Referee #1

Review of "Evaluating the ability of the operational High Resolution Rapid Refresh model version 3 (HRRRv3) and version 4 (HRRRv4) to forecast wind ramp events in the US Great Plains" by Bianco et al.

The manuscript evaluates the ability of the High Resolution Rapid Refresh (HRRR) numerical weather prediction model in forecasting wind ramp events in the U.S. Great Plains. The study focuses on two operational versions: HRRRv3 and HRRRv4. Utilizing the Ramp Tool and Metric (RT&M), it demonstrates that HRRRv4 outperforms HRRRv3 in skill, particularly in detecting up-ramp events during summer, which is vital for wind energy integration into the electric grid. The methodology includes using 10-m observational data from METAR stations and model outputs, with statistical analyses carried out for annual and seasonal variations.

The work is timely, addressing the critical need for reliable wind energy forecasting in the context of renewable energy integration. However, some areas require additional clarity or justification. The manuscript is well-written and accessible to a broad audience in meteorology and renewable energy fields.

We thank the Referee for the thoughtful and detailed comments. We hope we have addressed all of the Referee's concerns and we think that our manuscript did benefit from the constructive comments made by both Referees. In the following text, the Referee's comments are in black and our answers are in red.

Major Comments

1. While the manuscript provides a broad description of RT&M, it would benefit from a clearer explanation of its mathematical implementation. How is the skill score derived for different ramp characteristics (e.g., timing, amplitude, and duration)? Reference to specific equations (e.g., Bianco et al., 2016) is helpful but insufficient for readers unfamiliar with RT&M.

   More specifics on the way the skill score of the model at forecasting wind ramps is computed are included in the revised version of the manuscript (Section 2).

2. The decision to use 10 m wind speeds is justified by correlations with 80-m data and data availability. However, as acknowledged by the authors, this introduces uncertainty, particularly when converting the wind speed into power generation – small changes in wind speed can result in large changes in wind power and associated ramps. This limitation should be discussed in more detail. Since the focus of this study is the ramps, I would suggest also evaluating the ramp statistics (with wind power instead of wind speed) of those two levels to address the potential biases.

   To address both Referees' comment on the representativeness of 10 m wind speeds to evaluate model performances at 80 m agl, we included Appendix 1 to the revised version of the manuscript. Using the HRRR output over the 2020-2022 period, we show:

- high correlation values (R = 0.84) between wind speeds at 10 and 80 m;
- high correlation (R = 0.82 for up ramps and R = 0.84 for down ramps) between the total number of modeled ramps at the METAR weather stations at these 2 levels (new Fig. A1.1);
- consistency in the normalized geographical distribution of modeled ramps between the 10 m and 80 m levels (new Fig. A1.2).

Also, although 80 m wind speeds are not measured in many locations compared to the availability of METAR stations, we used observations collected routinely at the Central Site of the ARM Observatory in OK to show high correlation between the 10 m level and the next few levels above it (R = 0.94 for 10 m vs 80 m wind speed and R = ~0.8 for 10 m vs 80 m wind power capacity factor) for all 3 years (new Fig. A1.3).

We also included some reasoning on the purpose/implications of our study/results in Section 3: *"Ramp events can be divided into those that occur because of the strong diurnal variability within the boundary layer, and those that are associated with meteorological phenomena such as cold fronts, gust fronts, or other changes in forcing from transient mesoscale pressure gradient fields. Although the diurnal variation of wind speeds at 10 m and at several 100 m can be out-of-phase (with 10 m wind speeds decreasing during the night time hours while at 300-400 m they may increase at night due to the low-level jet) diurnal variations at both heights are driven by surface and boundary layer fluxes and turbulent mixing. If improvements to the model's parameterization of those diurnal processes increases forecast skill at 10 m, one would expect that improvements to forecast skill would also be found at greater heights within the boundary layer."*

3. Both HRRRv3 and HRRRv4 have longer periods of data than what is used in this study. Why was only one year of V3 data used? Additionally, although 2021 and 2022 were both simulated by HRRRv4, the large differences observed between these two years (Figure 5,6,7, and 9) indicate that the inter-annual variation in skill may not be fully explained by the model improvements alone. This raises concerns about the representativeness of the dataset. For instance, the conclusion that there is a 50% increase in skill for summer up-ramps; how much of this improvement can be attributed to the model improvement vs inter-annual variability? Can this conclusion apply to other years? Expanding the analysis to include multiple years and evaluate the interannual variability for both versions would strengthen the conclusions.

   Regarding inter-annual variability being a possible contribution to the skill of the model at forecasting wind ramps, we agree with the Referee's concern and we now mention this possibility in the main body of the manuscript (Section 5.2, discussion of Fig. 10) and also include Appendix 2 to investigate this possibility in more detail. In Appendix 2 we show that the wind speed field output at 80 m agl of the HRRR model are similar in winter months between years 2020 and 2021, but are indeed stronger in 2022, while they are stronger in summer 2020 compared to summer months of 2021 and 2022 (new Fig. A2.1). Although there is this variability in 80 m wind speed among the years, when we look at the skill score by individual years (new Fig. A2.2), we find that while there are some differences in skill score between years 2021 and 2020 (with the same HRRRv4 model), the skill score is still improved in both years with HRRRv4 (2021 and 2022), compared to HRRRv3 (2020). This confirms that even though inter-annual variability can impact the score of the model, HRRRv4 is still doing better than HRRRv3 as previously stated.

Also, we understand the Referee's point on using a larger dataset, but this would become a much larger effort computationally, and we do believe that the addition of the Appendix discussed above helps to confirm our results.

4. For many figures, the captions are repeated in the main text. I suggest removing this redundancy to save space and instead expanding on the discussion of the figure contents.

   The figure descriptions in the main text have been shortened as suggested and moved to the figure captions.

5. The geospatial distribution of results has not been sufficiently addressed. Most statistics are averaged over all sites, and there is little discussion of the spatial variability. This aspect could be tied to the physical developments in HRRRv4. Analyzing and discussing the spatial distribution would provide additional depth to the analysis.

   Additional discussion on the geographical distribution of the results is included in the discussion of Fig. 9 (Section 5.1).

Specific Comments

1. Lines 37-39, please add a reference for this statement.

   Added, as suggested.

2. Lines 85-88, this paragraph seems out of place and may connect better to the paragraph starting at Line 70.

   Thanks for this suggestion, the paragraph was moved to connect with the one starting at Line 70.

3. Lines 108-109: consider moving this sentence to figure caption.

   Modified, as suggested.

4. Line 111, please also note on the rated speed.

We don't understand the suggestion of the Referee, but would be happy to include it, if clarified. Thanks.

5. Lines, please see my major comment.

   More specifics on the way the skill score of the model at forecasting wind ramps is computed are included in the revised version of the manuscript (Section 2).

6. Line 180, since METAR data is assimilated, should good performance at those locations be expected? How this result applicable to the area without METAR station available should be discussed.

   Observational datasets are assimilated at the METAR weather stations as well as everywhere else when available. The assessment of the improvement in model performances independently from the fact that the data are being assimilated, would require performing data denial experiments, which is beyond the scope of this study. A data denial study performed during the first Wind Forecast Improvement Project has been presented in Bianco et al. (2016).

   In this study, since both HRRRv3 and HRRRv4 assimilate METAR weather station observations, as well as other observations available elsewhere, we can argue that our comparison of the skill of these models at forecasting wind ramp events is still valid.

7. Lines 186-188, many down ramps occur around 00 UTC (Figure 7) when artificial "ramp" are also expected. Why not use simulation start at other hours (e.g., 6 UTC) when ramp is less frequent?

   The 00UTC down ramps are due the reduction of wind speed that happens at sunset, not to the stitching of the model forecasts.

8. Figure 5, my understanding is this figure based on model results at site locations. How about the observations? Meanwhile, please indicate that the how the size of the circles was normalized. How does this normalization influence the results?

   We agree with the Referee that Fig. 5 was difficult to interpret and reproduced it using colorbars with appropriate ranges of variability. Also, the ramp numbers presented in this figure are not normalized. The observation values are used to produce Fig. 6.

9. Line 230, larger difference between 2021 and 2022 are observed compared to their difference from 2020, suggest that the interannual variability is more important than difference in model versions?

   See comment above on the inclusion of Appendix 2 to address inter-annual variability as a possible impact on the results. As mentioned above, while there are some differences in skill score between years 2021 and 2022 (with the same HRRRv4 model) due to inter-annual variability of the wind speed field, the skill score is still improved in both years with HRRRv4 (2021 and 2022), compared to HRRRv3 (2020). This confirms that even though inter-annual variability can impact the score of the model, HRRRv4 is still doing better than HRRRv3 as previously stated.

10. Figure 6, this figure suggests a good consistency between the years. However, the blue color spreads over a wide range of data from 0 to 100%, potentially masking large differences. Consider using more colors within the 0 to 100% range.

    We agree with the Referee that Fig. 6 was difficult to interpret and we reproduced it using colorbars with appropriate ranges of variability.

11. Lines 238-240, please move this to figure caption.

    Modified, as suggested.

12. Lines 245-248, Redundant with the figure caption; remove this repetition.

    Modified, as suggested.

13. Figure 7, please change the title to “Diurnal variability in ramps and wind at 10 m”. Visionally a better wind speed simulation in HRRRv3 compared to V4.

    We agree with the Referee that HRRRv4 seems to have some bias at nighttime in wind speed. This though, does not reflect in the ramp statistic results. In this study we are looking at improvements in forecasting wind ramp events, a different metric completely from standard statistical metrics.

    Also, the title of Fig. 7 has been changed, as suggested.

14. Line 257, could you elaborate how the statistics are calculated?

    This statistical evaluation has been explained in more detail in the revised version of the manuscript *“Although, as discussed in Fig. 6, the number of observed ramps is in general larger than the number of model ramps, we performed a statistical analysis for the matched wind ramp events (model and observed ramps are matched when the distance between their relative central time is less than the defined time window length, i.e. 2hr for the type of ramps defined as having a ΔP/ΔT 40%/2hrs). The correlation and root mean square error (RMSE) in ΔP for these matched events at all sites are presented in Fig. 8. For HRRRv4 we used the averaged correlation coefficient and RMSE of years 2021 and 2022. With the exception of winter, both the statistical metrics improve in HRRRv4 compared to HRRRv3”*.

15. Lines 258-259, this has already been included in figure legend.

    Modified, as suggested.

16. Lines 276-278, could you discussion the spatial distribution? we do see a larger improvement in the region with less ramps.

Some discussion on the geographical distribution of the results are included in the discussion of Fig. 9 (Section 5.1). Specifically, the improvement is found in all of the study area, despite the different geographical distribution of wind ramp events seen in Fig. 5.

17. Lines 281-284, already in figure caption.

    Modified, as suggested.