**Authors' comments in reply to the anonymous referee for "Evaluation of ACCMIP ozone simulations using a multi-constituent chemical reanalysis" by K. Miyazaki and K. Bowman**

We want to thank the referee for the helpful comments. We have revised the manuscript according to the comments, and hope that the revised version is now suitable for publication. Below are the referee comments in italics, with our replies in normal font.

***Reply to Referee #1***

*General remarks:*

*The paper would benefit from a more stringent focus on the ozone sonde sampling biases and its impact on the evaluation because this is the actual novelty aspect of the paper. The ACCMIP models have been evaluated. So only the differences of the new evaluation approach with previous work is of interest. I would recommend to add to the title "- focus on ozone sonde sampling biases" or similar. The sampling biases should be mentioned and discussed in abstract and introduction more clearly.*

The title has been revised as follows:
"Evaluation of ACCMIP ozone simulations and ozonesonde sampling biases using a satellite-based multi-constituent chemical reanalysis."

The abstract and introduction have been modified to mention the sampling biases.

*To get a better understanding of the sampling biases, i.e. the difference in the mean over area averages using all grid points at regular intervals or only the stations locations at the time of the observations, it is recommended to show the sampling biases not only for the differences between ACCMIP models and the reanalysis but also for the Re-analysis and the model runs, including the control run itself. It would be interesting to see to what extent they differ as the reanalysis may also be effected by the "sampling biases" of the assimilated observations. A strong sampling biases for model result will help to convince modellers to use reanalysis data for model evaluation.*

Table 9 has been added to discuss the impacts of the sampling biases in the reanalysis and control run comparisons. These results are discussed in Section 5.1 as follows:
"Further, ozonesonde sampling bias is evaluated for the control run and reanalysis comparisons. As summarized in Table 9, at 500 hPa, there are large differences (> 30 %) between the two evaluations in

many regions, especially in the NH mid latitude regions in winter and in the tropics throughout the year, as also found in the ACCMIP models and reanalysis comparisons (Table 8). The analysis increments introduced by data assimilation vary with space and time, reflecting the changes in coverage and uncertainty of assimilated measurements as well as in model errors. Nevertheless, observational information was propagated globally and integrated with time through forecast steps during the data assimilation cycles. This is true for ozone because of its relatively long lifetime in the free troposphere. Therefore, the spatial distribution is well constrained by data assimilation, and we do not expect large variations in the reanalysis quality within each analysis region."

*The authors should aim to provide a better understanding of the reasons of the sampling biases. Do they come more from spatial heterogeneity or the variable temporal sampling. The latter can be estimated by comparing re-analysis means at a 2 hourly resolution or only at the ozone sonde observing times.*
*For the sake of consistency the quantification of the sampling biases should be carried out for one set of latitude bands in the same way as in the more regional analysis presented in section 5.*

Table 8 has been revised to discuss the influences of temporal and spatial sampling errors separately. The following sentences have been added in Section 5.1:
"Our analysis using monthly reanalysis fields sampled at the ozonesonde locations (brackets in Table 8) suggests a greater impact of the spatial sampling bias than the temporal sampling bias for the NH polar east in DJF."
"At 500 hPa over Canada, the relative importance of the spatial and temporal sampling biases varies with season: the spatial (temporal) sampling bias is dominant in DJF (JJA), whereas both of them are important in MAM."
"Over the Western Pacific and East Indian Ocean, the sampling bias is not reduced by using monthly mean reanalysis fields (sampled at the ozonesonde locations) in DJF and JJA. This suggests that ozone varies with time and space in a complex manner, and a dense (in both space and time) network would be required to capture the regional and seasonally representative model biases in this region."
"The temporal sampling bias mostly dominates the difference in the SH high latitudes in MAM and JJA, whereas the spatial sampling bias is also important in the SH mid latitudes in DJF and MAM."

Table 8 has been modified to describe the sampling biases for four latitude bands. The following sentences have been added in Section 5.1:
"Table 8 also shows the model evaluation results for four latitudinal bands at 500 hPa. The observations used are shown in bold in Table 2. The differences between the two evaluations are small in the NH extratropics (30-90N) in all seasons, because of the relatively large number of observations. There are

large differences in the tropics of both hemispheres: the ozonesonde network reveals a large negative sampling bias in the model evaluation in the NH tropics (Eq-30N) in SON (-9 % in the complete sampling and -16 % in the ozonesonde sampling) and in the SH tropics (30S-Eq) in MAM (-14 % and -21 %) and a large positive sampling bias in the NH tropics in JJA (-7 % and -3 %). Large sampling biases (> 60 %) also exist in the SH extratropics (90-30S) in DJF and MAM due to the sparse ozonesonde network."

*Section 5 "Impact of Sampling on model evaluation" discusses the regional biases and the general problem in a lot of detail but sections 4.2 and 4.3 discuss already the sampling biases for the latitude bands. I recommend moving the introduction of the sampling biases to an earlier section (2).*

The introduction and methodology of the sampling biases have been moved to Section 2.4 (Section title: Ozonesonde sampling bias estimation).

*The discussions section, in particular 6.3, does not discuss the direct results of the paper but gives an outlook on other potential aspect of the usefulness of the evaluation with chemical re-analyses. However, the positive impact on species not directly assimilated has not be demonstrated in the paper. Also, the four year comparison is not long enough to infer trends and longer re-analysis of atmospheric composition are likely to suffer from temporal artefacts because of the changing observational system. I would therefore not discuss in detail these aspects in the paper as there is not enough evidence given to support them.*

Because this is the first study to use chemical reanalysis for model evaluation, it is worthwhile discussing its possible application in future studies. The positive impacts on non-assimilated species have been discussed in our previous studies, and this is described in the revised manuscript as follows:
"Miyazaki et al. (2012b, 2015) demonstrated that the multiple-species assimilation results in a strong influence on both assimilated and non-assimilated species."

The limitation of the evaluations using the five-year (2005-2009) reanalysis is discussed as follows in Section 5.1 of the revised manuscript:
"The five-year reanalysis (2005-2009) may cause biases in the estimated model errors in the evaluation of the 2000 decade ACCMIP simulations that used decadal-averaged SST boundary conditions and biomass-burning emissions averaged over 1997--2006 (Lamarque et al., 2010). It may neglect the influences of interannual and decadal changes in both anthropogenic and biomass emissions and meteorology. Longer-term reanalysis and time-consistent validation are required to obtain more robust

error estimations.

To discuss remaining issues with a longer-term reanalysis, the following sentences have been added in Section 6.3:
"However, any discontinuities in the availability and coverage of the assimilated measurement will affect the quality of the reanalysis and estimated interannual variability, which limit the usability of a long term reanalysis for model evaluation, as discussed in Miyazaki et al (2015) for chemical reanalyses and in Thorne and Vose (2010) for climate reanalyses. This also requires a bias-correction procedure for each assimilated measurement, in order to improve the reanalysis quality (Inness et al, 2013)."

*The used ozone sondes observations need to be clearer identified and their sampling discussed. A table of the used ozone sondes, their sampling frequency and outage in the period and mean should be summarised in a table not only for the regional areas but also for the latitude bands. It should be made clear which stations are used for global/hemispheric stratification in Figures 1 to 7 and the more regional stratification Figure 9-10.*

Table 2 has been added.

*Specific remarks,*

*PL1: 5 Please "the" before instrument names*

Added.

*P1L5: Please add a sentence on the advantages of using a 3D re-analysis rather than ozone sondes for the model evaluation.*

The following sentence has been added:
"The reanalysis provides comprehensive information on the weakness of the models, whereas we consider that the spatial and temporal coverage of individual measurements, such as ozonesonde measurements, is insufficient to capture the temporally and spatially representative model bias."

*P1L6: Please ad here or at L 12 the problem of the ozone sampling biases*
*P1L12: Please state more clearly the differences in the evaluation results when using the re-analysis as complete field and on the the ozone sonde observation locations and times only.*

The following sentence has been added:
"The ozonesonde sampling bias in the evaluated model bias for the seasonal mean concentration is 40-50 % over the Western Pacific and East India and reaches 110 % over the equatorial Americas in the middle troposphere."

*P1L24: better "transport"*

Replaced.

*P2L8: Please add some references for these evaluation studies*

Added.

*P2L13: there is a "First" and a "Third" (L19) but I did not find a "Second"*

Corrected.

*P2L18: The sentence starting with "However, . . ." is a strong motivation for the paper. Please elaborate and also mention that the climatologies do no capture the temporal variability of the observed ozone.*

The sentences have been rewritten as follows:
"However, the climatological data does not provide information on the temporal variability of the observed ozone. In addition, the current ozonesonde network does not cover the entire globe and is not homogeneously distributed between the hemispheres, ocean and land, and urban and rural areas, and its sampling interval is typically a week or longer. Model errors are also expected to vary greatly in time and space at various scales."

*P2L30: Please consider citing overview papers such as Bocquet et al. (ACP 2015) or Sandu et al. (Atmosphere, 2011)*

Added.

*P4L4: Please comment how this is related to resolution of the evaluated ACCMIP models.*

The following sentence has been added:
"The horizontal model resolution is comparable to the resolution of ACCMIP models (ranging from 1.24° to 5°)."

*P4 L22: Please explain how the ensemble is constructed, i.e. what parameters are varied to get a different ensemble members in the EnKF. This information is important because you later use the ensemble spread partially as indicator of the analysis error.*

The following sentence has been added:
"The ensemble perturbations were introduced to all the state vector variables as described below."

*P4 L23: "satellite retrieval operator"? This implies radiances i.e. Level 1 were assimilated, which is perhaps not the case. Please clarify.*

The sentence has been rewritten as:
"and an operator that converts the model fields into retrieval space"

*P5L4: Please mention if it could be shown that the modulation of the lifetimes was an improvement.*

The following sentence has been added:
"Miyazaki et al. (2015) demonstrated that the Northern/Southern Hemisphere OH ratio became closer to an observational estimate of Patra et al (2014) due to the multiple-species assimilation."

*P5 L14: Please provide a table with the assimilated retrievals and additional information such as assimilated height range, temporal data coverage and an indication of observation errors statistics.*

Table 1 has been added.

*P5L21: Please elaborate on the period and the meteorological input for this time-slice setup. It is important to know what sort of realism can be expected from the simulation if they are compared against observations.*

The following sentences have been added:
"The number of years that the ACCMIP models simulated for the 2000 decadal simulation mostly varied between 4 and 12 years for each model. Each model simulation was averaged over the simulated years."

"Meteorological fields were obtained from analyses in CICERO-OsloCTM2 and from climate model fields in MOCAGE. UM-CAM and STOC-HadAM3 simulated meteorological and chemical fields, but chemistry did not affect climate. In all other models, simulated chemical fields were used in the radiation calculations and hence provide a forcing effect on the general circulation of the atmosphere. Lamarque et al. (2013) indicated that most models overestimate global annual precipitation and have a cold bias in the lower troposphere."

*P6L9: Please clarify which station were used for the comparison. The ones listed in table 4 ? If so mention it here. Provide information about station numbers and individual temporal coverage as this may vary greatly and contribute to the ozone sonde sampling bias.*

The sentences have been rewritten as:
"All available data from the WOUDC database are used for the evaluation of reanalysis data (Section 3), as listed in Table 2. For the evaluation of ACCMIP models and ozonesonde sampling biases (Section 4 and 5), we use the ozonesonde sampling based on the compilation by Tilmes et al. (2012), which is shown in bold in Table 2. Because there is no observation after 2003 in Scoresbysund, this location has been removed from the compilation in this study."

Table 2 has been added to summarize the ozonesonde measurements.

*P6 L17-23: This description of the model changes may better put in the model description section.*

Moved.

*P6L30: Please clarify what the differences in the assimilated observations are between this data set and the previous one.*

The following sentence has been added:
"MLS retrievals have been updated from v3.3 in Miyazaki et al. (2015) to v4.2 in this study."

*P7L10: Please clarify what temporal averaging the temporal correlation is based on (i.e. monthly means, annual means, instantaneous values etc.). It is good practise to de-seasonalize the time series to get a more meaningful information about the temporal correlation. On the other hand, 5 years might be too short to obtain a robust information about seasonality and year-to-year variability.*

The sentence has been rewritten as follow:
"The tropospheric concentrations show distinct seasonal and year-to-year variations, for which the temporal correlation based on the monthly and regional mean concentrations is increased by the data assimilation globally, except at high latitudes in the lower troposphere."

Because the seasonal pattern varies with year especially in the tropics, we did not apply de-seasonalization.

The limitations of the five-year reanalysis data are discussed in the revised manuscript as follow:
"The reanalysis can be extended to a longer-term validation that will provide more information on seasonality and year-to-year variability."

*The numbers in figure 1 indicate a very good reduction in biases but far less so for variability measures. (The reduction in RMSE seems dominated by the bias component and temporal correlation is less improved). This seems to contradict the theoretical basis of data assimilation, which is meant to reduce the error variance assuming bias free model. A further discussion would be helpful.*

We confirmed that both the bias and RMSE are largely reduced compared with assimilated measurements (e.g., TES) due to data assimilation, as demonstrated in our previous study (Miyazaki et al., 2015). In the comparison against independent ozonesonde measurements in this study, spatial gaps between the model/analysis and observations (i.e., representativeness error) result in large RMSE even after data assimilation.

The relevant sentence has been rewritten as follows:
"Root-Mean-Square-Errors (RMSEs) are also reduced above the middle troposphere, although the reduction rate is relatively small compared to the bias, probably due to representativeness errors between the ozonesonde measurements and data assimilation analysis."

*P7L14: As you also show the ozone sondes in the sections on seasonal variation and hemispheric gradient, it seems odd not to show the ozone sondes observations in Figure 3. Please add a further panel with colour dots at the station location.*

Added.

*P7L18: It is confusing that you choose a different latitude bands for table 1 and table 2. Please use the*

*same selection of latitude bands through the paper.*

Corrected.

*P7L18: Please, clarify how spatial r was calculated (only using the 5 year mean, based on lat-long grid points or area-weighted grid-points etc. Consider filtering small scale noise by averaging over areas corresponding to the resolution of the reanalysis.*

The sentence has been rewritten as:
"As summarized in Table 4, the global spatial distributions are similar between the five-year mean reanalysis field and the ensemble mean when estimated at 2º×2.5º spatial resolution, with a spatial correlation (r) greater than...".

In addition, the following sentence has been added in Section 2.2:
"Both the ACCMIP models and chemical reanalysis are interpolated to at 2º×2.5º spatial resolution and 67 levels, following Bowman et al. (2013), and then compared each other. Spatial correlations are computed with consideration of weighting for the latitude."

*The spatial correlation coefficient presented here seems less suited to express agreement in spatial patterns, which would be meaningful for the understanding of the model performance. Spatial r might be too much effected by the underlying spatial variability of the actual fields, thereby penalizing fields with greater more random variability i.e. standard deviation.*

Because all model and reanalysis fields were interpolated into the same spatial resolution (2º×2.5º) before the comparisons, the estimated spatial correlation can provide information on the model performance on the spatial pattern at that spatial scale. Although more thorough evaluations would be required for more careful discussions of the spatial pattern, the present evaluation method has been widely used and is valid.

*P7L19: The lower spatial correlation coefficient at p=500 hPa in NH could simply be caused by a different transport patterns (winds) and larger heterogeneity than in SH. Good correlation at the surface could be simply because a good match of emission patterns. High correlation at 200hPa in extra-tropics could mean that the transition in to the stratosphere agreed reasonably well. So are the different spatial r really helpful to distinguish model performance?*

The different reasons for each region raised by the reviewer are discussed in the manuscript. Because the spatial correlation varies significantly among the models as discussed in Section 4.1, it is a useful diagnostic of model performance. Please also see our reply above.

*P8L4: Please clarify again how the statistical variables shown in the Taylor diagram were computed. Given my scepticism about the meaning of the spatial correlation, I would consider omitting Figure 4 and shortening the discussion.*

The following sentence has been added in Section 2.2:
"Both the ACCMIP models and chemical reanalysis are interpolated to at 2º×2.5º spatial resolution and 67 levels, following Bowman et al. (2013), and then compared each other. Spatial correlations are computed with consideration of weighting for the latitude."

We think the Taylor diagram plots are useful to measure the general performance of each model and are widely used in the climate model evaluation. Please also see our reply on the spatial correlation estimates above.

*P9L6: Please clarify how the seasonal amplitude was calculated. How was made sure that "noise", i.e. unstructured variability, was not attributed to the seasonal amplitude.*

The seasonal amplitude was estimated from the difference between maximum and minimum monthly mean concentrations, which could reflect noise in the seasonal variation. This is described in the revised manuscript.

*P9L12: In section5 you discuss the sampling bias with respect to the regional areas. You should also discuss the sampling biases w.r.t to the selected latitude bands. This is needed because you also discuss model performance for the latitude bands. As mentioned in the general remarks, please also indicate the difference between the model results sampled at ozone locations and observation times and the area averages.*

Please see our reply above.

*P9L13: Please confirm that the average is area weighted and not based simply on lat-long grid boxes, which decrease in size towards the poles.*

The average is area weighted, as described in the revised manuscript.

*P10L8: Please clarify how you exactly calculate the hemispheric gradient both for the gridded fields and for the ozone sonde observations.*

The following sentence has been added:
"For the estimation of the gradient using the ozonesonde observations, we made a gridded dataset from the ozonesonde observations based on the completion by Tilmes et al (2012) at $2^{o}$x$2.5^{o}$ spatial resolution, and then calculated area-weighted hemispheric mean concentrations using the gridded data."

*P10L15: add missing "At" before "Around "*

Added.

*P10L18: How do these value compare to values from the literature?*

To the best of our knowledge, there is no literature that shows an inter-hemispheric ozone gradient for different altitudes of the troposphere.

*P10L26: As you already discuss sampling biases it a bit inconsistent to put the section at this place. This very good introduction to sampling biases (p10L27 – p11L15) should come earlier in the paper, i.e. in the part when you discuss the methods (section2)*

Moved to Section 2.4.

*P10L15: The sampling biases depends on the averaging area and the selection of ozone sondes. The sampling biases estimated by using your re-analysis should be presented for the Tilmes regions as well as for the latitude bands (choose one set only) in and uniform way. As model results are often evaluated for the latitude bands, this information would be very interesting for the scientific community.*

Table 8 has been modified to include the results for four latitudinal bands.

*P10L24 Please add also the stations used for the latitude bands averages in table 3.*

This is mentioned in Section 2.3 and Table 2 in the revised manuscript.

*P12L1: I think there is would be very good to compute the sampling biases also directly for the re-analysis i.e. the difference between the re-analysis sampled as ozone and as area-time averages. This information would be in my opinion of more general meaning than the values for the ACCMIP error.*

Table 7 has been added to discuss the sampling bias for the reanalysis fields. The following discussion, regarding this table, has been added in Section 5:

"Table 7 demonstrates the regional and seasonal mean differences of the reanalysis concentrations between the complete sampling and the ozonesonde sampling. The ozonesonde sampling results have higher concentrations (by about 3 %) in the two NH polar regions for most cases, whereas the difference is smaller in NH polar west than in NH polar east. Among the NH mid-latitude regions, a large difference (about 14 %) exists between the two cases over the eastern United States in June-August (JJA), where the comparison using monthly reanalysis fields sampled at the ozonesonde locations (brackets in Table 7) suggests that the sampling bias is dominated by temporal variations. The tropical and subtropical regions exhibit large sampling biases, 4-12.3 % over the NH subtropics, -3.2-5.0 % over the Western Pacific and East Indian Ocean, 0--7.8 ¥% over the equatorial Americas, and -3.8-7.5 % over the Atlantic Ocean and Africa. In most of the tropical and subtropics regions, both the spatial and temporal sampling biases are important, because of large spatial and temporal variability of ozone and the sparse observation network. For the global tropics, the sampling bias reaches 13 % in the NH (Eq-30N) and 8 % in the SH (30S-Eq). Thus, the ozonesonde network has a major limitation when it comes to capturing ozone concentrations that are representative of seasonal and regional means for the entire tropical region. The sampling bias may not be negligible even in the SH (0.3-3.9 % in the SH mid-latitudes and 0.8-4.2 % in the SH high latitudes), and it is large (up to 13 %) when estimations are done for a large area (90-30S). The large sampling bias in 90-30S is primarily attributed to spatial variability. The impact of the sampling bias on the model evaluation is discussed in the following section."

*P12L1: Why does table 5 show the median whereas otherwise only the ensemble mean is discussed or shown. (Using only the median would be perhaps a better option overall)*

We present mean values in other estimates because we also discuss the standard deviation to the mean. Medians are shown only in this table (Table 8 in the revised manuscript), in order to provide more robust estimates of the model error and sampling bias for each region. The following sentence has been added to clarify this point:

"The sampling bias is evaluated using the median of the multiple models to provide robust estimates of the model performance."

*P13L8: The ozone network in SH high-latitudes is actually quite high because of the need to monitor the ozone hole. The launch frequency varies for some stations a lot because more sondes are launched during the ozone hole season.*

Yes, I agree. However, because only three stations were considered in the comparison following Tilmes et al. (2012), the ozonesonde network is not sufficient to capture the ozone variations. Our results suggest that the temporal sampling bias mainly causes the sampling bias in the SH high latitudes in MAM and JJA. This is discussed in the revised manuscript.

*P13L24: Please add also the sampling biases for the latitude bands.*

Added in Table 10. The following sentences have been added to discuss the results:
"Because the seasonal variations differ among different regions, the seasonal amplitude estimated for the entire NH extratropics (30-90N) is largely different between the two estimates throughout the troposphere."
"The sampling bias in the seasonal amplitude estimated for the entire tropics is larger than 60 % throughout the troposphere both in the NH (Eq-30N) and SH (30S-Eq)."

*P14 L14: Please clarify how much of the analysis ensemble spread depends on some-times arbitrary choices to cause spread between the ensemble members.*

The following sentence has been added:
"Note that the data assimilation setting influences the analysis uncertainty estimation in the reanalysis. In particular, the analysis spread was found to be sensitive to the choice of ensemble size (Miyazaki et al., 2012b). A large ensemble size is essential to capture the proper background error covariance structure (i.e., analysis uncertainty)."

*P14L15: Please clarify to what extent the analysis uncertainty is controlled by the uncertainty of the assimilated observations.*

The sentences have been rewritten as follows:
"Miyazaki et al (2015) investigated that the analysis spread is caused by errors in the model input data, model processes, and assimilated measurements, and it is reduced if the analysis converges to a true state. The analysis spread is smaller in the extratropical lower stratosphere than in the tropical upper

troposphere at 200 hPa, because of the high accuracy of the MLS measurements. In contrast, in the middle troposphere, the analysis spread is generally smaller in the tropics than the extratropics because of the higher sensitivities in the TES retrievals."

*P14L21: How does the ensemble spread relate to the spread of the ensemble in the EnKF. Could the ACCMIP ensemble spread be used to verify the EnKF ensemble spread?*

The simultaneous enhancement of the analysis uncertainty and the model spread indicates low robustness of the validation results, as discussed in the manuscript. Verification of EnKF ensemble spread using ACCMIP ensemble model spread would be an interesting research topic but requires careful discussion and is clearly out of the scope of the present study.

*P15L3: Please see my general comment on this chapter. Improvement on species not directly assimilated needs to be demonstrated. Long-term reanalysis could suffer from artificial jumps because of the change in the observing system (for example degradation of TES after 2010).*

Corrections have been made. Please see our reply above.

*P15L29: I don't understand this conclusion at all. Re-analysis are only valid for present day conditions when observations are available. They cannot be used to for pre- industrial estimates nor the differences with today's values.*

Assuming that the persistent systematic bias from the pre-industrial to present day can be attributed to time-independent model errors in chemical and transport processes, as suggested by previous studies (e.g., Young et al., 2013), the validation results using the reanalysis for the present day has the potential to evaluate preindustrial to present-day ozone radiative forcing. This is discussed in the manuscript.

*P16L4: Please mention that you (only) consider ozone sondes as reference in this paper.*

The following sentence has been added:
"The evaluation results are also used to quantify the ozonesonde network sampling bias."

*P16L6: Please mention the advantage of using a re-analysis, i.e. a gridded field. Please mention that the biases of the re-analysis against ozone sondes are small.*

The sentences have rewritten as follows:
"The reanalysis product provides comprehensive and unique information on global ozone distributions for the entire troposphere and on the weakness of the individual models and multi-model mean. Validation of the chemical reanalysis using global ozonesondes shows good agreement throughout the free troposphere and lower stratosphere for both seasonal and year-to-year variations."

*P16L17: Please add a statement if these finding are consistent with other evaluation studies, i.e. the Young et al. paper.*

The following sentence has been added:
"The performance of the ACCMIP model when compared with the reanalysis is qualitatively similar for most cases from that shown by Young et al. (2013) using the ozonesonde measurements but quantitatively different because of the ozonesonde network sampling bias."

*P16L20-30: Please give some numbers for the sampling biases. Also include the sampling bias w.r.t to latitude bands.*

The following sentences have been added:
"For the global tropics, the ozonesonde sampling bias is largely negative by 80 % in the NH (Eq-30N) in SON and by 50 % in the SH (30S-Eq) in MAM."
"Large sampling biases (> 60 %) exist in the SH extratropics (90-30S) in DJF and MAM"

*p17L4: Please add a statement that it will be a challenge to combine all these observations in a consistent way in a more long term re-analysis.*

The following sentence has been added:
"Combining many observations requires a bias correction procedure for each assimilated measurement to improve the reanalysis quality but needs to be carefully checked."

**Authors' comments in reply to the anonymous referee for "Evaluation of ACCMIP ozone simulations using a multi-constituent chemical reanalysis" by K. Miyazaki and K. Bowman**

We want to thank the referee for the helpful comments. We have revised the manuscript according to the comments, and hope that the revised version is now suitable for publication. Below are the referee comments in italics, with our replies in normal font.

***Reply to Referee #2***

*Specific Comments:*

*Abstract: Please mention that a 4-year reanalysis data is used to compare ACCMIP time-slice simulations, and that the evaluation itself can be biased because of this inconsistency.*

The abstract has been rewritten to describe the time period of the reanalysis and ACCMIP simulations.

*P1L17: Insert anthropogenic between "...important greenhouse.."*

Inserted.

*P2L1: There are several studies highlighting the use of CTMs/CCMs to assess the radiative impacts of tropospheric ozone prior to Bowman et al (2013). Please acknowledge those.*

Several studies are cited in the following sentence.

*P2: 2nd and 3rd paragraphs discuss uncertainties in measurements for evaluating chemistry-climate models. I think they can be combined and modified for clarity.*

Combined and modified.

*P3L12: Please provide motivation for evaluating only tropospheric ozone and not its precursors (e.g., NO2, CO). Presumably biases in ozone are driven by biases in its precursors.*

The following sentences have been added:
"Model errors in precursors can also be evaluated using the reanalysis product, and this could help

identify error sources in tropospheric ozone simulations. However, because no other study has shown the potential of reanalysis ozone for model evaluation, this study focuses on tropospheric ozone only."

*P4. How is photolysis calculated in the forecast model? Does the model represent methane - prescribed concentrations or emissions? Since the ACCMIP models used different emissions inventory for ozone precursors compared to what is used in this reanalysis product, how would the comparison be affected by this difference?*

The following sentences have been added:
"The radiative transfer scheme considers absorption within 37 bands, scattering by gases, aerosols, and clouds, and the effect of surface albedo. Detailed radiation calculations are used for photolysis calculation. Methane concentrations were scaled on the basis of present-day values with reference to the surface concentration."

Because the surface emissions of NOx and CO are optimized using data assimilation in the chemical reanalysis, the difference in the emission inventory should not affect the comparison. In our previous studies (Miyazaki et al., 2012, 2015, 2017), it was confirmed that the a priori emissions do not largely influence the a posteriori emissions through long data assimilation cycles.

*P4L13: What is the convection scheme of MIROC-AGCM? Please describe it in a sentence.*

The sentence has been rewritten as follows:
"Lightning NOx (LNOx) sources in MIROC-Chem were calculated based on the relationship between lightning activity and cloud top height (Price and Rind, 1992) and using the convection scheme of MIROC-AGCM developed based on the scheme presented by Arakawa and Schubert (1974)."

*P5L7: What is DOMINO data? What does the acronym stand for?*

DOMINO stands for Dutch OMI NO2. This is described in the revised manuscript.

*P5L21: According to Lamarque et al (2013), models were run for several years (up to 10 years) for each time slice. How are the model results for 2000 time-slice handled for comparison with the reanalysis?*

The following sentence has been added:
"The number of years that the ACCMIP models simulated for the 2000 decadal simulation mostly varied between 4 and 12 years for each model. Each model simulation was averaged over the simulated years."

To discuss the limitation of the short period of the chemical reanalysis used for the validation, the following sentences have been added in Section 6.1:
"The five-year reanalysis (2005-2009) may cause biases in the estimated model errors in the evaluation of the 2000 decade ACCMIP simulations that used decadal-averaged SST boundary conditions and biomass burning emissions averaged over 1997-2006 (Lamarque et al., 2010). It may neglect the influences of interannual and decadal changes in anthropogenic and biomass emissions and meteorology. Longer-term reanalysis and time-consistent validation are required to obtain more robust error estimations."

*P5L29: Young et al. (2013) note that although the models used the same anthropogenic and biomass burning emissions, model-to-model diversity in the implemented chemical scheme resulted in differences in precursor (especially VOCs) emissions across the models. So, the statement "same emissions were used in all the models" is not accurate. Please modify.*

Removed.

*P6L7: Please provide details on how the ACCMIP model monthly ozone concentrations were interpolated to 2 hour temporal resolution? What diagnostics from the reanalysis were used to compare with observations - monthly or hourly ozone? Some clarification is needed here.*

The sentences have been rewritten as:
"The two-hourly reanalysis and forecast model (i.e., control run) fields were linearly interpolated to the time and location of each measurement, with a bin of 25 hPa, and then compared with the measurements. For the ACCMIP models, the monthly model outputs were compared with the measurements at the location of each measurement."

*P6L29: Can you please elaborate on the model setting that caused this significant degradation of the representation of ozone in the UTLS?*

The clause has been rewritten as follows:

"because of different model settings, such as the upper boundary conditions of NOy, Cly, and Bry"

*P7L3: The terminology to refer to model output is somewhat confusing making it difficult to keep track of observations versus model output. Suggest referring to output from the forecast model and ACCMIP as "forecast model output" and "ACCMIP model output", respectively, and observations as "data".*

Corrected throughout the paper.

*P7L22-24: The simulation of the "wave-1" pattern by ACCMIP models has already been highlighted by Young et al. (2013). This reference needs to be cited.*

Cited.

*P7L32: It is not clear what "common reported" means here. Please clarify.*

Replaced by "Young et al. (2013) consistently revealed the positive bias in the NH and negative bias in the SH using OMI/MLS tropospheric ozone column measurements."

*P8L10: Which region is "In this region. . ." referring to?*

Replaced by "In the NH extratropics in the lower and middle troposphere".

*P8L14-15: These results are consistent with those discussed in Young et al. (2013).*

Young et al. (2013) is cited.

*P8L24: The differences could also be associated with the way biomass burning emissions are handled across models - whether they are emitted at the surface layer or distributed vertically in the model, as a larger role for biomass burning in the tropical mid to upper troposphere has been suggested recently (Anderson et al., 2016).*

The following sentence has been added:
"For instance, biomass burning emissions are handled differently across the models, which may lead to differences in ozone simulations in the tropics (Anderson et al., 2016)."

*P8L33-34: Why are results from a specific model highlighted? Is it because model 8 is driving the large model diversity at 200hPa? Please provide a figure to support this statement if this is indeed the case.*

The sentence has been removed.

*P9L3: It took me a while to understand the meaning of "reanalysis concentrations from the ozonesonde sampling". Please rephrase this to indicate that reanalysis output is sampled at ozonesonde sites instead of averaging the reanalysis at all grid cells.*

Replaced.

*P929-31: Here again the results are consistent with Young et al. (2013) - see their Figure 4 and its discussion.*

Cited.

*P10L23-25: Please clarify what is meant by "radiative heating distribution in chemistry-climate simulations are largely uncertain." Are you referring to the radiative heating due to tropospheric ozone? Please elaborate on how the O3 NH/SH ratio provides information on the radiative heating distribution.*

The sentence has been replaced by
"The large systematic error in the NH/SH ratio suggests that, for instance, the inter-hemispheric distribution of radiative heating due to tropospheric ozone in chemistry--climate simulations are largely uncertain in most models"

*P11L17: replace "completion" with compilation.*

Replaced.

*P11L19: By coarse resolution, do you mean the coarse horizontal resolution?*

Yes. Replaced.

*P11L29: Did you mean - "...less variabilities in the SH than in the NH"?*

Yes. Corrected.

*P12L9-11: Please quantify large in "Large negative model biases..." and larger in "...errors are larger than those. . ."*

The sentences have been rewritten as:
"Large negative model biases against the ozonesonde observations have been reported by Young et al. (2013) for 250 hPa (by about -13 % for the NH polar west and -18 % for the NH polar east for the annual mean concentration), whereas results from this study suggest that these errors based on the ozonesonde sampling (by -14 % for the NH polar west and -18 % for the NH polar east in DJF in our estimates) are larger than those from regional and seasonally representative model bias (by -3 % and +5 %, respectively)."

*Section 5.1 Please elaborate on how the comparisons discussed here may be influenced by inconsistency in the time period of the reanalysis and the ACCMIP simulations. The reanalysis output is for the 2005-2009 whereas the ACCMIP simulations are representative of 2000. Additionally, the precursor emissions used in ACCMIP simulations were decadal means and not specific to the year of simulation. For example, year 2000 biomass burning emissions were calculated as average over 1997–2006 (Lamarque et al., 2010), so they encompass the high emissions over Southeast Asia in 1998, an El Nino year. One would expect that biases due to sampling in time would occur similar to the biases due to sampling in space. How significant are the biases due to spatial sampling errors (highlighted here) compared to temporal sampling errors. The issue of temporal and spatial sampling was recently highlighted by Lin et al. (2016) in the context of tropospheric ozone trends.*

The limitation of the evaluation based on the 2005-2009 reanalysis fields is discussed in the revised manuscript. Please see our reply above.

The spatial and temporal sampling errors are separately evaluated in the revised paper, and the results are presented in Table 7 and discussed in Section 5.1. Lin et al. (2016) is cited in the revised paper.

*P14L24: Replace biogenetic with biogenic.*

Replaced.

*P15L6: I am not sure if the 2005-2009 can be considered a "long-record" Are the authors referring to*

*the possibility of a long-record reanalysis sometime in the future when observations have accumulated in time.*

In this sentence, we are referring to the possibility in the future. The sentence has been replaced by
"A long-record of the reanalysis will allow...."

*P15L7-9: From Lamarque et al. (2013): "This averaging was designed to reduce the effect of interannual variability and therefore provide optimal conditions from which average composition changes and associated forcings can be more readily computed." The ACCMIP simulations were designed to remove interannual variability, therefore, it is unjustified to state that "the influence of ENSO was not well-simulated in ACCMIP. . ."*

The sentence has been replaced by
"It is noted that the influence of ENSO was not included..."

*P33 Figure 5. Please provide statistics such as mean bias and correlation for the comparisons here.*

These statistics are provided in Table 4.

# Evaluation of ACCMIP ozone simulations and ozonesonde sampling biases using a satellite-based multi-constituent chemical reanalysis

削除：and ozonesonde sampling biases

削除：satellite-based

\begin{abstract}

The Atmospheric Chemistry Climate Model Intercomparison Project (ACCMIP) ensemble ozone simulations for the present-day from the 2000 decade simulation results are evaluated by a state-of-the-art multi-constituent atmospheric chemical reanalysis that ingests multiple satellite data including the Tropospheric Emission Spectrometer (TES), the Microwave Limb Sounder (MLS), the Ozone Mapping Instrument (OMI), and the Measurements of Pollution in the Troposphere (MOPITT) for 2005--2009. Validation of the chemical reanalysis against global ozonesondes shows good agreement throughout the free troposphere and lower stratosphere for both seasonal and year-to-year variations, with an annual mean bias of less than 0.9 ppb in the middle and upper troposphere at the tropics and mid-latitudes. The reanalysis provides comprehensive information on the weakness of the models, whereas we consider that the spatial and temporal coverage of individual measurements, such as ozonesonde measurements, is insufficient to capture the temporally and spatially representative model bias. The model evaluation using the reanalysis reveals that the ensemble mean overestimates ozone in the northern extratropics by 6--11 ppb while underestimating by up to 18 ppb in the southern tropics over the Atlantic in the lower troposphere. Most models underestimate the spatial variability of the annual mean concentration in the extratropics of both hemispheres in the lower troposphere. The ensemble mean also underestimates the seasonal amplitude by 25--70 \% in the northern extratropics and overestimates the inter-hemispheric gradient by about 30 \% in the lower and middle troposphere. A part of the discrepancies can be attributed to the five-year reanalysis data for the decadal model simulations. However, these differences are less evident with the current sonde network, which is shown to provide biased regional and monthly ozone statistics, especially in the tropics. The ozonesonde sampling bias in the evaluated model bias for the seasonal mean concentration is 40--50 \% over the Western Pacific and East Indian Ocean and reaches 110 \% over the equatorial Americas in the middle troposphere. These systematic biases have implications for ozone radiative forcing and the response of chemistry to climate that can be further quantified as the satellite observational record extends to multiple decades.

削除：

削除：These

\end{abstract}

\clearpage

\introduction

Tropospheric ozone is one of the most important air pollutants and the third most important anthropogenic greenhouse gases in the atmosphere (Forster et al., 2007; HTAP, 2010; Myhre et al., 2013; Stevenson et al., 2013) while also playing a crucial role in the tropospheric oxidative capacity through production of hydroxyl radicals (\chem{OH}) by photolysis in the presence of water vapor (Logan et al., 1981; Thompson, 1992). Global tropospheric ozone is formed from secondary photochemical production of ozone precursors including hydrocarbons or carbon monoxide (\chem{CO}) in the presence of nitrogen oxides (\chem{NO_x}) modulated by additional processes including in-situ chemical loss, deposition to the ground surface, and inflow from the stratosphere. These ozone precursors are largely controlled by anthropogenic and natural emissions, e.g., transport, industry, lightning, biomass burning sources. Representation of tropospheric ozone in chemical transport models (CTMs) and chemistry climate models (CCMs) is also important in estimating its impact on the atmospheric radiative budget. A number of chemical transport models (CTMs) and chemistry climate models (CCMs) have been developed and used to study variations in atmospheric environment and its impacts on climate (e.g. Bowman et al., 2013; Shindell et al, 2006, 2013; Stevenson et al., 2006, 2013; Wild, 2007, Kawase et al., 2011; Young et al., 2013). However, current tropospheric ozone simulations still have large uncertainties because of the incomplete representation of model processes, as well as the large uncertainty in precursor emissions. These in turn increase uncertainty in CCM projections.

Climate model evaluation has primarily been achieved by comparisons with observed concentrations or related variables, which requires a precise description of their geographical, vertical, and temporal variations. Various measurements have been employed for evaluating

削除：

削除：mobile

削除：(Bowman et al., 2013)

削除：, e.g., ground-based, ozonesonde, and satellite-retrieved measurements,

simulated fields (e.g., Huijnen et al., 2010; Parrish et al., 2014; Stevenson et al., 2006, 2013; Young et al., 2013). However, information obtained from individual measurements is limited, and evaluation of global ozone fields with a suite of satellite measurements and in situ measurements is challenging because of limited vertical sensitivity profiles that differ among measurements, different overpass times, and mismatches in spatial and temporal coverage between the instruments. First, surface measurements have a spatial representativeness that is much smaller than that of global models over polluted areas. Second, ozone climatology data sets have been established based on ozonesonde measurements for use in model evaluation (Logan et al., 1999; Considine et al., 2008). Tilmes et al. (2012) generated an ozone climatology using ozonesonde measurements obtained between 1995 and 2011, which mostly consists of the same station data described by Logan (1999) and Thompson et al (2003), but covering a longer time period. Using the compiled data of Tilmes et al. (2012), Young et al. (2013) conducted an intensive validation of tropospheric ozone from multiple model simulations in the Atmospheric Chemistry and Climate Model Inter-comparison Project (ACCMIP). However, the climatological data does not provide information on the temporal variability of the observed ozone. In addition, the current ozonesonde network does not cover the entire globe and is not homogeneously distributed between the hemispheres, ocean and land, and urban and rural areas, and its sampling interval is typically a week or longer. Model errors are also expected to vary greatly in time and space at various scales. Therefore, we consider that the spatial and temporal coverage of the ozonesonde network is insufficient to capture the temporally and spatially representative model bias. Third, satellite-retrieved measurements such as those from the Tropospheric Emission Spectrometer (TES) (Herman and Kulawik, 2013) and the Infrared Atmospheric Sounding Interferometer (IASI) (Clerbaux et al., 2009) have great potential for evaluating global ozone fields (e.g., Aghedo et al., 2011). However, information obtained from currently available satellite measurements are still limited. Their vertical sensitivity is not enough to resolve detailed vertical structures in the troposphere as appeared in current global models, and they measure at only a particular overpass time, thus the diurnal variation information is missing. Meanwhile, the characteristics of each measurement, such as observational error, vary with observational condition, but their influence is rarely taken into consideration in model evaluations.

削除: , ... [1]

削除: Ozone

Data assimilation is a technique for combining different observational data sets with a model, with consideration of the characteristics of individual measurements (e.g., Kalnay, 2003; Lahoz and Schneider, 2014). Advanced data assimilation allows the propagation of observational information in time and space and from a limited number of observed species to a wide range of chemical components, and provides global fields that are physically and chemically consistent and in agreement with individual observations (Sandu and Chai, 2011; Bocquet et al., 2015). Various studies have demonstrated the capability of data assimilation techniques in the analysis of chemical species in the troposphere and stratosphere (e.g. Stajner and Wargan, 2004; Jackson, 2007; Parrington et al., 2009; Kiesewetter et al., 2010; Flemming et al., 2011; Coman et al., 2012; Inness et al., 2013; Emili et al., 2014; Miyazaki et al., 2012a, 2012b, 2013, 2014, 2015, 2016; van der A et al., 2015; Gaubert et al., 2016).

Reanalysis is a systematic approach to creating a long-term data assimilation product. Meteorological reanalyses have been established at operational centers for many years and are widely used in climate and meteorological research (e.g., Hartmann et al., 2013). Tropospheric chemical reanalysis, however, is relatively new. Inness et al. (2013) performed an eight-year reanalysis of tropospheric chemistry for 2003--2010 using the integrated forecasting system with modules for atmospheric composition (C-IFS) with observations sensitive primarily to the upper troposphere, and highlighted the importance of estimating surface emissions. This chemical reanalysis is recently updated by Flemming et al. (2017). Miyazaki et al. (2015) simultaneously estimated concentrations and emissions for an eight-year tropospheric chemistry reanalysis for 2005--2012 obtained from an assimilation of multi-constituent satellite measurements, which had greater lower tropospheric sensitivity, using an ensemble Kalman filter (EnKF). Chemical reanalysis using the EnKF has been used to provide comprehensive information on atmospheric composition variability and elucidate variations in precursor emissions and to evaluate bottom-up emission inventories (Miyazaki et al., 2014, 2015, 2016).

削除: 2016

In this study, we explore the new potential of chemical reanalysis for evaluation of tropospheric ozone profiles in multi-model chemistry climate simulations from ACCMIP (Lamarque et al., 2013). Model errors in precursors can also be evaluated using the reanalysis product, and this could help identify error sources in tropospheric ozone simulations.

However, because no other study has shown the potential of reanalysis ozone for model evaluation, this study focuses on tropospheric ozone only. ACCMIP models have been used to calculate historic and future radiative and chemically important species and their coupling with the broader climate system (Bowman et al., 2013; Lee et al., 2013; Naik et al., 2013; Stevenson et al., 2013; Shindell et al., 2013; Voulgarakis et al., 2013; Young et al., 2013). We characterize ACCMIP models in simulating global distributions and the seasonal variation of ozone from the lower troposphere to the lower stratosphere. We further discuss the limitation of the current ozonesonde network for evaluating temporally and spatially representative model errors. To the best of our knowledge, this is the first study to apply chemical reanalysis to the evaluation of global chemistry-climate models and consequently offers a similar potential as meteorological reanalysis for evaluation of climate models (Ana4MIPS, https://esgf.nccs.nasa.gov/projects/ana4mips/Background).

削除:

\section{Methodology}

\subsection{Chemical data assimilation system}

The data assimilation system is constructed based on a global CTM MIROC-Chem (Watanabe et al. 2011) and an EnKF described in Miyazaki et al. (2016), which can be consulted for more detailed information. We use the two-hourly global chemical reanalysis data for the period 2005--2009 when tropospheric ozone fields are strongly constrained by TES tropospheric ozone measurements. The availability of TES measurements is strongly reduced after 2010, which led to a degradation of the reanalysis performance, as demonstrated by Miyazaki et al. (2015).

A major update from the system used in Miyazaki et al. (2015) to the system used in this study is the replacement of forecast model from CHASER (Sudo et al., 2002) to MIROC-Chem (Watanabe et al., 2011), which caused substantial changes in the a priori field and thus the data assimilation results of various species. Microwave Limb Sounder (MLS) retrievals have been updated from v3.3 in Miyazaki et al. (2015) to v4.2 in this study. In addition, we attempt to optimize the surface \chem{NO_x} emission diurnal variability using data

削除: MLS

assimilation of multiple \chem{NO_2} satellite retrievals obtained at different overpass times in the updated system (Miyazaki et al., 2016).

\subsubsection{Forecast model}

The forecast model, MIROC-Chem (Watanabe et al., 2011), considers detailed photochemistry in the troposphere and stratosphere by simulating tracer transport, wet and dry deposition, and emissions, and calculates the concentrations of 92 chemical species and 262 chemical reactions (58 photolytic, 183 kinetic, and 21 heterogeneous reactions). Its tropospheric chemistry considers the fundamental chemical cycle of \chem{O_x}-\chem{NOx}-\chem{HO_x}-\chem{CH_4}-\chem{CO} along with oxidation of non-methane volatile organic compounds (NMVOCs) to properly represent ozone chemistry in the troposphere. Its stratospheric chemistry simulates chlorine and bromine containing compounds, CFCs, HFCs, OCS, \chem{N_2O}, and the formation of polar stratospheric clouds (PSCs) and associated heterogeneous reactions on their surfaces. The radiative transfer scheme considers absorption within 37 bands, scattering by gases, aerosols, and clouds, and the effect of surface albedo. Detailed radiation calculations are used for photolysis calculation. Methane concentrations were scaled on the basis of present-day values with reference to the surface concentration. MIROC-Chem has a T42 horizontal resolution (2.8$^\circ$) with 32 vertical levels from the surface to 4.4 hPa. The horizontal model resolution is comparable to the resolution of ACCMIP models (ranging from 1.24$^\circ$ to 5$^\circ$). It is coupled to the atmospheric general circulation model MIROC-AGCM version 4 (Watanabe et al., 2011). The simulated meteorological fields were nudged toward the six-hourly ERA-Interim (Dee et al., 2011) to reproduce past meteorological fields.

削除：.

The a priori values for surface emissions of \chem{NO_x} and \chem{CO} were obtained from bottom-up emission inventories. Anthropogenic \chem{NO_x} and \chem{CO} emissions were obtained from the Emission Database for Global Atmospheric Research (EDGAR) version 4.2 (EC-JRC, 2011). Emissions from biomass burning were based on the monthly Global Fire Emissions Database (GFED) version 3.1 (van der Werf et al., 2010). Emissions from soils were based on monthly mean Global Emissions Inventory Activity

(GEIA) (Graedel et al., 1993). Lightning \chem{NO_x} (\chem{LNO_x}) sources in MIROC-Chem were calculated based on the relationship between lightning activity and cloud top height (Price and Rind, 1992) and using the convection scheme of MIROC-AGCM developed based on the scheme presented by Arakawa and Schubert (1974). For black carbon (BC) and organic carbon (OC) and other precursor gases, surface and aircraft emissions are specified from the emission scenarios for Greenhouse Gas and Air Pollution Interactions and Synergies (GAINS) model developed by International Institute for Applied System Analysis (IIASA) (Klimont et al., 2009; Akimoto et al., 2015).

\subsubsection{Data assimilation method}

Data assimilation used here is based upon on an EnKF approach (Hunt et al., 2007). The EnKF uses an ensemble forecast to estimate the background error covariance matrix and generates an analysis ensemble mean and covariance that satisfy the Kalman filter equations for linear models. In the forecast step, a background ensemble, $\vec{x}^b_i (i=1,...,k)$, is obtained from the evolution of an ensemble model forecast, where $\vec{x}$ represents the model variable, $b$ is the background state, and $k$ is the ensemble size (i.e., 32 in this study). The ensemble perturbations were introduced to all the state vector variables as described below. The background ensemble is then converted into the observation space, $\vec{y}^b_i=H(\vec{x}^b_i)$, using the observation operator $H$ which is composed of a spatial interpolation operator and an operator that converts the model fields into retrieval space, which can be derived from an a priori profile and an averaging kernel of individual measurements (e.g., Eskes and Boersam, 2003; Jones et al, 2003). Using the covariance matrices of observation and background error as estimated from ensemble model forecasts, the data assimilation determines the relative weights given to the observation and the background, and then transforms a background ensemble into an analysis ensemble, $\vec{x}^a_i (i=1,...,k)$. The new background error covariance is obtained from an ensemble forecast with the updated analysis ensemble.

In the data assimilation analysis, a covariance localization is applied to neglect the covariance among unrelated or weakly related variables, which has the effect of removing the influence

削除: satellite retrieval

of spurious correlations resulting from the limited ensemble size. The localization is also applied to avoid the influence of remote observations that may cause sampling errors. The state vector includes several emission sources (surface emissions of \chem{NO_x} and \chem{CO}, and \chem{LNO_x} sources) as well as the concentrations of 35 chemical species. The emission estimation is based on a state augmentation technique, in which the background error correlations determines the relationship between the concentrations and emissions of related species for each grid point. Because of the simultaneous assimilation of multiple-species data and because of the simultaneous optimization of the concentrations and emission fields, the global distribution of various species, including \chem{OH}, is modified considerably in our system. Miyazaki et al. (2015) demonstrated that the Northern/Southern Hemisphere \chem{OH} ratio became closer to an observational estimate of Patra et al (2014) due to the multiple-species assimilation. This propagates the observational information between various species and modulates the chemical lifetimes of many species (Miyazaki et al., 2012b; 2015; 2016).

\subsubsection{Assimilated measurements}

Assimilated observations were obtained from multiple satellite measurements (Table 1). Tropospheric \chem{NO_2} column retrievals used are the version-2 Dutch Ozone Monitoring Instrument (OMI) \chem{NO_2} (DOMINO) data product (Boersma et al., 2011) and version 2.3 TM4NO2A data products for Scanning Imaging Absorption Spectrometer for Atmospheric Cartography (SCIAMACHY) and Global Ozone Monitoring Experiment-2 (GOME-2) (Boersma et al., 2004) obtained through the TEMIS website (www.temis.nl). The TES ozone data and observation operators used are version 5 level 2 nadir data obtained from the global survey mode (Bowman et al, 2006; Herman and Kulawik, 2013). This data set consists of 16 daily orbits with a spatial resolution of 5--8 km along the orbit track. The MLS data used are the version 4.2 ozone and \chem{HNO_3} level 2 products (Livesey et al., 2011). We used data for pressures of less than 215 hPa for ozone and 150 hPa for \chem{HNO_3}. The Measurement of Pollution in the Troposphere (MOPITT) CO data used are version 6 level 2 TIR products (Deeter et al., 2013).

削除: OMI

削除: for Ozone Monitoring Instrument (OMI)

削除: Microwave Limb Sounder (MLS)

\subsection{ACCMIP models}

The Atmospheric Chemistry Climate Model Intercomparison Project (ACCMIP) focuses on chemistry-climate interactions needed to compute the proper climate forcing for Climate Model Intercomparison Project (CMIP5) climate simulations (Taylor et al., 2012) as well as the impact of climate change on chemical species. The ACCMIP consists of a series of time slice experiments for the long-term changes in atmospheric composition between 1850 and 2100, as described by Lamarque et al. (2013). The experimental design was based on decadal time-slice experiments driven by decadal mean sea surface temperatures (SST). This study uses the 2000 decade simulation results from 15 models (1. CESM-CAM, 2. CICERO-OsloCTM2, 3. CMAM, 4. EMAC, 5. GEOSCCM, 6. GFDL-AM3, 7. GISS-E2-R, 8. GISS-E2-TOMAS, 9. HadGEM2, 10. LMDzOR-INCA, 11. MIROC-CHEM, 12. MOCAGE, 13. NCAR-CAM3.5, 14. STOC-HadAM3, 15. UM-CAM). The number of years that the ACCMIP models simulated for the 2000 decadal simulation mostly varied between 4 and 12 years for each model. Each model simulation was averaged over the simulated years.

Meteorological fields were obtained from analyses in CICERO-OsloCTM2 and from climate model fields in MOCAGE. UM-CAM and STOC-HadAM3 simulated meteorological and chemical fields, but chemistry did not affect climate. In all other models, simulated chemical fields were used in the radiation calculations and hence provide a forcing effect on the general circulation of the atmosphere. Lamarque et al. (2013) indicated that most models overestimate global annual precipitation and have a cold bias in the lower troposphere.

Different models vary greatly in complexity. The calculated chemical species vary from 16 to 120 species. Photolysis rates are computed with offline or online methods, depending on the model. Many models include a full representation of stratospheric ozone chemistry and the heterogeneous chemistry of polar stratospheric clouds, but several models specify stratospheric ozone. Methane concentration is prescribed for the surface or over the whole atmosphere in many models. Ozone precursor emissions from anthropogenic and biomass burning sources were taken from those compiled by Lamarque et al. (2010), Natural emission sources such as isoprene emissions, and lightning and soil \chem{NO_x} sources were not

削除: , and the same emissions were used in all the models

specified and were accounted for differently between models. There is a large range in soil \chem{NO_x} emissions from 2.7 to 9.3 \unit{TgNyr^{-1}} and in \chem{LNO_x} sources from 1.2 to 9.7 \unit{TgNyr^{-1}} for the 2000 conditions. The range of natural emissions is a significant source of model-to-model ozone differences (Young et al. 2013). A complete description of the models along with the experiment design can be found in Lamarque et al. (2013).

Both the ACCMIP models and chemical reanalysis are interpolated to at 2$^\circ \times$2.5$^\circ$ spatial resolution and 67 levels, following Bowman et al. (2013), and then compared each other. Spatial correlations are computed with consideration of weighting for the latitude.

\subsection{Ozonesonde data}

Ozonesonde observations were taken from the World Ozone and Ultraviolet Radiation Data Center (WOUDC) database (available at http://www.woudc.org). All available data from the WOUDC database are used for the evaluation of reanalysis data (Section 3), as listed in Table 2. For the evaluation of ACCMIP models and ozonesonde sampling biases (Section 4 and 5), we use the ozonesonde sampling based on the compilation by Tilmes et al. (2012), which is shown in bold in Table 2. Because there is no observation after 2003 in Scoresbysund, this location has been removed from the compilation in this study. The accuracy of the ozonesonde measurement is about $\pm$5 \% in the troposphere (Smit and Kley, 1998).

To compare ozonesonde measurements with the data assimilation and ACCMIP models, all ozonesonde profiles have been interpolated to a common vertical pressure grid, with a bin of 25 hPa. The two-hourly reanalysis and forecast model (i.e., control run) fields were linearly interpolated to the time and location of each measurement, with a bin of 25 hPa, and then compared with the measurements. For the ACCMIP models, the monthly model outputs were compared with the measurements at the location of each measurement. The averaged profile is computed globally and for four latitudinal bands, SH extratropics (90--30{\degree}\,S), SH

削除: validation

削除: .

削除: and model

削除: using the two-hourly output data

削除: .

削除: five

tropics (30{\degree}\,S--Equator), NH tropics (Equator--30{\degree}\,N), and NH extratropics (30--90{\degree}\,N).

削除：SH high latitudes (55$^\circ$S--90$^\circ$S), SH mid-latitudes (15$^\circ$S--55$^\circ$S), tropics (15$^\circ$S--15$^\circ$N), NH mid-latitudes (15$^\circ$N--55$^\circ$N), and NH high latitudes (55$^\circ$N--90$^\circ$N)

\subsection{Ozonesonde sampling bias estimation}

The current ozonesonde network does not cover the entire globe and is not homogeneously distributed between the hemispheres, ocean and land, and urban and rural areas. Also, the sampling interval of ozonesonde observations is typically a week or longer, which does not reflect the influence of diurnal and day-to-day variations. Model errors are also expected to vary greatly in time and space at various scales. Therefore, the implications of model differences at ozonesonde locations to regional and seasonal processes is uncertain. Thus, we evaluate how changes in evaluated model performance could be obtained by using the complete sampling chemical reanalysis fields instead of the existing ozonesonde network on simulated regional ozone fields.

Sampling bias is an error in a computed quantity that arises due to unrepresentative (i.e., insufficient or inhomogeneous) sampling, which induces spurious features in the average estimates (e.g., Aghedo et al., 2011; Foelsche et al 2011; Toohey et al., 2013; Sofieva et al., 2014) and long-term trends (Lin et al., 2016). Sampling bias may occur when the atmospheric state within the time-space domain over which the average is calculated is not uniformly sampled. In regions where variability is dominated by short-term variations, limited sampling may lead to a random sampling error. The primary technique for sampling bias estimation is to subsample model or reanalysis fields based on the sampling patterns of the measurements and then to quantify differences between the mean fields based on the measurement sampling and those derived from the complete fields. Sampling bias cannot be negligible, even for satellite measurements (Aghedo et al. 2011; Toohey et al., 2013; Sofieva et al., 2014).

To estimate sampling biases of the ozonesonde network in the ACCMIP model evaluation, two evaluation results of mean model bias are compared using the chemical reanalysis. The first evaluation was conducted based on the complete sampling; the second evaluation used the ozonesonde sampling (in both space and time) that is based on the compilation by Tilmes

et al. (2012). By using the two-hourly reanalysis fields, we can address possible biases due to the limited model sampling (i.e., monthly ACCMIP model outputs were used). Note that the relatively coarse horizontal resolution of the reanalysis may lead to an underestimation of the sampling bias in the model evaluation, because the variability of a sampled field depends on the resolution of the measurement. Tilmes et al. (2012) stated that regional aggregates of individual ozonesonde measurements with similar characteristics are more representative for larger regions; however, this may not mean that evaluation results using the compiled data generate model errors that are representative of actual monthly mean for a surrounding area.

\section{Consistency between chemical reanalysis and ozonesonde observations}

Miyazaki et al. (2015) validated an older version of the reanalysis (http://www.jamstec.go.jp/res/ress/kmiyazaki/reanalysis/) and showed good agreement with independent observations such as ozonesonde and aircraft measurements on regional and global scales and for both seasonal and year-to-year variations from the lower troposphere to the lower stratosphere for the 2005-2012 period. The mean bias against the ozonesonde measurements in the older dataset is -3.9 ppb at the NH high-latitudes (55--90{\degree}\,N), -0.9 ppb at the NH mid-latitudes (15--55{\degree}\,N), 2.8 ppb in the tropics (15{\degree}\,S--15{\degree}\,N), -1.0 ppb at the SH mid-latitudes (55--15{\degree}\,S), -1.7 ppb at the SH high-latitudes (90--55{\degree}\,S) between 850 and 500 hPa (Miyazaki et al., 2015). Since the updated reanalysis ozone fields used in this study have not yet been validated in any publication, we first present the evaluation results of the chemical reanalysis using global ozonesonde observations for 2005--2009.

削除：A major update from the system used in Miyazaki et al. (2015) to the system used in this study is the replacement of forecast model from CHASER (Sudo et al., 2002) to MIROC-Chem (Watanabe et al., 2011), which caused substantial changes in the a priori field and thus the data assimilation results of various species. In addition, we attempt to optimize the surface \chem{NO_x} emission diurnal variability using data assimilation of OMI, SCIAMACHY, and GOME-2 retrievals in the updated system (Miyazaki et al., 2016).

Figs 1 and 2 compare the reanalysis and the global ozonesonde observations, and the comparison result is summarized in Table 3. In order to confirm improvements in the reanalysis, results from a model simulation without any chemical data assimilation (i.e., a control run) is also shown. The control run shows systematic biases, such as positive biases in the upper troposphere and lower stratosphere (UTLS) throughout the globe and negative biases in the lower and middle troposphere in the extratropics of both hemispheres. The positive bias in the UTLS is larger in the Southern Hemisphere (SH) than in the Northern

削除：1

Hemisphere (NH). The a priori systematic bias in this study is larger than that in our previous study (Miyazaki et al., 2015) in the UTLS, because of different model settings, such as the upper boundary conditions of \chem{NO_y}, \chem{Cl_y}, and \chem{Br_y}. However, the reanalysis fields were less sensitive to the a priori profiles in the UTLS than in the lower and middle troposphere because of strong constraints by MLS measurements and long chemical lifetime of ozone in the UTLS.

The reanalysis shows improved agreements with the ozonesonde observations over the globe for most cases. The data assimilation removed most of the positive bias in the UTLS throughout the year and reduced the negative bias in the lower and middle troposphere in the extratropics. In the NH extratropics in the lower and middle troposphere, the data assimilation reduced the annual mean negative bias of the forecast model by 55 \%, which is attributed to the reduced bias in boreal spring--summer. The mean bias in the new reanalysis dataset is smaller than that in the older reanalysis dataset (Miyazaki et al., 2015) for most cases (e.g., from -3.9 to -2.9 ppb at the NH high-latitudes (55--90{\degree}\,N), -0.9 to -0.1 ppb at the NH mid-latitudes (15--55{\degree}\,N), -1.0 to -0.1 ppb at the SH mid-latitudes (55--15{\degree}\,S) between 850 and 500 hPa). The mean bias in the new dataset is less than 0.9 ppb at the tropics and mid-latitudes between 500 and 200 hPa (not shown). The simultaneous optimization of concentrations and emissions played important roles in improving the lower tropospheric ozone analysis, associated with the pronounced ozone production caused by \chem{NO_x} increases, as demonstrated by Miyazaki et al. (2015). This advantage increases the ability of the chemical reanalysis to evaluate the simulated tropospheric ozone profiles, including the lower tropospheric ozone concentrations. Root-Mean-Square-Errors (RMSEs) are also reduced above the middle troposphere, although the reduction rate is relatively small compared to the bias, probably due to representativeness errors between the ozonesonde measurements and data assimilation analysis. The tropospheric concentrations show distinct seasonal and year-to-year variations, for which the temporal correlation based on the monthly and regional mean concentrations is increased by the data assimilation globally, except at high latitudes in the lower troposphere (Table 3). The reanalysis can be extended to a longer-term validation that will provide more information on seasonality and year-to-year variability.

\section{Evaluation of ACCMIP models}



削除: the entire troposphere

削除: At

削除: mid and high latitudes

削除: 45--90

削除: .

削除:

削除: 1

削除: Note that a longer-term validation could be useful, in order to obtain more robust information about the seasonality and year-to-year variability.

書式変更: フォント:斜体（なし）

書式変更: フォント:斜体（なし）

書式変更: フォント:斜体（なし）

\subsection{Global distribution}

We use the global chemical reanalysis to evaluate the global ozone profiles in ACCMIP simulations. Fig. 3 compares the global distribution of the annual mean ozone concentration between the five-year mean reanalysis and the ensemble mean of the ACCMIP models. The average over the multiple models can be expected to improve the robustness of the model simulation results, because some parts of the model errors may cancel each other out. As summarized in Table 4, the global spatial distributions are similar between the five-year mean reanalysis field and the ensemble mean when estimated at 2$^\circ \times$2.5$^\circ$ spatial resolution, with a spatial correlation (r) greater than 0.94 from the lower troposphere to the lower stratosphere, except for the NH extratropical middle troposphere (r=0.57). The reanalysis and multi-model mean commonly reveal distinct inter-hemispheric differences, associated with a stronger downwelling across the tropopause and stronger emission sources of ozone precursors in the NH. The wave-1 pattern in the zonal ozone distribution in the tropics, with a minimum over the Pacific Ocean and maximum over the Atlantic (Thompson et al., 2003; Bowman et al, 2009; Ziemke et al., 2011), can also be commonly found in the reanalysis and the multi-model mean and was also suggested by Young et al. (2013).

Large errors between the reanalysis and the multi-model mean in the troposphere are found in the NH extratropics and SH tropics (right panel in Fig. 3). The multi-model mean overestimates the zonal and annual mean concentrations by 6--11 ppb at 800 hPa and by 2--9 ppb at 500 hPa in the NH extratropics. The overestimation is larger over the oceans than over land at the NH mid-latitudes at 800 hPa. Both the mean RMSE and bias are larger at 800 hPa than at 500 hPa in the NH extratropics, whereas they are larger at 500 hPa in the NH tropics (Table 4). In the SH tropics, the multi-model mean underestimates the concentration over the eastern Pacific by up to 9 ppb, over the Atlantic by up to 18 ppb, and over the Indian Ocean by up to 8 ppb at 500 hPa. These negative biases are larger in the middle troposphere than in the lower troposphere for most places and also for the zonal means in the SH tropics (-15 \% in the middle troposphere and -10 \% in the lower troposphere) (Table 4). Young et al. (2013) consistently revealed the positive bias in the NH and negative bias in the SH using OMI/MLS tropospheric ozone column measurements. At 200 hPa, the multi-model mean underestimates

削除：2

削除：.

削除：2

削除：2

削除：The

削除：were common reported

削除：(Young et al., 2013)

the zonal mean concentration by 20--30 ppb at high latitudes in both hemispheres, with a larger error in the SH than in the NH (Table 4).

Fig. 4 shows the Taylor diagram of the ACCMIP models against the reanalysis for three latitudinal bands for three levels. The relevant statistics at 500 hPa are summarized in Table 5, for which the tropics is separated into two hemispheres. In the NH extratropics at 800 hPa, most models reproduced the spatial distribution (r = 0.8--0.95), while underestimating the spatial standard deviation (SD) by up to 50 \%. Three exceptional models (1, 7, 8) show relatively poor agreements (r = 0.45--0.6 and SD underestimations by 50--60 \%). At 500 hPa, there is a large diversity in the agreement. Only a few models (2, 4, 9, 11) show close agreement with the reanalysis (r > 0.8, SD error < 20 \%). Notably, two models (12, 15) reveal too large spatial variabilities (SD error > 80 \%), and five models (1, 6, 7, 8, 12) reveal small spatial correlation (r < 0.15). The regional mean bias is largely positive (> 10 ppb) in several models (7, 8, 12) (Table 5). In the NH extratropics in the lower and middle troposphere, ozone distributions are modified by various processes, including vertical transport by convection and along conveyor belts, inflow from the stratosphere, long-range transports, and photochemical production (e.g, Lelieveld and Dentine, 2000; Oltmants et al., 2006; Sudo and Akimoto, 2007; Jonson et al., 2010). The evaluation results indicate that these processes occur differently among models. At 200 hPa, all the models simulate well the spatial distribution (r > 0.95), whereas the spatial variability differs between the models (SD error ranges from -50 \% to +30 \%). There is relatively large variation in the stratospheric concentration, which results in the diversity in the UTLS, as also discussed by Young et al. (2013).

In the tropics, the spatial correlation is greater than 0.8 at all levels for most models (except for 12, 15), as they capture the wave-1 structure. When dividing the tropics into two hemispheres (Table 5), only a few models (4, 12) reveal low spatial correlation (r < 0.8) for the SH tropics (30$^\circ$S--EQ) at 500 hPa. The spatial correlation in the tropics is lower at 500 hPa than at 800 hPa for most models. The SD error is less than 40 \% for all the models at 800 and 500 hPa, while mostly overestimating the spatial variability at 800 hPa by up to 30 \%. The mean bias is negative for most models at 500 hPa in the tropics in both hemispheres, with larger negative biases in the SH tropics (Table 5). Young et al (2013) noted that

削除: 2

削除: 3

削除: 3

削除: In this region

削除: .

削除: 3

削除: 3

correlations between the biases for the NH and SH tropical tropospheric columns are strong. Similarly, our analysis using the reanalysis reveal a high correlation (0.91) between the NH and SH tropical biases at 500 hPa, suggests that similar processes are producing the model biases in the tropical middle troposphere between the hemispheres. For instance, biomass burning emissions are handled differently across the models, which may lead to differences in ozone simulations in the tropics (Anderson et al., 2016). At 200 hPa in the tropics, the SD error differs among models, which could primarily be associated with the different representations of convective transports and ozone production by \chem{LNO_x} sources (e.g., Lelieveld and Crutzen, 2007; Wu et al., 2007).

In the SH extratropics at 800 hPa, most models reproduce the spatial distribution ($r > 0.9$), while underestimating the SD by 15--70 \%, except for model 15. The model performance is similar between 800 hPa and 500 hPa, with a smaller SD error at 500 hPa for most models. These high spatial correlations may be related to a lack of local precursor emissions in the SH. At 500 hPa, a majority of the models underestimate the mean concentration (Table 5), with large negative biases (< -8 ppb) in several models (1, 2, 12, 14). At 200 hPa, the SD error varies from -80 \% to + 65 \%. The large diversity at 200 hPa may be related to the different representation of the tropopause and stratosphere--troposphere exchange (STE) among models.

\subsection{Seasonal variation}

Fig. 5 compares the seasonal variation of zonal mean ozone concentration between the ACCMIP models, the reanalysis, and ozonesonde observations. The comparison between the reanalysis concentrations sampled at ozonesonde sites/time (black dashed line) and the ozonesonde observations (blue solid line) shows that the reanalysis is in close agreement with the ozonesonde observations over the globe, as described in Sec. 3. However, in the NH extratropics at 800 hPa, the reanalysis concentration is too low from boreal spring to summer by up to 4 ppb, which leads to an underestimation of the seasonal amplitude (as estimated from the difference between maximum and minimum monthly mean concentrations). In the NH tropics at 500 hPa, the reanalysis overestimates the concentration except in April. In the SH tropics at 500 and 800 hPa, the reanalysis slightly overestimates the concentrations throughout the year by up to 5 ppb. In the SH extratropics at 800 hPa, the reanalysis

書式変更: フォント:斜体 (なし)
書式変更: フォント:斜体 (なし)
書式変更: フォント:斜体 (なし)
削除: 3
削除: The poor agreement in model 8 is attributed to too-high concentrations at SH mid-latitudes and too-low concentrations at SH high latitudes in austral spring.
削除: from the ozonesonde sampling

concentration is too low by up to 5 ppb from austral autumn to winter. The reanalysis concentration and seasonal variation differs largely between the complete sampling (black bold line, where the concentrations were averaged over all grid points) and the ozonesonde sampling (black dashed line) for the globe. The impact of using the reanalysis instead of the ozonesonde network in characterizing the ozone seasonal variation is discussed in Section 5.

The global ozone concentrations averaged over all grid points with area weights are compared between the ACCMIP models and the reanalysis (black solid line vs. red solid line for the multi-model mean and thin colored lines for individual models). There is considerable interannual variability in both the reanalysis and the ACCMIP models. We confirmed that the ACCMIP ensemble mean is mostly within the standard deviation (i.e., year-to-year variation) of the reanalysis (not shown). In the NH extratropics, the multi-model mean overestimates the monthly mean concentrations by 6--9 ppb at 800 hPa and by 3--6.5 ppb at 500 hPa. The multi-model mean reproduces the seasonal variation, whereas there is large diversity among the models. The increase from winter to spring differs among models at 500 hPa, which is probably associated with different representations of downwelling from the stratosphere. Fig. 6 compares the seasonal amplitude. Most models overestimate the seasonal amplitude in the NH lower and middle troposphere, with a mean overestimation of 50--70 \% at 800 hPa and 25--40 \% at 500 hPa at NH high latitudes. At 200 hPa, the multi-model annual mean concentration is in good agreement with that of the reanalysis, whereas the seasonal amplitude is underestimated by most models at NH high latitudes, with a mean underestimation of 15--25 \%.

In the NH tropics at 500 hPa, the multi-model mean underestimates the concentration by 1--4 ppb throughout the year, which can be attributed to the anomalously low concentrations in several models. There is a large diversity among the models in this region. In the SH subtropics, the multi-model mean is lower by up to 5 ppb at 800 hPa and by up to 11 ppb at 500 hPa, with the largest errors occurring in austral spring. A majority of models overestimate the seasonal amplitude in the NH subtropics at 800 hPa (by about 10--40 \%), whereas they mostly underestimate the amplitude in the SH tropics at 800 and 500 hPa. In the tropical upper troposphere in both hemispheres, a few models reveal anomalously high or low concentrations. Both the ozonesondes and reanalysis reveal a sharp increase in ozone between

March and April in the NH subtropics, which is not captured in the multi-model mean, as suggested by Young et al. (2013).

In the SH extratropics, the multi-model mean and the reanalysis are in good agreement at 800 hPa, whereas it largely underestimates the peak concentration in austral winter--spring at 500 hPa (by up to 7 ppb) and 200 hPa (by up to 35 ppb). The large diversity among the models and the large underestimation in the multi-model mean at 500 hPa in spring could be attributed to the differing influence of stratospheric air. The seasonal amplitude is overestimated at 800 and 200 hPa by most models at SH high-latitudes.

\subsection{Inter-hemispheric gradient}

Fig. 7 compares the inter-hemispheric gradient (NH/SH ratio) of the annual mean ozone concentration. We calculated the gradient of area-weighted ozone concentrations across the equator; however, recognize a more careful definition of the boundary between two hemispheres would be required to isolate air masses originated from each hemisphere (e.g., Hamilton et al., 2008). For the estimation of the gradient using the ozonesonde observations, we made a gridded dataset from the ozonesonde observations based on the completion by Tilmes et al (2012) at 2$^\circ \times$2.5$^\circ$ spatial resolution, and then calculated area-weighted hemispheric mean concentrations using the gridded dataset. The gradient is similar between the ozonesonde observations (blue solid line) and the reanalysis concentration from the ozonesonde sampling (black dashed line) throughout the troposphere. In these estimates, the NH mean concentration is higher than the SH mean by 60--70 \% in the lower troposphere, by 30--40 \% in the middle troposphere, and by 55--60 \% around 200 hPa. Near the surface, the reanalysis slightly overestimates the NH/SH ratio, mainly because of overestimated concentrations at the NH mid-latitudes.

By taking a complete sampling in the reanalysis (i.e., averaging over all model grid points for each hemisphere) (black solid line), the NH/SH ratio becomes smaller by about 25--30 \%, 7--10 \%, and 15--25 \% in the lower troposphere, the middle troposphere, and around 200 hPa, respectively, compared to the average at the ozonesonde sampling sites (black dashed line).

削除: .

The difference is a consequence of ozonesonde stations located near large cities at NH mid-latitudes, and therefore tend to observe higher ozone concentration than the hemispheric average. At around 200 hPa, the difference could also be attributed to the presence of atmospheric stationary waves and Asian monsoon circulation in the NH, which result in substantial spatial ozone variations in the UTLS (e.g., Wirth, 1993; Park et al., 2008) (c.f., Fig. 3). The annual mean NH/SH ratio based on the global reanalysis field estimated at the surface, 800 hPa, 500 hPa, and 200 hPa are 1.36, 1.42, 1.30, and 1.35, respectively.

Most models overestimate the NH/SH ratio compared with the reanalysis, with a mean overestimation (black solid line vs. red solid line) of 34 \% at the surface and 22--30 \% in the free troposphere, attributing to both too-high concentrations in the NH extratropics and too-low concentrations in the SH subtropics in most models (c.f., Figs. 3 and 5). The multi-model mean reveals annual mean NH/SH ratios of 1.71, 1.73, 1.54, and 1.49 at the surface, 800 hPa, 500 hPa, and 200 hPa, respectively. The large systematic error in the NH/SH ratio suggests that, for instance, the inter-hemispheric distribution of radiative heating due to tropospheric ozone in chemistry--climate simulations are largely uncertain in most models, and such comprehensive information for different altitudes in the troposphere cannot be obtained using any individual measurements, as is further discussed in Section 6.3.

\section{Impact of sampling on model evaluation}

As presented in the previous section, the chemical reanalysis provides comprehensive information on global ozone distributions for the entire troposphere which is useful for validating global model performance. It was also demonstrated that the inter-hemispheric gradient of ozone measured with the ozonesonde and complete sampling method produced different results, and the model-reanalysis difference strongly depended on the choice of the sampling method. As these networks have been the primary basis for CCM evaluation (e.g., Stevenson et al., 2006; Huijnen et al., 2010; Young et al., 2013), the implications of this sampling bias need to be quantified.

削除：Around

削除：distribution

This section evaluates how changes in evaluated model performance could be obtained by using the complete sampling chemical reanalysis fields instead of the existing ozonesonde network on simulated regional ozone fields.

The model evaluation results are shown for the 11 regions illustrated in Fig. 8 and summarized in Table 6. Japan was excluded from the evaluation because data from only one station was available for the reanalysis period. The 11 areas surrounding the ozonesonde stations were considered for complete atmospheric sampling (rectangles in Fig. 8), for which small margins were considered around the stations to prevent overestimation of the ozonesonde network limitation. It was confirmed that the discrepancy between the two evaluations generally increases with the size of the area. In contrast, for the SH mid- and high latitudes, the defined areas cover the entire range of longitudes, because of generally less variabilities in the SH than in the NH. Four latitude bands (90--30{\degree}\,S, 30{\degree}\,S--Equator, Equator--30{\degree}\,N, 30--90{\degree}\,N) were also considered in the sampling bias evaluation.

The reality of the reanalysis fields is important for reasonable estimates of the true sampling bias of the real atmosphere. As discussed in Section 3, there is good agreement in the evaluated model performance using the reanalysis and the ozonesonde measurements at the ozonesonde sampling, except for the lower troposphere. This result supports the use of the reanalysis data at the ozonesonde locations. The performance of the ACCMIP model as compared with the ozonesonde measurements is mostly consistent with that shown by Young et al. (2013), although the ozonesonde data periods differ -- 1997-2011 was used by Young et al (2013) and 2005-2009 was used in this study.

Table 7 demonstrates the regional and seasonal mean differences of the reanalysis concentrations between the complete sampling and the ozonesonde sampling. The ozonesonde sampling results have higher concentrations (by about 3 \%) in the two NH polar regions for most cases, whereas the difference is smaller in NH polar west than in NH polar east. Among the NH mid-latitude regions, a large difference (about 14 \%) exists between the

削除: The current ozonesonde network does not cover the entire globe and is not homogeneously distributed between the hemispheres, ocean and land, and urban and rural areas. Also, the sampling interval of ozonesonde observations is typically a week or longer, which does not reflect the influence of diurnal and day-to-day variations. Model errors are also expected to vary greatly in time and space at various scales. Therefore, the implications of model differences at ozonesonde locations to regional and seasonal processes is uncertain.

削除: Sampling bias is an error in a computed quantity that arises due to unrepresentative (i.e., insufficient or inhomogeneous) sampling, which induces spurious features in the average estimates (e.g., Aghedo et al., 2011; Foelsche et al 2011; Toohey et al., 2013; Sofieva et al., 2014). Sampling bias may occur when the atmospheric state within the time-space domain over which the average is calculated is not uniformly sampled. In regions where variability is dominated by short-term variations, limited sampling may lead to a random sampling error. The primary technique for sampling bias estimation is to subsample model or reanalysis fields based on the sampling patterns of the measurements and then to quantify differences between the mean fields based on the measurement sampling and those derived from the complete fields. Sampling bias cannot be negligible, even for satellite measurements (Aghedo et al. 2011; Toohey et al., 2013; Sofieva et al., 2014).

削除: To estimate sampling biases of the ozonesonde network in the ACCMIP model evaluation, two evaluation results of mean model bias are compared using the chemical reanalysis. The first evaluation was conducted based on the complete sampling; the second evaluation used the ozonesonde sampling (in both space and time) that is based on the completion by Tilmes et al. (2012). By using the two-hourly reanalysis fields, we can address possible biases due to the limited model sampling (i.e., monthly model outputs were used). Note that the relatively coarse resolution of the reanalysis may lead to an underestimation of the sampling bias in the model evaluation, because the variability of a sampled field depends on the resolution of the measurement. Tilmes et al. (2012) stated that regional aggregates of individual ozonesonde measurements with similar characteristics are more representative for larger regions; however, this may not mean that evaluation results using the compiled data generate model errors that are representative of actual monthly mean for a surrounding area.

削除: 3

削除: SH

削除: west

削除: from March--May (MAM) to September--November (SON)

削除: east

削除: west

two cases over the eastern United States in June--August (JJA), where the comparison using monthly reanalysis fields sampled at the ozonesonde locations (brackets in Table 7) suggests that the sampling bias is dominated by temporal variations. The tropical and subtropical regions exhibit large sampling biases, 4--12.3 \% over the NH subtropics, -3.2--5.0 \% over the Western Pacific and East Indian Ocean, 0--7.8 \% over the equatorial Americas, and -3.8--7.5 \% over the Atlantic Ocean and Africa. In most of the tropical and subtropics regions, both the spatial and temporal sampling biases are important, because of large spatial and temporal variability of ozone and the sparse observation network. For the global tropics, the sampling bias reaches 13 \% in the NH (Eq--30{\degree}\,N) and 8 \% in the SH (30{\degree}\,S--Eq). Thus, the ozonesonde network has a major limitation when it comes to capturing ozone concentrations that are representative of seasonal and regional means for the entire tropical region. The sampling bias may not be negligible even in the SH (0.3--3.9 \% in the SH mid-latitudes and 0.8--4.2 \% in the SH high latitudes), and it is large (up to 13 \%) when estimations are done for a large area (90--30{\degree}\,S). The large sampling bias in 90--30{\degree}\,S is primarily attributed to spatial variability. The impact of the sampling bias on the model evaluation is discussed in the following section.

\subsection{Mean error and its distribution}

The model evaluation results differ greatly for many regions between the complete sampling and the ozonesonde sampling, as shown by Fig. 9 and summarized in Table 8. The sampling bias is evaluated using the median of the multiple models to provide robust estimates of the model performance. For the NH Polar Regions, Tilmes et al. (2012) stated that separating the regions into eastern and western sectors reduces the variability in ozone within each region because long-range transports of pollution from low and mid-latitudes into high latitudes shows longitudinal variations in the NH (e.g., Stohl, 2006). Comparisons further suggest that, except for the UTLS in winter (December--February (DJF)), the evaluated model performance using the ozonesonde measurements are representative of the surrounding regional and seasonal mean model performance. For the two NH polar regions at 200 hPa in DJF, the validation based on the ozonesonde sampling reveals a large negative sampling bias in the model bias as compared with regional and monthly means. Large negative model biases against the ozonesonde observations have been reported by Young et al. (2013) for 250 hPa

削除：

削除：cannot

削除：for the two cases

削除：5

削除：

(by about -13 \% for the NH polar west and -18 \% for the NH polar east for the annual mean concentration), whereas results from this study suggest that these errors based on the ozonesonde sampling (by -14 \% for the NH polar west and -18 \% for the NH polar east in DJF in our estimates) are larger than those from regional and seasonally representative model bias (by -3 \% and +5 \%, respectively). At 500 hPa, the ozonesonde network reveals a negative sampling bias for the NH polar east in DJF. Thus, the positive bias reported in Young et al. (2013) for the NH polar east at 500 hPa may be lower than regional and seasonally representative model biases. Our analysis using monthly reanalysis fields sampled at the ozonesonde locations (brackets in Table 8) suggests a greater impact of the spatial sampling bias than the temporal sampling bias for the NH polar east in DJF. The large discrepancy between the two estimates in the UTLS model performance can be attributed to the large variability of ozone distribution and associated model errors on a regional and seasonal scale.

For Canada, large differences (>30 \%) exist in the two evaluations in the lower troposphere and for the UTLS in DJF and for the middle troposphere in March--May (MAM). The ozonesonde measurements reveal a large negative sampling bias in the model evaluation in DJF at 200 hPa (-4 \% in the complete sampling and -25 \% in the ozonesonde sampling), while they reveal a negative sampling bias (by about 50 \%) at 500 hPa in MAM. At 500 hPa over Canada, the relative importance of the spatial and temporal sampling biases varies with season: the spatial (temporal) sampling bias is dominant in DJF (JJA), whereas both of them are important in MAM. Similar differences between the two evaluations are found for Western Europe at 500 hPa and at 200 hPa in DJF. These results suggest that, for instance, the positive bias for Western Europe estimated by Young et al (2013) may be lower than regional and seasonally representative model bias, even for such a small area. The smaller discrepancy between the two estimates for Western Europe as compared for Canada for most cases could be associated with the better coverage of the ozonesonde measurements for Western Europe. Even for the small area of the eastern United States, the two validations differ largely in the UTLS (e.g., -9 \% in the ozonesonde sampling and +6 \% in the complete sampling at 200 hPa in MAM) and at 500 hPa in MAM, JJA, and September--November (SON). In the NH subtropics, the two evaluations disagree largely in the middle and upper troposphere in JJA and SON.

削除:

削除: .

削除: .

削除: June--August (

削除: )

The tropical stations were separated into the three sub-regions: Western Pacific and East Indian Ocean, equatorial America, and the Atlantic Ocean and Africa. These regions reflect the different dominant tropical processes including biomass burning and lightning over the Atlantic and Africa. The large variability of tropical ozone and its associated model error, together with the sparse ozonesonde network in these regions, results in large discrepancies between the two evaluations in the tropical regions. At 500 hPa, the ozonesonde measurements reveal a large (by 40--50 \%) negative sampling bias in MAM and a positive sampling bias in DJF over the Western Pacific and East Indian Ocean, whereas it shows a large negative sampling bias (by 110 \%) in MAM over the equatorial Americas. Over the Western Pacific and East Indian Ocean, the sampling bias is not reduced by using monthly mean reanalysis fields (sampled at the ozonesonde locations) in DJF and JJA. This suggests that ozone varies with time and space in a complex manner, and a dense (in both space and time) network would be required to capture the regional and seasonally representative model biases in this region. The probability distribution function (PDF) estimated using monthly mean reanalysis and model fields also differs largely between the two samplings (Fig. 10). Over the Western Pacific and East Indian Ocean in SON at 500 hPa, the multi-model mean shows a sharp peak around 54--58 ppb, in contrast to the broad distribution seen in the reanalysis with two peaks around 65 ppb and 35--45 ppb for the complete sampling (left bottom panel in Fig. 10). This information is useful to characterize model errors and for process-oriented model validation. On the other hand, the validation based on the ozonesonde sampling (left top panel) does not show any clear pattern and does not support model evaluation. Note that the influence of inter-annual variability was not considered in the analysis because the monthly climatological data were used by averaging over ten years for the models and five years for the reanalysis.

削除: March--May (

削除: )

削除: except in JJA

Although the variability of ozone is generally smaller in the SH than in the NH because of smaller local precursor emissions, large sampling biases exist even at SH mid- and high-latitudes due to the sparse ozonesonde network. In the SH mid latitudes, for example, the sign of the evaluated bias is opposite between the two cases at 200 hPa in DJF (-2.8 ppb in the complete sampling and +25.1 ppb in the ozonesonde sampling). In the SH high latitudes, evaluation results differ largely throughout the year in the middle troposphere. The temporal

sampling bias mostly dominates the difference in the SH high latitudes in MAM and JJA, whereas the spatial sampling bias is also important in the SH mid latitudes in DJF and MAM. Based on the complete sampling, the ozone PDF is broadly distributed with a peak around 38 ppb at 500 hPa in SON at the SH high latitudes (right bottom panel in Fig. 10), while the multi-model mean underestimates high concentrations (>47 ppb) and shows a sharp peak of about 35 ppb. The PDF generated by the ozonesonde sampling does not provide a strong information on the distribution of the ozone (right top panel). These results highlight the advantage of using the reanalysis data for evaluating regional and seasonally representative model performance, and for characterizing these distributions.

Table 8 also shows the model evaluation results for four latitudinal bands at 500 hPa. The observations used are shown in bold in Table 2. The differences between the two evaluations are small in the NH extratropics (30--90{\degree}\,N) in all seasons, because of the relatively large number of observations. There are large differences in the tropics of both hemispheres: the ozonesonde network reveals a large negative sampling bias in the model evaluation in the NH tropics (Eq--30{\degree}\,N) in SON (-9 \% in the complete sampling and -16 \% in the ozonesonde sampling) and in the SH tropics (30{\degree}\,S--Eq) in MAM (-14 \% and -21 \%) and a large positive sampling bias in the NH tropics in JJA (-7 \% and -3 \%). Large sampling biases (> 60 \%) also exist in the SH extratropics (90--30{\degree}\,S) in DJF and MAM due to the sparse ozonesonde network.

Further, ozonesonde sampling bias is evaluated for the control run and reanalysis comparisons. As summarized in Table 9, at 500 hPa, there are large differences (> 30 \%) between the two evaluations in many regions, especially in the NH mid latitude regions in winter and in the tropics throughout the year, as also found in the ACCMIP models and reanalysis comparisons (Table 8). The analysis increments introduced by data assimilation vary with space and time, reflecting the changes in coverage and uncertainty of assimilated measurements as well as in model errors. Nevertheless, observational information was propagated globally and integrated with time through forecast steps during the data assimilation cycles. This is true for ozone because of its relatively long lifetime in the free troposphere. Therefore, the spatial distribution is well constrained by data assimilation, and we do not expect large variations in the reanalysis quality within each analysis region.

削除:

削除: 5

削除: (>30 \%)

\subsection{Seasonal variation}

The seasonal cycle of tropospheric ozone is determined by various factors such as local photochemical production and atmospheric transport (e.g., Monks, 2000). Carslaw (2005), Bloomer et al. (2010), and Parrish et al. (2013) found multi-decadal changes in the amplitude and phase of the seasonal cycle at NH mid-latitudes. It was suggested that these changes can be attributed to changes in atmospheric transport patterns combined with spatial and temporal changes in emissions. CTMs have been used to explore the causal mechanisms; however, they failed to simulate several important features of the observed seasonal cycles (e.g., Ziemke et al., 2006; Stevenson et al., 2006; Parrish et al. 2014; Young et al., 2013). Accurate validation of the seasonal cycle is thus important for evaluating general model performance.

Table 10 compares the relative error in the seasonal amplitude obtained from the multi-mean model with that of the reanalysis for the complete and ozonesonde samplings. The evaluation based on the ozonesonde sampling results in a larger overestimation of the seasonal amplitude in the NH lower troposphere for most regions (+13.4--+63.4 \% in the sonde sampling and -19.0--+40.2 \% in the complete sampling). The large discrepancies can be attributed to large spatial variability in the seasonal variations of ozone and its model errors within each defined region and also the existence of short-term variability that is not completely captured by the ozonesonde sampling. For the Eastern US and Western Europe at 800 hPa, the sign of the bias is opposite between the two estimates. In contrast, at 200 hPa in the NH, results between the two evaluations are similar, suggesting spatial homogeneity in the seasonal cycle and its model errors within each region in the NH. Because the seasonal variations differ among different regions, the seasonal amplitude estimated for the entire NH extratropics (30--90{\degree}\,N) is largely different between the two estimates throughout the troposphere.

In the tropics, the estimated errors of the seasonal amplitude largely differ between the two samplings throughout the troposphere, suggesting that information obtained from the sparse ozonesonde network cannot be applied to characterize regional model errors in the seasonal

削除：6

削除：cases

削除：region

cycle, even within the small defined area. The sampling bias in the seasonal amplitude estimated for the entire tropics is larger than 60 \% throughout the troposphere both in the NH (Eq--30{\degree}\,N) and SH (30{\degree}\,S--Eq). Because of the large spatial variability, detailed validations using the chemical reanalysis (e.g., for each grid point) would be helpful. Also, in the SH high latitudes, large disagreements in the seasonal amplitude exist at 800 and 200 hPa.

削除: al

\section{Discussions}

\subsection{Reanalysis uncertainty}

Although the reanalysis dataset provides comprehensive information for global model evaluations, its performance still needs to be improved, especially for the lower troposphere, as also discussed by Miyazaki et al. (2015). Performance can be improved by ingesting more datasets including meteorological sounders such as IASI (Clerbaux et al., 2009), AIRS (Chahine et al., 2006), and CrIS (Glumb et al., 2002). Application of a bias correction procedure for multiple measurements, which is common in numerical weather prediction (e.g., Dee, 2005), is needed to improve reanalysis accuracy. Recently developed retrievals with high sensitivity to the lower troposphere (e.g. Deeter et al., 2013; Fu et al., 2016) and the optimization of additional precursor emissions would be helpful to improve analysis of the lower troposphere. The relatively coarse resolution of the model could cause large differences between the simulated and observed concentrations at urban sites and may degrade the reanalysis.

The statistical information obtained from the reanalysis and the multi-model simulations can be used to suggest further developments for the models and observations. The analysis ensemble spread from EnKF can be regarded as uncertainty information about the analysis mean fields, indicating requirements for additional observational constraints. As shown in Fig. 11 (left panels), the relative reanalysis uncertainty is large over the tropical areas of the oceans at 800 hPa (>20 \%), over the Southern Ocean at 500 hPa (10--20 \%), and over the tropics of the Pacific Ocean and the Antarctic at 200 hPa (>16 \%). Conversely, the reanalysis

uncertainty is small from the tropics to mid-latitudes in both hemispheres at 500 hPa (<11 \%). Miyazaki et al (2015) investigated that the analysis spread is caused by errors in the model input data, model processes, and assimilated measurements, and it is reduced if the analysis converges to a true state. The analysis spread is smaller in the extratropical lower stratosphere than in the tropical upper troposphere at 200 hPa, because of the high accuracy of the MLS measurements. In contrast, in the middle troposphere, the analysis spread is generally smaller in the tropics than the extratropics because of the higher sensitivities in the TES retrievals. Note that the data assimilation setting influences the analysis uncertainty estimation in the reanalysis. In particular, the analysis spread was found to be sensitive to the choice of ensemble size (Miyazaki et al., 2012b). A large ensemble size is essential to capture the proper background error covariance structure (i.e., analysis uncertainty).

削除： These variations may be related to changes in observation errors, the number of assimilated measurements, as to model errors.

削除： 。

The five-year reanalysis (2005--2009) may cause biases in the estimated model errors in the evaluation of the 2000 decade ACCMIP simulations that used decadal-averaged SST boundary conditions and biomass burning emissions averaged over 1997--2006 (Lamarque et al., 2010). It may neglect the influences of interannual and decadal changes in both anthropogenic and biomass emissions and meteorology. Longer-term reanalysis and time-consistent validation are required to obtain more robust error estimations.

書式変更： フォント:斜体（なし）

\subsection{Model uncertainty}

The variability across the ensemble models (i.e., ensemble spread) identifies where the models are most consistent or uncertain (center panels in Fig. 11). As discussed by Young et al. (2013), the relative spread among the ACCMIP models is large over the tropical areas of the oceans in the lower and middle troposphere, a reflection of the important differences among the models in various processes such as convective processes, lightning sources, biogenic emission sources with related chemistry. The large relative spread (>20\%) at the NH mid-latitudes and in the SH at 200 hPa may be associated with the different representations of the tropopause and STE among models. In contrast, the relative spread is small around 20--40$^\circ$N at 500 hPa (< 10 \%). The simultaneous enhancement of the analysis uncertainty (c.f., Section 6.1), together with the model spread, indicates low

削除： et

robustness of the validation results for some tropical regions over the oceans in the lower troposphere, and over the tropics in the Pacific Ocean as well as the Antarctic at 200 hPa. On the other hand, the ACCMIP model standard deviation with respect to the reanalysis could be used to identify the averaged uncertainty of ACCMIP models (right panels in Fig. 11). The standard deviation is large at NH high latitudes and over the tropical ocean areas at 800 hPa, over the SH tropics at 500 hPa, and in the SH extratropics at 200 hPa (> 25 \%).

\subsection{Implications into model improvements and climate studies}

Numerous studies have identified decadal-scale changes in global tropospheric ozone using observations, such as the shift in the seasonal cycle at NH mid-latitudes and trends observed over many regions (e.g., Parrish et al., 2014; Cooper et al., 2014). A long-record of the reanalysis will allow detailed structures in simulated inter-annual and long-term variations to be evaluated in association with changes in human activities and natural processes. However, any discontinuities in the availability and coverage of the assimilated measurement will affect the quality of the reanalysis and estimated interannual variability, which limit the usability of a long term reanalysis for model evaluation, as discussed in Miyazaki et al (2015) for chemical reanalyses and in Thorne and Vose (2010) for climate reanalyses. This also requires a bias-correction procedure for each assimilated measurements, in order to improve the reanalysis quality (Inness et al, 2013). It is noted that the influence of ENSO was not considered in ACCMIP due to a decadal-averaged SST boundary condition, which limits the evaluation of inter-annual variations and could lead to bias in the ACCMIP models and reanalysis comparisons.

Process-oriented validations using the reanalysis would be useful for understanding the uncertainty in simulated ozone fields and associated mechanisms. The ACCMIP models reveal large variations in short-lived species such as \chem{OH} and ozone precursors (Naik et al., 2013; Voulgarakis et al., 2013), whereas information obtained from direct in-situ measurements cannot be applied for investigating global distributions because of the limited coverage of the measurements and the large spatial variability of concentrations. Miyazaki et al. (2012b, 2015) demonstrated that the multiple-species assimilation results in a strong

削除：s

削除：as discussed in Miyazaki et al (2015)

削除：well simulated

influence on both assimilated and non-assimilated species. Validation of various species using the chemical reanalysis product can be used to identify potential sources of error in the simulated ozone fields. Meanwhile, the global monthly products of precursor emissions from the chemical reanalysis calculations (Miyazaki et al., 2012a, 2014, 2016) can be used to validate emission inventories and \chem{LNO_x} source parameterizations used in model simulations. As changes in tropospheric ozone burden associated with different future scenarios show a broadly linear relation to changes in \chem{NO_x} emissions (Stevenson et al., 2006), evaluations using up-to-date estimated emissions (Miyazaki et al., 2016) may prove useful to partly validate emissions for each scenario.

The performance of the simulated radiative forcing is largely influenced by representation of ozone in model simulations (Bowman et al., 2013; Shindell et al., 2013; Stevenson et al., 2013). Bowman et al (2013) suggested that overestimation of the OLR in the tropical seas of the east Atlantic Ocean and over Southern Africa is associated with model ozone errors, a persistent feature in all ACCMIP models, which was also found in this study using the reanalysis. Validation of short-lived species is also important for evaluating the radiative forcing because simulated \chem{OH} fields influence simulated climates through for instance their influences on methane (Voulgarakis et al., 2013). Thus, detailed information on model errors in ozone and other short-lived species could be used to improve estimates of radiative forcing in climate studies. Meanwhile, model biases for present-day ozone may be correlated with biases in other time periods. Young et al. (2013) showed that ACCMIP models with high, present day ozone burdens also had high burdens for the other periods of time, including the preindustrial period. Thus, the validation of present-day ozone fields using the reanalysis has the potential to evaluate preindustrial to present day ozone radiative forcing.

削除：have

\conclusions

We conducted a eight-year tropospheric chemistry reanalysis by assimilating multiple chemical species from the OMI, MLS, TES, MOPITT, SCIAMACHY, and GOME-2 to provide a gridded, chemically consistent estimate of concentrations and precursor emissions. This study explores the potential of atmospheric chemical reanalysis to evaluate global

削除：ten

tropospheric ozone of multi-model chemistry-climate model simulations. The evaluation results are also used to quantify the ozonesonde network sampling bias. Validation of the chemical reanalysis using global ozonesondes shows good agreement throughout the free troposphere and lower stratosphere for both seasonal and year-to-year variations.

The reanalysis product provides comprehensive and unique information on global ozone distributions for the entire troposphere and on the weakness of the individual models and multi-model mean. We found that the ACCMIP multi-model mean overestimates ozone concentration in the NH extratropics throughout the troposphere (by 6--11 ppb and 800 hPa and by 2--9 ppb at 500 hPa for the zonal and annual mean concentration), and underestimates it in the SH tropics in the lower and middle troposphere by about 9 ppb over the eastern Pacific, by up to 18 ppb over the Atlantic, and by up to 8 ppb over the Indian Ocean. Most models underestimate the spatial variability of the annual mean concentration in the NH extratropics at 800 hPa (by up to 50 \%) and in the SH extratropics at 800 and 500 hPa (by up to 70 \%). The multi-model mean overestimates the seasonal amplitude in the NH by 50--70 \% in the lower troposphere and by 25--40 \% in the middle troposphere, whereas the seasonal amplitude is underestimated by 15--25 \% at 200 hPa in the NH extratropics. The seasonal amplitude in the NH extratropics shows great diversity among models. The NH/SH ratio is overestimated by 22--30 \% in the free troposphere in the multi-model mean; this can be attributed to both a concentration high bias in the NH and a concentration low-bias in the SH in most models. The performance of the ACCMIP model when compared with the reanalysis is qualitatively similar for most cases from that shown by Young et al. (2013) using the ozonesonde measurements but quantitatively different because of the ozonesonde network sampling bias.

We quantified the ozonesonde network sampling bias and how reanalysis can help extend the range of that network as a kind of "transfer standard". For instance, the ozonesonde sampling bias in the evaluated model bias is largely negative (positive) in MAM (in DJF) by 40--50 \% over the Western Pacific and East Indian Ocean and largely negative by 110 \% in MAM over the equatorial Americas at 500 hPa. For the global tropics, the ozonesonde sampling bias is largely negative by 80 \% in the NH (Eq--30{\degree}\,N) in SON and by 50 \% in the SH (30{\degree}\,S--Eq) in MAM. Although the spatial and temporal variability is generally

削除: .

削除:

削除: .

削除: for most cases

削除: largely

smaller in the SH than in the NH, the ozonesonde sampling bias cannot be negligible for capturing the regionally and monthly representative model errors even in the SH. Large sampling biases (> 60 \%) exist in the SH extratropics (90--30{\degree}\,S) in DJF and MAM. The evaluation of the seasonal cycle of tropospheric ozone is also largely limited by the ozonesonde sampling bias. The evaluation based on the ozonesonde sampling introduces a larger overestimation of the seasonal amplitude than that based on the complete sampling for most of the surrounding areas in the NH lower troposphere, whereas the two estimates are largely different for the entire tropical regions. Therefore, there is an advantage of the reanalysis data for evaluating actual regionally and seasonally representative model performance required for model improvements. However, the network provides critical independent validation of the reanalysis, which can provide a much broader spatial constraint on chemistry-climate model performance.

The proposed model validation approach provides regionally and temporally representative model performance; this could ensure more accurate predictions for the chemistry--climate system. In future studies, validation of multiple species concentrations and precursor emissions from reanalysis would be useful in identifying error sources in model simulations. In particular, the response of tropospheric composition to changing emissions over decadal time scales is still not captured in CCMs relative to a few remote sites (Parrish et al, 2014). Recent increases in emissions from China have been linked to changes in tropospheric ozone concentrations (Verstraeten et al, 2015). Over the next decade, a new constellation of of low Earth Orbiting sounders, e.g., IASI, AIRS, CrIS, Sentinel-5p (TROPOMI), Sentinel-5 and geostationary satellites (Sentinel-4, GEMS, and TEMPO) will provide even more detailed knowledge of ozone and its precursors (Bowman, 2013). Assimilating these datasets into a decadal chemical reanalysis will be a more direct means of quantifying the response of atmospheric composition to emissions at climate relevant time scales, which should be a more direct test on chemistry-climate change scenarios. Combining many observations requires a bias correction procedure for each assimilated measurement to improve the reanalysis quality but needs to be carefully checked. We also plan to apply the proposed evaluation approach to a more recent model inter-comparison project, the Chemistry-Climate Model Initiative (CCMI).

削除: .

削除: s

\begin{acknowledgements}

We acknowledge the use of data products from the NASA AURA and EOS Terra satellite missions. We also acknowledge the free use of tropospheric \chem{NO_2} column data from the SCIAMACHY, GOME-2, and OMI sensors from www.temis.nl. We would also like to thank the editor and two anonymous reviewers for their valuable comments. This work was supported through JSPS KAKENHI grant numbers 15K05296, 26220101, and 26287117, and Coordination Funds for Promoting AeroSpace Utilization by MEXT, JAPAN.

\end{acknowledgements}

Table 1: Measurements used for data assimilation in the chemical reanalysis.

Table 2: Ozonesonde observation sites used in this study. All the data are used for the evaluation of reanalysis data (Section 3), whereas selected observations (shown in bold) based on the compilation by Tilmes et al. (2012) are used for the evaluation of ACCMIP models and to investigate ozonesonde sampling biases (Section 4 and 5).

Table 3: Chemical reanalysis (or control run in brackets) minus ozonesonde comparisons of mean ozone concentrations in 2005--2009. RMSE is the root-mean-square error. Units of bias and RMSE are ppb. T-Corr is the temporal correlation.

書式変更: タブ: 13.61 字, 左揃え

Table 4: ACCMIP model mean minus reanalysis comparisons of the mean ozone concentrations. Units of bias and RMSE are ppb. S-Corr is the spatial correlation coefficient.

Table 5: ACCMIP models minus reanalysis comparisons of the mean ozone concentrations at 500 hPa. Units of bias are ppb.

Table 6: Regions and observation sites used in model evaluation in Section 5. The 11 regions are defined following Tilmes et al. (2012). See also Fig. 8.

書式変更: タブ: 13.61 字, 左揃え

Table 7: The reanalysis ozone concentration differences between the ozonesonde sampling (for both time and space using two-hourly reanalysis fields) and the complete sampling at 500 hPa (in \% relative to the complete sampling). Results using monthly reanalysis fields sampled at the ozonesonde locations are also shown in brackets.

Table 8: Median of the ACCMIP models minus reanalysis at 500 hPa (in \% relative to the reanalysis concentrations). Results presented include the regional averages (Regional), for the ozonesonde temporal/spatial sampling using two-hourly reanalysis fields (Sonde), and for the ozonesonde spatial sampling using monthly reanalysis fields (in brackets). Relative differences between the two estimates larger than 30 \% are shown in bold.

Table 9: The control run minus reanalysis comparison of the mean ozone concentration at 500 hPa (in \% relative to the reanalysis concentrations). Results are the regional averages (Regional) and at the ozonesonde temporal/spatial sampling (Sonde). Relative differences between the two estimates larger than 30 \% are shown in bold.

Table 10: ACCMIP multi-model mean minus reanalysis comparisons of the seasonal amplitude of regional mean ozone concentration (in \%) for the regional average (Regional) and at the ozonesonde sampling (Sonde). The seasonal amplitude is estimated as a difference between maximum and minimum monthly mean concentrations.

Figure 1: Comparison of vertical ozone profiles from ozonesondes (black), control run (blue), and reanalysis (red) averaged for the period 2005--2009. Top row shows mean profile; middle and bottom rows show mean difference and RMSE between control run and observations (blue) and between the reanalysis and the observations (red). From left to right, results are shown for SH extratropics (30--90$^\circ$S), SH tropics (30$^\circ$S--Eq), NH tropics (Eq--30$^\circ$N), and NH extratropics (30--90$^\circ$N). All ozonesonde observations taken from the WOUDC database were used in the comparison.

削除: reanalysis concentrations

削除:

削除: high latitudes
削除: 55
削除: mid-latitudes
削除: 15
削除: --55
削除: tropics (15$^\circ$S--15$^\circ$N),
削除: mid-latitudes
削除: 15
削除: --55
削除: high latitudes
削除: 55

Figure 2: Time series of monthly mean ozone concentrations obtained from ozonesondes (black), control run (blue), and reanalysis (red) averaged between 850 and 500 hPa (top), 500 and 200 hPa (middle), and 200 and 90 hPa (bottom) for 2005--2009. From left to right the results are shown for SH extratropics (30--90$^\circ$S), SH tropics (30$^\circ$S--Eq), NH tropics (Eq--30$^\circ$N), and NH extratropics (30--90$^\circ$N).

削除：SH high latitudes (55--90$^\circ$S), SH mid-latitudes (15--55$^\circ$S), tropics (15$^\circ$S--15$^\circ$N), NH mid-latitudes (15--55$^\circ$N), and NH high latitudes (55--90$^\circ$N)

Figure 3: Global distributions of annual mean ozone concentrations obtained from reanalysis (left), ACCMIP model mean (2nd left), difference between ACCMIP model mean and reanalysis (3rd left), and the ozonesonde measurements used for the evaluation of ACCMIP models and ozonesonde sampling biases (right). From top to bottom, results are shown for global distributions at 200 hPa, 500 hPa, and 800 hPa. Units are ppb.

削除：center

削除：and

削除：right

Figure 4: Taylor diagrams showing standard deviation normalized with respect to that of reanalysis (x-axis) and spatial correlation coefficient (y-axis) for the comparison of annual mean ozone concentrations between ACCMIP models and reanalysis for SH extratropics (90$^\circ$S--30$^\circ$S, left), tropics and subtropics (30$^\circ$S--30$^\circ$N, center), and NH extratropics (30$^\circ$N--90$^\circ$N, right) at 200 hPa (top), 500 hPa (middle), and 800 hPa (bottom).

Figure 5: Comparison of seasonal variation of ozone concentration between the reanalysis (black lines), individual ACCMIP models (thin colored lines), ACCMIP ensemble mean (red solid line), and ozonesonde observations (blue solid line) averaged between 90$^\circ$S--30$^\circ$S (1st column from left), 30$^\circ$S--Eq (2nd column), Eq--30$^\circ$N (3rd column), and 30$^\circ$N--90$^\circ$N (4th column). From top to bottom, results are shown for concentrations at 200 hPa, 500 hPa, and 800 hPa. Individual model results are shown by colored thin lines. The reanalysis result is shown for the average over all model grid points (black solid line) and over the ozonesonde sampling sites/time (black dashed line).

削除：uator

削除：uator

削除：，

Figure 6: Seasonal amplitude (peak-to-peak difference) estimated from the reanalysis (black solid line) and ACCMIP models (thin colored lines). The $\pm 1 \sigma$ deviation among ACCMIP models (i.e., model spread) is shown in pink. The seasonal amplitude derived from the multi-model mean fields (red solid line) and the multi-model mean of the seasonal

amplitude from each model (red dashed line) are also shown. From top to bottom, results are shown for 200 hPa, 500 hPa, and 800 hPa.

Figure 7: Vertical profile of inter-hemispheric gradient of annual mean ozone concentrations estimated from the reanalysis (black lines), ACCMIP ensemble mean (red solid line), ACCMIP models (thin colored lines), and ozonesonde observations (blue solid line). The reanalysis result is shown for the average over all model grid points (black solid line) and over the ozonesonde samplings (black dashed line). The $\pm 1 \sigma$ deviation among the ACCMIP models is shown in pink.

Figure 8: Regions and observation sites used in model evaluation. The 11 regions are defined following Tilmes et al. (2012). See also Table 3.

Figure 9: Box plots of relative model--reanalysis difference for seasonal mean concentration for DJF (left) and MAM (right) at 200 hPa (top), 500 hPa (middle), and 800 hPa (bottom). Results are shown for ACCMIP model simulations for 11 regions (c.f., Table 3 and Fig. 8). Black box shows model minus reanalysis difference for regional mean concentration (averaged over all model grid points); red box shows model minus reanalysis at the ozonesonde samplings.

Figure 10: Probability distribution functions (PDFs) of ozone concentration obtained from the ACCMIP multi-model mean (blue) and the reanalysis (red) at 500 hPa for W. Pacific/E. India in SON (left) and for the SH high latitudes in MAM (right). The plots are shown for all model and reanalysis grid point (bottom) and for the ozonesonde sampling (top) within each defined region.

Figure 11: Global distributions of relative value (in \%) of reanalysis uncertainty (left), standard deviation among the ACCMIP models (center), and ACCMIP model standard deviation with respect to the reanalysis for the annual mean concentration (right). From top to bottom, results are shown for global distributions at 200 hPa, 500 hPa, and 800 hPa.