Community Comment

The manuscript addresses an important problem, namely, uncertainty in reconstructed inflow fields used for wind turbine model validation. The comparison of several inflow reconstruction methods using both real-world measurements and synthetic LES-based cases is a reasonable benchmark exercise. However, the paper is not fully convincing in its definition and use of “ground truth.” For real-world cases, the reference inflow is derived from lidar measurements, gridding, induction correction, time shifting, Taylor’s frozen-turbulence assumption, and other post-processing steps. The authors acknowledge this limitation, but the analysis still relies on the processed SpinnerLidar product to rank quantitative models. This weakens the significance of small differences among methods, especially given the reported discrepancies among instruments.

> We wholeheartedly agree that there are limitations with the "ground truth" in this study, and we do our best to alert readers to the shortcomings and to caveat our findings accordingly. “Each of these steps introduces uncertainty, and a more thorough accounting of errors can be read in Houck et al. (2026). 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 synthetic cases partly address this limitation by providing a fully known reference field, but they introduce a different issue. The benchmark then becomes partly a test of reconstruction performance in an idealized LES atmosphere rather than in the real atmosphere. The paper should more clearly separate conclusions supported by real-world data from those supported by synthetic cases. It should also give a sharper discussion of practical significance. Ranking inflow models by RMSE is useful, but the central question is whether these inflow differences materially affect turbine-load validation, controller validation, or design-code assessment. Since that connection is left mainly to future work, the present manuscript is incomplete.

> As you note, we ran benchmark analysis on an idealized LES atmosphere to remove certain sources of uncertainty associated with real-world reconstruction. As it turned out, the conclusions of the real-world analysis neatly align with the conclusions of the idealized LES analysis. The main discrepancies between real-world analysis and idealized LES analysis could be traced back to disagreements between the real-world constraint data (scanning lidar data) and target data (SpinnerLidar data), a type of misalignment that does not occur in an idealized LES

environment. The qualitative agreement between the LES analysis and real-world analysis adds confidence to the legitimacy of the real-world analysis, while still acknowledging the shortcomings there.

The paper is also longer and more descriptive than necessary. Some of the case setup, model descriptions, measurement details, and repeated discussion of inflow statistics could be compressed. The main text should focus more directly on reference-data uncertainty, robustness of model ranking, and the practical implications of the observed errors.

Thank you for sharing this perspective. Reviewer 2 has asked for more details in the manuscript. We did our best to strike a balance between sharing too many details and maintaining a tight narrative. We have offloaded many of the details to the benchmark website that is referenced in the manuscript.

Finally, the manuscript discloses that the lead author both organized the benchmark and submitted two models to it. This is a relevant point that should be clarified more fully, because benchmark organization and model participation are not fully independent roles. The manuscript would benefit from a short explanation of how impartiality was maintained, including whether the protocol, case definitions, evaluation metrics, post-processing procedures, and analysis scripts were fixed before model submission. Overall, I recommend a major revision.

This is a fair critique, and we have updated the text in the Competing Interest section.

"The lead author (Alex Rybchuk) both organized the benchmark and submitted two of the models (TurbSim and LER), which is a potential conflict of interest. Below we describe how the benchmark changed over time, and how we limited any advantage to these two models.

The benchmark protocol and evaluation were not fully fixed before models were submitted. The participants revised parts of the protocol together during the benchmark. For example, the synthetic LES target simulations were first planned at 10 m resolution, and after initial work the participants agreed to refine them to 1.25 m. The analysis in this paper also developed as results were examined, as is typical in research. It was not pre-registered.

The lead author did not tune either submission against the target data. The TurbSim inflows were generated using standard practice (Appendix B1), not procedures tailored to this benchmark. For LER, the main free choice is the training dataset. We took the simplest option: the training simulations use the same large-scale forcing

as the target LES cases (Appendix B7), and that forcing was not tuned. The only changes made to any submission after the initial analysis corrected clear errors, such as a reversed sign convention for veer. All participants were given the same opportunity to correct such errors after that initial analysis."