Journal of Science and Technology – Smart Systems and Devices
Volume x, Issue x, Month Year, ppp–ppp

Short-Term Wind Speed Forecasting from Public Gridded Weather Data at a Coastal Site in Southern Vietnam: a Leakage-Audited Validation Methodology

Melissa Cassan[1,2], Huong Pham Thi Lan[1,3,*], Dang Pham Quang[1], Thanh Bui Dang[1]

[1] *Hanoi University of Science and Technology, Ha Noi, Vietnam*
[2] *Arts et Metiers Institute of Technology, Paris, France*
[3] *Vietnam National University of Agriculture, Ha Noi, Vietnam*
**Corresponding author email: huongpham@vnua.edu.vn*

Abstract

Coastal microgrids in the ASEAN region increasingly rely on wind generation, and their operators schedule storage, dispatchable units and reserves on forecasts issued one to six hours ahead. This paper presents a short-term wind speed forecasting pipeline for a coastal site near Bac Lieu, southern Vietnam, and gives as much attention to the protocol used to validate it as to the pipeline itself. The study is restricted to wind speed; the translation of speed errors into power errors is left to future work. Only public gridded data are used (the Open-Meteo archive, which serves ERA5 reanalysis data for 2012–2016 and ECMWF IFS data from 2017 onwards), so that the method can be applied before any measurement mast is installed. The site has a moderate wind regime, with a mean speed of 5.37 m/s at 100 m over 2012–2023. Eight model families were compared against persistence; a gated recurrent unit network was tuned by a Bayesian hyperparameter search and followed by a gradient-boosted residual correction stage. On the 2023 test period (8,394 hours) the pipeline reaches a mean absolute error of 0.542 m/s, a root mean squared error of 0.753 m/s, 12 % lower than persistence, and a coefficient of determination of 0.861. Because model selection had used earlier periods that overlap 2023, the frozen pipeline was also applied to 2024–2025, never used in its development: its mean absolute error is 0.546 m/s and its root mean squared error 10 % lower than persistence. The methodological findings matter more than these figures. Four data leakage errors were found and corrected in our own code, and a further audit during revision found a unit error and comparisons made across different test periods; all results are therefore restated on explicitly identified test periods. Against a control fitted on shuffled residuals, the structured contribution of the correction stage is 0.003 m/s. A rolling ARIMA model, once evaluated correctly, is as accurate as the selected network on the architecture comparison, and applying a signal decomposition before rather than after the split lowers the apparent error by 41 % on identical test targets, whereas the causal decomposition brings no meaningful gain.

Keywords: Conformalized quantile regression, data leakage, gated recurrent unit, validation protocol, wind speed forecasting.

In press

1. Introduction

Wind generation is entering ASEAN distribution systems through units that are small enough to be operated as microgrids: coastal clusters of a few turbines combined with photovoltaic arrays, battery storage and a diesel or gas back-up, serving an industrial load or a weakly interconnected feeder. In such a system the operator has to decide, several hours in advance, how much energy to hold in the battery, whether to start a dispatchable unit, and how much operating reserve to carry. All three decisions are taken on a forecast rather than on a measurement, and errors in both directions have a cost: over-predicting the wind may leave too little dispatchable generation or operating reserve committed, whereas under-predicting it leads to excess reserve, unnecessary starts of the back-up unit and curtailment of wind that was in fact available.

The horizon that matters for these decisions is one to six hours. It is short enough that the persistence of the atmosphere still carries most of the information, and long enough that a purely reactive control strategy is no longer sufficient. It is also the horizon at which the published literature is most abundant, and, as shown in section 3, most difficult to trust.

This paper reports the construction of a forecasting pipeline for a coastal site in southern Vietnam, under two constraints. The first is that only publicly available data may be used, so that a candidate microgrid site can be assessed before any measurement mast is installed. The second is that every improvement claimed must survive a control experiment designed to make it disappear if it is not real. The scope is deliberately limited to wind speed: the conversion of speed errors into power errors, which depends on the turbine power curve, is discussed in section 8 but is not evaluated here.

None of the algorithmic components is new. The gated recurrent unit [1], the tree-structured Parzen estimator [2], gradient-boosted trees [3] and conformalized quantile regression [4] are used as published, and the long short-term memory network [5] and an ARIMA model [6] serve as comparators.

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.xxx.ssad.xxxx.xx.x.x
Received: Aug 30, 2026; Revised: Oct 19, 2026;
Accepted: Oct 22, 2026; Online: Nov 20, 2026

The contribution lies in the way they are validated, and it is fourfold. First, a five-zone chronological protocol gives each decision of the two-stage pipeline a period that no earlier decision has used, and we report the leakage errors that this protocol exposed in our own code. Second, the inflation produced by applying a signal decomposition before rather than after the train-test split is measured like-for-like on identical data, complementing earlier warnings about this practice [7–9]. Third, the contribution of the residual correction stage is measured against a control fitted on shuffled residuals rather than against an older version of the pipeline. Fourth, every result is tied to an explicit test period, and results are only compared when their test periods coincide. Fig. 1 summarises the workflow and separates the adopted components from these contributions.

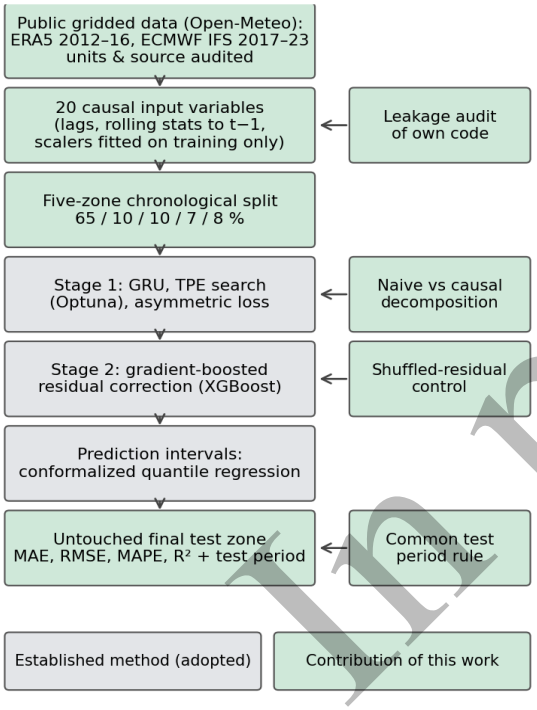


Fig. 1. Workflow of the study. Grey boxes are established methods used as published; green boxes are the validation elements that constitute the contribution of this paper

One characteristic of the study site conditions the reading of every error figure and is therefore stated at the outset: the mean speed at 100 m is 5.37 m/s, and 86 % of the hours lie between 3 and 12 m/s, that is, roughly between the cut-in and rated speeds of a typical turbine. Errors expressed in metres per second must be read relative to this level and must not be compared with those obtained at sites with a different mean speed without normalisation.

2. Site, Sata and Wind Regime

2.1. Operating Context

The grid point lies on the coast of Bac Lieu province, in the Mekong Delta region of southern Vietnam (requested coordinates 9.3° N, 105.8° E). It is representative of the coastal strip targeted by Vietnamese wind development, where near-shore turbines are increasingly connected to local distribution networks that can be operated in islanded mode. The model grid cells that serve these coordinates are centred on land, within the urban area of Bac Lieu (section 2.2), so the data describe a coastal rather than an offshore wind. The wind regime is dominated by the north-east monsoon, which brings the strongest and most regular winds of the year, with a peak in February.

2.2. Data Source

The data were retrieved from the Open-Meteo historical weather archive [10], which serves hourly fields from several products of the European Centre for Medium-Range Weather Forecasts (ECMWF). The retrieved series is a concatenation of two sources: for 2012–2016 the values are those of ERA5, the fifth-generation ECMWF reanalysis [11], at the 0.25° grid cell centred on 9.25° N, 105.75° E; from 1 January 2017 they are those of the ECMWF Integrated Forecasting System (IFS) at about 9 km resolution, at the cell centred on 9.31° N, 105.78° E. Both cells have an elevation of 2 m. Every hourly value matches one of the two products exactly, so all the test periods used here (2022–2025) lie in the IFS part of the record. The two parts are not statistically identical: the mean speed at 100 m is 5.60 m/s over 2012–2016 and 5.21 m/s over 2017–2023 (a difference that may reflect the change of model as well as interannual variability). Gridded data are not measurements: they give a consistent, gap-free description of the atmosphere, but as a grid-cell average (roughly 28 km for ERA5, 9 km for IFS) rather than the wind at a turbine.

Four variables were retrieved: wind speed at 10 m and at 100 m, wind direction at 100 m, and air temperature at 2 m. The target is the speed at 100 m, which corresponds to the hub height of the machines currently deployed in the region. The dataset covers the twelve years from 2012 to 2023, that is 105,192 hourly records with no missing value; no imputation was performed at any point. The archive returns wind speeds in km/h unless metres per second are explicitly requested; all speeds have been converted, and every statistic and error in this paper is expressed in m/s (section 3.5).

Two consequences follow. A model trained on gridded data learns a smoothed atmosphere, so short gusts are attenuated and the errors below cannot be transposed to mast measurements without a local calibration; in return, the pipeline can be applied at any coordinate, which makes the second-site test of section 8 possible.

In press

2.3. Wind Regime

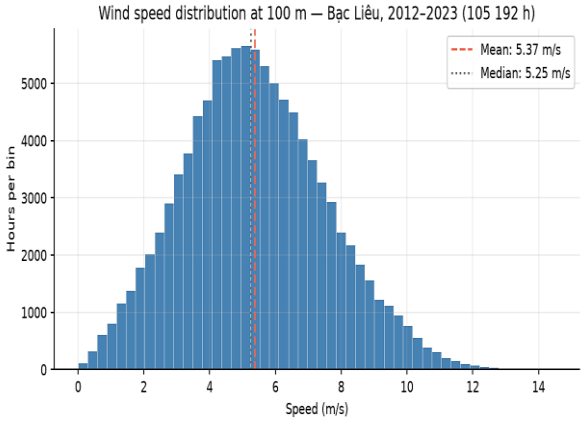


Fig. 2. Distribution of the wind speed at 100 m over 2012–2023 (105,192 hourly values)

The distribution of the wind speed at 100 m is unimodal and slightly right-skewed, as Fig. 2 shows. Over 2012–2023 the mean is 5.37 m/s, the median 5.25 m/s and the standard deviation 2.18 m/s; the maximum is 14.53 m/s, on 4 January 2019. 13.9 % of the hours lie below 3 m/s, 85.9 % between 3 and 12 m/s and 0.2 % above 12 m/s. An interquartile-range outlier test (1.5 × IQR) flags 570 hours, or 0.54 % of the record, all above 11.23 m/s. These hours lie on the continuous upper tail of the distribution and no physical reason was found to discard them, so they were kept: removing them would delete the strongest-wind situations from the training data.

The seasonal signal is clear: February is the windiest month, with an average of 7.06 m/s, May the calmest at 4.16 m/s, and October averages 4.26 m/s. The autocorrelation of the series is 0.949 at a lag of one hour and remains high over several hours, with a clearly visible 24-hour cycle.

Operationally, this regime means that most hours fall in the range where the power output of a turbine rises steeply with speed, so a given speed error translates into a comparatively large power error; a power-domain evaluation is therefore the natural extension of this work (section 8), although it is outside the scope of this paper.

2.4. Input Variables

Twenty input variables were built from the four raw ones. They are listed in Table 1.

Table 1. The twenty input variables, grouped by nature

Group	No.	Description
Lagged speed	4	Speed at 100 m at t-1, t-2, t-6, t-24 h
Rolling statistics	6	Mean, standard deviation and maximum over 6 h and 24 h windows
Cyclic time	4	Hour and month as sine and cosine pairs
Direction	2	Direction at 100 m as sine and cosine
Monsoon indicator	1	Binary flag for the north-east monsoon
Raw context	3	Speed at 10 m, direction, temperature at 2 m

Two design choices deserve comment. The first is the cyclic encoding of time and direction, so that 23:00 and 00:00, or 359 and 1 degrees, are close rather than maximally distant.

The second is causality, and it is the rule from which the rest of the paper follows. Every variable is constructed so that it is available at the instant the forecast is issued. Rolling statistics at each input hour cover windows ending at that hour, and the input window ends at $t-1$, so no statistic includes the hour being forecast; where a rate of change is needed it is computed as a backward difference and never as a centred one, since a centred difference uses the point that follows the current one; and the scalers are fitted on the training period only and then applied unchanged to the later periods. Each of these rules was violated at least once in an early version of our code, as reported in section 3.3.

3. Validation Protocol and Leakage Audit

3.1. Why the Split Must be Chronological

The series is cut chronologically and never shuffled. The reason is the autocorrelation reported above. With a coefficient of 0.949 at one hour, two consecutive hours are nearly identical; a random split places hour t in the training set and hour $t+1$ in the test set, so the model is not forecasting at all, it is interpolating between two points it has already seen. Any architecture then appears excellent and the comparison between architectures becomes meaningless.

3.2. What the Literature Does

Eight recent studies on short-term wind speed or wind power forecasting with deep or hybrid models [12–19] were reviewed against a common set of criteria: the nature of the validation split, the position of any pre-processing relative to that split, the completeness of the reported metrics, and the comparability of the sites used. Table 2 summarises the review.

In press

Table 2. Validation practice in eight recent studies; split proportions are training/validation/test

Ref.	Target and data	Decomposition and split	Metrics
[11]	Speed, 70 m, 15 min, one month	TVF-EMD on full record*; 70/10/20	R^2, MAE, RMSE, EV
[12]	Power, two farms, 10–15 min	MCEEMDAN before split*; 60/20/20; scalers fitted on training data	MAE, RMSE, R^2
[13]	Power, 10 min	None; 80/20, results on validation set	RMSE, MAE, R
[14]	Power, 15 min, four months	VMD, order not stated; 90/10, no validation set; input window chosen on test errors	RMSE, MAE, MAPE, R^2
[15]	Speed, buoy at 4.9 m, 10 min, mean 3.8–5.9 m/s	TVFEMD on full record, then 70/30	RMSE, MAE, MAPE, R
[16]	Speed, 50 m, monthly	None; 50/50	MAE, RMSE, MAPE, R^2, EC, COV
[17]	Speed, 10 m, hourly, mean 1.8–2.8 m/s	IVMD on training set only; 70/10/20	MAE, RMSE, MAPE
[18]	Speed, tower, hourly, mean 2.0 m/s	VMD, order not stated; scaling on global extremes; 80/10/10	MAE, RMSE, MSE, R^2

* As far as can be determined from the description.

Three findings stand out. First, six of the eight studies use a signal decomposition, most often a variant of empirical mode decomposition or variational mode decomposition [20]. Only one restricts it explicitly to the training data [18]; in three it is applied, as far as the descriptions allow us to judge, to the whole record before the split [12, 13, 16], and in two the order is not stated [15, 19]. The decomposition of each point then depends on the whole series, so information from the test period reaches the training inputs. The problem has been recognised [7–9]: Wang and Wu [7] showed that decomposition-based hybrids can lose their advantage once the decomposition is recomputed as in practical forecasting, and Li *et al.* [9] proposed a rolling decomposition that excludes future information. Second, the roles of the data subsets are often not fully separated: one study reports results on its validation set [14], one selects part of its configuration on test errors [15], and one fits the normalisation on the global extremes of the series [19], a mild form of leakage [21]. Third, reported errors are largely incomparable: targets, heights (4.9 to 100 m), time steps and mean speeds (about 2 to more than 9 m/s) all differ.

3.3. What the Audit Found in Our Own Code

The same severity was applied to the code written for this study, and four leakage errors were found. The first was a bootstrap resampling procedure that ignored temporal order and broke sequence continuity. The second was a correction model trained and evaluated on the same period. The third was a variable alignment error that gave the correction model the wind speed at 10 m for the very hour being predicted, which is very nearly the answer. The fourth was the centred-difference gradient mentioned in section 2.4, which uses one future point at every step. Each made the results look better, and none would have been caught by inspecting the training curves.

3.4. The Five-Zone Protocol



Fig. 3. Reconstructed validation protocol: five consecutive chronological zones, one per decision taken in the pipeline

The protocol was consequently rebuilt around the principle that each decision must be taken on a period no earlier decision has used. The series is divided into five consecutive chronological zones representing 65 %, 10 %, 10 %, 7 % and 8 % of the record, as shown in Fig. 3: training of the recurrent network, its early stopping, training of the correction model, its early stopping, and the final test. The final test zone covers 15 January–31 December 2023 (8,414 hourly rows, i.e. 8,394 forecast targets once the 20-hour input window is formed) and was not used for any decision of the five-zone pipeline, meaning no tuning, no early stopping and no variable selection. It is not strictly unseen, however: the architecture and input window (section 4.2) and the hypotheses and loss comparison of sections 4.3–4.4 were assessed earlier on test periods that overlap it (Table 3). The frozen pipeline was therefore also evaluated on 2024–2025, which was not used at any stage (section 8).

A single methodological rule accompanies the protocol and is applied without exception: every candidate improvement is evaluated against the current validated reference on the same test period, since comparing a new idea with an older, weaker version makes almost any idea look successful.

Not every result in this paper could be computed on that zone, because the study was carried out in stages. Table 3 lists the data set, split and test period behind every reported figure. Results are only compared when they share a test period, and the text says so explicitly where a figure from another period is quoted.

In press

Table 3. Data set, split and test period behind each result (first and last forecast targets, UTC)

Results	Data, split	Test period	Targets
Architecture comparison (Table 4, Fig. 4)	2022–2023, 70/15/15	14 Sep–31 Dec 2023	2,605
Tuned GRU (section 4.3); multi-horizon model (section 7)	2012–2023, 70/15/15	15 Mar 2022–31 Dec 2023	15,756
Correction stage, final results (Tables 5–6)	2012–2023, five zones	15 Jan–31 Dec 2023	8,394
Decomposition comparison (section 6, Fig. 5)	2012–2023 from hour 168, five zones	16 Jan–31 Dec 2023	8,380
Ouessant (Table 6)	2014–2023, five zones	15 Mar–31 Dec 2023	6,990
Unseen years (Table 6)	2024–2025, frozen pipeline	31 Dec 2023–31 Dec 2025	17,544

3.5. Errors Found During Revision

The same audit, repeated during the revision of this paper, found three further errors that did not involve leakage but did affect the reported figures. First, the archive had been queried without specifying the unit, so speeds were in km/h while being labelled m/s; every speed and error has been converted, and two thresholds that had been set in m/s but applied to km/h values (the 1 m/s mask of the percentage error and the wind-class boundaries of section 4.2) have been recomputed or are reported as they were actually applied. Second, the descriptive statistics had been computed on a two-year subset rather than on the twelve-year record used for training. Third, several comparisons combined results obtained on different test periods. None of them is visible in training curves or in a single results table: the unit, data set and evaluation period of every number need the same scrutiny as the code. Finally, the trained weights of the original runs could not be recovered, so the five-zone pipeline and the Ouessant experiment were re-run with identical code to report every metric of Tables 5 and 6 from a single set of predictions; the re-runs differ from the original runs by at most 0.008 m/s in mean absolute error.

4. Model Selection

4.1. Statistical Baseline

A baseline was established before any neural network, so that later improvements could be measured against something rather than presented in isolation. Two references are used: persistence, which forecasts the last observed value and is the standard reference at a one-hour horizon (skill is reported as the relative reduction of an error metric with respect to persistence), and an ARIMA model [6] whose order was selected automatically, the retained specification being ARIMA(1,1,2). The ARIMA is evaluated as a rolling one-step forecast: its parameters are estimated once on the training period and each forecast uses the observations up to $t-1$. A first evaluation as a single forecast of 2,625 steps issued at the end of the training period, which gave a coefficient of determination close to zero, is not used. The estimated first moving-average coefficient is close to -1, which indicates over-differencing.

Neither reference was engineered: the ARIMA has no seasonal term and no exogenous regressor. Both are nevertheless hard to beat at one hour ahead, because the series is strongly autocorrelated (section 2.3).

4.2. Architecture Comparison

Seven deep architectures were then trained on the same data, the same twenty variables and the same chronological split (the two-year data set of Table 3): a simple recurrent network, a long short-term memory network [5], a gated recurrent unit network [1], a one-dimensional convolutional network, two hybrids and an attention-based model. The input window was selected empirically between six and twenty-four hours; twenty hours gave the best trade-off, which is consistent with the daily cycle observed in the data, and the window was then frozen for every architecture so that the comparison remained fair. Table 4 gives the error metrics of the leading models and of the two references, and Fig. 4 the mean absolute error of the five single architectures and both references. The percentage error is not reported for this exploratory comparison, because it had been computed with a mask set in the wrong unit (section 3.5).

Table 4. Leading architectures compared with persistence and a rolling ARIMA on the two-year data set (test period 14 Sep–31 Dec 2023, 2,605 targets); MAE and RMSE in m/s

Model	MAE	RMSE	R^2
GRU	0.650	0.893	0.8436
LSTM	0.655	0.899	0.8415
CNN-1D	0.701	0.936	0.8281
Rolling ARIMA	0.645	0.899	0.8415
Persistence	0.654	0.935	0.8283

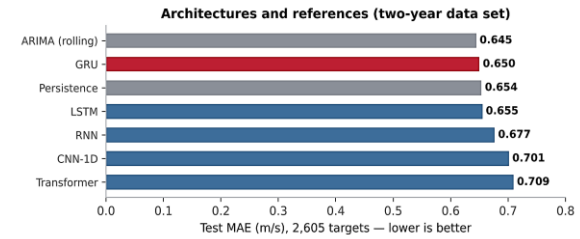


Fig. 4. Test mean absolute error of the five single architectures, the rolling ARIMA and persistence (two-year data set, Table 3)

The comparison with the references is sobering. The gated recurrent unit has the lowest root mean squared

In Press

error, but it is not significantly more accurate than the rolling ARIMA: a Diebold–Mariano test [22] with the small-sample correction of Harvey *et al.* [23] gives p = 0.52 for squared-error loss and p = 0.46 for absolute-error loss, the ARIMA having the lower mean absolute error. Against persistence the network is significantly better for squared-error loss ($p < 0.001$) but not for absolute-error loss (p = 0.67). On this short test period the choice of architecture therefore rests on the comparison between deep models; the gain of the full pipeline over persistence is established on longer periods in section 8.

Neither of the two hybrids improved on the gated recurrent unit alone. Feeding an ARIMA output into a recurrent network, an arrangement frequently proposed in the literature, propagates the bias of the linear model, and concatenating a recurrent and a convolutional branch adds about forty per cent more parameters for no gain, because both branches read the same inputs. This second result redirected the work towards correcting what the model gets wrong (section 5).

A final check split the test period into four speed classes: the ranking of the deep models is stable in every class. The class boundaries, intended as 5, 15 and 25 m/s, were applied to km/h values and therefore fell at 1.4, 4.2 and 6.9 m/s, so the check is a valid stratification but not one into calm, moderate, strong and storm conditions. The gated recurrent unit was retained because it is the best deep architecture on all three metrics and, with one gate fewer than the long short-term memory network, trains about thirty per cent faster, which made the search of the next subsection affordable.

4.3. Dataset Extension and Hyperparameter Search

At this point the results were not stable enough to build on: retraining the same model with a different random initialisation moved the metrics noticeably, which suggested that two years contain too few of the strongest-wind and near-calm hours. The dataset was therefore extended to the twelve years of section 2.2, and everything downstream was recomputed, including the scalers.

Six hyperparameters were then optimised with a Bayesian search using a tree-structured Parzen estimator [2], as implemented in Optuna [24]: hidden size, number of layers, dropout rate, learning rate, batch size and input window. The search converged in fifty trials, twenty-two of them pruned early. The retained configuration uses a hidden size of 256, two layers, a dropout rate of 0.059, a learning rate of 8.48e-4, a batch size of 64 and a window of twenty hours, for 608,513 trainable parameters; the learning rate dominated the importance analysis (0.61), ahead of dropout (0.19).

The training loss is an asymmetric squared error weighting under-prediction 1.5 times more than over-prediction. This was a design choice made early in the study, with a factor set a priori rather than tuned, and it is the opposite of the weighting that conventional dispatch suggests: since over-prediction is what can leave dispatchable generation and reserves short (section 1), a loss reflecting operational risk would penalise over-prediction instead. Its effect on the reported results is limited. In an early comparison on a 70/15/15 split it slightly increased the mean absolute error relative to a symmetric loss (0.585 against 0.579 m/s), and the positive bias it induces is removed by the correction stage of section 5, after which 49 % of the test hours are under-predicted. The tuned model reaches 0.577 m/s on the twelve-year 70/15/15 test period (Table 3), against 0.621 m/s for persistence, with a bias of +0.077 m/s; as the untuned network of Table 4 used the two-year data set, the two figures are not compared.

4.4. Hypotheses Tested and Rejected

Three hypotheses were tested and rejected on evidence. A monthly climate index of the El Nino-Southern Oscillation degrades the error from 0.577 to 0.585 m/s, because it is constant over a twenty-hour input window. Four variables from grid points 110 and 200 km up-monsoon leave the error practically unchanged (+0.001 m/s), because neighbouring grid points are outputs of the same physical model rather than independent measurements. A weighted resampling scheme favouring strong-wind hours degraded the results on two sites.

5. Residual Correction and Its Control

The tuned network leaves a structured error: a positive bias, largely induced by the asymmetric loss (+0.18 m/s on the five-zone test period), and residual patterns linked to the hour of the day. This is low-dimensional tabular structure, which is precisely what gradient-boosted decision trees handle better than a neural network [3]. A second stage was therefore added whose task is not to predict the wind but to predict the error of the first stage. It reads the twenty variables of hour $t-1$, the first-stage forecast and its change from the previous hour, so that no future information can reach it, and its number of trees is set by early stopping on its own zone of the split (70 trees in the run reported here).

Comparing the final pipeline with the reference of section 4.3 would suggest a gain of 0.036 m/s (from 0.577 to 0.542 m/s), but the two figures belong to different test periods (Table 3) and the comparison is not valid. Four configurations were therefore computed from the same first-stage predictions on the five-zone test period, to isolate the contribution of the correction stage; they are given in Table 5.

In press

Table 5. Contribution of the residual correction stage; all lines from one set of first-stage predictions on the five-zone test period (15 Jan–31 Dec 2023, 8,394 targets)

Configuration	MAE (m/s)	Bias (m/s)
GRU alone	0.5659	+0.1825
GRU with a constant bias correction	0.5448	+0.0187
GRU with boosting on shuffled residuals	0.5448	+0.0187
GRU with boosting on the real residuals	0.5415	+0.0139

The second and third lines are the important ones. A constant offset equal to the mean training residual brings the error from 0.566 to 0.545 m/s and the bias from +0.18 to +0.02 m/s, and a boosting model fitted on randomly shuffled residuals, that is on pure noise, gives exactly the same error, because all it can learn is the bias. Against this control, the structured contribution of the correction stage is 0.003 m/s, about 0.6 % of the error. A Diebold–Mariano test finds this difference significant on this test period ($p < 0.001$ for squared-error and $p = 0.002$ for absolute-error loss), but it is not stable: the original run of the same code gave 0.005 m/s, and on the unseen years of section 8 the constant offset is as accurate in 2024 (0.521 against 0.522 m/s) and the gap is 0.002 m/s over 2024–2025. Most of the gain of the stage is therefore a bias correction, which a constant offset achieves almost equally well. The stage was kept because it is computationally negligible and degrades no metric, but we do not claim that it captures structure in the error.

6. Causal Versus Naive Decomposition

The audit of section 3.2 identified decomposition before splitting as the most frequent weakness in this field, so both versions were implemented on identical data to measure the size of the effect.

In the naive version the whole series is decomposed by variational mode decomposition [20] (three modes, $\alpha = 2000$, $\tau = 0$, tolerance 10^{-7}) and the components are then split into the five zones, as in most of the decomposition studies reviewed. In the causal version the decomposition is recomputed at each time step using only the preceding 168 hours. Because the first 168 hours have no causal decomposition, they are removed from every variant before the zones are formed: all variants share the same zones (68,230, 10,480, 10,480, 7,330 and 8,380 sequences) and the same 8,380 test targets, from 16 January to 31 December 2023, and the identity of the target timestamps was checked programmatically. All variants use the same hyperparameters and random seed, at most 150 epochs with a patience of 20, and no learning-rate scheduler, which is why their errors differ slightly from Table 5. Fig. 5 places both decompositions against the pipeline without decomposition.

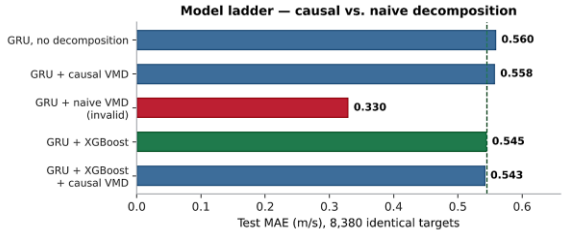


Fig. 5. Model ladder on 8,380 identical test targets (16 Jan–31 Dec 2023); the naive variant is invalid and shown for comparison only

The comparison that isolates the effect of leakage is between the two decomposition variants of the first-stage network, which differ only in the position of the decomposition relative to the split. The naive version reports 0.330 m/s and the causal version 0.558 m/s: decomposing the whole series lowers the apparent error by 0.228 m/s, or 41 %. The causal version itself brings no meaningful gain: 0.558 against 0.560 m/s for the first-stage network (−0.3 %), and 0.543 against 0.545 m/s once the correction stage is added (−0.4 %).

Methodologically, an improvement of about forty per cent reported for this family of techniques may be entirely due to the position of the decomposition relative to the split and cannot be read as forecasting skill. Practically, since the causal version brings nothing, the decomposition was rejected.

7. Multi-Horizon Forecasts and Prediction Intervals

The model was extended from one hour ahead to six using a direct strategy, with a separate network having one output per horizon rather than feeding predictions back into the model, which would accumulate error. This multi-horizon network has no correction stage and was trained and evaluated on the twelve-year 70/15/15 split (Table 3), so its one-hour error is not directly comparable with Table 6. Its mean absolute error grows from 0.605 m/s at one hour to 0.994 m/s at six hours. Persistence gives 0.621 m/s and 1.249 m/s on the same windows, so the skill of the network relative to persistence grows from 3 % at one hour to 20 % at six hours (Fig. 6).

A point forecast is of limited use to a microgrid operator on its own, because reserve sizing requires the plausible worst case rather than the average. Prediction intervals were therefore added at two nominal levels, 80 % and 95 %, with quantile models trained with a pinball loss and then calibrated by conformalized quantile regression [4]: a non-conformity score is computed on a held-out period and both bounds are shifted by a single offset. The held-out period is the validation zone of the 70/15/15 split, which also served for the early stopping of the quantile models, and the conformal guarantee assumes exchangeability, which an hourly wind series does not satisfy; the coverages below, measured on the test period with a re-run of the quantile models, are therefore empirical results. Compared at the

In press

same nominal level (Fig. 6), the calibration changes little: the mean coverage moves from 78.1 % to 78.6 % at 80 % and from 94.1 % to 94.0 % at 95 %, with offsets of at most 0.08 m/s. Neither level reaches its target, which points to a difference between the calibration period (2020–2022) and the test period (2022–2023) that a single offset cannot correct.

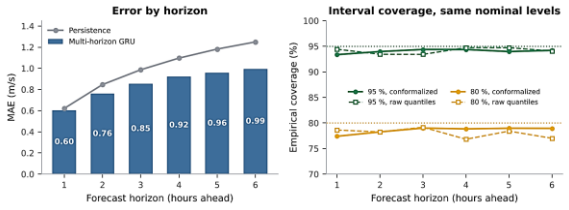


Fig. 6. Left: mean absolute error of the multi-horizon network and of persistence. Right: coverage of the raw quantile intervals (dashed) and of the conformalized intervals (solid) at the 80 % and 95 % nominal levels. Twelve-year 70/15/15 test period

The 95 % conformalized intervals are wide, from 2.93 m/s at one hour ahead to 4.77 m/s at six hours, that is between about half and nine-tenths of the mean wind speed. For microgrid operation the useful quantity would be the lower bound of the interval, which gives the wind speed that can be counted on with a stated confidence; translating it into a power that can be committed requires the turbine power curve and is left to future work.

8. Results and Transfer

The final pipeline combines the tuned gated recurrent unit with the residual correction stage. Table 6 gives its four metrics and those of persistence, each row computed from a single set of predictions: at Bac Lieu on the final zone of the five-zone split (2023, 8,394 targets) and on 2024–2025 (17,544 targets), and at Ouessant, France, where the same protocol was re-run (6,990 targets, 15 March–31 December 2023). For 2024–2025 the pipeline was frozen: the network weights, scalers and correction model are those of the 2023 evaluation, the inputs are built with the same code from the same IFS grid cell, and no decision was taken on these years.

Table 6. Final pipeline on each test period, followed by persistence on the same targets; all metrics of a row come from one set of predictions. MAE and RMSE in m/s, MAPE in per cent over hours above 1 m/s

Case	MAE	RMSE	MAPE	R^2
Bac Lieu 2023	0.542	0.753	12.83	0.8607
persistence	0.598	0.856	13.99	0.8196
Bac Lieu 2024–25	0.546	0.756	13.13	0.8799
persistence	0.589	0.836	14.12	0.8532
Ouessant 2023	0.623	0.950	8.40	0.9596
persistence	0.691	1.018	9.18	0.9536

The pipeline is significantly more accurate than persistence on every test period. Its root mean squared error is 12 % lower than that of persistence in 2023 and 10 % lower over 2024–2025, its mean absolute error 9 % and 7 % lower, and a Diebold–Mariano test rejects equal accuracy in each case for both squared and absolute losses ($p < 10^{-10}$). The frozen pipeline loses little on the unseen years (0.546 against 0.542 m/s), with a similar skill in 2024 (9 % in root mean squared error) and in 2025 (10 %), and it is also significantly more accurate than the rolling ARIMA, whose skill over persistence is only 2 %.

The Ouessant lines of Table 6 must not be read as a better or a worse model; there too the pipeline beats persistence (root mean squared error 7 % lower, $p < 10^{-10}$). Ouessant is markedly windier than Bac Lieu (9.36 m/s at 100 m over its record, against 5.37 m/s), so its absolute error is higher (0.623 against 0.542 m/s) while its relative error is lower: the mean absolute error is 6.4 % of the test-period mean speed at Ouessant (9.76 m/s) and 10.2 % at Bac Lieu (5.29 m/s). As at Bac Lieu, the archive serves ERA5 until 2016 and IFS from 2017 (cell 48.47° N, 5.14° W, elevation 23 m), so the test period lies in the IFS part. The model used the same protocol, five zones and hyperparameters, without a new search, but was trained from scratch on local data with a different set of twenty inputs (humidity and pressure added, different lags, a day-of-year encoding, no monsoon flag) and without a learning-rate scheduler. What the test establishes is therefore that the protocol, rather than a trained model, transfers to a site with a different climate; two sites are not enough to characterise its behaviour across climates.

At this site most hours lie between the cut-in and rated speeds of a typical turbine, where a speed error becomes a large relative power error. This paper therefore makes no claim about power-domain or operational performance; an evaluation weighted by the derivative of the power curve should precede any deployment.

9. Conclusion

This paper reported a short-term wind speed forecasting pipeline for a coastal site in southern Vietnam with a moderate wind regime, built entirely on public gridded data and therefore applicable to a candidate microgrid location without instrumentation. Eight model families were compared, the selected architecture was optimised, and a residual correction stage was added and tested against a control. On the 2023 test period the final mean absolute error is 0.542 m/s, with a coefficient of determination of 0.861; frozen and applied to 2024–2025, never used in its development, the pipeline reaches 0.546 m/s and remains about 10 % more accurate than persistence in root mean squared error. The same protocol, retrained on local data, also performs well at a windier site in a different climate.

The contribution we consider most useful is not the error figure but the procedure used to obtain it. Several

In Press

apparent improvements disappeared once leakage was removed, once a proper control was introduced, or once results were compared on a common test period. The statistical baseline is the clearest example: evaluated as a single long-range forecast it made the network look nearly three times more accurate, whereas evaluated as a rolling forecast it is as accurate as the network on the architecture comparison. Against a shuffled-residual control, the structured contribution of the correction stage is 0.003 m/s and is not stable across runs or years. The naive and causal versions of a widely used decomposition differ by 0.228 m/s on identical targets, an apparent gain of 41 % that is a measurement artefact rather than forecasting skill. The errors found during revision lead to the same conclusion: every number has to be tied to its unit, data set and evaluation period before it is compared with another.

Three directions extend the work: a power-domain evaluation weighted by the turbine power curve, including cut-in and cut-out events, which is required before operational use; the automation of the protocol, so that any wind series could be unit-checked, split causally, modelled and returned with its metrics, leakage checks and test period; and the use of independent buoy or mast measurements, to calibrate the gridded data and revisit the upstream information rejected in section 4.4.

Acknowledgment

The authors thank the members of the host laboratory for the discussion of intermediate results, and the teams who maintain the ECMWF ERA5 and IFS data and the Open-Meteo archive on which this work depends. The authors would like to express their sincere gratitude to Thac Mo Hydropower Joint Stock Company for their support during the research project titled "Application of Artificial Intelligence (AI) for Power Generation Forecasting at the Thac Mo Solar Power Plant".

References

[1] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, pp. 1724-1734, 2014. https://doi.org/10.3115/v1/D14-1179

[2] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, Algorithms for hyper-parameter optimization, in Advances in Neural Information Processing Systems 24, Granada, Spain, pp. 2546–2554, Dec. 2011.

[3] T. Chen and C. Guestrin, XGBoost: a scalable tree boosting system, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, San Francisco, CA, USA, pp. 785–794, Aug. 2016. https://doi.org/10.1145/2939672.2939785

[4] Y. Romano, E. Patterson, and E. J. Candès, Conformalized quantile regression, in Advances in Neural Information Processing Systems, Vancouver, Canada, vol. 32, pp. 3543–3553, Dec. 2019.

[5] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Computation, vol. 9, iss. 8, pp. 1735–1780, Nov. 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[6] G. Box, G. Jenkins, G. Reinsel, and G. Ljung, Time Series Analysis: Forecasting and Control, 5th ed. Hoboken, NJ, USA: Wiley, 2015.

[7] Y. Wang and L. Wu, On practical challenges of decomposition-based hybrid forecasting algorithms for wind speed and solar irradiation, Energy, vol. 112, pp. 208–220, Oct. 2016. https://doi.org/10.1016/j.energy.2016.06.075

[8] Z. Qian, Y. Pei, H. Zareipour, and N. Chen, A review and discussion of decomposition-based hybrid models for wind energy forecasting applications, Applied Energy, vol. 235, pp. 939–953, Feb. 2019. https://doi.org/10.1016/j.apenergy.2018.10.080

[9] K. Li, R. Shen, Z. Wang, B. Yan, Q. Yang, and X. Zhou, an efficient wind speed prediction method based on a deep neural network without future information leakage, Energy, vol. 267, Mar. 2023, Art. no. 126589. https://doi.org/10.1016/j.energy.2022.126589

[10] Open-Meteo, Historical Weather API. [Online]. Available: https://open-meteo.com/en/docs/historical-weather-api. Accessed on: Jun. 2, 2026.

[11] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, The ERA5 global reanalysis, Quarterly Jourrnal of the Royal Meteorological Society, vol. 146, iss. 730, pp. 1999–2049, Jul. 2020. https://doi.org/10.1002/qj.3803

[12] H. Zheng, Q. Wu, X. Lv, P. Zeng, X. Zhu, and Y. Zeng, Multi-scale wind speed prediction model based on improved escape algorithm for optimizing time-varying filtering empirical modal decomposition, Scientific Reports, vol. 16, Jan. 2026, Art. no. 4958. https://doi.org/10.1038/s41598-026-35505-6

[13] P. Dhaka, M. Sreejeth, and M. M. Tripathi, Wind power forecasting using multivariate signal decomposition and stacked GRU ensembles with error correction, Future Generation Computer Systems, vol. 175, Feb. 2026, Art. no. 108105. https://doi.org/10.1016/j.future.2025.108105

[14] X. Huang, Y. Zhang, J. Liu, X. Zhang, and S. Liu, A short-term wind power forecasting model based on 3D convolutional neural network–gated recurrent unit, Sustainability, vol. 15, iss. 19, Sep. 2023, Art. no. 14171. https://doi.org/10.3390/su151914171

In press

[15] Z. Zhao, S. Yun, L. Jia, J. Guo, Y. Meng, N. He, X. Li, J. Shi, and L. Yang, Hybrid VMD-CNN-GRU-based model for short-term forecasting of wind power considering spatio-temporal features, Engineering Applications of Artificial Intelligence, vol. 121, May 2023, Art. no. 105982. https://doi.org/10.1016/j.engappai.2023.105982

[16] L. Suo, T. Peng, S. Song, C. Zhang, Y. Wang, Y. Fu, and M. S. Nazir, Wind speed prediction by a swarm intelligence based deep learning model via signal decomposition and parameter optimization using improved chimp optimization algorithm, Energy, vol. 276, Aug. 2023, Art. no. 127526. https://doi.org/10.1016/j.energy.2023.127526

[17] M. Yaghoubirad, N. Azizi, M. Farajollahi, and A. Ahmadi, Deep learning-based multistep ahead wind speed and power generation forecasting using direct method, Energy Conversion and Management, vol. 281, Apr. 2023, Art. no. 116760. https://doi.org/10.1016/j.enconman.2023.116760

[18] J. Liu, C. Gong, S. Chen, and N. Zhou, Multi-step-ahead wind speed forecast method based on outlier correction, optimized decomposition, and DLinear model, Mathematics, vol. 11, iss. 12, Jun. 2023, Art. no. 2746. https://doi.org/10.3390/math11122746

[19] W. Jiang, B. Liu, Y. Liang, H. Gao, P. Lin. D. Zhang, and G. Hu, Applicability analysis of transformer to wind speed forecasting by a novel deep learning framework with multiple atmospheric variables, Applied Energy, vol. 353, Jan. 2024, Art. no. 122155. https://doi.org/10.1016/j.apenergy.2023.122155

[20] K. Dragomiretskiy and D. Zosso, Variational mode decomposition, IEEE Transaction on Signal Processing, vol. 62, iss. 3, pp. 531–544, Feb. 2014. https://doi.org/10.1109/TSP.2013.2288675

[21] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, Leakage in data mining: formulation, detection, and avoidance, ACM Transactions on Knowledge Discovery from Data, vol. 6, iss. 4, Dec. 2012, Art. no. 15. https://doi.org/10.1145/2382577.2382579

[22] F. X. Diebold and R. S. Mariano, Comparing predictive accuracy, Journal of Business & Economic Statistics, vol. 13, iss. 3, pp. 253–263, Jul. 1995. https://doi.org/10.1080/07350015.1995.10524599

[23] D. Harvey, S. Leybourne, and P. Newbold, Testing the equality of prediction mean squared errors, International Journal of Forecasting, vol. 13, iss. 2, pp. 281–291, Jun. 1997. https://doi.org/10.1016/S0169-2070(96)00719-4

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, pp. 2623–2631, Jul. 2019. https://doi.org/10.1145/3292500.3330701

In press