Reviewer 2

Dear Review 2,

Thank you for the feedback and volunteering your time for peer review. We have updated our manuscript following your comments, and we believe the manuscript is now stronger. Please see our updates below.

GENERAL ASSESSMENT

This manuscript presents a valuable international benchmark of wind-turbine inflow reconstruction methods using complementary real-world lidar measurements and fully observed synthetic large-eddy simulation cases. The combination of real and synthetic datasets is a particular strength. The manuscript is generally well written and is commendably transparent about several limitations of the measurements, models, and relatively small set of cases.

The real-world reference contains substantial uncertainty arising from spatial separation, induction correction, advection, Taylor's frozen-turbulence hypothesis, line-of-sight reconstruction, probe-volume averaging, gridding, synchronization, and signal processing. These uncertainties may be comparable to the differences reported among models. The measurement campaign and the associated measurement operators therefore require more detailed documentation and uncertainty analysis.

Major revision is recommended to improve the quality and readability.

MAJOR COMMENTS

1) The headline ranking is confounded by unequal spatial resolution and smoothing

SSM and LER were generated at 10 m resolution and linearly interpolated to 5 m, whereas the Kaimal- and conventional Mann-based fields were generated at 5 m. The models with the lowest RMSE are also described as producing the smoothest fields. Pixelwise RMSE can favor smooth predictions because small-scale variability and phase errors increase squared error.

All methods should be evaluated at a common effective resolution, preferably by filtering or coarsening every reconstruction and reference field to 10 m before comparison. Results should also be presented over several spatial scales. Otherwise, the conclusion that SSM and LER are more accurate may partly reflect spatial filtering rather than superior reconstruction.

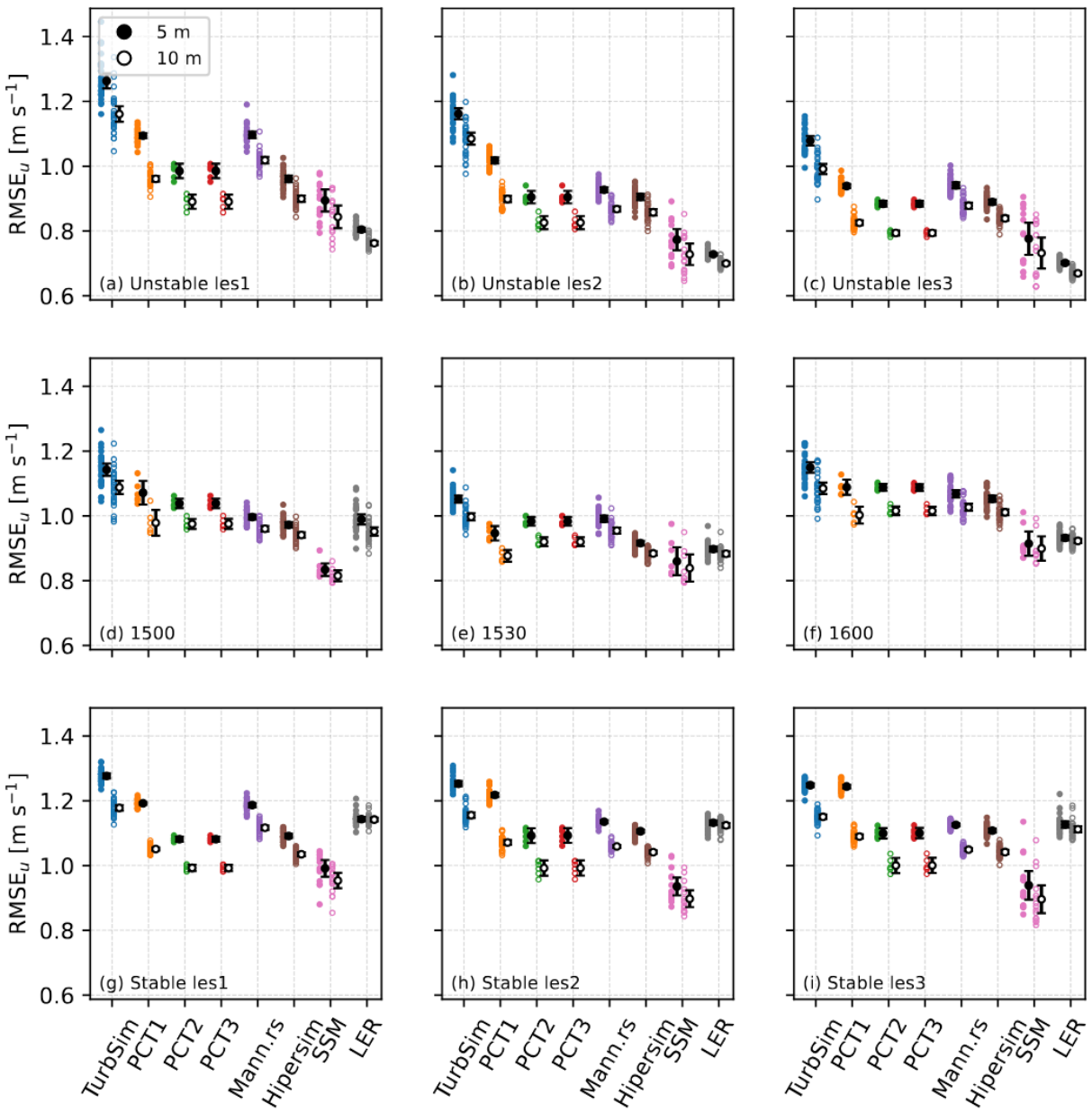
Thank you for this comment. We agree that the coarser resolutions of SSM and LER were artificially boosting their RMSE compared to other inflows, and the magnitude of this boost is now characterized more thoroughly. For context, we encouraged participants to submit inflows at as high a practical resolution as possible because of a planned follow on aeroelastic study. We have reasons to believe that 5 m inflows produce a notably more accurate turbine response than 10 m inflows for our turbines of interest. But, 10 m resolution was computationally infeasible for SSM and LER.

In an ideal world, we would have had benchmark participants re-submit their 5 m resolution inflows at 10 m resolution to study the question of RMSE impacts. However, that was not possible given limits from the benchmark participants. Instead, we did as you suggested and averaged the 5 m resolution inflows and to 10 m resolution. We wanted to use the native coordinates of the 10-m resolution models (e.g., z=10 m, 20 m, 30 m, ...), which necessitated the use of trapezoidal averaging/smoothing of the 5 m flows. For the sake of consistency, we re-processed the 5 m resolution target fields using trapezoidal averaging.

We recalculated all the quantities of interest at the coarser resolution (e.g., shear profiles, ensemble spread in time series, etc.). The only major changes were to RMSE. We updated Sec 2.2 to state that we conducted all analysis both at 5 m resolution and at 10 m resolution. The results in Sec 3 and 4.1 were fundamentally unchanged, so we maintain the 5 m analysis figures there. We updated Sec 4.2 to show results for both 5 m and 10 m analysis. Major updates include revisions to Figure 14, the addition of Table 4, and updates to the text. As expected, the inflows that were provided at 5 m resolution improved their RMSE when moving to 10 m. Improvements were on the order of 5-13%. However, because the reference fields were also coarsened, SSM and LER also saw RMSE improvements on the order of 2-3%. SSM still shows the best overall performance, but LER performs closer to PyConTurb and the Mann-based models when averaging across all cases. We have adjusted the text to slightly downgrade the performance of LER.

Below are the concrete, averaged numbers across all cases: (Table 4) (Figure 14)

Model	Mean RMSE$_u$ 5 m [m s^{-1}]	Mean RMSE$_u$, 10 m [m s^{-1}]
TurbSim	1.18	1.10
PCT1	1.09	0.97
PCT2	1.02	0.93
PCT3	1.02	0.93
Mann.rs	1.05	0.99
Hipersim	1.00	0.95
SSM	0.88	0.85
LER	0.94	0.92



2) The real-world reference uncertainty must be quantified

The SpinnerLidar-derived field is not a direct measurement of the true three-dimensional inflow. It involves observations near 1D rather than at the 3D reconstruction plane, advection using an average velocity, Taylor's frozen-turbulence hypothesis, induction removal, line-of-sight velocity reconstruction, neglected turbine tilt, spatial gridding, temporal processing, and gaps between data files.

Flow evolution over the approximately 2D separation can create an error even for a perfect reconstruction at 3D. The manuscript also shows disagreements of approximately 1 m/s among instruments, which are comparable to the reported reconstruction RMSE. The authors should quantify the uncertainty of the processed reference and perform sensitivity analyses for the advection velocity, time shift, induction correction, and reconstruction assumptions. Real-world RMSE should not be pooled or interpreted together with LES RMSE as though the two references had equal fidelity. The SpinnerLidar-derived field should be described as a processed reference reconstruction rather than unqualified ground truth.

We agree that the SpinnerLidar inflow is an imperfect ground truth, and we urge readers to keep this truth in mind "While throughout the manuscript we refer to the SpinnerLidar data as 'ground truth' data, we strongly emphasize that the SpinnerLidar-derived inflows are an imperfect measurement of the true winds" (L183). The text before this sentence details each major processing step between raw measurements and gridded inflow fields. We included that text because we wanted to explicitly highlight sources of uncertainty. We have made that goal more explicit by adding the statement "Each of these steps introduces uncertainty, and a more thorough accounting of errors can be read in (Houck et al. 2026)." (L181) The cited paper was published while our current paper was out for peer review, and it contains a more thorough analysis of SpinnerLidar wind fields.

We believe that if we had only used real-world measurement data as a target in this paper, then a more thorough accounting of sources of uncertainty would be apt. Our main reason for running the LES-style field campaigns was to circumvent these sources of error in the real world. The data from the LES cases is fully known without any approximations. We find that the behavior of the real-world reconstructions largely tracks the LES-style reconstructions, which builds confidence in the pattern seen in our real-world results.

As you note, there are disagreements of approximately 1 m/s between real-world instruments. Thus, to be thorough, we would not only need to further characterize uncertainty from the SpinnerLidar, but also uncertainty from every other instrument. We believe this kind of thorough accounting would be a valuable manuscript in its own right. In the wind energy literature, measurement intercomparison papers largely compare 10-min statistics over many bins (e.g., Letizia et al. (2024) in JRSE), but we are unaware of papers that do thorough instrument intercomparison analysis on a second-by-second basis. As such, we believe your request for more observational analysis is beyond the scope of the present manuscript. We note that

Letizia et al. (2024) shows that disagreements of 0.377 m/s in 10-min mean flow are common in their instrument intercomparison work, even for a correlation coefficient of 0.993 (L205). We also update the closing paragraph of the manuscript to read "We encourage further observational research on second-to-second timescales. In the RAAW campaign, instruments could disagree on these timescales, even though they were sampling approximately the same air mass." (L563)

We pool real-world RMSE with LES RMSE only in one place, and that is to provide a simple top-line value to refer to, which is helpful for clarity regarding the conclusions (e.g., Major Comment 3). Otherwise, we display RMSE values for the 9 individual cases and do not pool them.

3) Model rankings need uncertainty estimates and statistical assessment

The conclusion that SSM performs best is based on nine short and potentially correlated cases, with only three cases in each broad category. Differences among leading methods are small. For example, the average difference between SSM and LER is approximately 0.04 m/s, which may not be practically or statistically meaningful relative to case-to-case and reference uncertainty.

The authors should report paired case-level differences, uncertainty intervals, and sensitivity to the averaging and weighting procedure. A paired or hierarchical bootstrap could account for case and ensemble variability. The manuscript should distinguish consistent ranking differences from differences that are small relative to uncertainty.

We are uncertain about some of the specific terminology used in Comment 3, e.g., paired case-level differences and what "weighting procedure" refers to. We believe that your core critique here is that our original draft did not address the question: "Is model X more accurate (with statistical significance) than model Y for the case studies presented?". We addressed this question in two manners.

First, we revised Figure 14. In addition to showing the scatter of each individual ensemble member, we show the mean as well as the 95% confidence interval of the mean RMSE. Non-overlapping bars indicate that the difference is statistically significant. (Note that a small amount of overlap is permitted for statistical significance to hold). This figure enables interested readers to determine if one specific model confidently outperformed a different model in each case. The updated figure more clearly shows cases for which this has occurred e.g., the difference of case-specific RMSEs of SSM and LER is not statistically significant such as in cases 1530 and 1600. The caption of Figure 14 has been updated to

include "Adjacent black scatter points denote the ensemble mean for the model, and the bars denote the 95% confidence interval on that mean.".

Second, we adjusted the final, single RMSE number for each model. Previously, we were equally weighing each ensemble member equally across all cases, which led to a minor penalty for the real-world cases, as these sometimes had fewer ensemble members. Now, we report the final RMSE by averaging each case-specific RMSE from the 9 different cases. Each of the case-specific RMSEs came with uncertainty due to ensemble spread, and we propagate that uncertainty into the final RMSE value through quadrature. This change in methodology did not change the ranking of any models; it only changed the final reported RMSE.

Both these changes are reflected in the body of the text "For each case and model, we also calculate the ensemble mean RMSE_u, and the 95% confidence interval that only reflects the sampling variability of the finite ensembles.. We also calculate one, top-line RMSE_u value for each model by taking the ensemble-mean RMSE_u from each of the nine cases and then averaging." (L495)

4) Potential distribution advantage for LER in the synthetic cases

The LER neural networks were trained using AMR-Wind simulations with the same large-scale forcing used to generate the synthetic benchmark cases. Although the training and evaluation simulations may be separate, the stable and unstable evaluation cases appear closely matched to LER's training distribution.

The authors should clarify the independence of the training and evaluation datasets, including random seeds, simulated periods, boundary conditions, surface properties, and forcing. LER's strong performance in unstable LES cases should be described as in-distribution performance unless generalization to meaningfully different LES conditions is demonstrated.

In Appendix B7 of the original text, we noted "These simulations used the same large-scale forcing used to generate the synthetic cases, though these simulations maintained a uniform grid spacing of 10 m and did not employ grid refinement zones.". We have updated that text to read "These simulations used the same large-scale forcing and surface conditions used to generate the synthetic cases, though these simulations maintained a uniform grid spacing of 10 m and did not employ grid refinement zones. To avoid data leakage between the LER training simulations and the synthetic evaluation cases, the training simulations used a different initial random perturbation than the benchmark cases, despite sharing the same large-scale forcing and surface conditions.". Due to the small number of case studies, we

state that the performance of all models in the benchmark is specific to the periods analyzed here, and generalization performance is unclear (L560). We also added "To avoid the appearance of preferential treatment by the benchmark organizer, the networks were trained using the same hyperparameters used in (Rybchuk et al. 2025)." (L662)

5) The measurement campaign and instrument configurations require a more complete description

The manuscript should provide a self-contained description of the RAAW measurement campaign. For each instrument - the nacelle-mounted scanning lidar, SpinnerLidar, profiling lidar, and meteorological mast - the authors should report the make and model, pulsed or continuous-wave operating principle, wavelength, coordinates relative to the turbine, measurement heights and upstream distances, scan trajectory and angular range, sampling and scan-completion frequencies, range-gate or focal-distance settings, probe-volume dimensions or weighting functions, temporal and spatial averaging, velocity-retrieval assumptions, data availability, quality-control criteria, synchronization, and coordinate transformations.

A plan-view and side-view schematic should show the turbine, meteorological mast, profiling lidar, nacelle-lidar measurement plane at 3D, SpinnerLidar scan near 1D, wind direction, scan trajectories, and effective measurement volumes. Spatial separation and different measurement volumes directly affect the apparent reconstruction error and therefore must be clearly documented.

We have updated the manuscript (Sec 2.1.1, Sec 2.1.2) to provide more details of the measurement campaign.

Instrument make/model: The scanning lidars are Halo Streamline XR models by Lumibird. The SpinnerLidar is a custom instrument from DTU. The profiling lidar is a WindCube v2.1. The met mast had both sonic anemometers (Ultrasonic RM Young Anemometer) and cup anemometers (Thies Clima First Class Anemometer).

Pulsed/continuous: the scanning lidars and profiling lidars are based on a pulsed operating principle, and the SpinnerLidar is continuous wave.

Range gates: we clarified that the scanning lidar had 12 m range gates.

The original text of the manuscript provides the following information that was requested in Sec 2.1.1: coordinates relative to the turbine, measurement heights and upstream distances, scan trajectory and angular range, spatial separation and measurement volumes.

We intentionally omit the following information, as we believe it is not appropriate for this paper: wavelength, sampling and scan-completion frequencies, prove-volume dimensions or weighting functions, temporal and spatial averaging, velocity-retrieval assumptions, data availability, quality control criteria, synchronization, and coordinate transforms.

We do not add a schematic of the site here, but we now tell interested readers to consult https://rotor-inflow-benchmark.readthedocs.io/en/latest/, Doubrawa et al. (2025) and Houck et al. (2026) for schematics (L121).

6) Induction effects at the 1D SpinnerLidar reference plane require stronger treatment

The SpinnerLidar measurements are collected approximately one rotor diameter upstream, where rotor induction cannot generally be assumed negligible. The manuscript states that induction effects are estimated and removed, but the correction is not described sufficiently to demonstrate that the corrected field represents freestream inflow.

The authors should provide the induction model and assumptions, turbine operating variables used by the correction, treatment of radial and azimuthal induction variation, treatment of shear and yaw misalignment, validation of the corrected velocities, sensitivity to model parameters, and uncertainty propagated into the reported RMSE. The combined uncertainty from induction correction, advection from 1D to 3D, and turbulent evolution should be quantified. Recent induction-aware nacelle-lidar reconstruction studies should be discussed when explaining and validating this correction.

We now include a citation to Houck et al. (2026), which was published while our manuscript was in review. That manuscript focuses heavily on SpinnerLidar measurement processing and inflow modeling. For example, that manuscript states "this comparison showed that the processed SpinnerLidar V_los velocities had a root-mean-square error of 0.2-0.4 m/s for the conditions expected in the current study (i.e., measurements of unwaked inflow without significant effects from solid return interference)." Our updated text reads "Measurement data was converted to a gridded inflow, a process detailed in Houck et al. (2026). At a high level, the conversion process estimated induction effects" (L175).

7) Probe-volume averaging and spatial filtering should be included in the comparison

The SpinnerLidar is a continuous-wave lidar, so its observation represents a weighted average over a finite probe volume rather than a point velocity. Probe-volume averaging can attenuate small-scale turbulence, modify spectra, and smooth coherent structures.

The manuscript should explain how continuous-wave range weighting is treated when converting SpinnerLidar line-of-sight observations into gridded u, v, and w reference fields. It should also describe the probe-volume or range-gate characteristics of the nacelle lidar measuring near 3D. The reconstructed model fields should ideally be passed through instrument-specific virtual-lidar operators, including line-of-sight projection, scan trajectory, probe-volume weighting, gridding, and temporal sampling, before comparison with measurements.

The SpinnerLidar measures near 1D and its field is subsequently shifted to the 3D comparison plane; it does not directly measure at 3D. This geometry and the corresponding measurement operators should be stated precisely.

> Similar to Major Comment 6, we include a citation to Houck et al. (2026) to talk about probe-volume averaging effects (L175). That manuscript states, "The Spinner-derived inflows, however, also have uncertainty. Some of these have been quantified in aggregate in [48], but even higher resolution measurements would be required to fully quantify all uncertainties. The full set of uncertainties includes probe-volume effects in the raw measurement, the projection method, the induction removal, and the interpolation."

8) Instrument disagreement in Figures 1 and 2 needs further investigation

Figures 1 and 2 show substantial differences among the nacelle lidar, SpinnerLidar, profiling lidar, and meteorological mast. The manuscript attributes these differences to distinct sampled air masses and signal-processing procedures, but the explanation remains qualitative.

Figure 1 contains instantaneous disagreements exceeding 3 m/s. Exact measurement positions, scan times, effective averaging volumes, and temporal alignment should be reported directly in or near the figure. The authors should determine how much of the disagreement arises from spatial separation, probe-volume filtering, advection, flow evolution, measurement uncertainty, and data processing.

> We agree that these instrument disagreements are interesting and important, but as we stated in response to Major Comment 2, we have seen a dirth of observational studies that examine inter-instrument disagreements on the second-to-second timescale. We believe this topic warrants it own research study and is outside of the scope of our manuscript.

9) The fitted Davenport coherence model should be validated over multiple separations

Table 1 lists one pair of Davenport coherence parameters for all three real-world cases. Coherence may depend on atmospheric stability, velocity component, vertical and lateral separation, height, frequency, and averaging period. Parameters fitted using one pair of measurement heights may not represent coherence across the full rotor disk.

The authors should add figures comparing measured and fitted coherence curves for several vertical and, where available, lateral separations. The frequency range used for fitting, uncertainty of the fitted parameters, goodness-of-fit measures, and sensitivity of reconstructed fields to the selected parameters should be reported. If one parameter pair cannot describe all separations adequately, multiple parameter groups or a separation-dependent formulation should be considered. The authors should also justify applying parameters fitted from four hours of observations identically to all three approximately 10-minute unstable cases.

> We agree that more nuance could be displayed regarding the Davenport fitting. But to that extent, more nuance could have been displayed for every single modeling input provided to benchmark participants (e.g., hub-height wind speed, shear, veer). However, we intentionally kept it simple. Benchmark participants were limited by how much effort they could put into their submissions, and the benchmark organizers did not believe that participants had the bandwidth to simulate flows across a span of Davenport parameters, nor across different spans of shear, veer, etc.
>
> We added the following paragraph at the end of Sec 2.1.1 to clarify "We note an important limitation of our study regarding constraint data: benchmark organizers provided participants with a single deterministic value for each input parameter (Table 1), without uncertainty bounds. Ideally, a validation study would characterize and propagate uncertainty in the model inputs, alongside uncertainty in the target data and in the models themselves. The organizers deliberately kept the inputs simple to limit the workload on participants, which might otherwise have discouraged participation.".
>
> Coherence values were calculated from sensors located along a single met mast, and as such, lateral separation calculation values are not possible.

10) The literature review should include recent induction-aware and extreme-turbulence studies

The literature review should be updated to discuss recent studies that directly address limitations relevant to this benchmark. These include nacelle-lidar reconstruction of freestream inflow from measurements inside the turbine induction zone, methods that

jointly estimate wind speed, shear, veer, and induction, and nacelle-lidar characterization of extreme turbulence for constrained-field generation and aeroelastic load analysis.

Relevant recent examples include:

- "Identification of wind inflow characteristics from nacelle lidar measurements in the induction zone of a 9 MW wind turbine," Renewable Energy, DOI: 10.1016/j.renene.2025.124523.
- - "Reconstructing the upwind field of wind turbines using LiDAR data," Sustainable Energy Technologies and Assessments, DOI: 10.1016/j.seta.2025.104382.
- - "Extreme turbulence effects on wind turbine loads: A case study for the North China Plain using nacelle lidar," Renewable Energy, DOI: 10.1016/j.renene.2026.125774.

The authors should explain how the present benchmark differs from and advances these induction-aware and extreme-turbulence reconstruction studies.

> Thank you for sharing these three interesting recent papers. We are confused to hear the assertion that many of these listed items (e.g., joint estimation of wind speed, shear, veer, and induction) are weaknesses of the current paper in your final Major Comment. The prior Major Comments made no such assertions, and these statements here have come out of left field. We are also confused why "nacelle-lidar characterization of extreme turbulence" is seen as a weakness of our study, as we intentionally focus on normal conditions.
>
> We omit a reference to the first paper because they calculate freestream wind speeds with a nacelle-mounted lidar at "maximum measurement distance of 200 m in front of the rotor", but our study calculates freestream winds from a nacelle-mounted lidar at a further distance of 390 m.
>
> We omit a reference to the second paper because it develops a lidar processing method specifically focused on real-time flow processing (<0.25 s outputs), whereas our lidar processing pipeline is used for inflow validation, which has no such requirements for a quick turnaround.
>
> We omit a reference to the third because we do not see a connection between their lidar measurements and our study.

MINOR COMMENTS

1) The conclusion that accuracy generally improves with model complexity should be qualified because complexity is confounded with inputs, resolution, and implementation.

Our intended connotation behind the word “generally” was “typically”. We have adjusted the text “Model performance varied between case studies, but we find that RMSE typically improved as model complexity increased.” (L499)

2) Explain why SSM uses an intermittency coefficient derived from unstable observations for stable and synthetic cases.

For the provided measurement data in this rotor inflow benchmark, we could determine the intermittency coefficient only for the unstable measurement period with sufficient statistical significance: The stable period time series extract was not long enough for an accurate determination, but recent analysis of field measurements under stable nocturnal conditions, although at a different site, report an intermittency coefficient of μ=0.252 [J. Lengyel et al., "The spatial organization of wind turbine wakes." *arXiv preprint arXiv:2512.14804* (2025)]. For the LES benchmark, we also did not attempt to determine the intermittency coefficient, as an eddy-viscosity-type subgrid model in LES ultimately reduces intermittency at small scales. Keeping the intermittency coefficient fixed across case studies also improves the interpretability of the results. We have added a short comment on this in the main manuscript in B6.

3) Report the SSM realization rejection rate associated with the 10% standard-deviation acceptance criterion.

The rejection rate for both stable and LES case studies was approximately 1/3, i.e., 2/3 of the realizations possessed a global standard deviation within 10% of the local standard deviation provided by the cases studies. We emphasize that this acceptance criterion could only be applied a posteriori, i.e., after constraining the “free” inflow turbulence boxes and not for the free inflow turbulence boxes themselves (see also [J. Friedrich et al. "Stochastic interpolation of sparsely sampled time series via multipoint fractional Brownian bridges." Phys. Rev. Lett. 125.17 (2020): 170602] for the precise correction).

4) Clarify whether machine-learning processing of the nacelle-lidar constraint introduces an LES-like spectral shape that could favor LER or other smooth reconstruction methods.

We added the sentence "We speculate that this change in power spectra shape does not offer a distinct advantage to any individual inflow model, and some benchmark participants chose to postprocess the constraint data to adjust the power spectra (Appendix B3)." (L145)

5) Explain whether the constraint preprocessing used for PCT2 and PCT3 is included when interpreting their performance relative to other PyConTurb submissions.

We added the statement "The benchmark organizers did not evaluate the impacts of constraint postprocessing on inflow quality." (L146)

6) Add uncertainty bars or ranges to the instrument comparisons in Figures 1 and 2.

Per our responses to Major Comment 2 and Major Comment 8, we omit this work.

7) Add the exact instrument coordinates and scan planes to the captions or a dedicated measurement-layout figure.

We have added the statement “Schematics and further details regarding observation processing (e.g., averaging windows, coherence parameter fits) are available at https://rotor-inflow-benchmark.readthedocs.io/en/latest/ as well as Doubrawa et al. (2025) and Houck et al. (2026). ” (L121)

8) Correct the duplicated wording "differences between hub-height differences" in the conclusions.

Corrected to "occurred due to differences in hub-height measurements between".

9) The reference "Hannesdóttir, 2026" is incomplete and should include full bibliographic information.

We have corrected the citation.

SUGGESTED DECISION STATEMENT

This manuscript presents an important and thoughtfully designed inflow-reconstruction benchmark using complementary real-world and synthetic datasets. However, the current model ranking is confounded by substantial differences in constraint information, preprocessing, spatial resolution, ensemble size, parameter estimation, and realization selection. In addition, uncertainty in the real-world reference may be comparable to the reported differences among models, while the primary accuracy metric considers only the streamwise velocity. The measurement geometry, induction correction, probe-volume effects, coherence fitting, and inter-instrument disagreement also require more complete documentation and uncertainty analysis. I recommend major revision to introduce fairer or stratified comparisons, quantify reference and ranking uncertainty, adopt ensemble-size-robust and multicomponent metrics, and qualify conclusions regarding model accuracy.