Authors’ response to reviewer 2: Optimization of weather forecasting for cloud cover over the European domain using the meteorological component of the Ensemble for Stochastic Integration of Atmospheric Simulations version 1.0

Yen-Sen Lu[1,4], Garrett Good[2], and Hendrik Elbern[3]

[1]Institute of Energy and Climate Research – Troposphere (IEK-8), Forschungszentrum Jülich GmbH, 52425 Jülich, Germany
[2]Fraunhofer Institute for Energy Economics and Energy System Technology IEE, Königstor 59, 34119 Kassel, Germany
[3]Rhenish Institute for Environmental Research at the University of Cologne, Cologne, Germany
[4]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany

Correspondence: Yen-Sen Lu (ye.lu@fz-juelich.de)

First of all, we thanks reviewer 2 for giving us a nice and detailed review to improve our manuscript. We read the comments and will reply to the following comments.

1 Reply to the *General Comments*

My main concern is that there is no justification that the 6 cases provide enough information about the variety of cases that these parameterizations experience when in an operational model. Are the case characteristics representative of the variability in weather patterns across the domain? Does this collection of 6 cases contain passing fronts, extreme weather, and calm conditions? Why are there no winter cases included?

The reviewer is right to point out the compromise made in the number of physics simulated versus their representativeness due to computational expense. We discuss this in the revised manuscript and have been able to supplement the results with a long-term (6 month) simulation in now section 4.5 to help test the operational performance of the most promising physics configurations.

We have also tried to add clarity to the framing and objective of the paper regarding the (1) iterative approach of starting with a limited study of a very large number of physics and narrowing this down to fewer physics studied with more detailed simulations, and (2) the objective of this research, which is to optimize WRF specifically for cloud cover and the solar power application, which was the goal of the funding project.

Why only examine cloud cover? Simply using the fraction of a column covered by cloud could obscure important model deficiencies like putting the clouds too high, for example. Surely, the amount of light reaching the surface is different if the cloud cover comes in the form of cirrus instead of boundary layer clouds. The general conclusions of this work could be altered if, for example, column aerosol optical depth were considered instead of cloud cover. Colum AOD is crucial for modeling pollution transport and boundary layer physics packages might play a more significant role (of course scavenging in the microphysics parameterization will also be important).

We hope the motivation is clearer in the new text that we focus on clouds, as the aim of this research is to improve the choice of physics for the operational accuracy of WRF for solar power in Europe, without the complexity of assessing the specific deficencies or types of clouds of the many physics, though these are all good points for future research. The choice of cloud fraction in particular is also determined by the satellite observation data product.

The effect of aerosol on the cloud formation is very important and may change a lot. We'd planned to fully couple ESIAS-met and ESIAS-chem, i.e. WRF3.7.1 and EURAD-IM, to study the feedback between aerosols and cloud, but this task requires proper funding to complete.

The authors do not utilize a satellite simulator package in order to make fair comparisons between models and observations. I am concerned that model is looking "straight down" at each column's respective zenith when computing cloud cover but the SEVERI instrument is observing at a sharp angle (some observations are made above the arctic circle from geostationary orbit!). The lack of cloud height information could potentially lead to mis-placed clouds. Have the authors noticed any persistent biases or noise related to the zenith angle of the SEVERI observations? In addition to cloud height-related issues, every observation comes with a minimum detectable signal but models mostly do not. For example, truly-existing thin cirrus may be undetected by SEVERI due to weaknesses in infrared detection of clouds and algorithm deficiencies. A satellite simulator would alleviate these issues a great deal, if implemented correctly.

Due to the moderate resolution of 20km in our comparison and the use of a discrete cloud mask, we calculate that this effect is negligible for this study, though the use of a satellite simulator is a very good point for future, higher resolution studies. We have added a paragraph on this point to section 3.2.

Due to the viewing zenith angle, the satellite observations can indeed be offset by up to a few pixels for high clouds. As we aggregate the data to the 20 km grid, the inaccuracy for a high cloud could be up to one model grid point. If we however calculate how often the cloud mask has different values than its vertically neighboring grid point(s), only a few percent of the overall mask can be wrong/shifted. In the comparison to the simulation model cloud mask, where the clouds are indeed smoother than the 20km resolution, any effect on the matching rates is negligible.

The cloud product from CMSAF is corrected according to the methodology and algorithm by Stöckli et al. (2019), including and validated with level-2 data for instantaneous (hourly) data (Stöckli et al., 2017). For level-2 validation the difference between CFC and SYNOP data is -0.3 with a bias corrected RMSE around 30%. The application of the CFC production does not required spatial correction as well (Bojanowski and Musiał, 2020). We therefore believe this data to be suitable for our cloud mask comparison without additional calibration.

There are many small typographical errors, mostly related to plurals. I noted many in the "technical corrections" but I am confident I did not document them all.

Thank you for indicate the typo errors. We also check it during our correction for the manuscript.

2 Reply to the *Specific Comments*

16: Which recent events in 2021?

We have added more specifics about recent events to the revision.

95:96: "It is recommended that the surface layer physics be set with planetary boundary layer physics in WRF." Who is recommending this? Are you recommending it or is it the official recommendation from the WRF developers? It would be best if you would provide a source for this recommendation.

We've updated the recommendation from the developers in the users manual as: *We note that the official documentation recommends to set the surface layer physics with specific planetary boundary layer physics in WRF (WRF, 2015).*

102: I recognize the need for shortening the parameterization acronyms. However, these shortened acronyms are used throughout the paper and are important to the interpretation of most figures so Table A1 and Table A2 should be added to Table 1 and Table 2. Table 1 and Table 2 have plenty of space for the shortened acronyms in parentheses behind the full names, for example.

Thank you for the recommendation, but we did indeed have trouble fitting it in Table 2 keep the acronyms and the physics configuration tables separate. We now merge the acronyms with the physics and parameterization tables (Table 1 and Table 2) and update the text in the section 2.2 as: *The Set 1 combinations and the acronyms for the WRF physics and parameterizations are listed in Table 1.*

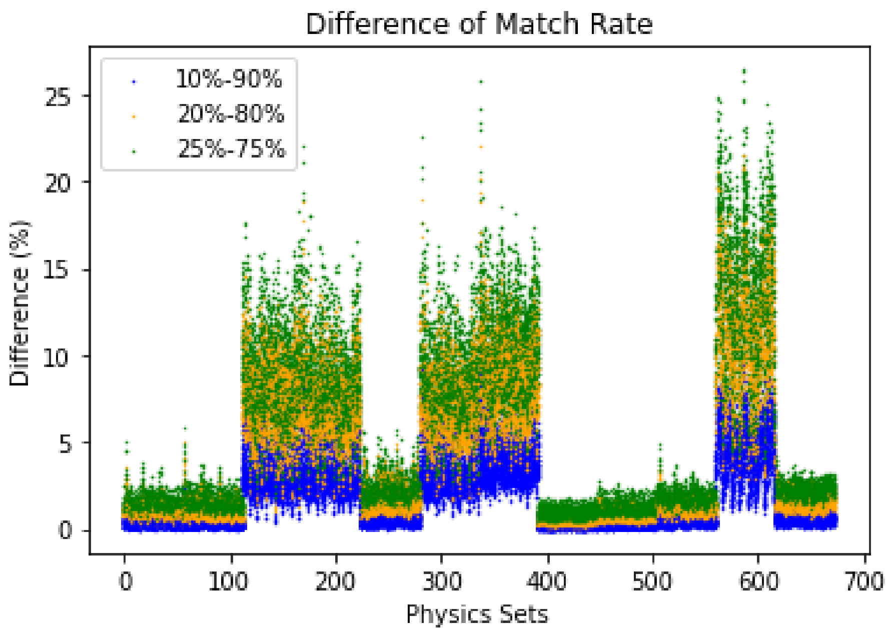

Figure 1. Sensitivity of threshold to matching rate for converting cloud fraction into cloud mask in the case 2015-08-23, which has the most dynamic of cloud information than other cases. The color of dot represents the difference of increase/decrease rate by applying different threshold. The base threshold is 5% and 95% for clear sky and full cloud cover, partial cloudy is in between.

123: Are your results sensitive to these near-arbitrary thresholds?

For the relative comparison of the model performances, we think the results are not dependent on this choice. The quantitative matching rates themselves are sensitive for some microphysics. As shown in Figure 1, the cloud mask is more sensitive to the microphysics of WSM3, WSM5, WSM6, Goddard, and CAM5.1, which can produce more cloud fraction between 0% and 100% than from other microphysics. The matching rate can increase up to 26.4%, which is produced by CAM5.1. However, with the increase of matching rate, the rate of miss (Figure 2 (a)) and over-predict (Figure 2 (b)) increase as well. The histogram (bin as 1%) shows that most of the cloud fraction is below 1% and overnon 100% and thus we use 5% and 95% to represent the threshold for clear sky and cloud cover, respectively.

124: The ASOS acronym needs a definition.

We've added the definition to the text as *Automated Surface Observing Systems*.

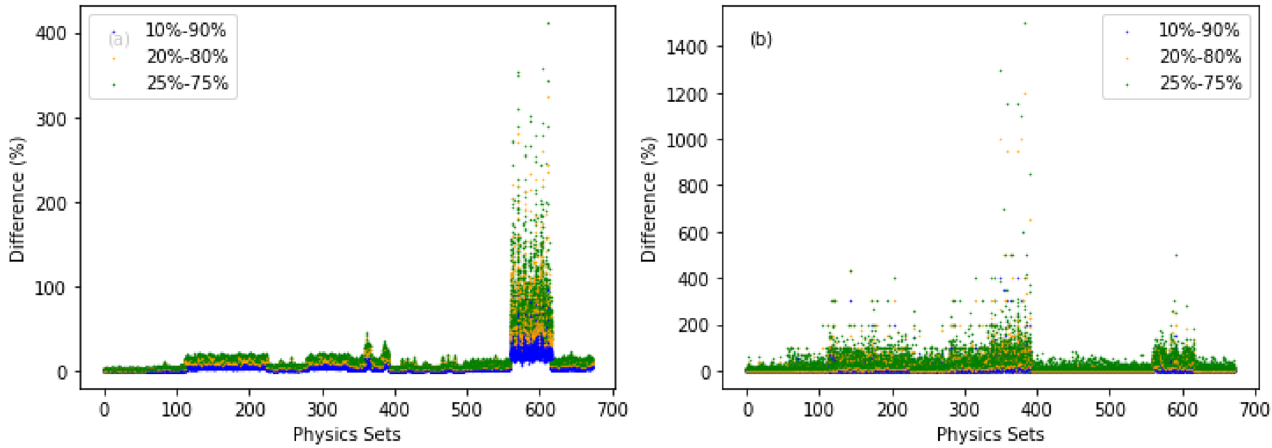

Figure 2. Sensitivity of threshold to the (a) missing rate and (b) over-predict rate for converting cloud fraction into cloud mask in the case 2015-08-23, which has the most dynamic of cloud information than other cases. The color of dot represents the difference of increase/decrease rate by applying different threshold. The base threshold is 5% and 95% for clear sky and full cloud cover, partial cloudy is in between.

135: The first model evaluation results utilize this Kappa score, but there is essentially no preview of what a low-Kappa or low-Kappa means in terms of agreement with observations. Please provide some interpretation of this metric.

We add the following text to the revision: *kappa has a maximum value of one 1 and can also be negative. The maximum kappa means a full match between two datasets. kappa between 0 and 1 indicates a partial match between the two datasets, while negative kappa indicates some anti-correlation in the matching (Pontius, 2001). A good model result should result in positive* κ.

135-140: What is N? Total number of subjects? I am also unsure what the "subjects" are. Please provide a definition for each variable in the equations.

We've clarified the text as : *where* $\bar{P}$ *is the sum of* P_i*, the matching rate of the* i^{th} *subjects or individuals being rated, for* k *categories.* $\bar{P}_e$ *is the sum of the category rate* p_j *over* j *and N is the total number of subjects.*

156: Why are these cases chosen? Does the domain experience considerable variability in these cases? Fronts with strong precipitation? Mesoscale convective systems? It is very important to explain why these days were chosen so please provide a short description of each and, more importantly, why simulations of these 6 cases are capable of summarizing the variety of weather conditions that these parameterizations are expected to simulate in an operational environment. Lines 172-181 provide a cursory description of what cloud cover patterns through each case, but not a justification of why these cases are *sufficient to understand the differences between the parameterizations.*

We did not have the resources to simulate all physics for a wide variety of weather conditions or to target outliers, but rather begin by taking 6 somewhat random or typical days from different months. We have supplemented the results with a long-term simulation in Set 4 of six months to include diverse conditions to confirm the general performance. The results are of course not absolutely definitive, but we take this as an economical approach to a very computationally expensive problem.

Section 3.2: Please elaborate on the description of the observational dataset. What instrument makes the observations? What techniques to they use in their cloud retrievals (BT-contrast, CO2 slicing, etc.)? What sort of processing takes the product from pixel-level to gridded, quality-controlled distribution?

We have updated the text as: *The data is corrected and generated from SEVIRI on METEOSAT-8, which uses the visable, near infrared, and infrared wavelengths to retrieve cloud information. The hourly CFC data has level 2 validation (Stöckli et al., 2017) for the accuracy of total synoptic cloud cover and the data is corrected by the algorithm from Stöckli et al. (2019) using the clear-sky background and diurnal cycle models for brightness temperature and reflectance.*

Figure 3: These are UTC times, right? Please state in the caption.

The caption is updated with "UTC".

Figure 3 caption: The caption says the colors represent both cloud cover and time of day. I think the second sentence should be removed.

We've removed this.

Figure 4 and Figure 5: These wallclock times would be more accessible to the reader if presented as hours, as is done with the Simulation Time. It would also be more convenient for the reader if the (a) and (b) plots had identical y-axis limits. They are very close now so why not make them identical?

We have updated the y-axis, but use seconds as unit for counting the time, as in computational matters seconds better represent the wallclock time instead of using hours. The simulation time is counted as hours because we record the wallclock time based on the hourly timestep.

Figure 4 caption and Figure 5 caption: There is no hourly simulation time, only total accumulated wallclock time

We have updated the caption to match the plots (the hourly plot was removed).

216-219: This mini-paragraph should be placed earlier in the manuscript because some science results have already been presented (Figure 6 and Figure 7). Near the first sentence in Section 4.2 or earlier would be good.

We have moved this mini paragraph to the beginning of Section 4.2.

255: "Accounting for the support of the simulation of the graupel mixing ratio for ESIAS-chem, we predominantly use the microphysics of WSM5, WSM6, and Goddard." is more understandable when written similar to, "We continue with the WSM5, WSM6, and Goddard microphysics parameterizations because they include treatments of graupel mixing ratio for ESIAS-chem.", unless I am misunderstanding the meaning of this sentence.

You have captured the meaning for sure. We have updated the text as: *Accounting for the support of the simulation of the graupel mixing ratio for ESIAS-chem, we predominantly use the microphysics of WSM5, WSM6, and Goddard.*.

*281: I'm confused about the "maximum of the boxplot". In Figure 14a, the boxplot endpoints do not appear to be 1.5*interquartile range greater than the third quartile (assuming you meant quartile instead of quantile). For example, the maximum boxplot edge for the W6-T combination is only a small amount greater than the third quartile.*

Yes it's quartile and we have updated the text as: *(as maximum or minimum of the data, or 3rd quartile*$\pm 1.5\times$*interquartile range)*. The top of the box is the 3rd quartile, and the maximum (the whiskers) of data is *limited* by the interquartile range (IQR). And therefore when the maximum/minimum of data does not greater or lower than the $1.5 \times IQR$, respectively, the whiskers does not have the length as $1.5 \times IQR$

284: Which cumulus parameterization can improve Kappa? Tiedke?

We have updated the text as: *Grell-3D and Tiedtke, which are more advanced than the Kain-Fritsch, can improve the Kappa*.

Figure 14 and Figure 15 do not really add much to the analyses because they present more data than can be reasonably interpreted by the reader. The most important data are the summary data, which are only shown as text. Also, these are only two of the days and there are no analyses that summarize the other four cases! I recommend banishing the time series plots to the supplemental material and replacing these two figures with heat-maps of RMSE, sigma_bar, and x_bar and span all 6 cases.

We believe the figures you mentioned are Figure 15 and Figure 16. We agree that the number are important to indicate the results and analysis. However, with time series we can see the comparison of spatial-average simulated cloud cover to the

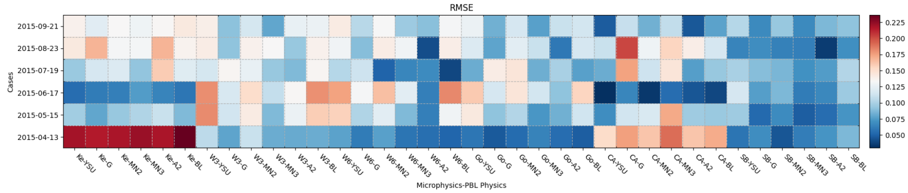

Figure 3. Heat map of the RMSE between the ensemble mean total cloud fraction and observation, averaged over the last 36 hours of the simulations, shown for each cluster of microphysics and PBL physics (columns) for all test cases (rows). In the colorscale, red represents higher RMSE and poorer performance.

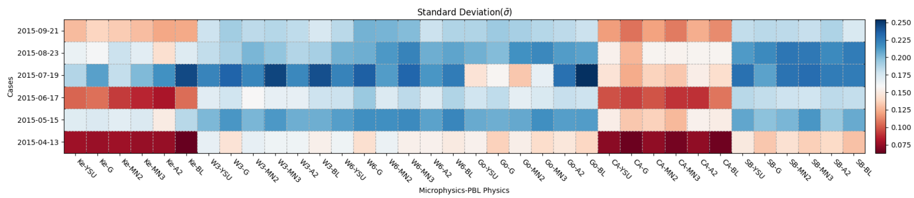

Figure 4. Heat map of the ensemble standard deviation ($\bar{\sigma}$) of the mean total cloud fraction, averaged over the last 36 hours of the simulations, shown for each cluster of microphysics and PBL physics (columns) for all test cases (rows). In this heat map, blue indicates larger $\bar{\sigma}$ and greater ensemble spread.

observation that changes with time. We remove Figure 16 and update Figure 15 and add the heatmap of $\bar{\sigma}$ and $\bar{x}$ to show all the results as following figures:

Also we've updated the text as: *Figure 3 summarizes the rmse performance of the ensemble mean against the observed domain total cloud fraction. Figure 4 illustrates the time-averaged ensemble spread with the standard deviation (std) of the domain total cloud fraction. From the rmse, SBU–Lin had the best mean total cloud fraction over all six cases, with a more variable cloud fraction according to the std. The WSM3, WSM6, Goddard, and SBU–Lin with MYNN3 and ACM2 produced more accurate average cloud fractions than the other combinations. Overall, the WSM series and SBU-Lin better represented the uncertainty than the other microphysics, while MYNN3 and ACM2 improved the simulation accuracy.*

310: Note that you've only investigated days during warm periods (mid-April to mid-September) so you cannot say with confidence that this is true for all time periods.

We hope this is addressed with the revisions above.

352:353: “but we should also consider the accuracy of the model physics”. Yes, good point and one that is often not considered in ensemble modeling

Thanks for appreciating this point.

354:355: I think this sentence is important but I am struggling to fully understand it. Please consider rewording for better clarity.

We have removed this sentence, as it does not address our point of view concisely. We have rephrased it as: *In our simulations, combining the ensembles into one multi-physics ensemble would enhance the spread, but this would be somewhat artificially due to the different biases of the model physics. The accuracies of the different model physics must then always been considered for multi-physics ensembles. We also note that the small ensemble spread reported in Jankov et al. (2017) may be due to the small number of ensembles, four for each physics configuration, yielding eight members in total.*

3 Reply to *Technical Corrections*

23: Remove the “to” in “study to the impact”

The sentence is rephrased and thus this error is not existed.

23-24: Parentheses around references

The wrong Parentheses is corrected.

Table 4 caption: Fix “Inaddition” to include a space

The missing space is inserted.

Figure 3: Move the legends upward into the white space and increase all text sizes

This figure has been updated.

210: “figures indicate” instead of “figures indicates”

The sentence is rephrased and thus this error is not existed.

216: “model is run” instead of “model is runs”

The misspelling has been corrected.

220: Please mention figure 8a instead of just figure 8

This has been corrected as "Figure 8a"

Figure 8: Please label the subplots in the figure.

The labels have been updated.

243: “wellby” should be “well by”

We have correctd this misspelling.

245: The word parameterizations is misspelled

We have correctd this misspelling.

265-266: You probably just want either “overall” or “over all” in this sentence and not both.

This has been corrected in the proofreading.

268: Missing space between sentences

We have updated the text to include the missing space.

Figure 15: Y-axes all say “Cloud Refraction” but they should say “Cloud Fraction”

This figure is updated and Figure 16 is removed according to the comment above.

308: There is a separate Conclusions section so there is no need to have "and conclusion" in this section heading.

Agreed, *and conclusion* has been removed.

339: Should say "not only changes the development of cloud cover fraction but also affects the"

Since the PBL physics is referring to the general term of "Physics", we use the singular verb as well.

References

User's Guide for the Advanced Research WRF (ARW) Modeling System Version 3.7, https://www2.mmm.ucar.edu/wrf/users/docs/user_guide_V3/user_guide_V3.7/users_guide_chap5.htm, 2015.

Bojanowski, J. S. and Musiał, J. P.: Dissecting effects of orbital drift of polar-orbiting satellites on accuracy and trends of climate data records of cloud fractional cover, Atmospheric Measurement Techniques, 13, 6771–6788, https://doi.org/10.5194/amt-13-6771-2020, https://amt.copernicus.org/articles/13/6771/2020/, 2020.

Jankov, I., Berner, J., Beck, J., Jiang, H., Olson, J. B., Grell, G., Smirnova, T. G., Benjamin, S. G., and Brown, J. M.: A Performance Comparison between Multiphysics and Stochastic Approaches within a North American RAP Ensemble, Monthly Weather Review, 145, 1161–1179, https://doi.org/10.1175/MWR-D-16-0160.1, https://journals.ametsoc.org/view/journals/mwre/145/4/mwr-d-16-0160.1.xml, 2017.

Pontius, R.: Quantification error versus location error in comparison of categorical maps (vol 66, pg 1011, 2000), Photogrammetric Engineering and Remote Sensing, 67, 540–540, 2001.

Stöckli, R., Duguay–Tetzlaff, A., Bojanowski, J., Hollmann, R., Fuchs, P., and Werscheck, M.: CM SAF ClOud Fractional Cover dataset from METeosat First and Second Generation - Edition 1 (COMET Ed. 1), https://doi.org/10.5676/EUM_SAF_CM/CFC_METEOSAT/V001, https://doi.org/10.5676/EUM_SAF_CM/CFC_METEOSAT/V001, 2017.

Stöckli, R., Bojanowski, J. S., John, V. O., Duguay-Tetzlaff, A., Bourgeois, Q., Schulz, J., and Hollmann, R.: Cloud Detection with Historical Geostationary Satellite Sensors for Climate Applications, Remote Sensing, 11, https://doi.org/10.3390/rs11091052, https://www.mdpi.com/2072-4292/11/9/1052, 2019.