REVIEWER 2

This manuscript presents a well-structured and valuable benchmarking exercise that leverages the unique AWAKEN dataset and a carefully designed multi-phase framework. The progressive release of observational data is a clear strength, as it enables a controlled evaluation of how models respond to increasing constraint. The paper provides important insights into model behavior across a wide range of fidelities, and I broadly agree with the central conclusions as well as with the points raised by the other reviewer, particularly regarding the dominant role of inflow characterization and the need to distinguish inflow-related errors from wake-model physics.

We thank the Reviewer for their thorough and constructive review. We have addressed the specific comments in the rest of this document.

However, from a systems and industrial perspective, the manuscript would benefit from a more explicit distinction between predictive capability and state-conditioned calibration. As currently framed, reductions in aggregate error metrics such as mean absolute error risk being interpreted as genuine improvements in model skill. In practice, these improvements are largely achieved after the full release of inflow and wake observations, and therefore reflect the model's ability to adapt to a well-constrained, fully observed state rather than to predict it. This is particularly relevant in Phase 3, where a significant reduction in MAE is reported for several models. Such improvements are best interpreted as conditional error minimization rather than as evidence of generalizable predictive capability.

We agree that the performance gains observed in Phase 3 reflect state-conditioned calibration rather than generalizable predictive skill.

We have added explicit clarifying text to Section 4.2.:
"It is important to note that the performance gains observed in Phase 3 should be interpreted as conditional error minimization rather than evidence of generalizable predictive capability. Models in this phase had access to the full observational dataset for the case study day, including all SCADA power outputs and inflow measurements. The resulting improvements therefore reflect the models' ability to adapt to a well-constrained and fully observed atmospheric state, not necessarily their ability to forecast that state independently. This distinction is critical for operational applications, where such comprehensive in-situ data are unavailable a priori."

And also, later in the same section:
"While these gains are scientifically valuable for diagnosing model behavior and calibration potential, they are best understood as upper bounds on achievable accuracy under ideal observational conditions rather than as estimates of operational forecast skill."

Finally, we have also mentioned this aspect in the revised abstract:
"In subsequent phases, access to progressively richer observational data enabled model refinement that reduced mean absolute error by up to 40%; however, these gains primarily reflect state-conditioned calibration to a well-observed atmospheric state."

Closely related to this is the question of calibration versus transferability. In several cases, improved agreement appears to be achieved through localized tuning or spatially heterogeneous parameter adjustments that allow models to absorb unresolved physical processes, such as terrain-induced flow modification or stability effects, into empirical corrections. While effective for reproducing the specific case studied, this approach does not necessarily transfer to different atmospheric conditions. The manuscript would therefore benefit from a clearer conceptual separation between reconstruction of an observed case and prediction of unseen conditions.

This is another great point that should be explicitly mentioned in the interpretation of the benchmark results.

We have added the following discussion in Section 4.2:
"Closely related to this point, the strong Phase 3 performance of models employing spatially heterogeneous parameter adjustments stems from localized tuning that effectively absorbs unresolved physical processes -- including terrain-induced flow modifications and stability effects -- into empirical corrections anchored to this specific case study. While highly effective for reconstructing the observed day, the transferability of such parameter sets to different atmospheric conditions, seasons, or wind directions has not been established here and represents an important open question."

The reliance on a single diurnal case study further reinforces this limitation. The benchmark focuses on one specific atmospheric realization, characterized by low-level jet dynamics and stability transitions. While this provides a rich and controlled test case, it also introduces a structural risk of overfitting, as model adjustments may implicitly encode features of this particular event. As a result, performance improvements observed within the benchmark cannot be directly extrapolated to longer time scales such as annual energy production or probabilistic metrics like P50 or P90. This does not reduce the scientific value of the study, but it defines a clear boundary on the interpretation of model performance.

Agreed!
We have explicitly discussed this limitation (and need for future work) in the revised Conclusions:
"Furthermore, a structural limitation of this benchmark that must be explicitly acknowledged is its focus on a single diurnal case study. While this provides a rich and controlled test environment, it introduces the risk that model adjustments and calibrations implicitly encode features of this particular event. Consequently, the

performance improvements demonstrated here cannot be directly extrapolated to annual energy production estimates or other probabilistic long-term assessments. Future benchmarks should target multi-year, climatologically representative datasets to rigorously assess model transferability."

Another important aspect concerns the interpretability of the model ensemble. The manuscript acknowledges that subjective modeling choices, or the "human factor," contribute significantly to variability in results. However, the absence of a structured control across key configuration elements, such as turbulence parameterizations, boundary layer schemes, and mesoscale setup, makes it difficult to attribute observed differences in performance to underlying physical mechanisms. In its current form, the ensemble represents a collection of plausible configurations rather than a fully controlled experiment, which limits its ability to isolate causal drivers of error.

We agree with the Reviewer that this aspect had been left implicit in the original draft.

We have added the following sentence in Section 3:
"It is important to note that the ensemble presented here represents a collection of plausible model configurations reflecting participant best practices rather than a fully controlled experiment."

The identification of inflow characterization as a primary determinant of accuracy is one of the most important outcomes of the study. From an industrial perspective, this result has an even stronger implication than currently stated. It indicates that inflow uncertainty effectively defines a lower bound on achievable model accuracy, independent of wake modeling fidelity. Even in cases where turbines are minimally affected by wakes, substantial errors persist, suggesting that upstream atmospheric state representation is the dominant source of uncertainty. This insight should be elevated from a supporting observation to a primary conclusion, as it has direct consequences for pre-construction assessment workflows and uncertainty quantification practices.

We have replaced the existing concluding remarks regarding inflow conditions with a strengthened paragraph that explicitly identifies inflow uncertainty as the dominant source of error in real-world terrain-driven flows.

The second paragraph of the Conclusions now reads:
"A primary conclusion is that accurate characterization of inflow and terrain effects is as important as accurate wake modeling. The results highlight that increased model complexity does not automatically guarantee improved accuracy. In Phase 1, simpler engineering and steady-state models often matched or outperformed higher-fidelity approaches in terms of bulk error metrics for this specific case. This result should not be interpreted as evidence that simplified wake physics are superior to high-fidelity approaches, nor that models performing well on aggregate metrics are necessarily physically accurate. Rather, it reflects a fundamental asymmetry in inflow strategy:

engineering models were site-calibrated, directly ingesting the provided in-situ lidar observations, whereas mesoscale and LES models were required to predict the inflow from coarser reanalysis products (e.g., ERA5, HRRR), which failed to capture specific atmospheric features such as the exact timing and height of the nocturnal LLJ or the spatial inhomogeneity of the inflow. Consequently, high-fidelity model errors are primarily inflow errors -- limitations of the boundary forcing -- rather than errors in the wake physics themselves. Moreover, many simpler models assumed spatially homogeneous inflow conditions; while this assumption can yield favorable aggregate energy production estimates, it does not capture the site's true spatial heterogeneity, and agreement at the farm-integrated level may mask compensating spatial biases rather than reflecting genuine physical fidelity. Correctly characterizing freestream inflow conditions is therefore a critical -- though not sufficient -- step for accurate farm-level modeling. More accurate reanalysis products would substantially benefit the class of simulation tools that depend on them.
From an industrial perspective, this finding has a direct and strong implication: inflow uncertainty defines a lower bound on achievable model accuracy that is independent of wake modeling fidelity. Even for turbines with negligible wake exposure, substantial prediction errors persisted throughout all phases of the benchmark, indicating that upstream atmospheric state representation -- not wake parameterization -- is the dominant source of uncertainty in real-world terrain-driven flows. This has immediate consequences for pre-construction energy assessment workflows and uncertainty quantification practices: investments in improving inflow characterization, whether through denser observational networks, improved data assimilation, or better mesoscale products, are likely to yield significantly larger gains in prediction accuracy than equivalent investments in wake model sophistication alone."

We have also added an explicit sentence on this to the revised abstract:
"Overall, the study demonstrates that inflow characterization defines a lower bound on achievable model accuracy that is independent of wake modeling fidelity -- a finding with direct implications for pre-construction energy assessment workflows."

I also concur with the other reviewer that the benchmark exposes several fundamental research challenges, including the breakdown of classical boundary layer assumptions in stable conditions, the difficulty of generating realistic turbulent inflow for high-fidelity models, and the strong coupling between terrain-induced flow features and wake dynamics. Rather than reiterating those points, I would emphasize that these challenges collectively define limits on model transferability, not just areas requiring incremental improvement. This reinforces the need to distinguish models that reproduce a single constrained state from those that generalize across atmospheric regimes.

To ensure this perspective is clear, we have integrated this concept into the new concluding paragraph (added in response to Reviewer 1), emphasizing that these challenges dictate whether a model can truly generalize across diverse atmospheric regimes versus simply reproducing a highly constrained state:

"Looking ahead, this benchmark exposes several high-priority, fundamental research challenges that the wind energy modeling community must address. Crucially, these challenges collectively define limits on model transferability, not simply areas requiring incremental improvement. First, accurate inflow characterization remains the primary limiting factor: improved data assimilation methods, denser observational networks, and more accurate mesoscale reanalysis products are needed to reduce the 'inflow error floor' that this benchmark clearly demonstrates. Second, non-equilibrium boundary layer physics -- including low-level jets, stability transitions, and conditions where MOST breaks down -- demands improved turbulence closures that extend beyond current equilibrium assumptions. Third, generating realistic turbulent inflow for LES in heterogeneous terrain, particularly under non-periodic and non-neutral conditions, requires dedicated methodological advances. Finally, the multiscale coupling of terrain effects and wake dynamics must be treated as a first-order modeling priority rather than a secondary detail, as terrain-induced flow modifications can rival or even exceed wake losses in magnitude. Ultimately, addressing these physical and structural boundaries is essential to distinguish models capable of truly generalizing across diverse atmospheric regimes from those that merely reconstruct a single constrained state."

In conclusion, this is a high-quality and important contribution that should be published. However, the manuscript would benefit from a reframing that explicitly separates predictive capability from calibration-driven reconstruction, clarifies the limits of generalization inherent in a single-case study, and more strongly emphasizes inflow uncertainty as a primary limiting factor. Addressing these points would significantly strengthen both the scientific interpretation and the applicability of the results in operational contexts.

Once again, we thank the Reviewer for recognizing the value of this benchmark, and for providing constructive feedback. As detailed in our specific responses above (and to the comments from the other Reviewer), we have reframed the manuscript and we believe these revisions have strengthened both the scientific rigor of our interpretations and the practical applicability of these results for the broader wind energy community.