Response to the reviewer

We thank the reviewer #2 for their useful comments and the time invested in reviewing our manuscript. We have addressed each of the referee comments as detailed point by point below, which we believe has significantly improved the quality of the manuscript.

Main Comments

Reviewer Point P1 — The topic addressed by the authors is very relevant to the wind energy industry and their "skill/cost" idea is certainly interesting. However, there is a lot of missing information about the organization and structure of the benchmark, making it difficult to assess if the model's comparison and its findings are robust. There are also some caveats described below about the scoring process definitions that prevent me to recommend the publication of this work without some major revisions that can provide more confidence in the conclusions drawn in the paper in its current form.

Reply: Regarding the point "There are also some caveats described below about the scoring process definitions", we have decided to restructure the paper accordingly. The new goal of the work is now "to build a simulation database that will help to develop and improve WRA decision tools". We have adjusted the introduction and the analysis section accordingly and the paper now focuses on the difficulties and learning outcomes for WRA decision systems, which could be used in the future. Furthermore, additional information about the "organization and structure of the benchmark" have been added to the article.

Reviewer Point P2 — First, I believe that the definition of the skill scores given "before" the simulations are not very clearly described in the manuscript. The reader is pointed to a previous work by the authors (Barber et al., 2022b), which was concluded with several suggested improvements and pending work related to the scoring conceptualization. However, those points don't seem to be implemented for the analysis of the present study, or at least this manuscript doesn't explain how the subjectivity in many of the scoring definitions can be mitigated.

Unlike the skills scores, some components of the cost scores are indeed more "quantifiable" (equation 1). Still, the values of some of them like the "cost of the staff training per project" can again be very subjective to the interpretation of the modeler. Other parameters such as the "hourly rate of the modeler" depend on the institution, country, etc. The cost scores assigned in this comparison can be biased towards the participants from countries with lower wages.

Reply: We have improved the description of skill scores "before" in Section 2.2. We haven't made the improvements suggested in the previous paper because the work was actually carried out in parallel. However, we have now restructured the paper with a goal of "to build a simulation database that will help to develop and improve WRA decision tools", rather than to present a fully-functioning decision-making method. This has allowed the difficulties with the method to be brought up in the analysis (including the ones you mention here as well as ones brought up by the other reviewer). We have adjusted the introduction and the analysis section accordingly. The analysis now avoids specific recommendations regarding the suitability of each model; rather, it now focuses on the difficulties and learning outcomes for WRA decision systems, which could be used in the future.

Reviewer Point P3 — On the other hand, besides the challenge of finding an optimal skill/cost model, equally relevant for the wind resource assessment community is the correct usage of the selected model, especially those models very sensitive to the user's expertise such as some research codes included in this comparison. In that regard, I find many technical details missing about the models' setup (see the specific comments), and for those details that are included, it is noticeable the important differences in their configurations beyond their different physics/numerics. For instance, one model includes the effects of the turbine wake (PCE-LES) while another one employed a long-term correction from ERA5 as input wind field (WAsP). Others included atmospheric stability (E-Wind) while others used time series instead of period averages (E-Wind, Fluent). While this is fine in benchmarking exercises such as the CREYAP series (Mortensen et al., 2015), it complicates the goal of the authors of finding an optimal WRA tool because it is very difficult to conclude if the differences in the skill (and costs) among the workflows are related to the model, its configuration or simply the methodology to compute the APE beyond simulating more or less directional sectors. Besides, this mix of factors would probably preclude extrapolating the optimal workflow found in this case study to other sites.

Reply: We agree that the correct usage of the selected model is highly relevant to this project, which is why "user skill" is included in the skill score. We have improved the description of this point in Section 2.2 and included a discussion in the analysis in Section 3.1.2.

Regarding the technical details of the setups, we completely agree with your point. However, in order to assess this correctly, we would have to carry out parameter studies for thousands, or millions, of combinations of different parameters, which was beyond the scope of this study. Instead, the idea was to let the users apply their models however they want, in order to demonstrate and highlight how these differences could impact the results. We have added a discussion about this in Section 2.3, see line 172. As well as this, We have added more details, links to further resources and citations describing the setups and hope this is sufficient to reproduce the simulations.

As stated above, the goal of this work was to apply and provide further results of the methodology developed in Barber et al. (2022b) by presenting the obtained results. These results can hopefully help improve the methodology in upcoming work.

Reviewer Point P4 — The description of the case study is also a bit superficial in section 2.1. In addition to the general description of the site and the measurement campaign, I would suggest that the authors could add more information about the met mast data preparation, filtering and selection. Useful information could be the data availability by direction from the three met masts and comments about the potential effects from the wake of the operating WTG on the mast measurements. The wind rose shown in Fig. 2. indicates less frequent but still important SW winds, for which the wake of the WTG is expected to occur.

Reply: As there is an extensive paper by Fernando et al. (2019), describing the measurement data and the flow phenomena of the site we restricted the description within this paper to the most important points. Furthermore, the used met mast data is referenced in the data availability section. No further data preparation or filtering was carried out. However, some additional information regarding wake effects and data availability have been added to the paper, see line 95.

Reviewer Point P5 — It would be also great if the authors could add information about the atmospheric stability during the 3-month period considered for the benchmark. This information

might help to explain some of the large errors obtained by some of the models. It is already interesting to see that mast 29 has a large deviation from the (neutral stability) logarithmic profile, whereas mast 25, located on the lee side of the hill, thus, with a more complex flow, has an excellent logarithmic fit.

Reply:

We agree that the thermal stability plays an important role in flow over non-flat terrain. A few paragraphs addressing this have been added to the text in Section 2.1.

Further Comments

Reviewer Point P6 — line 56-60: The authors are mentioning the analysis carried out by Barber et al., 2020c. but this work is only available as WECD since that paper has been withdrawn.

Reply: Reference has been changed to a published paper.

Reviewer Point P7 — line 72: By previous works are the authors referring to the Barber et al., 2022a-b articles instead of the Barber et al., 2020a?? Only the first ones compare several modeling tools on different sites.

Reply: Yes, the Barber et al., 2022a-b articles are referred to. An additional citation has been added to clarify.

Reviewer Point P8 — line 81: This citation also points the reader to the withdrawn paper of Barber et al., 2020c. Wouldn't it be "Barber et al., 2022b" the right citation in this case?

Reply: Yes, the Barber et al., 2022a-b articles are referred to. An additional citation has been added to clarify.

Reviewer Point P9 — As mentioned above, this section lacks many details about the model's technical setup. In the case of the RANS and LES-based models, information about their boundary conditions, especially the treatment of the ground, is critical for understanding and potentially allowing the repeatability of this work. Also very important is providing the values of the different models' constants used by the RANS models.

Reply: Where possible, the citation of work with further details are provided and more information was added for the different models. Furthermore, the simulation results themselves were made available on Zenodo `https://doi.org/10.5281/zenodo.8005621`.

Reviewer Point P10 — -the OpenFOAM model is set to match the wind speed and direction at the calibration mast at 100m, (line 212). Aren't all the other models set to match the wind at 80m (line 117)?

Reply: Thank you for noticing. The sentence stating that the models were set to match the wind speed at 80m was misleading and has been removed. More details about the exact calibration method for each model is given in the respective model section. The calibration process is highly dependent on the used tool.

Reviewer Point P11 — Line 223. Change "The direction variable" for "the wind direction". And, where is this wind direction obtained from?

Reply: The wind direction was obtained based on the given met mast data, providing the 'U' and 'V' wind speed components. This has been made clearer in the text.

Reviewer Point P12 — Line 223. Change "Wind speed variables" for "wind speed components"

Reply: Has been corrected.

Reviewer Point P13 — Line 227. I think that the phrase "constant line values for some variables" is not clear about which variables are they referring to, and what "constant line" means in this context.

Reply: Sentence has been changed to be more clear.

Reviewer Point P14 — Line 246. Does the E-Wind workflow simulate 24 wind directions? or 12 as described in Table 1?

Reply: Thanks for noticing. 24 wind directions were simulated and Table 1 has been corrected.

Reviewer Point P15 — Figure 4. Despite that is it expected to have some small errors in the RANS and LES models, shouldn't the WAsP model have no error at the calibration mast at 80m due to the way this model works??

Reply: No errors at the calibration mast are very difficult to obtain, because we are comparing simulated data with measured data, although this is at the same point. WAsP performs internal corrections, especially in complex terrain, which can slightly vary the results, as in this case.

Reviewer Point P16 — Line 364. Is "respective met mast" the calibration mast?? This phrase is a bit confusing, so, it is not very clear how this normalization is done.

Reply: This was done for each met mast, i.e. 20, 25 and 29. This has been clarified in the text.

Reviewer Point P17 — Figure 8. Is this the AEP at 80m height? Again, I'm not sure if I got how the AEP normalization was defined.

Reply: Correct, this is the AEP at height 80m. The AEP was normalised by diving the total AEP of the simulations at a met mast by the calculated AEP based on the measurements at that met mast. The clarification mentioned in the previews point hopefully helps to understand this better.

Reviewer Point P18 — Figure 11. It seems that the mast 20 is missing.

Reply: Unfortunately, the simulation results for the AEP at met mast 20 are not available. This is due to the original design of this case study, with the original goal to only compare met masts 25 and 29 for the AEP predictions. This information has been addded to the text, see line 394.