Atmos. Chem. Phys. Discuss., 15, C1162–C1169, 2015
www.atmos-chem-phys-discuss.net/15/C1162/2015/

Atmospheric Chemistry and Physics
Discussions
Open Access

Interactive comment on "Evaluation of the MACC operational forecast system – potential and challenges of global near-real-time modelling with respect to reactive gases in the troposphere" *by* A. Wagner et al.

Anonymous Referee #1

Received and published: 28 March 2015

GENERAL COMMENTS

The authors provide a validation study of reactive gases modelled by the MACC system using a variety of data sources. In particular they validate O3 using GAW and EMEP data, CO using GAW station data and MOPITT retrievals, and NO2 using SCIAMACHY and GOME-2 retrievals. Given the extensive current and expected future use of MACC/CAMS products, this is a highly welcome and important contribution. With MACC soon becoming operational as CAMS this is a very much needed study at this

point and it fits reasonably well within the scope of ACP although GMD probably would be more appropriate. The manuscript in general is structured consistently and well written, although some comprehensive editing by a native English speaker would be beneficial.

My main concern stems from the validation of modelled NO2 columns using satellite data. Satellite data of NO2 are extremely useful for comparing overall spatial patterns and providing an approximate qualitative assessment of the data. When using long-term averages they can even be used in a somewhat quantitative fashion to some extent. However, the uncertainty in the NO2 retrievals (both in terms of systematic biases and random errors!) themselves is too high to allow a full quantitative validation of model results. You are essentially comparing two similarly uncertain parameters with each other! Furthermore, NO2 is primarily relevant close to the surface and within the PBL (and the NO2 output from MACC/CAMS will be primarily used for such applications), whereas satellite-based validation of NO2 can only be carried out for tropospheric columns (and in addition the satellite instruments tend to be least sensitive near the surface!). It is thus impossible to draw robust quantitative conclusions from this, particularly for hourly/daily sampling and at the individual grid cell level (and this is very important for a full validation of the model results). A comprehensive validation of modelled NO2 with satellite data alone is not sufficient to draw accurate conclusions about the model performance. This is particularly relevant for validating the results from such a highly visible, high-profile, and heavily funded project as MACC/CAMS, whose model output will be used operationally for a wide variety of applications worldwide in future. As such, the validation methodology should be as robust as possible.

Therefore, in addition to the comparison against satellite data provided in the current manuscript, the authors really need to perform a solid quantitative validation of modelled surface NO2 against reliable station observations (and possibly a validation of modelled NO2 columns against ground-based MAX-DOAS data) before this manuscript can be published.

Me second concern is related to the extensive use of the MNMB in this study. This is a highly non-standard statistical metric and is not readily understandable by a general audience. It is entirely unclear why the MNMB is arbitrarily multiplied by a factor of 2, for example, and how the percentage values of a bounded index should be interpreted. Personally I think it would be preferable to stick to commonly used metrics such as for example the classic combination of mean bias and the standard deviation of the differences (representing systematic and random error, respectively) with RMSE as a measure of total error, possibly MAE, etc. I do realise that MNMB seems to have been adopted by the MACC validation team and is being used throughout several MACC-related papers in order to make statistics between species comparable. However, the vast majority of readers of MACC-related papers will not be familiar with this metric and will not know about its properties. If the authors insist on using this metric as intensively as in the given manuscript, I think they need to much better justify the use of such a non-standard validation metric and further should provide a detailed background regarding its statistical properties as compared to standard metrics.

SPECIFIC COMMENTS

P6279 L1: What about MACC-III? Wouldn't it be more sensible to call it something along the lines of a "series of MACC projects" or similar?

P6280-6281: This reads more like a textbook section on atmospheric chemistry than an introduction to a validation paper. Please be concise and focus on what is relevant for this study. It would also be useful here to discuss why we actually care about these gases and why we model them, i.e. what are some potential health effects or other impacts of these gases. See the submitted MACC validation paper by Eskes et al. (2015) in GMDD for an example on this.

P6281-6282 etc: Sometimes you talk about MACC/MACC-II, sometimes about MACC-II and sometimes about MACC. Please be consistent. I recommend introducing the series of MACC projects (including MACC-III) once in the beginning and then referring

to it simply as MACC in the remainder of the manuscript. Again, take a look at the submitted MACC validation paper by Eskes et al. (2015) in GMDD for finding out how to do this in a better way.

P6281 L19-20: This is worded a bit strangely. It is not the series of MACC projects that form the basis of CAMS, but rather the work that has been carried as part of MACC represents the preparatory activities that in the end are supposed to result in the operational CAMS.

P6281 L26: Are there more recent references on how data assimilation is being carried out within MACC/CAMS? If yes, cite them here. Maybe Inness et al. 2013 or similar?

P6282 L17-21: It is not clear how the availability of independent observations limits the period of this study to 2009-2012. For sure all the satellite datasets (MOPITT, SCIAMACHY, GOME-2) were available many years before 2009 and with exception of SCIAMACHY also continued after 2012. Surely GAW and EMEP data were available outside this period as well? Be precise about what is the limiting factor here.

P6282 L25: are -> is

P6282 L28: "encloses"? Better write something like "provides" or "contains"

P6283 L19: "MACC_osuite". Can you provide an explanation for this rather odd technical acronym?

P6283 L24: Be specific about the spatial resolution of the model. Is it 100 km x 100 km or irregular (and/or give it in degrees lat/lon)?

P6284 L9: What do you mean by "go back"? Do you mean the emissions are taken from or based on the RETRO-REAS inventory? Also, how exactly were the emissions merged?

P6284: Give more information about the spatial resolution of the various emission inventories

P6284 L26: "lists up" -> "lists"

P6285 L20: Has this been studied (if yes, provide results) or is this just an assumption?

P6286 L5: WMO 2010 is not included in the list of references

P6286 L6: Why specify "tropospheric" here? These are surface observations, right?

P6286 L24: Why didn't you use vertical interpolation between the two closest model levels. Discuss why the resulting error is negligible (or why not).

P6289 L1: The labels "Fires-Alaska" and "Fires-Siberia" look awkward compared to the other regions. Clarify why these specifically refer to fires and that they are only used for CO validation with MOPITT. Also in some of the Figures these labels are not used consistently. Please fix.

P6289 L11: "UV-VIS". Also I would recommend either writing "UV-VIS and NIR" or "ultraviolet-visible and near-infrared" and not mixed.

P6289: This section requires a discussion about the expected uncertainty of the satellite-based NO2 retrievals. Also, what is a reasonable minimum threshold of detection for the tropospheric NO2 column derived from SCIAMACHY and GOME-2?

P6290 L7: "linearly in time"

P6291 L8: Why does the MNMB used here range from -2 to 2 rather than -1 to 1? Why is this metric multiplied by 2? When using this metric in percent, as the authors do in this study you get a bounded range of -200% to 200%. How should this be interpreted? Please provide additional detail about the statistical properties of this non-standard evaluation metric.

P6291 L20: Keep the section headers consistent. Either spell out the species or not, but do not mix.

P6291 L23: It shows not one but two maps

P6291 L23: Figure 11? Figures 2-10 have not even been discussed yet. This also applies throughout the rest of the paper. Renumber Figures and Tables based on when they are introduced in the manuscript

P6292 L4: "far north" -> Better write "high latitudes in the northern hemisphere" or something similar to be specific

P6293 L6: better write "norther hemisphere winter months"

P6293 L24: This is not clear from Figure 14. It seems to show negative values of around -30% for Dec 2010?

P6293 L25-27: Can you provide an explanation for why Dec 2012 behaves so differently?

P6293 L27: "diurnal O3 cycle". This is misleading - Figure 15 does not really analyse the diurnal cycle but rather simply differentiates the result by day and night. Consider rewording this.

P6294 L21-23: Why do you need to refer to RMSEs and correlation coefficients in this sentence, when you are just talking about MNMBs? Please revise.

P6294 L24 "northern hemisphere"

P6295 L1: These correlation coefficients are indeed extremely low. A Pearson correlation coefficient of what is on average about 0.3 (Fig 2) translates to an Rˆ2 of 0.09! And this is even for monthly averages and not hourly/daily observations - so the random error should already be reduced to a large extent. If a model can explain less than 10% of the variability in monthly averages, I think quite a bit of explanation about possible reasons for the poor performance is necessary. Please add a discussion on this here.

P6295 L3: How was the subset of stations in Figure 3 selected? Were only those stations selected at which the model performed well, or was some other selection process used? Please add information about this in the text.

P6295 L25 to P6296 L10: This section discusses solely differences between MOPITT and IASI but not the relevance of these differences with respect to the model. Please revise to better indicate how these differences affect the model performances? Is it due to assimilation of IASI CO products in the model?

P6296 L24: Be careful about interpreting too much into satellite-based NO2 columns over the open oceans. The NO2 levels there tend to be below the detection limits of the instruments and the patterns observed there often represent no true geophysical signal.

P6298: Please clearly distinguish here between CO and NO2 here. These are intermixed in the discussion making it difficult to follow.

P6299: This section also requires a brief discussion of the potential uncertainties introduced by transitioning from SCIAMACHY to GOME-2 in 2012 and how it affects the validation of NO2.

P6300 L5: Again, you are not really studying/validating the diurnal cycle. Please reword.

P6316: The combined label/region field is a bit confusing. Do only the GAW stations have a label whereas the EMEP stations have a region acronym? For clarity please highlight this in the caption and list the region acronyms.

P6319: This table has unrealistically high number of significant digits. Please modify.

P6322: The panels in this plot are missing labels as a) b) c), yet the caption refers to them. Also, why does the caption only refer to a) and b) instead of all three. Please be consistent. Also the panels are very small, such that the legend is not readable.

P6324: The legend here does not list the region names as "Fires-Alaska" and "Fires-Siberia", as they were introduced previously. Please decide on a label for these regions and then stick to it consistently in text and Figures.

P6325: Same in this Figure.

P6326: It would be helpful to use different symbols/colours for SCIAMACHY and GOME-2 in this Figure.

P6328: Same in this Figure.

P6329: The caption says "daily" but the Figure shows monthly averages. Please correct.

P6331: This Figure has an unclear colour scale, making the interpretation of MNMBs close to zero challenging. Plots with divergent colour scale such as this should ideally have only one colour gradient for positive and negative values, respectively, with a neutral colour (white or grey) in between. I recommend shades of red for positive values and shades of blue for negative values with white or grey symmetrically around zero.

P6332: These plots are extremely busy and the legend is unreadable. Please consider ways of reducing the overplotting to increase the visual impact of the Figure. Also, once again, please consistently format and label the panels. Why does subplot a) consist of two panels and subplot b) of one panel. Why not have 3 separate subplots?

P6333: Describe either in the caption or in the text how this seemingly random subset of stations was selected.

Interactive comment on Atmos. Chem. Phys. Discuss., 15, 6277, 2015.