energies

MDPI

Article

Beyond Point Accuracy: Multi-Step Wind Power Forecasting and Two-Hour Ramp-Risk Early Warning: A Data-Driven Case Study

Haoxuan Dong and Weiqing Wang *

School of Electrical Engineering, Xinjiang University, Urumqi 830017, China; donghaoxuan1@stu.xju.edu.cn
* Correspondence: wwq59@xju.edu.cn

Abstract

Wind-power forecasting accuracy and ramp-risk warning may favor different models. We evaluate eight-step (two-hour) trajectories from 24-h histories using 35,040 observations at 15-min resolution from a publicly shared 2019 wind-farm dataset associated with Xinjiang, China. Fourteen input channels, a January–August/September–October/November–December chronological train/validation/test split, train-fitted scaling, and five-seed model comparisons are used. At the 20 MW absolute ramp criterion, DLinear provides the highest mean R^2 (0.8645) and lowest RMSE (25.10 MW), whereas iTransformer provides the highest window-level Average Precision (0.338, compared with event prevalence 0.158) and the lowest MAE among learnable models (14.77 MW). Persistence attains the lowest overall MAE (13.44 MW). Additional analysis shows that iTransformer has lower MAE in 4847 ordinary windows (9.99 versus 11.62 MW for DLinear), but DLinear has lower MAE in 906 ramp-positive windows (38.30 versus 40.37 MW). Thresholded auxiliary heads alter F1 without consistently improving ranking. A retrospective sensitivity analysis illustrates the impact of assumed false-alarm and missed-ramp costs. The single-site setting, two residual sentinel-coded meteorological records, and incomplete upstream cleaning provenance limit reproducibility and transferability claims. Regression accuracy, ramp-window ranking, and alarm decisions should therefore be assessed separately.

Keywords: wind power forecasting; multi-step forecasting; ramp event; early warning; iTransformer; DLinear; operational data; class imbalance; strict temporal evaluation

1. Introduction

The increasing penetration of wind generation places greater demands on short-horizon forecasting for reserve scheduling, energy management, and secure grid operation [1–5]. Wind power is driven by nonlinear and nonstationary atmospheric processes while also exhibiting strong short-term autocorrelation. As a result, very-short-term forecasting can be deceptively easy under aggregate regression metrics: a persistence forecast may remain highly competitive even when a complex model captures additional meteorological structure.

Deep-learning and Transformer-based forecasting methods have increasingly been applied to multivariate wind-power time series [6–12]. Informer reduces the computational burden of long-sequence attention through ProbSparse attention and attention distilling [13]. iTransformer reorganizes tokenization around variates rather than timestamps and has shown strong multivariate forecasting performance [14]. At the same time, DLinear demonstrated that simple linear architectures can remain highly competitive with Transformer models in long-horizon forecasting [15]. These developments make two methodological

Check for updates

Academic Editor: Lars Johanning

Received: 23 September 2026
Revised: 11 November 2026
Accepted: 17 November 2026
Published: 24 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

requirements particularly important for wind-power forecasting studies: strong baselines and leakage-free temporal evaluation.

A second operational challenge is Wind Power Ramp Events (WPREs), i.e., unusually large power variations over a short time interval. Ramp events can increase balancing requirements and complicate reserve and storage scheduling [16,17]. Their definition typically depends on ramp magnitude, time interval, direction, and timing [18]. Existing studies have considered both indirect ramp forecasting based on predicted power trajectories and direct ramp-oriented classifiers [19–23]. However, rare-event learning is strongly imbalanced, and a model optimized for average regression error may smooth precisely the transitions that matter most for operational warning.

Recent work has also emphasized multi-horizon and multi-step wind-power forecasting architectures [24–26]. This study addresses these issues by separating two questions: (i) which models provide reliable two-hour multi-step power trajectories under a chronologically partitioned protocol, and (ii) how much useful information those trajectories contain for advance ramp-risk warning. The analysis uses a full year of publicly shared wind-power and multisource meteorological data, includes persistence and DLinear as strong baselines, evaluates stochastic models over five random seeds, and distinguishes risk ranking from threshold-specific alarm performance.

The main contributions are as follows:

1. A strict chronological evaluation protocol is established for two-hour multi-step wind-power forecasting using a 24-h historical window, training-only normalization, a dedicated validation period, a chronologically separated November–December evaluation period, and five-seed reporting for stochastic models.
2. A two-hour window-level ramp-risk task is formulated while preserving the case-study event definition of a $\geq$20 MW change over one 15-min interval. This converts extremely sparse exact-timing detection into an operational early-warning problem without relaxing the event threshold.
3. A unified comparison of Persistence, DLinear, Informer, iTransformer, SE ablations, and auxiliary ramp heads is conducted using both regression metrics and imbalanced-event metrics. The results reveal that regression accuracy and ramp-risk discrimination are not equivalent objectives.
4. Stability and ablation analyses show that SE recalibration does not provide a reproducible gain in the present setting, while auxiliary risk heads primarily alter the Precision–Recall operating point rather than universally improving ranking and regression quality.

The remainder of the paper is organized as follows. Section 2 describes the data, strict temporal split, multi-step forecasting task, ramp-risk formulation, candidate models, and evaluation protocol. Section 3 reports forecasting, warning, stability, and ablation results. Section 4 discusses persistence dominance, the decoupling between regression and risk skill, and limitations. Section 5 concludes the study.

2. Materials and Methods

2.1. Publicly Shared Wind-Power Dataset

The dataset contains 35,040 chronological observations from the full year 2019 at 15-min resolution. The data were obtained from a publicly shared CSDN resource (resource ID 88485424), whose public resource page describes a Xinjiang wind-farm dataset at 15-min intervals and lists the same multi-height wind, meteorological, and active-power variables used in this study [27]. The downloaded archive contains a year-long workbook whose Chinese filename translates to “Xinjiang Wind Farm No. 1, 2019”; the site-name field in the workbook identifies “Guohua Jingxia North Wind Farm No. 1”. The accompanying

public description therefore associates the data with the Jingxia North wind farm in Hami, Xinjiang, China. Independent public information from the wind-farm operator reports an installed capacity of approximately 200 MW and 133 wind turbines rated at 1.5 MW each for this site [28]. Because the authors did not participate in the original data acquisition and the file was obtained through a third-party data-sharing platform, the linkage between the publicly redistributed workbook and the original SCADA acquisition chain could not be independently verified. Accordingly, the site identity and capacity are treated as source metadata supported by public records rather than as independently audited measurement provenance. The active-power series in the downloaded file ranges from 0.028 to 200.05 MW, with a mean value of approximately 70.3 MW. Power is reported in MW throughout.

The model input has 14 channels: 13 meteorological variables and historical farm-level active power. The meteorological variables comprise wind speed at 10, 30, 50, and 70 m and at the reported hub-height channel; wind direction at the same five levels; air temperature; atmospheric pressure; and relative humidity. Temperature, pressure, humidity, and the exact hub-height value are retained as described in the supplied data fields; sensor-installation details cannot be independently verified from the original acquisition record. Five calendar markers (month, day, weekday, hour, and quarter-hour) are provided separately as temporal covariates where supported by the forecasting architecture and are not counted among the 14 input channels. The 14 input variables and their measurement metadata are summarized in Table 1.

Table 1. Input variables used for forecasting. The manuscript standardizes active power to MW and temperature to °C. Sensor heights for temperature, pressure, and humidity are not available in the supplied data dictionary.

No.	Variable	Measurement Height	Unit	Source
1	Wind speed	10 m	m/s	Meteorological mast
2	Wind speed	30 m	m/s	Meteorological mast
3	Wind speed	50 m	m/s	Meteorological mast
4	Wind speed	70 m	m/s	Meteorological mast
5	Wind speed	Hub height	m/s	Hub-height channel
6	Wind direction	10 m	degree	Meteorological mast
7	Wind direction	30 m	degree	Meteorological mast
8	Wind direction	50 m	degree	Meteorological mast
9	Wind direction	70 m	degree	Meteorological mast
10	Wind direction	Hub height	degree	Hub-height channel
11	Air temperature	Not specified	°C	Meteorological measurement
12	Atmospheric pressure	Not specified	hPa	Meteorological measurement
13	Relative humidity	Not specified	%	Meteorological measurement
14	Farm active power (historical target)	Farm level	MW	Operational power record

2.2. Data Cleaning, Normalization, and Chronological Split

A direct audit of the publicly shared 2019 workbook identified 35,040 time-stamped rows, no structurally empty cells, and no spreadsheet error cells. However, the numeric sentinel −99 remains present in the downstream processed data: all 13 meteorological variables contain −99 at 19 September 2019, 22:30 and 25 December 2019, 05:00 (26 cells in two records), whereas the active-power target does not contain this marker. The final downstream loader did not explicitly convert these sentinel values to missing values or impute them, so we do not claim that these anomalies were corrected in the reported runs. No separately documented statistical outlier-removal cutoff or count is available. The final eight-step experimental code and saved normalization parameters support fitting min–max scaling on the training partition and applying those parameters to validation and

test. However, because the complete upstream preprocessing history is unavailable, full end-to-end preprocessing causality cannot be independently certified.

The year is divided chronologically into three non-overlapping partitions: January–August for training, September–October for validation, and November–December for final testing. With a 96-step historical window and an eight-step target horizon constructed independently within each partition, the resulting sample counts are

$$N_{\text{train}} = 23{,}225, \qquad N_{\text{val}} = 5753, \qquad N_{\text{test}} = 5753. \tag{1}$$

The final eight-step run metadata indicate validation-regression-MSE checkpoint selection and train-only normalization, with minibatch shuffling confined to training; they should not be confused with an older one-step script that fitted scalers globally and selected checkpoints using test loss. Model configurations and the primary 20 MW event criterion are documented, but an exhaustive pre-specified hyperparameter search and absence of all prior test inspection cannot be reconstructed from archived records. Supplementary analyses are explicitly retrospective. The validation sentinel can enter 96 input windows and may influence checkpoint selection; the test sentinel enters 96 test input windows. A post hoc exclusion sensitivity check cannot replace imputation and retraining. More rigorous causal cleaning and an independently untouched evaluation period remain necessary for deployment-oriented validation.

2.3. Two-Hour Multi-Step Forecasting Task

The historical input window is set to $W = 96$ samples, corresponding to the previous 24 h. Each model simultaneously predicts the next $H = 8$ samples, corresponding to a two-hour horizon at 15-min resolution. For a forecast origin at time t, the target and prediction vectors are

$$\mathbf{P}_{t+1:t+8} = [P_{t+1}, P_{t+2}, \ldots, P_{t+8}], \tag{2}$$

$$\widehat{\mathbf{P}}_{t+1:t+8} = [\widehat{P}_{t+1}, \widehat{P}_{t+2}, \ldots, \widehat{P}_{t+8}]. \tag{3}$$

The default regression loss is the mean squared error over all samples and forecast steps,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{NH} \sum_{i=1}^{N} \sum_{k=1}^{H} \left(P_{i,t+k} - \widehat{P}_{i,t+k} \right)^2. \tag{4}$$

2.4. Forecasting Models

Persistence. The deterministic persistence forecast assumes that future power remains equal to the latest observed power,

$$\widehat{P}_{t+k} = P_t, \qquad k = 1, \ldots, 8. \tag{5}$$

It is retained as a mandatory baseline because very-short-term wind power is strongly autocorrelated.

DLinear. DLinear decomposes the historical power sequence into trend and residual components by a moving-average kernel of 25 and applies shared linear mappings to the decomposed components [15]. The exported implementation is the channel-independent configuration (`individual=False`) with 1552 trainable parameters. For the selected farm-power output, the effective predictive input is historical power; auxiliary meteorological channels and calendar markers do not affect the selected DLinear output. It is therefore retained as a deliberately strong low-complexity persistence-like reference rather than a multisource fusion model.

Informer. The Informer baseline is a full encoder–decoder implementation with ProbSparse self-attention, full cross-attention, and sequence distillation [13]. The frozen configuration uses $d_{\text{model}} = 256$, eight attention heads, two encoder layers, one decoder layer, $d_{\text{ff}} = 512$, and dropout 0.2, with 2,063,363 trainable parameters. Historical calendar markers and known future calendar fields are embedded as temporal features. The decoder uses the final 48 historical steps as context followed by eight zero-valued prediction tokens.

iTransformer. iTransformer reorganizes multivariate forecasting by treating variables as tokens and applying full attention across variable representations [14]. The frozen configuration uses $d_{\text{model}} = 256$, eight heads, two encoder layers, $d_{\text{ff}} = 512$, and dropout 0.2, with 1,081,608 trainable parameters. Per-window, per-variable mean–standard-deviation normalization ($\epsilon = 10^{-5}$) is applied inside the model and reversed at the output. The five historical calendar covariates are supplied as additional tokens, whereas future calendar values are not used. Only the farm-power channel is projected to the reported eight-step output.

SE ablations. To test whether feature recalibration contributes reproducible value, an SE block [29] is inserted after the feature-embedding layer in the Informer branch. The embedded-SE ablation uses $d_{\text{model}} = 256$ and reduction ratio $r = 8$. Because the five-seed results do not show a stable gain, the SE configuration is treated as an ablation rather than as a universally superior architecture.

Auxiliary ramp-risk heads. An optional auxiliary ramp-classification head is trained jointly with the regression task. The head emits horizon-wise ramp logits; for window-level evaluation, the risk score is the maximum sigmoid probability across the eight forecast steps. For the weighted-BCE variant, a fixed task coefficient was used rather than adaptive uncertainty weighting, GradNorm, or another dynamic gradient-balancing procedure:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{ramp}}, \qquad \lambda = 0.05, \tag{6}$$

where $\lambda = 0.05$ was held constant throughout training and the positive-class weight was estimated on training data only. The archived Informer-SENet weighted-BCE experiment uses a positive-class weight of 17.24. An archived single-seed Informer–SENet weight scan over $\lambda \in \{0.05, 0.1, 0.2, 0.5\}$ favored 0.05 on validation MSE in that particular setting (Supplementary Table S7), but no multi-seed or cross-architecture weight search is documented. Thus, the reported multi-task trade-off applies to the tested fixed weighting; it does not establish that dynamic weighting would fail. For the final exported experiments, head-based alarms use a fixed probability threshold of 0.5, while trajectory-based alarms use the fixed 20 MW absolute threshold. Average Precision is threshold-independent and is treated as the primary ranking metric.

2.5. Hyperparameter Selection and Seed Protocol

The final configurations in Table 2 were frozen for reporting, and validation-period performance was used for early stopping (maximum 50 epochs, patience eight). The experiments use a common optimizer budget (batch size 128, initial learning rate 5×10^{-4}, weight decay 5×10^{-4}) and architecture-specific settings documented in the table. The available final configuration exports do not establish a systematic or exhaustive grid/Bayesian hyperparameter search across all architectures, and no such search is claimed; architectural comparisons should therefore be interpreted conditional on these reported configurations. The five distinct, predeclared integer seeds (42, 100, 2024, 2025, 3407) provide a small repeated-initialization and stochastic-training sensitivity sample, not five independent wind farms or a statistically powered random-effects experiment. Their numeric identities have no physical interpretation. The primary 20 MW criterion was fixed for the reported event definition. Because the archived record does not establish that all model choices

preceded every inspection of the November–December outcomes, the comparison and later threshold analyses must be interpreted as a retrospective case study rather than as proof of an entirely untouched prospective test.

Table 2. Frozen experimental settings used in the final matrix. Model-specific settings are reported only where supported by the supplied final configuration exports.

Item	Setting	Notes
Sampling interval	15 min	Fixed
Historical window	96 steps	24 h
Forecast horizon	8 steps	2 h
Chronological split	Jan–Aug/Sep–Oct/Nov–Dec	Train/validation/test
Window counts	23,225/5753/5753	Train/validation/test
Batch size	128	Learnable models
Learning rate	5×10^{-4}	Initial learning rate
Weight decay	5×10^{-4}	Fixed
Warm-up	5 epochs	Fixed
Maximum epochs/patience	50/8	Early stopping
Loss	MSE	Standard forecasting runs
Random seeds	42, 100, 2024, 2025, 3407	Stochastic models
iTransformer	$d_{\text{model}} = 256$, 8 heads, 2 encoder layers, $d_{\text{ff}} = 512$, dropout 0.2	1,081,608 parameters; per-window variable normalization
Informer	$d_{\text{model}} = 256$, 8 heads, 2 encoder/1 decoder layers, $d_{\text{ff}} = 512$, dropout 0.2	2,063,363 parameters; ProbSparse + distilling
Informer-family SE ablation	embedded SE, $d_{\text{model}} = 256$, $r = 8$	Skip connection enabled
DLinear	moving-average kernel 25, individual=False	1552 parameters; historical power effective input
Auxiliary ramp coefficient	$\lambda = 0.05$	Weighted-BCE head experiment

2.6. Ramp Definition and Two-Hour Window-Level Warning

A 15-min ramp event is defined when the absolute power change between adjacent measurements is at least $P_{\text{ramp}} = 20$ MW. The evaluated thresholds of 10, 20, and 30 MW are absolute MW differences, not percentages of rated power. An external public operator record associates the attributed site with approximately 200 MW nominal installed capacity; if that attribution is correct, the three thresholds correspond contextually to approximately 5%, 10%, and 15% of nominal capacity. Because the acquisition provenance of the third-party workbook was not independently authenticated, no capacity-normalized event evaluation is claimed. Sensitivity at 10/20/30 MW assesses dependence on the absolute threshold,

$$r_{t+k} = \mathbb{I}(|P_{t+k} - P_{t+k-1}| \geq 20\ \text{MW}), \qquad k = 1, \ldots, 8, \tag{7}$$

where P_t is the latest observed power for $k = 1$.

Exact localization of a rare event to one future 15-min interval was found to provide very weak discrimination under the strict evaluation protocol. Therefore, the operational warning task is formulated at the two-hour window level without changing the underlying 15-min event definition:

$$R_t^{(2\text{h})} = \mathbb{I}\left[\max_{k=1,\ldots,8} |P_{t+k} - P_{t+k-1}| \geq 20\ \text{MW}\right]. \tag{8}$$

In the final November–December test partition, 906 of 5753 forecast windows are positive, corresponding to a prevalence of 15.75%. This prevalence is also the expected average-precision level for a non-informative ranking.

For models without a dedicated classification head, the power-trajectory risk score is

$$s_t^{\text{power}} = \max_{k=1,\ldots,8} \left| \widehat{P}_{t+k} - \widehat{P}_{t+k-1} \right|, \tag{9}$$

with $\widehat{P}_t = P_t$. For head-based variants, the window risk score is $s_t^{\text{head}} = \max_k \sigma(z_{t+k})$, where z_{t+k} denotes the auxiliary ramp logit at forecast step k.

2.7. Evaluation and Statistical Protocol

Regression performance is evaluated using R^2, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE):

$$R^2 = 1 - \frac{\sum_i (P_i - \widehat{P}_i)^2}{\sum_i (P_i - \bar{P})^2}, \tag{10}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_i (P_i - \widehat{P}_i)^2}, \tag{11}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_i \left| P_i - \widehat{P}_i \right|. \tag{12}$$

Window-level risk ranking is evaluated primarily by Average Precision (AP), which summarizes the Precision–Recall curve and is appropriate for imbalanced outcomes. Precision, recall, and F1 are additionally reported at fixed operating thresholds: 20 MW for the power-trajectory score and 0.5 for the auxiliary-head probability. The primary reported alarm operating points are fixed in advance of the supplemental threshold-sensitivity calculations (20 MW for the trajectory score and 0.5 for the auxiliary-head output); the later retrospective decision-cost threshold comparisons are reported separately:

$$\mathrm{Precision} = \frac{TP}{TP + FP}, \tag{13}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}, \tag{14}$$

$$F_1 = \frac{2\,\mathrm{Precision}\,\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{15}$$

All stochastic models are trained with five random seeds (42, 100, 2024, 2025, and 3407), whereas Persistence is deterministic and is evaluated once. Mean and standard deviation are reported for seed-dependent regression and warning metrics. In addition, an archived exploratory paired bootstrap over the 5753 overlapping test windows provides descriptive resampling intervals and signs of paired differences. Because neighboring forecasting windows share input and target timestamps, these naive window-wise intervals do not account for temporal dependence and must not be interpreted as calibrated independent-sample confidence intervals or formal evidence of statistical significance.

3. Results

3.1. Two-Hour Multi-Step Power Forecasting

Table 3 summarizes the final five-seed matrix. DLinear achieves the highest mean R^2 (0.8645 ± 0.0004) and the lowest RMSE (25.10 ± 0.03 MW), demonstrating very low seed sensitivity. iTransformer achieves $R^2 = 0.8563 \pm 0.0020$ and the lowest MAE among the learnable models (14.77 ± 0.34 MW). The deterministic Persistence baseline attains $R^2 = 0.8550$, RMSE = 25.97 MW, and the lowest overall MAE (13.44 MW). Informer-based variants are less stable: the no-SE Informer reaches $R^2 = 0.8479 \pm 0.0158$, while the embedded-SE variant decreases to 0.8426 ± 0.0235. Thus, the final regression evidence does not support architectural dominance by the SE-enhanced Informer; instead, simple and Transformer-based models occupy different points on the accuracy–risk trade-off. Figure 1 shows the corresponding five-seed stability of the stochastic regression models.

Table 3. Five-seed performance matrix for two-hour multi-step forecasting and window-level ramp-risk warning. Values are mean ± standard deviation across five seeds except deterministic Persistence. AP for models without a head uses the power-trajectory score in Equation (9); head-based variants use predicted window probability. Bold indicates the best value in each metric column; Persistence is included when identifying the lowest overall MAE.

Model	R^2	RMSE (MW)	MAE (MW)	Window AP	Precision	Recall	Window F1
DLinear	**0.8645 ± 0.0004**	**25.10 ± 0.03**	15.82 ± 0.05	0.287 ± 0.015	–	0.000	0.000
iTransformer	0.8563 ± 0.0020	25.84 ± 0.18	14.77 ± 0.34	**0.338 ± 0.019**	0.491 ± 0.037	0.131 ± 0.019	0.206 ± 0.023
iTransformer + ramp head	0.8493 ± 0.0046	26.46 ± 0.40	15.27 ± 0.46	0.252 ± 0.009	0.188 ± 0.006	0.638 ± 0.065	0.290 ± 0.014
Informer (MSE, no SE)	0.8479 ± 0.0158	26.57 ± 1.37	17.61 ± 1.95	0.246 ± 0.032	0.331 ± 0.068	0.183 ± 0.075	0.222 ± 0.059
Informer + embedded SE (MSE)	0.8426 ± 0.0235	27.00 ± 1.96	17.99 ± 1.28	0.247 ± 0.040	0.334 ± 0.046	0.206 ± 0.141	0.230 ± 0.106
Informer + embedded SE + ramp head	0.8366 ± 0.0245	27.50 ± 2.08	18.28 ± 2.37	0.318 ± 0.031	0.313 ± 0.051	0.554 ± 0.177	**0.387 ± 0.036**
Persistence	0.8550	25.97	**13.44**	0.158 (prevalence)	–	–	–

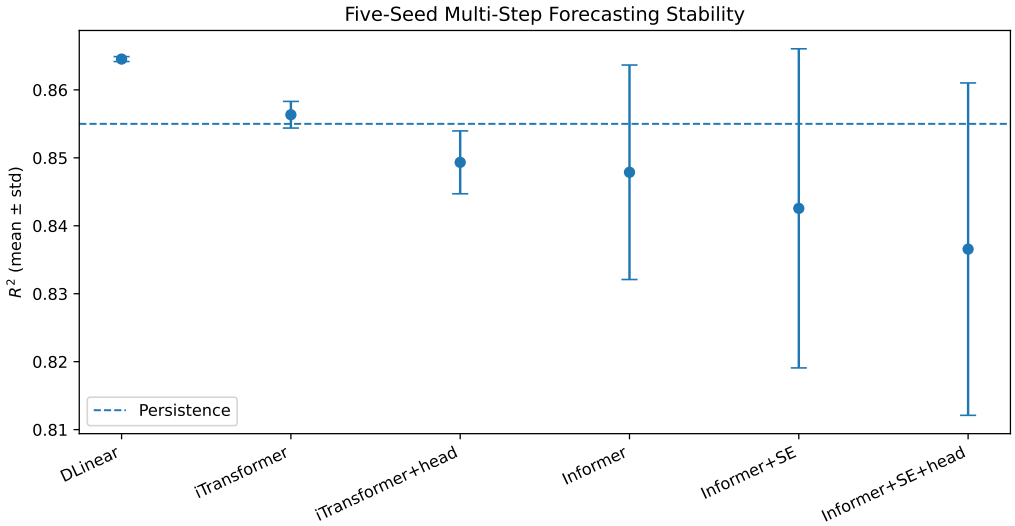


Figure 1. Five-seed stability of the two-hour multi-step regression models. Error bars denote one standard deviation across seeds. Persistence is deterministic and is shown as a horizontal reference.

3.2. Error Differences Between Ordinary and Ramp Windows

To examine the difference between DLinear's stronger squared-error metrics and iTransformer's lower neural-model MAE, the saved predictions were stratified using the *observed* maximum adjacent power change in each eight-step window, with the first difference anchored to the latest observed power. This descriptive stratification includes all five saved seeds for each model; it does not retrain or tune either model. Of 5753 test windows, 4847 have a maximum true adjacent change below 20 MW (ordinary) and 906 contain a change of at least 20 MW (ramp-positive). Table 4 shows that iTransformer has smaller errors in ordinary windows, whereas DLinear has smaller errors in ramp-positive windows. These ramp-positive windows contribute approximately 66.37% and 70.37% of the total squared error for DLinear and iTransformer, respectively. Consequently, a smaller overall neural-model MAE does not imply improved prediction of extreme changes. DLinear's effective target prediction uses historical power, whereas the iTransformer branch uses additional covariates; the findings are not an architecture-only causal ablation. The finer error stratification across maximum-change bins is shown in Figure 2.

Table 4. Ordinary and ramp-positive two-hour window errors from saved test predictions (mean ± sample standard deviation over five seeds). The true-event threshold is an absolute 20 MW adjacent change; all eight forecast-step errors within each selected window are included. Bold indicates the lower error for each observed-window class.

Observed Window Class	Model	Windows	MAE (MW)	RMSE (MW)
Ordinary (max change < 20 MW)	DLinear	4847	11.62 ± 0.05	15.86 ± 0.05
Ordinary	iTransformer	4847	$\mathbf{9.99 \pm 0.47}$	$\mathbf{15.32 \pm 0.55}$
Ramp-positive (max change ≥ 20 MW)	DLinear	906	$\mathbf{38.30 \pm 0.06}$	$\mathbf{51.52 \pm 0.06}$
Ramp-positive	iTransformer	906	40.37 ± 0.39	54.63 ± 0.40

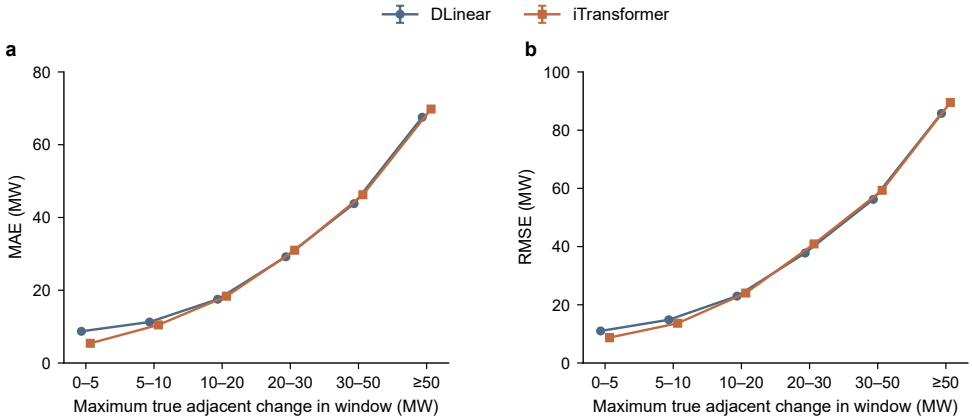


Figure 2. Forecast error by the maximum adjacent observed power change over the two-hour window: (**a**) MAE and (**b**) RMSE. Values show five-seed means and sample standard deviations. The six bins are $[0,5)$, $[5,10)$, $[10,20)$, $[20,30)$, $[30,50)$, and $\geq$50 MW. This post hoc analysis uses saved forecasts only.

3.3. Window-Level Ramp-Risk Early Warning

The test-window prevalence is 15.75%, giving a chance-level AP of 0.1575. Among head-free models, the iTransformer trajectory score reaches the highest AP, 0.338 ± 0.019, approximately 2.15 times the prevalence baseline. Its mean precision is 0.491 ± 0.037 and mean recall is 0.131 ± 0.019, corresponding to $F_1 = 0.206 \pm 0.023$. DLinear, despite the best regression R^2, reaches AP $= 0.287 \pm 0.015$ and does not cross the 20 MW alarm threshold at the evaluated operating point. This contrast provides direct evidence that aggregate regression accuracy and ramp-risk discrimination are distinct objectives. Figure 3 compares window-level AP and fixed-threshold F1 across the evaluated model variants.

Auxiliary risk heads shift the operating point but do not consistently improve global ranking. On iTransformer, the head increases mean F1 from 0.206 to 0.290 by increasing recall from 0.131 to 0.638, but AP decreases from 0.338 to 0.252 and regression accuracy also declines. On the Informer-SENet branch, the head produces the highest mean F1, 0.387 ± 0.036, with recall 0.554 ± 0.177, but its regression accuracy is the lowest and its cross-seed variance is large. Thus, the practical value of a head lies primarily in changing the alarm operating point rather than universally improving ranking quality. The five-seed precision–recall curves for the trajectory-based iTransformer score are presented in Figure 4.

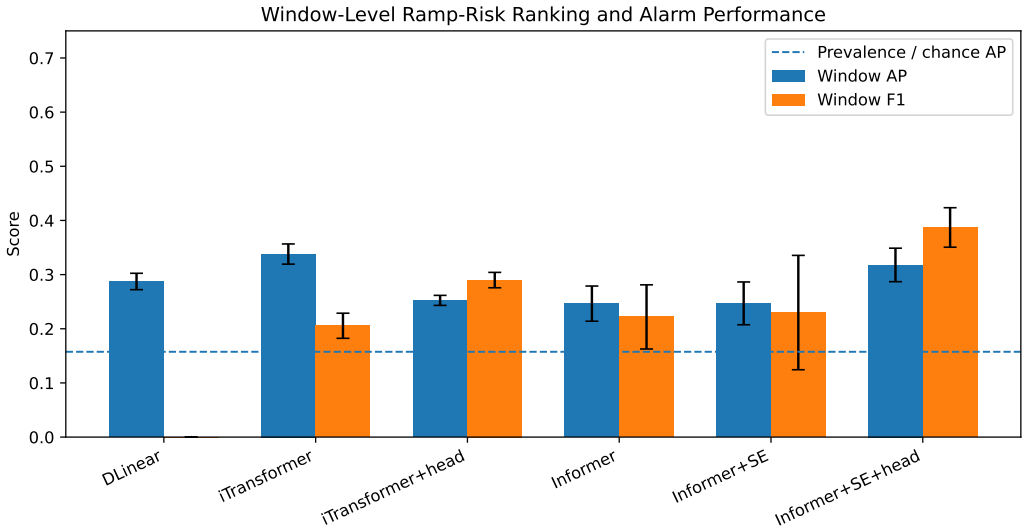


Figure 3. Five-seed window-level Average Precision (AP) and fixed-threshold F1. The dashed horizontal line denotes the test-window prevalence (0.1575), which is the expected AP of a non-informative ranking.

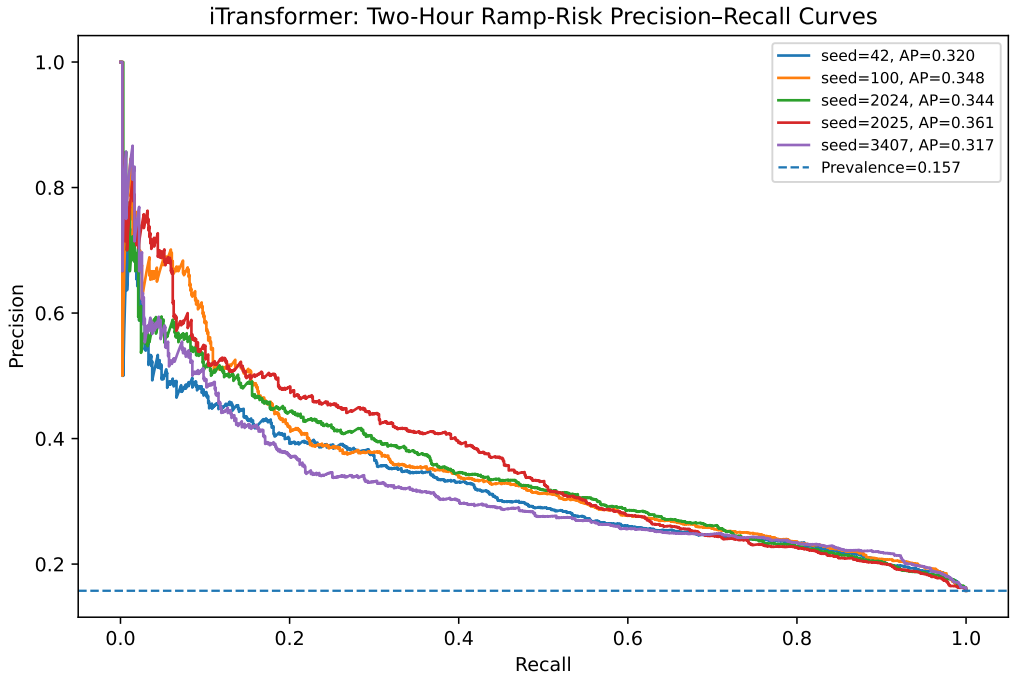


Figure 4. Precision–Recall curves for the trajectory-based iTransformer ramp-risk score across the five random seeds at the 20 MW/15-min event definition. The dashed horizontal line denotes the positive-window prevalence.

3.4. Sensitivity to the Ramp Threshold

To evaluate whether the window-level finding depends on the selected 20 MW criterion, the ramp threshold is varied to 10, 20, and 30 MW while keeping the forecasting models unchanged. Table 5 reports the five-seed iTransformer results using the trajectory-based score. As the threshold increases, the positive-window prevalence decreases from 38.14% to 7.07%, and the absolute AP decreases accordingly. However, AP remains above the corresponding non-informative prevalence baseline at all three thresholds. The AP-to-chance ratio increases from 1.58 at 10 MW to 2.15 at 20 MW and 2.46 at 30 MW, indicating that the ranking signal persists for rarer and more severe events. In contrast, fixed-threshold

F1 decreases with event rarity because the power-trajectory alarm becomes increasingly conservative. Figure 5 visualizes the corresponding AP-to-prevalence sensitivity.

Table 5. iTransformer sensitivity to the ramp threshold. AP and F1 are mean ± standard deviation across five seeds; precision and recall are five-seed means.

Threshold (MW)	Positive Rate	AP	AP/Chance	Precision	Recall	F1
10	0.3814	0.6028 ± 0.0266	1.58	0.6606	0.3015	0.4135 ± 0.0238
20	0.1575	0.3380 ± 0.0186	2.15	0.4908	0.1307	0.2056 ± 0.0232
30	0.0707	0.1738 ± 0.0180	2.46	0.3071	0.0501	0.0860 ± 0.0108

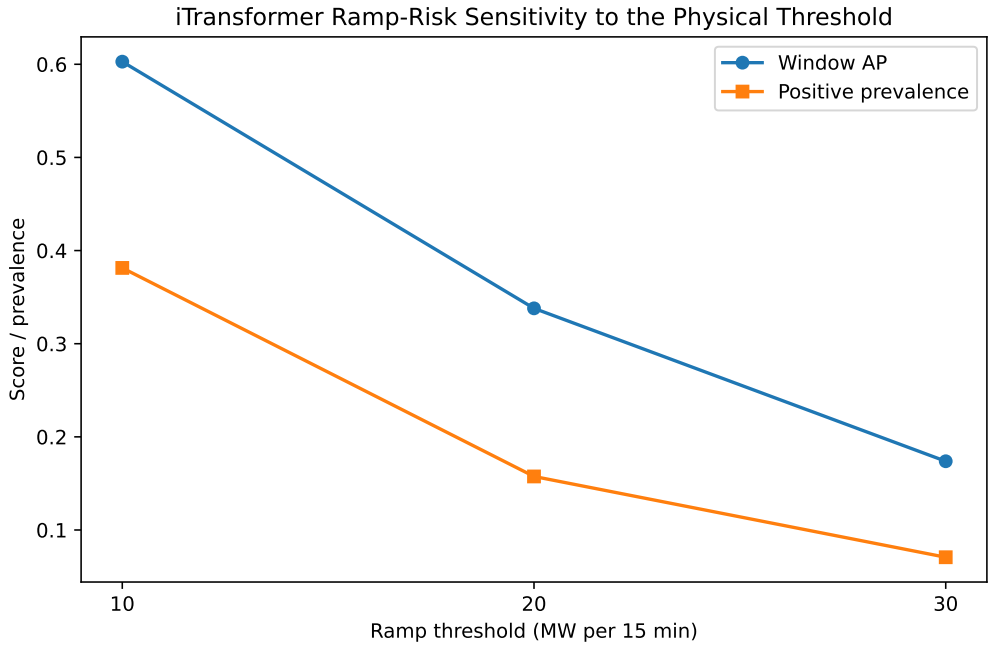


Figure 5. Sensitivity of iTransformer window-level Average Precision to the ramp threshold. The positive-window prevalence is shown for reference because it is the expected AP of a non-informative ranking.

3.5. Representative Two-Hour Forecast Windows

Figure 6 shows a representative positive test window selected by median eight-step trajectory MAE among all positive windows for seed 100, rather than by best-case visual fit. The observed trajectory contains a maximum adjacent 15-min change of 28.10 MW, while the iTransformer prediction has a maximum adjacent change of 30.22 MW. The predicted trajectory therefore correctly flags the two-hour window as high risk, although the timing and shape of the strongest transition are not reproduced exactly. This example illustrates why window-level warning can be useful even when exact intra-window ramp localization remains difficult.

For comparison, Figure 7 shows a representative non-ramp test window, also selected by median trajectory MAE within its class. The observed maximum adjacent 15-min change is 7.32 MW, and the predicted maximum change is 11.75 MW, both below the 20 MW alarm criterion.

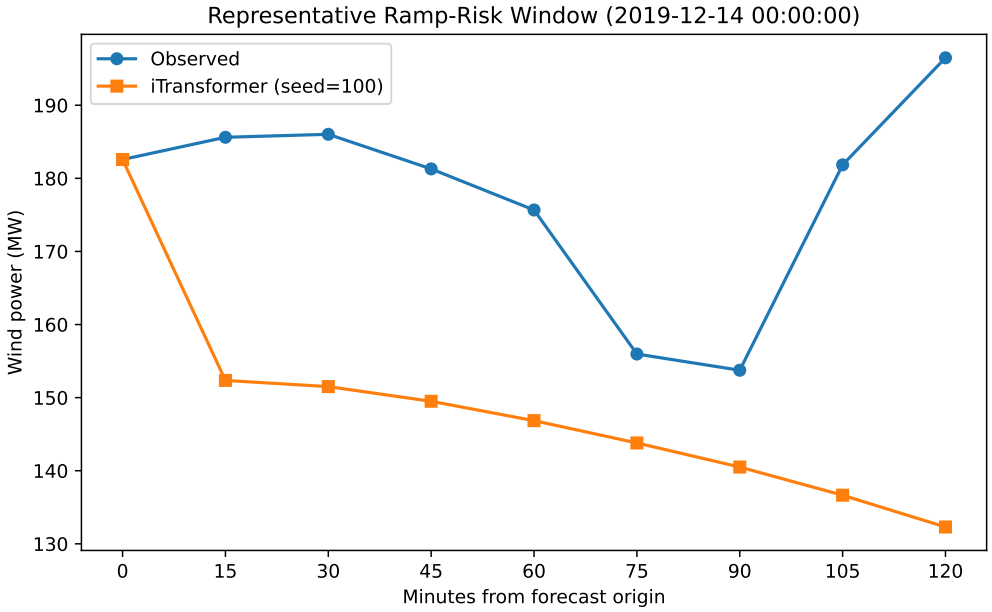


Figure 6. Representative positive ramp-risk window from the November–December test set (forecast origin 14 December 2019 00:00). The example is selected by median trajectory MAE among positive windows for seed 100.

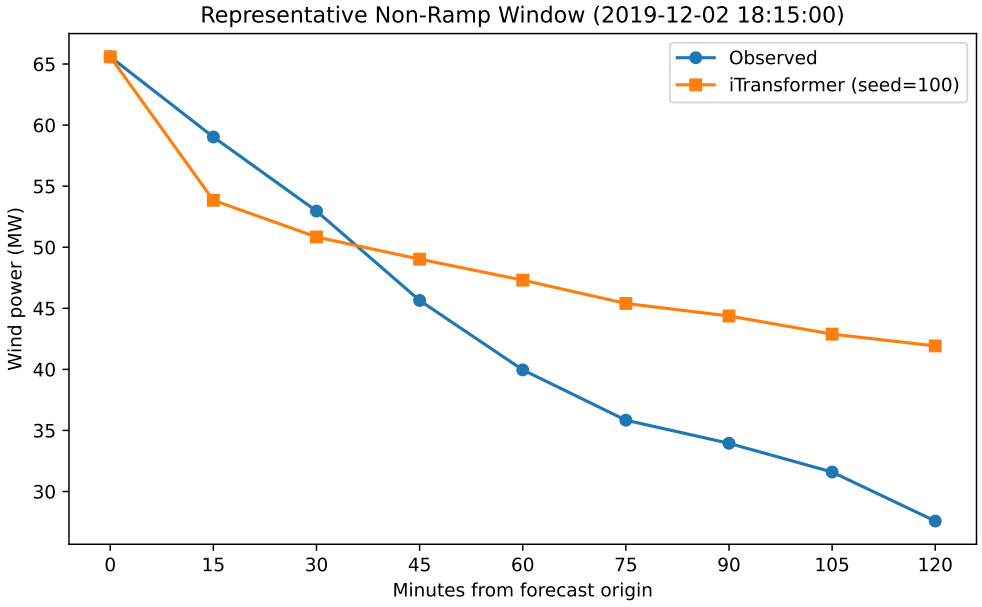


Figure 7. Representative non-ramp window from the November–December test set (forecast origin 2 December 2019 18:15), selected by median trajectory MAE among negative windows for seed 100.

3.6. Stability, Ablation, and Paired Comparisons

SE recalibration does not provide a stable cross-seed benefit in the present dataset. Relative to the no-SE Informer, mean R^2 decreases from 0.8479 to 0.8426 and the across-seed standard deviation increases from 0.0158 to 0.0235. The direction of the SE effect changes across seeds, indicating sensitivity to initialization rather than a reproducible gain.

The iTransformer auxiliary-head ablation provides a complementary result. In the seed-100 paired bootstrap over the 5753 test windows, adding the head changes R^2 by -0.00054 (95% CI $[-0.00146, 0.00038]$), while window AP decreases by -0.1043 (95% CI $[-0.1352, -0.0741]$) and window F1 increases by $+0.0902$ (95% CI $[0.0560, 0.1246]$). The archived resampling fraction for a positive AP difference is approximately zero and for a positive F1 difference is approximately one. These are descriptive under a window-wise

resampling scheme that does not preserve temporal dependence; they are not formal significance probabilities. The direction of the observed differences still illustrates the distinct objectives of ranking and fixed-threshold alarms. The exploratory paired-resampling estimates are reported in Table 6.

Table 6. Exploratory window-wise paired-resampling differences on seed-100 test predictions. These nominal intervals ignore dependence among overlapping windows and are not used as formal significance tests. Differences are first minus second; the final column is the resampling fraction with a positive difference.

Comparison	Metric	Mean Difference	95% CI Lower	95% CI Upper	$P(\Delta > 0)$
iTransformer+head vs. iTransformer	R^2	−0.00054	−0.00146	0.00038	0.1195
iTransformer+head vs. iTransformer	Window AP	−0.10432	−0.13518	−0.07405	0.0000
iTransformer+head vs. iTransformer	Window F1	+0.09021	+0.05602	+0.12461	1.0000
SE+head vs. SE (MSE)	R^2	−0.00064	−0.00237	0.00109	0.2290
SE+head vs. SE (MSE)	Window AP	+0.07856	+0.05314	+0.10438	1.0000
SE+head vs. DLinear	R^2	−0.00382	−0.00590	−0.00188	0.0000
SE+head vs. DLinear	Window AP	+0.04567	+0.01941	+0.07227	0.9990

Table 7 summarizes the principal ablation findings.

Table 7. Interpretation of the principal ablations.

Ablation	Observed Effect	Interpretation
Informer → Informer + SE	Lower mean R^2 and larger seed variance	No robust evidence that SE improves the present forecasting task
Power score → risk head	Higher operating-point F1 but lower AP and/or regression accuracy	Changes the alarm operating point rather than universally improving global ranking
Regression ranking vs. ramp ranking	DLinear best mean R^2/RMSE; iTransformer best AP	Regression accuracy and ramp-risk discrimination are distinct objectives

4. Discussion

4.1. Strong Baselines Are Essential in Very-Short-Term Wind Forecasting

The strong performance of persistence confirms that very-short-term wind power is highly autocorrelated. A high R^2 alone is therefore insufficient evidence that a complex model has learned additional predictive structure. DLinear further shows that a low-complexity model can achieve the highest mean R^2 and lowest RMSE in this dataset, consistent with broader time-series evidence that linear baselines can remain competitive with Transformer architectures [15]. The two architectures optimize the same power-regression objective but impose distinct inductive biases: DLinear models trend/seasonal decomposition and linear extrapolation, while iTransformer encodes variate-wise tokens and cross-variable relationships. The stratified results in Table 4 instead show lower iTransformer error in ordinary windows and lower DLinear error in ramp-positive windows. This observation is consistent with heterogeneous error distributions but does not prove that DLinear's trend decomposition caused a particular smoothing behavior or that variable-wise attention caused the differences. In particular, iTransformer's better window AP does not establish more accurate extreme-ramp amplitudes or timing. Moreover, the implemented DLinear farm-power output effectively depends on historical power only, while iTransformer uses multisource covariates; the comparison is therefore not a controlled test of architecture alone.

4.2. Regression Accuracy Is Not Equivalent to Ramp-Risk Skill

The clearest result of the study is the separation between aggregate regression accuracy and ramp-risk discrimination. DLinear provides the strongest mean regression metrics but not the highest risk AP. iTransformer provides slightly lower regression accuracy but the strongest trajectory-based risk ranking. This difference matters operationally because a model optimized for average error can under-represent rare transitions even when its point forecasts are accurate. Nevertheless, higher AP establishes stronger *ranking* of positive two-hour windows in these tests, not better timestamp-level localization or a uniformly superior prediction of every high-magnitude transition. Distinguishing these claims prevents architectural hypotheses about smoothing or channel independence from being presented as established experimental conclusions.

The two-hour window-level formulation reduces the extreme sparsity of exact 15-min event localization while preserving the same 20 MW/15-min case-study event definition. Instead of requiring the model to predict the exact future quarter-hour in which the event occurs, the task asks whether at least one critical transition will occur during the following two hours. For reserve scheduling and storage preparation, such a risk window can be more actionable than exact event timing.

4.3. Auxiliary Risk Heads Change the Operating Point

The auxiliary weighted-BCE and focal-style heads were introduced to address class imbalance, but the final experiments show that they should not be interpreted as universal improvements. On both iTransformer and Informer-SENet branches, adding a head can improve thresholded F1 while reducing regression accuracy and/or AP. This is a multi-objective trade-off: global risk ranking, power reconstruction, and a fixed alarm operating point are different objectives. The tested classification head was combined with MSE under a fixed weighted-loss coefficient ($\lambda = 0.05$ for the weighted-BCE configuration); dynamic uncertainty weighting and GradNorm were not evaluated. Head-derived probability scores and trajectory-derived MW scores also induce different operating-point definitions. Thus, improved F1 at a fixed cutoff is not equivalent to improved ranking; a validation-set recalibration of alarm thresholds might change F1 without modifying the regression model. Future work should evaluate multiple static coefficients, decoupled training, cost-sensitive calibration, and dynamic gradient-balancing methods before drawing broader conclusions about multi-task optimization.

4.4. Illustrative Operational Cost Sensitivity

For a two-hour advance warning, false alarms can prompt unnecessary reserve preparation, while missed ramps can lead to insufficient preparation. Without verified dispatch and market prices, we do *not* estimate monetary losses or optimal operator policy. Instead, a dimensionless illustrative misclassification cost per evaluated window is defined as

$$C_{\mathrm{rel}}(\tau; a) = \frac{FP(\tau) + a\,FN(\tau)}{N}, \qquad a = \frac{C_{\mathrm{FN}}}{C_{\mathrm{FP}}}, \tag{16}$$

where C_{FP} is normalized to one, a is an assumed relative cost, and τ is a fixed threshold on the iTransformer's trajectory-derived MW risk score. The calculation uses the five already-generated November–December test predictions and is strictly a *post hoc sensitivity description*; candidate score thresholds are not tuned on the held-out test set for any deployment claim. The test-window prevalence is 906/5753. Table 8 compares fixed warning score cutoffs with never- and always-alarm reference policies. For example, at $a = 5$ the 10 MW score cutoff has relative cost 0.600 ± 0.023 versus 0.787 for never warning, whereas at $a = 1$ never warning (0.157) is competitive with the conservative 20 MW score cutoff

(0.159 ± 0.003). At $a = 10$, always warning (0.843) incurs lower stipulated error cost than either of the two shown score cutoffs, demonstrating that a threshold cannot be declared operationally optimal from AP or F1 alone. Overlapping two-hour windows can trigger repeated alarms, so this per-window proxy should not be confused with a reserve-dispatch economic analysis.

Table 8. Exploratory normalized misclassification cost per test window (mean ± standard deviation over five iTransformer seeds for model-based alerts). Cost ratios are assumptions, not measured economic parameters; threshold values are fixed warning-score cutoffs, distinct from the 20 MW event definition.

C_{FN}/C_{FP}	Never Alarm	Always Alarm	Score $\geq$ 10 MW	Score $\geq$ 20 MW
1	0.157	0.843	0.212 ± 0.015	0.159 ± 0.003
2	0.315	0.843	0.309 ± 0.016	0.295 ± 0.004
5	0.787	0.843	0.600 ± 0.023	0.706 ± 0.012
10	1.575	0.843	1.086 ± 0.040	1.391 ± 0.026

Retrospective Month-Separated Threshold Check

To make the decision dependence more explicit without retraining, we also selected each model's alarm-score cutoff using November windows and froze it for evaluation on December windows. November forecasts whose targets crossed into December were excluded. The December subset contains 2968 windows, including 385 positive windows. The assumed cost ratio a is varied across 1, 2, 5, and 10; costs are reported per 1000 December windows (Table 9). DLinear has lower observed stipulated cost at $a = 5$ (449.46 versus 476.42), whereas iTransformer has lower cost at $a = 10$ (680.39 versus 864.15). The comparison does not demonstrate prospective economic savings: both months belong to the repeatedly inspected original test period, and warning windows overlap. The cost accounting excludes ramp direction, physical dispatch, reserve activation, and alert de-duplication. Figure 8 visualizes the illustrative cost comparisons for the two evaluation subsets.

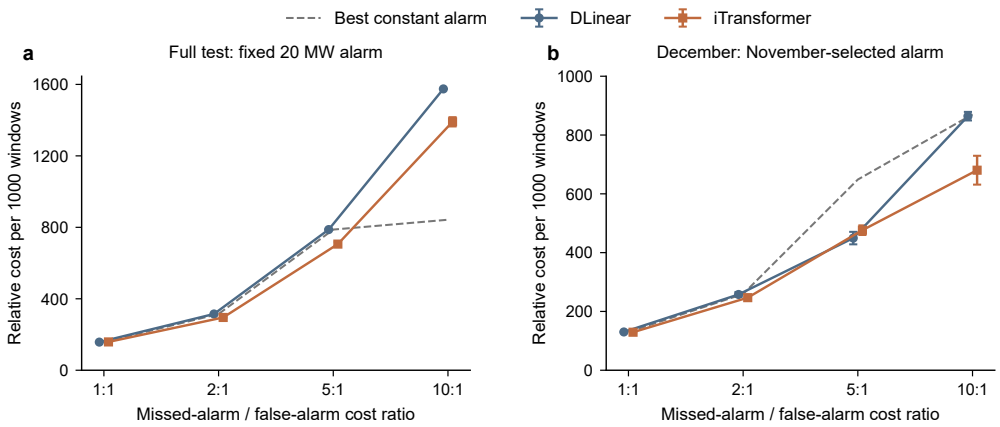


Figure 8. Illustrative relative false-alarm and missed-alarm costs for DLinear and iTransformer: (**a**) original fixed warning cutoff on all test windows; (**b**) November-selected cutoff evaluated on December windows. The plots use different evaluation sets and must not be interpreted as real grid financial returns or independent prospective testing.

Table 9. Retrospective November-selected threshold evaluation on December windows: hypothetical costs per 1000 windows (mean ± standard deviation across five seeds). The true event remains a ≥20 MW adjacent change; cost ratios are not empirical monetary prices. Bold indicates the lower hypothetical cost at each assumed cost ratio.

Assumed $C_{FN} : C_{FP}$	DLinear	iTransformer
1:1	130.19 ± 1.36	**129.31 ± 2.57**
2:1	257.68 ± 8.42	**246.90 ± 3.46**
5:1	**449.46 ± 21.23**	476.42 ± 16.35
10:1	864.15 ± 14.59	**680.39 ± 49.06**

This analysis identifies the decision trade-off but not its financial magnitude. A deployable warning system would require verified reserve-activation costs, the asymmetric consequence of missed ramps, calibration on separate historical periods, and assessment of how consecutive overlapping alerts are consolidated.

4.5. Limitations

Several limitations should be considered. First, the dataset was obtained through a third-party public data-sharing source. The site attribution is consistent with the accompanying description and with independent public operator metadata, but the authors did not participate in data acquisition and cannot independently verify the raw SCADA provenance. The findings should therefore be interpreted as a methodological case study rather than as site-certified operational validation. Second, the study uses one year of data from a single attributed site, so cross-site generalization is not evaluated. Transfer to offshore farms, mountain/complex terrain, or different wind regimes cannot be inferred from these data. A second, structurally distinct site and season-stratified external evaluation would be required to establish transferability. Third, although sensitivity analysis at 10/20/30 MW confirms above-chance ranking across all three thresholds, these are absolute power criteria; transfer to other farms should preferably use capacity-normalized definitions once reliable site metadata are available. Fourth, the exact positive prevalence varies seasonally, so season-stratified analyses would strengthen the operational interpretation. Fifth, a data-integrity audit of the archived processed file found two residual sentinel-coded meteorological records (−99), one in validation and one in testing. The downstream loader did not replace these markers. Excluding the 96 affected test-input windows in a retrospective sensitivity calculation changes overall DLinear MAE from 15.82 to 15.77 MW and iTransformer MAE from 14.77 to 14.69 MW, without reversing their overall ordering; this cannot rule out an effect of the validation anomaly on checkpoint selection or substitute for full retraining on causally cleaned data. The older one-step main script also contains whole-series preprocessing and test-based checkpointing that would be inappropriate for final evaluation. The saved eight-step runs document a distinct month-split, train-fitted scaling and validation-checkpoint pathway, but the full upstream cleaning history and a cryptographically verified historical code snapshot are unavailable. Moreover, later examinations of November–December outcomes make the supplemental threshold comparisons retrospective rather than independent. Sixth, some neural architectures show seed sensitivity, five-seed variation is not a cross-farm uncertainty interval, and overlapping forecast windows violate naive independent-window assumptions. Finally, the current study is deterministic; its point trajectories and derived scores are not calibrated occurrence probabilities. Quantile regression or probabilistic trajectory forecasting would permit direct estimation of ramp exceedance likelihood, prediction-interval coverage, and decision-theoretic reserve scheduling; such extensions require proper calibration and are

beyond the present experiments [17]. The illustrative cost ratios in Table 8 are not actual operator costs.

5. Conclusions

This study evaluated two-hour multi-step wind-power forecasting and window-level ramp-risk early warning under a chronologically partitioned protocol using a publicly shared year-long wind-power and multisource meteorological dataset.

The quantitative evaluation yields three principal findings. Relative to the 4847 ordinary test windows, the 906 ramp-positive windows produce substantially larger trajectory errors: iTransformer gives lower ordinary-window MAE (9.99 versus 11.62 MW for DLinear), but DLinear gives lower ramp-window MAE (38.30 versus 40.37 MW). This regime dependence must not be conflated with the risk-score AP ranking. First, strong simple baselines are indispensable. DLinear achieves the highest mean R^2 (0.8645 ± 0.0004) and lowest RMSE (25.10 ± 0.03 MW), while persistence remains highly competitive. Second, regression accuracy and ramp-risk discrimination are distinct objectives. iTransformer provides the lowest MAE among the learnable models (14.77 ± 0.34 MW) and the strongest trajectory-based window risk ranking at 20 MW (AP $= 0.338 \pm 0.019$), more than twice the 0.1575 prevalence baseline. Threshold sensitivity at 10/20/30 MW further shows that iTransformer AP remains 1.58/2.15/2.46 times the corresponding prevalence baselines. Third, auxiliary risk heads can improve a thresholded F1 operating point—up to 0.387 ± 0.036 in the tested Informer-SENet head variant—but this improvement is accompanied by trade-offs in regression accuracy, ranking quality, or stability.

Overall, the experiments support a task-driven rather than architecture-driven interpretation: multi-step power forecasting, global risk ranking, and threshold-specific alarm performance should be evaluated separately. For the present dataset, two-hour window-level warning provides a more useful operational target than exact 15-min ramp localization. A post hoc cost-sensitivity analysis using fixed warning score thresholds illustrates why the preferred alarm operating point changes with the assumed relative cost of misses and false alerts; no monetary savings or calibrated deployment threshold are claimed. The approximate 200 MW public capacity attribution only contextualizes but does not verify the absolute 10/20/30 MW criteria. The single-site, one-year setting and the unverified raw acquisition chain limit external validity. Future work should assess independently validated capacity-normalized events, fully causal preprocessing, probabilistic trajectories, season-aware threshold calibration, and cross-farm transferability.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/en19235708/s1; The Supplementary Materials contain Figures S1–S5, Tables S1–S7, and supporting analysis outputs derived from saved predictions. These materials are descriptive and do not constitute an independently untouched evaluation.

Author Contributions: Conceptualization, H.D. and W.W.; methodology, H.D.; software, H.D.; validation, H.D.; formal analysis, H.D.; investigation, H.D.; data curation, H.D.; writing—original draft preparation, H.D.; writing—review and editing, H.D. and W.W.; visualization, H.D.; supervision, W.W.; project administration, W.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52267005.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study was obtained from the publicly accessible CSDN resource “Xinjiang wind-farm data at 15-min intervals” (resource ID 88485424); the direct

download address and indexed community page are provided in Ref. [27]. The downloaded archive contains the year-long workbook used as the upstream source of the processed experimental data. A related, shorter subset has also been shared through the Alibaba Cloud Tianchi platform [30]. The workbook metadata and public description associate the records with Jingxia North Wind Farm in Hami, Xinjiang; however, because the authors did not participate in the original measurement campaign, the original SCADA acquisition chain cannot be independently verified. The processed-data transformation description, model configurations, and derived evaluation outputs used to reproduce the reported results are available from the corresponding author upon reasonable request; redistribution of the original or processed records remains subject to the terms of the public source.

Acknowledgments: During preparation of this manuscript, a generative AI tool was used for language editing and LaTeX organization. All scientific content, analyses, and conclusions were reviewed and verified by the authors, who take full responsibility for the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SCADA	Supervisory Control and Data Acquisition
WPRE	Wind Power Ramp Event
SE	Squeeze-and-Excitation
AP	Average Precision
PR	Precision–Recall
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
BCE	Binary Cross-Entropy

References

1. Foley, A.M.; Leahy, P.G.; Marvuglia, A.; McKeogh, E.J. Current Methods and Advances in Forecasting of Wind Power Generation. *Renew. Energy* **2012**, *37*, 1–8. [CrossRef]
2. Jung, J.; Broadwater, R.P. Current Status and Future Advances for Wind Speed and Power Forecasting. *Renew. Sustain. Energy Rev.* **2014**, *31*, 762–777. [CrossRef]
3. Tascikaraoglu, A.; Uzunoglu, M. A Review of Combined Approaches for Prediction of Short-Term Wind Speed and Power. *Renew. Sustain. Energy Rev.* **2014**, *34*, 243–254. [CrossRef]
4. Colak, I.; Sagiroglu, S.; Yesilbudak, M. Data Mining and Wind Power Prediction: A Literature Review. *Renew. Energy* **2012**, *46*, 241–247. [CrossRef]
5. Pinson, P. Wind Energy: Forecasting Challenges for Its Operational Management. *Stat. Sci.* **2013**, *28*, 564–585. [CrossRef]
6. Lim, B.; Zohren, S. Time-Series Forecasting with Deep Learning: A Survey. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200209. [CrossRef]
7. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; Volume 30, pp. 5998–6008.
8. Lim, B.; Arik, S.O.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
9. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34.
10. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview.net: Appleton, WI, USA, 2023.
11. Wang, S.; Sun, Y.; Zhang, W.; Chung, C.Y.; Srinivasan, D. Very Short-Term Wind Power Forecasting Considering Static Data: An Improved Transformer Model. *Energy* **2024**, *312*, 133577. [CrossRef]
12. Liu, L.; Wang, X.; Dong, X.; Chen, K.; Chen, Q.; Li, B. Interpretable Feature-Temporal Transformer for Short-Term Wind Power Forecasting with Multivariate Time Series. *Appl. Energy* **2024**, *374*, 124035. [CrossRef]
13. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [CrossRef]

14. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview.net: Appleton, WI, USA, 2024.
15. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers Effective for Time Series Forecasting? *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11121–11128. [CrossRef]
16. Wang, J.; Botterud, A.; Bessa, R.; Keko, H.; Carvalho, L.; Issicaba, D.; Sumaili, J.; Miranda, V. Wind Power Forecasting Uncertainty and Unit Commitment. *Appl. Energy* **2011**, *88*, 4014–4023. [CrossRef]
17. Pinson, P.; Madsen, H.; Nielsen, H.A.; Papaefthymiou, G.; Klockl, B. From Probabilistic Forecasts to Statistical Scenarios of Short-Term Wind Power Production. *Wind Energy* **2009**, *12*, 51–62. [CrossRef]
18. Gallego-Castillo, C.; Cuerva-Tejero, A.; Lopez-Garcia, O. A Review on the Recent History of Wind Power Ramp Forecasting. *Renew. Sustain. Energy Rev.* **2015**, *52*, 1148–1157. [CrossRef]
19. Cutler, N.; Kay, M.; Jacka, K.; Nielsen, T.S. Detecting, Categorizing and Forecasting Large Ramps in Wind Farm Power Output Using Meteorological Observations and WPPT. *Wind Energy* **2007**, *10*, 453–470. [CrossRef]
20. Bossavy, A.; Girard, R.; Kariniotakis, G. Forecasting Ramps of Wind Power Production with Numerical Weather Prediction Ensembles. *Wind Energy* **2013**, *16*, 51–63. [CrossRef]
21. Cui, Y.; Chen, Z.; He, Y.; Xiong, X.; Li, F. An Algorithm for Forecasting Day-Ahead Wind Power via Novel Long Short-Term Memory and Wind Power Ramp Events. *Energy* **2023**, *263*, 125888. [CrossRef]
22. Hu, J.; Zhang, L.; Tang, J.; Liu, Z. A Novel Transformer Ordinal Regression Network with Label Diversity for Wind Power Ramp Events Forecasting. *Energy* **2023**, *280*, 128075. [CrossRef]
23. Fu, J.; Ni, Y.; Ma, Y.; Zhao, J.; Yang, Q.; Xu, S.; Zhang, X.; Liu, Y. A Visualization-Based Ramp Event Detection Model for Wind Power Generation. *Energies* **2023**, *16*, 1166. [CrossRef]
24. Miele, E.S.; Ludwig, N.; Corsini, A. Multi-Horizon Wind Power Forecasting Using Multi-Modal Spatio-Temporal Neural Networks. *Energies* **2023**, *16*, 3522. [CrossRef]
25. Xu, S.; Wang, Y.; Xu, X.; Shi, G.; Zheng, Y.; Huang, H.; Hong, C. A Multi-Step Wind Power Group Forecasting Seq2Seq Architecture with Spatial-Temporal Feature Fusion and Numerical Weather Prediction Correction. *Energy* **2024**, *291*, 130352. [CrossRef]
26. Zhang, F.; Ren, X.; Liu, Y. A Refined Wind Power Forecasting Method with High Temporal Resolution Based on Light Convolutional Neural Network Architecture. *Energies* **2024**, *17*, 1183. [CrossRef]
27. CSDN Community. Wind-Farm Data from Xinjiang at 15-min Intervals for Wind-Power Forecasting, Wind-Speed Forecasting, and Control Studies; Resource ID 88485424. Available online: https://download.csdn.net/download/weixin_42671195/88485424 (accessed on 23 September 2026).
28. CHN Energy Group E-Bidding. Public Project Information for Jingxia Wind Farm Facilities, Hami, Xinjiang; Site Information Reports Approximately 200 MW Installed Capacity and 133 1.5 MW Wind Turbines. Available online: https://www.chnenergybidding.com.cn/bidweb/001/001002/001002003/20250311/4ef9da58-a886-4bc5-830e-9cbb169344df.html (accessed on 23 September 2026).
29. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2011–2023. [CrossRef]
30. Alibaba Cloud Tianchi. Wind Power Forecasting Dataset. Available online: https://tianchi.aliyun.com/dataset/159885 (accessed on 23 September 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.