilistic storm forecasts for wind farms in the North Sea, Tech. rep., Copernicus Meetings, https://doi.org/10.5194/ems2023-505, 2023.

Sørensen, P., Koivisto, M., and Murcia, J. P.: Elia - MOG II System Integration: Public version, no. E-0203 in DTU Wind Energy E, DTU Wind Energy, Denmark, 2020.

Vallance, L., Charbonnier, B., Paul, N., Dubost, S., and Blanc, P.: Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric, Solar Energy, 150, 408–422, https://doi.org/10.1016/j.solener.2017.04.064, 2017.

Van den Bleeken, D., Smet, G., Van den Bergh, J., Dehmous, I., Degrauwe, D., Van Ginderachter, M., and Deckmyn, A.: Improving wind power forecasts in the Belgian North Sea with a wind farm parameterization and a neural network, Advances in Science and Research, 22, 59–67, https://doi.org/10.5194/asr-22-59-2025, 2025.

Van Poecke, A., Finn, T. S., Meng, R., Van den Bergh, J., Smet, G., Demaeyer, J., Termonia, P., Tabari, H., and Hellinckx, P.: Self-attentive transformer for fast and accurate postprocessing of temperature and wind speed forecasts, Artificial Intelligence for the Earth Systems, 4, 240 127, https://doi.org/10.1175/AIES-D-24-0127.1, 2025.

Van Schaeybroeck, B. and Vannitsem, S.: Ensemble post-processing using member-by-member approaches: Theoretical aspects, Quarterly Journal of the Royal Meteorological Society, 141, 807–818, https://doi.org/10.1002/qj.2397, 2015.

Voyant, C.: A Tolerance-Based Framework for Spatio-Temporal Forecast Validation Using the gamma-Index, https://arxiv.org/abs/2601.21004, 2026.

Wan, Y.-h.: Analysis of Wind Power Ramping Behavior in ERCOT, Tech. rep., NREL, 2011.

Worsnop, R. P., Scheuerer, M., Hamill, T. M., and Lundquist, J. K.: Generating wind power scenarios for probabilistic ramp event prediction using multivariate statistical post-processing, Wind energy science, 3, 371–393, https://doi.org/10.5194/wes-3-371-2018, 2018.

https://doi.org/10.5194/wes-2026-125
Preprint. Discussion started: 27 August 2026

Yan, J., Möhrlen, C., Göçmen, T., Kelly, M., Wessel, A., and Giebel, G.: Uncovering wind power forecasting uncertainty sources and their propagation through the whole modelling chain, Renewable and Sustainable Energy Reviews, 165, 112 519, https://doi.org/10.1016/j.rser.2022.112519, 2022.