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

Title: 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

Author(s): Entao Yu et al.

MS No.: gmd-2022-53

MS type: Model evaluation paper

General Comments

This paper presents a sensitivity case study of WRF wind forecasts under calm and stable conditions with a systematic variation of planetary boundary layer (PBL), microphysics (MP), and radiation parameterizations. For a case study this work is result of an extensive computational effort. However, there are some aspects in the methodology that need clarification and better motivation. Considering the impressive volume of the generated model data, the presented evaluation is limited, and I suggest expanding on the analysis to improve the scientific quality of the paper. Several findings require more insightful interpretations and discussions; some of the presented conclusions require clarification or correction. The presentation of the data and results can be improved upon.

Response: Thank you for taking time out of your busy schedule to review this paper, I 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) This study aims to assess the ability of various physics parameterization configurations to predict **calm and stable weather conditions that favor air pollution**. However, the authors verify merely wind speed (and wind direction to a lesser degree). In order to gain insight what drives these differences in wind speed and to evaluate atmospheric properties that are crucial for air quality (such as static stability), it would be valuable to further assess vertical profiles.

Response: Thanks for the suggestion, according to this comment, we extended the analysis of wind direction by adding two figures (figure6 and figure8), and added the evaluation of vertical profile using a sounding station (location shown in figure1) in section 4.2, the revisions are as follows:

(1) Lines 227-230 and figure 6: "The sensitivity of wind direction to the MP schemes is also low, as the wind roses from simulations with different MP schemes are very similar (Figure 6). WDM6, NSSL2 and

ThompsonAA show the best CORR score of 0.52, followed by Thompson and CAM5. Meanwhile, WSM3 is the best scheme according to the BIAS score, and ThompsonAA is the best scheme according to the RMSE score".
(2) Lines 239-242 and figure 8: "Figure 8 shows the wind roses during 11-15 January 2019 from simulations with different SW-LW schemes and the corresponding statistic scores. The simulations indicate the wind is mostly from the southwest direction during the study period, which is different from the observation. According to the CORR score, Dudhia-RRTM is the best scheme with the highest value (0.55), meanwhile, RRTMG shows the best BIAS score of -16°, and Dudhia-RRTM shows the best RMSE score of 61°".
(3) Lines 331-340 and figure 14: "4.2 Vertical profile of wind speed
Figure 14 shows the observed and simulated vertical profile of wind speed at 08:00 and 20:00 during the study period, the location of the sounding station is illustrated in Figure 1. YSU reproduces the vertical structure of wind speed reasonably, with slightly larger model bias above the height of 15 km. Within the low levels below 2.5 km, simulated wind speed from the YSU scheme is close to the observation, with the bias lower than 2.5 m/s in most cases. Meanwhile, QNSE shows worse performance in reproducing the vertical structure of wind speed, with significant larger model bias compared to YSU. for example, QNSE overestimates the low-level (< 2.5 km) wind speed by about 10 m/s at 20:00 on 11 January 2019, and overestimate wind speed by 20 m/s at 20:00 on 11 January 2019. It is interesting to note that the simulation with QNSE is pretty similar to that with YSU at 08:00 during the study period, indicating that large difference between YSU and QNSE only occurs at specific time during the study period, which is also revealed in Figure 3a."

(2) This paper presents a **case study** for a specific location and event, and the manuscript should be framed accordingly (in the title, results, and discussion). However, the authors draw general conclusions from their findings and presume transferability of their results across the world and across variables (e.g., lines 81-83 and lines 337/338) and generalize shortcoming in the WRF model (lines 305/306). The current manuscript lacks an objective error discussion on the various sources of uncertainty and limitations of the study: The presented findings could be unique to the meteorological setup of the event, the location, the input dataset, the domain setup, other unchanged parameterization types or model settings, etc.
Response: Thanks for the suggestion, the points are well taken, we removed the general statements and focused our findings to the specific case and location, as follows:
(1) Lines 28-30: "Our results indicate the roles parameterizations play in wind simulation under stable weather conditions and provide a valuable reference for further research in the study area and nearby regions."
(2) Lines 91-92: "be helpful in the wind and air quality forecast in the study area and other coastal regions of China under stable weather conditions."
(3) Lines 377-379: "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."

(3) It is not clear how the authors arrive at the conclusion that "**wind direction** is insensitive to changes in physics parameterizations" (line 207 & 213-15). The manuscript only shows variations between PBL-scheme groups and differences are visible in figure 4 and table 5. The authors have not performed any hypothesis testing to show whether these differences are actually insignificant compared to the variation in wind speeds. The authors write that "the model RMSE and BIAS values for different PBL schemes are similar", yet table 5 reveals that the worst PBL scheme (namely, BouLac) has a wind-direction Bias that is 154% larger than the Bias of the best PBL scheme (namely, LES – excluding QNSE), see table 5. In comparison the wind- speed Bias of the worst PBL scheme (namely, TEMF) is 126% larger than the Bias of the best PBL scheme (namely YSU), see table 4.

Response: Thanks for the comments, the points are well taken, and we added the results of wind direction for MP and radiation schemes (Figure 6 and Figure 8), and we extended the analysis of wind direction in Figure 9, as follows:

(1) Lines 227-30 and figure 6: "The sensitivity of wind direction to the MP schemes is also low, as the wind roses from simulations with different MP schemes are very similar (Figure 6). WDM6, NSSL2 and ThompsonAA show the best CORR score of 0.52, followed by Thompson and CAM5. Meanwhile, WSM3 is the best scheme according to the BIAS score, and ThompsonAA is the best scheme according to the RMSE score".

(2) Lines 239-242 and figure 8: "Figure 8 shows the wind roses during 11-15 January 2019 from simulations with different SW-LW schemes and the corresponding statistic scores. The simulations indicate the wind is mostly from the southwest direction during the study period, which is different from the observation. According to the CORR score, Dudhia-RRTM is the best scheme with the highest value (0.55), meanwhile, RRTMG shows the best BIAS score of -16°, and Dudhia-RRTM shows the best RMSE score of 61°".

(3) Lines 252-253: "For the wind direction simulation, LES combined with Dudhia-RRTM shows the best CORR score, while TEMF is the best scheme according to BIAS and RMSE scores."

(4) A considerable amount of computational resources was spent to test **16 MP schemes** - this is the largest number of schemes within any of the parameterisation categories tested in this paper. Although the authors cite Cheng et al. (2013) to justify that MP can have an impact on wind fields, there is little justification to evaluate MP sensitivity to this extent under the dry and stable conditions of this case study. As opposed to the conditions described in this manuscript, Cheng et al. (2013) pointed out that MP can affect wind fields for convective weather phenomena during the summer (with weak large-scale forcing) that are associated with gust fronts / outflow boundaries from cold pools that result from strong downdrafts of thunderstorms. One would not expect to see significant impact from the choice of MP parameterization on wind forecasts under stable conditions, hence, these results are not surprising. Instead of assessing a vast amount of MP schemes without valid motivation, it could have been informative to expand on the number of radiative schemes and/or include an investigation on the impact of the land surface schemes.

Response: Thanks for the comments, we totally agree with the comments on the MP schemes, we added an investigation on the impact of land surface schemes in section 4.4, lines 350-360:

"4.4 Impact of land surface model

Figure 16 shows the evaluation of different land surface parameterizations with the same model configuration as the simulation with best the Taylor skill score, the land surface models (LSM) considered are the five-layer

thermal diffusion scheme (SLAB, Dudhia, 1996), the Noah scheme (NOAH, Chen and Dudhia, 2001), the Rapid Update Cycle scheme (RUC, Smirnova et al., 2000), the Noah-MP scheme (NOAHMP), and the Community Land Model Version 4 scheme (CLM4, Lawrence et al., 2011). The simulations with different LSMs reproduce the timeseries of wind speed well, with larger spread among the LSMs during 14-15 January 2019 (Figure 16a). NOAHMP shows the best CORR score of 0.94, CLM4 and NOAH are slightly worse according to this score. Meanwhile, according to the RMSE score, NOAHMP is the best scheme, followed by RUC and NOAH. In addition, RUC and NOAHMP show better BIAS scores than the other LSMs. Thus, NOAHMP shows the best performance among different LSMs in wind-speed simulation under stable conditions in this study, however, the large difference among LSMs indicates that we should take land surface parameterizations into consideration in future studies."

(5) **Tables and Figures:**

Tables 4-10 show verification data values that are not always easily to grasp from plain numbers in this quantity. The authors should consider visualizing this data for better presentation to the reader. For example, this could be in the form of bar plots, boxplots or heatmaps. I also encourage the authors to include the distribution across the 105 stations (e.g., the range across stations could be shown in box plots or with error bars on bar or line plots). The verification plots could be combined with the respective timeseries plot (e.g., figure 3 would have a subplot visualising the error metrics from table 4).

Response: Thanks for the comments, according to the suggestion, we plotted the data of Tables 4-10 as subplots of the timeseries figures, and the range across stations was included in the subplots. Thus, in the revision, Tables 4-10 were removed and shown as subplots in Figure 2-8, 10-12, the analysis of these figures are revised accordingly.

Figure 2: This figure shows identical data to other figures - i.e., 2a is repeated in figures 3, 5, 6, and 8, whereas 2b is identical to the top left panel in figure 4. This makes figure 2 redundant. An alternative could be to show the observational data of Tangshan city as a specific example in connection to the paragraph in line 96-101. When updating, please also include a more descriptive title to figure 2a.

Response: Thanks for the comments, figure 2 is redundant, thus we removed it, and added synoptic plots of the stable weather event as figure 2 in the revision, we also added the information of Tangshan city in figures 1 and 2.

The description of wind rose charts are revised to "for each wind rose chart, the circles represent the relative frequency (%), and the colors represent wind speed (m/s)" in lines 658-660.

(6) Lines 247-249: For wind speed Bias and RMSE, within each radiation group the same PBL schemes rank best; and within each PBL group, the same radiation schemes rank best. This indicates that a systematic variation of parameterizations as presented in this paper, is *not* necessary. (E.g., for wind-speed RMSE no matter which PBL scheme, Dudhia or Goddard radiation are always best.) However, for CORR and wind direction, this pattern is not always consistent, which indicates that a systematic variation of parameterizations *is* useful when focusing on these variables. (E.g., for wind-direction RMSE, the best radiation scheme depends on the choice of PBL scheme

- for TEMF PBL, RRTMG radiation is best; for BouLac PBL, Goddard radiation is best; for LES, CAM radiation is best.)

Response: Thanks for the comments, points are well taken, we revised in lines 255-261: "Overall, for BIAS and RMSE scores of wind speed, within each SW-LW group, the same PBL scheme ranks best (e.g., for wind-speed RMSE, no matter which PBL scheme, Dudhia-RRTM and Goddard are always best), and within each PBL scheme, the same SW-LW group ranks best, this indicates that a systematic variation of parameterizations is not necessary. However, for wind-speed CORR and wind direction, this pattern is not always consistent. For example, for wind-direction BIAS, the best SW-LW group depends on the choice of PBL scheme (e.g., for MYDM7 with TEMF, Dudhia-RRTM is best; for P3 MP scheme with TEMF, RRTMG is best), this indicates that a systematic variation of parameterizations is important when focusing on these variables".

(7) **Evaluation of model configurations with the best individual performance** (lines 254-269):

Lines 266/267: This statement is misleading. Perhaps correct to "the ensemble of the top *four* configurations reduces model bias by approximately half *compared to the ensemble that uses all configurations,* while the CORR value *of the super-ensemble mean* was highest *among* all the *ensembles*." Note that the highest CORR seen in table 8 (0.937) is result of either the single-model configuration using Goddard MP, or the ensemble using all configurations. The lowest BIAS (0.331) is result of the single-model configuration using MYDM7. And the lowest RMSE (0.524) is result of either the single-model configuration using NSSL1, or the ensemble using the 4 best configurations. This disagrees with the concluding statement in line 339-341.

Response: Thanks for the comments, it was revised in lines 273-277: "The result indicates that ensemble mean of four simulations with WDM6, Goddard, NSSL1 and MYDM7 shows the best BIAS and RMSE scores. Figure 10a shows the time series of wind speed from ensemble of 4 (ENS4) and all simulations (ENSall), the spread of ENS4 is significantly lower than ENSall, and ENS4 shows smaller difference with the observation compared to ENSall. According to the statistic scores, ENS4 reduces model bias by approximately half compared to ENSall".

Lines 369-373: "The ensemble mean of 576 simulations in our study shows the best CORR score (0.94) in wind-speed simulation, however, this best CORR score is also result of single-model simulation with Goddard MP scheme. At the same time, the best wind-speed 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. Thus, model ensemble does not always provide the best performance"

For simulations that struggle with systematic overprediction, it is implied that an ensemble of a subgroup of members with smaller biases improves the ensemble-mean bias compared to using all members. However, (a) systematic errors can be significantly reduced with bias- correction, and (b) ensembles generate probabilistic forecasts and the authors present no discussion on the probabilistic features of their ensembles (e.g., ensembles with narrow spread are often under-dispersive / overconfident). More on this in point 8).

Response: Thanks for the comments, according to them, we revised the discussion in lines 367-375: "As none

of the schemes is the best according to all the scores, model ensemble is used to provide optimizing model performance, at the same time, model ensemble also provides probabilistic evaluation of the simulations and ensembles with narrow spread are often overconfident. The ensemble mean of 576 simulations in our study shows the best CORR score (0.94) in wind-speed simulation, however, this best CORR score is also result of single-model simulation with Goddard MP scheme. At the same time, the best wind-speed 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. Thus, model ensemble does not always provide the best performance, and 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."

Please clarify if the SD in the Taylor skill scores was calculated over the various stations and averaged over time, or if SD was calculated over the time series and averaged over stations. This is important to understand why the MP scheme with best CORR is not also the scheme with best Taylor skill score considering that all MP schemes have very similar temporal patterns (as seen figure 5). If the variation across stations is significant among MP schemes, it would be informative to analyse this spatial variation.

Response: Thanks for the comments, SD is calculated over the time series and averaged over stations, and the variation across stations is significant among MP schemes. According to the comments, we added the range across the 105 stations in the revised figures, and we added an investigation on the spatial variation in section 4.1, lines 319-330: "Figure 13 shows the spatial distribution of observed and simulated wind fields during the study period, we choose 14:00 in local time as an example. The simulation using YSU, Dudhia-RRTM and WDM6 schemes is referred to as YSU, and the simulation using QNSE, Dudhia-RRTM and WDM6 is referred to as QNSE. YSU generally reproduces the wind field in the study area, especially in terms of wind speed. For example, the observed wind speed is lower on 13 January 2019, with values lower than 2 m/s in many stations, while on 15 January 2019, the observed wind speed is higher than 4 m/s in most of the stations. In the simulation with YSU, wind speed is about 2 m/s on 13 January 2019 and higher than 4 m/s on 15 January 2019 over the study area, which is close to the observation. On the contrary, simulation with QNSE fails to reproduce the distribution of wind speed, and shows strong overestimation, especially over the mountain areas of the study area (Figure 1a), for example, the peak wind speed in simulation with QNSE exceeds 20 m/s on 15 January 2019, which is more than five times greater than the observation, this overestimation is consistent with the large positive bias in previous investigation of Figure 3. For the wind-direction simulation, YSU shows degraded performance compared to wind speed, and generally fails to reproduce the wind-direction distribution for most of the stations, QNSE also fails to do so."

Table 8 – If the 10 best WRF configurations are based on the ranks of Taylor skill scores, please add the ranks with corresponding Taylor skill score values to the table.

Response: Thanks for pointing out this, it was revised in line 264: "the scores range from 0.2 to 1.0, with the best 10 WRF configurations having similar scores of about 1.0".

(8) The authors present correlation coefficients, biases, and RMSE without any insightful discussion what the different verification metrics represent and why it is conceivable that some schemes perform best according to one metric and worst according to another. It is also not discussed that the performance of raw model forecasts can be significantly enhanced by **post-processing** - in particular the systematic-errors component (biases), which appear to be the main issue in this case study. As already mentioned in point 7), the author's suggestion of the ensemble generation by picking a small number of members with lowest bias without considering the effects of calibration is problematic.

Response: Thanks for the comments, we revised them in lines 362-375: "In this study, CORR, BIAS and RMSE are used as verification scores, CORR is a measure of the strength and direction of the linear relationship between simulation and observation, BIAS is a measure of the 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, and some schemes perform best according to one metric and worst according to another in our previous investigation.

As none of the schemes is the best according to all the scores, model ensemble is used to provide optimizing model performance, at the same time, model ensemble also provide probabilistic evaluation of the simulations as ensembles with narrow spread are often overconfident. The ensemble mean of 576 simulations in our study shows the best CORR score (0.94) in wind-speed simulation, however, this best CORR score is also result of single-model simulation with Goddard MP scheme. At the same time, the best wind-speed 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. Thus, model ensemble does not always provide the best performance, and 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."

(9) Section 2.1: Please add a more comprehensive **synoptic analysis** of the event. Consider adding a weather analysis map and/or radio soundings to show the observed stable stratification (perhaps in figure 2). Considering the substantial testing of MP schemes, it would be useful to mention if there was any cloudiness. With regards to section 3.2.2 it would be interesting to analyse why observed wind speeds decrease with elevation.

Response: Thanks for the comments, we removed figure 2 and added the distribution of cloud, geopotential height and winds as a new figure. The descriptions are presented in lines 107-115: "Figure 2 depicts the distribution of geopotential height, winds, and cloud fraction during the study period, the geopotential height and winds data are from the ERA5 dataset (Hersbach et al., 2020), and the cloud fractions are from satellite observations of CLARA (CM SAF cLoud, Albedo and surface Radiation) product family (Karlsson et al., 2021). Figure 2 indicates that a weak high-pressure system persisted from 11 to 13 January, along with weak southwest wind in the study area, which would transport warm and wet air to the study area (Gao et al., 2016a), creating a favorable moisture condition for stable conditions and inhibiting pollutants dispersal (Zhang et al., 2014; Hua and Wu, 2022). Then the high-pressure system was replaced by strong northwest wind from 14 to 15 January 2019. The CLARA observations indicate cloud fraction exceeding 60% on 12

January at the study area, while for the rest of the time, cloud fraction is low. This stable event is used to investigate the impact of physical parameterizations of the WRF model".

(10) Ensemble spread usually grows with **forecast lead time** as predictive skill declines. Please discuss how the authors explain the narrowing ensemble spread in figure 5.

Response: Thanks for the comments, it is true that in weather forecasting operations, the ensemble spread of WRF model usually grows with forecast lead time as predictive skill declines, which is closely related to the growth of uncertainties in the global climate models that provide initial and lateral boundary conditions for the WRF model. In our study, the WRF model is driven by the ERA5 reanalysis dataset, so the simulations can be considered as dynamical downscaling simulations with imposed "perfect boundary conditions", thus the uncertainties of driving data will not grow significantly with time. At the same time, all the WRF simulations are initialized at the same starting points (00:00 UTC January 9th, 2019) with identical lateral boundary conditions, and they are configured as "climate simulations" with SST updated every 3 hours. All the aforementioned setups help to minimize the external uncertainties, making the variation in the results are only caused by the physical parameterization schemes. As the wind speed under stable conditions are insensitive to the MP parameterizations, the ensemble spread in that figure is narrow.

Technical Corrections

The title needs to state that (a) this is a case study and (b) calm wind speeds under stable conditions are investigated.

Response: Thanks for the comments, the title was revised to "Impact of physical parameterizations on wind simulation with WRF V3.9.1.1 under stable conditions at PBL gray-zone resolution: a case study over the coastal regions of North China".

There are inconsistencies with the tense, please make consistent everywhere.

E.g., line 25/26 "The MYNN scheme *showed* the highest correlation among all PBL schemes, while LES and YSU *has* the smallest model errors, the RRTMG and Goddard schemes *showed* the highest correlations …"

Response: Thanks for pointing out this, we revised it in lines 23-25: "for coastal stations, MYNN shows the best temporal correlation with observations among all PBL schemes, while Goddard shows the smallest bias out of SW-LW schemes, these results are different from that of inland stations."

We also checked the whole paper and corrected the errors.

Line 14: Technically wind speeds are always zero at the surface. The wind simulations in this paper likely correspond to the 10-m above surface level.

Response: Thank you for pointing out this, "surface wind" was replaced by "near-surface wind at 10-meter height" in Line 14 as: "different physical parameterization schemes impact simulated near-surface wind at 10-meter height over the coastal regions of North China", we also checked and revised them throughout the

paper.

Line 18: "The data show that *the* WRF model …"

Response: Thank you for pointing out this, the errors were corrected.

Line 25: "For example, for the weather stations located in coastal regions [*no new sentence*] the MYNN …"

Response: Thank you for pointing out this, the errors were corrected.

Line 35/36: "(Zhang et al., 2014; Cai et al., 2017; Zhang et al., 2015)" – please sort references either alphabetically or according to publication time (check journal guidelines) throughout the manuscript

Response: Thank you for pointing out this, the references were checked and the errors were corrected.

Line 36-39: Please provide a brief explanation on how increasing global temperatures relate to haze and low wind speeds.

Response: Thanks for the comment, the explanation was added as follows in lines 38-41: "Projections of future climate change suggest that global temperatures will increase, and the frequency of conducive weather conditions to severe haze is projected to increase substantially in response to the climate change, which in turn may increase the frequency of haze events over North China (Cai et al., 2017),".

Line 39/40: Before saying that it is crucial to improve wind predictions because haze events are hazardous and may become more frequent in future, inform the reader about the problems with WRF wind forecasts - i.e., include a short literature review on known biases and challenges in simulating wind fields in your region of interest.

Response: Thanks for pointing out this, we revised it in lines 40-43: "frequency of haze events over North China (Cai et al., 2017), however, numerical models always show large bias in wind prediction over China (Gao et al., 2016b; Zhao et al., 2016; Pan et al., 2021), thus it is crucial to improve wind prediction under stable weather conditions in order to minimize associated economic losses and environmental impacts".

Line 49: "choosing appropriate combinations is ~~extremely~~ important"

Response: Thanks for pointing out this, the errors were corrected.

Line 67: Start a new paragraph to separate the sections on MP vs radiation.

Response: Thank you for pointing out this, we revised the paper according to this comment.

Line 72/73: Add "to this extent", "in China" and "to our knowledge". WRF wind performance has been evaluated in a systematic manner in other studies, e.g., Fernández-González et al. 2018, Santos-Alamillos et al. 2013, Siuta et al. 2017.

Response: Thank you for pointing out this, according to the comment, it was revised as follows in lines 80-82: "Most of the aforementioned studies considered a small number of parameterization schemes, to the best of

our knowledge, the sensitivity of parameterizations on wind simulation has not yet been explored in a systematic way in China."

We also mentioned other systematic evaluations in lines 77-79: "The impact of parameterization combination on WRF performance has been investigated in previous studies (Santos-Alamillos et al., 2013; Fernández-González et al., 2018)"

Line 74/75: "error compensation among processes […] may predict incorrect wind patterns" If errors compensate each other, that would imply that the result would be more accurate; however, in the chaotic system of the atmosphere errors are usually amplified and grow from various imperfections in the model.

Response: Thanks for pointing out this, according to this comment, we revised in lines 76-77: "The interactions among physical parameterizations are also vital to wind simulation, as they may alter the processes of atmosphere-land interactions, radiation transport, and moist convection, and amplify the uncertainties in wind prediction."

Lines 77 & 81: "*the* WRF model" (also check everywhere else in the manuscript)

Response: Thank you for pointing out this, the errors were corrected. We also checked and corrected the errors in other places.

Lines 82: "founding"?

Response: Thank you for pointing out this, it was corrected by "finding" in the revision.

Line 86: The caption repeats all sub-captions. I suggest re-naming it to "Data and Methods" or similar.

Response: Thank you for pointing out this, the title was revised to "Data and Methods".

Line 87: "stable weather events in 2019" implies a series of events – this paper only discusses a single event!

Response: Thanks for the comment, according to it, we revised the title to "2.1. Study area and the stable weather event in 2019".

Line 92: "favorable weather conditions" – be more specific please

Response: Thanks for pointing out this, it was revised to "favorable weather conditions with lower wind speed" in line 101.

Line 94: "synoptic *forcing*"?; "vertical mixing *in* the atmosphere"; "increasing the stability of surface air" – air in the boundary layer?

Response: Thank you for pointing out this, according to the comment, it was revised to "anomalous southerly wind in the lower troposphere caused by the weak East Asian winter monsoon weakened the synoptic forcing and extent of vertical mixing in the atmosphere, thus increased the stability of air in the boundary layer and the local concentration of hazes (Zhang et al., 2014)" in lines 102-105.

Line 97: Could the location of Tangshan city (or the station measuring these values) please be included to figure 1? Thank you!

Thank you for pointing out this, the location of Tangshan city was included in Figures 1 and 2.

Line 98: Are any statistics available on the socioeconomic or health impacts of this event that could be included to exemplify the severity of the event (e.g., increased hospitalization rates)?

Response: Thanks, this is a very good suggestion, however, we did not get such economic information for this haze event from the public sources. Sorry for that.

Line 104: Plural "simulations"

Response: Thank you for pointing out this, the errors were corrected.

Line 109: Better "within the PBL" specifically for this event, how many levels were in the PBL on average and at the minimum?

Response: Thanks for pointing out this, according to the comment, we revised it to "levels existed <u>within the PBL</u> at any time". The eta values for the first 10 levels are 0.996, 0.988, 0.978, 0.966, 0.956, 0.946, 0.933, 0.923, 0.912, and 0.901, for the event in our study, about 6 levels on average were in the PBL during the first 4 days with low PBL height, for the last day with high PBL height, there are more levels on average.

Line 112: Why was such a long spin-up time selected? It is important to note that the model dataset was generated from a single initialization with a 7-day forecast horizon and that forecast skill is expected to degrade with lead time.

Response: Thank you for pointing out this, in our study, WRF is driven by ERA5 reanalysis data, not the forecast products, and we update SST data during the simulation period, thus the simulations can be considered as dynamical downscaling simulations driven by "perfect boundary conditions", and the model skill will not degrade obviously along with the simulation time. The study of Kleczek et al. (2014) indicated that longer spin-up time decreased the modelled wind speed bias, thus in our study, we selected a spin-up time of 40 hours, and the results are satisfactory.

References:

Kleczek, M.A., Steeneveld, GJ. & Holtslag, A.A.M. Evaluation of the Weather Research and Forecasting Mesoscale Model for GABLS3: Impact of Boundary-Layer Schemes, Boundary Conditions and Spin-Up. Boundary-Layer Meteorol 152, 213–243 (2014). https://doi.org/10.1007/s10546-014-9925-3

Line 112: What are the default parameterizations? Please name and reference them.

Response: Thank you for pointing out this, the parameterization and the corresponding references were listed in Table 2 in line 639, which is cited in line 127: "Firstly, the default physical parameterization schemes (Table 2) are applied in the single set of WRF simulations"

Line 113/114: Please clarify: Were the simulations first run for D01 and D02, then D03 was initialized with the

output from D02, or was each of the 640 simulations run with one-way-nested feedback across all three domains at each time step?

Response: Thanks for the comment, the simulations are first run for D01 and D02, then D03 is forced with the output of D02 using ndown. it was revised in lines 127-129: "Firstly, the default physical parameterization schemes (Table 2) are applied in the single set of WRF simulations for the outer two domains (D01 and D02), and then the output of D02 is used to drive inner domain simulations with different combinations of PBL, MP, and SW-LW schemes (see section 2.3)".

Line 117/118: "The lateral boundary *conditions* and sea surface temperature were updated …" also, which dataset did the SST come from?

Response: Thank you for pointing out this, it was revised to "The lateral boundary conditions and sea surface temperature are updated every three hours using the ERA5 reanalysis data" in lines 132-133.

Line 118: wind was *calculated,* or output retrieved from the model?

Response: Thank you for pointing out this, wind was from the model, it was revised to "the frequency of wind retrieved from WRF output was hourly, which matches the frequency of observations in the study area" in line 133.

Line 121 & 124: Add comma "e.g., …" & "i.e.," please check throughout the manuscript

Response: Thank you for pointing out this, the errors were corrected, and we also checked the whole manuscript.

Line 126: The gray zone is first mentioned at Line 77 and should be defined there. After definition, the quotation marks for this term can be removed.

Response: Thank you for pointing out this, in the revision, we introduced "gray zone" in lines 84-86: "which belongs to the PBL "gray zone" resolution that is too fine to utilize mesoscale turbulence parameterizations and too coarse for a large-eddy-simulation (LES) scheme to resolve turbulent eddies (Shin and Hong, 2015; Honnert et al., 2016)"

Then we mentioned it in lines 140-141 "As the horizontal grid spacing of 0.5 km is within the PBL gray zone resolution, both PBL and LES assumptions are imperfect".

Line 134: Both "atmospheric boundary layer" or "planetary boundary layer" are fine, but please be consistent throughout the manuscript.

Response: Thank you for pointing out this, we have checked the whole paper and revised them accordingly.

Lines 139-145:This paragraph should be revised for a more insightful summary.

The concept of single- vs double-moment and hydrometeor classes should be briefly explained.

Explain why some MP schemes are "suitable for high-resolution simulations".

The Goddard has a reference from 1989 and is described as one for the "new schemes".

NSSL1 has no reference in Table 2.

Response: Thank you for pointing out this, we have revised them in lines 151-161: "Sixteen MP schemes are applied in this study (Table 1), Lin, WSM3, WSM5, ETA, WSM6, Goddard, SBU, and NSSL1 schemes are the single-moment bulk microphysical scheme, which predicts only the mixing ratios of hydrometeors (i.e., cloud ice, snow, graupel, rain, and cloud water) by assuming particle size distributions. The other eight schemes (Thompson, MYDM7, Morrison, CMA, WDM6, NSSL2, ThompsonAA and P3) use a double-moment approach, predicting not only mixing ratios of hydrometeors but also number concentrations. Among them, two types of hydrometeors are included in WSM3 (cloud water and rain), three types of hydrometeors are included in ETA (cloud water, rain, and snow) and P3 (cloud water, rain, and ice), four types of hydrometeors are included in WSM5 and SBU (cloud water, rain, ice, and snow), five types of hydrometeors are included in Lin, WSM6, Goddard, Thompson, Morrison, CAM, WDM6 and ThompsonAA (cloud water, rain, ice, snow, and graupel), six types of hydrometeors are included in MYDM7, NSSL1, and NSSL2 (cloud water, rain, ice, snow, graupel, and hail). According to the user's guide of ARW (Skamarock et al., 2008), WSM6, Thompson, Morrison, WDM6, NSSL1, and NSSL2 are suitable for high-resolution simulations.", and the reference for NSSL1 was added in Table 1.

Line 146: Table 3 is not referenced anywhere in the manuscript. Tables 1, 2, and 3 could be combined.

Response: Thanks for the comments, we have combined Tables 1-3 to Table 1 in the revision.

Line 155: How did you define / identify "spurious jumps"?105 stations remained – out of how many?

Response: Thanks for the comments, during the stable event, some sensors were frozen and the data are all zeroes, so we skipped the stations. We revised them in 169-171: "All data are screened before analysis in order to remove stations with data showing spurious jumps (e.g., wind speed jumps to 0 m/s due to frozen sensor). After this filtering, 105 out of 132 weather stations (Figure 1a) remained, including 89 inland stations and 16 coastal stations."

Line 167: Also define "i"

Response: Thank you for pointing out this, the errors were corrected.

Line 170/171: Where in the paper is this metric considered?

Response: Thanks for pointing out this, the BIASs and RMSEs of wind direction are calculated based on this metric.

Line 174: This equation misses sums.

Response: Thanks for the comment, the errors were corrected.

Lines 177-180: Are there 105 or 106 stations? This paragraph can be skipped.

Response: Thank you for pointing out this, we have removed the paragraph.

Line 183: Wind speed data is not directly *produced by* PBL schemes. Wind speed is a dynamic variable that is adjusted by the PBL scheme. So perhaps the data "is produced *using* PBL schemes".

Response: Thanks for pointing out this, it was revised to "Figure 3 shows the time series of observed wind speeds and the corresponding simulations using different PBL schemes".

Figure 3 and all other figures with time series: Clarify in your manuscript whether these are UTC or local times. If times are in UTC, please mention to which local times these translate. Please also include tick marks for each date and possibly a vertical line separating each day as a reference for diurnal periods.

Response: Thank you for pointing out this, according to the comment, we change the timeseries plots to include tick marks and vertical lines for each day in the revision (Figure 3, 5, 7,10,11,12,15,16). The time in all the figures is local time, which is clarified in line 194: "Figure 3 shows the time series of observed wind speeds in local time and the corresponding simulations", Line 126: "The WRF model is initialized at 00:00 UTC (08:00 in local time) January 9, 2019"

Line 184: I am seeing that "The WRF model *exaggerates* the temporal variation of observed wind speed in the study area"

Response: Thanks for the comment, it was revised in lines 196-197: "The WRF model generally reproduces the temporal variation of observed wind speed in the study area with exaggeration".

Line 185: I disagree with the statement that "QNSE showed no obvious daily wind speed change during the simulation period" – The wind speed change is considerably larger than with all other schemes.

Response: Thanks for pointing out this, it was revised to "…except for QNSE, with which the wind speed change is considerably larger than with all other schemes during the simulation period".

Lines 186-189: This section needs revision. It is unclear which correlation the authors refer to; it needs to be clarified that the 10m/s bias applies to QNSE only; it should be elaborated what other studies found ("such models" – referring to the QNSE models or the general overprediction by all WRF models?) Note that a more thorough literature review is needed in the introduction which could be referred to here.

Response: Thanks for the comments, it was revised as "Almost all the PBL schemes overestimate wind speed by 1 m/s, however, for the QNSE scheme, the largest overestimation exceeds 10 m/s during the daytime on 11 and 15 January 2019."

In the Introduction, lines 57-62: "A lot of studies indicate an overestimation of wind speed in WRF simulation 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), for example, Gómez-Navarro et al. (2015) investigated the sensitivity of the WRF model to PBL schemes by simulating wind storms over complex terrain at a horizontal grid spacing of 2 km. In that study, the WRF model was configured with the Mellor-Yamada-Janjic (MYJ) scheme and overestimated wind speed by up to 100%,"

Line 194: "*on* January 15th" "partly due to faster observed wind speeds" – right, the bias looks multiplicative,

but probably also due to the general error growth in NWP with lead time.

Response: Thank you for pointing out this, it was revised to "In addition, the spread within the PBL schemes is larger on 15 January 2019, partly due to high wind speed (> 4 m/s) <u>or the general error growth in the model</u>" in lines 201-202.

Line 195/196: The description of what bold and italicized numbers mean belongs in the figure caption.

Response: Thank you for pointing out this, we have changed table4 to figure and the paragraph was removed.

Line 200, 224, 236: The authors often describe the next-best schemes as having "similar statistics".

Please be specific to avoid confusion. For example, "X1 is the best scheme, followed by X2 and X3" or "X1 shows the best verification score. X2 and X3 are slightly worse according to this verifications score."

Response: Thank you for pointing out this, they were revised as follows:

"<u>MYJ shows the best CORR score of 0.96, MYNN, ACM2 and UW are slightly worse according to this verification score. YSU is the best scheme in term of BIAS and RMSE with the lowest scores of 0.45 m/s and 0.61 m/s, flowed by MYNN (0.55 m/s and 0.70 m/s).</u>"

"while MYDM7 is the best scheme <u>according to BIAS and RMSE scores, followed by</u> P3 and ETA."

"Dudhia-RRTM is the best scheme according to BIAS and RMSE scores, followed by Goddard"

Line 200/201: A correlation coefficient of 0.643 would usually not be considered "extremely low". Please revise your wording.

Response: Thanks for the comment, it was revised to "<u>For the QNSE scheme, the maximum BIAS and RMSE scores for individual stations exceed 10 m/s and 16 m/s, indicating that it has problems in reproducing wind speed under stable conditions over the study area</u>" in lines 208-210.

Line 202/203: Please justify this assumption. What does other literature suggest?

Response: Thank you for pointing out this, it is only a speculation, and may only applied to the specific area in our study, to avoid misleading, the sentence was removed.

Lines 204/205: Note that QNSE is still included in figure 4 and table 5. Either exclude QNSE from there, or move the statement that QNSE will be omitted after referencing figure 4 and table 5.

Response: Thank you for pointing out this, the statements were moved to lines 218-219: "<u>Considering the large model bias in wind speed, all simulations using the QNSE scheme (64 simulations in total) are omitted from further investigation in order that these anomalous data do not affect our overall analysis".</u>

Line 206: Plural "wind roses" Wind roses in figures 2 and 4 need a legend description.

Response: Thanks for pointing out this, the errors were corrected, the legend description were added in line 659: "for each wind rose chart, the circles represent the relative frequency (%), and the colors represent wind speed (m/s)".

Line 210: "As the simulated wind direction was calculated using the wind speed *components*, the bias in modeled wind direction can be attributed to bias in the wind-speed *component* simulation." – this statement is somewhat redundant.

Response: Thank you for pointing out this, we have removed the paragraph.

Line 224: "P*3*" (not "P2")

Response: Thanks for pointing out this, the errors were corrected.

Line 230: "*ensemble* spread"

Response: Thanks for the comment, it was revised to "The SW-LW schemes have a larger model ensemble spread".

Line 233: "~~substantially~~ reduced this overestimation, and thus produced values that were ~~much~~ closer to weather station observations." – these differences are relatively small but consistent

Response: Thanks for pointing out this, the errors were corrected.

Line 238: "while simulations that employed the *Goddard* [not RRTMG] scheme showed the lowest BIAS values."

Response: Thank you for pointing out this, the errors were corrected.

Line 240: The authors use the term "physical components" interchangeably with "physics parameterizations". "Physical components" can be ambiguous, I recommend consistently referring to "parameterizations" throughout the manuscript.

Response: Thanks for pointing out this, we revised them according to this comment.

Line 243: "A total of 40 simulations" – it is actually only 36 because QNSE is missing.

Response: Thank you for the comment, it was revised to "thus for each MP scheme, a total of 36 simulations (excluding QNSE) are evaluated in this way"

Line 244: "which produced outcomes that were ~~fully~~ consistent with other results" –which results? Other MP groups; previously shown results in this manuscript; or different studies? Please be specific!

Response: Thank you for pointing out this, it was revised to "and the results are expected to be consistent with evaluations using other MP schemes".

Line 247: "further investigation"

Response: Thanks for the comment, the errors were corrected.

Lines 247/248: "Dudhia and RRTM schemes ~~together~~, or the Goddard *schemes* ~~by itself~~" – Goddard also has two schemes, LW and SW

Response: Thank you for pointing out this, the errors were corrected.

Lines 256/257 and Table 8: "The PBL and LW/SW radiation schemes used in the 10 best schemes were YSU and Dudhia and RRTM" – This wording sounds like the authors decided upon using these PBL and radiation schemes, rather than this being a results of their analysis. If I understand correctly, I suggest clarifying "The best 10 WRF model configurations have in common that they use the same PBL and radiation schemes, namely YSU and Dudhia-RRTM. Due to the slight differences between models using different MP schemes, the 10 best performing WRF configuration only vary in MP option." or similar

Response: Thanks for the comment, we have revised the paragraph accordingly in lines 265-267: "The timeseries and statistics are illustrated in Figure 10. The best 10 configurations have in common that they use the same PBL and SW-LW schemes, namely YSU and Dudhia-RRTM. Due to the slight differences between models using different MP schemes, the 10 best performing WRF configuration only vary in MP schemes."

Line 263: "plays an important role in ~~the~~ determining its performance"

Response: Thank you for pointing out this, the errors were corrected.

Table 8: Add a separating horizontal line between individual model configurations and ensembles.

Response: Thanks for the comment, Table 8 was revised to a subplot in Figure 10.

Line 254: "Evaluation of model configurations with the best *individual* performance"

Response: Thank you for pointing out this, the errors were corrected.

Line 276: Please explain how the classification between coastal and inland stations was conducted! Did you use an objective distance to the shoreline?

Response: Thanks for pointing out this, coastal and inland stations are classified by the distance to the shoreline, and it was revised in lines 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 5km from the shoreline, 16 stations in total), the locations of these stations are shown in Figure 1a."

Line 277: add that this is based on the ensemble spread

Response: Thanks for the comment, it was revised accordingly in line 286: "For both coastal and inland stations, the ensemble spread is largest among the PBL schemes, followed by SW-LW and MP schemes".

Line 278: "consistent with the results of previous analyses *in this study*"

Response: Thank you for pointing out this, it was revised to "which is consistent with the results of previous analyses in this study" in line 287.

Line 279: "generally by the same magnitude" – no, figure 9 and table 9 both show that the bias is larger for inland stations.

Response: Thanks for pointing out this, it was revised to "WRF reproduce the timeseries of wind speed reasonably, with larger overestimation for inland stations" in line 288.

Line 280/281: "the *ensemble* spread was relatively larger for coastal stations, especially *among MP schemes and* for the first three days of the simulation period that exhibited low wind speeds"

Response: Thanks for the comment, the errors were corrected.

Line 282: "source of model uncertainty" – although the authors observe different sensitivities for coastal vs inland stations between parameterizations, is that a source of model uncertainty? "generates greater model differences" – compared to what?

Response: Thank you for pointing out this, the sentence is misleading, we revised it to "As such, in addition to physical parameterizations, model performance is also influenced by ocean proximity, and WRF simulates wind speed less accurately for coastal stations compared to inland stations" in lines 290-291.

Line 285: Is this the temporal CORR averaged over stations? Otherwise the sample-size difference between the two groups (16 coastal stations vs 89 inland stations) needs to be considered.

Response: Thanks for the comment, the temporal CORR is averaged over the stations, so we revised in line 297: "consistent with those of previous investigations in this study, <u>considering most of the stations are inland stations (89 out of 105 stations),</u> this result…"

Line 285/286: What is meant by "as there are no clear differences in wind speed values and variations"?

Response: Thank you for pointing out this, the sentence is misleading, we revised in lines 292-293: "the CORR scores are consistently lower and the BIAS and RMSE scores are generally worse for coastal stations compared to inland stations, which indicates degraded model performance for coastal stations."

Line 289: "previous investigations" – in this study or in different studies?

Response: Thanks for the comment, it was revised to "are generally consistent with those of previous investigations <u>in this study</u>".

Line 294: "Dudhia and RRTM … showed *worst temporal agreement* with observational data *for coastal stations*"

Response: Thank you for pointing out this, the errors were corrected.

Line 298: Which subfigure was this info taken from? Perhaps the authors mean "the peak *observed* wind speed at high-elevation stations (>250 m) was *1.5* m/s slower than that for low-elevation stations (<50 m)." ?

Response: Thanks for pointing out this, it was revised in lines 304-305: "the peak <u>observed</u> wind speed of high-elevation stations (>250 m) is <u>1.5 m/s slower</u> than that of low-elevation stations (<50 m)."

Line 299: "the simulated peak values were generally *similar*"

Response: Thank you for pointing out this, it was revised according to the comment.

Lines 302/303: "Interestingly, model performances *of different parameterization types* were generally similar for

stations with different elevations, ...”

Response: Thanks for the comment, it was revised accordingly.

Line 304: “… smallest BIAS and RMSE values *at all elevation categories*” ?

Response: Thank you for pointing out this, it was revised to “MYJ is always best for <u>all elevation categories</u> according to the CORR score, and YSU is always best according to the BIAS and RMSE scores.”

Lines 305/306: It is plausible that physics-configuration performance depends on surface topography in other cases and locations with different topography. It is not appropriate to generalize these findings to overarching WRF performance like this. The authors provide no foundation to suggest that the limited configuration dependency seen in figure 10 is result of the terrain-following coordinates.

Response: Thank you for pointing out this, points were well taken, and we removed the statements.

Line 314: “underlying land type and topography” – better “coast proximity and elevation” – land types include factors that were not investigated (e.g., soil texture, vegetation, roughness, canopy, etc.); topography includes aspects that were not investigated (e.g., slope steepness and directional angles of slopes)

Response: Thanks for pointing out this, we revised it according to the comments.

Lines 315/316: “the WRF model reproduced the temporal variation of wind speed and direction over the study area *well*” – I don’t agree that “to a high degree of accuracy” is an accurate description considering the biases presented.

Response: Thanks for the comment, we revised it to “the WRF model reproduces the temporal variation of wind speed over the study area <u>well</u>”.

Line 322/323: “The combined Dudhia and RRTM *radiation* schemes, and the ~~individual~~ MYDM7 *MP* scheme both show the best *wind-speed* performances…”

Response: Thank you for pointing out this, the errors were corrected.

Line 325: “~~and can either increase or decrease the accuracy of the simulation~~.” (redundant)

Response: Thanks for the comment, the errors were corrected.

Line 327/328: “… substantially reduced overestimation of wind speed *compared to the ensemble of all 640 configurations* ~~and provided the best combined performance of all simulations”~~

Response: Thank you for pointing out this, the errors were corrected.

Line 329: Most (85%) of the 105 stations are inland stations, so it is implied that the overall pattern matches the inland stations most. These conclusions are not “applicable to the inland station” but they were mostly derived from them.

Response: Thanks for the comment, we revised it in lines 296-298: “Our comparison indicates that the

parameterization schemes with the best performance for inland stations are generally consistent with those of previous investigations in this study, considering most of the stations are inland stations (89 out of 105 stations), this result is not surprising".

Lines 332/333: The best-configuration ranking might be similar, but the model results are different.
Response: Thank you for pointing out this, we revised it according to the comment.

Line 335: These are not "all possible combinations that are available". (1) There are other parameterization types that were not investigated (e.g., cumulus convection, land surface, etc.); (2) within the parameterization types that were assessed there are more available (e.g., Kessler and WDM5 for MP, GBM and MRF for PBL, Fu–Liou–Gu and GFDL for radiation)
Response: Thanks for the comment, our statements are not correct and we removed them.

Line 337: "which has rarely been used in previous simulation studies *in China*"
Response: Thank you for pointing out this, the errors were corrected.

Lines 339-343: This paragraph needs to be revised considering the major comments above.
Response: Thank you for the comment, we revised it in lines 369-375: "The ensemble mean of 576 simulations in our study shows the best CORR score (0.94) in wind-speed simulation, however, this best CORR score is also result of single-model simulation with Goddard MP scheme. At the same time, the best wind-speed 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. Thus, model ensemble does not always provide the best performance, and 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"

Referee 2:

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.

This paper examines wind forecasts during a relatively long period of stable conditions when a haze event affected China. Surface meteorological observations are used to evaluate the WRF model's ability to predict the evolution of winds during the event. The authors conduct a number of WRF simulations (640 total), altering the PBL, radiation, and microphysics schemes to determine the sensitivity of wind speed and direction forecasts to choice of model physics. Pearson's correlation coefficient, bias, RMSE, and Taylor skill score are utilized to perform the model evaluation. Overall, the study shows the largest spread in wind speed within the PBL schemes tested, followed by radiation, and then microphysics schemes. Delineation between coastal/inland stations as well as stations at different elevation are examined to understand any model biases specific to land type and characteristics. An important finding is that WRF predicts wind speed less accurately for coastal stations compared to inland stations, and error metrics tended to degrade with increasing elevation.

Overall, this study has interesting components and would be a nice contribution to the literature especially due to the very large ensemble that was run. However, there are several aspects that should be addressed before the paper is suitable for publication. My main concerns are related to the authors' model setup, lack of some background information about the case, and insufficient physical explanations for some of their results. My general and specific comments are listed below.

Response: Thank you for taking time out of your busy schedule to review this paper, I really appreciate all your comments and suggestions. Please find my responses in below and my revisions/corrections in the re-submitted files. Thanks again.

<u>Major/general comments:</u>

1. WRF model configurations: I wonder why the authors chose to run all of the physics parameterizations as default except for YSU, which was run using a topographic correction for surface winds and the top-down mixing option. Were the impacts of these YSU options tested? I believe that YSU is not run this way by default, so it would be good to know the impact, especially since your results show that YSU is one of the best performers. For instance, MYNN has a number of namelist tuning options, so why not modify these? Also, what is the motivation for running with the top-down mixing option in YSU if this is a statically stable case? Are stratus/fog conditions expected in some of the coastal regions? Please explain.

Response: Thank you for pointing out this, according to this comment, we added experiments considering the effects of YSU options, and the results are presented in section 4.3, lines 341-349:

"4.3 Impact of options in the YSU scheme

The impact of different options in YSU on wind-speed simulation is illustrated in Figure 15, the simulation with the best Taylor skill score in previous investigation is selected and referred to as YSU, three extra simulations with top-down mixing option turning off (No_mix), topographic correction option turning off (No_topo), and both options turning off (No_topo_mix) are conducted for comparison. The simulated wind speed increases when we turn off the individual or both options, which enlarges the overestimation of wind

speed under stable conditions in our study (Figure 15a). Further investigation indicates that turning off the two options in YSU mainly degrades model performance with worse evaluation metrics, for example, the BIAS score increases from 0.36 m/s to 0.67 m/s in No_topo, to 0.43 m/s in No_mix, and to 0.69 m/s in No_topo_mix, the RMSE scores show similar degradation to BIAS by turning off the options in YSU.".

In the past few years, over North China, haze events are always accompanied by fog over the coastal areas, thus the top-down mixing option was turned on in the simulation.

As a separate but related issue, the authors use different surface layer schemes between the PBL schemes. This means that it is impossible to attribute all of the differences in results specifically to the PBL scheme. There is no discussion about this at all in the paper, although it is definitely important considering the station observations likely fall within the first model grid cell.

Furthermore, why not use the revised MM5 scheme for all of the PBL schemes? I believe that it is compatible with all of them (I may be wrong here). Regardless, it would be good to run an additional simulation to determine the impact of the surface layer scheme (which I suspect is more important than the microphysics scheme under stable conditions).

Response: Thanks for the comment, we totally agree with the reviewer that it is important to determine the impact of surface layer scheme, however, in the WRF model, the surface layer scheme is somehow tied to the PBL scheme, it is impossible to conduct systematic experiments using different surface layer schemes with the same PBL schemes, for example, YSU cannot run with surface layer scheme other than revised MM5 scheme, MYJ PBL scheme can only be used with MYJ surface layer scheme. In this study we designed the PBL experiments including both surface layer schemes and PBL schemes, thus the results of PBL schemes are actually the results of the combination of PBL and surface layer schemes.

For the revised MM5 scheme, it is compatible with a lot of PBL schemes, but not all, for example, it is not compatible with the TEMP PBL schemes.

2. Case study: The authors select a 4-day study period when stability conditions were stable to evaluate the WRF model; however, there is only a few sentence discussion about the case in Section 2.1. Although the authors do conduct many simulations, this is still a case study, and unfortunately, the authors do not present any large-scale meteorological information. It would be good to know the synoptic pattern and what type of evolution occurred; I imagine there is a pattern change over the course of the event since the regional wind speeds went from ~1 m/s to ~5 m/s according to Fig. 2. Moreover, the authors consider the impact of microphysical schemes in WRF even though this is a stable case. Are the authors anticipating cloud effects? Despite the relatively small impact of the microphysics options (e.g., Fig. 5), there are noticeable differences on 2019-01-12. I think the authors should include some metric of observed clouds (e.g., satellite images) since clearly the model is producing clouds.

Response: Thank you for pointing out this, we removed figure 2 and added the distribution of cloud, geopotential height and winds as a new figure. The descriptions are in Lines 107-115: "Figure 2 depicts the distribution of geopotential height, winds, and cloud fraction during the study period, the geopotential height and winds data are from the ERA5 dataset (Hersbach et al., 2020), and the cloud fractions are from satellite observations of CLARA (CM SAF cLoud, Albedo and surface Radiation) product family (Karlsson et al., 2021). Figure 2 indicates that a weak high-pressure system persisted from 11 to 13 January, along with weak southwest wind in the study area, which would transport warm and wet air to the study area (Gao et al.,

2016a), creating a favorable moisture condition for stable conditions and inhibiting pollutants dispersal (Zhang et al., 2014; Hua and Wu, 2022). Then the high-pressure system was replaced by strong northwest wind from 14 to 15 January 2019. The CLARA observations indicate cloud fraction exceeding 60% on 12 January at the study area, while for the rest of the time, cloud fraction is low. This stable event is used to investigate the impact of physical parameterizations of the WRF model."

3. Physical explanations: By and large, this paper reports on the model performance with respect to near-surface wind speed and direction. However, I think the authors do not provide any physical explanations for any of their results. Ultimately, this ends up limiting the applicability of the study to other stable events in different seasonal periods and geographical locations. It would be good to address questions such as: why is QNSE so different from the other PBL schemes? How does the YSU topographic correction affect the forecast? Additionally, linking the low-level wind results to the PBL vertical structure (e.g., wind, temperature/stability, and moisture), as well as x-y spatial variability in the wind field, would be useful. For the planview, the authors could zoom in on a region where they see the largest differences in PBL schemes and overlay their observations. I think these comparisons are especially important for the different PBL schemes because the authors highlight the gray zone, which is where the choice of turbulentmixing tends to dominate the solution.

Response: Thanks for the comments, according to them, we added the comparisons in sections 4.1 (lines 318-330) and 4.3 (lines 341-349):

"4.1 Spatial distribution of wind field

Figure 13 shows the spatial distribution of observed and simulated wind fields during the study period, we choose 14:00 in local time as an example. The simulation using YSU, Dudhia-RRTM and WDM6 schemes is referred to as YSU, and the simulation using QNSE, Dudhia-RRTM and WDM6 is referred to as QNSE. YSU generally reproduces the wind field in the study area, especially in terms of wind speed. For example, the observed wind speed is lower on 13 January 2019, with values lower than 2 m/s in many stations, while on 15 January 2019, the observed wind speed is higher than 4 m/s in most of the stations. In the simulation with YSU, wind speed is about 2 m/s on 13 January 2019 and higher than 4 m/s on 15 January 2019 over the study area, which is close to the observation. On the contrary, simulation with QNSE fails to reproduce the distribution of wind speed, and shows strong overestimation, especially over the mountain areas of the study area (Figure 1a), for example, the peak wind speed in simulation with QNSE exceeds 20 m/s on 15 January 2019, which is more than five times greater than the observation, this overestimation is consistent with the large positive bias in previous investigation of Figure 3. For the wind-direction simulation, YSU shows degraded performance compared to wind speed, and generally fails to reproduce the wind-direction distribution for most of the stations, QNSE also fails to do so."

"4.3 Impact of options in the YSU scheme

The impact of different options in YSU on wind-speed simulation is illustrated in Figure 15, the simulation with the best Taylor skill score in previous investigation is selected and referred to as YSU, three extra simulations with top-down mixing option turning off (No_mix), topographic correction option turning off (No_topo), and both options turning off (No_topo_mix) are conducted for comparison. The simulated wind speed increases when we turn off the individual or both options, which enlarges the overestimation of wind speed under stable conditions in our study (Figure 15a). Further investigation indicates that turning off the

two options in YSU mainly degrades model performance with worse evaluation metrics, for example, the BIAS score increases from 0.36 m/s to 0.67 m/s in No_topo, to 0.43 m/s in No_mix, and to 0.69 m/s in No_topo_mix, the RMSE scores show similar degradation to BIAS by turning off the options in YSU."

Minor/specific comments:

1. Abstract, L14: Should be, "near-surface" rather than "surface".

Response: Thank you for pointing out this, the errors were corrected in lines 14-15:"how different physical parameterizations impact simulated near-surface wind at 10-meter height over the coastal regions of North China"

2. Abstract, L24: Please be more specific by what you mean by "land type".

Response: Thanks for pointing out this, it was revised to "model sensitivity is also impacted by ocean proximity and elevation"

3. Abstract, L25: "For example, for the weather stations located in coastal regions." This is an awkward sentence; it would be good to combine with another sentence or rewrite.

Response: Thanks for the comment, it was revised in lines 23-25: "for coastal stations, MYNN shows the best temporal correlation with observations among all PBL schemes, while Goddard shows the smallest bias out of SW-LW schemes"

4. L47: Should add, "that occur at the sub-grid scale" after you say "key physical processes". Also, some of the processes listed (i.e., planetary boundary layers and cloud microphysics) aren't really processes. Please be more specific or re-phrase the beginning of the sentence.

Response: Thank you for pointing out this, it was revised in line 50: "instead, parameterizations are needed to represent the effect of key physical processes, such as radiative transfer, turbulent mixing, and moist convection that occur at the sub-grid scale."

5. L51: The impact of PBL schemes on wind simulations has been studied for many years (probably >60 years at this point).

Response: Thanks for the comment, it was revised to "The impact of planetary boundary layer (PBL) schemes on wind simulations has been studied for many years".

6. L55: Should be, "horizontal grid spacing" rather than "horizontal resolution".

Response: Thank you for pointing out this, the errors were corrected in line 61, we also revised the Abstract accordingly.

7. L77: I would say, "high resolution mesoscale simulation" rather than "very high resolution" considering dx=500 m is not high resolution with respect to typical LES scales (10s of meters).

Response: Thanks for the comment, it was revised to "The investigation is conducted using the WRF model at a grid spacing of 0.5 km" in lines 83-84.

8. L77: Please explain briefly what is mean by "gray zone" and provide references.

Response: Thank you for pointing out this, it was revised in lines 84-86: "which belongs to the PBL "gray zone" resolution that is too fine to utilize mesoscale turbulence parameterizations and too coarse for a large-eddy-simulation (LES) scheme to resolve turbulent eddies (Shin and Hong, 2015; Honnert et al., 2016)"

9. L82: Should be, "findings" rather than "foundings".

Response: Thank you for pointing out this, it was corrected in line 91.

10. L88-92: Some references would be nice here, especially for an audience who is not familiarwith this region.

Response: Thanks for the comment, it was revised in lines 96-102: "The study area is located in the central section of the "Bohai Economic Rim", which is bordered to the southeast by the Bohai Sea and to the northwest by the Yan Mountains (Figure 1a). This region traditionally hosts heavy industry and manufacturing businesses and is a significant region of economic growth and development in North China (Song et al., 2020; Zhao et al., 2020). Air quality in this area has declined over the past decades, and the frequency of winter haze events has increased due to increased pollutant emissions and favorable stable weather conditions with lower wind speed (Gao et al., 2016a; Cai et al., 2017)."

11. L99-100: Please add references for why this wind direction indicated a weakened East Asiawinter monsoon.

Response: Thank you for pointing out this, the figure was removed and we added the synoptic description of this event in lines 107-115: "Figure 2 depicts the distribution of geopotential height, winds, and cloud fraction during the study period, the geopotential height and winds data are from the ERA5 dataset (Hersbach et al., 2020), and the cloud fractions are from satellite observations of CLARA (CM SAF cLoud, Albedo and surface Radiation) product family (Karlsson et al., 2021). Figure 2 indicates that a weak high-pressure system persisted from 11 to 13 January, along with weak southwest wind in the study area, which would transport warm and wet air to the study area (Gao et al., 2016a), creating a favorable moisture condition for stable conditions and inhibiting pollutants dispersal (Zhang et al., 2014; Hua and Wu, 2022). Then the high-pressure system was replaced by strong northwest wind from 14 to 15 January 2019. The CLARA observations indicate cloud fraction exceeding 60% on 12 January at the study area, while for the rest of the time, cloud fraction is low. This stable event is used to investigate the impact of physical parameterizations of the WRF model."

12. L103: Why not use a newer version of WRF? Version 3.9.1.1 was released years ago (August2017, I believe).

Response: Thank you for the comments, WRF V3.9.1.1 was released on August 28, 2017, and we used this version for a long time in different areas across China, the overall performance is satisfactory to us, thus we continue to use it in this study to support comparative studies with previous simulations. we assume that the WRF version does not affect our results.

13. L108-109: Were eta levels specified? It would be important to report the vertical grid spacingnear the surface.

Response: Thanks for pointing out this, it was revised in lines 123-124: "and the eta values of the first 10 levels are 0.996, 0.988, 0.978, 0.966, 0.956, 0.946, 0.933, 0.923, 0.912, and 0.901."

14. L126: Discussing the "gray zone" here is fine, but please reference this section in the introduction so that the reader is not confused by what you mean by the terminology.

Response: Thank you for the comments, we added the description of "gray zone' in the section 1 lines 84-86: "the PBL "gray zone" resolution that is too fine to utilize mesoscale turbulence parameterizations and too coarse for a large-eddy-simulation (LES) scheme to resolve turbulent eddies (Shin and Hong, 2015; Honnert et al., 2016)"

We also revised the sentences here as "The horizontal grid spacing of 0.5 km is within the PBL gray zone resolution, both PBL and LES assumptions are imperfect".

15. L126-130: Please provide references for the "gray zone" discussion as well as the Deardorff SGS TKE scheme.

Response: Thanks for the comments, the references for the "gray zone" are added in lines 85-86: "the PBL "gray zone" resolution that is too fine to utilize mesoscale turbulence parameterizations and too coarse for a large-eddy-simulation (LES) scheme to resolve turbulent eddies (Shin and Hong, 2015; Honnert et al., 2016)", and the reference for the Deardorff SGS TKE scheme is added in line 142: "the 1.5-order turbulence kinetic energy closure model is used to parameterize motion at the sub-grid scale (Deardorff, 1985)".

16. L139: I disagree that the "Lin scheme is a sophisticated scheme", considering it is single moment and there are several double (or higher order) moment schemes.

Response: Thanks for pointing out this, we removed it and rewrite the whole part in lines 151-160: "Sixteen MP schemes were applied in this study (Table 1), Lin, WSM3, WSM5, ETA, WSM6, Goddard, SBU and NSSL1 schemes are the single-moment bulk microphysical scheme, which predicts only the mixing ratios of hydrometeors (i.e., cloud ice, snow, graupel, rain, and cloud water) by assuming particle size distributions. The other eight schemes (Thompson, MYDM7, Morrison, CMA, WDM6, NSSL2, ThompsonAA and P3) use a double-moment approach, predicting not only mixing ratios of hydrometeors but also number concentrations. Among them, two types of hydrometeors are included in WSM3 (cloud water and rain), three types of hydrometeors are included in ETA (cloud water, rain and snow) and P3 (cloud water, rain and ice), four types of hydrometeors are included in WSM5 and SBU (cloud water, rain, ice and snow), five types of hydrometeors are included in Lin, WSM6, Goddard, Thompson, Morrison, CAM, WDM6 and ThompsonAA (cloud water, rain, ice, snow and graupel), six types of hydrometeors are included in MYDM7, NSSL1 and NSSL2 (cloud water, rain, ice, snow, graupel and hail)."

17. L142: These microphysics schemes are not new, as they are over 10 years old now.

Response: Thank you for the comments, we removed this and rewrite the whole part in lines 151-160.

18. L144: Please define "ThompsonAA".

Response: Thank you for the comments, ThompsonAA is short for "Aerosol-Aware Thompson", and we explained it in Table 1.

19. L154-155: How exactly are the data screened? Please explain.

Response: Thanks for the comments, it was explained in lines 169-171: "All data are screened before analysis in order to remove stations with data showing spurious jumps (e.g., wind speed jumps to 0 m/s due to frozen sensor). After this filtering, 105 out of 132 weather stations (Figure 1a) remained, including 89 inland stations and 16 coastal stations."

20. L187: Only in the extreme case of QNSE did the absolute difference exceed 10 m/s; however, as written currently it sounds like many simulations have large differences.

Response: Thank you for the comments, it was revised in lines 196-199: "reproduced by all schemes except for QNSE, with which the wind speed change is considerably larger than with all other schemes during the simulation period. Almost all the PBL schemes overestimate wind speed by 1 m/s, however, for the QNSE scheme, the largest overestimation exceeds 10 m/s during the daytime on 11 and 15 January 2019."

21. L202-204: Do you have proof for the claim about QNSE? Are there other studies that support this claim? I wonder if MYNN used EDMF or just ED in these simulations. I know that in newer versions of WRF, EDMF is default for EDMF.

Response: Thank you for pointing out this, the claim about QNSE is only a speculation, to avoid misleading, the sentence was revised in lines 208-210 as "For the QNSE scheme, the maximum BIAS and RMSE scores for individual stations exceed 10 m/s and 16 m/s, indicating that it has problems in reproducing wind speed under stable conditions over the study area.", For MYNN, in version 3.9.1.1, EDMF is turned off by default.

22. L209-210: I don't understand this sentence. Wind direction depends upon the u and v- components, so one can have the correct wind speed but not the correct wind direction becausethe components are incorrect.

Response: Thank you for pointing out this, it should be "the simulated wind direction was calculated using the u and v- components, the bias in modeled wind direction can be attributed to bias in the u and v- component simulations", this is redundant, so we deleted it.

23. Fig. 5: There are repeat colors in this figures, please fix this.

Response: Thanks for pointing out this, the errors are corrected in figures 5, 11 and 12.

24. L252-253: It would be good here to briefly confirm/reiterate the best combination schemesfor wind direction.

Response: Thank you for pointing out this, we added the description of wind direction in lines 252-253: "For the wind direction simulation, LES combined with Dudhia-RRTM shows the best CORR score, while TEMF is the best scheme according to BIAS and RMSE scores".

25. L265-266: The difference in statistical measures between the ensemble and WDM6 is very small, suggesting the ensemble is not improving things too much.

Response: Thanks for pointing out this, we added discussion of this in line 373: "Thus, model ensemble does not always provide the best performance".

26. L266-267: "reduces model bias by approximately half", compared to what?

Response: Thank you for pointing out this, it was revised to "According to the statistic scores, ENS4 reduces model bias by approximately half compared to ENSall" in lines 276-277.

27. L271-272: Please provide references.

Response: Thanks for the comment, it was revised in lines 279-280: "which impacts local low-level circulation patterns and the wind distribution (Yu et al., 2013; Barlage et al., 2016)."

28. L287: This is a broad statement. What exactly do you mean by it with respect to future model developments? Perhaps some discussion should be added to Section 4.

Response: Thank you for pointing out this, we made it clear in lines 294-295: "This degradation may be caused by the uncertainties from the prescribed SST in our simulation, which may require a better description of atmosphere-ocean coupling process in future model development".

29. L289: "previous investigations", please add references.

Response: Thanks for the comment, previous investigation refers to the investigation in this study, so we revised it to "…for inland stations are generally consistent with those of previous investigations in this study," in lines 296-297 .

30. L305: "surface topography does not induce new uncertainties into the simulation", this is simply not true. What about errors associated with horizontal diffusion and topographic data sets? These actually worse as terrain becomes steeper/more complex.

Response: Thanks for pointing out this, the statement is misleading, so we removed it.

31. L321-322: In general, the QNSE finding is surprising, given that it was originally developed for stable conditions. Do your results agree with other studies that have used QNSE under stable conditions? Discussion here is needed.

Response: Thank you for the comments, we conducted further investigation of QNSE scheme in sections 4.1 and 4.2 (318-340):

"4.1 Spatial distribution of wind field

Figure 13 shows the spatial distribution of observed and simulated wind fields during the study period, we choose 14:00 in local time as an example. The simulation using YSU, Dudhia-RRTM and WDM6 schemes is referred to as YSU, and the simulation using QNSE, Dudhia-RRTM and WDM6 is referred to as QNSE. YSU generally reproduces the wind field in the study area, especially in terms of wind speed. For example, the observed wind speed is lower on 13 January 2019, with values lower than 2 m/s in many stations, while on 15 January 2019, the observed wind speed is higher than 4 m/s in most of the stations. In the simulation with YSU, wind speed is about 2 m/s on 13 January 2019 and higher than 4 m/s on 15 January 2019 over the study area, which is close to the observation. On the contrary, simulation with QNSE fails to reproduce the distribution of wind speed, and shows strong overestimation, especially over the mountain areas of the study area (Figure 1a), for example, the peak wind speed in simulation with QNSE exceeds 20 m/s on 15 January 2019, which is more than five times greater than the observation, this overestimation is consistent with the large positive bias in previous investigation of Figure 3. For the wind-direction simulation, YSU shows

degraded performance compared to wind speed, and generally fails to reproduce the wind-direction distribution for most of the stations, QNSE also fails to do so.

4.2 Vertical profile of wind speed

Figure 14 shows the observed and simulated vertical profile of wind speed at 08:00 and 20:00 during the study period, the location of the sounding station is illustrated in Figure 1. YSU reproduces the vertical structure of wind speed reasonably, with slightly larger model bias above the height of 15 km. Within the low levels below 2.5 km, simulated wind speed from the YSU scheme is close to the observation, with the bias lower than 2.5 m/s in most cases. Meanwhile, QNSE shows worse performance in reproducing the vertical structure of wind speed, with significant larger model bias compared to YSU. for example, QNSE overestimates the low-level (< 2.5 km) wind speed by about 10 m/s at 20:00 on 11 January 2019, and overestimate wind speed by 20 m/s at 20:00 on 11 January 2019. It is interesting to note that the simulation with QNSE is pretty similar to that with YSU at 08:00 during the study period, indicating that large difference between YSU and QNSE only occurs at specific time during the study period, which is also revealed in Figure 3a."

32. L336-337: This broad statement about these "gray zone" resolutions being used "rarely" in previous simulation studies is not true. There are dozens (if not hundreds) of papers looking at gray zone modeling, especially in the last ~5-10 years. Please be more specific with your statement.

Response: Thank you for pointing out this, we revised it to "to the PBL gray zone, which has rarely been used in previous simulation studies <u>in China</u>" in line 406.

33. L342: To which PBL scheme parameters are you referring? It would be good to give someexamples.

Response: Thank you for pointing out this, we revised it in lines 375-377: "further tuning of the parameters within the PBL schemes, <u>such as turbulent kinetic energy (TKE) dissipation rate, TKE diffusion factor, and turbulent length-scale coefficients</u> is needed."

34. Figs. 9 and 10: Please note in the caption that the y-axis range is not the same (alternatively, make them the same).

Response: Thanks for the comment, we make the range of y-axis identical in the revision.