Reviewer Comments on "A terrain-stratified comparison of regional MPAS-A (v8.3.1) and nested WRF (v4.7.1) for short-term near-surface wind forecasting"

By Ruhi Deniz Yalcin, Mustafa Tugrul Yilmaz, Ismail Yucel, Omer Lutfi Sen, Sertac Oruc, Ali Ulvi Galip Senocak

This manuscript presents a comparison of short-term near-surface wind forecasts over Türkiye using regional MPAS-A and nested WRF configurations, with particular emphasis on terrain complexity and nesting strategies. Using observations from 509 meteorological stations, the authors evaluate model performance across different terrain conditions and wind-speed regimes. The results suggest that MPAS-A generally achieves lower aggregate forecast errors, although the relative performance of the two models varies with terrain complexity and local wind conditions. Overall, the study addresses a relevant scientific question and provides a potentially useful assessment of the strengths and limitations of these two modeling approaches for near-surface wind prediction over complex terrain. However, I have several major concerns and would like to ask for further clarification from the authors to help me better understand the study and assess its scientific value. I therefore **recommend major revision**, and my detailed comments are as follows.

Major Comments

1. The concluding statement in the Abstract (lines 29–31) seems overly general and does not sufficiently highlight the key scientific insights gained from this study. More broadly, the authors state in lines 71–72 and 90–92 that the study aims to investigate how model design, refinement strategy, and terrain complexity interact and influence short-range wind prediction. Given these stated objectives, I would expect the authors to provide clearer insights into the respective roles of these factors and their contributions to differences in model performance, as well as how the findings could inform model selection and reveal the potential advantages and limitations of WRF and MPAS-A for near-surface wind prediction over complex terrain. Although Section 5 provides more specific findings regarding horizontal resolution, nesting strategy, and terrain-dependent performance, the underlying physical or numerical mechanisms are not always adequately demonstrated to support the proposed interpretations. I encourage the authors to strengthen the connection between the scientific questions, analyses, and conclusions, while distinguishing findings directly supported by the experiments from potential explanations requiring further investigation.
2. **Section 2:** I suggest reorganizing and streamlining this section to improve its clarity and logical flow. The main text could focus on the key experimental design, model configurations, and methodological information necessary to understand the subsequent analyses, while detailed technical descriptions, such as mesh-generation procedures and specific computational settings, could be moved to an Appendix or Supplementary Material. In particular, the experimental design could be described more clearly, including how the daily forecasts are initialized, how different model configurations are compared, and how overlapping forecast periods are handled in the verification. In addition, Sections 2.5.1 and 2.5.2 could be integrated with Section 2.6 into a unified verification methodology section. This would make the methodology more concise and focused on the scientific objectives of the study.
3. **Sections 2.5.1 and 2.5.2:** The terrain-complexity classification is important to the study, but two aspects need further clarification. First, the TRI is calculated over an approximately 900 m neighborhood, while the model resolutions range from 4 to 12 km. Is this spatial scale appropriate

for characterizing terrain complexity relevant to the model simulations, and are the conclusions sensitive to the selected scale and classification thresholds? Second, given the focus on near-surface wind prediction, do the three terrain classes exhibit distinct characteristics in the observations, such as mean wind speed, variability, or wind-speed distributions? I think addressing these questions would help establish the physical relevance and robustness of the terrain classification and provide a stronger basis for interpreting the terrain-dependent differences in model performance.

4. **Tables 4–6:** I noticed some apparent inconsistencies between the overall and terrain-stratified verification statistics. For example, Table 4 reports an overall ME of 1.38 m s^{-1} for WRF-d01-1way, whereas the station-count-weighted average of the three terrain classes in Table 6 is approximately 1.24 m s^{-1}. Interestingly, the terrain-stratified statistics appear to agree more closely with the June-only results than with the combined January–June results. I suggest that the authors check the consistency of the reported statistics and clarify how these metrics are calculated and aggregated. In particular, providing the sample sizes (e.g., the number of stations and matched forecast–observation pairs) used for each set of statistics, along with a clearer description of the aggregation procedures, would help readers better understand and interpret the results.
5. **Sections 3.2 and 4.1:** The authors attribute the improved performance of MPAS-A over complex terrain partly to its smooth resolution transition and the absence of internal nesting boundaries. While this is a plausible explanation, I wonder whether the current experiments provide sufficient evidence to support this interpretation, since WRF and MPAS-A also differ in their numerical formulations, vertical coordinates, implementations of physical parameterizations, and potentially terrain representation. In particular, is there evidence that the differences in forecast performance are directly related to the mesh refinement and nesting strategies? Additional diagnostics or discussion would help clarify the underlying mechanisms. Otherwise, I suggest that the authors present these explanations more cautiously as potential contributing factors.
6. **Sections 3 and 4 (Results and Discussion):** A major concern is the lack of uncertainty quantification throughout the results and discussion. Without uncertainty estimates or statistical significance assessments, it is difficult to determine whether the reported differences between model configurations represent meaningful improvements or fall within sampling variability. For example, although the authors emphasize a 41% reduction in ME for MPAS relative to WRF-d02-2way at high-complexity sites, the corresponding RMSE improvement is only about 6% (2.12 versus 1.99 m s^{-1}). Similarly, lines 473–475 interpret a small CC difference (0.545 versus 0.534) as evidence that the finer WRF domain captures temporal variability more accurately, which I find difficult to justify without uncertainty estimates. For the wind-speed-dependent analysis in Section 3.6, sample sizes and uncertainties should also be reported for each wind-speed bin, particularly under high-wind conditions. I strongly encourage the authors to assess the statistical robustness and practical significance of these differences, especially given that several key conclusions regarding the relative advantages of WRF and MPAS-A rely on relatively small differences in verification metrics. **I would ask the authors to pay particular attention to this issue**, as it directly affects the strength and reliability of the study's main conclusions.

Minor Comments

1. **Abstract:** Would it be better to define the abbreviations ME, MAE, RMSE, and CC at their first occurrence in the Abstract? Although these terms are defined later in the manuscript, the Abstract should be self-contained and understandable without referring to the main text. Defining these abbreviations would improve clarity and readability, particularly for readers who may not be familiar with these verification metrics.

2. **Throughout the manuscript:** The notation of wind-speed units is inconsistent, with formats such as m s^{-1} and m/s used throughout the manuscript. Please check and standardize the unit notation for consistency.
3. **Lines 116–119:** The justification for excluding the first 6 h as model spin-up is not entirely convincing and would benefit from supporting references or additional analysis. As I understand it, model spin-up refers to the initial adjustment associated with physical and dynamical inconsistencies in the initial conditions, and its duration can be model-dependent. Therefore, a fixed 6 h spin-up period may not necessarily be equally appropriate for WRF and MPAS-A. Rather than relying primarily on a general justification, why not examine the spin-up behavior directly using the simulations? For example, presenting the evolution of near-surface wind forecast errors with lead time could help assess the initial adjustment and support the choice of the 6 h threshold.
4. **Figure 5:** Please clarify how the statistics presented in the Taylor diagram are calculated and aggregated across stations and forecast periods, and whether the RMSE contours represent normalized centered RMSE. In addition, the figure suggests that differences among terrain-complexity classes are more pronounced than differences among model configurations within each class. It would be helpful to discuss the significance of these differences and what they reveal about the relative performance of WRF and MPAS-A under different terrain conditions.
5. **Lines 299–317:** I suggest reconsidering the paragraph organization in this section. Several short paragraphs discuss closely related aspects of the spatial differences in model performance and do not seem to represent sufficiently distinct ideas to warrant separate paragraphs. Combining some of these paragraphs and improving the transitions between them could make the discussion more coherent and easier to follow.
6. **Lines 315–317:** The reasoning here is not entirely clear to me. Are the authors suggesting that MPAS-A would otherwise be expected to outperform WRF, but its stronger underprediction at high wind speeds leads to better WRF performance in coastal regions? I wonder whether the current analysis provides sufficient evidence to support this interpretation, as differences in grid geometry, coastline representation, and other model characteristics may also contribute to the spatial differences in performance. Moreover, Section 3.6 primarily presents a statistical relationship between forecast bias and observed wind speed rather than identifying the underlying physical processes responsible for these differences. Therefore, I find the statement "The physical mechanism underlying the coastal reversal" somewhat misleading. I suggest that the authors clarify their reasoning and provide stronger evidence to support the proposed explanation, or present it more cautiously as a possible interpretation rather than an established physical mechanism.
7. **Sections 3 and 4:** The authors frequently use terms such as low, moderate, and high wind speeds, but do not provide clear definitions or criteria for these categories. For example, line 486 describes low-to-moderate wind speeds as below approximately 5 m s^{-1} in high-complexity terrain, while lines 425–429 discuss different zero-crossing thresholds for WRF and MPAS across the three terrain classes. It is unclear whether these categories refer to fixed wind-speed ranges, observed wind-speed distributions, or model-dependent bias characteristics. I suggest clarifying these definitions and using the terminology consistently throughout the manuscript.
8. **Sections 2.6 and 3:** Since the authors conduct daily 48-hour forecasts and verify lead times 6–47 h, there are overlapping verification periods in which the same observations may be used for different forecast initializations. Could the authors clarify how these overlapping samples are handled in the verification and whether their dependence is considered when assessing statistical uncertainty? This would help establish the robustness of the reported differences between model configurations, particularly for metrics showing relatively small differences.

9. **Sections 4 and 5:** The authors suggest that grey-zone boundary-layer limitations and lateral boundary constraints may explain why increasing WRF resolution from 12 to 4 km does not consistently improve near-surface wind predictions. While these explanations are plausible, they are not directly demonstrated by the current analysis and should be presented more cautiously. In addition, since the evaluation covers only January and June 2023, I suggest that the authors acknowledge the limitations of this relatively short evaluation period when drawing broader conclusions about seasonal differences and model performance.