In this document, the reviewer's comments are in black, the authors' responses are in red.

The authors thank the reviewer for their thoughtful and productive comments.

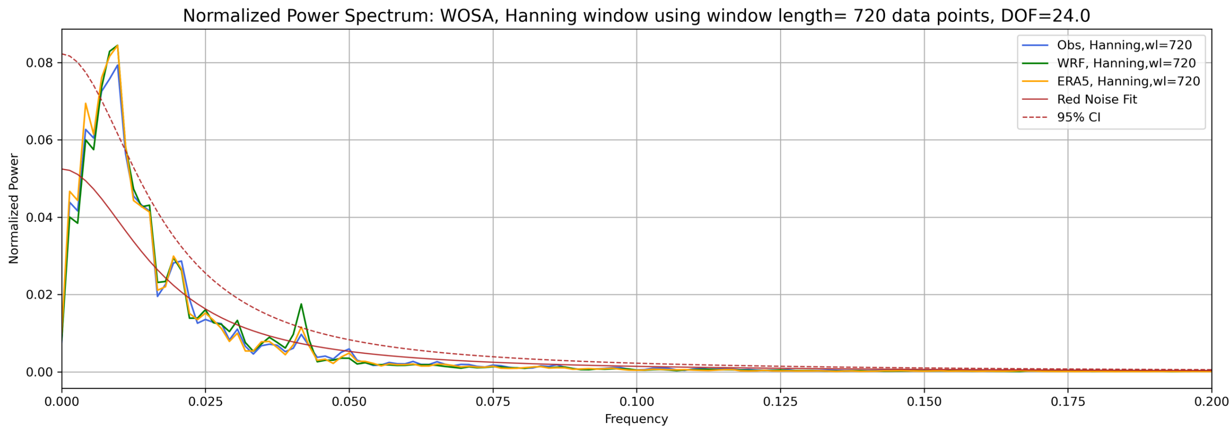
Summary

This study investigates the question whether a current state-of-the-art re-analysis product ERA-5 is sufficiently good to replace mesoscale models for wind resource assessments in simple terrain. Although the study doesn't provide a definite answer to the question, it provides a good contribution to the scientific community dealing with these type questions. The manuscript is well written and well structured. The figures shown are well prepared highlighting the most important results. Overall I recommend publishing this manuscript with some minor revisions.

Comments

I find the methods used appropriate for this type of study. However, part of the analysis could be summarized in a Taylor diagram (using the cRMSE). This has the benefit of adding the standard deviation to the evaluation in a format which is easily evaluated graphically. This metric is otherwise not analysed. So I would like to see either adding the standard deviation to the analysis separately or included in a Taylor diagram.
While we agree that Taylor diagrams are an effective way of summarizing multiple metrics on a single plot, the fact we are analyzing the performance of WRF and ERA-5 with height would require a large number of Taylor diagrams, so we prefer to stick with the vertical profiles of each error metric as in Figures 5 and 6.
Instead, we have added the standard deviation as an additional performance metric as suggested. We have updated Figures 5 and 6 (and the corresponding ones in the SI), and the text in many parts of the paper. In summary, WTK-LED outperforms ERA-5 at the offshore site, whereas no clear winner is observed at SGP, with similar considerations holding in all stability conditions.

The authors study the diurnal cycle in more detail and shows that the WRF simulations yields a larger diurnal variability compared to observations, whereas ERA-5's diurnal variation is underestimated. Variations on additional time scales could also be added to the analysis by e.g. computing the Fourier spectrum for the time series for the three different datasets and the two sites. Please consider this in the revision.
The following plot shows the spectrum at the E05 lidar location. The peaks that are statistically significant are the 24-hour one (already analyzed in the paper), and a region corresponding to 4-6 day timeframe, which might be connected to the impact of weather systems. As such, it would be hard to calculate an "average weather-related cycle" from the various data sources to build a new plot parallel to Figures 7 and 10.



Minor comments

L42, Molina reference lacks year
We have added the year (2021) to the reference.

L47, Please add a few sentences commenting the Sheridan (2020) results here. Similar studies has also been performed for the North Sea and the Baltic Sea (Kalverla et al 2019, Wind Energy Sci. 2019, 4, 193–209, Hallgren et al. 2020, Energies, 13, 3670; doi:10.3390/en13143670).
We have added the following sentences: "Sheridan et al. (2020) recently validated three reanalysis products using data at one single height from a floating lidar in the U.S. Eastern Seaboard, and found that ERA-5 had the best performance out of the four considered reanalysis products. Similar validations have been performed in Northern Europe, especially focusing on low-level jet events (Kalverla et al., 2019; Hallgren et al., 2020)."

L93, Please elaborate on the sensitivity of assuming w=0 for the horizontal wind sped estimate using this assumption? Will this be significant e.g. during strong convection?
Thanks for catching this. As shown in Werner et al. 2005, assuming w = 0 is actually not necessary in this context, so we have removed that sentence. On the other hand, assuming horizontal homogeneity is necessary, but this is expected to mostly impact wind speed retrievals in complex terrain (see for example Bingöl et al. 2008), and not offshore or in flat terrain as considered in our analysis.

- Werner C. (2005) Doppler Wind Lidar. In: Weitkamp C. (eds) Lidar. Springer Series in Optical Sciences, vol 102. Springer, New York, NY. https://doi.org/10.1007/0-387-25101-4_12
- Bingöl, F., Mann, J., & Foussekis, D. (2009). Conically scanning lidar error in complex terrain. *Meteorologische Zeitschrift*, *18*(2), 189-195. https://doi.org/10.1127/0941-2948/2009/0368

L95, Not sure I understand the details here. Is it correct that you get 1 wind speed sample from the lidar every 15 minutes? The hourly estimate is then an average of 4 15 minutes estimates?
Yes, this is correct.

L107, definition of near neutral: with your definition this leaves near neutral to L=0 or L>200 or L<-200. Normally, you would define a range around z/L=0 (see e.g. Sorbjan and Grachev. Boundary-Layer Meteorol 135, 385–405 (2010). https://doi.org/10.1007/s10546-010-9482-3). Please comment and revise

We have now included L=0m in the stable classification (however, that never occurred in the datasets, so the analysis itself did not change). For the rest of the classification, this is consistent with the "range around z/L=0" you pointed out.

L113, What type of lidars where deployed at the offshore location and how was the wind speed evaluated from these? Did you also here get hourly average?

We have added the following information: "The New York State Energy Research and Development Authority (NYSERDA) recently deployed ZephIR ZX300M lidars and made their data publicly available \citep{nyserda2020}." and "The lidars measure at a nominal 50-Hz resolution, and observations are provided as hourly averages, after proprietary quality checks are applied to the data."

L134, Please comment on the different model setup for the land and offshore location

We have added the following comment: "All the main setups that have been shown to have a major impact on modeled wind speed (e.g., the choice of the planetary boundary layer scheme and of the atmospheric forcing) are the same between the offshore and land-based domains. For some other setups, different choices were made between the two domains in order to optimize and tailor the numerical simulations to the specific needs of each domain."

L263, strictly, the conclusions part is more written as "summary and conclusion".

We have changed the title of this section as suggested.