Atmos. Chem. Phys. Discuss.,
doi:10.5194/acp-2017-257-RC3, 2017

Atmospheric Chemistry and Physics Discussions

Interactive comment on "Advanced error diagnostics of the CMAQ and Chimere modelling systems within the AQMEII3 model evaluation framework" *by* Efisio Solazzo et al.

Anonymous Referee #3

Received and published: 10 May 2017

Review of "Advanced error diagnostics of the CMAQ and Chimere modelling systems within the AQMEII3 model evaluation framework, but Solazzo et al. Section 4.3:

General comments: My overall rating for this paper is "major revisions" – the paper is valuable from the standpoint of showing how the errors in air pollution models may be linked to particular time scales, but some of the methodology appears to be circular, and they attribute some of the errors to specific processes without sufficient evidence. Further model simulations or rewriting portions of the paper to provide more caveats on their conclusions is necessary. There were issues with regards to formatting of the paper (ordering of the figures not matching the text) and inconsistent terminology, which should have been addressed prior to submission.

My main concerns are given a "**" symbol in the following detailed comments

Abstract: ** The abstract makes the conclusion that the representation of the PBL is pivotal, but I see little direct evidence in the paper that this is the case – rather, that the PBL representation, based on its diurnal variability and a version of CMAQ which was not tested by the authors analysis, may have a role in the errors. That role has not been established in the authors work – which suggests that the errors have time scales similar to the PBL variation. The potential for any other sources of error with a diurnal timescale to have this effect is insufficiently addressed in the main body of the text and the conclusions. I'll return to this point here and there throughout my review.

** Point (i) made on line 27 suggests that much of the total ozone quadratic error occurs in the component with time scales > 1.5 days – this makes the contribution of identifying a time scale which contains much of the error. However, I'm concerned that this conclusion is the result of a circular / potentially bias-inducing process in selecting the stations for comparison. Later in the paper, the regions for comparison have been selected based on hierarchical clustering which the authors mention select for longer time scales, and spatial averaging is also used, which will tend to minimize the shorter time scales. The finding that much of the error lies in the larger time scales may thus be a result of the selection of stations for analysis. If the entire dataset is used, within NA and within EU respectively, will the same finding still hold?

Point (iv): CMAQ ozone error has a negligible dependence on errors in NO2, while NO2 errors are important in Chimere – a sentence on why this might be the case would be useful (e.g. are the NOx emissions in the European domain much more poorly characterized than in the North American emissions data, hence the European values are more influenced by that error? Point (v): did zeroing CMAQ's winter anthropogenic emissions have no impact on the ozone concentrations (this is how I'm interpreting the word "null" on line 34) – that's a surprising result; there should have been some response due to the removal of NOx via the NOx titration path, at the least. Can the authors explain why one domain is much more sensitive to anthropogenic emissions

than the other, or can these differences be ascribed to differences in the two modelling frameworks?

Comments on the main part of the manuscript: **Page 3, lines 109 to 115: the first of the five enumerated results in the abstract is that 70 to 85% of the total ozone quadratic error is may be attributed to fluctuations associated with time scales > 1.5 days in duration. Here, the sections chosen for analysis are described as being those which cluster closely for time scales of > 1.5 days in duration. The regions for analysis have thus been selected based on the level of similarity in that time scale range, so perhaps the result that most of the error lies within that time scale is not surprising – it does however raise the issue of whether the finding is broadly applicable, or is the result of this selection procedure. If all stations are included in the analysis, rather than the subregions indicated in Figure 1, is finding (i) from the abstract still true? Can the authors provide an argument regarding why the finding is more broadly significant beyond a possible selection bias in how points were selected for comparisons?

Line 128: I think the writing might be a bit unclear – it sounds like the model values being subjected to a two-hour average when the obs are a one-hour average? Might be better to use "The model concentrations were assumed to be linear between the instantaneous on-the-hour reporting times; the integration (average) between those times was used to construct hour starting (or ending) values in order to more directly compare to the averaging used in the observations." I suspect that what is meant in the sentence, please confirm and rewrite for clarity.

**Lines 132 to 133: Why were the data spatially averaged? This needs to be discussed: how does this aid in or improve the analysis? This will tend to smooth out local sources of short term variations, which will reinforce the tendency of the resulting data to have more error associated with the longer time scales (point (i) issue noted above) – are there other reasons why this is done? Line 136: "Missing values have not been imputed" – please clarify: if you are spatially averaging data at each hour, does that not mean that missing data are being imputed spatially?

Line 145: shouldn't that be "BEIS"? Lines 142 to 152: somewhere, mention the extent of the model domain in each case (or add that to Figure 1 as an outline, perhaps?).

**Lines 161 to 165: why was the same lateral boundary condition not used for both domains? Please include an explanation. A zero or non-zero value may have a large impact on the model results; the response of either model O3 may not be linear with respect to the magnitude of the boundary ozone value chosen. For example, a model initialized with 35 ppbv O3 would have more OH on the model inflow boundaries than a 0 ppbv boundary condition. Were any tests conducted showing the potential impact of the choice of boundary condition level on the simulations?

Line 173: Ah – I had been wondering what had been meant by "total ozone quadratic error" in the abstract. Please change the abstract to use either "mean square error" or "total ozone quadratic error (mean square error)"; more people are familiar with that term than total ozone quadratic error, and the terms have not been introduced in the abstract.

Section 3.1: The difference between North American and EU responses to zero emissions is very interesting. A few words describing the flow field in the relevant months in each case would be worthwhile (e.g. I'm assuming that the spring maximum in North America is likely due to mixing events over the Rockies; trop-strat folds and the like – are the EU O3 highs for zero emissions similarly a reflection of transport, or are this local formation)?

Figures 2 and 3: modify the captions: "Average monthly (right column of panels) and diurnal (left column of panels) curves constructed. . ."

**Lines 208 – 225: The authors mention the work of Appel here as a possible explanation for the more rapid decrease in ozone concentrations in NA1 and NA2. The timing seems less off in NA3, EU1 and EU2, but again the model leads the observations in EU3. So, the question: why is this effect more prominent in some areas rather than others? For example, one possibility is that the stomatal conductance and heat ca-

pacity values are more incorrect in regions NA1, NA2 and EU3 than elsewhere, if one assumes that the main cause of the problem is related to those two terms. Is there evidence to suggest that this might be the case (e.g. did Appel describe vegetation types where these differences would be the most noticeable, and did those correspond to the same regions with the largest differences)?

Section 3.1: How does this rule out other possible causes of diurnal error in the simulations?

Section 3.2: This section was very interesting – I'm wondering if the authors can suggest possible reasons for their results. For example, the NA3 subregion had little impact of emissions on night-time model performance for total MSE, but daytime performance showed a more significant impact. I'm wondering, for example, if there are sufficient natural emissions of NO in the San Joaquin valley to titrate the available ozone, hence little impact?

Lines 260 – 272: Implications, please? The anthropogenic emissions affect the bias, variance and covariance of ozone formation. This seems to have taken a back seat to PBL as a source of error elsewhere in the paper.

**Line 267-268: The absence of a variance error in the base case tells us that the standard deviation magnitude is the same for both base case and observations – the authors ascribe this to the emissions being of the correct intensity, but the reasoning for this conclusion is unclear – please explain. For example, if two time series are identical aside from a bias offset at every time, they will have the same net standard deviation from the mean, and much of the error in the zero emissions case is in the bias term – which is what Figure 4 seems to show. That to me implies that the temporal / spatial distribution of the emissions, potentially coupled with the variability due to the meteorology, results in the two time series having the same standard deviation – but not that the emissions have the right intensity (I'm interpreting intensity here as magnitude). Maybe I should ask, "What do the authors mean by 'intensity', in this

context, and how specifically does this explain the absence of a variance error?"

Lines 273 to 276: yes, this makes sense; the boundary condition being a constant in this case so only the bias should be affected. However, you'd want to be careful about assuming that the behavior of the error would be the same with a time-varying boundary condition, which would presumably have variance and covariance errors associated with how well that boundary condition actually captures the inflow conditions. Lines 277 to 285: Interesting: the bias is the dominant part of the deposition effect, which is what I would expect. The covariance term contributes at a few orders of magnitude lower – why? Spatial variability of the deposition field due to changes in ground cover, responses to meteorology of the resistance terms, e.g. temporal variation of the stomatal size? One implication is that getting the deposition magnitude "right" is more important than, e.g., the time dependence of surface resistance. Worth mentioning in the abstract/conclusions?

Line 299: Please clarify this a bit, viz: "resemble those of the daylight base case (Figure 6a, left column), but reduced in magnitude during winter. . ."

**Figure 8 and 9 and similar figures: a general comment: These figures would benefit from a common (perhaps logarithmic) colour scale being used, a scale based on something other than an even division from maximum to minimum in each figure alone (e.g. -80, -50, -30, -10, -8, -5, -3, -1, 0, 1, 3, 5, 8, 10, 30, 50, 80, 100, 300), The graphics quality is poor (images are very fuzzy in the version provided to reviewers). It's difficult to cross compare the different panels to get a feeling for the relative magnitude of the changes in each case, due to these combined issues.

** There are two places where the figure ordering doesn't follow the standard convention of numbering the figures and presenting them in the order in which they are discussed in the text, this being the first. This should have been addressed prior to submission. Given that Figures 4, 6 and 8 are described together in the text, followed by Figures 5, 7, and 9 – they should be reordered and numbered in the order in which

they are discussed.

Lines 308 to 310: what is the (potential) physical significance of the differences between the different regions? Much more of the error seems associated with the covariance than with the NA simulation. Assuming that the two meteorological models are approximately the same in accuracy and timing of events, would this imply that the Chimere model has more issues with the timing of emissions than with the magnitudes?

**Line 311: "Removing the anthropogenic emissions had almost no effect on the covariance share of the MSE". This might be suggested by Figure 5, though since the scale is logarithmic the differences may not appear large. But Figure 7 (a) compared to Figure 7(c) implies that the share of the covariance has decreased significantly for the zero emissions case, going from less than 10% for the continent, for the base case mean ozone, to over 75% for the zero emissions mean ozone. I agree that the variance portion is unchanged, though. Please correct or clarify the statement starting on line 311.

Line 327-328: Another reason to wonder why choose 35 ppbv as the boundary condition for one model simulation and 0 for the other. It's unlikely that a model would use 0 as its boundary condition... it would have been better to use a constant value for CMAQ, perhaps chosen based on upwind observations. Why was 35 ppbv chosen for Chimere, and zero for CMAQ?

Line 333-334: Interesting. This implies that the variability of the boundary conditions themselves become more important in winter, as well.

Line 335: RMSE or MSE? Please use consistent terminology throughout the paper. I think this should be MSE in this section. Line 336: "... and bias error during the night-time (Figure 5e)."

Line 341: Clarify with an additional reference to the figure demonstrating the point

being made, viz. "a drastically reduced covariance error compared to the mean ozone (Figure 7a); the timing error is now shifted..."

Figure 6e versus Figure 7e. Why is a much larger portion of the daily maximum 8hr ozone MSE for Chimere in variance, while it's almost non-existent in CMAQ? I suspect that this may represent issues with the timing of the emissions in the European domain. It would be useful for the authors to state the values of sigma-m and sigma-o for both models for 8hr daily max O3. My guess here is that the Chimere value of ïĄşo is larger than the European value of sigma-m, which would indicate a larger variability of conditions in the observations than in the model (given that I'm assuming the meteorological variability is likely to be the same, this makes the emissions/AQ model variability the likely culprit). Actually stating the values of sigma-o and sigma-m would help. It would be good to compare Figure 6 e and Figure 7e in this regard, too, in the text, since they are so very different, and to provide some possible explanation for these differences.

Figure 9: all figures look alike to casual inspection, since the colour scales differ in extent. See the comment above regarding Figures 8 and 9 and the figures of this type in the text.

Lines 348 to 352: Would the authors be able to explain why the impact of deposition might be so different between the two models? E.g. perhaps a paragraph in which how deposition is implemented at the code level might explain this. Are both models using deposition as a flux boundary condition on the vertical diffusion equation, for example, or is one electing to remove the mass from the lowest model layer? Are the magnitude of the deposition fluxes larger in Chimere than in CMAQ (e.g. compared the total O3 deposition from the two models)?

**Section 4.1 and Figures 10,11: This is interesting, but I'm not sure of the relevance to one of the main issues with regards to correctly predicting air pollution: accurate prediction of high concentration ozone episodes, which are usually of a few days duration. I think this may need a bit of rewriting to focus on the implications of the analysis

towards these episodes. For example, having more energy in the differences in the longer time scales implies that boundary conditions are important again (line 377); the analysis shows that much of the energy in the differences between model and obs is associated with longer time periods; whole year, 1 to 2 months, etc. However, the analysis elsewhere in the paper suggests that boundary conditions are more important only during the winter. I'm not sure, based on the discussion, what this adds to the overall analysis. A suggestion: can you expand the y-axis to focus on time scales of hourly to synoptic (e.g. 5 days), perhaps as a second column of panels in Figures 10 and 11, modify the colour scale accordingly, and focus the discussion on that part of the power spectrum? It also might help if the x-axes of Figures 10 and 11 were done in months of the year. The key question to be discussed in this section should be the extent to which this analysis can suggest deficiencies in the models on the time scales associated with ozone episodes. E.g. an expansion of lines 386 to 391, with less of a focus on the longer time scales, unless the authors can make an argument that the longer time scales are in some way influencing high ozone days, and add more to the issue of boundary conditions than has already been discussed.

Line 393: How well can a model with the relatively low resolution of Chimere in these simulations (25x18 km) be able to capture the daily variation in upslope/downslope winds in the Alps/Po valley – and have there been any higher resolution simulations in the vicinity of EU3 which have better comparisons to observations than attempted here?

Line 398: yes, fold events might be an issue as well – or their combination with upslope/downslope winds. The ozone high in the springtime in North America has been linked to Troposphere/Stratosphere fold events as well, and the events tend to be better captured at higher resolution (though some low resolution models include parameterizations for the events). Transport of fold events to the surface sometimes requires coupling with smaller scale processes such as convection or upslope/downslope winds; the resolution may thus play a factor in this issue.

**The second place where the figure ordering does not match the order of appearance in the text occurs at this point in the analysis. This is disturbing in a multi-author paper which has seen one level of pre-peer-review at the ACPD stage and presumably peer review in some of the writers' institutions; the figures should be ordered and numbered according to their appearance in the text. Figures 12, 13 and 14 appear after Figures 15 and 16 in the text – please renumber the figures to appear in the sequence they appear in the text, and reorder and re-caption the figures accordingly. Perhaps these were last minute changes to the manuscript?

Section 4.2, 4.3:

**Line 413: "explains" is too generous here. "The majority of the variance is accounted for in the 24 hour lag autocorrelation" might be a better way to phrase this. This sort of analysis identifies the time scales associated with the model error – which is a very useful thing to do – but does not describe the causes of those errors. There are a number of variables listed (lines 414-417) which have been known to affect ozone formation for several decades, many of which are interlinked; e.g. the incoming shortwave radiation may be affected by the cloud cover in turn affected by the relative humidity, etc.. The variables are not independent. The key finding of these sections is that a 24 hour (or multiple of 24 hour) periodicity exists within the ozone errors. A similar periodicity is found for the wind speed and temperature errors – but the authors imply a causal link to the meteorological modules (line 433-434), which I don't think is justified based on their analysis ("correlation /= causation"). Here's two alternative possibilities (neither of which may be the "actual" cause, my point here is that unless there's direct evidence, the linkage suggested is hypothesis rather than the result of analysis):

(1) Suppose the diurnal temporal allocation of the emissions data is incorrect, so that the model emissions peak at the wrong time of day – this would result in a diurnal error variation which has nothing to do with the meteorology. The authors later acknowledge this possibility later in section 4.3, line 483 – but in the Conclusions (lines 605 to 609) focus only on the PBL possibility – where really, it could be either, or something else

altogether. The only solid conclusion is that the error does have that periodicity – which is very valuable information for model developers! However, the evidence for PBL being the main culprit is not clear from the analysis presented here.

(2) Suppose that the actual radiation reaching the surface was attenuated more than in the model, in a fashion where the amount of attenuation depends on the solar zenith angle? This would also result in a diurnal signal to the error (and possibly influence some of the other meteorological processes noted as well). And again have little to do with the PBL.

The authors go on to link the analysis (which only shows the temporal nature of the errors) with the PBL dynamics in both models, and present that link as one of the conclusions of the paper (line 605-609) and in the abstract. This link is not justified by their analysis, which identifies time scales only, not specific processes. The suggestion that the issue observed in the current work has been addressed in a more recent version of CMAQ (line 436) is unfounded – unless the authors make use of that specific version of CMAQ and repeat their analysis, and find that the error is completely removed. The conclusions and the 4.2/4.3/4.4 analysis need the text modified to make it clear that there are, based on the analysis carried out thus far, multiple possibilities for the time offset noted in the errors (e.g. modifying their text to remove phrases like "likely" with regards to the cause of the errors, using instead, "possibly, subject to further investigation"). Alternatively, they should redo the analysis (if not for the whole period, for a shorter test period in order to better make their case) with the new version of CMAQ mentioned, and/or modify Chimere in order to test out the hypothesis that the PBL parameterization.

** The authors mention that the European measurement data themselves may have timing errors, referencing an earlier paper (lines 480 to 481), as a possible explanation for the grouping by country of the model errors which can be seen in Figure 13(a). Explain why timing issues in the observations doesn't invalidate or weaken the analyses in this section. This is a serious issue – the authors have made an effort to ensure that

the model values are specific and representative of specific hours – but if the measurement data time stamps can't be trusted, that adds additional uncertainty to the analysis and its conclusions. For example, if the time series in the observations are all off by 2 hours in some parts of the domain, would that not show up as an "error" signal with a strong diurnal variation? I also wonder about time zone errors in reporting (local time, standard time, daylight savings time, etc.).

Line 480: harmonization misspelled.

Section 4.4:

**Line 518-519: The authors need to explain why the data were not limited to co-located measurements of air pollutants and meteorology, in this section's analysis. The authors are averaging different sets of locations between air pollutants and meteorology – and making the assumption that the resulting averages will both represent the same region to the same degree. Worst case scenario, for example: air pollution measurements in valley bottoms and meteorological measurements on mountain-tops.

Line 525: The authors use the phrase "overwhelming daily memory of the error" – please explain. The term implies that the results of day 1 are in some way affecting day 2 – but the analysis thus far has dealt with periodicity, which does not necessarily imply a causal link between one day and the next.

Line 531-532: At what point would these collinearities invalidate the analysis?

Line 538-540: The authors again attribute the errors to PBL issues, without sufficient cause. All that can safely be concluded by the analysis thus far is that the errors have a significant diurnal component, and that the PBL parameterization is one of many possible contributors to that diurnal component.

Lines 544-546 and lines 554-558: Filtering out the shorter time-scales removes much of the collinearity of the fields examined – which says that those fields are interrelated for the shorter time scales more than the longer time scales. This does not imply that

errors with CMAQ are associated with longer time scales, however (line 555) – the reasoning behind this statement made by the authors is not clear.

Line 566: Anything with a diurnally repeating signal in the error could be the root cause here.

Conclusions: The authors state (line 577) that other methods of analysis do not help target the causes of model error. I'm not completely sure that the work they have presented does a better job – the main contribution (and a very valuable one!) is that the time scales of the errors may be identified with the methodologies the authors present. However, the causes of those errors remain speculation at this point in time, since a causal link between specific processes and the time scale analysis has not been demonstrated. Nevertheless, the analysis methods introduced here are an excellent route by which to reach that endpoint, and worthy of publication in ACP on that basis alone. Coupled with process analysis (their suggestion for further work), it may be possible to tease out the causes of the errors: e.g. output the change in concentration due to each operator at each time step, and analyzing the rates of change of the different components.

Interactive comment on Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2017-257, 2017.