Response to referee comments

We are grateful for the thoughtful and constructive feedback from the reviewer. We have revised the text in the manuscript to answer the referee's points and we believe this revision has improved the clarity and quality of our manuscript. This response provides a complete description of the changes that have been made in response to each comment. Referee comments are shown in plain text, author responses are shown in bold blue text. All line numbers in the responses refer to locations in the revised manuscript.

Referee 1:

Referee Comments

Second round review of "Impact of physical parameterizations on wind simulation with WRF V3.9.1.1 over the coastal regions of North China at PBL gray-zone resolution" by Yu et al.

In general, the authors have responded quite well to my feedback. I thank them for their attention to details. However, there are two remaining points that the authors should address before the paper is ready for publication.

Response: Thanks, we really appreciate all your comments and suggestions. Please find my responses in below and my revisions/corrections in the re-submitted files. Thanks again.

Specific comments:

1. Impact of different surface layer (SL) schemes: I understand that it is not possible to run all of the simulations with the same SL scheme, but it still would be nice to see the impact. For example, they could choose one configuration that is compatible with multiple SL schemes and run one more simulation while modifying the SL scheme. If the authors do not wish to run any additional simulations at this point, then they should at least provide some brief discussion on this issue.

Response: Thank you for pointing out this, according to this comment, we added the experiment considering the effects of SL options using UW as the PBL scheme, the results are added in Lines 350-359:

"3.3.4 Impact of surface layer schemes

In the WRF model, the surface layer (SL) schemes are somehow binding with PBL schemes, it is not possible to run all PBL schemes with the same SL scheme. However, it is meaningful to conduct simulations using a specific PBL scheme that can work with multiple SL schemes to investigate the effect of SL schemes on wind simulation. Figure 16 compares the simulations results of different SL (MM5, Janjic, GFS, MYNN and PX, Table 4) schemes using UW as the PBL scheme, the other model configurations are the same as the simulation with the best Taylor skill score. Simulations with different SL schemes generally reproduce the timeseries of wind speed well, with CORR scores of about 0.93 for most schemes. However, all simulations overestimate the wind speed, especially for the Janjic scheme. At the same time, according to the BIAS and RMSE scores, MYNN shows the best performance, followed by GFS and PX schemes. Thus, SL schemes also have major influence on the wind simulation."

2. Figure 13: Please indicate at which height you are plotting the wind field.

Response: Thanks for the comment, we added "the height of the wind is 10 meters" in the caption of Figure 13, in Line 713.

Referee 2:

The authors have addressed most of my comments and I am happy to see that many of my suggestions were adapted in the revised manuscript. The quality of the analysis has improved; however, the discussion and figures can still be improved upon, some corrections are required (in some parts the text does not seem to agree with what the figures show), and a few concerns remain to be addressed and clarified.

Response: Thank you for reviewing this paper, we really appreciate all your comments and suggestions. Please find my responses in below and my revisions/corrections in the re-submitted files. Thanks again.

Specific Comments

- "The stable weather event":
  - Based on Fig. 2 to me it seems that the study area is under the influence of the high- pressure system only on Jan 11th. Afterwards the study area experiences *north*west winds (lines 110/111 state "southwest") as the high-pressure system weakens and moves west. Except on Jan 11th, I am not convinced that the conditions are stable based on Fig. 2. The subplot for Jan 13th even indicates a weak shortwave and a shift towards positive vorticity over the study area. Please also frame the area covering D03 to guide the reader where to find the "study area" in this figure. Consider assessing other variables that indicate static stability (e.g., vertical profiles of potential temperature).
  - Line 111: "which would transport warm and wet air to the study area" – weak winds do not transport air masses very far and there is no major body of water to the west of the study area. Where does the "wet air" come from? And why are moist conditions favorable for stable conditions (line 112)?
  - Under stable conditions, wind direction is (a) more variable, hence, more challenging to predict accurately, and (b) less important, as wind speeds are weak. The manuscript should frame the results accordingly.

Response: Thank you for pointing out this.

1. ***The previous figure covers too large an area and does not show the location of the study area, which makes it difficult for the reader to accurately identify the weather conditions over the study area. We have made the following changes to Fig. 2: (1) The outline of D03 was added, (2) the cloud fraction from ERA5 was used and (3) the surface wind at 10m height was used to illustrate the stable weather condition. We revised the analysis of the weather condition in Lines 104-112: "Figure 2 depicts the distribution of geopotential height at 500 hPa, surface winds at 10 meters and total cloud fraction from the ERA5 dataset (Hersbach et al., 2020) during the study period. A weak high-pressure system persisted from 11 to 14 January 2019 over the study area, with the geopotential height at 500 hPa of about 5400 gpm, at the surface level (10 meters), weak southwest winds occurred at the south side of the study area on 11, 12 and 14 January 2019. The surface wind speeds over the study area were weaker than 5 m/s during the first four days of the study period, then the geopotential height decreased and strong northwest winds occurred over the study area on 15 January 2019. Although there were slight differences between ERA5 and satellite products (e.g., CLARA, Karlsson et al., 2021), both datasets indicated higher cloud fraction on 11 and 14 January 2019, while for the rest of the time, the cloud fraction was low. This stable weather event is used to investigate the impact of physical parameterizations of the WRF model."***
2. ***We agree with the comments that weak winds do not transport air masses very far during the study period, so we remove the statement of "which would transport warm and wet air to the study area" in the***

manuscript.

3. ***Points are well taken and we added the following in Lines 242-244: "As wind direction is more variable but less important under stable conditions with weak wind speed, the subsequent investigations mainly focus on wind speed."***

- This manuscript includes numerous figures with subplots. I think the authors should reconsider which figures are meaningful to include in the manuscript. I.e., is it essential for the reader to see Fig 5b vs. a short description that CORR scores are indistinguishable among MP schemes at a precision of .XX? For all figures, ensure that Figure labels are readable on a printed paper size. (Larger font size is needed for most figures, but in particular, Fig 2, 11 & 12.)

Response: Thanks very much, all the figures are replotted according to the comments, we used larger font size for Figures 2-17, we revised Figure 1, 2 and 13 according to the latter valuable comments.

We also revised Lines 224-226 for Figure 5: "The CORR scores are very similar for all the MP schemes at a precision of 0.01, while MYDM7 is the best scheme according to BIAS and RMSE scores, followed by P3 and ETA."

- Figure 2:
  - Is cloud fraction data available from ERA5? Although the observational data product is interesting, for your study it might be also (maybe even more) important to see whether the initial condition model produced clouds. Several cloud products do not necessarily have to be shown in the manuscript, but it would be nice if differences among observed, ERA5 and WRF cloudiness would be explained and discussed in the manuscript.
  - Could you please add a box framing the "study area" (i.e., D03)? Is it necessary to show these maps at this scale? Perhaps the area could be a more zoomed in on the study area. (Also note that the labels are not readable.)
  - With respect to the existing topography in the area, 925hPa is a very low level. Consider showing 750 or 500hPa geopotential heights instead, which can be more descriptive for synoptic situations.

Response: Thank you for pointing out this, we have made the following changes to Figure 2: (1) The cloud fraction from ERA5 was used, and the cloud fraction from different dataset was discussed in the manuscript, (2) we added the box indicating the boundary of D03, and zoomed to the regions of the study area, (3) the geopotential height at 500 hPa was used.

The manuscript was revised in Lines 104-112: "Figure 2 depicts the distribution of <u>geopotential height at 500 hPa,</u> surface winds at 10 meters and <u>total cloud fraction from the ERA5 dataset (Hersbach et al., 2020)</u> during the study period. A weak high-pressure system persisted from 11 to 14 January 2019 over the study area, with the geopotential height at 500 hPa of about 5400 gpm, at the surface level (10 meters), weak southwest winds occurred at the south side of the study area on 11, 12 and 14 January 2019. The surface wind speeds over the study area were weaker than 5 m/s during the first four days of the study period, then the geopotential height decreased and strong northwest winds occurred over the study area on 15 January 2019. <u>Although there were slight differences between ERA5 and satellite products (e.g., CLARA, Karlsson et al., 2021), both datasets indicated higher cloud fraction on 11 and 14 January 2019, while for the rest of the time, the cloud fraction was low.</u> This stable weather event is used to investigate the impact of physical parameterizations of the WRF model."

We also revised the caption of Figure 2 in Lines 658-660: "Figure 2: The daily averaged geopotential height (contour, units: gpm) at 500 hPa, total cloud fraction (shading) and surface winds at 10 meters (vectors, units: m/s) from ERA5 during 11 to 15 January 2019, the box indicates the D03 domain in Figure 1"

- Figures with error bar plots (Fig. 3-8): Please explain why the error bars (blue) sometimes have values outside of the range of station values (orange dots). Consider whether the metrics should be calculated over all stations (and lead times?), then averaged over simulations, or whether the metrics should be calculated over all lead times, then averaged over stations and simulations. Accordingly, show error bar ranges either across stations or simulations or both.

Response: Thank you for pointing out this, in Figures 3-8, the blue bars indicate the statistic scores (CORR, BIAS and RMSE) of a specific physical scheme, which are calculated by first averaging over all stations, then over the simulations using that physical scheme; while the orange dots indicate the spread of the statistic scores across all the stations. In some cases (e.g., RMSEs of MP schemes in Figure 5), the results of station mean (ensemble) are better than any of the individual station, thus the statistic scores for the station mean (blue) are outside of the spread for all stations (orange dots). According to the comments from the first round, the orange dots were added to illustrate the distribution across the 105 stations for different physical schemes, thus for each station, the metrics are averaged over simulations using the same physical scheme.

- It is challenging for the reader to remember which acronym belongs to each parameterization type. Therefore, it would be helpful to mention parameterization types more often throughout the manuscript please. E.g., in Fig. 11: Please use same scales on y-axis and add a label on the left for the corresponding parameterization types.

Response: Thanks for the comments, we revised Figure 11 according to the comments, and we mentioned the parameterization types more often in the manuscript in the following:

Line 272: The result indicates that ensemble mean of four simulations with WDM6, Goddard, NSSL1 and MYDM7 MP schemes shows the best BIAS and RMSE scores.

Line 315: "the same simulation but with QNSE PBL scheme (i.e., using QNSE, Dudhia-RRTM and WDM6 schemes) is used for comparison between the simulations with good and poor performances"

Line 385: "The simulation using YSU PBL, WDM6 MP and Dudhia-RRTM SW-LW schemes shows the best performance with the highest Taylor skill score."

- Figure 13: The many overlapping vectors are hard to decipher. Consider using wind barbs instead of vectors or less dense vector distributions with smaller reference vector. At the scale of this figure, it is also difficult to identify direction and size of the simulated wind vectors.

Response: Thanks for the comments, in the revision, we used less dense vectors so that there were fewer overlapping vectors, and we made the simulated wind vectors larger and clear.

- In figure 13 and 14 it is unclear why QNSE is shown when it was previously determined to perform poorly, and the authors decided to disregard it from further investigation.

Response: Thanks for the comments, according to the comments from the first round, we added comparisons of

YSU and QNSE to illustrate the differences between schemes with good and poor performances.

- Figure 14: Why do you show and discuss the wind speeds up to a height of 20 km (stratospheric altitudes)? How relevant is that for your study and analysis? I further suggest re-ordering the subfigures and showing all profiles at 8:00 at the top and 20:00 at the bottom.

Response: Thanks for the comments, we re-ordered the subfigures in Figure 14, and the profiles at 08:00 were shown at the upper panel and profiles at 20:00 were shown at the lower panel, and we limited the profiles to the height of 5km. we revised the manuscript in Lines 334-340: "YSU reproduces the vertical structure of wind speed reasonably, for example, within the low levels below 2.5 km, the simulated wind speed from the YSU scheme is similar to the observation, with model bias lower than 2.5 m/s in most cases. Meanwhile, QNSE shows worse performance in reproducing the vertical structure of wind speed, with large model bias compared to YSU. QNSE overestimates the wind speed by almost 20 m/s at 20:00, 11 January 2019, and by 30 m/s at 20:00, 12 January 2019. It is interesting to note that at 08:00, the simulations using QNSE show smaller differences with that using YSU, thus the largest differences between YSU and QNSE generally occur at specific time during the study period, which is also revealed in Figure 3a."

- Lines 214-216: "Further comparison indicates that all PBL schemes strongly overestimate the speed of north wind compared to the observations, which is the main cause of positive bias in wind speed (Figure 3)." Please elaborate. The wind speed bias could well be attributed to several wind directions. When and where does north wind occur in the simulations? Could this value be from an isolated (e.g., high elevation) station?

Response: Thanks for the comments, in the wind rose charts, the circles represent the relative frequency (%), and the colors represent wind speed (m/s), we can find in Figure 4 that for the observations, the majority of the north winds are lower than 2 m/s, at the same time, in most of the simulations, the frequency of north winds is similar to the observations, but the speed is higher than 5m/s, which may be the main cause of positive bias in wind speed. To avoid misleading, we revised it as "Further comparison indicates that all PBL schemes strongly overestimate the speed of north wind compared to the observations, which may be the main cause of positive bias in wind speed (Figure 3)" in Lines 214-216.

- Line 170: "wind speed jumps to 0 m/s due to frozen sensor" - Is there a clear "jump" when a sensor freezes? How do you distinguish between real 0m/s in wind speed vs 0m/s resulting from a frozen sensor? Do you consider measured temperatures?

Response: Thank you for the comments, the quality control of station observations was completed by Heibei Climate Center (two authors work at this agency), they were able to distinguish the jumps of the wind speed resulting from a frozen sensor because the sensors were operated by them.

- Results and conclusions: Instead of discussing each metric separately and listing the best schemes, I think it would be better to present the big picture to the reader, and discussing the schemes that have better overall scores across metrics.

Response: Thanks very much, according to the comments, we added the following in Lines 385-387: "The simulation using YSU PBL, WDM6 MP and Dudhia-RRTM SW-LW schemes shows the best performance with the

highest Taylor skill score."

- Line 291-293: "WRF simulates wind speed less accurately for coastal stations compared to inland stations." and further "the BIAS and RMSE scores are generally worse for coastal stations compared to inland stations" – actually Fig. 11 shows that bias is consistently worse for inland stations and RMSE is a tie. Revise! Is the larger ensemble spread for coastal stations (the ensemble being less confidence) maybe a better representation for the forecast than the narrower spread for inland stations that have a larger bias (the ensemble being overconfident)?

Response: Thanks very much and sorry for the error, and we totally agreed with the comments, we revised in Lines 292-294: "The statistic scores are also illustrated in Figure 11, <u>the CORR scores are consistently lower for coastal stations compared to inland stations, while the BIAS scores are generally worse for the inland stations. Thus, the model performance tends to degrade for the inland stations according to the BIAS scores.</u>"

- Lines 302-311: Could the distribution of station elevations also be included in Fig. 1? Since different marker shapes are already used for inland vs coastal station, different marker colors could be used for the elevation categories. I would still like to see a synoptic discussion explaining why the high elevation station have lower wind speeds. It is interesting that WRF wind speeds are similar or increase while observations decrease with increasing elevation.

Response: Thanks for the comments, we made changes to Figure 1, and added the information of station elevations with different marker size. The caption of Figure 1 was revised to "Solid yellow and blue circles in (a) represent coastal (16 stations in total) and inland stations (89 stations in total), the size of circles represents the station elevations, and white circle represent the sounding station."

Further investigations were needed to confirm the mechanism for lower wind speeds of the high elevation stations, and the mismatch between the observations and the WRF simulations. We added the following in Lines 305-307: "Further investigations are needed to reveal the underlying mechanisms for lower wind speed of high elevation stations and the mismatch between observations and model simulations"

- The current conclusions are more of a summary. And the current discussion section contains mainly content that belongs into the results section. Since the actual discussion is limited (I only see two discussion points: the need for bias correction and other potential sources of errors), I suggest combing the discussion with the current conclusions and renaming this section to "summary and discussion".

Response: Thanks for the comments, we revised the manuscript accordingly in Lines 369-404:

"4. Summary and discussion

……

As in our study, the model ensemble does not always provide the best performance, model post-processing, especially the bias correction techniques are needed to be taken into consideration, which can significantly reduce the systematic errors in model simulation. In addition, the PBL schemes play a dominant role in wind simulation, further tuning of the parameters within the PBL schemes, such as turbulent kinetic energy (TKE) dissipation rate, TKE diffusion factor, and turbulent length-scale coefficients is needed. In addition to PBL and SW-LW schemes, LSM and SL schemes also has unneglectable influence on the wind simulation, which should be taken into consideration in future studies. Finally, it is worth pointing out that the presented findings in this study could be

unique to the meteorological setup of the event, the location, the input dataset, the domain setup, and other unchanged parameterization types or model settings."

Technical Corrections

Technical Corrections

Line 16: "We performed 640 ~~ensemble~~ simulations using ~~multiple~~ combinations of"

Response: Thanks for the comments, it was revised to "We performed 640 simulations using combinations of 10 planetary boundary layer (PBL), 16 microphysics" in Line 16.

Lines 17/18: "Model performance is evaluated using measurements from *105* weather station observations"

Response: Thanks for the comments, it was revised to "Model performance is evaluated using measurements from 105 weather station observations" in Line 17.

Line 19: Where does the sensitivity to land surface models fit in?

Response: Thanks for the comments, the effect of LSM and SL schemes was not investigated in a systematic way as other schemes, thus we were not able to compare the sensitivity of LSM and SL to that of PBL, MP and SW-SW schemes. We added the discussion in Lines 401-402: "In addition to PBL and SW-LW schemes, LSM and SL schemes also has unneglectable influence on the wind simulation, which should be taken into consideration in future studies."

Lines 37/38: To avoid repetition: "The haze events are most frequent in boreal winter and are closely related to local weather conditions~~, with haze forming in regions~~ with low wind speeds"

Response: Thanks for the comments, it was revised to "The haze events are most frequent in boreal winter and are closely related to local weather conditions with low wind speeds" in Lines 37-38.

Lines 38-42: "Projections of future climate change suggest that global temperatures ~~will increase, and the frequency of conducive~~ and weather conditions conducive to severe haze ~~is projected to~~ will increase~~ substantially in response to the climate change, which in turn may increase the frequency of haze event over~~ affecting North China (Cai et al., 2017). *H*owever, numerical models ~~always~~ *often* show large bias in wind predictions over China […]"

Response: Thanks for the comments, it was revised to "Projections of future climate change suggest that global temperatures and weather conditions conducive to severe haze will increase affecting North China (Cai et al., 2017). However, numerical models often show large bias in wind prediction over China (Gao et al., 2016b; Zhao et al., 2016; Pan et al., 2021)," in Lines 38-41.

Lines 44-46: "In recent years, numerical models have been used extensively to study *and forecast* the weather and climate over China, as they have high spatial and temporal resolutions, and employ sophisticated physical parameterization schemes that can reproduce ~~detailed~~ atmospheric and land surface processes"

Response: Thanks for the comments, it was revised to "In recent years, numerical models have been used extensively to study and forecast the weather and climate over China, as they have high spatial and temporal

resolutions, and employ sophisticated physical parameterization schemes that can reproduce atmospheric and land surface processes (Wang et al., 2011; Zhou et al., 2019; Kong et al., 2021)." in Lines 43-45.

Lines 46/47: "However, *studies* mostly focus on temperature or precipitation, and only a few ~~studies~~ have attempted to simulate winds over China" The models simulate all variables, but *studies* focus on certain variables.

Response: Thanks for the comments, it was revised to "However, studies mostly focus on temperature or precipitation, and only a few have attempted to simulate winds over China (Li et al., 2019; Xia et al., 2019; Pan et al., 2021)." in Lines 45-47.

Line 52: "[..] as it *can* strongly influence~~s the~~ model results" Your study shows that it does not always influence model results "strongly"

Response: Thanks for the comments, it was revised to "Different physical parameterization schemes depict natural phenomena to different degrees of accuracy and choosing appropriate combinations is important, as it can strongly influences model results (Yu et al., 2011; Gómez-Navarro et al., 2015; Stegehuis et al., 2015; Gao et al., 2016b; Yang et al., 2017; Taraphdar et al., 2021)." in Lines 49-52.

Line 57: "~~A lot of~~ *Many* studies"

Response: Thanks for the comments, it was revised to "Many studies indicate an overestimation of wind speed in WRF simulations with different PBL schemes (Jiménez and Dudhia, 2012; Carvalho et al., 2014a, b; Pan et al., 2021; Gholami et al., 2021; Dzebre and Adaramola, 2020)," in Lines 56-58.

Line 59: [New sentence] "*F*or example […]"

Response: Thanks, we revised the paper according to the comments in Line 58.

Line 63: "is used instead. [New sentence] *T*he YSU scheme also shows […]"

Response: Thanks, we revised the paper according to the comments in Line 62.

Line 65: "~~There are also some~~ Other studies suggest~~ing~~ that MYNN and ACM2 are more appropriate […]"

Response: Thanks, we revised the paper according to the comments in Lines 64-65.

Lines 67-69: The word "affect" is used 3 times in two sentences – see if you can alter language more.

Response: Thanks for the comments, it was revised to "The performance of wind simulation is also influenced by the choice of cloud microphysics (MP) parameterizations. Cloud microphysical processes, such as moisture evaporation and condensation, can change thermodynamic and dynamic interactions in the atmosphere (Rajeevan et al., 2010; Santos-Alamillos et al., 2013; Li et al., 2020), then affect the vertical distribution of heat and wind fields close to the surface." in Lines 66-69.

Lines 70/71: The citation of Cheng et al. (2013) feels out of place – it considers a different season and location than this study. Maybe use it as general citation for the sentence in line 67 or with the other citations in line 69 if appropriate.

Response: Thanks for the comments, we added Cheng et al. (2013) to the general citations, now it was revised to "Cloud microphysical processes, such as moisture evaporation and condensation, can change thermodynamic and dynamic interactions in the atmosphere (Rajeevan et al., 2010; Cheng et al., 2013; Santos-Alamillos et al., 2013; Li et al., 2020)" in Lines 66-68.

Lines 76/77: "The *combination of* physical parameterizations are also vital to wind simulation, as ~~they may alter~~ the processes of atmosphere-land interactions, radiation transport, and moist convection *interact*, and *may* amplify the uncertainties in wind prediction."

Response: Thanks for the comments, it was revised to "The combinations of physical parameterizations are also vital to wind simulation, as the processes of atmosphere-land interactions, radiation transport, and moist convection interact, and may amplify the uncertainties in wind prediction." in Lines 74-75.

Line 81: "parameterization schemes. [New sentence] *To* ~~the best of~~ our knowledge, the sensitivity"

Response: Thanks, we revised the paper according to the comments in Line 79.

Line 105: New paragraph

Response: Thanks, we revised the paper according to the comments.

Figure 1: Why is there such a large difference in station number between TangShan City and QuinHuangDao City?

Response: Thanks for the comments, we got the observations form Heibei Climate Center, which belongs to the China Meteorology Administration (CMA), during the study period, there are more observations available at QinHuangDao City than other cities.

Figure 2 caption: "The daily averaged geopotential height (contour *lines*, units: gpm) and winds (vectors, units: m/s) *at 925 hPa* and cloud fraction (shading, units: %) ~~at 925 hPa~~ during 11 *through* 15 January 2019"

Response: Thanks for the comments, the caption was revised to "The daily averaged geopotential height (contour lines, units: gpm) at 500 hPa, total cloud fraction (shading) and surface winds at 10 meters (vectors, units: m/s) from ERA5 during 11 through 15 January 2019, the box indicates D03 in Figure 1"

Line 130: "All ~~the~~ simulations …"

Response: Thanks, we revised the paper according to the comments in Line 127.

Lines 148/149: "In this study, ETA, QNSE, MYNN, Pleim-Xiu, and TEMF *SL* schemes are chosen separately for PBL schemes of MYJ, QNSE, MYNN, ACM2, and TEMF, *respectively(?)*." All the SL schemes need references too.

Response: Thanks, we revised the paper according to the comments in Lines 145-146: "In this study, ETA, QNSE, MYNN, Pleim-Xiu, and TEMF SL schemes are chosen separately for PBL schemes of MYJ, QNSE, MYNN, ACM2, and TEMF, respectively". References were listed in Table 4.

Table 1: Either clarify in the caption that the schemes that share rows are not specifically assigned to each other

(except for SW-LW), or change back to the previous layout, with columns “Parameterization Type”, “Scheme” and “Reference”, then in the rows list all PBL, MP, SW-LW schemes.

Response: Thanks, we revised the caption of Table 1 to “Table 1: List of microphysics (MP), planetary boundary layer (PBL), and shortwave-longwave radiation (SW-LW) schemes investigated in the 640 simulations, schemes that share rows are not specifically assigned to each other (except for SW-LW).”

Line 176/177: “including ~~the~~ Pearson’s correlation coefficient (CORR)” or “including the Pearson~~’s~~ correlation coefficient (CORR)”

Response: Thanks for the comments, it was revised to “Several metrics are employed for evaluating the performance of each model configuration, including the Pearson correlation coefficient (CORR)” in Line 172.

Lines 183/184: Perhaps also mention what the perfect scores for the other metrics are.

Response: Thanks for the comments, we added “while higher CORR, lower BIAS and RMSE scores indicate better model simulations.” in Line 184.

Line 188: “… vector notation approach. *[New sentences] C*ircular correlation coefficient …”

Response: Thanks, we revised the paper according to the comments in Line 188.

Line 194: “Figure 3a shows the time series of observed and simulated wind speeds at local time. *Model wind speeds are shown for different PBL schemes averaged over all other parameterization types.*” Also please clarify whether the day starts with the background shading or at the tick marks.

Response: Thanks for the comments, it was revised to “Figure 3a shows the time series of observed wind speed in local time. Model wind speeds are shown for different PBL schemes averaged over all other parameterization types.” in Lines 194-195.

We clarified in the caption: “Figure 3: (a) Time series of observed and simulated wind speeds (m/s) and the corresponding statistics of (b) CORR, (c) BIAS and (d) RMSE for the PBL schemes. In a, the frequency of wind speed is hourly, and the tick marks in x-axis indicate 12:00 in local time of that day, for each PBL scheme, the average is calculated over the 105 stations and then over all the simulations with that scheme; the dots in b, c and d represent the range across the stations, for each station, the metrics are calculated by averaging all the simulations with the specific PBL scheme”.

Line 200: “*smaller* difference” or better “more similar to measurements”

Response: Thanks for the comments, it was revised to “while YSU is more similar to the measurements during 14-15 January 2019.” in Line 200.

Line 203: “… Figure 3b-d. *[New sentence]* MYJ shows the best CORR score of 0.96*; [semicolon]* MYNN, ACM2 and UW are *next best*…”

Response: Thanks for the comments, it was revised to “The statistics of CORR, BIAS, and RMSE are illustrated in Figure 3b-d. MYJ shows the best CORR score of 0.96; MYNN, ACM2 and UW are next best according to this verification score” in Lines 203-204.

Line 223/224: They are not "the same", but "very similar" (perhaps use "to a precision of 0.XX [fill in the correct precision number]")

Response: Thanks for the comments, it was revised to "The CORR scores are very similar for all the MP schemes at a precision of 0.01, while MYDM7 is the best scheme according to BIAS and RMSE scores" in Line 225.

Line 235: "Strong overestimation" seems extreme considering that total errors are often within a 1m/s margin and not much different from the other two schemes. Maybe: "RRTMG and CAM show larger overestimation than Dudhia-RRTM and Goddard at daytime peaks."

Response: Thanks for the comments, it was revised to "RRTMG and CAM show a larger overestimation than Dudhia-RRTM and Goddard at daytime peaks." in Lines 235-236.

Line 242: "BIAS of -15.7" – be consistent with rounding precisions (please check everywhere in the manuscript)

Response: Thanks for the comments, it was revised to "the RRTMG scheme shows the best BIAS of -15.69°, and Dudhia-RRTM shows the best RMSE of 61.13°." in Line 242.

In Lines 206-207: "for the schemes except for QNSE, the range of CORR is 0.18-0.88, the range of BIAS is -2.10-2.91 m/s, and the range of RMSE is 0.79-3.85 m/s"

Line 218: "while TEMF shows the best BIAS and RMSE scores of -11.33° and 56.19°."

Line 248/249: "the results are expected to be consistent with evaluations using other MP schemes" – why? If you see differences between Fig 9a and 9b, is it not plausible that there would also be differences with other MP schemes?

Response: Thanks for the comments, as we cannot show all the combinations of PBL, MP and SW-LW schemes, the results of combinations using MP scheme of MYDM7 and P3 were chosen as examples, and we can find that the interactions among parameterization schemes are also important. We expected that there should be differences between MYDM7 and P3 results (fig 9a and 9b), however, the results from the investigation should be consistent (the interactions among parameterization schemes are also important). To avoid misleading, we revised as "thus for each MP scheme, a total of 36 simulations (excluding QNSE) are evaluated. For wind speed simulations with both MYDM7 and P3, MYJ shows the best CORR score, while YSU shows the lowest BIAS and RMSE scores".

Line 251: "… Dudhia-RRTM or Goddard. *[Period]"*

"when [YSU?] is applied with RRTMG, the BIAS and RMSE scores show an obvious increasement compared with that with Dudhia-RRTM" – I can't follow.

Response: Thanks for the comments, it was revised to "However, it is worth noting that YSU shows pretty low BIAS (< 0.4 m/s) and RMSE (<0.6 m/s) scores only when combined with the Dudhia-RRTM or Goddard schemes, when it is combined with RRTMG schemes, the BIAS and RMSE scores increase a lot" in Lines 257-259.

Line 252: "[New paragraph] For the wind direction simulation …"

Response: Thanks, we revised the paper according to the comments in Line 252.

Line 257/258: "this indicates that a systematic variation of parameterizations is not necessary" & "this indicates that a systematic variation of parameterizations is important when focusing on these variables." (line 261) – These are my words from the previous review – it was meant as a comment on the authors' motivation for this study, however, to the reader this wording is confusing here. Please use your own words to rephrase this depending on what you wish to say.

Response: Thanks for the comments, it was revised to "Overall, <u>for BIAS and RMSE scores of wind speed, within each PBL scheme, the same SW-LW group ranks best, and within each SW-LW group, the same PBL scheme ranks best. For example, no matter which SW-LW group, YSU is always the best, which indicate the good performance of YSU</u>. However, it is worth noting that YSU shows pretty low BIAS (< 0.4 m/s) and RMSE (<0.6 m/s) scores only when combined with the Dudhia-RRTM or Goddard schemes, when it is combined with RRTMG schemes, the BIAS and RMSE scores increase a lot. <u>For the wind direction simulation, the pattern is different from that of wind speed. For example, for BIAS and RMSE scores, the best PBL scheme depends on the choice of SW-LW schemes,</u> which indicates the influence of scheme interaction on model performance in wind direction simulations." in Lines 255-261.

Figure 10: Please note in the figure caption that only MP schemes are labeled because all configurations use the same PBL and SW/LW schemes. You may also highlight in your discussion that the best individual schemes can reduce the bias of the ensemble by ~50%, which is significant.

Response: Thanks for the comments, the caption was revised to "Time series of wind speed (m/s) from observation and different ensembles, and (b) CORR, (c) BIAS and (d) RMSE scores for the best 10 simulations along with different ensembles. The shading in a shows the spread for ENS(4) and ENS(576). As the best 10 simulations use the same PBL (YSU) and SW-LW (Dudhia-RRTM) schemes, only the MP schemes are labeled in the figure."
Lines 274-276: "According to the statistic scores, ENS(4) reduces model bias by approximately half compared to ENS(576). At the same time, the best individual schemes (NSSL1, MYDM7, P3 and ETA) can also reduce the bias by ~50%."

Line 275: Be consistent with abbreviations ENSall in test body vs ENS(576) in Fig. 10. (I prefer ENS(576).)

Response: Thanks, we revised the figure and text according to the comments, we used ENS(576). The revision are in Lines 271-276: "Figure 10 also shows the time series of wind speed from ensemble of 4 [ENS(4)) and all 576 simulations [ENS(576)], the spread of ENS(4) is significantly lower than that of ENS(576), and ENS(4) shows smaller difference with the observation compared to ENS(576). According to the statistic scores, ENS(4) reduces model bias by approximately half compared to ENS(576). At the same time, the best individual schemes (NSSL1, MYDM7, P3 and ETA) can also reduce the bias by ~50%."

Line 281/282: "~~The effects of these parameters on the model results are presented below."~~ Redundant. Skip.

Response: Thanks, we revised the paper according to the comments.

Line 284/285: "Figure 11 compares the results of wind speed for coastal stations (closer than 5 km from the shoreline, 89 stations in total) and inland stations (over 5 km from the shoreline, 16 stations in total), the locations of these stations are shown in Figure 1a." – In Fig 1a looks there are more inland stations than coastal stations – should it be 89 inland & 16 coastal station?

Response: Thanks for the comments and sorry for the mistakes, we corrected them as "Figure 11 compares the results of wind speed for coastal stations (closer than 5 km from the shoreline, 16 stations in total) and inland stations (over 5 km from the shoreline, 89 stations in total), the locations of these stations are shown in Figure 1a.".

Line 289/290: ", ~~especially among the MP schemes during 11-13 January 2019 with lower wind speed~~" It looks like this is valid everywhere.

Response: Thanks, we revised the paper according to the comments.

Sections 4.1-4.4 should be part of the results (under section 3)

Response: Thanks, we revised the paper according to the comments.

Line 325: "which is *similar* to the observations"

Response: Thanks, we revised the paper according to the comments in Line 325.

Line 337/338: Do you mean Jan 12th for one of the two? To me it looks almost like 30m/s on the 12th, and like 20m/s on the 11th.

Response: Thanks for the comments and sorry for the mistakes, we revised in Lines 337-338: "QNSE overestimates the wind speed by almost 20 m/s at 20:00, 11 January 2019, and by 30 m/s at 20:00, 12 January 2019."

Line 336: This sentence states generally "significant" differences between YSU and QNSE with an example; later the text says they are "pretty similar".

Response: Thanks for the comments and sorry for the mistakes, we revised in Lines 338-340: "It is interesting to note that at 08:00, the simulations using QNSE show smaller differences with that using YSU, thus the largest differences between YSU and QNSE generally occur at specific time during the study period, which is also revealed in Figure 3a."

"For example" (capitalize at the beginning of a new sentence)

Response: Thanks, we revised the paper according to the comments.

Lines 362-365: This fits in better in the methods sections where the metrics are first introduced.

Response: Thanks, we revised the paper according to the comments in Lines 173-176: "CORR is a measure of the strength and direction of the linear relationship between simulation and observation, BIAS is a measure of mean difference between simulation and observation, and RMSE is the square root of the average of the set of squared differences between simulation and observation, thus each of this score gives a partial view of the model performance."

Lines 367-370: This fits in better with section 3.1.5 where the ensemble results are shown.

Response: Thanks, we revised the paper according to the comments in Lines 276-279: "It is worth to mention that the best CORR score of ENS(576) is also result of single model simulation with Goddard MP scheme. At the same

time, the best BIAS score (0.33 m/s) is result of the single model simulation with MYDM7 or ETA, and the best RMSE score (0.52 m/s) is result of either the single model simulation with Goddard, NSSL1, MYDM7, or the ensemble using the 3, 4 or 5 best simulations."

Line 381: "during a relatively long period of stable conditions" – see my comment above

Response: Thanks for the comments, it was revised to "we investigate wind simulations under stable conditions when a haze event affected North China" in Line 370.

Line 391: "*followed* by MYNN"?

Response: Thanks for the comments, it was revised to "YSU is the best scheme according to the BIAS and RMSE scores (0.45 m/s and 0.61 m/s), followed by MYNN (0.55 m/s and 0.70 m/s)" in Lines 383-384.

The authors should try to vary their language more. Many sentences start with "further investigation" and "further comparison". This phrase can be omitted.

Response: Thanks for the comments, we checked the manuscript and delete the phrase of "further investigation" and "further comparison".