Response to Reviewers

Manuscript WES-2026-75: *Methods for high-accuracy wind resource assessment to support distributed wind turbine siting*

Summary of the revision. The revision (i) adds height-stratified and terrain-stratified validation, (ii) replaces the "uncorrelated errors" framing with measured cross-dataset error correlations, (iii) reframes the interannual-variability argument against model error and adds the requested sensitivity experiments, (iv) adds a single-source ablation demonstrating that the improvement comes from fusion rather than from bias correction, and (v) substantially shortens and consolidates the manuscript. The body text is approximately 17 % shorter than the discussion paper despite these additions, the figure count is reduced from 25 to 14 plus one full-width figure, and two tables are removed. In several places the added analyses changed our reading of the results; we have followed the data rather than our original interpretation, and those cases are noted below. Three requests we have not implemented as asked: in each we performed the underlying analysis and report its results, but present them differently from the form requested, with the reasoning given in full. Line, section, and figure references are to the discussion-paper version unless stated otherwise.

---

1 Referee #1 (RC1)

We thank Referee 1 for a detailed and technically substantive review. The central request, an explicit demonstration that the method improves *hub-height* estimates rather than only near-surface winds, is well taken, and the revision addresses it directly with height-stratified metrics and a terrain-stratified analysis.

RC1.G: General framing (10 m vs. hub height; point vs. grid-cell representativeness)

> *A central methodological issue is the relationship between 10 m wind and hub-height wind-resource assessment. Ten-meter winds observationally are strongly influenced by local exposure, roughness, obstacles. In models such as WRF, 10 m wind is a diagnostic quantity and may behave differently from winds on other vertical levels. Therefore, good performance at 10 m does not necessarily imply good performance at hub height. Because this paper combines ASOS 10 m observations with GS measurements, it needs to demonstrate explicitly that the method improves hub-height estimates, not merely near-surface winds.*
> *From a different angle, I would like to reflect on the relationship between observation that is a point-like measurement and model output that represents the average wind speed over a grid-cell. This distinction is less important in flat terrain where the wind speed is spatially similar. However, this is very important when talking about complex terrain or coastal regions. In addition, measurements are not randomly distributed across grid cell. People carrying out measurements will put the met masts on top of hills and not in the valleys.*

Response. We agree on the first point, and the validation is now structured so that hub-height skill is demonstrated separately from near-surface skill (RC1.M1).

On the second, we agree that the point-versus-grid-cell distinction matters in complex terrain, and we now address it empirically by stratifying errors by terrain complexity (RC1.M6, RC3.M6). The referee's specific hypothesis, that masts are preferentially sited on high ground rather than in

valleys, we were able to test directly, and it is not borne out in this network. Comparing each GS site's true elevation against the mean elevation of its ERA5 cell, computed from a 278 m elevation grid, the median offset is −1.4 m and only 39.5 % of masts sit above their cell mean, slightly less often than chance. Restricting to the 170 cells that are entirely land, in case coastal cells distort the comparison, gives a median of −0.8 m and 47.1 %. At a finer 2 km scale the offset turns positive but is negligible in magnitude (+0.3 m). Only in complex terrain does a suggestion of the expected pattern appear, at fine scale (+11.8 m, 70 % of sites above their surroundings), and with 30 sites and $p = 0.058$ we regard that as underpowered rather than established.

The mechanism the referee describes is nonetheless real, and it discriminates between products. Sites sitting further above their cell mean are more strongly underestimated, as expected, but ERA5 is more than twice as sensitive to this offset as WEM (Spearman $\rho = -0.32$ versus -0.14; difference in $|\rho|$ of 0.175, bootstrap 95 % CI $[0.071, 0.277]$, with ERA5 more sensitive in every one of 5,000 resamples). Whatever representativeness penalty siting imposes, the ensemble absorbs most of it rather than inheriting it.

RC1.M1: Height-stratified validation, especially WEM vs. GWA

> *My main concern is that the final verification is performed on the full GS dataset without sufficient stratification by measurement height. From Fig. 3, a large fraction of the GS measurements appears to be below typical wind-energy hub heights, and in my estimation 85% are below 50 m. Therefore, the aggregate performance metrics may be dominated by lower-height measurements. This is problematic because the main application is wind-energy assessment at hub height. For example, Fig. 17 indicates that WEM improves over GWA in approximately 65 % of cases, but without height-stratified metrics it is impossible to know whether this improvement occurs at hub-relevant heights or primarily at lower levels. The authors should report WEM versus GWA performance separately for height bins, for example <20 m, 20–50 m, 50–80 m, and >80 m.*

Response. We confirm the referee's estimate (85.6 % of GS site–heights fall below 50 m; 31.2 % below 20 m) and now report height-stratified results. WEM significantly outperforms GWA both overall and, critically, *within the hub-relevant 50–80 m band* (mean |bias| 0.33 vs. 0.65 m s^{-1}, $p = 2 \times 10^{-5}$, $n = 44$, with WEM better at 73 % of site–heights). The improvement is therefore not an artifact of low-height measurements.

Above 80 m the ranking reverses (GWA 0.25 vs. WEM 0.52 m s^{-1}), and we now treat this openly rather than setting it aside on sample size alone. That bin holds only nine site–heights: too few for a firm statistical comparison, but more importantly too few for the model to *learn* from. The height-balancing experiment requested in RC1.M2 supports this reading directly: up-weighting the scarce tall-tower rows lowers the >80 m error from 0.52 to 0.34 m s^{-1}, most of the way to GWA. The revision therefore states that WEM's tall-tower weakness is largely an artifact of training-data scarcity rather than a limitation of the approach, and that additional tall-tower observations, not a change of method, are what would close it.

We would add one point about scope. This is a paper about *distributed* wind, whose hub heights fall largely within the 20 to 80 m band rather than the 100 m and above of utility-scale machines. WEM attains the lowest error of any product throughout that band, and it is above it that the >80 m bin sits. We do not offer this as a reason to discount the referee's concern, which is why we report the reversal plainly and explain its cause, but it does bear on how much weight the tall-tower result should carry for the paper's claims.

Manuscript changes.

- Height-stratified WEM-versus-GWA results reported in the ML results section, including the 50–80 m significance test and the >80 m reversal with its training-scarcity explanation.
- The site-level $\Delta|\text{bias}|$ figure is now faceted by height bin with a summary panel giving WEM's win rate against each baseline per bin; win rates remain 65–80 % through the hub-relevant band.
- A second figure resolves bias jointly in height and wind speed (RC3.M4), so height dependence is addressed by two complementary figures.
- Exact height-bin counts added where the GS archive is introduced (also answers RC3.m9).
- [We did not add a separate height-stratified table.] The stratified values are reported in the text and the two figures above display the height dependence directly; a table duplicated them and worked against RC2's length request. We are glad to add one if the referee prefers a tabular presentation.

RC1.M2: Why not balance GS by height, as ASOS/GS are balanced?

> *To reiterate, the question of how 10m winds can be used to improve wind assessment at hub heights is quite a complicated one. However, because the GS dataset bins together a lot of measurements close to 10 m (one third of measurements is below 20m), the methodology used completely sidesteps this question. Authors argue that GS and ASOS data should be balanced (Section 4.2.3). Why shouldn't the GS dataset be similarly balanced between low heights and high heights? It is quite easy to improve the 10 m wind speeds in wind energy datasets, because 10 m wind speeds are of limited direct relevance for wind-energy applications, they are provided mostly for reference.*

Response. We ran exactly this experiment, and the result is more interesting than a null. Equalising GS training weight across the four height bins leaves overall performance unchanged ($\Delta = +0.000\,\mathrm{m\,s^{-1}}$), but it is not inert: it trades a substantial gain in the data-poor >80 m bin (from 0.518 to $0.336\,\mathrm{m\,s^{-1}}$) against small degradations in the better-sampled bins.

We retain the production configuration, but not because balancing "does not work." We decline it because the >80 m bin contains nine site–heights, too few to establish that the gain is systematic rather than sampling noise, and we do not think it sound to tune a model to a bin that small. The experiment is nonetheless informative and we report it as such: it indicates that WEM's weaker tall-tower performance is partly a consequence of the training-set height distribution rather than of the method. We are grateful for the suggestion, which changed our understanding of the tall-tower behavior.

Manuscript changes.

- Height-balancing experiment reported in an appendix with a per-bin table.
- Methods clarifies the rationale distinguishing cross-archive (ASOS/GS) balancing from within-GS height balancing.
- The tall-tower interpretation is carried into the RC1.M1 discussion of the >80 m bin.

RC1.M3: "Uncorrelated error structure" / model independence

> *Another important issue that is present throughout the paper is the idea that combination of different model datasets is useful because the datasets contain independent information ("uncorrelated error structure"). It is not obvious to me, that using different equations should lead to uncorrelated error structures. In my experience, mesoscale models inherit biases from global reanalysis. As for microscale models, they*

explicitly use model chain. GWA is based on ERA5 and then downscaled using WRF. If authors want to stress the model independence, they should show it explicitly, using the data available. The results already shown in the paper suggest quite the opposite: Figure 3 shows that adding more models decreases the performance metrics.

Response. The referee is correct on both counts and we have changed the paper accordingly. The "uncorrelated" characterization is removed: the products are demonstrably *not* independent, and we now state the shared lineage explicitly (GWA is an ERA5-based downscaling; the mesoscale products inherit reanalysis structure). We report the measured cross-dataset error correlations (mean pairwise $r = 0.60$, rising to 0.65–0.91 within the NWP block, with GWA only 0.21–0.31 against them) and reframe the benefit as complementarity rather than independence: the leading principal component accounts for 70 % of the variance, leaving roughly 30 % of the structure off the dominant shared axis.

On the referee's second observation, the sweep confirms the pattern, and we have made it explicit. The production model already used a selected subset of wind products rather than all of them, chosen by the feature sweep, but the discussion paper did not state clearly why. The revision now does: the best-performing configurations are compact subsets, and the lowest error is reached by a three-dataset combination rather than by adding every product.

Manuscript changes.

- "Uncorrelated error structures" and equivalent independence phrasing replaced with "partially correlated but complementary" throughout; the model chain stated explicitly in the Background.
- New cross-dataset error-correlation heatmap at GS sites; correlation and PCA statistics reported in the benchmark results.
- The 30 % residual structure is no longer asserted to be "potentially learnable." It is now tied to the ablation (RC3.M5), which demonstrates empirically that a substantial part of it is genuinely learnable rather than noise.
- The feature-selection text states that the lowest-error configuration is a compact subset rather than the full combination.

RC1.M4: Interannual variability and GS record lengths

Similarly, I think that the reasoning for omitting interannual variability is not well established (Section 3.5). First of all, the time series length of GS dataset is not reported, even in summarized form. In my experience, tall mast campaigns tend to be quite short and I am not sure if I would trust annual variability estimate from two or three years of data, because of questions how representative the data are, and maybe there is autocorrelation in the data. I find Figure 4 unconvincing. Only the first 50 stations are of interest, because the measurements there are over 50 m. More importantly, the spatial distribution is quite irrelevant here because nobody is arguing that whole CONUS has the same wind speed. To show improvement, interannual variability should be compared to the difference between models and observations or a similar metric.

Response. We have reframed the argument exactly as the referee suggests, against the model error itself. The sampling uncertainty of the observed long-term mean is small (median standard error $0.06\,\mathrm{m\,s^{-1}}$) relative to the biases being corrected (median $0.48\,\mathrm{m\,s^{-1}}$for the strongest baseline), and model bias exceeds the sampling standard error at 86 % of GS sites (85–89 % across datasets). We report GS and ASOS record-length statistics and the completeness policy, and reconcile the argument with L22–24 by distinguishing systematic bias, which recurs every year and propagates

into energy estimates, from interannual fluctuation, which is a sampling effect an order of magnitude smaller.

The referee also raises autocorrelation, which bears directly on that standard error, since $\mathrm{SE} = \mathrm{sd}/\sqrt{n_{\mathrm{years}}}$ assumes years are independent. We have tested it. Across the 242 GS site–heights with at least six valid years, computing lag-1 correlation only across genuinely adjacent years, the median coefficient is +0.08, and correcting the standard error for the implied loss of effective sample size raises its median from 0.058 to $0.063\,\mathrm{m\,s^{-1}}$. Applying in addition the standard small-sample correction for the low bias of ρ at short records, which we treat as a conservative bound rather than an estimate, gives +0.23 and a corrected standard error of $0.074\,\mathrm{m\,s^{-1}}$; even then the sampling uncertainty is roughly six times smaller than the median model bias. The result is not an artifact of short records, since restricting to the 104 site–heights with ten or more valid years gives +0.085.

Manuscript changes.

- Valid-year distributions for GS and ASOS and the minimum-year policy reported; all quality-control criteria consolidated into a single paragraph placed after both archives are introduced.
- Interannual section rewritten: the "as informative as any subset" claim removed and the bias-versus-standard-error comparison added. The section now also states that the standard error treats years as independent, reports the measured lag-1 correlation, and gives the corrected standard error.
- Year-threshold ($\geq$2/3/5 yr) and restricted-window sensitivity results added (summary in text, detail in appendix).
- Explicit reconciliation with L22–24 added.

RC1.M5: L204 quantile mean vs. arithmetic mean; "smooth"

> *L204: "The empirical quantile representation also provides a consistent way to recover long-term mean wind speed for each site and height." For a complete empirical sample, the mean computed from the empirical distribution should be equivalent, up to numerical approximation, to the arithmetic mean of the time series. If there are significant differences, I would expect an error in programming code. The word "smooth" is also unclear in this context, since the result is a single scalar mean for one site and height. I am not sure if spatial smoothness is a desirable result for this application.*

Response. The referee is correct, and we verified the equivalence: the quantile-integrated mean matches the arithmetic mean to a median of $0.027\,\mathrm{m\,s^{-1}}$, consistent with numerical integration error rather than a defect. The sentence now gives the real motivation, namely a single mean estimator applied uniformly across all sources, including products delivered only as means or quantiles, and "smooth" is deleted.

Manuscript changes. "Smooth" is removed and the sentence reworded to the cross-source-consistency rationale.

RC1.M6: Is WEM useful? WEM vs. GWA aloft; ERA5-grid resolution vs. 250 m GWA

> *I have provided a number of theoretical objections to the methodology used. However, "all models are wrong, but some of them are useful". Is the WEM dataset useful? I would argue that it is not established in the paper because of GS binning over heights (Point (1)). In addition, in Figure 18 WEM looks indistinguishable from GWA. If I had to construct the worst possible interpretation based on Points (1)*

and (2), I would say that WEM is basically the same as GWA in upper levels with all improvement over GWA (Figure 17) concentrated on lower levels. Authors could persuade me otherwise by changing Figure 17 to height-stratified metrics and showing how the WEM datasets differs from GWA in upper levels. On top of that, WEM is provided on ERA5 grid (30 km) which is a significant "upscaling", i.e., lowering the resolution with respect to 250m of GWA. I am not sure that the increase in WEM quality (that remains to be shown for higher levels) offsets the representativeness error in complex terrain introduced by this drastic loss of resolution.

Response. The height-stratified results (RC1.M1) address the first concern directly: WEM significantly improves on GWA within the hub-relevant 50–80 m band (0.33 vs. 0.65 $\mathrm{m\,s^{-1}}$, $p = 2 \times 10^{-5}$), so the improvement is not confined to low levels. As noted under RC1.M1, that band is where distributed-wind hub heights sit, so the worst-case reading that WEM merely reproduces GWA does not hold in the regime the paper targets.

On resolution, a distinction the discussion paper left implicit is now explicit. WEM is a point-evaluable function: it maps features at a location (driving-dataset quantiles, the GWA climatology, terrain descriptors, coordinates, and height) to a quantile curve, and nothing ties those features to ERA5 cell centers. The ∼31 km grid is where we currently evaluate the model for the public product, not a property of the model.

This matters for reading our validation. The terrain-stratified analysis evaluates WEM *at station locations*, with features sampled at those points, so it measures the model's skill at an arbitrary point, which is what a developer needs. On that measure WEM has lower mean |bias| than GWA in every terrain class, with its largest margin in the most complex terrain (0.57 vs. 0.91 $\mathrm{m\,s^{-1}}$), and it degrades far less steeply than ERA5 as terrain roughens (ERA5 0.79 → 1.51 $\mathrm{m\,s^{-1}}$, Spearman $r = 0.26$, $p < 10^{-5}$; WEM ≈0.40 → 0.57 $\mathrm{m\,s^{-1}}$, with no significant monotonic trend, $r = 0.04$, $p = 0.43$).

We agree with the referee about the served product, and say so. A field rendered on the ERA5 grid *is* limited in complex terrain, because a 31 km cell cannot resolve ridge- and valley-scale structure that GWA captures at 250 m; for micrositing in rugged terrain the higher-resolution product remains complementary and we do not suggest otherwise. But that limitation belongs to the present deployment, not to the method. Because the features can be sampled anywhere, the same trained model can be evaluated on an arbitrarily fine mesh, bounded in practice only by the finest input field, and the 3DEP terrain data are finer still than GWA. Running the same trained model's inference at higher resolution is therefore a natural extension requiring no retraining, and we identify it as such. We also state plainly that the current product is a screening-scale tool rather than a substitute for microscale modeling or on-site measurement where sub-grid terrain dominates.

Manuscript changes. A terrain-stratified |bias| figure is added (shared with RC3.M6) and the resolution discussion expanded.

RC1.m1 (minor): De-quantization, uniform versus Weibull

Regarding the "de-quantization step" (L152). I understand the need for smoothing the wind speed distribution, however, it would seem to me that the uniform distribution is a bad approximation to use. I would expect wind speeds to obey Weibull distribution. For distributions with small median wind speeds the shape of the distribution is quite steep for small wind speeds. I hypothesize that using uniform distribution could result in unphysical or distorted distributions.

Response. The point is structural: de-quantization is applied to the *observations*, which are the common reference against which every dataset is scored, so it shifts all datasets' signed biases by a

common amount and leaves their signed ordering unchanged. Its effect is also confined to the 10 m ASOS comparison and does not touch the hub-height GS validation on which the paper's claims rest.

On the uniform draw specifically, the method assumes no distributional form anywhere: quantiles are computed empirically and no parametric family is fitted at any stage, so nothing rests on wind speeds being Weibull or any other shape. Within the calm bin the sub-bin shape is unknowable, since rounding to integer knots destroys exactly the information that would distinguish speeds inside the interval. A Weibull fitted to the *resolved* part could be extrapolated downward, but that imports an assumption from outside the interval, and the low-speed tail is where such a fit is least dependable, since calms are shaped by stable stratification, obstacles, and anemometer stall behavior that a two-parameter form does not represent. Uniform introduces no structure we cannot support.

We have also measured what the choice is worth. Calm reports are not rare, at 21.4 % of hourly observations network-wide, so this cannot be dismissed as affecting negligible mass. What makes it immaterial is that the competing fills are nearly the same distribution over a $1.03\,\mathrm{m\,s^{-1}}$interval. Fitting a Weibull per station to the resolved ($> 2\,\mathrm{kt}$) portion, extrapolating into the calm bin, and comparing conditional means against uniform's changes a station's long-term mean by at most $\pm 0.017\,\mathrm{m\,s^{-1}}$for 90 % of stations, reaching $0.095\,\mathrm{m\,s^{-1}}$only at the most extreme site. Against the mean absolute biases the benchmark measures (0.54 to $0.84\,\mathrm{m\,s^{-1}}$at ASOS), that is smaller by a factor of about thirty typically and about six at worst.

Manuscript changes.

- The data section states the de-quantization procedure, its analytic bound, the invariance argument above, and its restriction to the ASOS comparison.
- No parametric distribution is assumed or fitted anywhere in the method; the quantile construction remains fully empirical.
- The sub-bin sensitivity check summarized above can be added to the appendix if the referee would find it useful.

RC1.m2 (minor): Slope and aspect, real versus model-resolved terrain

> *The authors find that slope and aspect are useless as ML parameters (Figure 11(b)). I would argue that this represents a missed opportunity from methodological point of view. I would argue that it is more likely that improvement will be found when comparing the "real terrain" parameters with "what model thinks" is the parameters at that grid cell. Not sure how well that would work for large grid-cells like ERA5.*

Response. We agree, and think this is the more promising formulation for the reason the referee implies. Each model carries its own coarsened orography: an ERA5 cell represents roughly 31 km of landscape by a single elevation, so a station on a ridge may sit several hundred meters above the elevation the model effectively assigns it. Raw slope and aspect describe the true landscape but say nothing about that discrepancy, which is plausibly why they carried no signal in our sweep. The difference between the true high-resolution terrain and the terrain the model resolves is diagnostic in a way the raw descriptors are not, since it indicates directly where and by how much a product's effective surface departs from the real one. The revision names this *representativeness gap* as the quantity worth encoding, retains the negative result for raw descriptors, and notes the construction difficulty for coarse grids that the referee anticipates.

Manuscript changes. Limitations proposes terrain features constructed as the difference between true and model-resolved orography as future work. The negative result for raw slope and aspect is

unchanged.

2 Referee #2 (RC2)

We thank Referee 2 for the positive assessment of the framework, the observational archive, and the WindWatts deployment. The requests concern length, which we have taken seriously.

RC2.1: Shorten; move ML details and Sect. 6.1 to an appendix

> *The manuscript is quite long and could benefit from significant shortening. Some sections, particularly methodological descriptions and supporting details, could be condensed without sacrificing clarity. I suggest moving the details of the ML algorithm and section 6.1 to an appendix. This would improve the readability of the main text and help maintain focus on the key contributions and results.*

Response. The manuscript has been shortened substantially. Because the other referees requested several new analyses, we performed the compression after those additions so that net length is controlled: the body is now roughly 17 % shorter than the discussion paper *despite* the added content. The reduction came from three subsection-by-subsection passes removing restatement and redundancy, from consolidating figures, and from removing material that did not earn its place.

We have moved Sect. 6.1 to an appendix as suggested. The machine-learning methodology we have compressed rather than relocated, because Referees 1 and 3 press hard on precisely these details: the rationale for balancing the two observational archives, the evidence on cross-dataset error correlation, the cross-validation and neighbor-exclusion scheme, and the single-source ablation. Moving that material to an appendix would make the paper harder to evaluate on the points those referees identified as central. We have instead reduced its length in place, chiefly by deleting prose that restated equations in words, and would relocate it if the editor judges the balance differently.

Manuscript changes.

- Storage and access architecture (Sect. 6.1) moved to an appendix.
- Figure count reduced from 25 to 14 plus one full-width figure; two tables removed. Several figures were merged rather than deleted, with the underlying statistics retained in the text.
- Three compression passes through every subsection, including removal of cross-section redundancy between the Introduction, Background, Results, and Discussion.
- Machine-learning methodology condensed in place rather than relocated.

RC2.2: Too dense; simplify or relocate equations

> *While the paper is written in a scientific manner, it may be overly dense for the distributed wind energy audience. In particular, some equations and derivations in the main text could be simplified or moved to an appendix.*

Response. We have simplified where possible while keeping the main narrative self-contained. Prose that merely restated an equation in words has been removed throughout; we found this accounted for more of the density than the equations themselves.

Manuscript changes. Equation-restating prose is removed from the metrics and problem-formulation subsections, redundant metric equations are consolidated, and the trapezoidal-integration detail is reduced to its result.

3 Referee #3 (RC3)

We thank Referee 3 for an exceptionally thorough review. Several comments prompted new analyses that materially changed the paper, and in two cases changed our interpretation of the results.

RC3.M1: L228 "as informative as any subset"; contradiction with L22–24; how many years?

> *I am very interested in the statement in line 228 that "a quantile summary over all available years is essentially as informative as a summary restricted to any subset of years". This is problematic for a number of reasons. Firstly, the pdf of interannual wind speed range (inset in figure 4) shows a peak in the distribution at 0.5 m/s, and many sites are to the right of this with larger differences. Claiming that this is not important is in direct contradiction to lines 22-24, which state that "biases of order 0.5-1.0 m/s . . . can translate into substantial changes in annual energy production". Moreover, it is not clear how many years are actually used at each site: Line 159 states that "only years with at least 95% completeness are retained". Is there any threshold for how many years should be retained? Would a station with only a single valid year be included?*

Response. Both points are well taken: the L228 claim was asserted rather than demonstrated, and as written it sits awkwardly against L22–24.

The tension dissolves once the two quantities are distinguished, which the revision now does explicitly. A *systematic* bias recurs in every year of record and propagates directly into an energy estimate. Interannual *variability* is a sampling effect: it describes how much a single year departs from the long-term mean, and it shrinks as years are averaged. What matters is therefore not the interannual range itself but the sampling uncertainty of the multi-year mean we actually estimate. That uncertainty is small: a median standard error of $0.06\,\mathrm{m\,s^{-1}}$, roughly an order of magnitude below the median model bias of $0.48\,\mathrm{m\,s^{-1}}$, with bias exceeding sampling error at 86 % of sites. Pooling is now justified on that comparison rather than by assertion, and the overstated L228 sentence is removed.

On the year policy, we agree it was insufficiently specified and now state it in full. Years are retained at $\geq$95 % completeness; ASOS additionally requires at least five valid years per station, whereas GS applies no multi-year floor, so a site–height with a single complete year is retained. That is a real exposure, and we quantify it: 26 % of GS site–heights have two or fewer valid years. The sensitivity experiments requested in RC3.M3 test whether this distorts the benchmark, and they indicate it does not.

Manuscript changes.

- Interannual section rewritten; the L228 claim removed; bias-versus-standard-error comparison added.
- Valid-year distribution and minimum-year policy reported (shared with RC1.M4).
- L22–24 reconciliation added.

RC3.M2: ENSO variability (Hamlington et al. 2015); representative years; guidance on number of years

> *Following on from point 1, there have been a small number of studies looking at the interannual variability in wind speed in the US. For example, Hamlington et al. 2014 found up to 20% variabilty in the wind resource with ENSO, particularly in the western part of the United States. This suggests that a*

representative number of years that includes all three phases of ENSO should be included in the analysis. I think this is an important point because it can provide concrete guidance on how many years of wind data the industry should use for wind resource assessment. I'm not sure it is correct to be claiming that the number years does not really matter.

Response. The referee is right, and we no longer make that claim. The number of years does matter; our error was in not saying what it matters *for*. Two questions are involved, and they have different answers.

Characterizing a site's absolute resource is the case Hamlington et al. (2015) speaks to, and there the referee's reasoning holds: if ENSO modulates the resource by up to 20 %, a record that samples only one phase will misestimate the long-term mean, and spanning a full cycle is advisable. Our own climatology is built on 2007–2024, which covers multiple ENSO phases.

Ranking datasets is the other question, and it turns on the distinction drawn in RC3.M1: the quantity compared is a systematic bias recurring in every year, not a climate signal varying between them. Restricting the benchmark to validation sites with progressively longer records (≥ 2, ≥ 3, ≥ 5 valid years) leaves the rankings unchanged and, if anything, widens WEM's margin through the ≥ 3-year threshold (RC3.M3). That establishes robustness to including short-record sites, not a minimum anyone should adopt: we ourselves retain GS site–heights with a single complete year, since the biases being measured, of order 0.5 to 1.2 $\mathrm{m\,s^{-1}}$, are much larger than the year-to-year variability of a site's own mean. That reasoning does not carry over to absolute resource assessment, where the site mean is itself the quantity of interest and the tolerance far tighter. There the referee's standard is the right one, and we offer no threshold of our own.

Manuscript changes. Hamlington et al. (2015) is added and discussed in the interannual-variability appendix.

RC3.M3: Requested experiments (IAV map, year-threshold, restricted-years)

To address these concerns, as a minimum it would be useful to see: (a) A map of the interannual (max − min) variability plotted on a map, to see whether the spatial variability matches with that of Hamlington et al. (2014). (b) A sensitivity experiment where only stations that have above a threshold number of years. (c) A sensitivity experiment where all stations are restricted to a small number of years, versus including all years.

Response. All three analyses were performed. Experiments (b) and (c) are reported: dataset rankings are unchanged across the $\geq$2/3/5-year thresholds and WEM's margin, if anything, widens through the $\geq$3-year threshold; the fixed common sub-period (2014–2024) reproduces the all-years ranking exactly; and the spread across random two-year sub-windows is only $\approx 0.02\,\mathrm{m\,s^{-1}}$, far smaller than the gaps between datasets.

For (a) we performed the analysis but concluded the map should not be published. Our archive cannot support the spatial test the map is meant to enable. Only 49 western GS sites have three or more valid years, and among those multi-year sites there is no statistically significant west–east difference in interannual range (0.50 vs. 0.49 $\mathrm{m\,s^{-1}}$; Mann–Whitney $p = 0.38$). A larger west–east contrast does appear if all sites are included, but it is an artifact of record length, since one- and two-year sites have poorly defined interannual metrics, and it disappears once short records are excluded. A map would therefore have displayed a spatial pattern driven by uneven record length rather than by climate, and invited exactly the comparison with Hamlington et al. that our network cannot legitimately support.

We therefore characterize interannual variability quantitatively (median range $0.48\,\mathrm{m\,s^{-1}}$; 88 % of sites below $1\,\mathrm{m\,s^{-1}}$) and state plainly that we neither confirm nor contradict the ENSO geography with this network. We would rather report the limitation than present a figure whose most visible feature is a sampling artifact.

Manuscript changes.

- Year-threshold and restricted-window experiments added (summary in text, detail in appendix).
- Interannual variability characterized quantitatively; the west–east test and its record-length confound reported in the appendix.
- [The IAV map is not included], for the reasons above.

RC3.M4: Redundancy across Table 2 / Figs. 6, 8, 9; recolor Fig. 9 by height; bias heat map

> *It looks like table 2 and figure 6 are essentially showing the same information. Moreover, many of the numbers in the table are also repeated in the text (lines 408-430). I suggest this material can be condensed, for example, show the median numbers on the box plots, and get rid of table 2, and then just pull out the key points briefly in the text. This same data is then repeated in a spatial sense in figure 8, and as scatter plots in figure 9. There are many different ways that this data could be combined, but one idea would be to retain figure 9, but instead of colour coding the dots by bias (which is shown in the scatter plot anyway, colour code the points by height. This would remove the need for figure 7, and also be quite interesting. Another idea would be a heat map of bias, with wind speed on the horizontal axis and height on the vertical axis. This might actually bring out some additional structure in the data.*

Response. This comment drove the largest structural change in the revision, and of the two consolidations the referee proposed we took the second, which proved more valuable than we anticipated. Table 2 is removed and the boxplots are consolidated into a single figure showing signed and absolute bias for all datasets together with WEM. Rather than recoloring the scatter plots by height, we resolve bias jointly in height and wind speed in one figure, which achieves the same economy (it likewise removes the need for the separate height figure) while exposing structure that neither alternative would have shown on its own.

That figure changed our interpretation. In the discussion paper we described a wind-speed "compression" effect and, separately, a positive offset for the WTK-LED products. The joint view shows these are the same effect: bias is organized along a rising diagonal, so the controlling variable is not absolute wind speed but wind speed *relative to the value expected at that height*, the signature of regression toward the height-conditional mean. Products differ mainly in where that diagonal sits. The joint view also corrected a factual error: we had written that both WTK-LED ensembles are positively biased almost everywhere, whereas WTK-LED Climate turns to underestimation at the highest wind speeds and only WTK-LED CONUS is warm nearly everywhere.

We did not annotate medians on the boxplots: with seven products across two panels the labels crowded the figure, and the median values are given in the accompanying table.

Manuscript changes.

- Table 2 removed; three separate boxplot figures consolidated into one (signed and absolute bias, all datasets and WEM, ordered worst to best).
- New joint height × wind-speed bias figure (the referee's heat map, rendered as per-dataset scatter panels), replacing both the separate height figure and the earlier bias heat map.

- Height-dependence text rewritten around the diagonal, height-conditional framing; the WTK-LED Climate error corrected.
- Scatterplot discussion restructured to treat ASOS and GS separately, and the overstated "compression" language moderated to match what the data support.
- ASOS benchmark results presented in a compact table, with GS results shown later alongside WEM and forward-referenced from the benchmark section.

RC3.M5: "ML always zeroes bias"; spread matters more; train on a single dataset

> *The authors mention several times that the mean bias across all stations in the new dataset is reduced to zero. Is it not the case that a ML model will always reduce the bias to virtually zero, since this is what it is optimising for? Therefore, it is less important that the mean bias is reduced to zero, and more important that the spread about this mean is reduced. I believe that if you just trained the ML model on one input dataset (eg ERA5 only), then it would also reduce the mean bias to zero. Did you do this test? For example, in figure 10, could you add the model error if you trained it only on WTK or only on WTK-LED CONUS. This would add weight to your argument that it is the combination of the multiple datasets that leads to the model improvement, not just that you're essentially bias correcting a dataset.*

Response. The referee is right on both counts, and we tested the specific case asked for. We swept every combination of wind products, from single sources through to the complete set, retraining under the identical protocol. Every configuration, including each single-dataset variant, drives the aggregate mean bias to near zero. Bias-zeroing is therefore generic to the objective, exactly as the referee says, and on its own it demonstrates nothing about the value of combining datasets. We have removed the framing that treated near-zero mean bias as evidence of skill.

We also agree with the referee's implied standard: a near-zero mean bias is necessary but not sufficient, because a model can be unbiased on average while still being badly wrong site by site in compensating directions. What matters for a user siting a turbine is the *spread* of site-level error, and this is where the products separate. In the site-level bias distributions, WEM has both the narrowest interquartile range of any product and a signed-bias distribution tightly centered on zero and nearly symmetric, in contrast to the wider, skewed distributions of the individual datasets. Its mean and median signed bias agree to $0.01\,\mathrm{m\,s^{-1}}$, indicating a genuinely centered error distribution rather than one that averages to zero through cancellation.

The ablation then shows that this contraction requires the combination. No single dataset performs as well as a fusion of several: single sources occupy the high-error end of the sweep, error falls in steps as complementary products are added, and although WTK is the strongest individual source, every configuration in the better-performing half of the sweep pairs it with other datasets. The lowest error of all is reached not by any single product, nor by the full set, but by a compact subset.

Manuscript changes.

- Single-source and all-combination ablation added as one presence-matrix figure with cross-validated MAE per configuration.
- The subsection is retitled "Fusion versus single-source correction" and merged with the feature-sweep outcomes, since both analyze the same sweep (also RC3.M8).
- The "mean bias reduced to ∼zero" framing tempered throughout in favor of error and spread reduction and the flattening of quantile-dependent bias.

- The separate marginal-feature-effect figure and its accompanying statistics are removed; feature selection is now reported from the single sweep figure.

RC3.M6: Representativeness vs. actual model error

> *Can the authors comment on the representivity of the datasets, versus actual model errors? For example, is ERA5 really providing a poor result, or is it providing a result that is reasonable but at a larger scale? Is it reasonable to compare ERA5 to an observations that have been influenced by small-scale obstacles and changes in surface roughness?*

Response. We address this empirically. ERA5's site-level |bias| grows significantly with terrain complexity ($r = 0.26$, $p < 10^{-5}$), the signature expected if much of its error is representativeness error rather than a defect of the model physics at its own resolution. We now say explicitly that we do not read the benchmark as a claim that ERA5 is "wrong": at its native scale it is a skillful product, and part of its site-level bias is the unavoidable cost of representing sub-grid terrain in a ∼31 km cell.

Two considerations nonetheless make the comparison fair and useful, and we state both. First, the representativeness error is real from the user's standpoint: a developer querying ERA5 at a prospective site inherits exactly this point-versus-grid discrepancy, so the reported site-level bias is the operationally relevant quantity for desktop assessment, whatever its origin. Second, WEM's error does not grow with complexity in the same way, indicating that the ensemble absorbs much of the representativeness-driven component rather than passing it through.

Manuscript changes. The terrain stratification uses fixed physical complexity classes rather than terciles, so the complex class isolates genuinely rugged terrain.

RC3.M7: Per-station uncertainty in the output

> *Can the authors include uncertainty measurements for each station in the model output? This seems like it should be within reach, given that multiple inputs go into each station, and also that the ML algorithm presumably gives some metric of uncertainty.*

Response. We attempted this and decided against it. What we built was a stratified characterization formed from cross-validation residuals binned by height and terrain class, and on review we judged it unsound to present as uncertainty. Those bands are conditional-average residual spreads, not calibrated predictive intervals, so labeling them "predictive uncertainty" would invite readers to place them around a new site's estimate, which they do not support. They were also thinnest where they were most interesting, the widest band resting on ten site–heights, and stripped of the noisy cells they largely restated the mean absolute error already reported.

The honest position is that WEM does not produce an uncertainty estimate. It is a point-prediction model: the quantiles it predicts describe the wind speed distribution at a site, not the uncertainty of that prediction. We therefore state plainly that quantile targets and predictions are treated as deterministic and that per-site predictive intervals are not provided. Genuine uncertainty quantification requires a modeling approach that supplies it, such as uncertainty-aware loss functions, ensemble or Bayesian variance, or conformal prediction, which we identify as the natural next step.

Manuscript changes.

- [No stratified uncertainty section is included], for the reasons above.

- Limitations states that WEM emits point predictions without per-site predictive intervals, and identifies uncertainty-aware losses, ensemble and Bayesian methods, and conformal prediction as the route to genuine uncertainty quantification.
- The legitimate error characterization (typical held-out error, its behavior by height, and its increase in complex terrain) is retained where it belongs, in the benchmark and terrain diagnostics.

RC3.M8: Relationship between Figs. 10 and 11

> *Please clarify the relationship between figures 10 and 11, and associated text. Is figure 10 the result of the "combinatorial sweeps", and then after that a separate experiment was run where all datasets were included, but then withheld one by one? Are both these figures really necessary, and why are the results presented in different ways (model error in figure 10, and box plots in figure 11)?*

Response. They were two views of the same sweep, not separate experiments, and the confusion the referee identifies was a genuine defect in the presentation. Rather than clarify two figures we have replaced them with one, and merged the two subsections that discussed them, since both analyzed the same experiment.

Manuscript changes. The feature-sweep and ablation subsections are merged and placed after the row-level metrics they depend on, which also fixed a forward reference to a metric not yet defined.

RC3: Minor comments

m1 ("national" / US not stated). **Response.** Corrected.

Manuscript changes. Abstract and first use now specify the contiguous United States (CONUS).

m2 (L33 "intermittency"). **Response.** Agreed; the study uses quantiles and does not examine intermittency.

Manuscript changes. "Intermittency" removed.

m3 (L43 "at a single dataset"). **Response.** Agreed. Learning the quantile *distribution* is the core methodological step and applies even to a single dataset, distinct from multi-source fusion.

Manuscript changes. Introduction reworded to foreground distribution learning and separate it from the fusion contribution, pointing to the ablation.

m4 (L104 surface-layer parameterization). **Response.** Added.

Manuscript changes. "Surface-layer parameterization" inserted.

m5 (L135 "energy-producing or hazard-prone"). **Response.** Adopted.

Manuscript changes. The passage now reads "... robust across the whole distribution, including the energy-producing or hazard-prone tails that govern both expected yield and extreme-wind loading."

m6 (L155 uniform within $(0, 2]$ kt; discontinuities?). **Response.** Within $(0, 2]$ kt each speed is treated as equally likely, the maximum-entropy choice given that rounding maps all such values to 0 kt.

Manuscript changes. Covered by the RC1.m1 clarification, which also states the scope and the invariance argument.

m7 (L161 actual number; 1842 before/after filter; "long-term"). **Response.** We retrieved 1972 ASOS stations and retained 1842 after completeness filtering and restriction to the ERA5 land grid.

Manuscript changes. Retrieved and retained counts and the valid-year distribution now stated; "long-term" wording adjusted.

m8 (L166 306 vs. 368). **Response.** Correct. 368 counts distinct site–height combinations and 306 distinct sites; the difference is towers instrumented at more than one height.

Manuscript changes. Clarifying clause added.

m9 (L170 "fewer sites sample only near-surface", be precise). **Response.** Made precise.

Manuscript changes. Exact height-bin counts given (shared with RC1.M1).

m10 (L240 coastal interpolation over sea). **Response.** The four-nearest-neighbor IDW applies no land–sea mask. We bound the exposure (∼188 stations, 8.9 %, sit within one source grid cell of the coast) and show the impact is immaterial: none of the datasets informing the ensemble is significantly worse at coastal stations. Only ERA5 degrades ($+0.15\,\mathrm{m\,s^{-1}}$, $p < 10^{-4}$), consistent with its coarser resolution rather than with the interpolation, and ERA5 is not among the final wind inputs.

Manuscript changes. Documented in the interpolation subsection, with a land-masked reinterpolation noted as straightforward future confirmation.

m11 (L345 subsampling, how many remain; once or repeated?). **Response.** Clarified, with a seed-stability check.

Manuscript changes. Methods states the per-fold down-sampling (median ≈34,500 rows, ≈340 stations, matched one-to-one to the held-in GS row count) and reports that held-out MAE varies by less than $0.004\,\mathrm{m\,s^{-1}}$across five random seeds.

m12 (L395 Optuna citation). **Response.** Added.

Manuscript changes. Akiba et al. (2019) cited and added to the references.

m13 (consistent dataset ordering). **Response.** Adopted.

Manuscript changes. All figures and tables now order datasets by performance (worst to best) with WEM last, using a consistent color assignment.

m14 (L415 "perform slightly worse", significance). **Response.** Added Holm-corrected paired Wilcoxon tests on site-level |bias|. At ASOS the two leading products (HRRR, WTK) are statistically indistinguishable ($p = 0.24$), while GWA sits significantly above WTK ($\Delta = 0.04\,\mathrm{m\,s^{-1}}$, $p < 0.001$) and ERA5 above GWA ($\Delta = 0.05\,\mathrm{m\,s^{-1}}$, $p = 0.03$); the ordering is real but the products are tightly bunched, all four within roughly $0.1\,\mathrm{m\,s^{-1}}$.

Manuscript changes. Significance results reported, with the test, the meaning of Δ, and the multiple-comparison correction explained; "slightly worse" replaced by explicit levels and significance statements.

m15 (L456–462 stability; wind-shear predictor). **Response.** We agree the pattern is consistent with stability effects, but have moved this from an explanation to a hypothesis: the joint height × wind-speed figure (RC3.M4) shows the structure is organized by wind speed relative to the height-expected value, which our earlier shear explanation did not capture.

Manuscript changes. The stability explanation is removed from the Results. Limitations proposes a time-resolved stability or shear descriptor drawn from the mesoscale models' native multi-height hourly output as future work, noting that a static climatological shear cannot represent the time-varying distribution such a pattern would require.

m16 (Fig. 10 colors). **Response.** Colors denote dataset identity.

Manuscript changes. Caption clarified as part of the figure merge described under RC3.M8.

m17 (Eq. 23 tautology). **Response.** A typesetting error, now corrected.

Manuscript changes. The row-level error is defined once, with the absolute value taken in the aggregation.

m18 (L572 both-sign correction with one dataset?). **Response.** Yes. A single-dataset model also corrects both signs, and we now say so directly rather than presenting two-sided correction as distinctive.

Manuscript changes. Reworded to state that the telling property is the contraction of the *spread* of site-level biases, cross-referencing the ablation.

m19 (Sect. 6 vs. Figs. 18–19). **Response.** Figs. 18 and 19 are the full-grid product described in Sect. 6.

Manuscript changes. A clarifying sentence anchors the maps to the deployment section, and the deployment description no longer repeats the feature construction already given in the Methods.

m20 (Figs. 4–5 axis scales; inset labels). **Response.** Adopted.

Manuscript changes. Axis scales unified across panels and insets, and inset axes labeled.

m21 (Fig. 7 density and trend fits). **Response.** We added per-dataset trend fits, but the analysis did not support presenting them: the marginal height trends are weak, noisy, and inconsistent in sign across datasets ($|r| \leq 0.18$). The joint height × wind-speed figure shows why. There is no systematic dependence on height *alone*, but a strong joint organization of bias by height and wind speed together, which the marginal view averages away.

Manuscript changes. Fig. 7 is replaced by the joint figure and height dependence described through the diagonal structure rather than trend lines; no slope table is reported.