Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2019-351-AC2, 2020

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Sensitivity of *CHIMERE* to changes in model resolution and chemistry over the northwestern Iberian Peninsula" *by* Swen Brands et al.

Swen Brands et al.

swen.brands@gmail.com

Received and published: 10 January 2020

Response to Anonymous Referee #2

Referee comment: Quite a few different models are currently in use for chemistry transport modelling on the regional scale. Still many questions concerning the validity of the model results with respect to the necessary complexity of the chemistry mechanisms, the needed quality of underlying emissions or a sufficient grid resolution are not finally answered, yet. On the other hand these type of models are more and more applied for short term air quality forecasting. The present manuscript offers a sensitivity study conducted with the model CHIMERE, which was set up for the north-western Iberian

Peninsula, a region with complex topography and a long, structured coast line. Meteorological data from the WRF model was fed into CHIMERE, for the emissions the HTAP v2.2 inventory was used. Two different horizontal and two different vertical resolutions were tested as well as two chemistry mechanisms (SPARC-07 A and Melchior mechanism). Model derived nitrogen dioxide, PM10, PM2.5 and ozone concentrations are discussed and compared to observational data from a regional air quality network based on statistical measures. The comparisons were done for daily minimum and maximum values of those substances. The underlying investigation for the article is a straightforward sensitivity study with a pragmatic choice of varied parameters (âĹij12 km andâĹij4 km horizontal grid resolution and 10 and 20 vertical layers to 500 hPa). The comparisons between the model set-ups is done (for daily extreme values only) by using the bias, Pearson correlation and standard deviation ratio for the chemicals under investigation in relation to respective observed values. In addition, a mean absolute error is chosen to compare the runs to a chosen reference case (the computationally cheapest). The results of this quite "technical" study may be interesting for those intending to set up CHIMERE for purposes, for which computing resources are a limiting factor. The result section is dominated by describing point for point in words, what the figures show anyway. No deeper investigation and discussion of possible reasons for the discrepancies among model runs for the different the set-ups are offered. E.g. for the strong statement in the conclusions section that "CHIMEREâĂšs performance is very poor" it is in my opinion not sufficient to just speculate that the used emission inventory has deficiencies for the region: This definitely should be investigated (e.g. by consulting other inventories).

Response: We would like to thank you very much for taking the time to review our manuscript and for sharing your thoughts with us. During the last months, we have made large efforts to take into account your valuable critics and suggestions and this is why it took a little more time to complete this review. Most importantly, we ran all model experiments once again with another emission configuration (EMEP 2017 downscaled with several proxies) and conducted even more sensitivity experiments

to assess the specific roles of 1) the population proxy used for downscaling raw anthropogenic emissions, 2) the resolution of the meteorological input data from WRF and 3) biogenic emissions. We added new figure and table material, rewrote nearly the entire manuscript and now provide a more profound interpretation of our findings. Also, after an reassessment of the corresponding results, the text passage stating that "CHIMERE's performance is very poor..." was removed since it no longer holds.

General comments

Referee comment: I doubt that the paper in its current form would be of great interest for the typical ACP readership. It fits not well into the journals scope. Neither the used procedures are sufficiently innovative nor the analysis of the results is deep enough to provide transferable insights. The results, which might be interesting from a technical point of view, are addressing a specific region only, they could have wider implications for the modelling community in atmospheric sciences, if the analysis would look closer at the influence of the heterogeneous terrain and coastal flow effects on the findings. I leave a final consideration to the editor. In general, a more thorough discussion of reasons for the presented deviations between results from runs for the different model setups and from the measurements is needed. Since the results have some value for air quality modelling, I would suggest to the authors considering a submission, though in a revised and extended form, to a journal, which is more devoted to technical analyses for modelling.

Response: The revised manuscript has been nearly complete rewritten, adding new figure and table material as well as an interpretation of our findings whenever we think this is justified from our experimental analyses (e.g. in lines 265-271, 376-378, 481-492 and 516-523) However, despite our efforts to thrive away from a purely technical paper in order to take into account your suggestions, the principal aim of our study is not to go deep with specific meteorological phenomena or the with differences between specific experiment pairs, but to estimate the role of as many parameters as possible on CHIMERE's average performance when compared to observations in order to make

recommendations on the best model configuration. This is crucial for setting up reliable model simulations, be it for real-time forecasts or in analysis mode, i.e. whenever CHIMERE is assumed to realistically reproduce the real world. To our knowledge, such an endeavour has not been undertaken yet for this type of study region (mid-latitudes, coastal, complex orography, relatively low pollution) albeit regional air quality prediction schemes are demanded there by politics, scientists and the general public.

Due the "broad" approximation or our study, an in-depth analysis for each of the many differences found between the experiments cannot be provided in one single study. However, thanks to the large extension of applied numerical experiments we are now in the position to provide an extended list of conclusions, some of which are not only of purely technical but rather general interest from our point of view (see Section 4 of the revised manuscript).

In summary, we are confident that our study is of interest for ACP and also would like to mention that this journal has published similar studies in the past (e.g. Bessagnet et al. 2016, see reference list of the revised manuscript). We would of course also welcome handing the manuscript over to GMD, should the editor decide so.

Some mayor points

Referee comment: In addition to the remarks made above some further issues (shortcomings) of the manuscript need to be mentioned.

Emissions: The backbone of air quality studies, especially when compared to observations, are suited emissions. The authors use HTAP v2.2 for the year 2010, while the study period are two summer months of 2018. A discussion of implications of this mismatch is missing. If the necessary observational data would be available, this technical study could have been performed for 2010 using appropriate meteorology. Or the 2010 emissions from HTAP could have been compared to more recent emission data and may be scaled (2010 compared to 2018).

Response: To solve this problem, a total of 8 new experiments were run with the anthropogenic emission inventory EMEP 2017, which is the most up-to-date inventory provided by the European Union (see Section 2.2). It is then straightforward to assume that the effects arising from the one year difference to our the study period are negligible.

Referee comment: HTAP v2.2 emissions are provided on 0.1âŰęx 0.1âŰęgrid, which does not directly fit the used resolution (0.15âŰęx 0.15âŰęand 0.05âŰęx 0.04âŰę). The regridding was done without downscaling. The authors do not explain what this mismatch means. But they should, since the resolution of the emissions may affect the results differently on the two grids, certainly a limitation of the study.

Response: We are afraid this is a misinterpretation arising from a too short description of the emission postprocessing in the first version of the manuscript. In fact, the 8 experiments shown in the first version already had been downscaled to the 0.05° x 0.04 grid using land-use categories, which is the basic option of the emiSURF emission postprocessor shipped with CHIMERE. We never tried to run CHIMERE with emissions on another grid and think the model would probably crash when trying to do so.

Referee comment: HTAPv2.2 emissions are provided with a monthly time resolution. The authors do not inform the reader how they dealt with this coarse resolution when feeding the emissions into CHIMERE. Did they use time profiles on the emissions or did they feed in just the monthly means? A higher time resolution is needed, when comparing to daily maximum and minimum values of the observations (i.e. NO/NO2 and O3 relations are strongly dependent on the daily emission time profile).

Response: Similarly, already for the experiments in the first version of the manuscript, monthly HTAPv2.2 emissions had been disaggregated to hourly timescale using the standard information in the CHIMERE pre-processors (Mailler et al. 2017). In the revised manuscript, this is stated in lines 151-154. Referee comment: Any way, it is not adequate to state “CHIMERE’s performance for NO2 is very poor” and point

at the same time to possible deficiencies in the emissions (in that case it is not CHIMEREâĂšs performance). Page 17/Line 16 (1 on that page). To improve their manuscript, the authors need to devote an entire section to these emission issues.

Response: You are absolutely right, this conclusion has been removed from the manuscript. Most of the large percentage bias values for NO2 are located at background stations where the observed mean concentrations for this species are generally very low, meaning that an absolute bias of only a few ug/m^3 (which is obviously not important) translates into a large percentage value. This has to be considered when interpreting the results and is stated in lines 376-378 of the revised manuscript.

Referee comment: Meteorology The study area is characterized by structured terrain and a complex coastline. The question appears whether the meteorological model in use (here WRF) is resolving the local flow features sufficiently well (terrain effects and summer sea breeze circulations)? The 12 km and 4 km runs might produce different results here, not unimportant, since quite a few of the observational stations are located near the coast or in hilly areas, where local flow fields might dominate the dispersion of emitted substances. Differences between modelled and observed concentration maxima/minima could partly be due to the quality of the meteorological simulations rather than entirely ascribed to CHIMERE. A discussion of the quality of the meteorological fields and reproduction of local features (best against observations) needs to be provided.

Response: The WRF configuration used to force the CHIMERE experiments has been in operational use at the Galician Meteorological Service (MeteoGalicia) for now more than a decade. Its performance is supervised by a team of operational forecasters as part of their day-to-day business and any large model deficiencies, if possible, have been corrected in the course of time. In Supplementary Figure 1, the performance of the model is illustrated for a typical summertime heat day for the high-resolution domain (4 km), showing that local wind and temperature features are reproduced fairly well by the model. In the revised manuscript, this is pointed out in lines 95-99.

Furthermore, to assess the influence the "meteorological resolution" has on CHIMERE's performance, an additional experiment was designed in which the fine CHIMERE domain (0.05° x 0.04°) was not run with the 4 km WRF simulation (as in all other experiments) but with the coarse 12 km simulation (see Table 3). If the regional orography, the structure of the coastline or the land-sea contrast was key for CHIMERE's performance, then this "coarse meteorology" experiment should produce considerably worse results than the reference experiment run with fine-scale meteorological input (compare FM20H-C with FM20H in the boxplots). However, our results indicate that the performance differences between these two experiments are generally smaller than expected and that other factors are more influential on model performance (see lines 304-307, 336-337, 401-402 and 516-523).

Referee comment: Data handling for results section. The evaluation of the modelling results using observational data only considers maximum and minimum values. No information is provided how the values are taken from the respective series. There are several possibilities. Is the maximum taken from the observational time series and compared to the model result for the same time stamp? Or is the maximum taken from the observational time series and compared to the maximum of that day in the model series, which could occur at a different time (a considerable time shift might be possible). Or is the model output leading the selection? The same questions holds for the comparison of minimum values. It is interesting to learn whether maxima/minima are missed in general or whether there is a certain time shift.

Response: The modelled daily extreme values are calculated on the hourly model data for a given day and the observed daily extreme values are calculated on the observed hourly data for this day, i.e. the calculation for the model and observations is independent from each other meaning that the extremes can occur at different times of day. In the revised manuscript, this is stated in lines 223-225. Albeit a detailed assessment of the modelled vs. observed daily cycle is out of the scope of the present study, in the revised manuscript we now as well include the verification results of the

hourly data (see Section 3.3 and Figure 10).

Referee comment: Although maximum values are important for air quality and health related studies, it would have been instructive to additionally analyse better time resolved concentrations to assess the models ability to reproduce daily cycles in different regions (i.e. O3) and the variability in the model and observational data. Both should be available with an hourly resolution. This could help also to discuss reasons for the deviations between the different set ups. Show a few selected time series (for different quantities, different locations) of modelled versus measured concentrations (hourly resolution); may be more of that in the supplement. This would be very instructive.

Response: In Section 3.3 and Figure 10 of the revised manuscript, we have included the verification results for hourly data which provides insight on the average performance of the distinct model experiments on sub daily timescale. In many aspects, the results for the hourly data are similar to those obtained for the maxima. As argued at the start of this response letter, an in-depth analysis of the daily cycles is one of the efforts to be undertaken in future studies since the manuscript size is already at its upper limit.

Referee comment: In an additional evaluation step the Mean Absolute Skill Score (using the reference run CS10) was provided separately for background, industry and traffic locations. In general no bad idea. It should be discussed, why for the background and industry stations NO2 and O3 perform so differently for the different settings? Are they really that much decoupled? Are in case of the background stations O3 concentrations influenced predominantly by BVOCs from MEGAN? This needs a more thorough discussion.

Response: For the revised manuscript, we have performed an additional model experiment in which the biogenic emissions (comprising biogenic VOCs and NO) were intentionally turned off. We then specifically looked at the model performance for the background O3 and NO2 maxima (see Supplementary Figure 2). At these sides, both

the O3 and NO2 maxima are reduced when biogenic emissions are intentionally turned off, meaning that the concentration of both species is affected by this emission type. In the revised manuscript, this is pointed out in lines 477-493.

Referee comment: Quite a few of the stations used as bases for the statistical analysis are so called “traffic stations”. These stations are often hotspots for some of the considered substances (NO2, PM10), because local traffic emissions are dominating (not resolved by HTAP). The authors should inform the readers about these traffic stations. Are some of them located within street canyons, which channel the flow and dispersion near the ground?

Response: The traffic stations from the Galician air quality monitoring network are not located in street canyons but rather in relatively open terrain, as required by the Spanish legislation. We are well aware that our CHIMERE experiments run at a resolution of 0.05° x 0.04° do not resolve concentration differences on the street scale but it is nevertheless interesting to know what the model suggests for these stations, particularly taking into account that the downscaling procedure applied to the EMEP emissions uses road traffic density as a proxy to re-allocate the raw emissions on the sub grid scale. Hence, we do not rule out the corresponding results unless they produce such large outliers that they hamper the visualization, in which case these outliers are simply not shown (as in the boxplots). We also use outlier-resistant statistics such as the median instead of the arithmetic mean (in the boxplots and Figure 10).

Referee comment: This very local data is compared to model results obtained with a relatively coarse resolution (4 km and more horizontally). This seems not to be appropriate. It would be recommended to nit consider traffic station in the statistical evaluation (or do it separately to see the effect).

Response: A separate evaluation of model performance for each station type (background, traffic, industry) is provided in Section 3.3 and Figure 10 of the manuscript. The corresponding text passages have been largely extended to discuss the corresponding

differences.

Referee comment: Also for the other measurement stations, it would be useful to know, how they are located within the modelling grid. A station located close to the grid boundary or grid corner might be better represented by the neighbouring grid cell(s), dependent on local terrain effects. A study of this localisation effects for the 12 km and 4 km grid resolution would be helpful.

Response: In the revised manuscript, a total of 60 maps are provided (in 4 Figures) in which the observed concentrations are plotted on top of the modelled ones. These “overlay maps” show that the modelled concentrations change relatively smoothly in space even on the 4 km grid. Hence, a hypothetical shift of the mesh by a few gridboxes in any direction would return similar results.

Referee comment: Point measurements are compared to grid box means. The issue of the spatial representativeness of the stations needs to be addressed. Which of the background stations are in forested regions (BVOC emissions)?

Response: The fact that areal mean values from a model are compared to point measures is a common issue in the evaluation of these models and can’t be overcome unless the modelled values are statistically downscaled to match the observations , in which case the bias could be probably further reduced. In our study area, almost all background stations are in forested area and thus influenced by BVOC emissions. In the revised manuscript, this is stated in lines 68-69 and line 103.

Minor points

Referee comment: Since computational costs are one of the parameters of interest/motivation of this study, it would be helpful for the reader to get quantitative information on the model runtime for the different experiments.

Response: We have included the runtimes for the 8 experiments run with EMEP in the last column of revised Table 3. The runtimes for the respective experiments run with

HTAP are nearly identical, but were unfortunately not saved. This is clarified in lines 191-194 of the revised manuscript.

Referee comment: On page 8, line 18, CS10 is flagged as the computational cheapest experiment. In table 3, in which the experiments are ordered according to their computational costs, CM10 is the first mentioned and here apparently the cheapest. Clarify.

Response: Thanks for careful reading, this error has been corrected. CS10 was the computationally cheapest experiments. In the revised manuscript, the ordering according to the computational costs only applies within the same emission configuration. Within the EMEP group (ending on "E"), CS10E is the cheapest experiment. Within the HTAP group (ending on "H") CS10H is the cheapest one, and so on.

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2019-351, 2019.

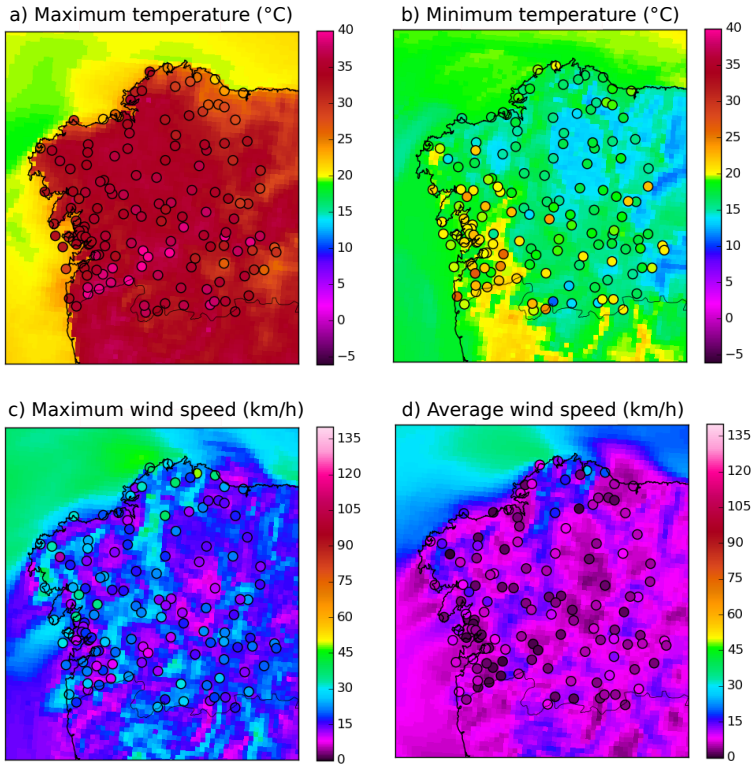

Fig. 1. Suppl. Fig. 1) Daily observations vs. WRF forecasts for a August 5, 2018, (a) max. and(b) min. temperature, (c) max. and (d) average wind speed

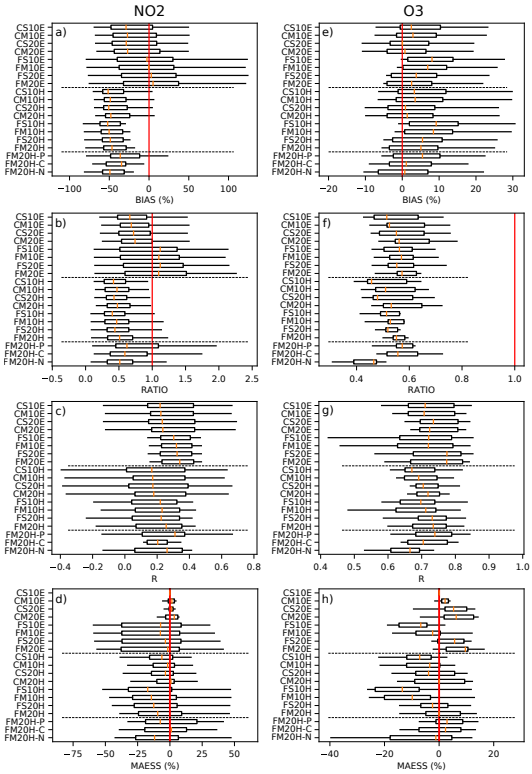

Fig. 2. Suppl. Fig. 2) As Figure 4, but for background sides only