https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

Accelerating regional wind energy assessment with deep-learning surrogates of WRF wind farm simulations

Tianxia Jia[1], Mike Optis[2], Adam H. Monahan[1], and Slim Ibrahim[1]

[1]University of Victoria, Victoria, British Columbia, Canada
[2]Veer Renewables, Courtenay, British Columbia, Canada

Correspondence: Tianxia Jia (tianxiajia@uvic.ca)

Abstract. Downwind wake effects are becoming increasingly important as wind farms grow larger and turbine capacities increase. Mesoscale weather models with wind farm parameterizations have emerged as a key tool in modelling long-distance wakes between wind farms. However, they can be too computationally expensive for evaluating dozens or more layouts considered in regional planning and optimization. In this study, we develop deep-learning surrogate models that reproduce the power losses and wind speed deficits caused by turbine layouts in mesoscale simulations at a fraction of the computational cost. The models combine atmospheric inputs from free-stream Weather Research and Forecasting (WRF) model simulations with turbine layouts to predict spatial power fields produced by WRF when the wind farm parameterization is activated. First, convolutional neural networks (U-Net) are developed as deterministic surrogates and achieve strong accuracy on two unseen scenarios. Second, diffusion-based models are developed to generate predictive ensembles and quantify uncertainty, including a residual diffusion model that learns the error of a deterministic U-Net prediction. Overall, the all models show a strong ability to predict wind power, both on a per-grid cell basis and aggregated across wind farms. The U-Net model strength shows sensitivity to the predictand (capacity factor vs. normalized power output), the combination of predictors (wind speed, wind direction, turbulence, and temperature), the number of training scenarios, and the type of loss function. Among the probabilistic models, DDPM provides the best calibrated ensembles, whereas residual diffusion yields more accurate point predictions and better farm-level bias control. These results demonstrate that deep-learning surrogates can enable rapid and cost-effective evaluation of candidate wind farm layouts, while also supporting uncertainty-aware planning-stage assessment.

1 Introduction

As wind farms grow in size and turbine ratings increase, external wake effects play an increasingly important role in wind energy assessment and planning. In many regions, new projects must be evaluated in the presence of existing or concurrently planned neighbouring farms, so developers must consider not only within-farm losses but also wake interactions across multiple sites. This creates a need for layout-aware models capable of evaluating many configurations under realistic atmospheric conditions.

Traditional engineering wake models are attractive due to their speed and simplicity, but they have recognized limitations in representing long-range external wakes and their dependence on atmospheric stability (Fischereit et al., 2022; zum Berge

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

et al., 2024). These limitations are particularly consequential for regional planning, where location-specific atmospheric conditions, terrain, and inter-farm spacing can strongly influence wake propagation. For such applications, mesoscale numerical weather prediction models with wind farm parameterizations (WFPs) provide a physically more consistent alternative by representing the interaction between large-scale atmospheric conditions and turbine-induced effects, including momentum sinks and turbulence kinetic energy sources (Fischereit et al., 2021; Schicker et al., 2023; Pryor et al., 2023). As a result, mesoscale WFP-based approaches are increasingly being used for regional planning purposes.

Among these approaches, the Weather Research and Forecasting (WRF) model with WFP, hereafter WRF-WFP, has become a useful tool for simulating layout-dependent wake impacts under realistic meteorological conditions. However, this improved physical realism comes at substantial computational cost, especially when many candidate layouts must be screened during planning and optimization. This motivates surrogate models that can emulate expensive simulations at a much lower cost. Surrogate modelling has been widely used in wind energy, including turbine-level emulation of aerodynamic and aeroelastic responses such as loads and fatigue under varying inflow conditions (Dimitrov, 2019; Slot et al., 2020; Ti et al., 2020), as well as farm-level prediction of annual energy production and wake-related losses (Padrón et al., 2019; King et al., 2020; Bleeg, 2020; Li et al., 2024; Wang et al., 2024).

Recent studies show that deep learning models perform well for wind speed and power prediction by capturing complex nonlinear relationships (Wang et al., 2021; Annau et al., 2023; Abdelsattar et al., 2025; Liu et al., 2025; Rajaperumal and Christopher Columbus, 2025). However, most existing work focuses on operational forecasting from observations or SCADA data (Pandit and Wang, 2024), where the goal is to predict future power production at individual sites. In contrast, planning-stage analysis poses a different problem, where the objective is to emulate the layout-dependent power response produced by a physics-based simulation model under many candidate turbine configurations over a broader geographical region

This distinction is important because the layout assessment task is not a time-series forecasting problem, but a spatial surrogate modelling problem. The inputs consist of gridded atmospheric fields together with turbine layout information, and the output is a gridded waked wind speed or power field. This setting naturally lends itself to a computer-vision formulation, in which convolutional neural network (CNN) architectures learn a mapping from atmospheric covariate and layout to layout-dependent power responses. CNN-based models have shown strong performance in wind-related prediction tasks, particularly in multi-site settings (Harrison-Atlas et al., 2021; Wu et al., 2022; Arslan Tuncar et al., 2024; Liu and Zhang, 2024). Among these architectures, U-Net (Ronneberger et al., 2015) is particularly well suited to this task given its abilities to capture multi-scale spatial structure while preserving local detail. A U-Net therefore provides a strong deterministic surrogate for this problem, learning the conditional central tendency of layout-dependent power fields.

However, such surrogates do not quantify predictive uncertainty. A single prediction cannot fully characterize how surrogate error varies across atmospheric regimes and turbine layouts, particularly for unseen configurations. Moreover, for computational efficiency, the surrogate is typically conditioned on a reduced set of gridded inputs, rather than full three-dimensional atmospheric state resolved by the mesoscale model across multiple variables and vertical levels. This reduced representation introduces indeterminacy in the relationship and makes the mapping from inputs to waked power fields inherently uncertain. For planning-stage applications, where decisions depend on both expected performance and associated risks, it is therefore

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

important to complement deterministic predictions with uncertainty estimates. Generative models, such as diffusion models, provide a flexible framework for probabilistic modelling of wind power fields by producing multiple predictions from the conditional distribution, forming ensembles that reflect predictive uncertainty (Zhang et al., 2025).

To this end, we develop diffusion-based probabilistic surrogate models that learn the conditional distribution of layout-dependent power fields. We consider conditional denoising diffusion probabilistic model (DDPM) (Ho et al., 2020; Song et al., 2020), which have recently shown strong performance in conditional generation of atmospheric fields (Ling et al., 2024; Price et al., 2024). In addition, a residual diffusion formulation in considered, in which the diffusion model is trained to represent the residual between the true power field and a deterministic prediction. This approach allows the probabilistic model to focus on learning corrections to the deterministic surrogate rather than modelling the full field directly, and has been shown to improve bias control and calibration in related applications (Yu et al., 2024; Chen and Gao, 2025; Mardani et al., 2025).

1.1 Intent of Study

This study develops layout-conditioned surrogate models for wind power prediction using paired free-stream and waked WRF simulations over a single domain. High-resolution free-stream WRF simulations without turbines, together with 12 layout-dependent WRF-WFP simulations, are used to learn the corresponding spatial wind power fields over a typical meteorological year (TMY). By pairing free-stream and waked simulations at the same resolution, the framework isolates turbine-induced wake effects under realistic mesoscale forcing while avoiding confounding from downscaling errors. The task is therefore formulated as learning the layout-dependent wake response conditioned on background atmospheric flow, rather than as atmospheric downscaling.

Within this framework, the goal is to develop a surrogate framework capable of emulating layout-dependent wake responses for practical use in regional planning. We approach this in two phases. In the first, we develop a deterministic U-Net as a spatial surrogate for predicting layout-conditioned power fields from free-stream atmospheric covariates and turbine configurations. The U-Net serves not just as a final model but as a diagnostic platform: we systematically examine the role of different atmospheric predictors, the sensitivity of results to the choice of predictand, the effect of training set size, and the selected loss function (L_1 or L_2). The aim is to establish what a sound deterministic surrogate requires before introducing additional complexity. In the second phase, we build on these lessons to develop diffusion-based probabilistic surrogates that generate ensembles of predictions and quantify model uncertainty. We evaluate both a standard conditional diffusion formulation and a residual diffusion variant that learns corrections to the U-Net prediction rather than the full power field directly. This residual approach concentrates the probabilistic model's capacity on correcting the deterministic surrogate rather than modelling the full power field directly, which is important for bias control in risk-aware planning decisions. Together, the two phases yield a framework designed with deployability in mind, balancing predictive accuracy, training and inference cost, and robustness to unseen layouts.

The remainder of the paper is organized as follows. Section 2 presents the proposed models. Section 3 describes the data and preprocessing procedures. Section 4 first reports the deterministic U-Net results together with a sensitivity analysis of the

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

deterministic surrogate, and then presents the results for the diffusion-based models. Finally, Section 5 concludes the paper and discusses future work.

2 Models

2.1 Problem formulation

We model the waked wind power field over a fixed spatial domain on an $H \times W$ WRF grid, where H and W are the numbers of grid cells along the two spatial grid axes, respectively. Let $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ denote the free-stream atmospheric covariates from WRF simulations without turbines, where C is the number of atmospheric input variables used for surrogate modelling. Let $\boldsymbol{\ell}_l \in \mathbb{R}^{1\times H\times W}$ denote a layout encoding for $l \in \{1,\dots,L\}$ layouts over the grid. Since capacity is spatially sparse, we define the corresponding binary mask $\mathbf{m}_l \in \{0,1\}^{1\times H\times W}$ that identifies active wind farm grid cells:

$$m_{l,ij} = \mathbb{I}[\boldsymbol{\ell}_{l,ij} > 0], \qquad i = 1,\dots,H,\ j = 1,\dots,W. \tag{1}$$

All training losses are evaluated only on masked grid cells having positive capacity to avoid learning from regions with no installed capacity. The learning target is the corresponding layout-dependent wind power field $\mathbf{y}_l \in \mathbb{R}^{1\times H\times W}$ from WRF-WFP simulations.

A deterministic surrogate learns a function f_θ such that

$$\widehat{\mathbf{y}}_{l,\text{det}} = f_\theta(\mathbf{x}, \boldsymbol{\ell}_l), \tag{2}$$

which produces a single prediction for layout $\boldsymbol{\ell}_l$ given the conditioning inputs . Under L_1 or L_2 loss, this prediction approximates the conditional central tendency, corresponding respectively to the conditional median or conditional mean.

For probabilistic modelling, the goal is instead to learn the conditional distribution

$$p(\mathbf{y}_l \mid \mathbf{x}, \boldsymbol{\ell}_l), \tag{3}$$

which describes the layout-conditioned power response given atmospheric conditions $\mathbf{x}$ and layout $\boldsymbol{\ell}_l$. In contrast to a deterministic surrogate that produces a single point prediction, a probabilistic model generates multiple predictions by sampling from the learned conditional distribution. A set of K ensemble samples can then be generated from the learned conditional distribution:

$$\widehat{\mathbf{y}}_l^{(k)} \sim p_\theta(\mathbf{y}_l \mid \mathbf{x}, \boldsymbol{\ell}_l), \qquad k = 1,\dots,K. \tag{4}$$

2.2 Deterministic U-Net Emulator

We first construct a deterministic surrogate using an encoder–decoder architecture based on U-Net. The model takes the concatenated conditioning input $[\mathbf{x}, \boldsymbol{\ell}_l]$ and predicts a spatially structured power field $\widehat{\mathbf{y}}_{l,\text{det}}$ at the same resolution. Figure A1 illustrates the model architecture.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

The deterministic surrogate is a fully convolutional U-Net composed of residual blocks with RMS normalization and sigmoid linear unit (SiLU) activations. Spatial resolution is reduced through successive $2\times$ downsampling operations implemented by space-to-depth rearrangement followed by 1×1 convolutions. Linear attention is applied at the shallower resolutions, whereas full self-attention is used at the deepest scale and bottleneck to capture longer-range spatial dependencies. The decoder mirrors the encoder, using bilinear upsampling and skip connections to restore spatial resolution while preserving fine-scale features.

The U-Net model is trained by minimizing an L_p loss,

$$\mathcal{L}_{\text{det}}(\theta) = \mathbb{E}_{\boldsymbol{\ell}}\mathbb{E}_{\mathbf{y}_l}\left\{\frac{1}{N_l}\left\|\mathbf{m}_l \odot (\mathbf{y}_l - \widehat{\mathbf{y}}_{l,\text{det}})\right\|_p^p\right\}, \tag{5}$$

where $\odot$ denotes the Hadamard (element-wise) product and $N_l = \sum_{i=1}^{H}\sum_{j=1}^{W} m_{l,1ij}$ is the number of nonzero-capacity grid cells under layout $\boldsymbol{\ell}_l$. In this study, we consider both $p = 1$ and $p = 2$, corresponding to L_1 and L_2 training, respectively. These losses target different measures of the conditional distribution over active grid cells, where L_1 training estimates the conditional median and L_2 training estimates the conditional mean. Accordingly, the deterministic U-Net provides a flexible surrogate that can be used either as a direct point predictor or as the mean component in the residual diffusion formulation.

In preliminary experiments, the L_1 objective consistently gave smaller deterministic errors, making it the stronger choice for direct point prediction. However, an L_2-trained U-Net is also required in the residual diffusion formulation, where it serves as the mean predictor used to define the residual field. For this reason, both L_1 and L_2-trained U-Net models are considered and compared in this study.

2.3 Probabilistic Diffusion Models

2.3.1 Conditional Diffusion Model

U-Net provides an efficient point emulator of the paired free-stream-to-waked mapping but does not directly quantify predictive uncertainty. To model predictive uncertainty, diffusion models are well suited because they can learn complex conditional distributions and generate ensembles from them. We adopt a conditional denoising diffusion probabilistic model (DDPM) (Ho et al., 2020), which learns the conditional distribution by gradually corrupting the data with Gaussian noise and training a neural network to reverse this noising process. Mathematically, the forward noising process defines a discrete Markov chain over index $s \in [0, S]$:

$$q(\mathbf{y}_{l,s} \mid \mathbf{y}_{l,s-1}) = \mathcal{N}\left(\mathbf{y}_{l,s}; \sqrt{1-\beta_s}\mathbf{y}_{l,s-1}, \beta_s\mathbf{I}\right), \tag{6}$$

where $\{\beta_s\}_{s=1}^{S}$ is a predefined linear noise schedule. Letting $\bar{\alpha}_s = \prod_{t=1}^{s}(1-\beta_t)$, the equivalent closed-form expression for the noisy sample is

$$\mathbf{y}_{l,s} = \sqrt{\bar{\alpha}_s}\mathbf{y}_{l,0} + \sqrt{1-\bar{\alpha}_s}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where $\mathbf{y}_{l,0} \equiv \mathbf{y}_l$.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

The reverse-time generative process is parameterized by a neural network conditioned on atmospheric covariates $\mathbf{x}$ and layout $\boldsymbol{\ell}_l$, given by

$$p_\theta(\mathbf{y}_{l,s-1} \mid \mathbf{y}_{l,s}, \mathbf{x}, \boldsymbol{\ell}_l). \tag{8}$$

In this study, conditioning is implemented by concatenating $[\mathbf{x}, \boldsymbol{\ell}_l]$ with the noisy sample $\mathbf{y}_{l,s}$ along the channel dimension, together with a learned embedding of the diffusion step s. The denoiser backbone uses the same U-Net model as the deterministic emulator, augmented with time embeddings of the denosing steps.

Rather than considering prediction error $\boldsymbol{\epsilon}$ (noise prediction) or clean prediction $\mathbf{y}_{l,0}$, we train the denoiser using the $\mathbf{v}$-prediction parameterization (Salimans and Ho, 2022), which improves training stability and is better suited for few-step sampling with fast solvers. This parameterization leads a reparameterization of the denoising target as a linear combination of $(\mathbf{y}_{l,0}, \boldsymbol{\epsilon})$:

$$\mathbf{v}_{l,s} = \sqrt{\bar{\alpha}_s}\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_s}\mathbf{y}_{l,0}. \tag{9}$$

The denoising network outputs $\widehat{\mathbf{v}}_\theta(\mathbf{y}_{l,s}, s, \mathbf{x}, \boldsymbol{\ell}_l)$ and is trained with

$$\mathcal{L}_v(\theta) = \mathbb{E}_{\boldsymbol{\ell}}\mathbb{E}_{s,\mathbf{y}_{l,0},\boldsymbol{\epsilon}}\left[\frac{\bar{\alpha}_s}{N_l}\,\|\mathbf{m}_l \odot [\mathbf{v}_{l,s} - \widehat{\mathbf{v}}_\theta(\mathbf{y}_{l,s}, s, \mathbf{x}, \boldsymbol{\ell}_l)]\|_2^2\right], \tag{10}$$

with s sampled uniformly from $[1, S]$. Without masking, this objective is equivalent to the original noise-prediction loss (Ho et al., 2020) up to a linear reparameterization between $\boldsymbol{\epsilon}$ and $\mathbf{v}$, with the weighting factor $\bar{\alpha}_s$ ensuring consistency between the two parameterizations (Hang et al., 2023).

At inference time, ensemble members are generated by numerically solving the reverse-time dynamics starting from Gaussian noise. We employ the DPM-Solver++(2M) sampler (Lu et al., 2025) with 20 sampling steps. This method is a second-order multi-step solver designed for fast sampling in a log signal-to-noise ratio parameterization. Repeated sampling yields an ensemble $\{\widehat{\mathbf{y}}_l^{(k)}\}_{k=1}^K$ approximating $p_\theta(\mathbf{y}_l \mid \mathbf{x}, \boldsymbol{\ell}_l)$. Since wind power is physically meaningful only at active wind farm grid cells, we additionally enforce the spatial mask $\mathbf{m}_l$ by fixing zero-capacity grid cells to zero at each solver step.

2.3.2 Residual Diffusion Model

To improve bias control and calibration, we further consider residual diffusion (Mardani et al., 2025). Let $\widehat{\mathbf{y}}_{l,L_2}$ denote the deterministic prediction from an auxiliary U-Net trained separately with an L_2 loss. The residual field is then defined as

$$\mathbf{r}_l = \mathbf{y}_l - \widehat{\mathbf{y}}_{l,L_2}. \tag{11}$$

Instead of modelling $\mathbf{y}_l$ directly, we train a conditional diffusion model that approximates the residual distribution via

$$p_\theta(\mathbf{r}_l \mid \mathbf{x}, \boldsymbol{\ell}_l). \tag{12}$$

Because the auxiliary deterministic model is trained with L_2 loss, $\widehat{\mathbf{y}}_{l,L_2}$ approximates the conditional mean, so the residual field primarily represents the remaining unresolved variability. As a result, the residual formulation typically yields a lower-variance target distribution. Empirically, this often simplifies the conditional distribution, making $p(\mathbf{r}_l \mid \mathbf{x}, \boldsymbol{\ell}_l)$ easier to learn than $p(\mathbf{y}_l \mid \mathbf{x}, \boldsymbol{\ell}_l)$ (Mardani et al., 2025).

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

At inference, we sample a set of K ensemble members $\widehat{\mathbf{r}}_l^{(k)}$ using the same diffusion setup as above and reconstruct the final ensemble prediction as

$$\widehat{\mathbf{y}}_{l,\mathrm{r}}^{(k)} = \widehat{\mathbf{y}}_{l,L_2} + \widehat{\mathbf{r}}_l^{(k)}, \qquad k = 1, \ldots, K. \tag{13}$$

This decomposition allows the diffusion model to focus on modelling corrections to the deterministic predictor rather than the full target field. We adopt the same diffusion framework as in DDPM to model the residuals and refer to this variant as DDPM-Res.

3 Data preparation

This study uses a paired simulation dataset generated by high-resolution WRF (version 4.6.1) over an anonymized domain in mid-latitude North America. A three-level nested domain configuration with horizontal grid of spacing of 9 km, 3 km, and 1 km is employed, where the innermost 1 km domain covers the wind farm region of interest. The simulations are forced using ERA5 reanalysis data at 3-h intervals. The physical parameterizations used in the WRF simulations are summarized in Table 1. This paired design uses free-stream WRF and WRF-WFP simulations at the same spatial resolution, with the learning target defined as the corresponding layout-dependent wind power response from the WRF-WFP simulations.

Table 1. Physical parameterization schemes used in the WRF simulations (version 4.6.1).

Model component	Configuration
Radiation (shortwave and longwave)	RRTMG radiation scheme (Iacono et al., 2008)
Planetary boundary layer	MYNN level-2.5 turbulence closure (Nakanishi and Niino, 2004, 2009)
Surface layer	Revised MM5 Monin–Obukhov similarity scheme (Jiménez et al., 2012)
Land surface processes	Noah land surface model (Chen and Dudhia, 2001)
Deep convection	Kain–Fritsch cumulus parameterization (Kain and Fritsch, 1990; Kain, 2004)
Cloud microphysics	WRF Single-Moment 3-class microphysics (Hong et al., 2004)
Wind farm parameterization	Fitch wind farm scheme (Fitch et al., 2012), with TKE generation factor set to 1.0

A baseline free-stream WRF simulation without turbines is paired with 12 WRF-WFP waked simulations corresponding to the layout scenarios shown in Fig. 1. These scenarios are designed to reflect a realistic regional planning setting rather than random perturbations of a single layout. In practice, a developer seeks to evaluate new wind farm proposals within a region that already contains existing and in-construction farms. In this context, it is important to assess both (i) how existing farms affect the production of proposed layouts, and (ii) how the proposed layouts, in turn, impact the performance of existing assets, for example in the context of inter-farm wake interactions and potential compensation agreements.

To capture this two-way interaction, the simulation set is constructed around a baseline configuration representing the existing farms (Scenario 1) and an extended configuration including existing and in-construction farms (Scenario 2). For each

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

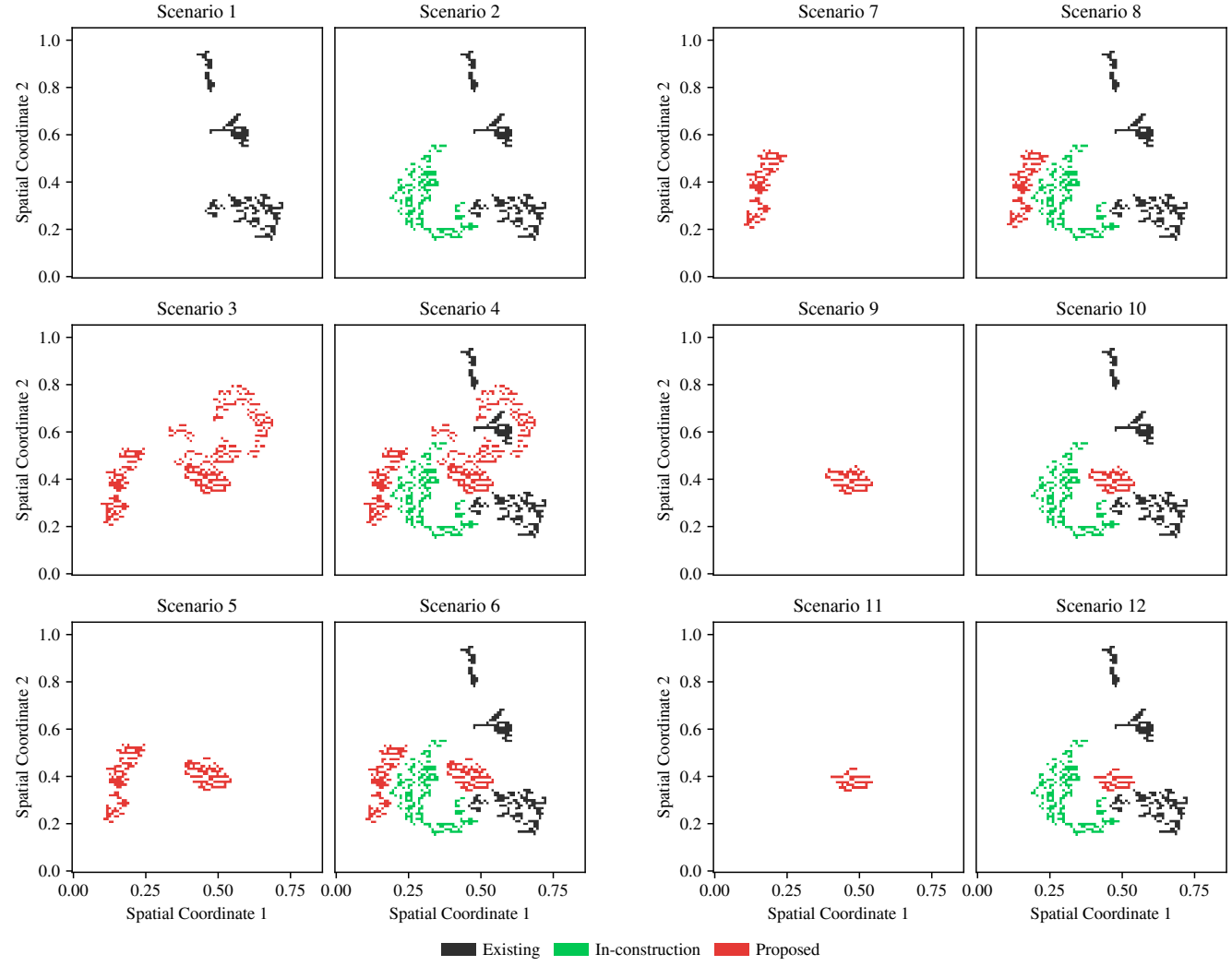


Figure 1. Layouts of the 12 study scenarios, including existing and existing-plus-construction baselines (Scenarios 1 and 2) and paired proposed-farm layouts in standalone and combined multi-farm settings (Scenarios 3–4, 5–6, 7–8, 9–10, and 11–12). Scenarios 1–10 are used for training, while Scenarios 11 and 12 are reserved for validation. Spatial coordinates are anonymized while preserving relative separation, and domains are shown in arbitrary orientations for confidentiality.

proposed wind farm layout, we then generate paired scenarios consisting of (i) the proposed farm in isolation and (ii) the proposed farm combined with the existing and in-construction farms (e.g., Scenario 3 vs. 4). Repeating this construction across multiple proposed layouts (Scenarios 5, 7, 9, and 11) yields a collection of paired scenarios that explicitly encode both directions of wake interaction between proposed and existing farms. This pairing strategy allows the surrogate model to learn how wake effects propagate across layouts under realistic multi-farm configurations, rather than treating each layout independently.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

For model development, Scenarios 1–10 are used for training, while Scenarios 11 and 12 are reserved for evaluation. The spatial distribution of wind farm grid cells, coloured by farm, is shown in Fig. B1 (see Appendix B).

All scenarios are produced on the same spatial grid and time axis, with $1\,\mathrm{km} \times 1\,\mathrm{km}$ horizontal resolution and 10-min temporal resolution over a TMY, yielding 52,560 samples per simulation. The TMY is constructed following Xia et al. (2025), where for each calendar month, a representative year is selected from 2000-2024 ERA5 data at the domain centre by minimizing a normalized Wasserstein distance between each candidate month and the long-term distribution of $100\,\mathrm{m}$ wind speed and direction and $2\,\mathrm{m}$ temperature.

A variety of wind-related atmospheric variables are used as covariates extracted from the free-stream WRF run, as shown in Table 2. The baseline covariates include hub-height ($100\,\mathrm{m}$) wind speed and wind direction, with direction represented by sine and cosine encoding. Additional wind levels at $80\,\mathrm{m}$ and $120\,\mathrm{m}$ are also considered. Beyond wind, we consider hub-height air density and turbulent kinetic energy, as well as the potential temperature gradient between $200\,\mathrm{m}$ and $20\,\mathrm{m}$. These covariates are concatenated channel-wise to form the atmospheric input tensor. Land surface elevation, normalized to $[0, 1]$, is included as an additional static covariate. All continuous covariates are standardized to have zero mean and unit variance, except the sine and cosine direction channels which are already bounded in $[-1, 1]$.

Table 2. Atmospheric and static covariates used as model inputs. Wind direction is encoded by its sine and cosine components.

Covariate	Height / representation
Wind speed	80 m, 100 m, 120 m
Wind direction	80 m, 100 m, 120 m; sine and cosine
Air density	100 m
Turbulent kinetic energy	100 m
Potential temperature gradient	200 m – 20 m
Elevation	Static; normalized to $[0, 1]$
Layout encoding	Normalized installed capacity map

The layout encoding $\boldsymbol{\ell}$ is defined as the normalized installed capacity map on the WRF grid. Given the turbines' locations for each scenario, each turbine is assigned to its nearest WRF grid cell and the installed capacity in a grid cell is then computed by summing the rated capacities of all turbines assigned to it. To obtain $\boldsymbol{\ell}$, the resulting capacity map is normalized by a global reference capacity $P_{\max} = 20\ \mathrm{MW}$, chosen to be comfortably above the maximum installed capacity within any single grid cell across all scenarios. From the same capacity grid we derive the binary farm mask $\mathbf{m}_{\boldsymbol{\ell}}$ that identifies active wind farm grid cells to which the loss, sampling, and evaluation are restricted. Both $\boldsymbol{\ell}$ and $\mathbf{m}_{\boldsymbol{\ell}}$ are invariant in time within a scenario but vary across scenarios.

The learning target is the layout-dependent wind power field from WRF-WFP simulations. For deterministic U-Net models, we consider two representations: globally normalized wind power (WP), normalized by $P_{\max}$, and capacity factor (CF), normalized by the maximum capacity at each grid cell. CF is attractive because it removes variation in absolute power arising

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

only from differences in installed capacity across layouts, so the target more directly reflects production efficiency under the local flow and wake conditions. This can simplify the learning problem within a single domain when turbine type and installed capacity are similar. However, CF also removes information about absolute production scale, which becomes important when comparing farms of different sizes or extending the framework across heterogeneous domains or turbines. Moreover, because CF is normalized by the layout-specific maximum capacity, typical operating values are concentrated closer to the upper boundary of $[0, 1]$, making additive residual corrections more prone to artifacts near this effective ceiling. Normalizing by the global reference capacity $P_{\max}$ instead keeps most grid cell values away from this upper limit, giving the residual model more room to operate. For these reasons, both WP and CF are considered for deterministic U-Net models under L_1 loss, whereas diffusion models are trained only on WP. In addition, an L_2 trained U-Net using WP is included for the DDPM-Res formulation to provide the mean predictor for defining the residual field. For clarity, the surrogate models considered in this study and their corresponding target representations are summarized in Table 3.

Table 3. Summary of surrogate models and target representations.

Model name	Response	Training target	Loss / objective	Output
U-Net-WP-L_1	WP	$\mathbf{y}_l$	L_1 loss	Point prediction
U-Net-CF-L_1	CF	$\mathbf{y}_l$	L_1 loss	Point prediction
U-Net-WP-L_2	WP	$\mathbf{y}_l$	L_2 loss	Point prediction
DDPM	WP	$\mathbf{y}_l$	Diffusion ($\mathbf{v}$-prediction)	$\widehat{\mathbf{y}}_l^{(k)\ddagger}$
DDPM-Res	WP	$\mathbf{r}_l{}^{\dagger}$	Diffusion on residuals	$\widehat{\mathbf{y}}_{l,L_2} + \widehat{\mathbf{r}}_l^{(k)\ddagger}$

$^{\dagger}\mathbf{r}_l = \mathbf{y}_l - \widehat{\mathbf{y}}_{l,L_2}$, where $\widehat{\mathbf{y}}_{l,L_2}$ denotes the prediction from U-Net-WP-L_2.
$^{\ddagger}k = \{1, \ldots, K\}$.

Within each training scenario, time steps are randomly split into 80% training, 10% validation, and 10% testing, with only the training split used for model fitting. For withheld scenarios, results are evaluated on all time steps unless otherwise noted. Model performance across these data splits for all scenarios is reported in Appendix C.

All deep-learning models are trained on 4 NVIDIA H100 GPUs, with training taking approximately 8 hours for U-Net and 52 hours for the diffusion models. Inference on a single H100 GPU takes about 5 minutes per scenario for the U-Net and about 14 hours for the diffusion models to generate 10 ensemble members. For comparison, a 3-day WRF-WFP simulation on this domain required about 4 hours on a single H100 GPU or on 16 vCPUs. This corresponds to a total runtime of approximately 488 hours (20 days) for the full TMY on a single H100, or about 122 hours (5 days) when distributed across 4 H100 GPUs. Once trained, the U-Net reduces full-TMY inference time by a factor of roughly 1,500 relative to direct WRF-WFP simulation run across 4 H100 GPUs, while even the more computationally demanding diffusion model ensemble achieves a speedup of approximately 9×.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

4 Results

We evaluate the proposed models on two withheld scenarios not used during training: Scenario 11 (single-farm extrapolation) and Scenario 12 (multi-farm configuration). All metrics are computed over “active" grid cells that contain one or more wind turbines. In Section 4.1, we first evaluate the deterministic U-Net variants, namely U-Net-WP-L_1, U-Net-CF-L_1, and U-Net-WP-L_2, followed by a sensitivity analysis examining the effects of target representation and several modelling choices.

In Section 4.2, we then evaluate the diffusion models DDPM and DDPM-Res using 10 ensemble members. Probabilistic performance is assessed using fair Continuous Ranked Probability Score (CRPS) (Ferro, 2013), empirical ensemble coverage, rank histograms, spread–skill relationships, and farm-level uncertainty summaries.

4.1 Deterministic U-Net Models

4.1.1 Grid-Level Accuracy

Table 4 summarizes the performance for the two withheld layouts, computed over active wind turbine grid cells. For both scenarios, all three models achieve relatively small deterministic errors, with absolute bias below 0.02 MW, MAE below 0.1 MW, and RMSE below 0.28 MW. The percentages in parentheses contextualize these errors relative to the scenario-specific mean true power over active wind farm grid cells. On this scale, absolute bias remains below 0.5%, MAE below 4%, and RMSE below 10% for all models in both scenarios.

Table 4. Deterministic error (MW) on withheld scenarios for U-Net models. The percentages in parentheses are the corresponding relative errors, expressed relative to the scenario-specific mean true power over active wind farm grid cells.

	Scenario 11			Scenario 12		
Model	Bias	MAE	RMSE	Bias	MAE	RMSE
U-Net-WP-L_1	-0.004 (-0.12%)	0.081 (2.72%)	0.272 (9.13%)	0.005 (0.18%)	0.071 (2.76%)	0.232 (8.97%)
U-Net-CF-L_1	-0.009 (-0.29%)	0.074 (2.48%)	0.268 (9.01%)	0.005 (0.19%)	0.066 (2.56%)	0.230 (8.90%)
U-Net-WP-L_2	-0.013 (-0.45%)	0.091 (3.07%)	0.276 (9.27%)	0.000 (0.00%)	0.082 (3.19%)	0.235 (9.08%)

In terms of systematic error, biases are generally small, indicating limited systematic over- or underprediction. In Scenario 11, all models exhibit slight negative bias, whereas in Scenario 12 the biases are near zero overall. Among the U-Net variants, the L_1-trained models consistently achieve lower MAE and slightly lower RMSE than the L_2 model in both withheld scenarios. Between the two L_1 models, using CF as the target yields slightly lower MAE and RMSE than using WP in both scenarios, while both retain very small mean bias. This suggests that CF simplifies the learning problem in this single domain setting by removing variation due to installed capacity, allowing the model to focus more directly on layout-induced wake effects.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

4.1.2 Farm-Level Accuracy

While grid cell-level performance is informative, planning-stage decisions are typically made at the wind farm scale, where stakeholders care about the layout-induced impact on annualized farm production. Accordingly, we also evaluate farm-wise relative error defined as

$$\text{relative error } (\%) = 100 \times \frac{\widehat{P} - P}{P}, \tag{14}$$

where $\widehat{P}$ is the predicted total power and P is the true total power. Table 5 reports farm-wise relative error for the withheld scenarios. Scenario 11 contains a single wind farm (Farm A), whereas Scenario 12 contains eight farms (Farms 1–8), with Farm 5 corresponding to the same physical site as Farm A in Scenario 11. For confidentiality reasons, we can not disclose which turbines in Scenario 12 correspond to which assigned wind farm.

Table 5. Farm-wise relative error (%) on the withheld scenarios for the U-Net models. Negative values indicate underprediction and positive values indicate overprediction.

	Scenario 11	Scenario 12							
Model	Farm A[†]	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5[†]	Farm 6	Farm 7	Farm 8
U-Net-WP-L_1	-0.12	0.06	0.39	0.26	0.25	-0.31	0.23	0.21	0.29
U-Net-CF-L_1	-0.29	0.21	0.33	0.09	0.26	-0.23	0.27	0.24	0.28
U-Net-WP-L_2	-0.45	0.56	0.11	-0.64	-0.14	-0.86	0.35	0.20	0.28

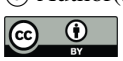
[†] Farm 5 in Scenario 12 represents the same physical farm as Farm A in Scenario 11, but under a combined layout that includes existing and in-construction farms.

At the farm scale, all three U-Nets show small relative errors overall. In Scenario 11, all models slightly underpredict annual farm power, with U-Net-WP-L_1 closest to zero. The small negative bias may be related to the fact that Farm A has lower installed capacity than the corresponding layout seen during training. Although the site itself is represented in Scenario 9, the Scenario 11 layout contains fewer turbines and therefore differs in both capacity and spatial configuration. This shift may cause the model to slightly underpredict production for the withheld layout and makes the case less well aligned with the training examples.

In Scenario 12, the two L_1-trained models maintain small relative errors across all farms, whereas U-Net-WP-L_2 show somewhat larger deviations at several sites, most notably Farms 1, 3, and 5. The largest deviations across both withheld scenarios occur at Farm A in Scenario 11 and the corresponding Farm 5 in Scenario 12, suggesting that this shared site is more difficult to emulate accurately under changing layout context. Overall, the L_1-trained U-Net models provide the most consistent farm-level agreement.

For completeness, additional farm-wise comparisons are provided in Section E1. Figs. E1 to E4 show scatter plots of predicted versus true spatially aggregated farm power for each model and farm at 10-minute timescales, further demonstrating the high accuracy of the U-Net models.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026



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

4.1.3 Spatial Error Patterns

Figure 2 shows the mean bias for Scenarios 11 and 12, computed at each grid cell as predicted minus truth. These maps show where in the domain each model tends to over- or underpredict, and provide a spatial diagnostic complementary to the aggregated metrics reported above. For both layouts, error patterns differ between models, indicating location-dependent strengths and weaknesses. Consistent with the lower deterministic errors reported in Table 4, the L_1-trained models generally exhibit smaller bias than the L_2 model. In Scenario 12, the most spatially coherent errors occur around Farm 5, consistent with the greater difficulty of this withheld layout discussed above. Overall, these diagnostics suggest that the withheld scenario

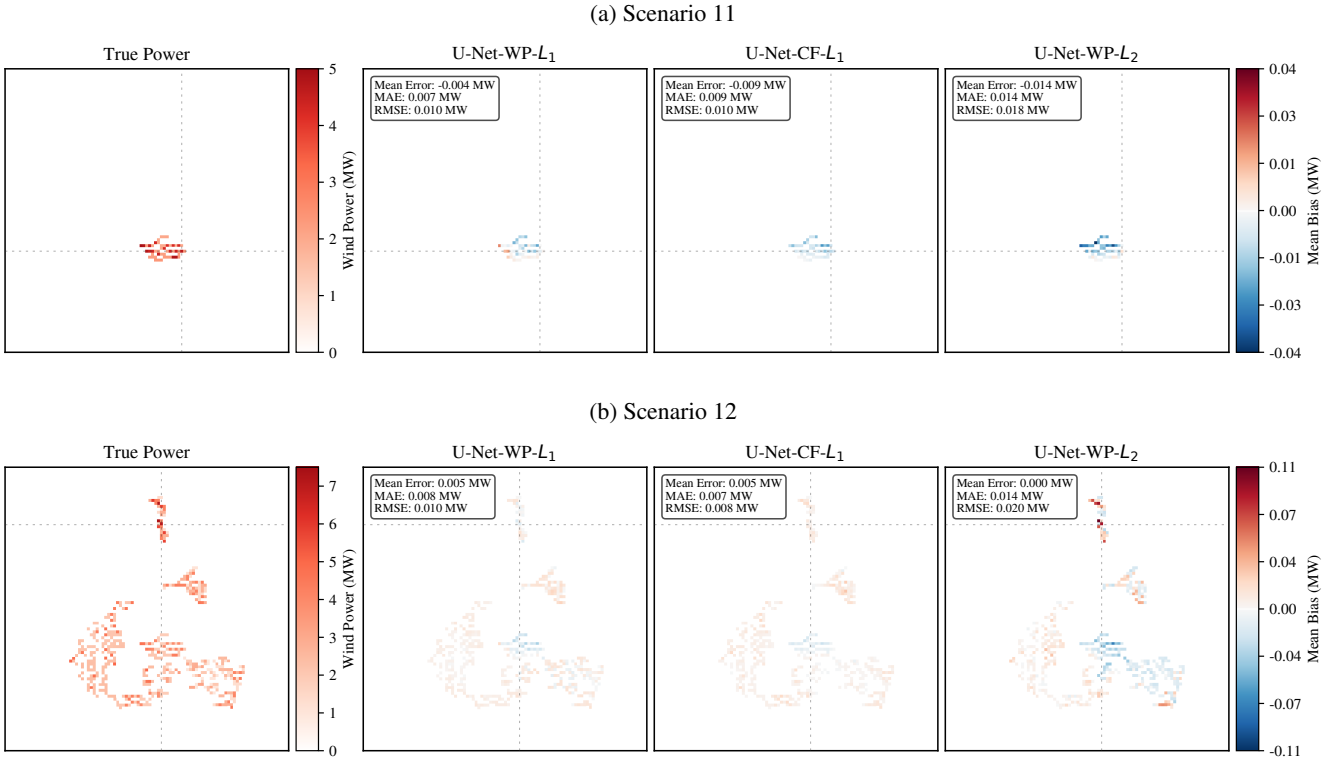


Figure 2. Ground-truth mean WRF power and mean bias fields for the U-Net models under (a) Scenario 11 and (b) Scenario 12. Reference lines mark the grid cell with the highest installed capacity used for the time-series analysis. A subset of the corresponding time series is shown in Fig. 8 and Fig. 9 for Scenario 11 and Scenario 12, respectively.

errors depend on the scenario design. Scenario 11 represents the proposed farm in isolation, whereas in Scenario 12 the same site is embedded in a combined layout with existing and in-construction farms. Errors are somewhat larger in Scenario 11, indicating that the standalone case is harder to emulate than the corresponding combined-layout case. This differs from the pattern observed within the training scenarios (see Table C1 in Appendix C), where multi-farm configurations tend to be more challenging. As a temporal robustness check, Table C1 also reports results separately for the training, validation, and test splits.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

To summarize, the deterministic U-Net provides strong predictive performance across both withheld scenarios, with low errors at both the grid cell and aggregated wind farm scales and small systematic bias. Even under layout extrapolation, the model captures the main spatial structure of waked power fields and preserves aggregated farm-level production with high accuracy. Given its much lower computational cost than WRF-WFP, the U-Net provides a practical surrogate for layout-conditioned power prediction in planning-stage analysis.

4.1.4 Sensitivity Analysis

Next, we use a U-Net-based sensitivity analysis to examine how several modelling choices affect deterministic surrogate performance. Since L_1 gave the best deterministic performance above, all models in this analysis are trained with L_1. The analysis assesses the importance of training diversity (i.e., number of scenarios modelled), wind representation, and additional atmospheric variables. To isolate the effect of these design choices, we perform a one-factor-at-a-time sensitivity analysis varying four components:

1. Number of training scenarios: 4, 6, 8, and 10.
2. Target representation: normalized wind power (Power) vs. capacity factor (CF),
3. Wind encoding: vector components (UV) vs wind speed and direction using sine and cosine (WSA), and
4. Additional covariates beyond the baseline 100 m winds:
   (a) air density at 100 m (AD),
   (b) turbulent kinetic energy at 100 m (TKE),
   (c) potential temperature gradient between 200 m and 20 m (PTG),
   (d) additional wind levels at 80 m and 120 m (Winds (80 & 120 m)).

All models in this sensitivity analysis are evaluated on the withheld Scenarios 11 and 12. Figure 3 summarizes the effect of the number of training scenarios on MAE under different target and wind-encoding choices. Across all configurations, MAE on withheld scenarios generally decreases as more training scenarios are included, with a small uptick at 8 scenarios for the Power and WSA setting. Crucially, we observe that model error does not plateau with the added scenarios, suggesting higher accuracy could still be achieved by training on additional diverse scenarios. Models trained on capacity factor consistently outperform those trained directly on power, and, within each target representation, wind speed with sine and cosine direction encoding yields lower errors than using wind vector components. The combined effect is most pronounced when both CF and WSA are used, achieving the lowest MAE across all training sizes.

In terms of output representation, although WP is better suited to modelling across heterogeneous domains, the sensitivity analysis shows that CF can yield lower MAE than WP in the present single domain setting. This suggests that, within a single domain containing farms with similar installed capacities and turbine types, CF can be a better target for deterministic

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

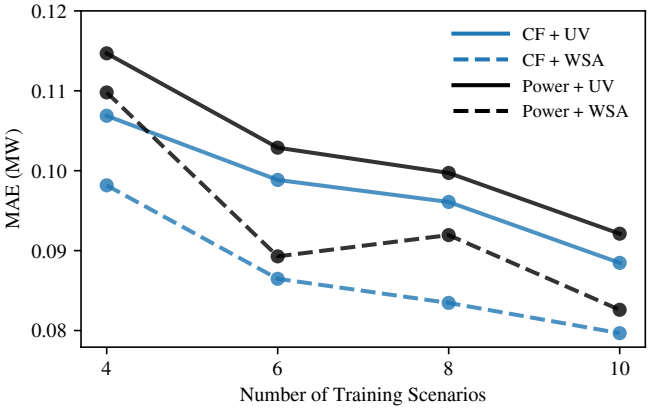


Figure 3. MAE as a function of the number of training scenarios, comparing target representations (CF vs. Power) and wind encodings (UV vs. WSA). CF results are shown in blue and Power in black; solid and dashed lines denote UV and WSA inputs, respectively.

prediction because the normalization removes simple capacity scaling and lets the model focus more directly on layout-driven wake effects.

Figure 4 reports the relative MAE change when adding additional atmospheric covariates to the 100 m wind-only baseline. Most additions reduce MAE, but the absolute improvements are small, on the order of 10^{-3}–10^{-2} MW, indicating only limited

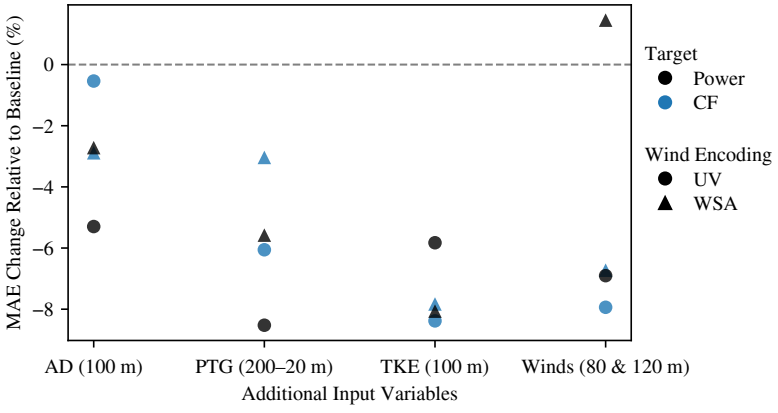


Figure 4. Relative MAE Change From the 100 m Wind-Only Baselines for Each Configuration With Additional Atmospheric Covariates

incremental benefit from the added atmospheric variables in this single domain setting. TKE and the additional wind levels (80 and 120 m) provide the most consistent improvements, whereas PTG and AD show weaker and less uniform effects across targets and wind encodings. Relative improvements are generally larger under UV than WSA, which is consistent with the lower baseline MAE of the WSA models and therefore their reduced room for further improvement. Only the Power with

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

WSA configuration shows slight degradation when adding the 80 m & 120 m wind levels, with an increase in MAE by just 0.0012 MW.

The differences across added covariates are modest and may be comparable to variability from from model stochastic training, such as random initialization and minibatch ordering. Thus, relative metrics can exaggerate their practical importance when baseline errors are already low. The limited sensitivity may reflect the single-domain design, where all scenarios share the same underlying meteorological conditions, and larger gains may be more likely in extensions to multiple domains with more diverse atmospheric regimes. Nevertheless, all additional covariates are retained in the main model comparisons, as they are physically relevant and do not degrade performance.

4.2 Probabilistic Diffusion Models

Moving beyond deterministic surrogates, we next evaluate the diffusion-based models. Since these models produce predictive distributions rather than single point estimates, deterministic metrics computed from the ensemble mean alone do not fully characterize performance. We therefore assess both deterministic and probabilistic skill.

4.2.1 Distributional Accuracy and Coverage

Table 6 summarizes deterministic and probabilistic metrics for the withheld scenarios. In terms of deterministic performance,

Table 6. Error metrics on withheld scenarios for diffusion models. Bias, MAE, and RMSE are reported in MW. Fair CRPS and empirical ensemble coverage are evaluated using 10 ensemble members.

	Scenario 11				
Model	Bias	MAE	RMSE	CRPS	Coverage
DDPM	-0.036 (-1.20%)	0.103 (3.44%)	0.266 (8.94%)	0.075	76.8%
DDPM-Res	-0.023 (-0.78%)	0.084 (2.81%)	0.267 (8.96%)	0.068	53.7%
	Scenario 12				
Model	Bias	MAE	RMSE	CRPS	Coverage
DDPM	-0.019 (-0.75%)	0.088 (3.41%)	0.224 (8.65%)	0.061	82.7%
DDPM-Res	-0.001 (-0.03%)	0.072 (2.80%)	0.231 (8.94%)	0.054	74.7%

the ensemble means of both diffusion models achieve errors of similar magnitude to the U-Net. Absolute bias remains below 0.04 MW, MAE below 0.11 MW, and RMSE below 0.28 MW in both scenarios. Relative errors remain below 1.3% for bias, 4% for MAE, and 9% for RMSE. Compared with the U-Net results reported earlier, the diffusion models show somewhat larger negative bias and slightly higher MAE, but comparable or slightly lower RMSE. This pattern suggests that diffusion models

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

tend to reduce large errors while introducing more moderate deviations overall, likely because ensemble averaging smooths predictions, damping extremes at the expense of sharpness.

Comparing the two diffusion models, DDPM-Res achieves lower bias, MAE, and CRPS than DDPM in both scenarios, while maintaining similar RMSE. By contrast, DDPM attains higher empirical coverage, reflecting its broader predictive distributions. This indicates a trade-off between sharpness and dispersion: DDPM-Res produces sharper and more accurate predictions, whereas DDPM yields wider ensembles that more often contain the truth. Coverage is higher in Scenario 12 than in Scenario 11 for both models, a pattern that is also consistent across the training, validation, and test subsets reported in Table C2 in Appendix C. More broadly, CRPS and coverage reflect different aspects of probabilistic performance: sharper ensembles may achieve lower CRPS yet remain underdispersed, whereas broader ensembles may attain higher coverage at the expense of resolution.

4.2.2 Farm-Level Accuracy

At the farm level, Table 7 shows that the largest deviations for the diffusion models occur at Farm A in Scenario 11 and the corresponding Farm 5 in Scenario 12, indicating that this shared site is the most difficult for both the U-Net and diffusion models under both standalone and multi-farm configurations. In Scenario 11, both diffusion models underpredict total power, with DDPM-Res errors closer to zero. In Scenario 12, DDPM shows systematic underprediction across all farms, whereas DDPM-Res errors remain closer to zero for most farms, with a notable negative bias persisting at Farm 5.

Table 7. Farm-wise relative error (%) on the withheld scenarios for diffusion models. Negative values indicate underprediction and positive values indicate overprediction.

	Scenario 11	Scenario 12							
Model	Farm A[†]	Farm 1	Farm 2	Farm 3	Farm 4	Farm 5[†]	Farm 6	Farm 7	Farm 8
DDPM	-1.20	-0.53	-0.70	-0.83	-0.46	-1.39	-0.77	-0.68	-0.70
DDPM-Res	-0.78	0.08	0.20	0.06	0.02	-1.10	0.22	0.05	0.04

[†] Farm 5 in Scenario 12 represents the same physical farm as Farm A in Scenario 11, but under a combined layout that includes existing and in-construction farms.

Although DDPM-Res does not exhibit a consistent advantage over the deterministic U-Net in farm-level accuracy across the withheld layouts, it does improve accuracy for some farms in Scenario 12, where its relative errors are closer to zero. U-Net, however, performs better in Scenario 11 and remains more accurate at the most difficult shared sites such as Farm A and Farm 5. Further farm-level comparisons are reported in Section E2. Figures E5 to E7 display scatter plots of predicted versus true spatially aggregated power for each farm and model.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

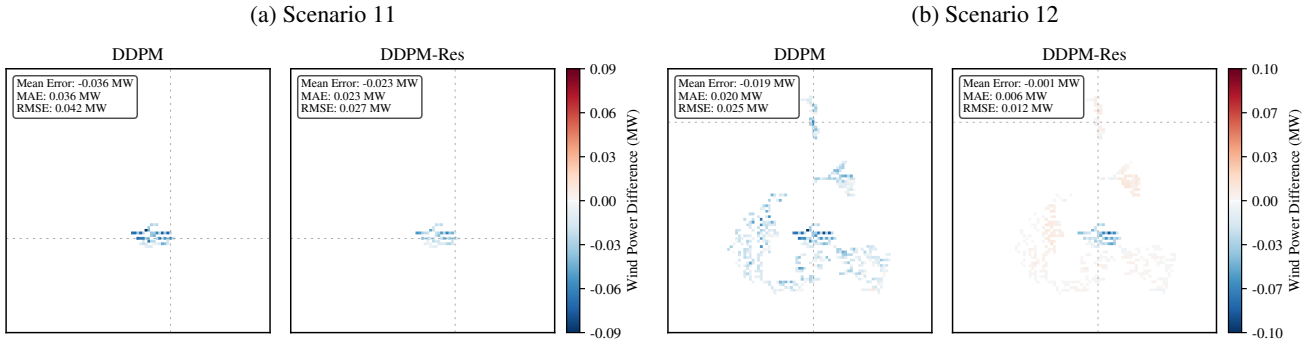


Figure 5. Ground-truth mean WRF power and mean bias fields for DDPM and DDPM-Res under (a) Scenario 11 and (b) Scenario 12. Reference lines mark the grid cell with the highest installed capacity used for the time-series analysis. A subset of the corresponding time series is shown in Fig. 8 and Fig. 9 for Scenario 11 and Scenario 12, respectively.

4.2.3 Spatial Error Patterns

Figure 5 shows the mean bias for the diffusion models on withheld scenarios. Similar to the U-Net results, the bias maps reveal clear location-dependent error patterns, with the most spatially coherent errors concentrated around Farm 5. In Scenario 12, DDPM exhibits broader and more systematic underprediction, whereas DDPM-Res shows lower-magnitude and more spatially localized errors. For both models, the largest deviations remain concentrated around Farm 5, consistent with the farm-level results and shows the greater difficulty of this site under the multi-farm configuration. The persistence of these errors suggests that the difficulty arises from changes in the surrounding wake-interaction context, rather than the mere presence of the farm in the training layouts.

4.2.4 Calibration and Reliability Diagnostics

Rank histograms provide additional diagnostics of ensemble calibration. Figure 6 presents the rank histograms and corresponding empirical CDFs (ECDF) for DDPM and DDPM-Res. For a well-calibrated ensemble, the rank histogram should be approximately uniform and ECDF should follow the identity line. Here rank 1 corresponds to the truth value being smaller than all ensemble members and 11 to truth being greater than all ensemble members.

For Scenario 11, both DDPM and DDPM-Res show clear departures from uniformity, with elevated frequencies at the extreme ranks. This indicates that the truth frequently falls outside the ensemble range, consistent with the reduced coverage reported in Table 6. In Scenario 12, while both diffusion models display non-uniform rank distributions, that of DDPM-Res is closer to uniform across non-extreme ranks, though peaks at boundary ranks demonstrate that the generated ensembles tend to be too narrow. The corresponding ECDF for DDPM-Res tracks the ideal diagonal line more closely than DDPM, showing improved calibration in the sense that the truth more frequently falls with the ensemble range.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

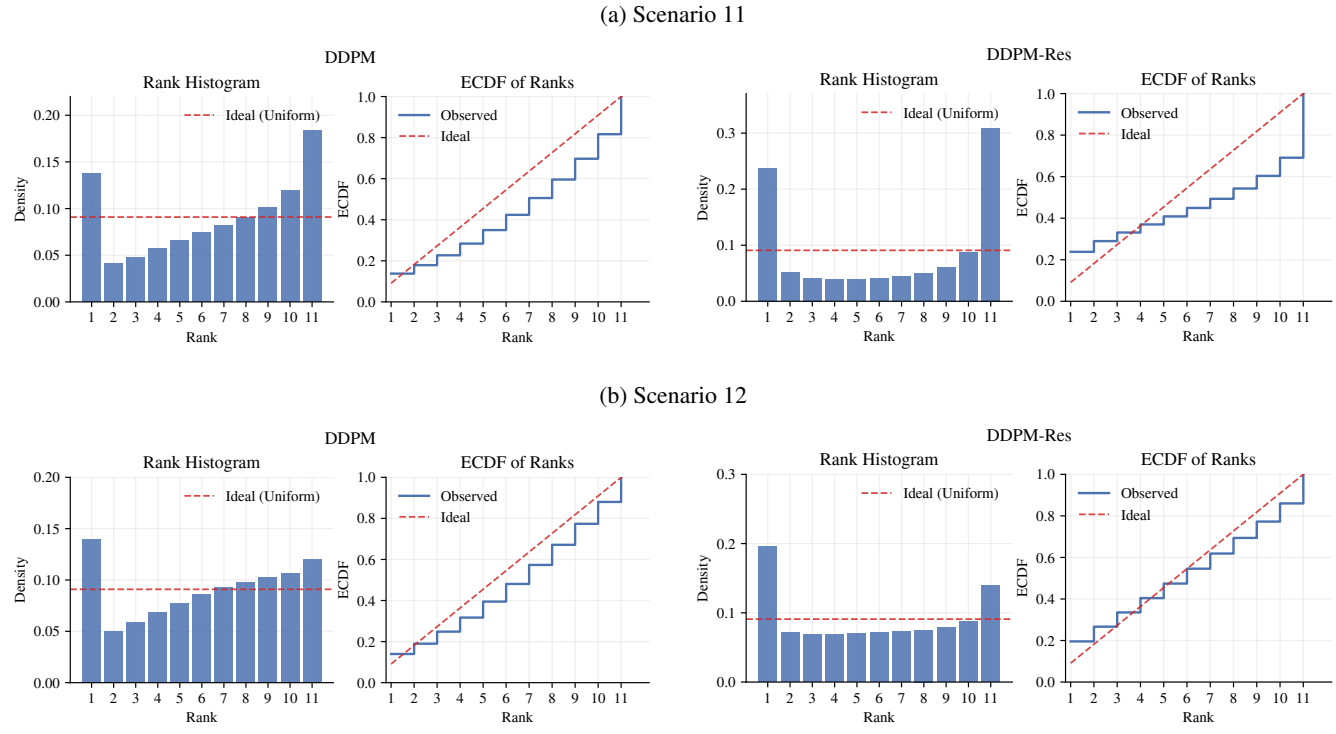


Figure 6. Rank histograms and empirical cumulative distribution functions (ECDFs) for DDPM (left) and DDPM-Res (right) under (a) Scenario 11 and (b) Scenario 12.

Across both scenarios, asymmetry of the DDPM rank histograms can also be interpreted as evidence of systematic ensemble bias relative to the truth, which may partly explain its slightly larger CRPS. Taken together with the lower CRPS for DDPM-Res, these diagnostics indicate that residual diffusion provides better overall probabilistic performance, particularly in the multi-farm evaluation, although perfect calibration is not achieved. More broadly, these results show the challenge of representing uncertainty when generalizing to unseen layouts. This interpretation is also consistent with the validation and test results for the training scenarios in Table C2, where probabilistic errors are generally smaller than for the withheld layouts.

We further assess ensemble calibration using spread-skill relationship, which compare ensemble dispersion with the corresponding prediction error. Figure 7 plots the relationship between RMSE and ensemble SD for DDPM and DDPM-Res. The plots show that DDPM follows the identity line more closely in both scenarios, indicating better consistency between ensemble spread and realized error. By contrast, DDPM-Res lies systematically above the line, showing the underdispersed ensembles and tend to underestimate the true error. This is consistent with the earlier probabilistic diagnostics, where DDPM-Res produces sharper predictive distributions with lower CRPS but reduced coverage. At the same time, DDPM-Res attains lower RMSE across most spread bins, showing improved deterministic accuracy despite poorer spread calibration. Together, these results

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

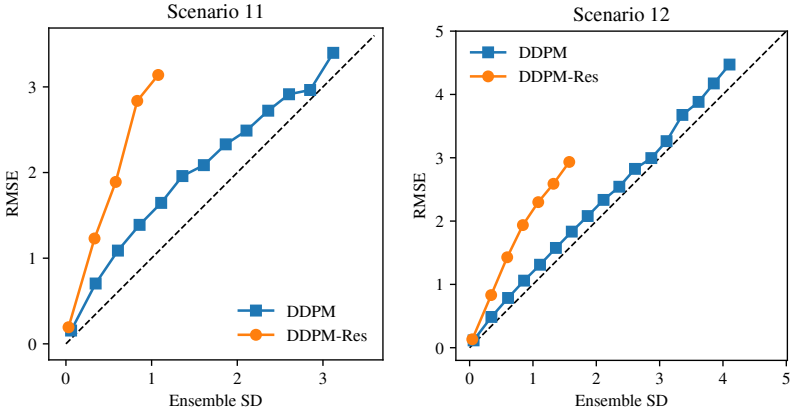


Figure 7. Spread–skill relationship for Scenario 11 (left) and Scenario 12 (right).

indicate a trade-off between calibration and sharpness, where DDPM provides better spread-error consistency and DDPM-Res yields more accurate but less dispersed productions, with the mismatch most evident in Scenario 11.

4.2.5 Temporal Uncertainty Dynamics

To further examine the temporal behaviour of predictive uncertainty, we show example time series at a representative high-capacity grid cell during a selected period of the year. Figures 8 and 9 show examples for Scenario 11 and Scenario 12, respectively.

The top panels show the corresponding wind speed and wind direction encoded as sine and cosine components at 100 m height. The lower panels display the ground-truth power output together with predictions for the U-Nets and diffusion models. For the probabilistic models, the ensemble mean is plotted together with grey bands corresponding to the full ensemble range across 10 members. These bands represent pointwise uncertainty conditional on the inputs at each time step, as the generative samples are independent across time and do not capture joint temporal dependence beyond that induced by the conditioning covariates.

It is observed that under both scenarios, all predicted trajectories follow the truth closely for much of the selected interval. The ensemble ranges from DDPM and DDPM-Res are often so narrow that they are effectively indistinguishable from the deterministic prediction in parts of the plots, especially when wind speed and direction evolve gradually, indicating strong agreement among ensemble members under relatively steady conditions. The most evident local deviations occur during short intervals of rapidly varying wind conditions, notably around 09-15 12:00–18:00, 09-16 16:00–21:00, and 09-17 14:00–18:00 in both scenarios, suggesting that forecast difficulty is associated with rapid changes in the atmospheric states.

The diffusion models are more informative during these rapidly varying intervals. In general, the ensemble spread increases when the deterministic prediction is less accurate and contracts during quieter periods, indicating that predictive uncertainty is closely coupled to prediction difficulty. Across both scenarios, DDPM generally produces the broader envelope, whereas

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

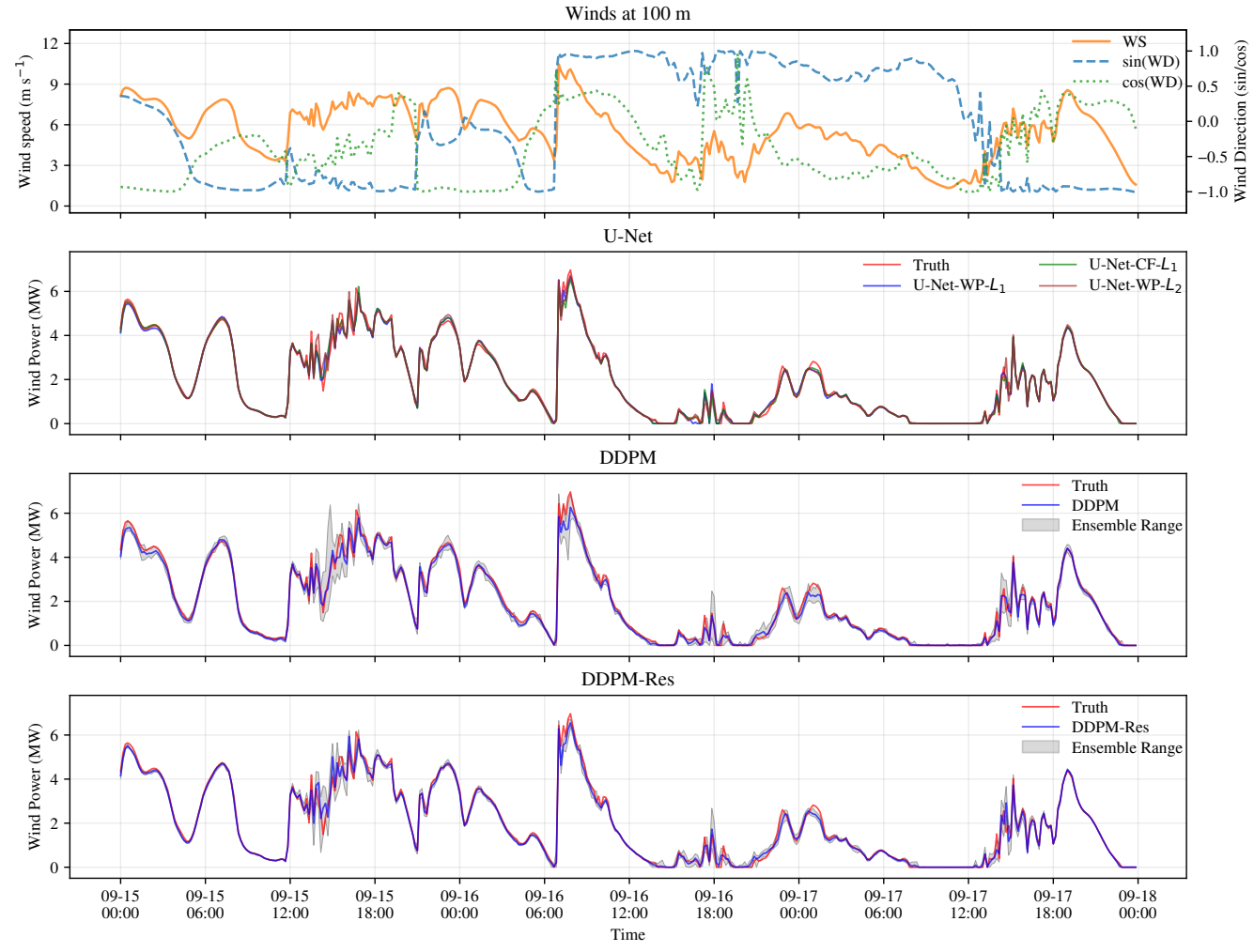


Figure 8. Time series over a selected interval for the highest-capacity grid cell in Scenario 11 (grid cell location indicated in the top row of Fig. 2 and Fig. 5). Top: 100 m wind speed and direction encodings. Bottom: truth, U-Net predictions, and diffusion ensemble means; the grey band shows the min–max 10 realization ensemble range for DDPM and DDPM-Res.

DDPM-Res remains sharper while still expanding during the same episodes. In many cases the truth remains within the ensemble range, although it occasionally falls outside the range during the most rapidly varying periods, indicating that the generative models capture meaningful uncertainty without achieving perfect calibration.

To examine these temporal uncertainty patterns more systematically, we next summarize probabilistic performance across broader time scales. Seasonal diagnostics further illustrate how predictive uncertainty varies across time scales. Figure 10 show monthly-averaged CRPS and ensemble spread, measure in standard deviation (SD), for both scenarios. The monthly statistics show a marked seasonal cycle, with both CRPS and ensemble spread generally higher during the summer months

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

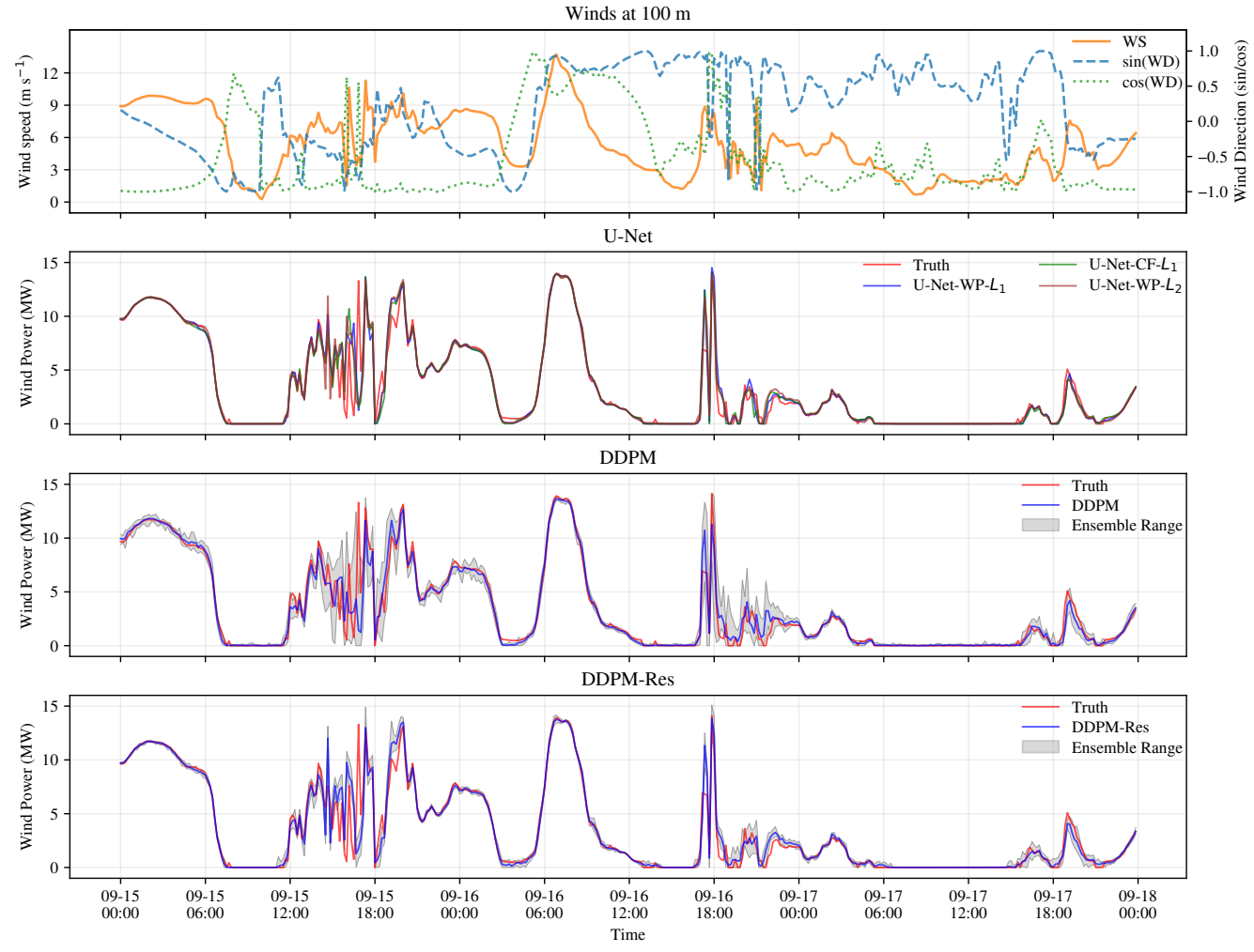


Figure 9. Time series over a selected interval for the highest-capacity grid cell in Scenario 12 (grid cell location indicated in the bottom row of Fig. 2 and Fig. 5). Top: 100 m wind speed and direction encodings. Bottom: truth, U-Net predictions, and diffusion ensemble means; the grey band shows the min–max ensemble range for DDPM and DDPM-Res.

(June–September). This suggests that the fast sub-daily fluctuations associated with increased ensemble range may themselves have a seasonal dependence. These patterns should, however, be interpreted with caution, since the analysis is based on a TMY constructed from representative months rather than a continuous physical year. The corresponding diurnal diagnostics show comparatively weaker variation and are provided in Appendix D (Figure D1).

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

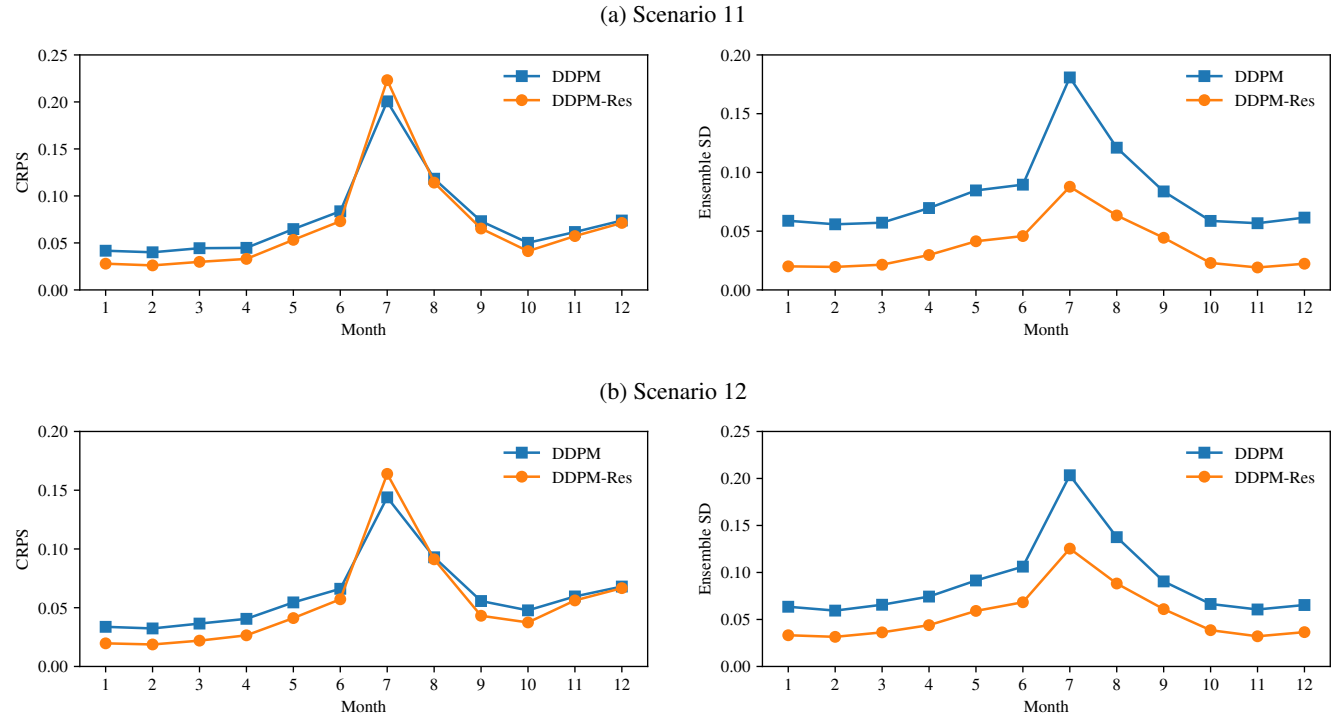


Figure 10. Monthly mean CRPS (left) and ensemble standard deviation (right) under (a) Scenario 11 and (b) Scenario 12.

4.2.6 Farm-Level Uncertainty Bands

To evaluate uncertainty at an operationally relevant scale, we aggregate power predictions to the farm level. For each wind farm, the spatially aggregated power time series is normalized by the farm-specific maximum true power to allow direct comparison. Figure 11 summarizes the resulting farm-wise wind power predictions across models. Each deterministic U-Nets provide single point estimates, whereas diffusion models produce ensembles whose spreads characterize predictive uncertainty. Shaded bands represent the ensemble interquartile range (IQR) together with full ensemble range.

Across farms, the ensemble means of DDPM and DDPM-Res remain close to the U-Net predictions, indicating that the diffusion models largely preserve point prediction skill. The ground truth generally falls within the ensemble ranges, showing that the probabilistic models capture meaningful farm-level uncertainty, although DDPM-Res shows a small miss at Farm A. Relative to DDPM, DDPM-Res produces slightly narrower uncertainty bands, consistent with its residual formulation, in which the diffusion model focuses on correcting the remaining errors of the deterministic emulator. Farm 5 in Scenario 12 stands out, with broader uncertainty bands than Farm A in Scenario 11, which may reflect the greater difficulty of this site under the multi-farm wake-interaction setting.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

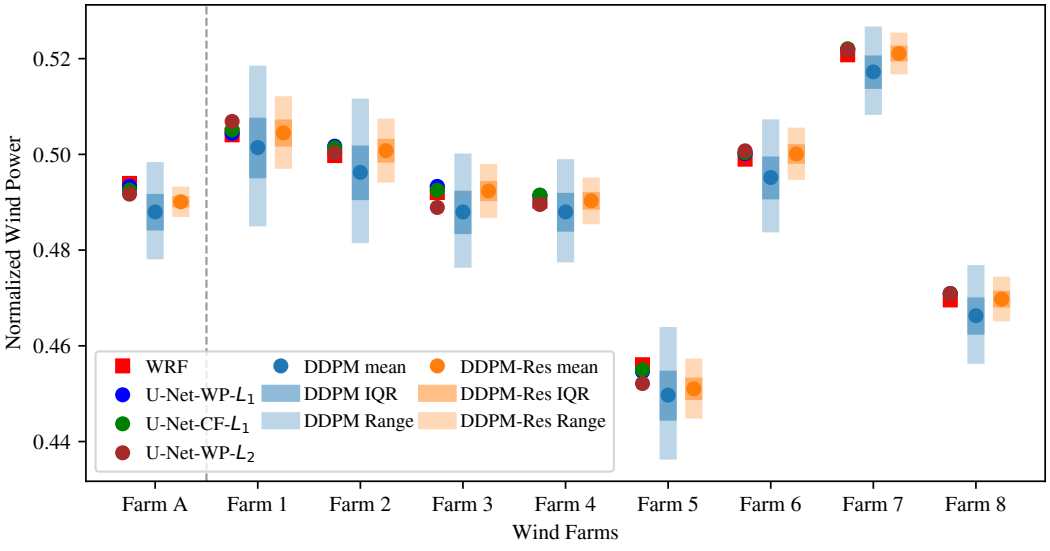


Figure 11. Farm-wise comparison of observed and predicted normalized power for the withheld scenarios, with U-Net shown as a deterministic prediction and DDPM and DDPM-Res summarized by the ensemble mean, IQR, and range.

Taken together, these results indicate that the main benefit of the diffusion models is probabilistic rather than purely deterministic. Although they do not deliver a uniform improvement over U-Net in farm-level aggregated accuracy, they retain competitive point predictions while providing ensemble-based uncertainty estimates under unseen layouts. Between the two diffusion models, DDPM provides better ensemble calibration and spread–skill consistency, whereas DDPM-Res gives slightly sharper and, in some cases, more accurate mean predictions. Thus, DDPM is preferable when calibrated uncertainty estimates are the primary goal, while DDPM-Res may be useful when sharper predictions are desired.

5 Conclusion and Discussion

This work develops layout-conditioned surrogate models for wind power prediction using paired WRF simulations with and without wind farm parameterization. The results show that CNN-based surrogates can effectively predict spatial power fields from atmospheric covariates and turbine layout information in the presence of wake effects. The deterministic U-Nets perform strongly in predicting spatial power fields, showing that layout-dependent wake interactions can be captured through data-driven spatial regression when trained on physically consistent simulations. Critically, generating results with a trained U-Net model was found to be 1,500 times faster than running WRF-WFP directly, removing a key cost barrier for wind energy stakeholders who need to run dozens of simulations for layout optimization.

Diffusion-based models extend the deterministic emulators by producing ensembles that represent predictive uncertainty. The resulting uncertainty is adaptive, such that the ensemble spread increases with deterministic error. In particular, the gener-

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

ated ensemble is found to be larger when atmospheric covariate changes rapidly and smaller under steadier conditions. While diffusion-based surrogates do not uniformly outperform U-Net in deterministic accuracy, they retain comparable predictive skill while providing additional information about uncertainty.

Within this probabilistic framework, residual diffusion provides an alternative to standard DDPM by leveraging the deterministic U-Net as a baseline and modelling the remaining error structure. This leads to lower MAE, better farm-level bias control, and lower CRPS compared with DDPM, but not uniformly better probabilistic performance. Instead, the two diffusion formulations exhibit a clear trade-off: DDPM produces broader ensembles with better coverage and spread–skill consistency, whereas DDPM-Res yields sharper and more accurate mean predictions but is somewhat underdispersed.

Overall, these results demonstrate that diffusion-based surrogates can extend strong deterministic emulators by providing ensemble-based uncertainty estimates. From a practical perspective, these models are well suited for wind farm planning and layout optimization, enabling rapid evaluation of candidate layouts while also supporting risk-aware decision-making through uncertainty quantification.

A key surrogate modelling choice concerns how turbine layouts are represented. In this work, wind turbines are represented simply as a capacity value per grid cell, which may contain one or more turbines. This representation seems sufficient for this preliminary analysis which worked with broadly similar turbine types (capacity ranges from 2.3 to 4.5 MW). In reality, different turbines with their own power and thrust curves, hub heights, and rotor diameters demonstrate very different power/wind speed relationships, especially in the core linear part of the power curve. In fact, in ongoing work involving multiple domains and a much broader range of turbine types (i.e., 0.66 MW to 8 MW), it is observed that this simple representation becomes insufficient. In such settings, incorporating turbine-specific information such as discrete thrust and power curves, and extrapolating winds to actual turbine hub heights, more accurately captures wake behaviour across turbine types and domains.

Moreover, different methods for incorporating conditioning information are not explored in this work. Conditioning is implemented through simple channel-wise concatenation of the inputs. While effective in the current setting, more expressive conditioning mechanisms, such as feature-wise modulation (Perez et al., 2018), cross-attention (Chen et al., 2021), or classifier-free guidance (Ho and Salimans, 2022), may improve how layout and atmospheric features influence the prediction process, particularly in cross-domain settings.

Another important direction is to move beyond a surrogate model trained and applied at a single domain to a generalized model that can provide accurate results at an unseen domain. For this work, the single domain was appropriate for an existing industry partner, who was developing significant wind assets in the region and needed a model highly tuned to these local conditions. More broadly, wind energy stakeholders are interested in layout optimization and wake analysis across many different domains, each with unique atmospheric conditions, terrains, and turbine configurations. Therefore, there is value in a surrogate model trained on multiple domains and which can generalize across unseen domains and layouts. In such cases, high-resolution 1-km “free-stream" WRF simulations may not be available, and therefore alternate data sources may be required (e.g. ERA5, 2-3 km WRF-based wind atlases, etc.). For such applications, we expect the inclusion of additional atmospheric variables, especially those characterizing atmospheric stability, to be of higher importance than they were in this single domain analysis. We are currently implementing this multi-domain model and will share results as part of a follow-up paper next year.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

Finally, alternative model architectures may further improve generalization across domains. The current framework relies on CNNs that operate on a fixed grid, which align naturally with the gridded atmospheric inputs within a single domain. However, wind farms span domains of different spatial extent and turbine layouts vary in scale. Graph neural networks (GNN) provide a flexible representation that is not tied to a fixed grid and can explicitly model turbine-to-turbine interactions under arbitrary layouts. Therefore, GNN-based surrogates offer a promising direction for improving layout and cross-domain generalization in heterogeneous wind farm settings.

Code availability. This study builds on publicly available implementations of diffusion models. Specifically, the U-Net backbone and DDPM training framework are adapted from https://github.com/lucidrains/denoising-diffusion-pytorch, and sampling is performed using the DPM-Solver++ implementation from https://github.com/LuChengTHU/dpm-solver. We modify these implementations to incorporate conditional layout information and a masked loss defined over active wind farm grid cells.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

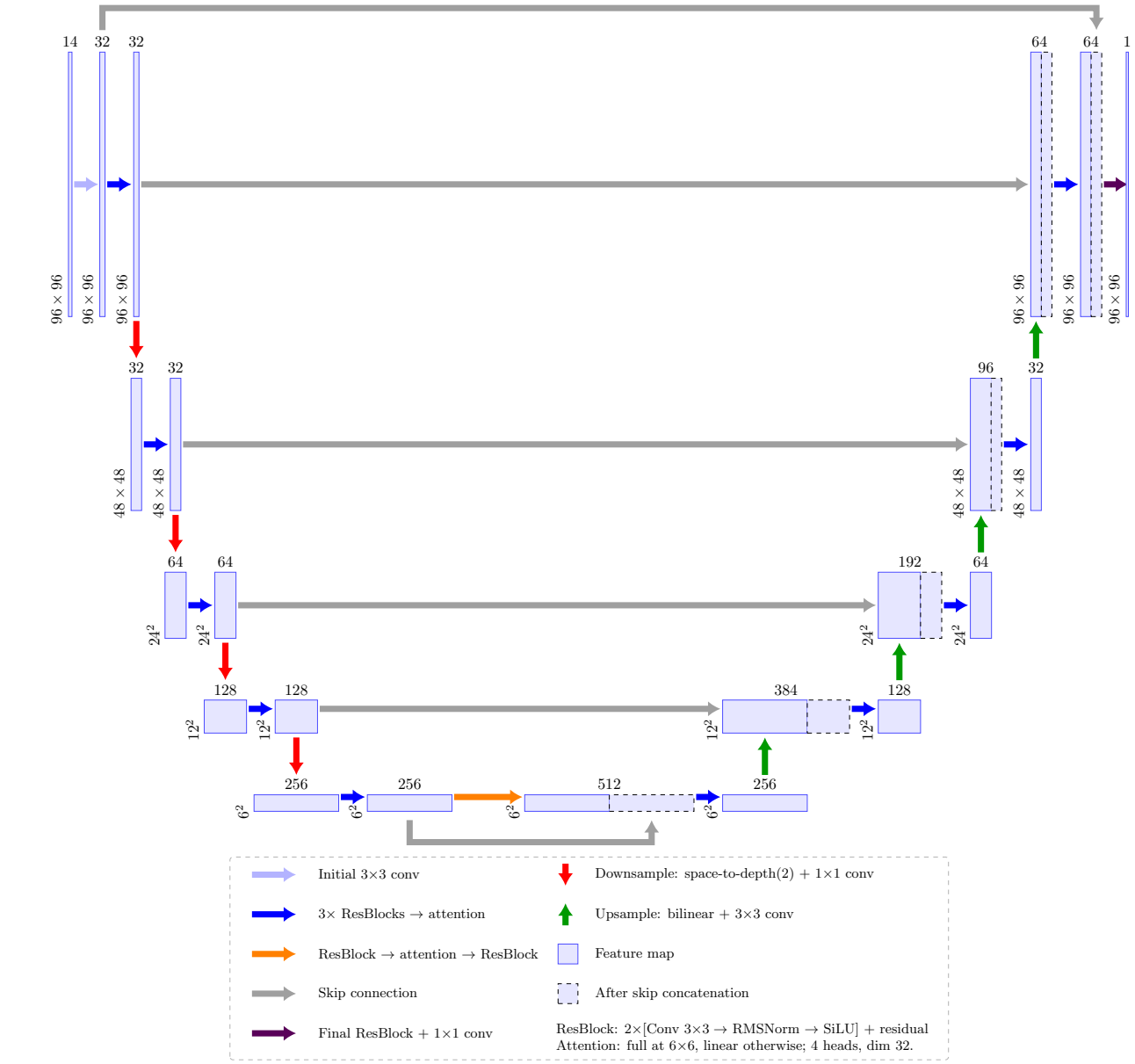


Figure A1. Detailed architecture of the deterministic U-Net. The model takes a 14-channel input on a 96×96 grid and predicts a single-channel power field at the same resolution. Each encoder level contains three residual blocks followed by attention, with downsampling implemented by space-to-depth rearrangement and 1×1 convolutions. Linear attention is used at the shallower resolutions, whereas full self-attention is used at the deepest resolution and bottleneck. The decoder mirrors the encoder using bilinear upsampling, convolution, and skip connections, with each scale containing three skip features in the implementation.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

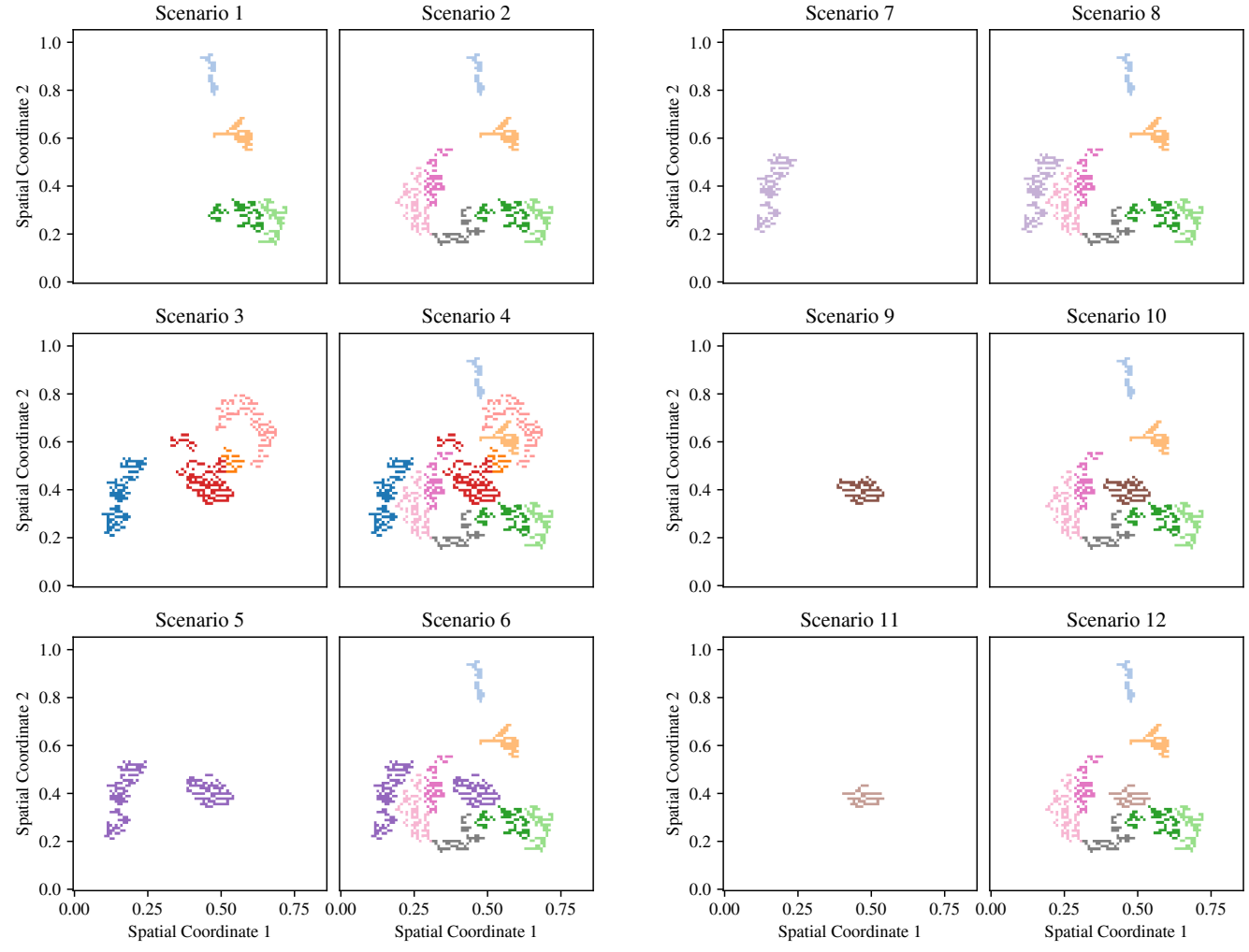


Figure B1. Wind farm partitioning across scenarios. Spatial coordinates are anonymized while preserving relative separation and domains are shown in arbitrary orientations for confidentiality.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

Appendix C: Model Performance Across Data Splits for All Scenarios

Table C1. Deterministic errors (in MW) for U-Net models on the training, validation, and testing subsets across all scenarios.

		Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5			Scenario 6		
Split	Model	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE
Train	U-Net-WP-L_1	0.009	0.048	0.112	0.007	0.051	0.127	0.008	0.056	0.140	0.003	0.064	0.156	0.009	0.050	0.134	0.005	0.058	0.146
	U-Net-CF-L_1	0.041	0.107	0.006	0.045	0.123	0.011	0.008	0.051	0.137	0.004	0.058	0.151	0.006	0.045	0.131	0.006	0.052	0.142
	U-Net-WP-L_2	-0.000	0.060	0.113	0.003	0.061	0.124	0.010	0.066	0.138	0.002	0.075	0.152	0.009	0.058	0.132	0.003	0.068	0.142
Valid	U-Net-WP-L_1	0.009	0.048	0.112	0.006	0.051	0.122	0.007	0.056	0.130	0.002	0.063	0.149	0.007	0.049	0.117	0.003	0.057	0.138
	U-Net-CF-L_1	0.006	0.041	0.108	0.006	0.045	0.118	0.007	0.051	0.127	0.004	0.057	0.143	0.010	0.044	0.114	0.005	0.051	0.134
	U-Net-WP-L_2	-0.001	0.060	0.114	0.002	0.061	0.119	0.009	0.066	0.129	0.001	0.074	0.146	0.008	0.058	0.114	0.001	0.067	0.135
Test	U-Net-WP-L_1	0.009	0.047	0.111	0.007	0.050	0.125	0.008	0.055	0.136	0.003	0.063	0.152	0.009	0.049	0.131	0.005	0.057	0.144
	U-Net-CF-L_1	0.006	0.041	0.106	0.006	0.044	0.122	0.008	0.050	0.134	0.004	0.057	0.148	0.010	0.044	0.129	0.006	0.051	0.140
	U-Net-WP-L_2	0.000	0.060	0.112	0.003	0.060	0.123	0.009	0.065	0.135	0.002	0.073	0.151	0.009	0.058	0.129	0.003	0.067	0.141
		Scenario 7			Scenario 8			Scenario 9			Scenario 10			Scenario 11			Scenario 12		
Split	Model	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE	Bias	MAE	RMSE
Train	U-Net-WP-L_1	0.010	0.047	0.137	0.006	0.054	0.140	0.009	0.044	0.100	0.005	0.054	0.134	-0.004	0.082	0.277	0.005	0.072	0.234
	U-Net-CF-L_1	0.010	0.042	0.134	0.006	0.049	0.136	0.010	0.040	0.098	0.006	0.049	0.131	-0.009	0.075	0.274	0.005	0.067	0.232
	U-Net-WP-L_2	0.012	0.057	0.138	0.004	0.064	0.136	0.008	0.051	0.099	0.002	0.065	0.131	-0.014	0.092	0.281	0.000	0.083	0.237
Valid	U-Net-WP-L_1	0.008	0.046	0.107	0.005	0.054	0.132	0.010	0.044	0.100	0.004	0.054	0.129	-0.004	0.080	0.259	0.004	0.071	0.229
	U-Net-CF-L_1	0.009	0.040	0.104	0.006	0.048	0.128	0.011	0.040	0.096	0.006	0.048	0.125	-0.008	0.072	0.254	0.005	0.066	0.228
	U-Net-WP-L_2	0.010	0.056	0.109	0.003	0.064	0.128	0.008	0.051	0.100	0.002	0.064	0.127	-0.014	0.089	0.262	-0.001	0.082	0.232
Test	U-Net-WP-L_1	0.010	0.047	0.137	0.006	0.053	0.139	0.009	0.043	0.096	0.005	0.054	0.134	-0.002	0.076	0.245	0.005	0.069	0.218
	U-Net-CF-L_1	0.010	0.041	0.134	0.006	0.047	0.134	0.010	0.039	0.096	0.006	0.048	0.130	-0.007	0.068	0.237	0.005	0.063	0.214
	U-Net-WP-L_2	0.012	0.057	0.137	0.004	0.064	0.135	0.008	0.051	0.099	0.002	0.064	0.131	-0.012	0.087	0.249	-0.000	0.080	0.220

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

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

Table C2. Error metrics for diffusion models on the training, validation, and testing subsets across all scenarios. Bias, MAE, and RMSE are reported in MW. Fair CRPS and empirical ensemble coverage are evaluated using 10 ensemble members.

		Scenario 1					Scenario 2					Scenario 3					Scenario 4				
Split	Model	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage
Train	DDPM	-0.006	0.071	0.175	0.048	87.3%	-0.013	0.076	0.196	0.052	86.8%	-0.020	0.083	0.205	0.057	86.2%	-0.024	0.094	0.226	0.065	84.8%
	DDPM-Res	-0.001	0.047	0.105	0.033	79.7%	-0.000	0.051	0.120	0.035	78.8%	0.000	0.058	0.135	0.040	78.5%	-0.000	0.067	0.151	0.047	79.2%
Valid	DDPM	-0.007	0.071	0.175	0.048	87.5%	-0.014	0.076	0.192	0.052	86.9%	-0.021	0.083	0.196	0.057	86.2%	-0.025	0.093	0.218	0.064	84.9%
	DDPM-Res	-0.001	0.047	0.106	0.032	80.2%	-0.000	0.050	0.115	0.035	79.3%	-0.000	0.057	0.126	0.040	78.8%	-0.001	0.066	0.144	0.046	79.6%
Test	DDPM	-0.006	0.069	0.168	0.047	87.4%	-0.012	0.074	0.187	0.051	87.0%	-0.019	0.081	0.193	0.056	86.3%	-0.023	0.091	0.215	0.063	84.9%
	DDPM-Res	-0.001	0.046	0.104	0.032	79.9%	-0.000	0.050	0.119	0.035	79.0%	-0.000	0.056	0.133	0.039	78.5%	-0.001	0.066	0.149	0.046	79.2%

		Scenario 5					Scenario 6					Scenario 7					Scenario 8				
Split	Model	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage
Train	DDPM	-0.018	0.076	0.204	0.053	87.2%	-0.018	0.085	0.217	0.059	85.9%	-0.016	0.069	0.179	0.048	87.6%	-0.016	0.081	0.210	0.056	86.3%
	DDPM-Res	0.000	0.047	0.127	0.034	76.7%	-0.000	0.058	0.139	0.041	78.7%	0.001	0.043	0.130	0.032	76.8%	0.000	0.054	0.133	0.038	78.6%
Valid	DDPM	-0.018	0.075	0.191	0.052	87.3%	-0.019	0.084	0.209	0.058	86.0%	-0.019	0.067	0.155	0.047	87.7%	-0.018	0.081	0.204	0.056	86.4%
	DDPM-Res	-0.001	0.046	0.109	0.033	77.2%	-0.001	0.057	0.132	0.040	79.1%	-0.001	0.041	0.100	0.030	77.4%	0.000	0.054	0.124	0.038	79.1%
Test	DDPM	-0.017	0.074	0.191	0.051	87.3%	-0.017	0.083	0.205	0.057	86.0%	-0.016	0.068	0.174	0.047	87.7%	-0.015	0.079	0.201	0.054	86.3%
	DDPM-Res	-0.000	0.046	0.125	0.033	76.8%	-0.000	0.057	0.138	0.040	78.7%	0.001	0.042	0.130	0.031	76.9%	0.000	0.053	0.131	0.037	78.7%

		Scenario 9					Scenario 10					Scenario 11					Scenario 12				
Split	Model	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage	Bias	MAE	RMSE	CRPS	Coverage
Train	DDPM	-0.015	0.064	0.147	0.044	88.5%	-0.016	0.081	0.205	0.056	86.3%	-0.036	0.103	0.271	0.076	76.8%	-0.019	0.088	0.225	0.062	82.7%
	DDPM-Res	-0.000	0.039	0.090	0.028	76.1%	0.000	0.055	0.128	0.038	78.6%	-0.023	0.085	0.271	0.069	53.7%	-0.001	0.073	0.233	0.054	74.7%
Valid	DDPM	-0.015	0.063	0.144	0.044	88.5%	-0.017	0.080	0.200	0.055	86.5%	-0.037	0.100	0.251	0.073	77.0%	-0.020	0.088	0.222	0.061	82.7%
	DDPM-Res	-0.000	0.039	0.090	0.028	76.7%	-0.001	0.054	0.123	0.038	79.1%	-0.023	0.082	0.252	0.067	53.9%	-0.001	0.072	0.229	0.054	75.1%
Test	DDPM	-0.014	0.062	0.140	0.043	88.5%	-0.015	0.079	0.196	0.054	86.4%	-0.034	0.098	0.241	0.072	76.9%	-0.019	0.085	0.214	0.059	82.9%
	DDPM-Res	-0.001	0.039	0.090	0.028	76.1%	-0.001	0.054	0.127	0.038	78.6%	-0.022	0.080	0.242	0.065	53.6%	-0.001	0.069	0.217	0.051	74.8%

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

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

Appendix D: Hourly Probabilistic Performance Diagnostics

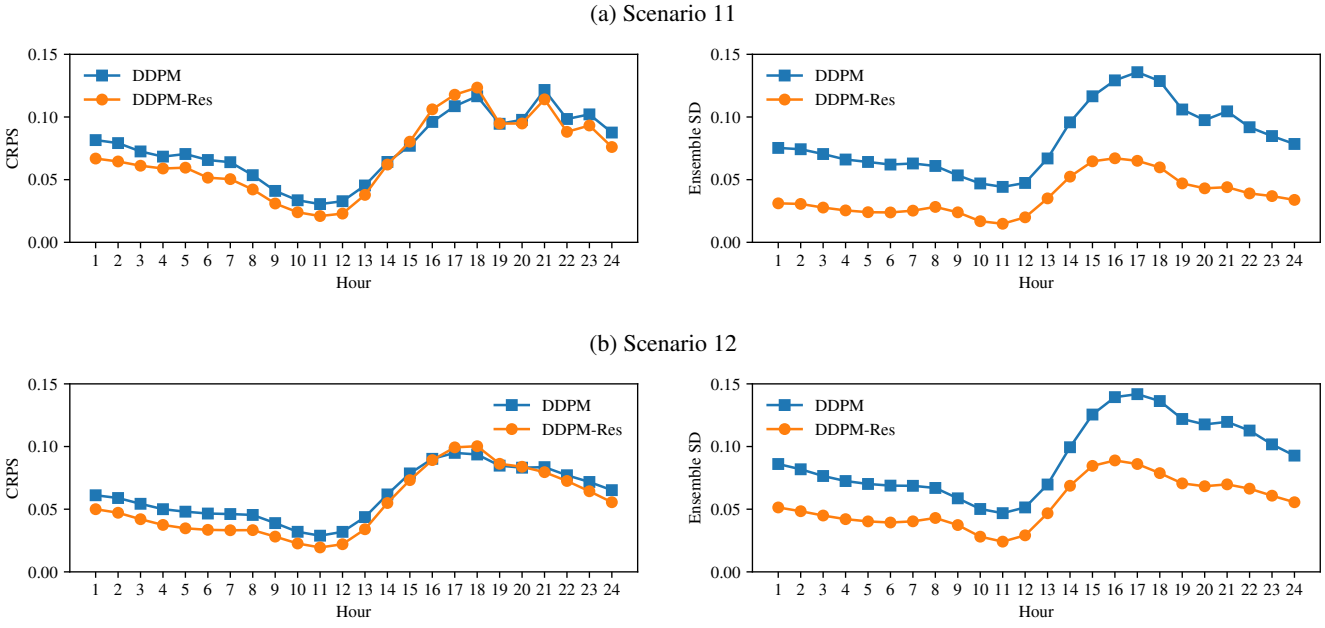


Figure D1. Hourly mean CRPS (left) and ensemble standard deviation (right) under (a) Scenario 11 and (b) Scenario 12.

Appendix E: Additional Farm-Level Prediction Comparisons

E1 Deterministic U-Net Models

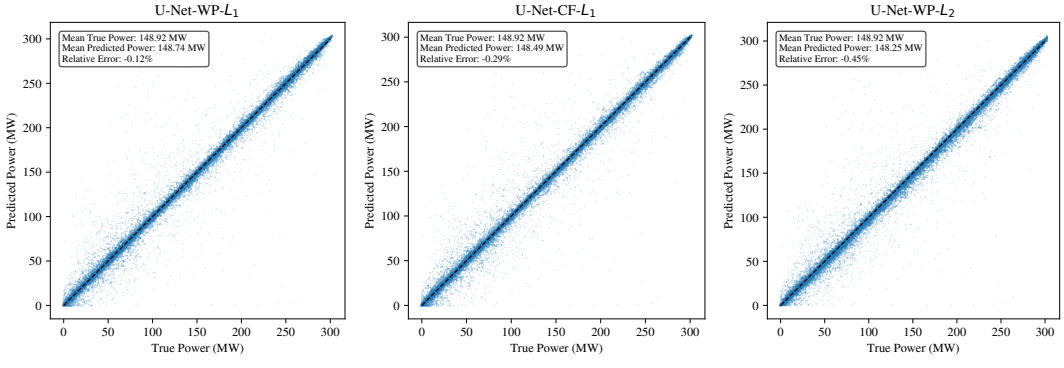


Figure E1. Farm-level scatter plots for Scenario 11 for U-Net models. Each point denotes a 10-min time step, comparing spatially aggregated predicted and ground-truth power. Text annotations show the annual mean power and the corresponding farm-level relative error.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

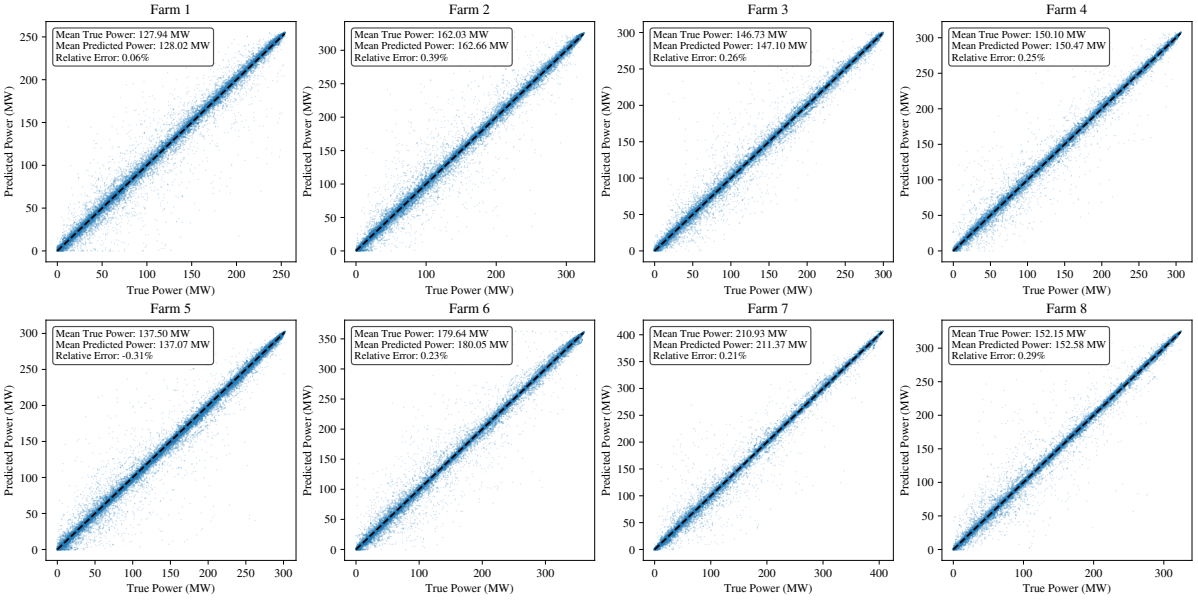


Figure E2. Farm-level scatter plots for Scenario 12 for U-Net-WP-L_1. Each point denotes a 10-min time step, comparing spatially aggregated predicted and ground-truth power. Text annotations show the annual mean power and the corresponding farm-level relative error.

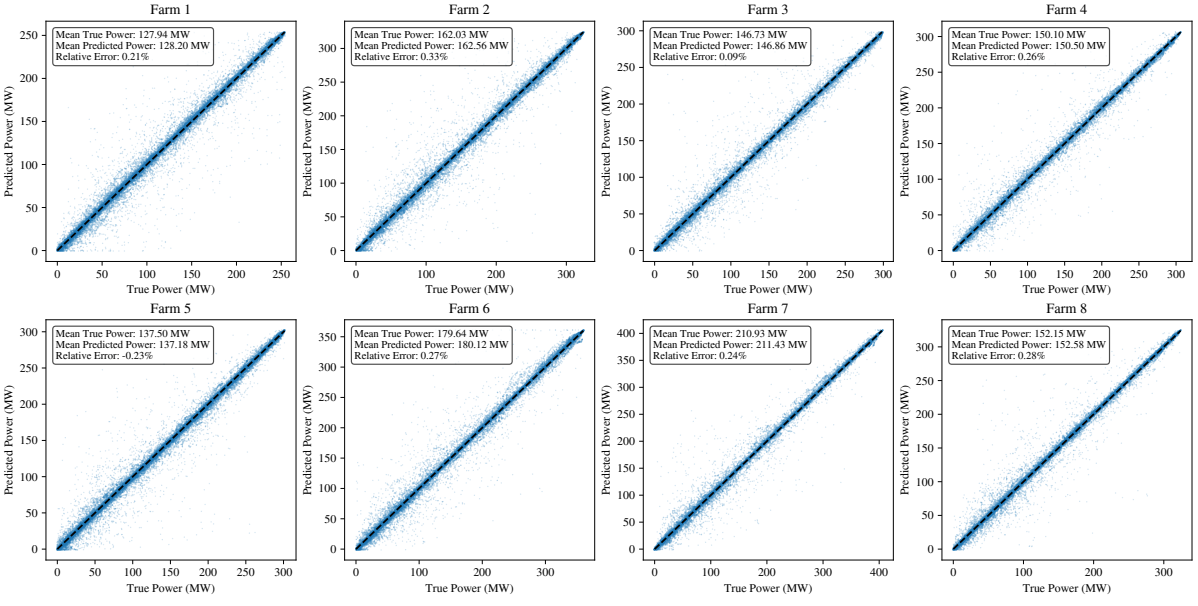


Figure E3. Farm-level scatter plots for Scenario 12 for U-Net-CF-L_1. Each point denotes a 10-min time step, comparing spatially aggregated predicted and ground-truth power. Text annotations show the annual mean power and the corresponding farm-level relative error.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

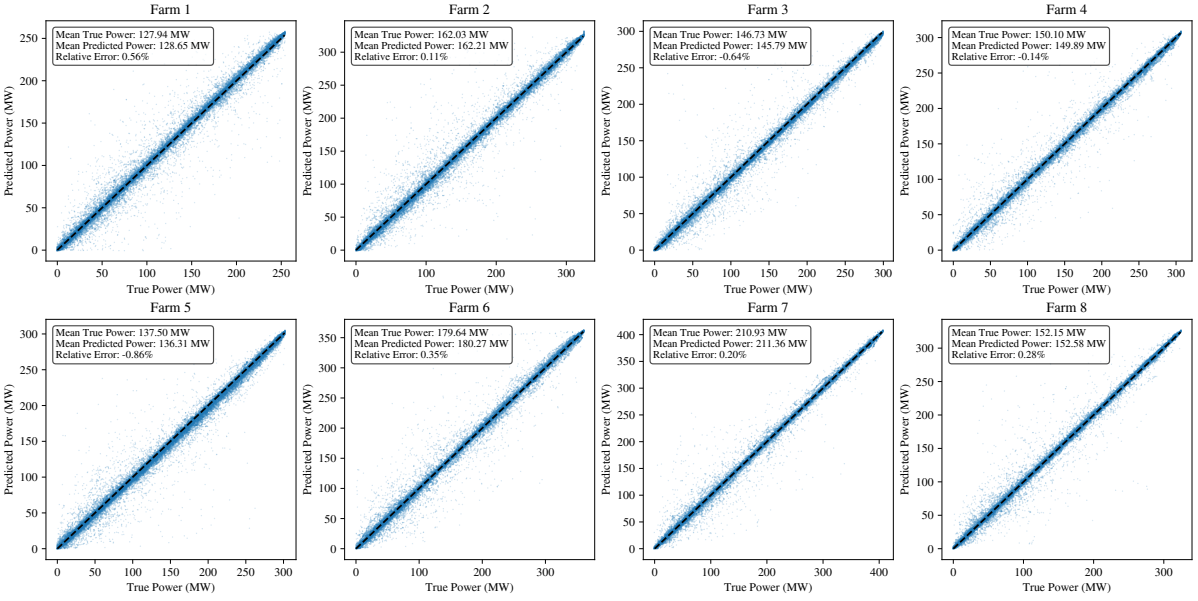


Figure E4. Farm-level scatter plots for Scenario 11 for U-Net-WP-L_2. Each point denotes a 10-min time step, comparing spatially aggregated predicted and ground-truth power. Text annotations show the annual mean power and the corresponding farm-level relative error.

E2 Probabilistic Diffusion Models

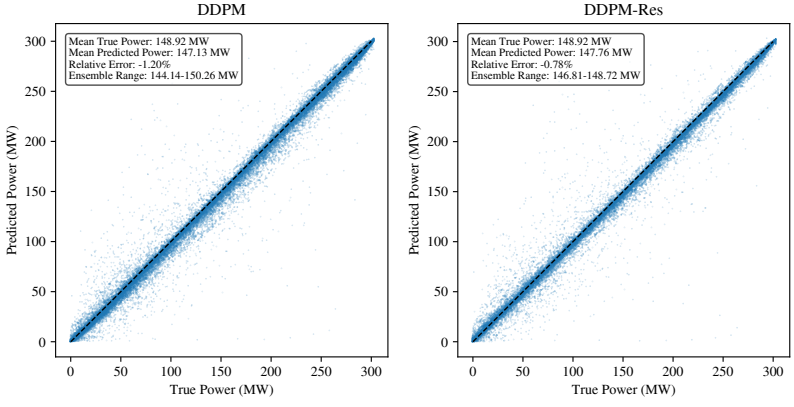


Figure E5. Farm-level scatter plot for Scenario 11 for diffusion models. Each point corresponds to a 10-min time step, showing spatially aggregated predicted power versus ground-truth power. Text annotations report the annual-mean power and the resulting farm-level relative error.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

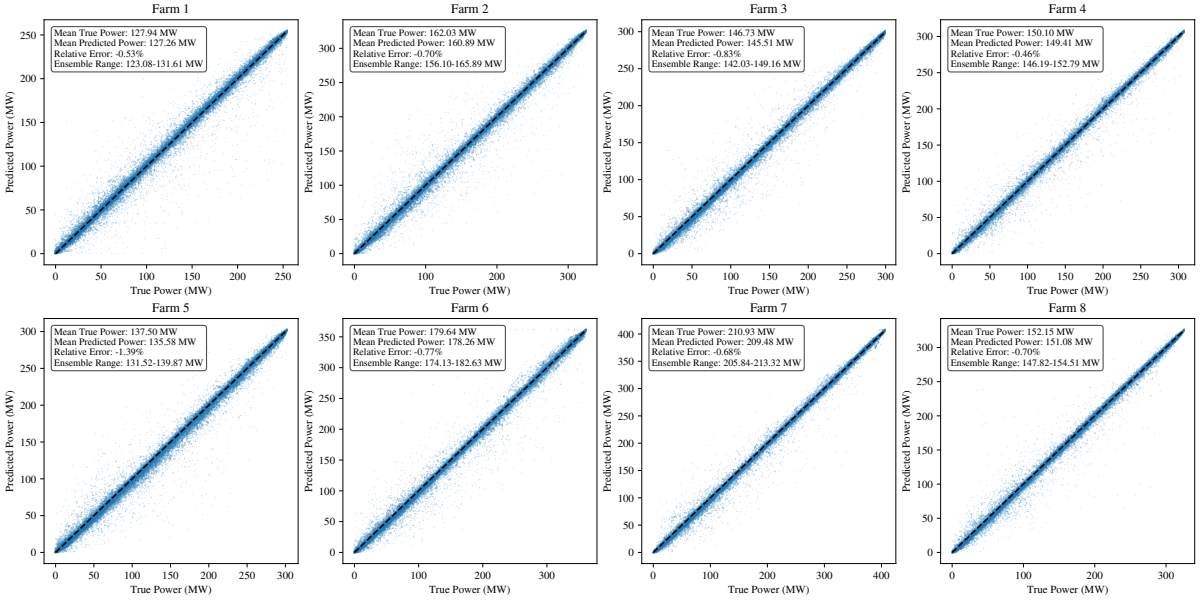


Figure E6. Farm-level scatter plots for Scenario 12 for DDPM. Each point corresponds to a 10-min time step, showing spatially aggregated predicted power versus true power. Text annotations report the annual-mean power and the resulting farm-level relative error.

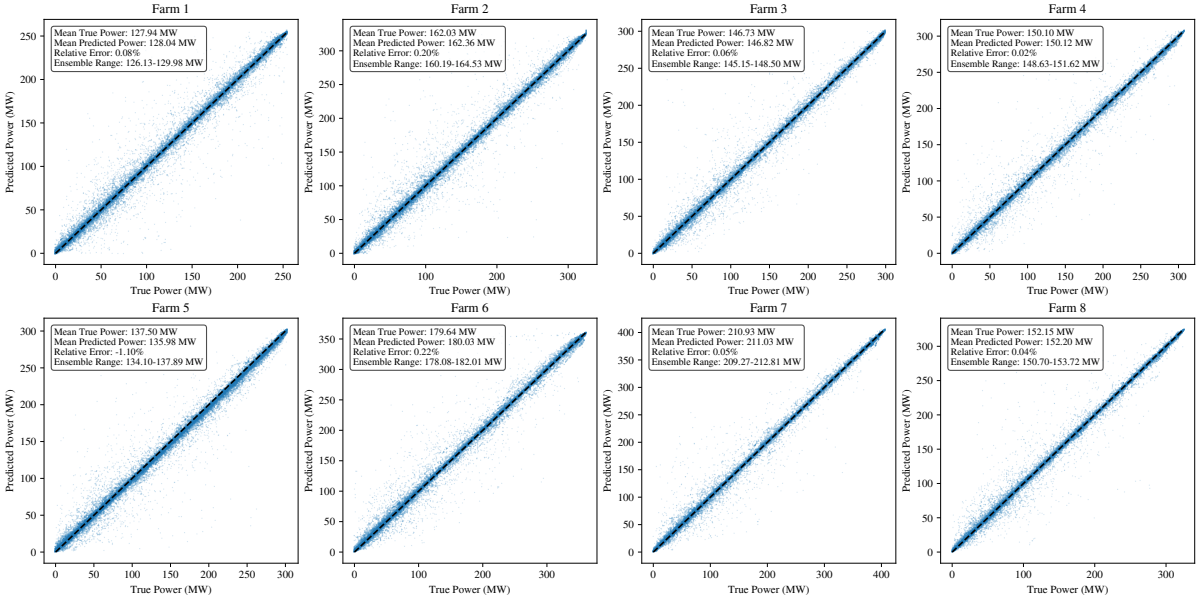


Figure E7. Farm-level scatter plots for Scenario 12 for DDPM-Res. Each point corresponds to a 10-min time step, showing spatially aggregated predicted power versus true power. Text annotations report the annual-mean power and the resulting farm-level relative error.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

Author contributions. All co-authors contributed significantly to this manuscript. Their individual contributions are described below using the CRediT taxonomy: TJ: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing; MO: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – review & editing; AHM: Conceptualization, Formal Analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – review & editing; SI: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Validation, Writing – review & editing.

Competing interests. This study uses WRF simulations generated through the WakeMap product developed by Veer Renewables, a for-profit consulting company. Mike Optis is the founder and president of this company. The other authors declare that they have no competing interests.

Acknowledgements. The authors gratefully acknowledge support from the Pacific Institute for the Mathematical Sciences. This work was also supported by Mitacs through the Mitacs Accelerate program. Computational resources for this research were provided in part by the Digital Research Alliance of Canada (alliancecan.ca), including access to the Fir cluster hosted at Simon Fraser University.

To improve the readability of the manuscript, the authors used ChatGPT during preparation of the text. All text was subsequently evaluated and revised by the authors, who take full responsibility for the final manuscript.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

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

References

Abdelsattar, M., A. Ismeil, M., Menoufi, K., AbdelMoety, A., and Emad-Eldeen, A.: Evaluating Machine Learning and Deep Learning models for predicting Wind Turbine power output from environmental factors, PLOS ONE, 20, e0317619, https://doi.org/10.1371/journal.pone.0317619, 2025.

Annau, N. J., Cannon, A. J., and Monahan, A. H.: Algorithmic Hallucinations of Near-Surface Winds: Statistical Downscaling with Generative Adversarial Networks to Convection-Permitting Scales, Artificial Intelligence for the Earth Systems, 2, https://doi.org/10.1175/aies-d-23-0015.1, 2023.

Arslan Tuncar, E., Şafak Sağlam, and Oral, B.: A review of short-term wind power generation forecasting methods in recent technological trends, Energy Reports, 12, 197–209, https://doi.org/10.1016/j.egyr.2024.06.006, 2024.

Bleeg, J.: A Graph Neural Network Surrogate Model for the Prediction of Turbine Interaction Loss, Journal of Physics: Conference Series, 1618, 062 054, https://doi.org/10.1088/1742-6596/1618/6/062054, 2020.

Chen, C.-F. R., Fan, Q., and Panda, R.: CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification , in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 347–356, IEEE Computer Society, Los Alamitos, CA, USA, https://doi.org/10.1109/ICCV48922.2021.00041, 2021.

Chen, F. and Dudhia, J.: Coupling an Advanced Land Surface–Hydrology Model with the Penn State–NCAR MM5 Modeling System. Part I: Model Implementation and Sensitivity, Monthly Weather Review, 129, 569–585, https://doi.org/10.1175/1520-0493(2001)129<0569:caalsh>2.0.co;2, 2001.

Chen, F. and Gao, L.: Learning Residual Distributions with Diffusion Models for Probabilistic Wind Power Forecasting, Energies, 18, 4226, https://doi.org/10.3390/en18164226, 2025.

Dimitrov, N.: Surrogate models for parameterized representation of wake-induced loads in wind farms, Wind Energy, 22, 1371–1389, https://doi.org/10.1002/we.2362, 2019.

Ferro, C. A. T.: Fair scores for ensemble forecasts: Fair Scores for Ensemble Forecasts, Quarterly Journal of the Royal Meteorological Society, 140, 1917–1923, https://doi.org/10.1002/qj.2270, 2013.

Fischereit, J., Brown, R., Larsén, X. G., Badger, J., and Hawkes, G.: Review of Mesoscale Wind-Farm Parametrizations and Their Applications, Boundary-Layer Meteorology, 182, 175–224, https://doi.org/10.1007/s10546-021-00652-y, 2021.

Fischereit, J., Schaldemose Hansen, K., Larsén, X. G., van der Laan, M. P., Réthoré, P.-E., and Murcia Leon, J. P.: Comparing and validating intra-farm and farm-to-farm wakes across different mesoscale and high-resolution wake models, Wind Energy Science, 7, 1069–1091, https://doi.org/10.5194/wes-7-1069-2022, 2022.

Fitch, A. C., Olson, J. B., Lundquist, J. K., Dudhia, J., Gupta, A. K., Michalakes, J., and Barstad, I.: Local and Mesoscale Impacts of Wind Farms as Parameterized in a Mesoscale NWP Model, Monthly Weather Review, 140, 3017–3038, https://doi.org/10.1175/mwr-d-11-00352.1, 2012.

Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., and Guo, B.: Efficient Diffusion Training via Min-SNR Weighting Strategy, https://doi.org/10.48550/ARXIV.2303.09556, 2023.

Harrison-Atlas, D., King, R. N., and Glaws, A.: Machine learning enables national assessment of wind plant controls with implications for land use, Wind Energy, 25, 618–638, https://doi.org/10.1002/we.2689, 2021.

Ho, J. and Salimans, T.: Classifier-Free Diffusion Guidance, https://arxiv.org/abs/2207.12598, 2022.

Ho, J., Jain, A., and Abbeel, P.: Denoising Diffusion Probabilistic Models, https://doi.org/10.48550/ARXIV.2006.11239, 2020.

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

Hong, S.-Y., Dudhia, J., and Chen, S.-H.: A Revised Approach to Ice Microphysical Processes for the Bulk Parameterization of Clouds and Precipitation, Monthly Weather Review, 132, 103–120, https://doi.org/10.1175/1520-0493(2004)132<0103:aratim>2.0.co;2, 2004.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D.: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2008jd009944, 2008.

Jiménez, P. A., Dudhia, J., González-Rouco, J. F., Navarro, J., Montávez, J. P., and García-Bustamante, E.: A Revised Scheme for the WRF Surface Layer Formulation, Monthly Weather Review, 140, 898–918, https://doi.org/10.1175/mwr-d-11-00056.1, 2012.

Kain, J. S.: The Kain–Fritsch Convective Parameterization: An Update, Journal of Applied Meteorology, 43, 170–181, https://doi.org/10.1175/1520-0450(2004)043<0170:tkcpau>2.0.co;2, 2004.

Kain, J. S. and Fritsch, J. M.: A One-Dimensional Entraining/Detraining Plume Model and Its Application in Convective Parameterization, Journal of the Atmospheric Sciences, 47, 2784–2802, https://doi.org/10.1175/1520-0469(1990)047<2784:aodepm>2.0.co;2, 1990.

King, R., Glaws, A., Geraci, G., and Eldred, M. S.: A Probabilistic Approach to Estimating Wind Farm Annual Energy Production with Bayesian Quadrature, in: AIAA Scitech 2020 Forum, American Institute of Aeronautics and Astronautics, https://doi.org/10.2514/6.2020-1951, 2020.

Li, S., Robert, A., Faisal, A. A., and Piggott, M. D.: Learning to optimise wind farms with graph transformers, Applied Energy, 359, 122 758, https://doi.org/10.1016/j.apenergy.2024.122758, 2024.

Ling, F., Lu, Z., Luo, J.-J., Bai, L., Behera, S. K., Jin, D., Pan, B., Jiang, H., and Yamagata, T.: Diffusion model-based probabilistic downscaling for 180-year East Asian climate reconstruction, npj Climate and Atmospheric Science, 7, https://doi.org/10.1038/s41612-024-00679-1, 2024.

Liu, H. and Zhang, Z.: Development and trending of deep learning methods for wind power predictions, Artificial Intelligence Review, 57, https://doi.org/10.1007/s10462-024-10728-z, 2024.

Liu, Z., Guo, H., Zhang, Y., and Zuo, Z.: A Comprehensive Review of Wind Power Prediction Based on Machine Learning: Models, Applications, and Challenges, Energies, 18, 350, https://doi.org/10.3390/en18020350, 2025.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J.: DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models, Machine Intelligence Research, 22, 730–751, https://doi.org/10.1007/s11633-025-1562-4, 2025.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Nabian, M. A., Ge, T., Subramaniam, A., Kashinath, K., Kautz, J., and Pritchard, M.: Residual corrective diffusion modeling for km-scale atmospheric downscaling, Communications Earth & Environment, 6, https://doi.org/10.1038/s43247-025-02042-5, 2025.

Nakanishi, M. and Niino, H.: An Improved Mellor–Yamada Level-3 Model with Condensation Physics: Its Design and Verification, Boundary-Layer Meteorology, 112, 1–31, https://doi.org/10.1023/b:boun.0000020164.04146.98, 2004.

Nakanishi, M. and Niino, H.: Development of an Improved Turbulence Closure Model for the Atmospheric Boundary Layer, Journal of the Meteorological Society of Japan. Ser. II, 87, 895–912, https://doi.org/10.2151/jmsj.87.895, 2009.

Padrón, A. S., Thomas, J., Stanley, A. P. J., Alonso, J. J., and Ning, A.: Polynomial chaos to efficiently compute the annual energy production in wind farm layout optimization, Wind Energy Science, 4, 211–231, https://doi.org/10.5194/wes-4-211-2019, 2019.

Pandit, R. and Wang, J.: A comprehensive review on enhancing wind turbine applications with advanced SCADA data analytics and practical insights, IET Renewable Power Generation, 18, 722–742, https://doi.org/10.1049/rpg2.12920, 2024.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A.: FiLM: visual reasoning with a general conditioning layer, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference

https://doi.org/10.5194/wes-2026-87
Preprint. Discussion started: 1 June 2026

and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18, AAAI Press, ISBN 978-1-57735-800-8, 2018.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M.: Probabilistic weather forecasting with machine learning, Nature, 637, 84–90, https://doi.org/10.1038/s41586-024-08252-9, 2024.

Pryor, S. C., Coburn, J. J., Barthelmie, R. J., and Shepherd, T. J.: Projecting Future Energy Production from Operating Wind Farms in North America. Part I: Dynamical Downscaling, Journal of Applied Meteorology and Climatology, 62, 63–80, https://doi.org/10.1175/jamc-d-22-0044.1, 2023.

Rajaperumal, T. A. and Christopher Columbus, C.: Enhanced wind power forecasting using machine learning, deep learning models and ensemble integration, Scientific Reports, 15, https://doi.org/10.1038/s41598-025-05250-3, 2025.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, https://doi.org/10.48550/ARXIV.1505.04597, 2015.

Salimans, T. and Ho, J.: Progressive Distillation for Fast Sampling of Diffusion Models, https://doi.org/10.48550/ARXIV.2202.00512, 2022.

Schicker, I., Ganglbauer, J., Dabernig, M., and Nacht, T.: Wind power estimation on local scale—A case study of representativeness of reanalysis data and data-driven analysis, Frontiers in Climate, 5, https://doi.org/10.3389/fclim.2023.1017774, 2023.

Slot, R. M., Sørensen, J. D., Sudret, B., Svenningsen, L., and Thøgersen, M. L.: Surrogate model uncertainty in wind turbine reliability assessment, Renewable Energy, 151, 1150–1162, https://doi.org/10.1016/j.renene.2019.11.101, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations, https://doi.org/10.48550/ARXIV.2011.13456, 2020.

Ti, Z., Deng, X. W., and Yang, H.: Wake modeling of wind turbines using machine learning, Applied Energy, 257, 114 025, https://doi.org/10.1016/j.apenergy.2019.114025, 2020.

Wang, Y., Zou, R., Liu, F., Zhang, L., and Liu, Q.: A review of wind speed and wind power forecasting with deep neural networks, Applied Energy, 304, 117 766, https://doi.org/10.1016/j.apenergy.2021.117766, 2021.

Wang, Z., Tu, Y., Zhang, K., Han, Z., Cao, Y., and Zhou, D.: An optimization framework for wind farm layout design using CFD-based Kriging model, Ocean Engineering, 293, 116 644, https://doi.org/10.1016/j.oceaneng.2023.116644, 2024.

Wu, Z., Luo, G., Yang, Z., Guo, Y., Li, K., and Xue, Y.: A comprehensive review on deep learning approaches in wind forecasting applications, CAAI Transactions on Intelligence Technology, 7, 129–143, https://doi.org/10.1049/cit2.12076, 2022.

Xia, G., Optis, M., Deskos, G., Sinner, M., Mulas Hernando, D., Lundquist, J. K., Kumler, A., Sanchez Gomez, M., Fleming, P., and Musial, W.: Understanding Cluster Wake-Induced Energy Losses off the U.S. East Coast, https://doi.org/10.5194/wes-2025-154, preprint, 2025.

Yu, D., Li, X., Ye, Y., Zhang, B., Luo, C., Dai, K., Wang, R., and Chen, X.: DiffCast: A Unified Framework via Residual Diffusion for Precipitation Nowcasting, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), p. 27758–27767, IEEE, https://doi.org/10.1109/cvpr52733.2024.02622, 2024.

Zhang, X., Glaws, A., Cortiella, A., Emami, P., and King, R. N.: Deep generative models in energy system applications: Review, challenges, and future directions, Applied Energy, 380, 125 059, https://doi.org/10.1016/j.apenergy.2024.125059, 2025.

zum Berge, K., Centurelli, G., Dörenkämper, M., Bange, J., and Platis, A.: Evaluation of Engineering Models for Large-Scale Cluster Wakes With the Help of In Situ Airborne Measurements, Wind Energy, 27, 1040–1062, https://doi.org/10.1002/we.2942, 2024.