Response to Referee #2

Thank you for the comprehensive comments. Your comments, together with those of referee #1, led to a thorough revision of the paper.

The most general comments regarding the revisions to the manuscript are:

1. At the start of the research project typically there are high expectations placed on the sensitivity experiments, however, reality always brings some corrections and caveats. Given the enormous possibilities in setting up WRF, an "optimal" configuration is unreachable. We have tried to revise the introduction to convey that the paper focuses on finding the "best possible" model configuration **constrained** by the practical issues in running the model simulations and the ultimate goal to use the simulations for a **wind atlas**.

2. The manuscript aims to tell the story of how the NEWA wind atlas came to be. Therefore, further analysis of the model results will make the flow of the paper less clear. We have tried to enhance this structure in the revised manuscript.

3. We have replaced some of the figures (6, 9–12, 13) to homogenise the analysis of the results. We have also added new figures including the RMSE and circular EMD for wind direction.

4. We strengthened the connection to the companion paper, https://www.geosci-model-dev-discuss.net/gmd-2020-23/, which is now available.

The reviewers' comments are in black and our responses in blue.

General Comments

1. The paper summarizes an exhaustive sensitivity analysis performed to inform the final model setup of the New European Wind Atlas. This surely must be the most extensive such analysis to date and overall is an impressive achievement. The novel use of the Earth Mover's Distance is also applauded and clearly offers a much-needed complimentary metric alongside the typical timeseries-based performance metrics.

   Thank you. As described above, we have expanded the use of the EMD and Circular EMD (CEMD) for wind direction in the manuscript.

2. I believe this paper should ultimately be published; however, I have several comments and concerns about the work that have not been addressed in the paper. First, all of the critical validation was performed in Northern Europe, despite the NEWA being produced for Europe and Turkey as a whole. I realize that computational expense and data availability/quality were probably a factor, I can't help but feel that with such collaboration across European institutes that

a more regionally diverse validation campaign could have been performed. Of course NEWA has already been produced, but I think some critical commentary on how validation in Northern Europe (with its unique climatology) would apply across other climates in Europe with their own unique climatologies is needed here. Otherwise, the paper reads as if the idea of more extensive validation was overlooked.

In the second part of this study [1], the final wind atlas is validated against masts over all of Europe. However, at the time we did the sensitivity simulations and we needed to decide on a final configuration, further evaluation with data besides the 8 sites in N. Europe was not possible. The public data from tall masts needed for evaluation are scarce over Europe. In a recent paper and database [3], where a global database of tall masts was compiled, there is only a handful of mast over Europe where data are available and only a couple lie in the region chosen for the sensitivity study. Further, even if the data used in part 2 would have been available, the evaluation would not have been possible. There are no masts in Denmark and very few in Germany and the masts in Poland have data for only a few months in 2015. (see Figure 4 of Part 2)

3. Furthermore, I did not find sufficient presentation of results to justify selection of the final model setup. Rather, a wind profile plot and two heat maps of bias and EMD were provided, and it seemed very quickly the section was wrapped up with the final model selection. I think some further synthesis is required, such as a table of figure showing mean bias, RMSE, EMD, etc. across all validation sites. Without this, in my opinion, the selection of the final model setup seems unjustified.

   Agreed. The new manuscript includes further figures with all the statistics including BIAS, RMSE, EMD for wind speed and CEMD for the wind direction.

4. Finally, as far as I can tell, ERA-interim was used in the sensitivity analysis, but ERA-5 was used in the final production run. This point is not discussed in this paper but I think it's an important one. Does existing research suggest bias or EMD differences between the two data sets? If so, what are the implications on selecting the best model setup using one large-scale forcing but pivoting to a new product for the actual production runs?

   There seems to be some confusion here and we will clarify the data used in the manuscript. The sensitivity simulations described in sections 5.1, 5.2 and 5.5 used ERA-Interim as forcing. All other simulations (excluding the simulation named "ERAI" in section 5.4) used ERA5 data. Actually the sensitivity test of replacing ERA5 with ERA-Interim is described in Table 5 and the results are shown in the heatmap plots. The differences in BIAS and EMD (figures 11 and 12 in the manuscript) show very small differences. Since ERA-Interim was scheduled to be discontinued in 2019, we chose to continue our sensitivity studies and production run with the ERA5 dataset.

5. In conclusion, I think this is a valuable contribution to the literature. However, several key limitations of this study need to be sufficiently addressed and discussed before final publication. In addition, a couple summary figures and tables would help justify final model selection.

   Thank you. We believe the document is much improved after this round of revisions.

Specific Comments

1. Page 1, Line 9: Why were sensitivity experiments only conducted in Northern Europe when the data set was for Europe as a whole? Surely tall masts must be available elsewhere? If this was a decision based on computational restrictions, this should be stated and the implications of this smaller validation domain, in the context of regional wind climates, should be discussed.

   The sensitivity analysis was done only for the domain over Northern Europe. As mentioned in our answer to item 2 above, tall masts of good quality data publicly available are very scarce. The implications are the focus of section 5.1, where we argue that the behaviour of the mean wind speed relative to various PBL/SL parameterisations is similar among the five domains in very distinct wind climates.

2. Page 2, Line 15: Can 'linearized model' be described more, or at least a couple references listed to provide background?

   The paragraph where this statement appears has now been rewritten. We refer here to the whole wind atlas method, which is now described in a little more detail. Therefore, "linearised model" is now "linearised method".

3. Figure 1: As in comment in Line 9, validation only in Northern Europe poses a problem for a product that covers Europe as a whole. This key study limitation needs to be discussed in detail.

   This is now better explained in section 5.1. It is a limitation of the study, but we had no alternatives. In retrospect, the validation in paper 2 [1] shows that the resulting wind atlas provides good estimates not only in N. Europe, but also in other regions.

4. Table 1: What is the time resolution of the observed data used to indicate sample size? I'd assume hourly but please make this clear.

   We have added a description of the temporal resolution of the data, which was 10-min means that were filtered to hourly using the period closest to the top of the hour. Additionally, Table 1 has been extended to include wind directions, and show the data availability as a percentage rather than number of samples.

5. Page 6, Line 9: Given the known impact of turbine wakes at these measurement sites, why not filter the data by wind direction to ensure the data are free stream? Especially in such a detailed sensitivity analysis where performance metrics between different model setups can be on the order of 0.1 m/s, allowing wakes to affect the measurement data seems inappropriate.

   We now explain in the text that the impact of the wind farm is difficult to quantify. For example, at some of the sites, wind farms were being built and tested in 2015, and without operational data, we cannot know when a wind farm was curtailed or otherwise not operating. Additionally, filtering for the wind farm possible perturbation can severely decrease the number of samples for some sites. We have now added the centre of the filtering wind direction and the fact that there is an additional wind farm near the FINO2 mast. The text in the revised manuscript has been changed to "...the data has not been filtered or corrected for the turbine wakes. However, the presence of the wind farm can impact the evaluation of the model results and should be kept in mind."

6. Page 7, Line 14: I'd use 'interpreted' rather than 'understood' when describing EMD as a measure of physical work.

   Agreed. We have replaced "understood" by "interpreted"

7. Page 7, Line 15: Given the novelty of the EMD metric, I wonder if a new Figure showing the area between cumulative distribution functions would be useful, given this is how the metric is actually computed.

   Agreed. This is a very good suggestion. A new panel has been added to Figure 2 showing the cumulative distribution functions.

8. Page 7, Line 16: What are circular variables and why are they relevant here? Are you validating wind direction?

   Circular variables are variables, like wind direction, where there is an apparent discontinuity at between 0 and 360°. In the updated manuscript we use a version of EMD metric adapted for circular variables (CEMD), which was used to evaluate the wind direction distributions in the model simulations.

9. Page 8, Line 23: Why was WRF 3.6.1 used, given it is 6 years old and the significant advances made since then? Was this part of an older study that is now being published?

   At the start of the project (summer 2015), WRF V3.6.1 was not that old (it was released August 14, 2014) and it provided good evaluation against observations in other regions, for example South Africa [2]. Using the latest version of a model is not always advantageous as seen in the changes to the MYNN parameterisation in WRF V3.8.1, that heavily impacted the validation statistics. Later in the large ensemble we moved to WRF V3.8.1. We acknowledge that the model version used in the various simulations was not clearly stated. This situation is fixed in the revised manuscript.

10. Page 10, Line 3: But MYNN winds are higher in the NW offshore domain and lower in the SW domain. Can you discuss? Is NW offshore domain generally more stable?

    What was meant by the statement was that normally the winds in the YSU scheme are larger than those in the MYNN scheme (see Figure 1 in the answer to the comments from reviewer #2). But when conditions are mostly unstable, as it is in the French Atlantic coast (50–60% of the time), Mediterranean sea (60–70% of the time) and some coastal areas Turkey, the situation reverses and the 100m mean winds are higher in the simulations using the MYNN scheme than the YSU scheme. Yes, conditions are mostly stable or neutral over the North Sea and the Baltic Sea. The sentence in the text has been expanded to clarify this issue.

11. Figure 6: Given the detailed justification of EMD earlier, why is it not being used here?

    Totally agree. At the time that these analyses were originally made we had yet to discover the advantages of the EMD metric. But now we show this metric throughout the manuscript and also in Figure 6.

12. Figure 8: I'm struggling trying to distinguish the different model runs. Multiple setups seem to have identical markers (at least to the naked eye). Also the lines are so tightly clustered that it's generally not possible to discern one profile from another. As such the Figure does not provide much useful information and I would recommend revising or deleting.

    The objective of the figure was to show that the wind profiles from the simulations clustered, not to be able to differentiate between them. We have redone the figure with a single grey colour, highlighting only the results from two relevant simulations. Hopefully it will reflect better our intention.

13. Figure 9a: Would an additional column showing average across sites be useful in identifying the best performing model setup?

    Thanks. It is a good suggestion. However, in this case we would argue against adding the averaging over the stations. Please see the long discussion in the response to Referee #1, item 25 (P18 L5). We don't want to give the impression that the decision on final configuration was just based on a raw evaluation of the numbers. Adding the average over the stations will convey that impression in our opinion.

14. Figure 9b: I'm not sure I see the value of performance metrics relative to the 'base' setup. In my mind this base setup is just another member of the ensemble

and not otherwise special. So why compare all ensembles against this one? Do we know it to be the most accurate? If not, I don't see the value in this relative comparison. Please justify.

Thank you, this is an important question. As we searched for the "best" model configuration, we kept asking "Is there another different configuration that will be better than our base?" The relative heatmaps help answer that, while also showing how small the differences between simulations are, which is sometimes hard to spot in the BIAS or EMD plots alone. This is because the differences between the stations are often more pronounced, for absolute values of metrics, than the differences between the ensemble members at the same station. It is important to note, that this method of examining results does not assume that the "BASE" setup is the most accurate, it is just a more convenient way of identifying differences between different models.

15. Figure 10b: Likewise to comment above. I'm not seeing the value of this relative comparison.

    Please see our reasoning above.

16. Page 18, Line 5: This is a big jump to conclude the best performing model setup based on the figures shown in this section. For example, the improved performance of MO over the Base and MM5 setups isn't clear from the profile plots or the heat maps. I think some final figure or table is needed showing key performance metrics averaged across all sites in order to justify this model choice. It also seems that the multi-physics sensitivity analyses and the selection of final production run in Section 5.3 was done using ERA-interim as the large scale forcing in WRF. However, ERA-5 was used in the final NEWA. This seems problematic given potential differences (e.g., biases) between the two data sets. I understand that ERA-5 was not available at the time these simulations were performed; however, some discussion around the implications of changing the large scale forcing without sensitivity analysis needs to be provided.

    Agreed. It is a very fair question. In conclusion the simulations show that many parameters usually though to be important for NWP or climate modelling have little or no influence. So it would probably be fair to choose any of them, except for some PBL/LS/LSM combinations that definitely degrade the validation metrics. A long discussion on the matter was given in the answer to referee #1 (item 25, P18, L5). In the revised manuscript we try to convey this in a more direct manner.

17. Page 19, Line 8: Unclear how ERA5 reanalysis slow down of winds relates to a sensitivity analysis of ERA-interim, FNL, and MERRA2. Was ERA5 part of this comparison?

    All experiments in these tables used ERA5. The comparison is then from ERA5 to MERRA2, FNL, and ERAI. As the table reveals the differences between ERAI and ERA5 are small.

18. Figure 11a and 12a: What is the difference between BASE and ERAI? I thought the base run was done using ERA-interim.

    All the experiments, except for “ERAI”, were carried out using ERA5.

19. Figure 11b and 12b: Same comment as previous.

    Same response as above. Sorry about the confusion.

References

[1] Martin Dörenkämper et al. “The Production of the New European Wind Atlas, Part 2: Production and Validation”. In: *Geosci. Model Dev. Discuss.* in review (2020).

[2] Andrea N. Hahmann et al. *Mesoscale modeling for the wind atlas for South Africa (WASA) Project*. Tech. rep. DTU Wind Energy, p. 77. URL: http://orbit.dtu.dk/services/downloadRegister/107110172/DTU_Wind_Energy_E_0050.pdf.

[3] J. Ramon et al. “The Tall Tower Dataset: a unique initiative to boost wind energy research”. In: *Earth System Science Data* 12.1 (2020), pp. 429–439. DOI: 10.5194/essd-12-429-2020. URL: https://essd.copernicus.org/articles/12/429/2020/.