Reply to RC1: reanalysis and planned revisions

Author comment on wes-2026-154, https://doi.org/10.5194/wes-2026-154

We thank the referee for a careful report and for working through the archived data and code. We agree with the four main concerns. We have reproduced every published table and figure from the public code and data; the published numbers follow correctly from the code, but the new analyses below confirm each concern. They also found a further error in the chronological splits (point 4), and an analysis over five calendar test years changes the main empirical conclusion of the preprint (see "Revised conclusion"). Unless stated otherwise, numbers refer to the 2018 data used in the preprint.

1. Operating states and the use of one year

The 71 zero-output hours at Kelmarsh WT1 are explained by the status log. A manual on-site stop (Scheduled Maintenance) at 13:32 UTC on 6 March 2018 was followed by a "Battery test" stop (Technical Standby) lasting 70.1 hours, until 12:53 on 9 March. The raw ten-minute power values in these hours are slightly negative and were clipped to zero by our preprocessing.

This was not an isolated case. In 2018, every hour with zero output and wind speed above 5 m/s coincides with a logged stop: 62 of 62 hours at Kelmarsh and 19 of 19 at Penmanshiel. Stops touch 2.8 % of the pre-test hours at Kelmarsh and 2.0 % at Penmanshiel. The referee's count of 20 to 146 hours per retained bin above 10 m/s matches Penmanshiel exactly. Removing these hours from the curve and scatter lowers the 1-h benchmark (as in Table 2) from 10.65 % to 9.68 % of rated power at Kelmarsh and from 11.61 % to 11.21 % at Penmanshiel. For $h \geq 12$ h the change is below 0.2 percentage points (pp).

The check also exposed two preprocessing faults. At Penmanshiel, every hour with a missing value in any column was removed, including nacelle temperature, which the analysis does not use. Its sensor was missing for most of February to May 2018, so 2,317 hours with complete wind and power records (26 % of the year) were discarded. Of these, 1,910 have positive power and only 177 overlap a logged stop, so the gap was not a maintenance period. Restoring them without a state filter raises the 1-h benchmark to 14.25 %, because these months include a 134-h icing stop; with logged stops removed it is 11.01 %. In addition, the two sites clipped negative power at different stages; the revised analysis averages first and clips negative hourly means to zero at both sites.

A benchmark built from all operating states therefore depends on how many outage hours fall in the fitting period, and its curve is a statistical wind-to-power mapping for the processed power, not an aerodynamic power curve. The revision will state the forecasting target explicitly. The main target will be power in hours without a recorded stop: the curve and scatter are fitted without hours that overlap a logged stop, and only target hours without such an overlap are evaluated. Power limitations logged as warnings, without a stop, are not excluded. The evaluation is therefore conditional on the status recorded at the target hour. An unplanned stop is not necessarily known when the forecast is issued, so these scores do not describe performance under all operating conditions; the analysis of all operating

states covers that case. The AR wind model is fitted on all valid hours, because wind speed is recorded during stops as well. All operating states will be a secondary analysis. Both sites will use the same preprocessing.

We used 2018 alone to keep the demonstration small, and point 3b(i) shows that this was not a sound choice. The revised analysis uses 2016–2023 for both turbines from the current Zenodo releases (Penmanshiel WT01 has no 2024 data, and its 2016 record starts in June).

2. Figure 1 and the Taylor comparison

The referee is right. The parabola in Fig. 1 uses finite-difference derivatives of the piecewise-linear curve at one grid point. At Kelmarsh this point lies 0.02–0.03 m/s from a curve knot, and the estimated curvature is 725, 424 and 0 kW $(\mathrm{m\,s^{-1}})^{-2}$ on grids of 400, 800 (as published) and 1,600 or more points. At Penmanshiel the point lies inside a linear segment, so the dashed curve in the figure is the tangent line.

The percentages in the text come from a different calculation and contain a unit error. The 22–61 % in the abstract, Sect. 3.4 and the conclusions are relative errors in the standard deviation of the curve-variation term, not in its variance. On the variance scale they are 48–70 % at Kelmarsh and 109–158 % at Penmanshiel.

We repeated the comparison with two smooth curves fitted to the same bins, a four-parameter logistic function and a cubic smoothing spline, using analytic derivatives, quadrature of the same function as reference, and an average over all pre-test AR states. The quadrature has converged (less than 0.007 % change between 128 and 256 nodes). The second-order expression in Eq. (4) is the variance of the quadratic Taylor polynomial; it is not the full σ^4-order expansion of the variance, which also contains a term in $m'm'''\sigma^4$, and the revision will describe it that way. Without logged stop hours, this expression overestimates the standard deviation of the curve-variation term as follows:

	h = 1 h	h = 48 h	h = 48 h, variance scale
Kelmarsh	3.3–3.4 %	31–38 %	73–90 %
Penmanshiel	5.4–5.5 %	61–77 %	161–213 %

The overestimate grows with horizon for both curves at both sites. The revision will redraw Fig. 1 with the smooth curve, add the Taylor error by horizon, and give the result as a range over both curves with the unit stated.

3a. Length of Section 4

Section 4 is too short. The revised section will describe Table 2 and Figs. 1 and 2: how curve variation and scatter change with horizon at each site, the plug-in penalty, the matched comparison of point 3b(ii), and the multi-year and seasonal results. We will also remove "theoretical minimum RMSE" (Sect. 4.1) and the statement that NWP is the "mathematically mandated path" (Sect. 5.2), which

conflict with the model-conditional definition in Sect. 3.1. Two statements in Sect. 2 are also inaccurate and will be corrected: the bin width is 0.36 m/s at Kelmarsh and 0.51 m/s at Penmanshiel (not 0.33 m/s), and the bin means are interpolated directly, not onto a 2,000-point grid.

3b(i). Training and test seasons

We agree. With 2018 alone, the forecasters are trained on data up to late September or early October and tested on 20 November to 31 December, and the benchmark is fitted on 1 January to 20 November.

Before opening the additional years, we fixed a protocol in writing. Each test year Y from 2019 to 2023 is a calendar year; forecasters are trained on 2016 to Y−2, validated on Y−1 and evaluated on Y. Coverage is incomplete in some months, and the Penmanshiel 2023 record starts on 1 February. The benchmark is fitted on all hours known before the first test forecast is issued and evaluated on the same test origins. Persistence through the curve, linear regression and the MLP are evaluated for all years and horizons; the other four networks will be evaluated for 2019. We also added the model's own conditional-mean forecast, $E[m(V_{t+h}) \mid F_t]$, where F_t is the wind history up to the forecast origin t. If the model is calibrated, the error of this forecast should match the benchmark. In each test year, the method with the lowest validation error is selected, and the results for all years together pool the methods selected year by year. Intervals come from a moving-block bootstrap on the full test calendar (48-h blocks, 1,000 resamples) that also refits the benchmark. All numbers use the harmonized preprocessing of point 1. Some steps were changed after the first multi-year results: the harmonized power rule, selection within each test year instead of on the mean of the five validation years, and bootstrap blocks that follow the calendar. Each change is recorded in the protocol with its date and reason, and the protocol will be released with the code. Together these changes move the pooled gaps of the selected method by at most 0.16 pp.

The benchmark is stable across years: for the target without recorded stops, the 1-h value lies between 9.7 % and 10.6 % at Kelmarsh and between 11.0 % and 11.6 % at Penmanshiel, and the 48-h value between 28.2 % and 29.1 % and between 32.9 % and 33.4 %. With more data, BIC selects AR(4) or AR(5) at Penmanshiel.

For the five test years together, the method selected on validation lies between −0.90 and +0.31 pp from the benchmark across the 18 site–horizon cases (gap of the selected method and 95 % interval, pp):

	h = 1 h	**h = 6 h**	**h = 24 h**	**h = 48 h**
Kelmarsh	−0.05 [−0.23, 0.14]	+0.19 [−0.28, 0.63]	−0.31 [−1.01, 0.46]	+0.31 [−0.55, 1.16]
Penmanshiel	−0.22 [−0.54, 0.10]	−0.19 [−0.70, 0.34]	−0.90 [−1.59, −0.20]	−0.70 [−1.43, 0.01]

No interval lies above zero. Three lie below, all at Penmanshiel (18, 24 and 36 h), where the method selected on validation is linear regression in four of the five test years; at 48 h the interval of the selected method reaches +0.01 pp. Linear regression evaluated on its own, which we single out after

seeing the test results, lies 0.76–0.94 pp below the reference at Penmanshiel from 18 to 48 h, and all four of its intervals exclude zero. The model's own forecast lies between −0.42 and +0.92 pp from the benchmark and within about ±0.4 pp up to 24 h; only one of its intervals excludes zero (Kelmarsh, 48 h: +0.92 [0.11, 1.65] pp). For this comparison only the test hours are resampled, because the forecast and the reference come from the same fit. These intervals therefore leave out the uncertainty of that fit. Among the individual methods, linear regression on the last hourly wind speed or the model's own forecast has the lowest test error in 17 of the 18 cases, and the MLP in the remaining one (Penmanshiel, 6 h). This count describes the test results only: it plays no part in the selection and does not include the year-by-year selection as a method.

Single years and seasons differ. For one test year at a time, the 48-h gap ranges from −2.4 to +3.1 pp, positive in 2020 and mostly negative otherwise. With one AR model for the whole year, the model's own forecast exceeds the benchmark in winter by 3.8–6.7 pp at 48 h and falls below it in summer by 2.4–4.2 pp. A season-dependent AR model reduces the winter excess to 1.3–3.1 pp but overcorrects in summer at Penmanshiel. The preprint's test window lies mostly in winter.

3b(ii). Benchmark and forecast errors from different periods

Tables 2 and 3 do compare different periods. We recomputed the benchmark at the same 1,000 test origins as the archived forecasts, with the published fit, and rescored all seven methods from the archived files (all 126 published scores reproduced to within 1.4×10^{-7} pp). At Kelmarsh the test-origin benchmark is higher than the pre-test average at every horizon (11.61 % against 10.65 % at 1 h); at Penmanshiel the two differ by −0.12 to +0.20 pp.

On matched origins the gaps for the best method range from −0.07 to +4.17 pp, compared with 0.89–4.52 pp in the preprint; at Kelmarsh 1 h the best method lies below the benchmark. With 95 % moving-block bootstrap intervals, only one of the 18 gaps excludes zero when only the test sample is resampled, and none when the benchmark is also refitted. Six weeks of hourly data hold only about 21 non-overlapping 48-h blocks, and the errors are autocorrelated over about a day, so the effective sample is small.

The model's own forecast helps to explain the long-horizon gaps. If the model were calibrated for the test window, its error would be close to the benchmark. At 48 h it is 32.84 % at Kelmarsh against a benchmark of 28.51 %, almost the same as the best method in Table 3 (32.68 %), with a mean error of +149 kW; at Penmanshiel it is 35.35 % against 33.12 %. The test weeks were windier than the training mean towards which the AR forecast reverts. These results are consistent with a wind model miscalibrated for this window contributing to the long-horizon gaps in Table 3.

The abstract sentence stating that every method lies above its benchmark in all eighteen combinations is therefore not supported, and we will remove it. The best methods in Table 3 were also selected on the test set. The revision will compare benchmark and forecasts on identical origins, select methods on validation data, include benchmark-fitting uncertainty, and use the multi-year test of point 3b(i).

4. A further error: chronological boundaries for h > 1

In the data, the hourly value labelled t is the mean of the ten-minute records labelled t to t + 50 min. Assuming that these labels mark the start of each ten-minute interval and that values are available without delay, an hourly value is known at t + 1 h; we have not verified the telemetry latency. This convention affects the exact count below, not the existence of the problem. No test target enters training, validation or the benchmark fit. However, samples are split by forecast origin, so for h > 1 the last h − 1 training targets fall after the first validation forecast is issued, and the last h − 1 validation targets, used for early stopping, fall after the first test forecast is issued. The benchmark fit also ends h − 1 hours after the first test forecast. The statements in Sect. 2 and Appendix A1 that the setup prevents leakage are therefore wrong for h > 1.

Removing the affected samples and refitting changes persistence by −0.15 to +0.17 pp, linear regression by −0.02 to +0.11 pp and the Table 2 benchmark by −0.02 to +0.05 pp. We retrained the five networks at 48 h with ten seeds on the same machine, with original and purged splits. With original splits they reproduce the published scores within 0.13 pp; purging changes their error by +0.33 to +0.60 pp at Kelmarsh and by −0.07 to +0.56 pp at Penmanshiel. The revision will state the purging rule in the methods and retrain all networks under it.

Revised conclusion

These analyses change the main empirical claim of the preprint. Over five calendar test years, the wind-only methods selected on validation lie within about ±1 pp of the model-conditional benchmark at every horizon, and the pooled RMSE of the model's own forecast differs from the fitted reference by approximately ±0.4 pp up to 24 h. The gaps of 0.9–4.5 pp reported in the preprint were measured on a six-week winter test window, and the seasonal results are consistent with a wind model that ignores season contributing to them. The fitted reference is also not a universal lower bound on forecast errors for the physical wind process: at Penmanshiel, the method selected on validation, mostly linear regression, falls 0.70–0.90 pp below it at horizons of 18–48 h, with intervals that exclude zero at 18, 24 and 36 h.

The revised manuscript will present the benchmark as a reference conditional on the model assumptions. It is the RMSE that the model's own forecast would have if hourly wind speed followed the fitted Gaussian AR model, power scattered around the binned curve with a variance that depends only on the wind speed in the target hour, and the wind history were the only information. Over full years it is close to the achieved wind-only error, but it is miscalibrated by season. All lower-bound language will be removed. Reaching this reference does not show that the wind history has been fully exploited. The revision will name richer wind-only models, for example with seasonal or regime-dependent dynamics, and numerical weather predictions as directions to test, not as results of this study. The Taylor result of point 2 and the variance decomposition are not affected.

Planned revision

- A revised central claim and the removal of all lower-bound language.

- An explicit forecasting target, with hours without a recorded stop as the main target; the same preprocessing at both sites (point 1).
- Data from 2016–2023, with calendar-year test periods from 2019 to 2023 and results by season (points 1 and 3b(i)).
- Purged chronological splits, retrained networks and model selection on validation data (point 4).
- Benchmark and forecasts on identical origins, with intervals that include benchmark-fitting uncertainty (point 3b(ii)).
- A new Fig. 1 and a Taylor comparison on smooth curves, with the unit corrected (point 2).
- A rewritten Sect. 4, with revised abstract, discussion and conclusions (point 3a).

The referee's checks were possible because the processed data and code were public, and the revised analysis will remain fully public. The current archive starts from processed hourly files. The revised archive will also contain the processing of the raw SCADA and status-log files, the forecasts for every test origin behind each table, and the scripts for all analyses in this reply, released as a new versioned Zenodo record and cited in the revised manuscript.

We recognise that this is a substantial reanalysis and would welcome the editor's view on whether it fits within the present review.

Blerant Ramadani and Vangel Fustic