Dear Authors, thank you for your detailed responses to my original comments. You have made many improvements and clarified most of my questions. However, some issues remain, which should be revised before final publication.

My comments to your “Author responses” are given below in blue

Comments to Author responses

1. P4L119-120: It’s not clear to me what “it’s” refers to here
   In the sentence “The jet's acceleration is driven by the isallobaric term, with Coriolis torque and advective tendency terms contributing to its propagation perpendicular to the FLLJ.”, it refers to the FLLJ itself. However, the additional terms “perpendicular to the FLLJ” mislead the reader. It should be “parallel to the frontal surface”. The entire sentence has been rephrased as below for clarity, and is modified in the revised, at P4L119-120.
   The jet's acceleration is driven by the isallobaric term, while the Coriolis torque and advective tendency terms influence its propagation along the frontal surface and adjust its alignment relative to the frontal boundary
   Excellent, it's much clearer now.

2. P5L145: Why did you choose the two cases from the five as you did? What criteria did you use? The choice of cases is arbitrary. We selected two cases for a complete analysis while the rest of three cases for selected analysis
   Thank you for the clarification; it should be mentioned in the paper also.

3. P6L162: I assume the “forecast” experiment is using GFS? Perhaps it’s obvious, but consider stating it explicitly here already
   In the revised manuscript, the forecast experiment is explicitly mentioned in the methods section as presented below.

   At P8L161-163:
   For the sensitivity analysis, we chose to vary the initial and boundary conditions, using ERA5 and CERRA datasets to represent hindcast experiments and GFS (Global Forecasting System) to represent the forecast experiment.

   At P10L195-201:
   The initial and boundary conditions provided by reanalysis datasets (e.g., ERA5 and CERRA) tend to be very accurate due to extensive data assimilation. However, in the context of real-time forecasting, such high-fidelity boundary conditions are not available, thus operational forecast data from global models (e.g., GFS) can be used. In the forecast experiment GFS-3d1kmMYFP, cases 1 and 2 are forecasted using a similar configuration as ERA5-3d1kmMYFP, except with GFS real-time forecast data provided as initial and boundary conditions. The GFS IC/BCs during the simulation period are available at a three-hourly resolution and a horizontal

grid spacing of about 30km.

At P11L243-244:
The GFS IC/BCs initialized at 1200 UTC on 21st February 2016 for case 1 and at 1200 UTC on 3rd March 2016 for case 2 are obtained for the forecast experiment.

Great. That clears it up.

4. MYNN2.5: what "bl_mynn_*" settings were used? The defaults of WRF v4.4? MYNN2.5 can be quite different depending on this.
   In our simulations we used the following settings for bl_mynn. bl_mynn_tkebudget = 1, bl_mynn_tkeadvect = .true., On the other hand, the other settings are unaltered, thus taken as their default settings.
   Thank you for the clarification. You should describe at least these non-default settings in the paper to allow others to reproduce the results.

5. In a study focussing on capturing the timing of an event, it's surprising to me that you don't consider the influence of data assimilation (except to say that is one reason for the accuracy of re-analysis datasets). Why did you not test data assimilation and/or discuss this in the paper?
   It is true that data assimilation can be a crucial factor, especially when it comes to real-time forecasting, where accurately initializing the model states is key to improving short-term predictions. However, in this study, our main objective was to examine the sensitivity of modeling parameters and assess the ability of the WRF model to simulate frontal low-level jets and associated extreme ramp events. By focusing on the modeling aspects, we aim to better understand the dynamics and physical processes involved in the event simulation, separate from the impacts of assimilation techniques.
   Your paper strongly emphasizes modeling the timing of ramp events, so data assimilation remains a relevant missing factor. I will maintain that you should do more to discuss the possible consequences of not including DA here as it relates to the conclusions about choosing your nesting and model domains.

6. You use 51 levels, why not more? Are you sure it's not sensitive to this?
   In the revised manuscript, we have examined the influence of vertical levels on the simulations by selecting 101 vertical levels, with approximately 27 vertical levels within the first 1 km of height. Using this configuration and the CERRA-1km1dMYFP simulation strategy, the first two cases were simulated. From the analysis, it was found that the vertical levels exert noticeable influence on the ramp timing and post-ramp wind speed. However, the performance of the 51 vertical levels configuration was better than that of 101 vertical levels, further justifying our choice of vertical levels. The analysis is presented in Appendix B in the revised manuscript and is presented below for reference.
   We have examined the influence of vertical levels on the simulations by selecting 100 vertical levels, with approximately 27 vertical levels within the first 1 km of height. Using this vertical resolution and the CERRA-1km1dMYFP configuration,

the first two cases were simulated. In our study, the choice of 50 vertical levels was adopted from Nunalee and Basu (2014), where in the case of coastal LLJs, the authors reported reduced jet strength and lower jet core height with increased vertical levels, deviated from the observations. Similar to their findings, we noticed the vertical levels exert considerable influence on the ramp timing and marginal influence on the ramp intensity, as shown in Fig. B1. However, we also recognize the limitations of our findings, since they are based on two simulations focused specifically on FLLJ cases.

This is a nice addition to the paper, providing clear evidence that adding more levels matters, but not with a clear positive on your application.

7. You used 6 hr of spin-up. Are you sure the model has enough time to develop these strong weather events? Perhaps the poor performance of some of the experiments is due to insufficient spin-up time. See e.g. Lui et al. (2023)

   To evaluate the sensitivity with respect to spin-up time, we conducted simulations for the two cases using the CERRA-1d1kmMYFP configuration, one with a 6-hour spin-up and another with a 12-hour spin-up. The results revealed that the wind power output from both simulations was identical, indicating that a 6-hour spin-up was sufficient for the development of the FLLJ cases. This confirms the adequacy of the chosen spin-up duration for accurately capturing the dynamics of these events. The following analysis is presented in Appendix C in the revised manuscript.

   To evaluate the sensitivity of model accuracy to the choice of spin-up time, we conducted simulations for two cases using the CERRA-1d1kmMYFP configuration: one with a 6-hour spin-up (original simulation duration) and another with a 12-hour spin-up (starting 6 hours prior to the original duration). Figure C1 presents wind power time-series from both simulations, compared with the measured power output for Cases 1 and 2. The results show that the wind power output from both simulations is identical, indicating that a 6-hour spin-up is sufficient for the development of the FLLJ cases. However, it is highly unlikely for simulations initialized with different IC/BCs to produce identical results. To further verify this, we compared the time-height crosssection of wind speed from the two simulations, along with lidar observations at the LOT2 location, for Cases 1 and 2, as shown in Fig. C2. The comparison reveals that the simulations are nearly identical, with no discernible differences in wind speed, except for minor variations marked by sky-blue circles in both cases. Since no differences are observed in wind speed below the 100 m level, the wind power outputs from the two simulations also remain identical, as shown in Fig. C1. These findings confirm that the chosen 6-hour spin-up duration is adequate for accurately capturing the dynamics of these events.

   Save as above. This is a helpful addition to the paper, ruling out influences of spin-up time.

8. Figures 4-10: Why not show the results from the datasets used for forcing data: ERA5, CERRA, and GFS? It would be more convincing to show that downscaling is needed if I

could see the reference data as well.

We appreciate the reviewer's suggestion to include the results from the forcing datasets (ERA5, CERRA, and GFS) in Figures 4–10. However, we are unable to include these results in the manuscript for the following reasons. We did analyze wind speeds from ERA5 and CERRA at 100 m and from GFS at the 975 hPa pressure level, comparing them with lidar observations, which are shown in Fig. D1, and are presented here for reference. Our analysis revealed that while GFS follows the overall trend of observed wind speeds, it exhibits a clear overestimation in magnitude. Though the GFS data at the current time exist at an hourly temporal resolution, it is only available at a 3- hourly resolution for 2016-17 years. Due to the coarse temporal resolution, it is challenging to quantify ramp statistics, particularly as extreme ramps associated with frontal low-level jets occur on scales of minutes. This further demonstrates that dynamical downscaling is indispensable for understanding extreme wind ramps at sub-hourly scales, which are critical for wind power applications

Furthermore, and most importantly, CERRA and ERA5 provide wind speeds at 100 m above ground level, while GFS provides wind speeds at the 975 hPa pressure level. However, four of the five wind turbine types operating in the wind farm have hub heights below 100 m, making it challenging to accurately approximate wind power production using the forcing datasets.

Additionally, most of the analysis from WRF simulations in this study was conducted at temporal scales of 10–15 minutes, whereas the temporal resolutions of the forcing datasets range from 1 to 3 hours. This mismatch in temporal resolution further limits the utility of the forcing data in capturing rapid changes and short-term wind fluctuations associated with extreme ramps.

The ERA5 reanalysis exists at a spatial resolution of 0.25°, the CERRA analysis exists at a spatial resolution of 5.5 km, and the GFS forecast exists at a spatial resolution of 0.25°. Due to these coarse spatial resolutions, the entire wind farm fits within a single grid cell for ERA5 and GFS, while only a few grid cells from CERRA cover the wind farm. These spatial limitations hinder the representation of the wind power output generated by 182 turbines within the wind farm. This is critical, as the heterogeneity in turbine locations and wind conditions cannot be resolved with the coarser spatial scales of the forcing data.

Including the analysis from the forcing datasets would also increase the number of outputs by three, significantly complicating the visual presentation of Figures 4–10. We believe excluding the forcing data maintains the clarity and focus of the manuscript while strengthening the argument for the necessity of dynamical downscaling to capture subhourly wind variability. The justification regarding the necessity of dynamical downscaling is incorporated at PL in the revised manuscript

I agree that the forcing datasets are probably less appropriate for extreme wind ramps due to the reasons you list. However, that is precisely why it would be valuable to show the significant improvements from your dynamical downscaling relative to the forcing data. Both ERA5 and CERRA are available hourly at many levels. See, e.g., https://cds.climate.copernicus.eu/datasets/reanalysis-cerra-height-levels?tab=download, and https://cds.climate.copernicus.eu/datasets/reanalysis-era5-complete?tab=d_download. I agree with you that adding more datasets to the figures could lead to cluttering. However, you could add them to Tables 2 and 4, which would indicate the improvements from downscaling.

9. The range of values in the colormap makes some of the plots difficult to it difficult to interpret. For example, do you really need to include values up to 30 m/s in figures D1-4. In Fig. C1, the breaks in the colormap seem to be inconsistent (sometimes 2 m/s, sometimes 1 m/s)
   In the revised manuscript, the color levels of Figures F1-4 are set according to the minimum and maximum of the wind speed, at an interval of 1 m/s. For Fig. E1, the color levels are set from 4 to 14 m/s at an interval of 2 m/s and from 15 to 28 m/s at an interval of 1 m/s. This gives clarity in visualizing the wind speed contours for cases 1 and 2.
   The new levels help, but I recommend choosing a different colormap without so many discontinuities/breaks. See, e.g., https://www.fabiocrameri.ch/colourmaps/.

10. Do you use the ERA5 pressure-levels or model-levels?
    We have used the ERA5 pressure level data as initial and boundary conditions.
    Thank you for the clarification here. You should also clarify this in the paper.

11. In P14L298-300: What exactly are you arguing here? that WRF cannot generate sharp gradients from coarse ERA5 boundary-condition data? You say CERRA constitutes better BCs, but it is itself based on ERA5.
    In the referenced passage, we argue that while CERRA is derived from ERA5, it incorporates additional data assimilation and a significantly higher horizontal resolution (5.5 km vs. 31 km), which enhances its representation of mesoscale features. This improved resolution and assimilation process allow CERRA to better capture sharp gradients, such as those associated with the Frontal Low-Level Jet, providing more accurate boundary conditions for WRF simulations.
    So, one way to see the results is that, since you don't include data assimilation, the best option is to rely mainly on the high-resolution CERRA IC/BCs to obtain good features and timing by using only 1 WRF domain and reducing the fetch from the boundary to the region of interest. Your innermost domain is centered on what looks like approx. 30 km northeast of the wind farms in the direction opposite to the prevailing wind direction, leaving approx. 120 km of fetch for the prevailing wind direction.

12. P23L423: Why are you surprised that the model captured the event? Please elaborate on why it's surprising
    The intention was not to express surprise at the WRF model's ability to forecast the ramp and extremity of the FLLJ but rather to highlight the contrasting performance of the Elia forecast. Specifically, while the WRF model captured the ramp events in both cases, albeit with some timing mismatch, the Elia forecast either predicted a gradual ramp-down (case 1) or did not capture the ramp at all (case 2). This discrepancy was noteworthy because it suggests potential challenges in operational forecasting of such extreme

events. We have revised the text to clarify this point and to better reflect the intended message without any exaggeration. The revised manuscript is updated with the following text, at P22L443-444.
We note that the WRF model was able to forecast the extreme ramp event in both cases, though with some degree of timing mismatch.
Thank you for the clarification.

13. Figure 11: Unfortunately, only the CERRA-1d1kmMYFP was included here, it would have been more convincing to show that the trend from cases 1 and 2 continues here
We appreciate your interest in extending the comparison to demonstrate the trends from Cases 1 and 2 within this figure. The three additional cases have been simulated using the CERRA-based configurations, namely CERRA-2d1kmMYFP, CERRA-2d1kmMYnoFP, CERRA-2d1kmSH, and CERRA-1d1kmMYFP. The following analysis, figure, and table are incorporated in the revised manuscript, at P26L463-476.
The results demonstrate that the WRF model successfully simulates the extreme wind power ramps associated with FLLJs, though the specific modeling configurations significantly influence the wind power time series. All simulations capture the strong pre-ramp wind power at the 720 MW rated capacity, likely due to peak wind speeds during the FLLJ, as well as the post-ramp wind power output, which aligns with observed values. The observed strong drop in wind power, representing the extreme ramp, is consistently simulated across cases; however, discrepancies in ramp intensity and timing persist, varying across configurations. The ramp statistics from the four CERRA-based configurations for three cases are computed at a 1-hour time scale and are presented in Table 4. The CERRA-1d1kmMYFP simulation shows superior performance, with the power ramps surpassing 50% within just 1 hour, signifying the extremity of the power ramps. In terms of ramp timing, the simulated ramps are in advance by 2 hours, 45 minutes, and 15 minutes in cases 3, 4, and 5, respectively. This indicates the robustness of CERRA-1d1kmMYFP in accurately capturing ramp timing and intensity for cases 4 and 5. In case 3, while the timing is simulated with a 2-hour lead, the ramp intensity is closely represented compared to other configurations. Other simulations exhibit larger temporal shifts in cases 4 and 5, and underpredicted intensities in cases 3 and 4. Nonetheless, these findings corroborate the robustness of the WRF model in simulating the extreme ramp events associated with the FLLJ, and the better predictability of the CERRA-1d1kmMYFP modeling configuration.
It's highly appreciated that you extended your analysis to the three remaining cases. The trend appears reasonably consistent, with CERRA-1d1kmMYFP performing best (at least for 4 out of 5 events). Perhaps you could highlight (bold text, perhaps?) the best-performing simulation in Tables 2, 3, and 4 for intensity and timing.

14. P27L463: Same as above, you say CERRA provides better BCs than ERA5 to capture timing and intensity, but CERRA is based on downscaling from ERA5, so perhaps the problem is not ERA5 BCs but the downscaling and e.g. lack of data assimilation?

In continuation, as stated in response 5, while CERRA is based on downscaling from ERA5, it provides better boundary conditions (BCs) due to its higher resolution (5.5 km) and the frequent incorporation of observational data through data assimilation techniques on a regional scale. These factors enhance CERRA's ability to represent mesoscale processes, allowing for more accurate reproduction of the timing and intensity of atmospheric phenomena like frontal low-level jets. That said, we acknowledge that data assimilation could significantly impact the results, particularly for real-time simulations and improving the initial state of the atmosphere. However, data assimilation is out of scope for the current study, since many different approaches exist, such as 3DVAR, 4DVAR, Kalman filter, which perform differently as per the literature. In this study, our primary focus was on evaluating the model's sensitivity towards different modeling configurations and examining the differences between using higher and lower resolution IC/BCs
See my answer under 5)

Additional comments

- Please complete the “data availability” section related to all the data used and produced in your study. What about the availability of your simulations, the LiDAR data, and so on?
- Consider providing more detail in the “author contribution” section. See, e.g., https://publications.copernicus.org/services/contributor_roles_taxonomy.html