“Simulation performance of different planetary boundary layer schemes in WRF V4.3.1 on wind field over Sichuan Basin within “Gray zone” resolution”

The authors undertake a “gray zone” WRF simulation campaign in an understudied region of China (Sichuan Basin) using different PBL schemes compared to one airport meteorological measurement. Results using common statistical error metrics for wind speed and direction are shown, where results for the different PBL schemes show good agreement for wind direction but poor agreement for wind speed. A k-means clustering technique is leveraged to help group different error metrics together and gauge PBL performance.

Overall, while I appreciate the study the authors are trying to undertake, I feel the analysis is underwhelming in breadth and justifications for modeling choices made are weak. Only one observation site is chosen for comparison, and yet it is believed to be representative of the entire region. Additionally, I am still left questioning why such a high spatial resolution WRF simulation was conducted, especially when the comparison was only performed against one observation site. Discussion of relevant meteorological phenomena is vague. For example, stability is often mentioned and used to understand the results, but no mention of a stability metric is used or referenced.

2 Data and Methods – general comments/questions

What is the temporal output of the WRF data, and how often is the model updated? I don’t think this is every mentioned.

Why use such a high-resolution inner domain? Is it to prove that such simulations are possible with a mesoscale model in this region? It is unclear why such a high resolution WRF simulation is performed, especially when only considering one measurement site.

Only one reference measurement is used, yet strong claims are made about PBL scheme performance for just one 10 m wind tower measurement.

2 Data and Methods – specific comments/questions

pg. 5, line 166: A spin-up period of 3-hours is short, especially with a domain with complex topography. What was the reason for such a short spin-up time? I’m concerned this could affect results for the case studies, at least in the first few hours after spin-up are thrown out.

3 Overview of historical cases and evaluation of simulations results – general comments/questions

Throughout the results stability is mentioned many times by the authors, but it is never made clear how stability is defined in this study. If a discussion of model results compared to observations is going to take place, stability needs to be defined and/or referenced.

3 Overview of historical cases and evaluation of simulation results – specific comments/questions

pg. 8, line 240: A more thorough description of the dominate atmospheric circulations for each event is needed. Where is the "cold air" coming from? Is it a frontal passage, low-level jet, local terrain flows, etc.? Just saying "cold air" is not informative.

Figure 2b: I appreciate and understand what the authors are trying to convey here, as trying to plot 28 different time-series in one plot is not easy. I would emphasize in the figure caption though that this is not a continuous time-series, as upon first glance, the figure can be misleading. Also, what is the significance of the 5 m/s dotted line?

Figure 3: Why is the color bar range for wind speed values different than those of Figure 2a? This makes it difficult to compare observations and model results. It would be more beneficial visually if the observational wind rose from Figure 2 is combined into one figure with the model results of Figure 3 to more easily compare.

pg. 11, line 285: Again, it's hard to compare the differences in wind speed with a different color bar range and not having the plots side-by-side. Additionally, what are these other studies showing similar results? Cite them at the very least, and perhaps include some number ranges for reference.

pg. 14, line 356: Quantitatively state what these deviations are instead of using qualitative language. This advice goes for the entire paper, where qualitative statements are often more common than quantitative.

Figure 7: There is a lot of information being shown here, which is tricky to do, but would this be better as a line plot where each line is a different PBL scheme, and the error bars are shading around those lines? That might be easier to read than ~100 bar charts.

pg. 15, line 390: Perhaps the wrong word is being used here, but if the authors are going to make claims of significance, the authors should back up this statement with statistical significance tests. Otherwise, remove this statement and/or reword this sentence.

pg. 16, line 406: Unclassified results? What does this mean?

pg. 16, line 414: Are seasonal results not shown because there are no obvious seasonal differences?