Reply to anonymous referee #1

*Summary: This paper presents a thorough overview of the validation activities relating to the MACC global composition forecast model. The program of activities developed under MACC/MACC-II will form the basis of validation for the new Copernicus Atmosphere Service (CAMS) and hence represents an important framework for future European activities in atmospheric monitoring. The paper is well-structured and written to a high standard, using precise language and grammatically accurate English. Only a small number of equations appear in the paper, but these are accurately transcribed, with all symbols described in the text. The large number of references provided is consistent with the wide scope of the paper and reliance on modelling and datasets produced by other workers.*

We thank the referee for his/her appreciation of our paper!

*Specific comments*

*Section 5: The use of the mean-field metrics 'modified normalized mean bias' and 'fractional gross error' is fully justified and provides a consistent reference scale which allows a forecast skill for a wide range of species to be meaningfully compared. In the course of time it is likely that the model resolution will increase. However with increasing model resolution mean-field metrics are susceptible to the 'double penalty' problem. Other metrics, for example relating to model skill in predicting the magnitude of elevated pollutant levels, will typically improve with increasing model resolution. It is useful therefore to also include these types of metrics in order to give a more balanced picture of model performance. This is not necessary for the present paper but is suggested as a comment for future evolutions of the work.*

We are aware of the "double penalty" issue related to the metrics used. This is something that will be picked up in the future by the validation subproject in the Copernicus Atmosphere Monitoring Service, where the introduction of new metrics is one of the work packages, and where steps in increasing the resolution of the model are foreseen. Our paper describes the current status of the validation work, and the MNMB and FGE are the metrics we have been working with in the past three years. We are not in favour of including explicit formulas for metrics we have not been using so far.

The following sentence is included in the modified paper at the end of section 5 to discuss this point:
"In the coming years, the resolution of the CAMS system is expected to increase to below one degree. The MNMB and FGE scores scores in this case become less appropriate to monitor the model improvements. Small filaments of polluted air may be slightly displaced, and the mean norms will lead to a "double penalty" for the higher resolution model, even though the simulated peak values are more realistic. The introduction of new metrics is needed for a more appropriate evaluation of the improvements, and this is one of the tasks of the future validation subproject of CAMS."

*Section 6.2 and elsewhere: Please clarify whether the results in the paper relating to C-IFS refer to the free-running model or with data assimilation.*

The C-IFS modelling system is used with, and without data assimilation, and results are presented for both. The main product of MACC is the o-suite analysis and forecast assimilation system.

This sentence has been added to section 6.2:
"We remind the reader that "o-suite" always refers to the IFS-based analysis and forecast system including the assimilation of the full suite of aerosol, chemical and meteorological observations."

*Section 7. In future it would be useful to also include ozone measurements from suitable rural/remote surface air quality measurement sites.*

We fully agree, and in fact these observations are used.

The following sentence is added at the end of section 7:
"Apart from the GAW and ESRL in-situ observations, also measurements from rural and remote surface air quality measurement sites are considered. The sites have to be carefully selected, because they should be representative for a larger area of the size of the model resolution. Furthermore, validated datasets are typically only available after a few years and only unvalidated data can be used for the near-real time evaluations. In particular, observations from the European Monitoring and Evaluation Programme (EMEP, http://www.emep.int), and the European air quality database "AIRBASE" (http://www.eea.europa.eu/themes/air/air-quality/map/airbase) are used to evaluate the reanalysis results. Also evaluations based on the USA "AirNow" observations (http://www.airnow.gov) are in preparation."

*Section 8, page 1135 line 20 onwards: The large negative bias in the model extinctions is presumable partly due to the difficulty in estimating the source strength of the biomass burning emissions.*

We agree, and added a line to section 8:
"Many aspects influence the quantitative comparison, including uncertainties in the source strenght (fire radiative power observation and aerosol mass produced) uncertainties in the transport over several days, removal processes, resolution of the model and local representativity issues. Part of these modelling errors may have been corrected by the assimilation of the MODIS observations."

*Section 9.2, 17: The under-estimation of NO2 columns may be partly due to underestimates in the emissions, but is probably also partly related to the 'low' model resolution, which unavoidably spreads emissions over a minimum of a grid box.*

We agree and added a line to section 9.2:
" The relatively low model resolution will lead to an underestimate of strong localised emission sources."

*Section 9.3, page 1140, line 1,2: It would be helpful if these correlation coefficients could also be added to the caption of Figure 6.*

This is added to the caption in the new manuscript.

*Section 9.5, line 23 onwards: The bias figures quoted in the piece of text do not seem consistent with Figure 7. Also, does the term 'bias' imply the 'Modified Normalised Mean Bias' used and defined earlier or is this a different 'bias'? If it is the MNMB then units of % seem incorrect. Please review this text and figure carefully and ensure consistency / clarity.*

The two panels of figure 7 will be replaced in the revised manuscript and will contain the MNMB.

*A confusion may arise because in the text of Section 9.5 (line 23 onwards) the described bias figures use AERONET 2.0 level as a reference, while the bias is plotted using AERONET Level 1.5 as reference in Figure 7. Given that AERONET Level 1.5 has a bias of + 20% over AERONET level 2.0, this explains the apparent inconsistency. Even though it is mentioned in the caption of Fig.7, it is confusing for the readers. Note also that the end of section 9.5 discusses dust, while the first part of the section deals with all aerosol types.*

Because only AERONET 1.5 is available in NRT, we decided to keep the data from this version for figure 7 (which has been updated). The difference between the two versions is an important feature one should be aware of when using the NRT data, so we think it is useful to have this discussed. The text has been adjusted to make the difference clearer.

*Section 9, page 1141, line 20: As for ozone, for future work please consider using the surface air quality networks for deriving aerosol composition measurements.*

We have added a line to section 7:
"Apart from ozone, also the aerosol composition measurements from these networks will be considered, as well as other compounds like CO and NO2."

*Minor typographic corrections:*
*page 1131, line 6: suggest replacing 'in-line' with 'on-line'.*
*line 22: 'longterm' –> 'long-term'*
*page 1144, line 24 'Ceilometer' –>'ceilometer'*

Corrections included in revised manuscipt.

Response to anonymous referee #2

*The manuscript by Eskes et al. provides an outline of the global validation component of the European MACC project. The MACC project seeks to establish and evaluate atmospheric composition modelling tools. Basic model configurations are described, an overview of metrics is given, and the recent performance of operational part of the system is evaluated. Particular weight is attached to the importance of having easily accessible metrics of model performance so that end-users of the MACC (and subsequently CAMS) data can assess the quality of that data.*
*The format and content of the manuscript are appropriate, and when published the paper will provide a very useful reference point for users of MACC / CAMS data. I would recommend addressing the general and specific comments listed below, prior to publication, however.*

We thank the reviewer for her/his positive judgement, and for the useful detailed comments made.

*I would also recommend a detailed reading by a native English speaker to pick up on a few instances of slightly awkward phrases.*

We have carefully gone through the whole document and made small modifications to the text. In particular, we have included the textual suggestions from all three referees.

*General comments*

*Many acronyms are not defined, e.g. ECMWF, IFS, FTIR, SDs*

We have checked all acronyms and expanded them in the revised manuscript.

*I would like to see more development / justification / discussion of the metrics used – this would seem to be a focal point of the manuscript. The formulae behind the calculation of the metrics is presented, but there is very little justification given for why those metrics are suited. What do they tell us, what are their limitations?*

We assume that the reader is familiar with the basic concepts of bias, rms and correlation. The MNMB and FGE are are special forms of bias and rms type of measures, which indeed need some explanation. The main reasons for adopting these metrics is provided in the text (p 1128, l22 and following paragraph). The limit values of the expressions were also discussed.

In the new manuscript this paragraph is extended:
"The MNMB and FGE are alternatives for the more commonly used mean bias and the root-mean-square error respectively. The normalised approach in the MNMB and FGE provides errors in a relative sense, which is easier to comprehend by users not very familiar with the concentration ranges and their units. The fractional gross error is a linear measure, and has the advantage compared to the more common root-mean-square measure that it is not dominated by outliers. Both MNMB and FGE are defined relative to the mean of the observation and the model value, $( f_i + o_i )/2$, which improves over expressions where the observation alone is used as reference. For

instance, surface ozone observations do in practice give readings equal to 0, which causes the division by $o_i$ to become infinity."

*Furthermore, the presented model / obs comparisons summaries don't appear to make full use of the metrics outlined. In fact in many parts of the manuscript a qualitative description of model performance is given instead of the metrics. Why describe the metrics if they're not going to be used in the evaluation summaries?*

We can sympathise with this remark, but do not fully agree. The metrics are used extensively in the MACC validation activities. One example of this is is Fig. 4. The present paper shows just a couple of examples from the validation work, and more examples can be found in the validation reports (see references). In practice, validation should always start by making more simple direct comparisons in the form of maps and time series. The metrics quantify the comparisons, but to interpret the quantitative results it is crucial to display the modelled and measured values. This is also mentioned on p1127 and is one of the scoring recommendations, so it is already covered by the original manuscript.

In the revised manuscript figure 7 has been replaced. It now displays both correlation and MNMB.

*Specific comments*

*Page 1121, line 18: 'Haiden et al., 2014' does not appear in Reference list*

Reference has been added to the revised manuscript.

*Page 1122, line 7: 'Copernicus' (unavoidably) refers to one of two different entities in the context of the manuscript. I would suggest adding 'EGU' before 'Copernicus' in this instance*

Done.

*Page 1122, line 20: Author name not given in citation*

"Eskes" has been added.

*Page 1123, lines 12-14: Better word for 'aspects' might be 'species' or 'quantities'? Otherwise the meaning of the sentence is not clear*

Agree. Text now mentions "species".

*Page 1126, line 26: '...this report...'. Not clear which report is being referred to – the present manuscript, or the 'living document on the evaluation methodology'?*

It is the latter. Text has been modified to make this explicit.

*Page 1128, equations 1-3: 'N' not defined*

Definition of N has been added to the revised manuscript.

*Page 1130, paragraph beginning on line 18: this paragraph seems a bit out of place – better placed in Section 9.6?*

We agree. The paragraph has been moved to 9.6. The introductory line on the three models in the first paragraph of sec 9.6 has been removed.

*Page 1131, line 14: in the model simulations without data assimilation, where are the initial conditions derived from?*

Initial conditions are taken from the run of the previous day. The total length of the simulations is such that the impact of the original initial state is negligible.

*Page 1135, line 13: give value of 'low bias'*

The sentence has been modified:
" The correct timing of the dust event in the MACC o- suite is further confirmed by the time series at the available AERONET sites (black dots), although the modelled optical depth has a moderate low bias of about 0.1 compared to the observations. "

*Page 1140, Section 9.4: Is there any value in presenting a figure (time-series) of HCHO? As per the other quantities?*

We considered this. Such a decision is a matter of finding a good balance between the amount of detail and keeping the length of the paper within reasonable limits. There are a few motivations not to include it: 1. Figure 5 already shows an example of the use of SCIAMACHY/GOME-2 for validation, and 2. there are no HCHO observations assimilated. The figures in the paper serve as examples for the range of activities in the validation subproject. Much more detail, including HCHO figures, can be found in the validation reports.

*Page 1140, Section 9.5: (Aerosol evaluation) gives percentage biases – are these based on the MNMB? It would be preferable to retain the same metric for all modelled quantities. If percentages are considered appropriate please list formula for reference (as per other metrics)*

In the new manuscript figure 7 will be replaced, and the lower panel will show the MNMB. Aso the text will be adjusted accordingly.

*Page 1143: line 7: '>5%' should be '<5%'? Or perhaps better to present MNMB rather than percentage?*

'>5%' is replaced by '<5%'.

*Page 1159, Caption for Table 1: 'quantities' instead of 'aspects'?*

Has been replaced.

*Page 1161, Figure 2: Extra AERONET sites marked on maps but not referred to?*

The caption of the figure includes the names of the stations corresponding to the dots. Results are shown for two stations. We feel this is clear enough. We could have shown results for the other stations, but this would not have changed the conclusions.

*Page 1161, Caption for Figure 2: Date given in caption different to that on maps (22 June vs 25 June)*

Caption is adjusted to 22 June.

*Page 1162, Figure 3: MACC o-suite values are daily means?*

No, these are 12 utc snapshots. Caption text has been adjusted accordingly.

Response to Anonymous Referee #3

*This paper provides an extensive overview of the MACC global forecast system. The main aim of the paper is to document the data that is assimilated into the MACC system and to provide a detailed description of the methods and the data used for the validation of the modeling suite (VAL). The paper will be of particular use to the community that will be using the MACC data analysis and forecast system and those who are involved in the development of the MACC system. However, I feel it will also be of interest to wider modeling community. I believe the content of the paper is ideally placed for publication in GMD and recommend publication after the authors have addressed the relatively minor corrections below.*

We would like to thank the reviewer for the kind words, positive recommendation, and the very careful reading of the manuscript which has led to many small improvements.

*Main comments:*
*There is a large amount of data being assimilated into the MACC system and being used for validation so I believe the paper would benefit from a table that gives the reader an overview of the measurement uncertainties and an idea of the temporal and spatial frequency of the observations. This has been discussed during some sections in reference to specific observational datasets; however there are many observations whose uncertainties are not discussed specifically.*

We have discussed this among the authors of the paper and with the modellers. We agree that especially the information on the uncertainty is useful. A table 2 has been added to the revised manuscript, section 7, listing the measurements, the spatial and temporal characteristics and the estimated uncertainty. However, to our opinion it is outside of the scope of the present paper to list also the datasets which are assimilated. We will refer to the relevant papers for this.

*My other main comment is that sometimes the language/sentence structure is a little mixed up and that the paper would benefit from a careful read through to check the English and grammar used. I have noted a few of these occasions below.*

We have carefully gone through the whole document and made small modifications to the text. In particular, we have included the textual suggestions from all three referees, and in particular the ones listed below.

*Minor comments:*

*Pg 1121, L22-28: Refer to the section numbers.*

References to the section numbers have been included.

*Pg 1123, L12-25: This paragraph is hard to follow. Please restructure and shorten to make a bit clearer.*

The first sentence has been reformulated, and a short list has been introduced. The text in the revised manuscript now reads:

For a good understanding of the quality of the MACC system it is important to consider which species in the global assimilation system are constrained by the observations, and which species are covered by the validation datasets used. This is summarized in Table 1. The MACC aerosol and reactive gas models contain on the order of 100 species with global coverage and ranging from the surface into the mesosphere. Clearly, only a small fraction of this is observed and constrained by the available observations.

- Assimilation: The MACC assimilation is focusing on aerosol optical depth (AOD), ozone, CO, NO2 and SO2. Note that the species are treated in a univariate way and correlations in background errors of different species are neglected (Inness et al., 2014). An analysis update of one trace gas will nevertheless influence others through the chemical reactions.
- Validation: The validation is also constrained by the limited amount of trace gas and aerosol properties for which validation data is available. Furthermore, validation is limited by the amount of external data that is available in real time or at least within a few weeks after measurement, and with a reasonable global coverage.

*Pg 1124, L7-25: This paragraph is a bit hard to follow. It may be better condensed as a table?*

We have turned the paragraph into a list in the revised manuscript, which improves the readability.

*Pg 1129, L7: Add a reference for MOCAGE*

This is dealt with in the revised manuscript by referring to the paper by Flemming where a list of references is provided for the three models. Introducing all these models is beyond the scope of our paper.

*Pg 1135, L8: 'out through the Sahel,' – Check grammar.*

Line replaced by:

A first example of a case study is shown in Fig. 2. In June 2014 a huge desert dust plume occurred that originated in the Sahara and traveled more than 6000 km over the Sahel and the North Atlantic, impacting the Amazon and the Caribbean.

*Pg 1135, L10-11: Is it surprising that the MACC system can capture the MODIS AOD when it assimilates MODIS AOD? Maybe mention here that the comparison to MODIS is not totally independent whereas the surface sites are.*

There is a subtle difference, because figure 2 contains MODIS DeepBlue data over land, which provides observations over bright land surfaces. The DeepBlue data is not assimilated in this version of the C-IFS system. So, in fact it is in part an independent check.

The following line is added to the text of the manuscript:
"Note that the MODIS DeepBlue data, which is providing aerosol observations over bright land surfaces, is used in the figure but not in the assimilation."

*Pg 1135, L24: Add reference for uncertainty of ceilometer. Figure 3: Make labels clearer.*

A reference has been added, and the figure is improved in the new manuscript.

*Pg 1136, L4-6: What do you mean by 'representativity' issues? Do you mean concentration bias or location bias? What improvements are planned?*

This line is replaced by the following text:

"Many aspects influence the quantitative comparison, including uncertainties in the source strenght (fire radiative power observation and aerosol mass produced) uncertainties in the transport over several days, removal processes, resolution of the model and local representativity issues. Part of these modelling errors may have been corrected by the assimilation of the MODIS observations."

*Pg 1136, L19: Add reference to the other POLMIP studies – Arnold et al (2014) and Monks et al., (2015).*

We feel that three references for POLMIP is out of balance. MACC has made significant contributions to the paper by Emmons.

*Pg 1138, L4: Do you mean the number of observations being assimilated are more sparse in the SH so the model bias is larger or do you mean that the model has undergone little previous evaluation and therefore model improvements that benefit the SH?*

We mean that there are only few GAW observations available for the evaluation.

The line is modified into:
" The model is scarcely evaluated by the GAW network over the Southern Hemisphere."

*Figure 4: You discuss comparisons to the other model simulations without data assimilation but they are not included in Fig 4. It would be interesting to see these model runs also.*

These figures are available in the MACC validation reports. For the current paper we chose to show the regional dependence of the o-suite as example. Adding more curves would make the figure very crowded and we do not think this is a good idea.

*Pg 1138, L14-17: You say 'The comparisons with SCIAMACHY/GOME-2 show that spatial distributions of tropospheric NO2 columns are well reproduced by all three NRT model runs throughout all seasons, indicating that emission patterns and NOx photo- chemistry are generally well represented.' I don't see this from the figures included (Fig 5). I'd say the models capture the seasonality, however, I wouldn't say*

*they capture the emission patterns as SCIAMACHY indicates larger NO2 over Asia compared to Europe whereas the model indicates larger NO2 over Europe. Are you referring to comparisons that aren't included or from one of the other scientific papers in the special issues? If so, say 'not shown' or reference the paper.*

The text has been reformulated:
"Comparisons to SCIAMACHY/GOME-2 monthly mean tropospheric NO2 columns on a global map (Eskes et al., 2014a) shows that spatial distributions of tropospheric NO2 columns are well reproduced by all three NRT model runs throughout all seasons, indicating that emission patterns and NOx photochemistry are generally well represented. A general feature is the underestimation of NO2 columns over the continents in general and particularly in China (the latter is also evident from Fig. 5), ..."

*Pg1139, L21-28: You say there is an improvement when assimilating data (o-suite) . However when you look at Fig. 6, it seems the C-IFS run does a better job at capturing CO. The correlation coefficients are also better for this run than the o-suite run (pg1140, L1-2). Can you please check this paragraph and clarify why you think the o-suite run gives a better performance.*

The improvement occurs for the o-suite, based on IFS-MOZART, and the free running IFS-MOZART. The general statements are of course not based on one location, but summarise mean results from the entire set of GAW stations.

C-IFS is a new and entirely different model which for the period shown in the plot was only operated in free running mode. The figure shows one example where the C-IFS free run improves the correlation as compared to IFS-MOZART free running configuration.

*Figure 7: Say what data the correlations coefficients have been calculated for. Is it daily/hourly data within each month?*

The correlation coefficients are based on consistent daily mean values, from all stations and when observations are available.
This explanation is added to the text of section 9.5.

*Figure 8: Make plot lines and text thicker to ensure quality of figures when printing.*

New figures will be produced for the revised paper.

*Pg 1141, L27: Check sentence structure: MACC o-suite captures almost all dust out-breaks tracking fairly well their spatiotemporal evolution over the North Atlantic and the Mediterranean.*

Sentence has been somewhat reformulated

*Figure 9: Caption – Define SD.*

Replaced by "standard deviation".

*Pg 1142, L23: Check sentence grammar: 'The impact of data assimilation at other locations is confirmed'. Do you mean 'the impact of data assimilation at other locations can be seen'?*

Replaced by "can also be seen".

*Pg 1144, L13: Check sentence: 'More research and technical work is needed to use e.g. the climatological aerosol composition and variation as used for AeroCom model'*

Reformulated: "Additional research will be based on the climatological aerosol composition and variation (as used for AeroCom model evaluations) to obtain relevant information on the quality of the IFS forecast system."

*Technical corrections:*
*Pg 1120, L1: remove 'even'. Pg 1120, L14: remove 'to' before respond.*

Both done.

*Pg 1120, L: Define IFS properly in the following sentence - 'the numerical weather prediction forecasting system of ECMWF (IFS)'.*

Added " Integrated Forecasting System"

*Pg 1123, L16: ranging -> range*

Done

*Pg 1128, L8&9: in case -> in the case*

Done (2x)

*Pg 1129, L5: insert comma after MACC.*

Done

*Pg 1130, L25: Insert 'of' after 'all'*

Done

*Pg 1131, L21: profiles -> Profiles*

Done

*Pg 1132, L 2-3: Check sentence structure and grammar.*

This part was replaced by: "The validation activities in GEMS and MACC have been using ozone and CO from MOZAIC and IAGOS for ten years. Both the take off and landing profiles and the UTLS cruise part of the flights at northern mid-latitudes have

been compared to the different model runs on a regular basis. Special events such as the summer 2003 heat wave over Europe (Ordonez et al., 2010) and summer 2004 Canadian boreal forest fires (Elguindi et al., 2010) have been studied."

*Pg 1132, 9: The second and final -> the second is the final*

Not changed.

*Pg 1132, L23: Assimilation O3 results –> O3 results*

Done

*Pg 1135, L19: I don't see the need for starting a new paragraph.*

Done

*Pg 1136, L8: which -> that*

Done

*Pg 1136, L20: Remove 'e.g'*

Done

*Pg 1138, L3: Insert comma after stations.*

Done

*Pg1140, L21: , see Fig. 7 -> (see Fig. 7)*

Done

*Pg 1141, L15: Aeronet -> AERONET*

Done

*Pg 1141, L18: , see Fig. 8 -> (see Fig. 8)*

Done

*Pg 1142, L4: and -> an*

Done

*Pg 1142, L6: Add units (0.08 to 0.24)*

Optical depth is dimensionless.

*Pg 1142, L26: show always good agreement -> show good agreement Pg 1145, L10: remove 'model'.*

Done