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.

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.

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.

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.

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.

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).
   - 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.

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.)

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.
   - 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).
   - 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.
   - 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.

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.

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.

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.

Technical Corrections

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

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 …"

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.

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

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

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

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

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.

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

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

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.

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.

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

Lines 82: "founding"?

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

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

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

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

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

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)?

Line 104: Plural “simulations”

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

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.

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

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?

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

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

Line 121 & 124: Add comma “e.g., …” & “i.e.,” – please check throughout the 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.

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

Lines 139-145:
This paragraph should be revised for a more insightful summary.
1) The concept of single- vs double-moment and hydrometeor classes should be briefly explained.
2) Explain why some MP schemes are “suitable for high-resolution simulations”.
3) The Goddard has a reference from 1989 and is described as one for the “new schemes”.
4) NSSL1 has no reference in Table 2.

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

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

Line 167: Also define “i”

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

Line 174: This equation misses sums.

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

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".

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.

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

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.

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.

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

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

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."

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

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

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.

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

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.

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

Line 230: "*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

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

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.

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

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!

Line 247: "*further* investigation"

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

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

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

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

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

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

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

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

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

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"

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?

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.

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

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

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

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).” ?

Line 299: “the simulated peak values were generally *similar*”

Lines 302/303: “Interestingly, model performances *of different parameterization types* were generally similar for stations with different elevations, ...”

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

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.

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)

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.

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

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

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”~~

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.

Lines 332/333: The best-configuration ranking might be similar, but the model results are different.

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)

Line 337: “which has rarely been used in previous simulation studies *in China*”

Lines 339-343: This paragraph needs to be revised considering the major comments above.