Reviewer 1

Dear Review 1,

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.

The motivation of the paper is three folded it is stated in the beginning of the paper, but there is no clear answer in the conclusion part of the paper. It would be an improvement if the main objective was clearer and not divided into three parts. The paper gives a very detailed list and figures of the differences, but brings in so many details that it is hard to follow. An overview in the conclusion part pinpointing the "common strength and weaknesses" stated in line 531 would be helpful so that it is clear what is found in the study. Here it is also stated that "context on how large inflow errors can be …" I don't think that there is inflow errors? Is this in the simulations or in the measurement or is it the differences? I think that there should put in a bit more work on the conclusion to show that the objective (or motivation) is responded to in a clear manner.

> Upon reflection, we opted to cut the text about the 3-fold motivation in the Introduction. As you noted, we didn't provide a clean resolution to these motivations in the Conclusion. We believe that the two guiding questions in the Introduction give sufficient framing (i.e., do different models produce substantially different fields, and how accurate are the reconstructed inflows?), and we added the phrase "motivating questions" to highlight that these questions are our main motivation.
>
> We used the phrase "common strengths and weaknesses" in paragraph 5 of the Conclusion. That was intended to be a nod to the preceding paragraph 2 (mostly successes) and paragraph 3 (largely misses) of the Conclusion. These preceding paragraphs summarize the main results of the study. To aid in clarity, we changed the sentence at the start of paragraph 5 to "To summarize, we identified common dynamics and quantified reconstruction errors among modern turbine inflow models."
>
> We adjusted the phrase "context on how large inflow errors can be" to "context on how large inflow reconstruction errors can be" to clarify that we are talking about the simulated inflows, not the ground truth inflows.

Following your feedback, we provide a more direct answer to our motivating questions in the Conclusion. The first paragraph previously read "In this paper, we leverage submissions to an international turbine inflow benchmark to probe the questions: Do modern inflow models produce substantially different inflows from one another, and how accurate are these models in our small number of case studies? We are interested in these questions in the context of turbine design code validation, both from a statistical perspective and on a second-by-second basis." and it now reads "In this paper, we leverage submissions to an international turbine inflow benchmark to study two motivating questions in the context of turbine design code validation. First, do modern inflow models produce substantially different inflows from one another? As summarized below, the answer to this question depends on the quantity of interest. Second, how accurate are these models in our small number of case studies? We quantified this error for both instantaneous and time-averaged streamwise velocity fields".

I find the paper to be very interesting, and think it would be of broad international interest, and the simulations and comparisons are of good quality. However, I am missing information on who are the participants? Is there a reason why the institutes are not mentioned? There are also a lot of times where “we” are used, and for clarity of the paper, the usage of “we” should maybe be omitted. An example is in line 161 “… participants reconstructed inflows …” and then in the next sentence “ .. we compared to data …” To the reader it seems like all the authors compared the data, and the participants are not a part of this group? For clarity try to keep “we” out of the Methods section.

We have updated Table 2 to specify which of the co-authors were responsible for which submission. Their corresponding institution can be identified on the first page under author information.

That is a good catch regarding the ambiguity of the word "we". We believe that most instances of using that word is sufficiently clear, e.g., "To investigate inflow reconstruction, we use two complementary datasets.". There were four instances in which "we" was used next to "participants" in the Methods section, and the text has been updated to clarify: (~L122, ~L154, ~L161, ~L232). Specifically, we changed "we" to "benchmark organizers", or we restructured the sentence to use the passive voice.

In many sections is referred to Supplementary Material. Where is this found? I also see that there is a to appendices, is it necessary with both Supplementary Material and appendices?

> The Supplementary Material should be available to you on the reviewer portal. Please reach out to the editor if you cannot locate it.
>
> The Supplementary Material and Appendices both contain additional information, but of a different variety. Our Supplementary Material primarily contains figures that are somewhat identical to those in the main body in the text, but the Supplementary Material shows data for other cases (e.g., Case U1500 was visualized in the main text, and we include visuals for Cases U1530 and U1600 in the Supplementary Material). Our Appendices contain some qualitative views of flow phenomena of interest (e.g., coherent flow features) and more details about each inflow model.

I also miss an illustration or information about at which height is the met mast measuring, what is the participants given as information (constraint information?), and exactly what measurements are the “ground truth”. Consider making some of the information tabular so it is easier to find.

> At the start of Sec 2, we added the sentence "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/." The website has schematics and details about the measurement processing that we deemed were too niche to include in the text of the manuscript. The schematics on that website also show the height of met mast measurements, the constraint information, and the ground truth information. We omit these schematics from the manuscript because we need two different schematics (one for real world, one for the LES world), and we believe that would take up too much space. We also added the statement: "The met mast had cup anemometers (Thies Clima First Class) and sonic anemometers (Ultrasonic RM Young) at heights between 52.6 m and 183.5 m." (LXX) to give some sense of the heights associated with met mast measurements.

In line 324 and onward the model ensembles and how often they fall within the ground truth is discussed, but there is no mentioning that the number of ensembles are different. Is this taken into account when discussing the results anywhere? In Figure 7 the black line is referred to as Spinner, should this have been ground truth?

> Yes the varying ensemble size is taken into consideration, and we state up front in Table 3 that different models provided different numbers of ensemble members. Near line 324, we also state "The narrower spread in PCT2 and PCT3 is likely due to the inclusion of 6 flow realizations, whereas the other models had between 15 and 30 realizations." Ideally all participants would have submitted 30 members (for statistical robustness), but that number was prohibitive for some benchmark

participants. We believe dropping all models to 6 ensemble members would have decreased some of the statistical power. Our compromise was to present all available ensemble members and discuss the effects of varied ensemble sizes where relevant.

Thank you for catching the inconsistent label in Figure 7. We have corrected it.

In line 454, failure modes are mentioned. Is this the correct use of words? Are the models failing, or are they missing some features that exists in real life? Is it possible for these models to generate the coherent gust or the veer, would it really help to run more ensembles?

The goal of the models to reconstruct real-world flow/high-resolution LES flow, and they fail to capture certain features that exist in real life/in LES. Given that, we believe the phrase "failure mode" is appropriate, as they fail to achieve aspects of that goal.

Minor corrections:

Define Davenport coherence parameters in line 153 and the parameters in line 155

We updated the text to add the missing definitions: "Benchmark organizers calculated Davenport coherence parameter values (IEC, 2019) for use by the Kaimal-based inflow models by fitting parameters to 4 hours of 1 Hz met mast measurements centered on the convective periods. The coherence model is

$$\mathrm{Coh}(r,f) = \exp\left[-a_k\sqrt{\left(\frac{fr}{U_{hub}}\right)^2 + (b_k r)^2}\right]$$

with *r* the separation distance in meters, *f* the frequency in Hz, and *U_{hub}* the hub-height wind speed. The IEC 61400-1 standard assumes *a_k = 12* and *b_k = 0.12 / L_c*, with *L_c = 8.1 \Lambda_1*. *\Lambda_1* is the longitudinal turbulence scale parameter. We calculate period-specific parameters for our convective benchmark (Table 1), and the best fit for parameters came from wind measurements at 110.5 m and 183.5 m." (~L155)

Line 177, Fig 1 does not contain a gap

Thank you for catching that. We removed "Fig 1" from that sentence.

Line 275, give a clear definition of the 2-dimensional wind angle

We rephrased the text as "As a result of the differences in v, the wind angle fields in the plane parallel to the ground..."

Line 292 Based on previous

We changed "Based off previous" to "Based on previous"

It would be helpful with grid lines in figure 14

We have added grid lines to Fig. 14