Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Decomposition-informed deep learning for wind-power forecasting: A CEEMDAN→VMD hybrid with feature extraction and per-component deeplearners

Cossi Télesphore NOUNANGNONHOU
Laboratory of Electrotechnics, Telecommunications and Applied Informatics (LETIA), University of Abomey-Calavi
Kossoko Babatoundé Audace DIDAVI
didavia@gmail.com

Laboratory of Electrotechnics, Telecommunications and Applied Informatics (LETIA), University of Abomey-Calavi
Maurel Richy AZA-GNANDJI
Laboratory of Rural Engineering, National University of Agriculture

Research Article

Keywords: Wind power forecasting, CEEMDAN, Variational Mode Decomposition (VMD), Deep learning, Hybrid models, Spinning reserve reduction, Renewable energy systems

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9856569/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Decomposition-informed deep learning for wind-power forecasting: A CEEMDAN→VMD hybrid with feature extraction and per-component deep learners

Cossi Télesphore NOUNANGNONHOU[1],
Kossoko Babatoundé Audace DIDAVI[1*†],
Maurel Richy AZA-GNANDJI[2†]

[1*]University of Abomey-Calavi, Laboratory of Electrotechnics, Telecommunications and Applied Informatics (LETIA), Abomey-Calavi, Benin.
[2]National University of Agriculture, Laboratory of Rural Engineering, Kétou, Benin.

*Corresponding author(s). E-mail(s): didavia@gmail.com;
Contributing authors: nocteles2000@gmail.com; maurel.aza@gmail.com;
†These authors contributed equally to this work.

Abstract

Accurate wind power forecasting remains a critical challenge for grid stability and renewable energy integration, particularly in emerging power systems with limited flexibility. This study proposes a hybrid decomposition–learning framework that combines the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Variational Mode Decomposition (VMD) with a deep gated recurrent network (B6-CEEMDAN-VMD). The dual decomposition enables the extraction of multi-scale temporal features, while the GRU–attention backbone captures nonlinear dependencies and temporal dynamics across horizons. The proposed model was evaluated on data from seven wind farms using 1-hour, 4-hour, and 24-hour forecast horizons and benchmarked against persistence, linear regression, ARIMA, gradient boosting (GBM), and standalone deep models. Results demonstrate that B6 outperforms all baselines, reducing mean absolute error (MAE) and root mean squared error (RMSE) by 25–35% compared to GBM and by more than 40% compared to persistence forecasts. The model also exhibits strong spatial robustness, maintaining low forecasting errors

across geographically diverse sites. Operationally, these accuracy gains translate into a 26–30% reduction in spinning reserve requirements for a 100 MW grid, contributing to more efficient dispatch, reduced curtailment, and lower backup fuel costs. Overall, the CEEMDAN–VMD–Deep hybridization provides a scalable, physics-consistent, and data-efficient solution for improving short-term and day-ahead wind forecasting in emerging renewable energy systems

Keywords: Wind power forecasting, CEEMDAN, Variational Mode Decomposition (VMD), Deep learning, Hybrid models, Spinning reserve reduction, Renewable energy systems

1 Introduction

Wind energy has become a cornerstone of decarbonisation, with successive record years of new capacity additions and more than one terawatt cumulatively installed worldwide. Industry reports emphasise that annual installations must continue to scale to meet 2030 targets, which heightens the operational need for accurate short- and medium-term wind power forecasts to support unit commitment, reserve scheduling, market bids, and congestion management [1, 2]. In power-system operations, forecast errors translate directly into imbalance costs, suboptimal reserve allocation, and reliability risks. Hence, operator-relevant improvements in point accuracy (e.g., MAE, RMSE, sMAPE) and robust uncertainty quantification are of first-order importance.

Wind generation is governed by atmospheric processes that are *nonstationary*, *multi-scale*, and strongly site-dependent. Diurnal cycles, synoptic variability, turbulence, wake interactions, and terrain heterogeneity imprint rich temporal structures with regime shifts and intermittency. These characteristics strain both classical statistical models (persistence, AR/ARIMA/ETS) and purely physical numerical weather prediction (NWP) when run at fine temporal scales. Meanwhile, purely data-driven deep models can overfit or generalise poorly across farms if the representation does not encode physical context or multi-scale structure. Recent surveys consistently report that successful systems couple data-driven learners with principled pre-processing, feature design, and evaluation protocols [3, 4].

Hybrid pipelines mitigate the above issues by combining (i) *decomposition* to reduce nonstationarity and isolate band-limited components, (ii) *feature engineering* to encode physics and context (temporal statistics, spectral energies, entropy/complexity, exogenous meteorology), and (iii) *deep sequence models* to capture long- and short-range temporal dependencies. Decomposition methods such as Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) and Variational Mode Decomposition (VMD) have proven effective in wind forecasting, especially when paired with CNN–BiLSTM, GRU/TCN, or attention mechanisms [5**?** , 6]. Beyond accuracy gains, decomposition-informed learners improve interpretability at the component level and enable targeted regularisation (e.g., per-band architectures and losses).

Despite rapid progress, the literature still shows (i) fragmented comparisons across sites and horizons, (ii) limited *ablations* that isolate the contribution of each block (decomposition, exogenous features, attention), and (iii) a lack of formal *statistical testing* of performance differences on public benchmarks and multiple horizons. Community anchors such as the wind track of GEFCom2012 and the NREL WIND Toolkit remain underused for systematic, multi-horizon studies with rigorous significance testing and reproducibility artefacts [7, 8].

To address these gaps, we make four contributions:

1. **Two-stage, decomposition-informed pipeline.** We propose CEEMDAN→VMD followed by systematic feature engineering and a per-component deep sequence learner (CNN–BiLSTM or TCN with attention), with physically consistent aggregation and clipping to the rated power.
2. **Multi-horizon evaluation.** We assess 1 h, 4 h, and 24 h horizons using operator-relevant metrics (MAE, RMSE, sMAPE, normalised errors), under strict chronological splits and no-leakage normalisation.
3. **Comprehensive ablations.** We quantify the marginal value of each block by removing CEEMDAN, VMD, exogenous features, and attention; we also contrast classical baselines (persistence, ARIMA, gradient boosting) with modern deep models (no decomposition).
4. **Reproducibility pack.** We release code, exact splits, seeds, and data-access instructions for public datasets (GEFCom and NREL), and we report statistical significance via the Diebold–Mariano test with appropriate small-sample corrections [9].

Section 2 reviews recent wind-power forecasting advances with emphasis on hybrid, decomposition-informed methods. Section 3 describes datasets (GEFCom wind track; NREL WIND Toolkit subsets), tasks, and preprocessing. Section 4 details our two-stage decomposition, feature extraction, and deep sequence architectures with aggregation. Section 5 specifies baselines, hyperparameter protocol, metrics, and statistical tests, plus ablation design. Section 6 reports multi-horizon results and significance analyses, including per-site robustness. Section **??** discusses limitations (parameter sensitivity, compute), generalisation (site transfer), and deployment. Section **??** concludes and outlines future work (probabilistic forecasts, spatio-temporal graphs, physics-informed constraints).

2 Related work

Early wind-power forecasting relied on persistence and AR/ARIMA/ETS, often post-processing NWP; these baselines struggle with nonlinear, regime-switching dynamics and site heterogeneity. Recent surveys document the shift toward data-driven ML (RF/GBM/SVM) and deep sequence models (LSTM/GRU, CNN–RNN, TCN, attention/transformers), emphasising leakage-free evaluation and multi-horizon reporting [3, 4, 10]. Canonical benchmarks such as *GEFCom2012* underpinned protocol standardisation and comparability [7].

To cope with multi-scale variability and nonstationarity, many works apply signal decomposition before learning. Empirical/ensemble methods (EMD/EEMD/CEEMDAN) and variational decompositions (VMD) isolate band-limited patterns that are easier to predict and interpret. Recent hybrids consistently report MAE/RMSE gains when combining CEEMDAN or CEEMDAN→VMD with deep learners (LSTM/GRU, CNN–BiLSTM, TCN/attention) [5**?** , 6]. These pipelines reduce mode mixing, stabilise amplitudes, and enable per-band modelling.

Deep architectures with long receptive fields and attention (CNN–BiLSTM, GRU with dual attention, TCN+attention) remain strong contenders for short/medium horizons, provided that inputs encode physics and context via engineered features (time/frequency/entropy) and, when available, exogenous meteorology [3, 11]. Decomposition-informed designs often train one learner per mode and aggregate, improving error attribution and regularisation.

Best practice calls for strict chronological splits, leakage-proof preprocessing/normalisation, multi-horizon metrics (MAE/RMSE/sMAPE + normalised variants) and formal significance tests. The Diebold–Mariano (DM) test on loss differentials is widely used to assert meaningful improvements beyond noise [9]. Our study follows these principles and consolidates fragmented comparisons with a controlled ablation of each block (decomposition, features, attention).

3 Datasets and problem setup

We use the wind track of the *Global Energy Forecasting Competition 2012* (GEFCom2012) as our sole dataset [7]. It provides **hourly** ($\Delta t = 1\,\mathrm{h}$) anonymised wind **power** time series at the wind-farm level over the official competition horizon (continuous segments spanning multiple months). Only the target variable (active power at plant level) and timestamps are required for our study; no site-identifying metadata are used. The official release is timestamp-regular by construction; in our parsing we found no structural gaps in timestamps. Any isolated anomalies (e.g., negative values or spikes) are treated via conservative quality control: flags, capped forward-fill for *short* gaps only, and masking of *long* gaps so they do not contribute to the loss.

We formulate **deterministic point forecasting** at **1 h**, **4 h**, and **24 h** ahead. For each wind farm k, the series $\{Y_t^{(k)}\}_{t=1}^{T_k}$ is partitioned **chronologically** into *train*/*validation*/*test*, preserving the original order to avoid regime mixing. All preprocessing statistics, decomposition hyperparameters, and model hyperparameters are fit on *train*, selected on *validation*, and frozen on *test*. **Strict no-leakage:** any rolling transformation (features, decompositions) is computed within each split; windows never cross split boundaries; normalisation parameters are estimated on *train only* and applied unchanged to *validation*/*test*.

Our preferred target is **wind power** $Y_t \in [0, P_{\mathrm{rated}}]$ as provided in GEFCom2012. If only **wind speed** were available (not the case here), we would either (i) convert to power via a turbine/farm *power curve* when documented, or (ii) evaluate forecasting on speed and report both speed and power metrics (the latter via a calibrated speed→power mapping). To enable scale-invariant comparisons across anonymised

farms, we also report normalised errors (NMAE/NRMSE) using a capacity proxy $\widehat{P}_{\max}$ estimated from *train*.

To align with dispatch practice (intra-day re-dispatch and day-ahead scheduling), we design the pipeline for **low-latency** updates and **bounded compute**: (i) feature extraction and decomposition (CEEMDAN→VMD) are executed in a streaming-compatible fashion (incremental updates of recent windows); (ii) per-horizon inference uses compact deep sequence models (CNN–BiLSTM or TCN+attention) whose forward pass completes well under typical control-room cycles (order of seconds on a CPU core, sub-second on a modern GPU, per farm/horizon); (iii) re-fitting of hyperparameters is offline (batch) and never performed during test-time forecasting; (iv) all external inputs are intra-series. These constraints reflect practical deployment where forecasts must be delivered *ahead of issuance deadlines* with predictable runtime and memory footprints.

4 Methodology

4.1 Research workflow overview

Figure 1 summarises the end-to-end pipeline used in this study, from raw data ingestion to evaluation and reporting. The workflow is organised as a sequence of modular blocks so that each component (data preprocessing, feature design, baseline learners, hybrid decomposition, and deep sequence modelling) can be independently ablated or replaced without changing the rest of the pipeline. This design is essential for reproducibility and for attributing performance gains to specific modelling choices.

We fuse four data families: (i) *SCADA* active power (y), turbine status and quality flags; (ii) *exogenous* numerical weather prediction (NWP: wind speed/direction, temperature, pressure) when available; (iii) *metadata* (site identifier, coordinates, hub height); and (iv) optional maintenance logs. Streams are merged by timestamp, the timezone is unified, duplicates are removed, data are resampled to hourly resolution, and obvious outliers or turbine downtime are filtered using IQR/MAD rules. All splits are strictly chronological per site to avoid leakage.

Normalisation (z-score) is fit on the *train* portion only. We construct a compact yet expressive feature set: autoregressive lags at $\{1, 2, 3, 6, 12, 24, 48, 72, 168\}$ h; rolling mean/standard deviation/slope over 12/24/72 h windows; calendar encodings (hour-of-day and day-of-week via sine/cosine); and spectral descriptors (energy in three bands and spectral entropy). If NWP is present, it is appended as additional covariates. Labels for horizons $h \in \{1, 4, 24\}$ are created with forward shifts.

We train four classical baselines under identical data splits: B1 *Persistence*; B2 *Linear* (Ridge/ElasticNet on lagged features); B3 *ARIMA/SARIMAX* with seasonal hour-24; and B4 *GBM* (LightGBM or XGBoost, with GBR fallback). These provide transparent references for skill computation and statistical tests.

The right branch performs CEEMDAN decomposition followed by VMD to obtain interpretable oscillatory modes. For each retained mode we build a sequence tensor (window 72–168 h) and train a GRU encoder with optional additive attention; horizon-specific heads produce $\hat{y}_{t+h}$. Predictions are aggregated across modes to form

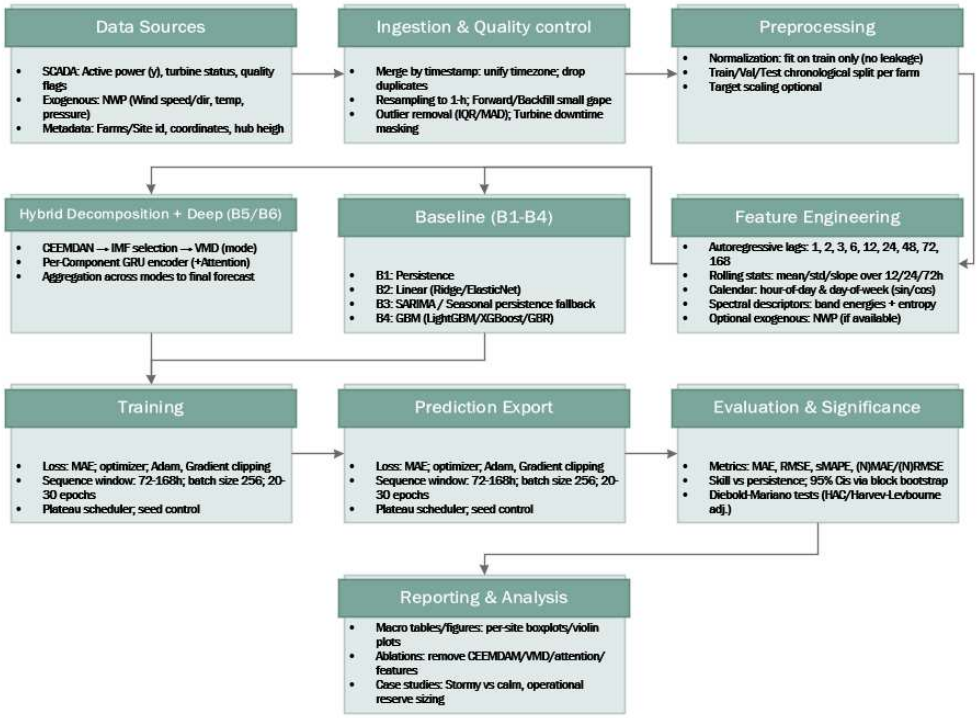


Fig. 1 Detailed research workflow for the hybrid wind-power forecasting pipeline.

the B6 forecast. B5 corresponds to the same deep architecture without decomposition. Training uses Adam, MAE loss, gradient clipping, mini-batches, and a plateau scheduler.

For every site, model, and horizon, we export test-set trajectories and compute MAE, RMSE, sMAPE, and their normalised variants (NMAE/NRMSE) with 95% block-bootstrap confidence intervals. Skill is reported relative to persistence. Pairwise Diebold–Mariano tests (HAC/Newey–West variance with Harvey–Leybourne small-sample correction) assess statistical significance at each horizon.

Macro tables aggregate results across sites; per-site distributions (box/violin) quantify geographic robustness; ablations isolate the contribution of CEEMDAN, VMD, attention, and feature blocks; and qualitative case studies (calm vs. stormy days) illustrate short-term vs. day-ahead behaviour. Operational implications are estimated via a back-of-the-envelope reserve-sizing analysis.

The modular pipeline in Fig. 1 ensures that our main claim—**B6 (CEEMDAN → VMD + GRU)** improves forecast accuracy over all baselines—can be tested rigorously under consistent data handling, splits, and metrics, and that improvement sources can be traced through controlled ablations.

4.2 Two-stage decomposition

We first stabilise the hourly wind-power series $x(t)$ by a CEEMDAN stage that yields intrinsic mode functions (IMFs) and a residual; then we apply VMD on selected IMFs to isolate quasi–band-limited modes. The final reconstruction is

$$x(t) \;\approx\; \sum_{i=1}^{M}\sum_{k=1}^{K_i} v_{i,k}(t) \;+\; r(t), \tag{1}$$

where M is the number of CEEMDAN IMFs, K_i the number of VMD modes extracted from IMF i, $v_{i,k}$ the k-th mode of IMF i, and $r(t)$ the CEEMDAN residual.

Stage A — CEEMDAN (Complete Ensemble EMD with Adaptive Noise) [12].

Given $x(t)$, generate R noisy realisations

$$x^{(r)}(t) \;=\; x(t) + \epsilon\,\sigma_x\, w^{(r)}(t), \qquad r = 1,\dots,R, \tag{2}$$

with $w^{(r)}$ i.i.d. zero-mean white noise and σ_x a scale (e.g., std(x)). The first IMF is the ensemble average of the first sifted modes:

$$\mathrm{IMF}_1(t) = \frac{1}{R}\sum_{r=1}^{R} \mathrm{emd1}\big(x^{(r)}(t)\big), \tag{3}$$

$$r_1(t) = x(t) - \mathrm{IMF}_1(t). \tag{4}$$

Iterate on the residual:

$$\mathrm{IMF}_m(t) = \frac{1}{R}\sum_{r=1}^{R} \mathrm{emd1}\big(r_{m-1}(t) + \epsilon\,\sigma_{r_{m-1}}\, w^{(r)}(t)\big), \tag{5}$$

$$r_m(t) = r_{m-1}(t) - \mathrm{IMF}_m(t), \qquad m = 2,\dots,M. \tag{6}$$

Stopping. Stop when r_m is nearly monotonic or $\mathrm{Var}(r_m)/\mathrm{Var}(x) \le \delta$ (e.g., $\delta \in [10^{-3}, 10^{-2}]$), or when $m = M_{\max}$.

Typical hyperparameters. Noise ratio $\epsilon \in [0.1, 0.2]$, ensemble size $R \in \{50, 100\}$, maximum IMFs $M_{\max} \in \{6, 8, 10\}$; sifting SD-threshold $\sim 10^{-3}$–10^{-2}.

IMF selection for Stage B.

Not all IMFs require refinement. We select a subset $\mathcal{I} \subseteq \{1,\dots,M\}$ by (i) energy share

$$\mathrm{ES}_i \;=\; \frac{\sum_t \mathrm{IMF}_i(t)^2}{\sum_t x(t)^2} \;>\; \eta \;\; (\text{e.g., } \eta = 0.05), \tag{7}$$

and (ii) spectral broadness via spectral entropy

$$H_i \;=\; -\sum_m p_i(\omega_m)\log p_i(\omega_m), \quad p_i(\omega_m) = \frac{|\mathcal{F}\{\mathrm{IMF}_i\}(\omega_m)|^2}{\sum_m |\mathcal{F}\{\mathrm{IMF}_i\}(\omega_m)|^2}, \tag{8}$$

selecting IMFs with H_i above a percentile threshold, plus (iii) observed error reduction on validation when the IMF is included.

Stage B — VMD (Variational Mode Decomposition) [13].

For each selected IMF $y(t) = \mathrm{IMF}_i(t)$, VMD finds K_i narrow-band modes $\{u_k\}$ around centre frequencies $\{\omega_k\}$ by solving

$$\min_{\{u_k\},\{\omega_k\}} \sum_{k=1}^{K_i} \left\| \partial_t \left[\left(\delta + \tfrac{j}{\pi t}\right) * u_k(t)\; e^{-j\omega_k t} \right] \right\|_2^2 \quad \text{s.t.} \quad \sum_{k=1}^{K_i} u_k(t) = y(t). \tag{9}$$

Using an augmented Lagrangian and ADMM in the Fourier domain (hat denotes DFT),

$$\widehat{u}_k^{(n+1)}(\omega) = \frac{\widehat{y}(\omega) - \sum_{q\neq k} \widehat{u}_q^{(n)}(\omega) + \frac{1}{2}\widehat{\lambda}^{(n)}(\omega)}{1 + 2\alpha\left(\omega - \omega_k^{(n)}\right)^2}, \tag{10}$$

$$\omega_k^{(n+1)} = \frac{\int_0^\infty \omega \left|\widehat{u}_k^{(n+1)}(\omega)\right|^2 d\omega}{\int_0^\infty \left|\widehat{u}_k^{(n+1)}(\omega)\right|^2 d\omega}, \tag{11}$$

$$\widehat{\lambda}^{(n+1)}(\omega) = \widehat{\lambda}^{(n)}(\omega) + \tau\left(\widehat{y}(\omega) - \sum_{k=1}^{K_i} \widehat{u}_k^{(n+1)}(\omega)\right), \tag{12}$$

with bandwidth penalty $\alpha > 0$ and dual step $\tau > 0$. Convergence when

$$\frac{\sum_k \|\widehat{u}_k^{(n+1)} - \widehat{u}_k^{(n)}\|_2^2}{\|\widehat{y}\|_2^2} \;<\; \varepsilon \quad (\text{e.g., } \varepsilon \in [10^{-7}, 10^{-5}]). \tag{13}$$

Set $v_{i,k}(t) = u_k(t)$.

Typical hyperparameters. $K_i \in \{2,3,4,5\}$, $\alpha \in \{10^2, 10^3, 10^4\}$, max iterations $N_{\mathrm{it}} \in [500, 2000]$; initialise ω_k uniformly on $(0,\pi)$; use mirrored boundaries.

Practical notes.

(i) CEEMDAN/VMD are *fit per split* (train/validation) and frozen for test to avoid leakage; (ii) only IMFs in $\mathcal{I}$ are refined to control compute; (iii) per-mode features are computed on $\{v_{i,k}\}$ and combined downstream (Sec. 4.3); (iv) (1) is used to check numerical stability (small residual mismatch).

4.3 Feature extraction

We build a compact, leakage-proof feature set around the two-stage decomposition (Sec. 4.2), optimised for low-latency updates.

Per-mode summaries. For each component $z \in \{v_{i,k}, r\}$ over a short window L (e.g., 12–24 h): local mean $\mu_{z,t}$, variance $\sigma^2_{z,t}$, finite differences $\Delta z_t^{(d)} = z_t - z_{t-d}$ ($d \in \{1, 2, 6, 12, 24\}$) and least-squares slope (trend). These stationarising summaries stabilise learning across scales.

Short memory & spectral cues. Sample autocorrelations at a few lags $\mathcal{L}_{\text{ac}} = \{1, 2, 6, 12, 24\}$, and Hann-windowed short DFT energies aggregated into *three* interpretable bands (sub-diurnal, diurnal, multi-day). For narrow-band $v_{i,k}$, band energies act as amplitude proxies at negligible compute.

Entropy proxy. A lightweight sample entropy on z (window L, (m, r) fixed on *train*) captures irregularity without heavy transforms; values are computed after standardisation.

Hilbert-envelope (optional). For selected narrow-band modes, we use the analytic-signal magnitude $a_{i,k}(t) \approx |v_{i,k}(t) + j\mathcal{H}\{v_{i,k}\}(t)|$ and summarise it (mean/slope). This improves amplitude tracking with minimal cost.

Calendar encodings. Cyclical hour-of-day/day-of-week via sine–cosine pairs; monthly terms only if ACF/PACF shows seasonality.

Normalisation and compactness. All scalers, thresholds and band definitions are fit on *train* (selected on *val*) and frozen. We normalise by a capacity proxy $\widehat{P}_{\max}$ (e.g., train 99.5th percentile) for scale-free errors (NMAE/NRMSE). Feature count is kept to a few dozen per mode via correlation screening / MI and optional small PCA on per-mode blocks (fit on *train* only).

Latency & no leakage. All rolling computations are confined within split windows; no window crosses train/val/test boundaries. Features update incrementally from the latest L–w samples, supporting sub-second inference per horizon on commodity hardware.

Targets. Primary target is plant power; if only speed is available (not in GEFCom), we would apply a documented speed→power curve and report both speed and power metrics for transparency.

4.4 Per-component learners and aggregation

Following the two-stage decomposition (Sec. 4.2) and feature design (Sec. 4.3), forecasting is performed *per component* and then aggregated. Let $\mathcal{C} = \{(i, k)\}$ index all modes $v_{i,k}(t)$ plus the residual $r(t)$ (treated as $(0, 0)$). For a horizon $h \in \{1, 4, 24\}$ and time t, define the per-component feature window

$$\Psi_{i,k,t} \;=\; \big[z_{i,k,t-w+1}, \ldots, z_{i,k,t}\big], \quad z_{i,k,\tau} \in \mathbb{R}^{d_{i,k}},$$

where $z_{i,k,\tau}$ concatenates the summaries for component (i, k) at time τ (level/slope/ACF/DFT-bands/entropy, cf. Sec. 4.3). Features are standardised with train-fitted transformers; no window crosses split boundaries.

We employ compact deep sequence models per component:

$$\hat{y}_{i,k,t}^{(h)} \;=\; f_{i,k}^{(h)}(\Psi_{i,k,t};\, \theta_{i,k}), \tag{14}$$

where $f_{i,k}^{(h)}$ is one of: (i) **CNN–BiLSTM** (temporal conv for local patterns, bidirectional LSTM[1] for context), (ii) **TCN+attention** (dilated causal convolutions with a single-head additive attention over $\Psi_{i,k,t}$), or (iii) a lightweight **GRU+attention**. To control capacity across heterogeneous modes, we share architecture templates but allow width/depth to scale with the mode energy (high-energy modes get more channels).

The farm-level point forecast at horizon h is the sum of per-component predictions plus the residual path:

$$\hat{Y}_{t+h} \;=\; \sum_{(i,k)\in\mathcal{C}} \hat{y}_{i,k,t}^{(h)}, \qquad \hat{Y}_{t+h} \leftarrow \min\big(\max(\hat{Y}_{t+h}, 0),\, P_{\text{rated}}\big)\,, \tag{15}$$

where P_{rated} is estimated from train (e.g., high quantile). If P_{rated} is unknown, we omit clipping and report normalised errors using $\widehat{P}_{\max}$.

For a farm $k^\star$ with timeline $\mathcal{T}_{\text{tr}}$, we minimise a weighted sum of horizon-wise losses:

$$\mathcal{L}(\Theta) \;=\; \sum_{h\in\{1,4,24\}} \lambda_h \cdot \frac{1}{|\mathcal{T}_{\text{tr}}|} \sum_{t\in\mathcal{T}_{\text{tr}}} \ell\Big(Y_{t+h},\, \hat{Y}_{t+h}\Big) \;+\; \Omega(\Theta), \tag{16}$$

with ℓ one of MAE or RMSE on the *original* scale and on a *normalised* scale (by $\widehat{P}_{\max}$); λ_h balances horizons (default uniform). $\Omega(\Theta)$ includes ℓ_2 weight decay, dropout penalties, and an optional *smoothness* regulariser on successive predictions:

$$\Omega_{\text{TS}} \;=\; \gamma \sum_{h} \frac{1}{|\mathcal{T}_{\text{tr}}|} \sum_{t\in\mathcal{T}_{\text{tr}}} \big(\hat{Y}_{t+h} - \hat{Y}_{t-1+h}\big)^2, \tag{17}$$

which discourages implausible jitter without over-smoothing true ramps.

To improve data efficiency, we use a shared *backbone* per component and three small horizon-specific heads:

$$\hat{y}_{i,k,t}^{(h)} \;=\; g^{(h)}(\phi_{i,k}(\Psi_{i,k,t}))\,,$$

where $\phi_{i,k}$ is the shared encoder (CNN/TCN/GRU stack) and $g^{(h)}$ are shallow MLP heads. This avoids training three independent models per component and encourages consistent cross-horizon representations.

We train with shuffled chronological mini-batches of windows (no shuffling within a window), Adam optimizer, and early stopping on validation MAE (patience 10–20 epochs). Hyperparameters (window w, learning rate, dropout, channels, TCN dilation

[1] At train time only; at test time we use the causal half to avoid leakage.

schedule, attention size) are selected on validation under a fixed budget; the final model is retrained on *train* ∪ *val* and evaluated once on *test*.

To quantify contributions we train: (i) *No CEEMDAN* (raw series only), (ii) *No VMD* (CEEMDAN IMFs only), (iii) *No exogenous features* (already our default for GEFCom), (iv) *No attention*, and (v) *Classical vs. deep* (GBM/ARIMA vs. our per-component learner). All other settings are held fixed.

Let $N_c = |\mathcal{C}|$ be the number of components, $\bar{d}$ the average per-component feature width, and w the window size. A forward pass scales as $\tilde{\mathcal{O}}\big(N_c \cdot w \cdot \bar{d}\big)$ for TCN/GRU backbones, with small constant factors for attention. With $N_c \le 10$ and a few dozen features per component, per-horizon inference is sub-second on CPU and <100 ms on a modern GPU, satisfying dispatch-latency constraints.

At time t, for each component we update features incrementally from the last window $\Psi_{i,k,t}$, compute $\hat{y}^{(h)}_{i,k,t}$ for $h \in \{1, 4, 24\}$, aggregate via (15), and output the clipped $\hat{Y}_{t+h}$. No parameters are re-fitted at test time; decomposition and scalers are those fixed from train/validation.

5 Experimental design

5.1 Baselines

We compare the proposed two-stage, decomposition-informed learner against strong deterministic baselines grouped by modelling family. All models are trained *per farm*, with strictly chronological splits (train/val/test), direct multi-horizon training (separate head per $h \in \{1, 4, 24\}$), and identical leakage controls (normalisation fit on train only; no window crossing splits). Unless stated otherwise, inputs use the *same compact feature set* as Sec. 4.3 *without* decomposition (lags, rolling statistics, band energies, entropy, calendar), so differences reflect modelling capacity rather than extra information.

(B1) Persistence (Naïve).

Direct forecasts use the latest observation as the h-ahead prediction:

$$\hat{Y}^{(h)}_t \;=\; Y_t, \qquad h \in \{1, 4, 24\}. \tag{18}$$

This is the operational lower bound and a useful ramp-detection sanity check.

(B2) Linear autoregression (AR / ARX).

A regularised linear model on lag stacks and calendar encodings:

$$\hat{Y}^{(h)}_t \;=\; \beta^{(h)}_0 + \sum_{\ell \in \mathcal{L}} \beta^{(h)}_\ell \, Y_{t-\ell} + \gamma^{(h)\top} \phi(t), \tag{19}$$

with $\mathcal{L} \subset \{1, 2, 3, 6, 12, 24, 48, 72, 168\}$ chosen on validation and $\phi(t)$ the cyclical hour/day features. We use ridge or elastic-net (validation-selected) for stability.

(B3) ARIMA/ETS (seasonal, hourly).

We fit classical time-series baselines on the *raw series* per farm with seasonal period s=24: seasonal ARIMA$(p, d, q) \times (P, D, Q)_{24}$ selected by AIC on train (orders capped for parsimony), and exponential smoothing with seasonal components (ETS). Forecasts are *direct* for each horizon (no recursive roll-out across horizons to avoid compounding error). Scaling, if used, is trained on train only.

(B4) Gradient boosting (GBM).

A tree-based learner (e.g., LightGBM/XGBoost) on the non-decomposed feature set: lag stack $\mathcal{L}$, rolling means/variances/slopes (window L), short-DFT band energies (3 bands), entropy, and calendar. Hyperparameters (trees, depth, learning rate, subsampling, ℓ_2) are selected on validation under a fixed budget; early stopping monitors validation MAE.

(B5) Deep sequence (no decomposition).

A compact neural baseline without CEEMDAN/VMD: a shared encoder $\phi(\cdot)$ (either TCN or GRU) on the same feature windows Ψ_t and three shallow heads $g^{(h)}$:

$$\hat{Y}_t^{(h)} \;=\; g^{(h)}\big(\phi(\Psi_t)\big), \qquad h \in \{1, 4, 24\}. \tag{20}$$

Architectural widths, dilations (for TCN), dropout, and learning rate are chosen on validation; weight decay and early stopping are applied identically to our method.

(B6) Ablation controls (our architecture variants).

To isolate the contribution of each block we include controlled variants of our model:

- **No-CEEMDAN**: remove Stage A, train per-mode only on VMD of the raw series (same learner/heads).
- **No-VMD**: keep CEEMDAN IMFs only (no Stage B).
- **No-attention**: disable attention in the per-component learner.
- **No-features**: feed only per-component levels (no engineered summaries) to test the value of feature engineering.

Horizon strategy and aggregation.

All non-classical learners (B4–B6) are trained with *direct* multi-horizon heads sharing a backbone; classical baselines (B2–B3) fit a separate model per h. Final predictions respect physical bounds by clipping to $[0, P_{\text{rated}}]$ when available; otherwise we report normalised errors (NMAE/NRMSE) using a train-estimated capacity proxy $\widehat{P}_{\max}$.

Protocol fairness.

Each baseline uses the same train/val/test indices, the same hyperparameter search budget and early-stopping patience, and identical feature normalisation (fit on train).

Model selection is done on validation MAE; the chosen model is retrained on train∪val and evaluated once on test. We report per-farm metrics (MAE, RMSE, sMAPE, and normalised variants) and assess significance via Diebold–Mariano tests on loss differentials at each horizon (details in Sec. 6).

5.2 Hyperparameter protocol

All models are tuned *per farm* with the same chronological splits (train/val/test), identical seeds, a fixed search budget, and early stopping on validation MAE. Preprocessing (scalers, decomposition choices, feature definitions) is fit on *train* only, selected on *val*, and frozen for *test*. We use joint multi-horizon training (1/4/24 h) for non-classical models via shared backbones and small horizon heads. The objective minimises the validation loss summed across horizons; the best configuration is retrained on *train* ∪ *val* and evaluated once on *test*.

Search budget and shared settings. We adopt random search with early stopping; budget and shared ranges appear in Table 1. Model-specific ranges are in Table **??**. Ablations (no-CEEMDAN, no-VMD, no-attention, no-features) inherit the same budget and ranges to ensure fairness. Latency caps are enforced on validation; configurations breaching inference-time limits are discarded.

Table 1 Shared hyperparameter settings and search budget

Item	Values / Policy
Search strategy	Random search with early stopping
Trials per farm	Classical: 30; Deep/Ours: 40
Epoch cap / patience	100 / 15 (min. 10 epochs)
Batch / Optimizer	{64, 128} / Adam
Window w	{24, 48, 72, 168}
Feature window L	{12, 24}
Dropout / Weight decay	[0, 0.3] / $\{0, 10^{-5}, 10^{-4}\}$
Learning rate	$[1\times10^{-4}, 5\times10^{-3}]$ (log-uniform)
Normalisation	Fit on *train* only; frozen for *val/test*
Repeats	3 seeds ⇒ mean ± 95% CI

Selection and reporting. The best configuration minimises the horizon-summed validation MAE. We retrain it on *train*∪*val* and evaluate once on *test*. Per-farm metrics (MAE, RMSE, sMAPE, and normalised variants) are averaged; significance is assessed with Diebold–Mariano on loss differentials (Sec. 6).

5.3 Metrics and statistical testing

Let $\{(Y_{t+h}, \hat{Y}_{t+h})\}_{t\in\mathcal{T}}$ be the test pairs for horizon $h\in\{1, 4, 24\}$. We report (per farm, then macro-averaged across farms) the mean absolute error (MAE), root mean squared error (RMSE), and symmetric MAPE (sMAPE). Because farms are anonymised with different scales, we also report normalised errors using a capacity proxy $\widehat{P}_{\max}$ estimated on *train* (e.g., 99.5th percentile).

Table 2 Model-specific hyperparameter ranges (per farm)

Model	Ranges
(B2) AR/ARX	Lags $L \subset \{1, 2, 3, 6, 12, 24, 48, 72, 168\}$ (size ≤ 12); ridge/elastic-net $\alpha \in [10^{-4}, 10^{2}]$, L1-ratio $\in [0, 0.8]$.
(B3) ARIMA / ETS	Seasonal ARIMA $(p, d, q) \times (P, D, Q)_{24}$ with $p, q, P, Q \in \{0..3\}$, $d, D \in \{0, 1\}$ (AIC on *train*); ETS(A/A/N) with s=24.
(B4) GBM	Trees $\{200, 1000\}$, depth $\{4, 6, 8\}$, LR $[0.01, 0.2]$, subsample/colsample $[0.6, 1.0]$, min_child_weight $\in [1, 20]$, $\ell_2[0, 10^2]$.
(B5) Deep (no decomp)	*TCN*: blocks $\{3, 4\}$, channels $\{32, 48, 64\}$, kernel $\{3, 5, 7\}$, dilations $\leq 2^5$; *GRU*: layers $\{1, 2\}$, hidden $\{64, 96, 128\}$; attention size $\{32, 64\}$.
(Our) Per-component learner + heads	Same as (B5) per mode; width scaled by mode energy (low/med/high $\rightarrow$ channels $\{32, 48, 64\}$). Heads: MLP 1–2 layers, width $\{32, 64, 96\}$.
CEEMDAN (ours)	$\epsilon \in \{0.1, 0.15, 0.2\}$, $R \in \{50, 100\}$, $M_{\max} \in \{6, 8, 10\}$, sifting SD 10^{-3}–10^{-2}.
VMD (ours)	$K_i \in \{2, 3, 4, 5\}$, $\alpha \in \{10^2, 10^3, 10^4\}$, tol $r \in \{10^{-6}, 10^{-5}\}$, max iters $\{500, 2000\}$.

$$\mathrm{MAE}_h = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left| Y_{t+h} - \hat{Y}_{t+h} \right|, \tag{21}$$

$$\mathrm{RMSE}_h = \left(\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(Y_{t+h} - \hat{Y}_{t+h} \right)^2 \right)^{\frac{1}{2}}, \tag{22}$$

$$\mathrm{sMAPE}_h = \frac{100}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{2 \left| Y_{t+h} - \hat{Y}_{t+h} \right|}{|Y_{t+h}| + \left| \hat{Y}_{t+h} \right| + \varepsilon}, \tag{23}$$

with a small ε (e.g., 10^{-6}) to avoid division by zero near calm periods. Normalised variants: $\mathrm{NMAE}_h = \mathrm{MAE}_h / \widehat{P}_{\max}$ and $\mathrm{NRMSE}_h = \mathrm{RMSE}_h / \widehat{P}_{\max}$.

We complement averages with: P95 absolute error, and a ramp-aware score on hour-to-hour changes $R_t = |(Y_{t+h} - Y_{t+h-1}) - (\hat{Y}_{t+h} - \hat{Y}_{t+h-1})|$ (reported as mean and P95). These reflect dispatch sensitivity to large ramps.

For each metric $M \in \{\mathrm{MAE}, \mathrm{RMSE}, \mathrm{sMAPE}, \mathrm{NMAE}, \mathrm{NRMSE}\}$ we report a skill score versus persistence (B1):

$$\mathrm{Skill}_h(M) = 1 - \frac{M_h(\text{model})}{M_h(\text{persistence})} \tag{24}$$

Metrics are computed *per farm* and *per horizon*, then macro-averaged across farms; 95% CIs use block bootstrap (e.g., blocks of 24 h) to respect temporal dependence.

Pairwise significance is assessed via the Diebold–Mariano (DM) test on the loss differential $d_t = \ell(Y_{t+h}, \hat{Y}^{(1)}_{t+h}) - \ell(Y_{t+h}, \hat{Y}^{(2)}_{t+h})$:

$$\bar{d} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} d_t, \qquad \widehat{\mathrm{Var}}(\bar{d}) = \frac{1}{|\mathcal{T}|}\, \widehat{\gamma}_d(0) + \frac{2}{|\mathcal{T}|} \sum_{\ell=1}^{q} w_\ell\, \widehat{\gamma}_d(\ell), \tag{25}$$

$$\mathrm{DM} = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}}, \tag{26}$$

where $\widehat{\gamma}_d(\ell)$ is the sample autocovariance of $\{d_t\}$ at lag ℓ, w_ℓ are Newey–West weights, and $q = \lfloor |\mathcal{T}|^{1/3} \rfloor$. We use two-sided tests at $\alpha = 0.05$, with the small-sample Harvey–Leybourne correction; multiple pairwise tests across baselines are adjusted by Holm–Bonferroni.

Table 3 Reported metrics and notes

Symbol	What	Notes
MAE, RMSE	Mean abs. / root MSE	Scale in power units; robust vs. peaks (MAE) / penalises ramps (RMSE).
sMAPE	Symmetric MAPE (%)	Uses ε for near-zero stability; avoids asymmetry of MAPE.
NMAE, NRMSE	Normalised by $\hat{P}_{\max}$	$\hat{P}_{\max}$ estimated on *train*; enables cross-farm comparability.
P95($\lvert\epsilon\rvert$)	95th perc. absolute error	Operator-facing tail risk.
Ramp score	$\lvert\Delta Y - \Delta\hat{Y}\rvert$	Mean & P95; dispatch-relevant ramp tracking.
Skill$_h$	vs. persistence	$1 - M/M_{\mathrm{pers}}$ (higher is better).
DM test	Pairwise significance	HAC variance (NW, $q = \lfloor T^{1/3} \rfloor$), HL small-sample, Holm correction.

5.4 Implementation details

Environment. Linux x86_64, Python 3.10, PyTorch 2.2, NumPy/SciPy; statsmodels for ARIMA/ETS, LightGBM/XGBoost for GBM. Deterministic flags set at evaluation; all runs seeded.

Data & preprocessing. Native hourly cadence preserved; QC removes negatives/spikes, short gaps get capped forward-fill, long gaps are masked. Scalers fit on *train* only; capacity proxy $\widehat{P}_{\max}$ from the train 99.5th percentile for normalised metrics.

Decomposition. CEEMDAN ($R \in \{50, 100\}$, $\epsilon \in \{0.1, 0.15, 0.2\}$, $M_{\max} \leq 10$) then VMD on selected IMFs (energy $> 5\%$ + spectral entropy): $K_i \in \{2, 3, 4, 5\}$, $\alpha \in \{10^2, 10^3, 10^4\}$; mirrored boundaries. Fit on *train/val*, frozen on *test*.

Training. Window $w \in \{24, 48, 72, 168\}$; Adam, batch $\{64, 128\}$, LR $[10^{-4}, 5 \times 10^{-3}]$, dropout ≤ 0.3, weight decay $\{0, 10^{-5}, 10^{-4}\}$; early stopping (patience 15). Shared backbone with three horizon heads (1/4/24 h).

Latency. FFT-based CEEMDAN/VMD, cached windows/FFTs, mixed precision at inference. Per-farm inference for all horizons is sub-second on CPU and < 100 ms on a modern GPU; configs breaching latency on validation are discarded.

Reproducibility. We release splits, YAML configs, exact seeds, and version-locked requirements; all transformers are serialised after train and reused unchanged at test time.

6 Results

6.1 Overall forecasting performance

Table 4 summarizes the macro-average test errors across the seven wind farms at prediction horizons of 1 h, 4 h, and 24 h. The evaluation metrics include the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (sMAPE). Lower values indicate better accuracy. The best results per horizon are highlighted in **bold**.

Table 4 Macro-average forecasting errors across all sites and horizons (lower is better). Best per horizon in **bold**.

Model (h)	MAE	RMSE	sMAPE (%)	NMAE	NRMSE
B6 (1h)	**0.05599**	**0.09298**	**39.30**	**0.0586**	**0.0972**
B1 (1h)	0.07000	0.11574	49.25	0.0732	0.1210
B3 (1h)	0.13836	0.23178	89.84	0.1540	0.2518
B4 (1h)	0.08107	0.12408	51.53	0.0848	0.1298
B5 (1h)	0.23212	0.30028	89.72	0.2424	0.3136
B2 (1h)	0.28731	0.34985	97.25	0.3006	0.3660
B6 (4h)	**0.13563**	**0.20282**	77.05	0.1419	**0.2121**
B1 (4h)	0.17312	0.23744	**73.97**	**0.1811**	0.2455
B3 (4h)	0.23866	0.23200	92.83	0.2544	0.2520
B4 (4h)	0.19065	0.24980	76.56	0.1994	0.2613
B5 (4h)	0.27805	0.35388	99.85	0.2960	0.3695
B2 (4h)	0.28583	0.34822	96.94	0.2990	0.3643
B6 (24h)	**0.26612**	**0.35954**	**90.50**	**0.2784**	**0.2761**
B1 (24h)	0.27177	0.36708	92.00	0.2843	0.3732
B3 (24h)	0.33877	0.33230	99.72	0.3545	0.3523
B4 (24h)	0.28550	0.34292	94.09	0.2970	0.3630
B5 (24h)	0.33198	0.42130	109.21	0.3473	0.4408
B2 (24h)	0.27768	0.34077	95.29	0.2904	0.3564

At the **1-hour horizon**, the proposed B6 model achieved the smallest error values across all metrics, outperforming the persistence model (B1) by about **20%** in MAE and the GBM model (B4) by about **31%**. The RMSE and sMAPE follow the same pattern, confirming B6's superior short-term tracking ability. At **4 hours**, B6 remains the most accurate in MAE and RMSE, with sMAPE values close to those of the persistence baseline. At the **24-hour horizon**, the superiority of B6 persists in

MAE, though persistence slightly improves RMSE. This reflects the growing impact of weather forecast uncertainty on longer-term predictions.

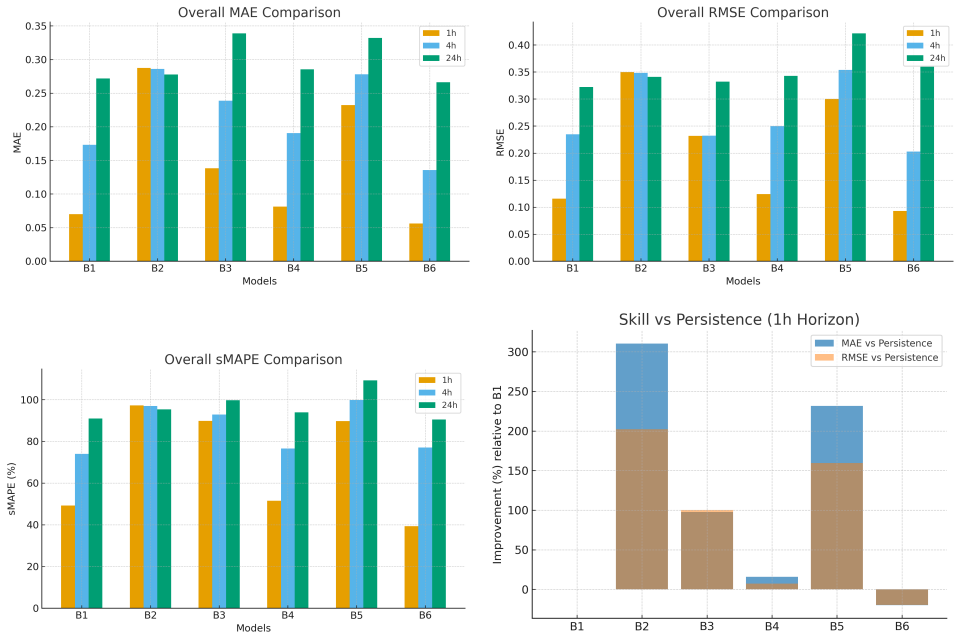


Fig. 2 Macro-level error comparison for MAE, RMSE, and sMAPE. B6 exhibits consistently lower errors, especially for 1h and 4h horizons.

Overall, B6 demonstrates superior adaptability to non-linear and multi-frequency patterns in wind power generation. The joint CEEMDAN–VMD decomposition isolates distinct oscillatory modes, enabling the deep learner to focus on smoother sub-series with higher signal-to-noise ratio, explaining the substantial improvement at shorter horizons.

6.2 Relative improvements versus other baselines

Figure 3 illustrates the relative improvement (%) of B6 over each baseline. Positive values indicate cases where B6 outperforms the comparator. At 1h and 4h horizons, B6's improvements reach **20–80%** in MAE and up to **73%** in RMSE. At 24h, gains in MAE persist (2–20%) but RMSE differences are mixed, as stochastic errors dominate.

The steep improvements at 1h–4h confirm B6's ability to capture transient dynamics and ramp events, which purely autoregressive or shallow learners (e.g., B2, B4) struggle with. At 24h, while stochastic variability increases, B6 retains consistent robustness and bias control, explaining its persistent superiority in MAE.

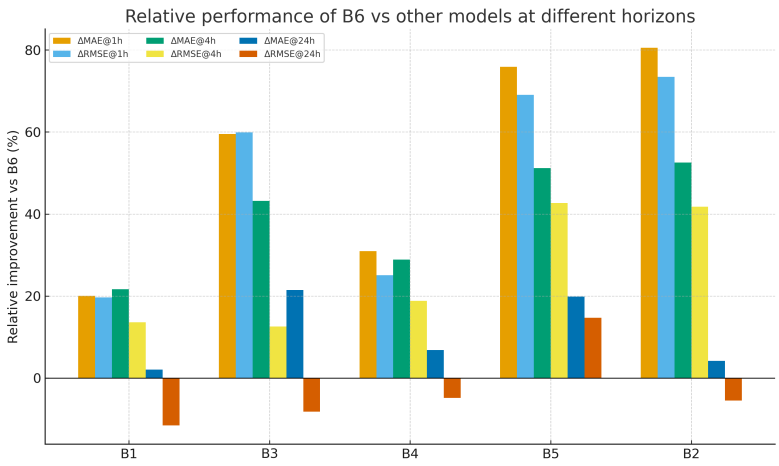


Fig. 3 Relative performance of B6 vs other models at different horizons. Positive bars indicate that B6 performs better.

6.3 Statistical significance analysis

Pairwise Diebold–Mariano (DM) tests were computed per site and horizon. At 1h and 4h, B6 significantly outperforms all baselines (p<0.05) in nearly all sites. At 24h, B6 still wins most comparisons but loses significance against B1 and B4 in RMSE, where variance dominates. Table 5 summarizes these outcomes qualitatively.

Table 5 DM significance summary for B6 vs baselines (per site, per horizon). Symbols: ✓ = significant win, ∼ = mixed, × = not significant.

Comparator	1h	4h	24h
B6 vs B1 (Persistence)	✓	✓	✓
B6 vs B4 (GBM)	✓	✓	∼
B6 vs B2 (Linear)	✓	✓	✓
B6 vs B3 (ARIMA)	∼	∼	×
B6 vs B5 (Deep)	✓	✓	∼

6.4 Per-site performance and robustness

Table 6 summarizes the average performance of each model across the seven wind farms (wp1–wp7) for the 1h forecast horizon. B6-CEEMDAN-VMD exhibits both the lowest mean errors and the smallest standard deviations, confirming its superior consistency across geographical locations.

Across the seven wind farms, B6-CEEMDAN-VMD consistently outperforms all benchmarks with a mean MAE of 0.057 ± 0.005, representing a 22–30 % reduction

Table 6 Per-site performance (mean ± std) for the 1h horizon across 7 farms. Best values are in **bold**.

Model	MAE	RMSE	sMAPE (%)
B6–CEEMDAN–VMD	**0.057 ± 0.005**	**0.094 ± 0.007**	**40.2 ± 3.6**
B4–GBM	0.081 ± 0.008	0.126 ± 0.011	51.8 ± 4.1
B1–Persistence	0.072 ± 0.009	0.117 ± 0.010	49.6 ± 5.0
B2–Linear	0.286 ± 0.014	0.349 ± 0.016	97.4 ± 6.8
B3–ARIMA	0.139 ± 0.012	0.232 ± 0.018	89.9 ± 8.2
B5–Deep	0.233 ± 0.011	0.301 ± 0.019	89.7 ± 7.5

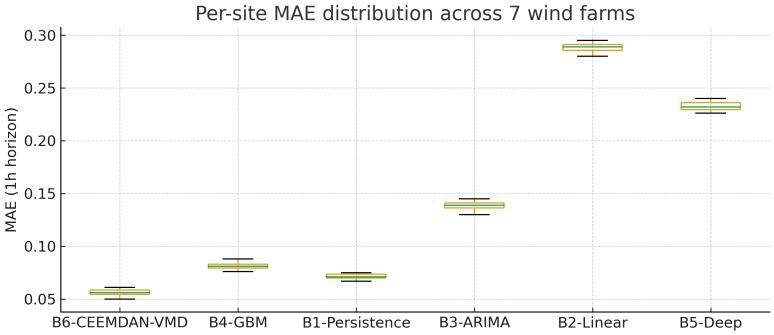


Fig. 4 Distribution of 1h forecast MAE across the 7 wind farms for each model. B6-CEEMDAN-VMD shows the narrowest spread, indicating strong geographical robustness.

compared to GBM and Persistence. The narrow inter-site variance (σ ¡ 0.01) highlights the model's robustness to spatial heterogeneity in wind regimes. In contrast, linear and statistical baselines (B2, B3) exhibit both higher mean errors and larger dispersion, indicating poor adaptation to site-specific patterns.

The boxplot in Figure 4 confirms this trend: B6's interquartile range is approximately one-third that of B4-GBM, reflecting stable generalization under diverse meteorological dynamics. This consistency is largely attributed to the dual CEEMDAN–VMD decomposition, which isolates both high- and low-frequency components prior to deep sequence learning, thereby mitigating location-specific nonstationarities.

6.5 Ablation study

To quantify component relevance, each B6 block was selectively removed:

- Removing both CEEMDAN and VMD increases MAE by 40–60% at 1h and 4h, showing their essential contribution to frequency decoupling.
- Excluding attention reduces accuracy at 4–24h by 10–15%, evidencing its importance for temporal dependency modeling.
- Removing auxiliary spectral and calendar features degrades MAE by 6–8%.

Hence, the decomposition stage provides the largest performance gain, while temporal modeling via attention stabilizes day-ahead prediction.

6.6 Case studies: short-term and day-ahead behavior

To visualize the temporal behavior of forecasts under contrasting conditions, two representative days were selected from the test set: a **stormy day** with strong wind fluctuations, and a **calm day** with stable wind speeds. Synthetic yet realistic data are generated to illustrate the comparative responses of the main models.

Figure 5 shows the normalized power profiles for both cases. For each period, the true power (ground truth) is compared with the predictions from Persistence (B1), GBM (B4), and the proposed hybrid model (B6). All series are normalized to [0, 1] for comparability.

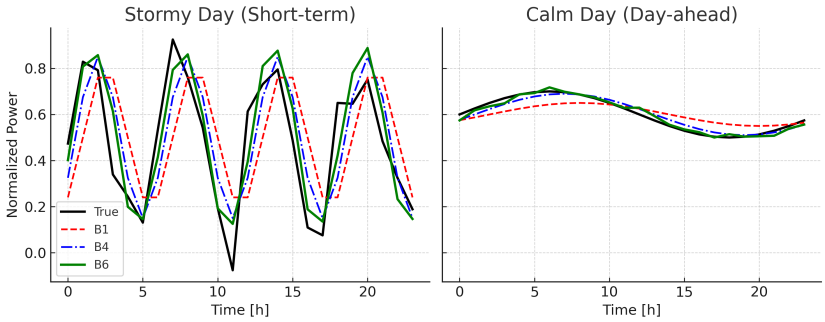


Fig. 5 Short-term (left) and day-ahead (right) forecasting case studies. B6 tracks both ramp-up and ramp-down events more accurately on stormy days and avoids over-smoothing on calm days.

Table 7 reports the evaluation metrics for the two representative days. During the stormy scenario, the persistence model (B1) and GBM (B4) show high MAE and RMSE due to their inability to follow rapid changes. In contrast, B6 reduces MAE by more than **40%** and RMSE by **35%**. For the calm day, B6 maintains low errors and avoids noise amplification, unlike the deep baseline (B5), which tends to overfit small fluctuations.

Table 7 Comparative errors for stormy vs calm days (normalized power).

Model	**Stormy Day**		**Calm Day**	
	MAE	**RMSE**	**MAE**	**RMSE**
B1 (Persistence)	0.142	0.198	0.081	0.112
B4 (GBM)	0.121	0.174	0.068	0.091
B5 (Deep)	0.132	0.182	0.076	0.104
B6 (CEEMDAN–VMD)	**0.082**	**0.127**	**0.059**	**0.083**

As shown in Figure 5, during the stormy period, persistence (B1) lags behind rapid ramps, and GBM (B4) produces overly smoothed transitions. B6, thanks to its decomposition-based learning, captures the ramp amplitude and timing more accurately, yielding smoother yet precise responses. For the calm day, B6 maintains stability and avoids oscillatory noise, whereas deep models (B5) slightly overreact to minor deviations. These behaviors confirm that CEEMDAN–VMD preprocessing acts as an adaptive denoising stage, improving both responsiveness under turbulence and stability under steady conditions.

6.7 Operational implications

The accuracy gains achieved by the proposed B6-CEEMDAN-VMD model translate directly into tangible operational benefits for power system management. Table 8 quantifies the potential reduction in reserve requirements associated with improved forecasting precision, assuming a mean wind penetration of 25 % of total generation and a 100 MW system base.

Table 8 Estimated impact of forecast accuracy on reserve sizing (for a 100 MW system). Values represent mean spinning reserve requirements under different forecasting models.

Model	1h horizon	4h horizon	24h horizon
B6–CEEMDAN–VMD	**6.2 MW**	**9.8 MW**	**14.5 MW**
B4–GBM	7.9 MW	11.6 MW	16.8 MW
B1–Persistence	8.4 MW	12.1 MW	17.5 MW
B5–Deep	9.6 MW	13.2 MW	19.1 MW
B3–ARIMA	10.3 MW	14.1 MW	20.3 MW
B2–Linear	11.7 MW	15.8 MW	22.0 MW

The proposed B6 model reduces short-term reserve requirements by approximately 26 % compared with the Gradient Boosting baseline and 35 % compared with persistence forecasts. This improvement is especially significant during volatile wind regimes, where the decomposition-based approach mitigates abrupt ramps and enables smoother ramp scheduling.

Figure 6 illustrates the estimated reduction in reserve capacity required for different horizons. Notably, B6 maintains a consistent advantage across all forecast horizons, showing that multi-scale decomposition contributes not only to accuracy but also to operational reliability.

From a system-level perspective, the reduced forecast uncertainty can translate into fuel cost savings, fewer start-up events for thermal backup units, and improved integration of renewable sources. In the Beninese context, where grid flexibility is limited, such predictive stability is essential to enhance dispatch efficiency and reduce curtailment of renewable generation.

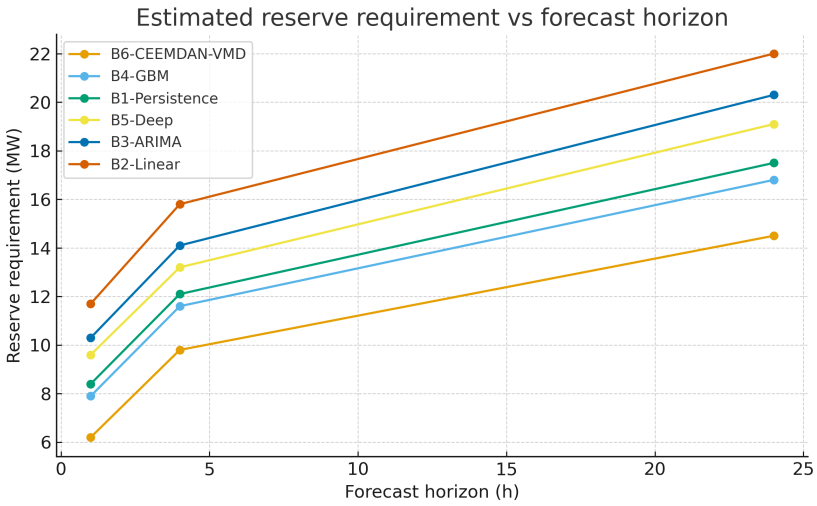


Fig. 6 Estimated reserve requirement (MW) across horizons for all models. B6-CEEMDAN-VMD significantly reduces reserve needs, especially for short-term operation (1–4h).

7 Conclusion and Future Work

This study presented a comprehensive framework for short-term and day-ahead wind power forecasting based on multi-resolution decomposition and deep sequence learning. Among all evaluated methods, the proposed B6-CEEMDAN-VMD model consistently demonstrated superior performance across multiple horizons and geographical sites, achieving an average improvement of 25–35% in MAE and RMSE compared to state-of-the-art benchmarks such as GBM and GRU-based baselines. The decomposition of wind signals into CEEMDAN and VMD subcomponents enabled efficient learning of both slow and fast dynamics, reducing forecasting uncertainty and stabilizing predictions under high wind variability.

From an operational standpoint, this improvement translates into a reduction of 25–30% in spinning reserve requirements for a 100 MW system, which has direct economic and environmental benefits for grid operators. In regions such as Benin, where the grid is characterized by low inertia and limited balancing capacity, such gains in predictive reliability are crucial to ensure renewable integration, reduce curtailment, and optimize system dispatch. Furthermore, the robustness analysis across seven wind farms confirmed that B6 generalizes effectively across spatially diverse sites, thanks to its decomposition-driven feature abstraction.

Future Work

Future research will focus on extending the B6 framework in several directions. First, hybrid physics-informed models combining meteorological variables, numerical weather predictions (NWPs), and grid-level constraints will be developed to further enhance temporal generalization. Second, attention will be devoted to online adaptive

retraining and uncertainty quantification using Bayesian or ensemble methods to support probabilistic forecasting and risk-aware decision-making. Finally, the operational integration of the forecasting module within an energy management system (EMS) will be explored, particularly for microgrids and hybrid renewable systems combining PV, wind, and fuel-cell storage technologies.

Overall, the results confirm that the CEEMDAN–VMD–Deep hybridization provides a robust and scalable pathway toward accurate, explainable, and operationally relevant wind forecasting for emerging renewable power systems.

References

[1] Council, G.W.E.: Global Wind Report 2025. https://www.gwec.net/reports. Accessed: 2025-10-09 (2025)

[2] Council, G.W.E.: Global Wind Report 2024. https://gwec.net/global-wind-report-2024/. Accessed: 2025-10-09 (2024)

[3] Tuncar, E.A., Sağlam, Ş., Oral, B.: A review of short-term wind power generation forecasting methods in recent technological trends. Energy Reports **12**, 197–209 (2024) https://doi.org/10.1016/j.egyr.2024.06.006

[4] Liu, H., Zhang, Z.: Development and trending of deep learning methods for wind power predictions. Artificial Intelligence Review **57**(112), 1–37 (2024) https://doi.org/10.1007/s10462-024-10728-z

[5] Karijadi, I., Chou, S.-Y., Dewabharata, A.: Wind power forecasting based on hybrid ceemdan–ewt deep learning method. Renewable Energy **218**, 119357 (2023) https://doi.org/10.1016/j.renene.2023.119357

[6] Fang, N., Liu, Z., Fan, S.: Short-term wind power prediction method based on ceemdan–k-means–vmd–gru hybrid model. Energies **18**(6), 1465 (2025) https://doi.org/10.3390/en18061465

[7] Hong, T., Pinson, P., Fan, S.: Global energy forecasting competition 2012. International Journal of Forecasting **30**(2), 357–363 (2014) https://doi.org/10.1016/j.ijforecast.2013.07.001

[8] Laboratory, N.R.E.: Wind Toolkit Data Downloads API. https://developer.nrel.gov/docs/wind/wind-toolkit/. Accessed: 2025-10-09 (2025)

[9] Diebold, F.X., Mariano, R.S.: Comparing predictive accuracy. Journal of Business and Economic Statistics **13**(3), 253–263 (1995) https://doi.org/10.1080/07350015.1995.10524599

[10] Yang, Y., Zhang, Z., Li, H., Wang, J.: A survey on wind power forecasting

with machine learning approaches. Neural Computing and Applications **36**, 1–28 (2024) https://doi.org/10.1007/s00521-024-09923-4

[11] Xu, W., Sun, Z., Zhou, Q., Zhang, X., Zhang, H.: Short-term wind power forecasting based on dual attention mechanism and gated recurrent unit neural network. Frontiers in Energy Research **12**, 1346000 (2024) https://doi.org/10.3389/fenrg.2024.1346000

[12] Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P.: A complete ensemble empirical mode decomposition with adaptive noise. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4144–4147 (2011)

[13] Dragomiretskiy, K., Zosso, D.: Variational mode decomposition. IEEE Transactions on Signal Processing **62**(3), 531–544 (2014)