REVIEWER 1

This paper presents a highly valuable and extensive benchmark of wind farm flow models against the unique AWAKEN dataset. The multi-phase, blind approach is a significant strength, offering critical insights into model performance and the value of data for model improvement. The finding that inflow characterization is a primary prerequisite for accuracy is an important, well-supported conclusion. However, the framing of model performance, especially regarding the comparison between engineering and higher-fidelity tools, can be misleading and risks underrepresenting the fundamental research challenges that this benchmark uniquely exposes. Specific comments are as follows:

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

1. The manuscript makes statements such as "initial blind predictions showed that higher-fidelity models did not uniformly outperform simpler simulation tools" (Abstract) and "simpler engineering and steady-state models often matched or outperformed higher-fidelity mesoscale approaches" (Conclusions). While factually correct in terms of the bulk error metrics (e.g., MAE) for this specific case, this framing can be misleading without critical context.

We agree with the Reviewer that the Abstract and Conclusions lack sufficient context around this finding. Since this comment is closely related to Comment 2, which addresses the same underlying issue in greater depth, we discuss our specific edits in the response to that comment below.

2. The authors correctly note that engineering models directly ingested high-quality, single-point observations (from site A1) in Phase 1. Their performance is therefore not a triumph of simplified physics, but a demonstration of the effectiveness of empirical calibration against a known inflow. In contrast, the higher-fidelity models (WRF, LES) were tasked with a much harder problem: predicting the inflow from coarser boundary conditions. Their errors are primarily "inflow errors," not necessarily "physics errors" within the wake model. The text should more clearly distinguish between the performance of a model's inflow characterization strategy and its wake physics fidelity. Suggesting that an engineering model "outperforms" an LES model conflates a site-calibrated tool with a predictive one.

As stated above, we agree with the Reviewer that this is the most important framing issue in the original draft. We have revised the text to better highlight this aspect across the abstract, Section 4.1, and the Conclusions.

Revised text in the abstract:

"Initial blind predictions showed that higher-fidelity models did not uniformly outperform simpler simulation tools in terms of aggregate error metrics; however, this largely reflects differences in inflow strategy rather than wake physics fidelity -- simpler models directly ingested high-quality observations, while higher-fidelity models were tasked with predicting the inflow from coarser reanalysis boundary conditions."

Revised text in Section 4.1:
"It is important to note, however, that this comparison conflates inflow characterization strategy with wake physics fidelity. Engineering models benefited from direct ingestion of high-quality point observations at Site A1, effectively site-calibrating their inflow, while mesoscale and LES models were required to derive inflow conditions from reanalysis products without such grounding. The apparent underperformance of higher-fidelity tools is therefore primarily a consequence of the inflow specification challenge rather than pure inadequacies in their wake physics representation."

Revised text in the Conclusions:
"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."

3. The results of this benchmark expose profound, fundamental research challenges for high-fidelity modeling that are mentioned but not centered as key findings. The paper should more forcefully articulate these challenges as critical outcomes of the study:

We thank the Reviewer for this insightful comment. We agree that the fundamental challenges for high-fidelity modeling exposed by the AWAKEN benchmark should be centered as key findings. We have revised the manuscript to place greater emphasis on these issues, as detailed in our responses to the specific comments below.

4. The stable case (06:00 UTC) demonstrates the breakdown of Monin-Obukhov Similarity Theory (MOST) (as noted for Participant 6), placing the turbine rotor layer outside the surface layer. Standard RANS models, which rely heavily on equilibrium boundary layer assumptions, are fundamentally challenged by such non-canonical conditions (e.g., low-level jets). The manuscript should explicitly state and discuss the crucial need for improved turbulence closures for wind energy applications.

We agree with the Reviewer that the breakdown of MOST and the failure of equilibrium boundary layer assumptions are major takeaways from this stable case. While low-level jets can be considered canonical within idealized stable boundary layers (in this region), the specific, terrain-modulated dynamics observed during the benchmark represent non-idealized, real-world conditions that challenge standard RANS models. To explicitly articulate these modeling challenges, we have added the suggested discussion to Section 4.1.

"Steady-state and most RANS approaches generally struggled to reproduce the specific nosed-deficit profile shape characteristic of the LLJ. This difficulty is rooted in a fundamental limitation: the stable period was characterized by a very shallow atmospheric boundary layer, placing the turbine rotor layer largely above the surface layer to which Monin-Obukhov Similarity Theory strictly applies. Standard turbulence closures for both RANS and engineering wake models rely heavily on equilibrium boundary layer assumptions that break down under these non-idealized, real-world conditions. The PyWakeEllipSys RANS approach was able to maintain the nosed-deficit profile, though at heights above those observed, possibly because the inflow ABL height was set to a higher value compared to the inflow data to maintain numerical stability, as discussed in Sect. 3.6. Consequently, the results of this benchmark highlight a critical research need for the community: the development and validation of improved turbulence closures capable of resolving non-equilibrium dynamics and real-world low-level-jet interactions that are frequently encountered in operational wind energy environments."

5. LES is often driven by mesoscale simulations, which do not have turbulence content. This underscores the challenge of generating realistic, turbulent, site-specific inflows for LES, particularly in complex terrain where standard periodic boundary conditions or simplified precursor methods fail. It is suggested to discuss this issue in the revised paper.

We agree that the absence of resolved turbulence in mesoscale forcing poses a major structural hurdle for LES, particularly in non-idealized terrain.

To ensure this is recognized as a major limitation, we have expanded our discussion of LES performance in Section 4.1 to explicitly highlight this inflow generation challenge:
"The apparent underperformance of higher-fidelity tools is therefore primarily a consequence of the inflow specification challenge rather than pure inadequacies in their wake physics representation. In fact, LES models face an additional structural challenge: they are frequently driven by mesoscale model outputs that lack resolved turbulence content. Generating realistic, turbulent, site-specific inflow conditions for LES -- particularly for terrain-driven flows where periodic boundary conditions or simplified precursor methods are inadequate -- remains an open and important research problem. The variable performance of LES submissions in this benchmark partly reflects these inflow generation challenges rather than deficiencies in the LES wake physics itself."

6. The observation that terrain-induced flow acceleration frequently outweighed wake losses (e.g., at Site H) and caused specific waked turbines to overproduce is a critical finding. This demonstrates that resolving microscale terrain features is not just a detail but a first-order priority on par with wake parameterization. In LES of atmospheric turbulent flows, the near-wall region is often modelled rather than directly resolved. The classic wall model depends on the logarithmic law of the wall, which, however, fails in complex. This challenge should be highlighted in the paper as a fundamental research challenge.

We appreciate the Reviewer highlighting this detail regarding near-wall modeling. We agree that the log-law breaks down in these scenarios. We have added a discussion of this specific LES limitation to Section 4.1 to underscore that resolving these terrain-driven features is a first-order priority:

"Beyond the challenge of terrain representation, this result also exposes a fundamental difficulty in LES: the near-wall region over terrain is typically modeled rather than directly resolved. Classical wall models that rely on the logarithmic law of the wall are known to break down in terrain-driven flows such as the terrain-induced accelerations documented here, highlighting the critical need for improved near-wall models for LES in wind energy applications."

7. The conclusion section can be strengthened by distilling the key results from the points above into a clear summary of high-priority, fundamental research needs, including such as inflow characterization and modeling, non-equilibrium and non-canonical boundary layer physics, and multiscale coupling of terrain and wake effects.

We agree that distilling these takeaways into a clear summary can strengthen the impact of the conclusion. To explicitly outline these fundamental research needs (and also what noted by Reviewer 2) for the community, we have added the following dedicated paragraph to the end of Section 5 (Conclusions):
"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."