Response to Referee 2:

We thank the reviewer for the critical review of the paper, which has contributed to an improved version of this manuscript. Comments by the referee are shown in *italic* font. In addition to the changes described below, note that the figures and Table 1 have changed slightly from the previous version of the manuscript due to having to redo some simulations to correct a minor inconsistency. None of the statements in the text had to be changed because of this correction. The simulations that were redone are CAM5-chem, CAM5-chem*, CAM5-chem MAM4, and SD-CAM4-chem.

This manuscript presents an overview assessment of the performance of tropospheric chemistry within the CESM1.2. The presentation makes extensive comparison to observation for validation and assessment of model results. In the Abstract, the authors state that the system is "well suited as tools for atmospheric-chemistry modeling studies in the troposphere and lower stratosphere, whether with internally derived "free running" (FR) meteorology, or "specified dynamics" (SD)" and follows with "The main focus of this paper is to compare the performance of these configurations against observations from surface, aircraft, and satellite, as well as understand the origin of the identified differences." Unfortunately, the paper doesn't seem to sufficiently address the second part of the statement well enough to support the first.

We agree that the model configurations are not "perfect" in the sense that they reproduce available observations within the range of uncertainty. Indeed many shortcomings still exist in the model and further development is required. We drop the sentence in the abstract that is of concern to the reviewer and instead focus on describing the performance of the different configurations in comparison to observations.

We therefore modify the abstract as follows:

"The Community Atmosphere Model (CAM), version 5, is now coupled to extensive tropospheric and stratospheric chemistry, called CAM5-chem, and is available in addition to CAM4-chem in the Community Earth System Model (CESM) version 1.2. The main focus of this paper is to compare the performance of configurations with internally derived "free running" (FR) meteorology and "specified dynamics" (SD) against observations from surface, aircraft, and satellite, as well as understand the origin of the identified differences. We focus on the representation of aerosols and chemistry. All model configurations reproduce tropospheric ozone for most regions based on in-situ and satellite observations. However, shortcomings exist in the representation of ozone precursors and aerosols. Tropospheric ozone in all model configurations agrees for the most part with ozonesondes and satellite observations in the Tropics and the Northern Hemisphere within the variability of the observations. Southern Hemispheric tropospheric ozone is consistently underestimated by up to 25%. Differences in convection and stratosphere to troposphere exchange processes are mostly responsible for differences in ozone in the different model configurations. Carbon monoxide (CO) and other volatile organic compounds are largely underestimated in Northern Hemisphere mid latitudes based on satellite and aircraft observations. Nitrogen oxides (NOx) are biased low in the free tropical troposphere, whereas peroxyacetyl nitrate (PAN) is overestimated in particular in high northern latitudes. The present-day methane lifetime estimates are compared among the different model configurations. These range between 7.8 years in the SD configuration of CAM5-chem and 8.8 years in the FR configuration of CAM4-chem and are therefore underestimated compared to observational estimations. We find that differences in tropospheric aerosol surface area between CAM4 and CAM5 play an important role in controlling the burden of the tropical tropospheric hydroxyl radical (OH), which causes differences in tropical methane lifetime of about half a year between

CAM4-chem and CAM5-chem. In addition, different distributions of NOx from lightning explain about half of the difference between SD and FR model versions in both CAM4-chem and CAM5-chem. Remaining differences in the tropical OH burden are due to enhanced tropical ozone burden in SD configurations compared to the FR versions, which are not only caused by differences in chemical production or loss, but also by transport and mixing. For future studies, we recommend the use of CAM5-chem configurations, due to improved aerosol description and inclusion of aerosol-cloud interactions. However, smaller tropospheric surface area density in the current version of CAM5-chem compared to CAM4-chem results in larger oxidizing capacity in the troposphere and therefore a shorter methane lifetime."

It is recognized that a truly thorough evaluation of a system as complex as this is difficult and time consuming, but the content of this paper is a pretty extensive information dump that for the most part lacks anything more than speculative and unsupported guesses regarding the results, their main drivers and why they are what they are relative to observation.
We agree with the reviewer that the submitted version of the manuscript was lacking in supportive arguments for some of the conclusions drawn. We also agree that a lot of information is given in the paper. To improve the paper, we have now included what we feel are better arguments for the drawn conclusions and removed those conclusions that could not be supported, as discussed in more detail below. We have tried to make the statements more quantitative, and also improved the structure of Section 4.2 as suggested by the other reviewer, and perform the evaluations grouped by chemical species.

- Why and what is the impact of the new reactions added?
The reason for the change in reactions are described in the paper
line 8882: "In CESM1.2 CAM4-chem, the lumped aromatic ("TOLUENE") was replaced with the specific species benzene, xylene and toluene, along with simplified oxidation products for the two new species, to accommodate the 2-product formation of SOA (new reactions listed in Appendix A)."
The 2-product formation of SOA in CAM4-chem requires additional species that were not included in earlier versions of the model. This change has little impact on the chemistry, but allows the new SOA calculation. We have added the sentence:
"These changes do not have an impact on the chemical performance of the model."

- Can the impacts of clouds, dynamics, aerosols, etc. be more explicitly evaluated to address the inter-model differences and differences between model and observation?
In this paper, we focus on the performance and evaluation of aerosols and chemistry. Differences in clouds and dynamics between the different configurations are discussed briefly in Section 4.1 but were evaluated in detail in earlier studies. In the revised version of the manuscript we try to improve the paper and identify differences in chemistry as a result of convection and transport, and to some degree clouds, by including an additional figure, new Figure 3 (see below) and add the following text:

"The comparison of chemical constituents in the two model configurations further supports a stronger tropical vertical transport in CAM5-chem compared to SD-CAM5-chem and stronger STE in high latitudes (Fig.3). Stronger tropical vertical transport (mostly in deep convection) in CAM5-chem is evident due to higher mixing ratios in CO and lower mixing ratios of nitric acid in the upper tropical troposphere. The resulting higher CO mixing ratios in the upper troposphere together with increased LNOx in mid-latitudes lead to greater ozone production, while reduced LNOx in the tropical belt reduces ozone production. Furthermore, increased nitric acid in addition to higher ozone mixing ratios in high northern latitudes point to more STE. Additionally, lower

NOx and CO values in the boundary layer in CAM5-chem indicate that increased STE rather than chemical processing results in larger ozone mixing ratios in CAM5-chem than SD-CAM5-chem. Differences in low clouds between CAM5-chem and SD-CAM5-chem also impact chemistry and result in reduced ozone production in the boundary layer in CAM5-chem. Similar differences are present between CAM4-chem to SD-CAM4-chem, however, with smaller differences in STE in high latitudes compared to the CAM5-chem configurations (not shown)."

It seems that this paper should go in one of two directions:
1) It should be extensively revised to include a more explicit statistical analysis of the results and their comparison to observation. This would be perhaps be best done with a smaller set of observations that highlight some of the main driving factors that impact results: e.g. tropospheric O3 is often a clear signal of vertical transport. Methane lifetimes reflect OH abundance and chemistry, but this pits one of the slowest reacting species as an indicator of one of the fastest reacting. Is it possible to more directly assess their interaction?

We have attempted to improve the presentation of the statistical evaluation of tropospheric ozone in the revised version of the paper. In addition to the discussion of the Taylor-like diagrams for ozone evaluation based on ozonesonde observations, we also show seasonal cycle plots, including information of the variability of the observations (adding standard deviations). Further, we included the evaluation of the probability distribution function of surface ozone observations.

In additions, we have discussed the connection between methane lifetime and OH, and correlations with other factors in detail in Section 5.

2) The paper should be retracted and resubmitted as two papers: One as a technical description paper with a lot more technical discussion and evaluation. And a second that evaluates the model's performance against observation, including a proper statistical analysis.
We agree with the reviewer that this paper does not cover the technical aspects of the various model configurations since many of those are available in previous publications. The reviewer might have been misled in the introduction by the statement: "The performance of CAM4-chem and CAM5-chem is discussed in Sect. 4, which includes model-to-model comparisons of chemistry and aerosol budgets". We have changed this sentence to: "Model-to-model differences in dynamics, chemistry and aerosols, and global budgets are discussed in Sect. 4. A comprehensive evaluation of chemistry and aerosols, based on satellite and in-situ observations is performed in Sect. 4.2."

The focus of this paper is on evaluating the inter-model performance and the performance against observations. We believe that our statistical analysis is for the most part proper, but we agree that improvements can be made, as described in the comment above.

The authors have clearly done a lot of work preparing and testing the model system, compiling observations, and developing tools for comparison. They are well regarded for their effort, but this paper, unfortunately, is sub-par. I add, lastly, that with the density of acronyms it is somewhat difficult to read.
Unfortunately many acronyms are used, but there is not really a good solution in changing this. We have added a table for the purpose of making the paper more readable due to the acronyms used, which is updated in the revised version of the manuscript.

Please consider my comments below for more detail.

P8878, L1: The intro starts off rather abruptly. Please include a general introduction to the CESM.
We agree with the reviewer and modify the first paragraph of the introduction:

"The Community Earth System Model (CESM) is a comprehensive model that couples different independent models for atmosphere, land, ocean, sea-ice, land ice, and river runoff (e.g., Neale et al., 2013; Lamarque et al., 2012). It can be used in various configurations, depending on the use of different components and the coupling between them. The atmospheric component of CESM, the Community Atmosphere Model (CAM), has the capability of including chemistry of varying complexity. Default CESM configurations used for long-term climate model simulations usually include prescribed chemical fields in the atmosphere using monthly averages. …"

and the first sentence of the second paragraph:
"In CESM version 1.2, CAM version 5 (CAM5) extensive tropospheric and stratospheric chemistry, referred hereafter to as CAM5-chem, has been successfully implemented."

P8879, L27: With fixed SST, what impact might this have on truly evaluating the interactions between chem., physics, and dynamics in the results?
We purposely performed the experiments with prescribed present-day SST, to eliminate an additional factor in differences between the various model simulations. For specified dynamics runs, calculated SST would not make a difference since the atmospheric dynamics is nudged. For the free running simulation, calculated SST would add additional variability to the atmospheric fields, which would make it more difficult to identify the reason for the differences in different chemical fields. The larger variability in the specified dynamics runs for example in lightning NOx compared to the free-running version may be caused by the fact that the SST are prescribed for all model years, which may reduce the variability in the FR version, but not in the SD version where atmospheric dynamics are from analyses. We will change this sentence in Section 5:
"The spread in the annual LNOx production is much larger in the SD simulations compared to the FR configurations, which is likely introduced by the use of climatological SSTs in the FR configurations."

P8880 – Since the system uses chemistry coupled to the modal aerosol scheme, comparison to results from Long et al. (2013, GMD; 2014, ACP) with more explicit multiphase chem would perhaps be valuable to this MS.
This study is focused on evaluating the different configurations of CESM with chemistry. We think that it would be beyond the scope of the paper to compare our results to a different model.

P. 8882, L15-17: Small particles DO impact het. chem. Why do the authors believe they don't? This would only be the case if the specific chemical mechanism used specifically neglected the chemistry associated with smaller particles.
We rephrase the sentence to clarify that heterogeneous chemistry on small particles is not included in the current model version, even though it has been shown that those reactions matter.
"The contribution of very small particles, such as the Aitken mode in MAM3 and the primary carbon mode in MAM4, are neglected in the model calculation of surface area density. Further, sea-salt and mineral dust aerosols do not contribute to SAD in either model version, as heterogeneous reactions are not assumed to occur on these surfaces. Since reactions on very small particles are important, this may lead to an underestimation of SAD in the model."
The conclusions also state:
"An underestimation of SAD in the model is possible, because BC plumes are significantly underestimated over source regions. Since background aerosols are in general overestimated, shortcomings may exist in the calculation of SAD. For example, sea-salt and dust provide

surfaces for heterogeneous reactions that have not been taken into account in any of the simulations (Evans et al., 2005)"

P. 8888, L 20-24: At the resolution used, is CAM even able to resolve STE well enough to permit diagnostic analysis of the impact of stratospheric chemistry and exchange on the results?
The purpose of the performed diagnostics is to identify the differences in STE in the models. Comparisons with ozonesonde and aircraft data indicate that differences in STE can lead to a slight overestimation of ozone (in CAM5-chem) or in a slight underestimation (SD-CAM versions) in mid- to high latitudes, but on average to reasonable values of ozone, as described in the text.

P8889, L10: SAD is used here but defined on p. 8890
SAD is now introduced when it is used the first time.

P8891, L13-15: This statement is self-contradicting. It appears you're using the term SAD in place of the more appropriate term surface-to-volume ratio. Perhaps this should be changed.
Surface area density in microns^2/cm^3 is indeed the surface area to volume ratio.
"Heterogeneous reactions on aerosol particles in the model do not directly relate to the aerosol burden, but rather depend on the amount of tropospheric SAD. SAD depends not only on aerosol burden or mass, but also on their size distribution. For the same aerosol burden, smaller particles provide a larger SAD than larger particles."

P8892 and in general: Since this is a chemistry paper, more or less, it is important that the appropriate symbols and terminology are used. Is SO2 meant to refer to SO2 gas? If so, it should be SO2(g). SO4, as presented, suggests some non-ionic sulfur oxide radical. If it is meant to be particulate or aqueous SO42-, it should be presented as such.
To clarify in the paper we added "SO2 is referred to as gas-phase species throughout the paper" and further, define SO4 as "sulfate aerosol".

P8892, L17-20: Unfortunately, for a publication seeking to present a proper analysis of a chemistry model system, the terms "reasonable agreement" and "agrees well with" are insufficient. This work should be a proper analysis based on observational and modeled statistics. Avoiding this type of analysis completely undermines the utility and ability to extract meaningful information from this as a scientific work. The work that the authors have done to date compiling data and creating unbelievably useful tools for making comparisons and analysis possible should be leveraged to do at least something to step beyond the current state. IF, on the other hand, it is the authors intention to perform a more complete analysis for presentation in another manuscript for submission elsewhere – which may be the more suitable way forward – then this should be stated and much of this analysis should be removed from the manuscript.
To address the concerns of the reviewer, in the revised version of the manuscript we have added more quantitative statements to describe how "well" the model compares to the observations. We further added two new figures, Figures 13 and 16 (see below) showing seasonal cycle comparisons of the model to ozonesonde observations, additions the statistical comparison in from of a Taylor-like diagram, to visualize the performance of the model with regard to the variability of observations. We also added a comparison of probability distribution functions between surface measurements and model results (new Figure 14).
The sentence pointed out by the reviewer:
"Comparisons to aircraft observations over the US and high northern latitudes (Fig. 8), show a reasonable agreement of SO2 over the US for all model configurations. Further, SO4 agrees well in the troposphere over the US, while boundary layer values are overestimated."

Is changed to:
"Comparisons to aircraft observations over the US (Fig. 8, left), show very good agreement for SO2 that are very close to the observed values for two of the campaigns, while simulated values are slightly larger for ARCTAS-CARB. Further, the model configurations reproduce observed sulfate aerosol for some campaigns, with the exception of CAM4-chem, while boundary layer values are more than double to those observed."

P8896, L6, L8; P8897, L1; P8904, L9: Invoking cloud impacts on O3 seems speculative here. Further analysis is both possible and necessary.
We agree and have removed the discussion and corresponding Figure.

P8893, L28 – P8894, L1: If the AOD bias is the result of too much seasalt and/or dust, wouldn't this mean that the winds may be too high? This would be a straightforward analysis, wouldn't it? Also, given that AOD is based on the MIRAGE system, wouldn't a sensitivity to RH also be just as likely to affect AOD?
We give some more information on sea salt and dust in the model configuration and experiment section: "The dust emissions are calibrated so that the global dust AOD is between 0.025 and 0.030 (Mahowald et al, 2006). Further, sea salt emissions are calibrated in present-day conditions so that the global mean AOD (for all species) are within the reasonable range. Those values have been evaluated in Liu et al., 2012, who show that the difference between model simulations and observations are generally within a factor of two.

Further we add the following text to Section 4.2.1: Aerosols and Aerosol Optical Depth (AOD):

"The AOD bias in the subtropical ocean (mostly from coarse mode sea salt) can be due to the model deficiency representing the sea salt emission or sedimentation (scavenging) process that requires further investigation. Using reanalysis winds do not reduce this bias (not shown)."

P8904, L3-4: This is improper English.

We change the sentence: "Differences in the H_2O-tape recorder and in AOA point to a too strong Brewer–Dobson circulation in the FR model configurations, while it is reasonably reproduced in the SD configurations." to
Differences in the H_2O tape recorder and in AOA indicate that the Brewer–Dobson circulation is too strong in the FR model configurations, while both diagnostics are reasonably reproduced in the SD configurations."

Table 1 (and most of the rest of the data shown): What are the statistics? Are they means? If so, what about std. dev. If they're medians, what about range? In any case, are the statistics appropriate? For example, if the optical depth data are means, it should be pointed out that AOD is not normally distributed and therefore means are meaningless (no pun intended). In fact, most of the quantities presented in this table and elsewhere are known to not only NOT be normally distributed, but the distribution and resulting statistics are known to differ with model horizontal resolution. In most cases, there are clear modes within the distributions that can be attributed to individual geographic regions, features, or locations in the atmosphere. As such, changes in one area may appear to affect the whole from a statistical standpoint, when in fact there is no change at all elsewhere. It is possible that any analysis or comparison based on these data are insupportable.

The purpose of Table 1 is to show global values and budgets that are meant to give an overview of the performance of the different model configuration. The total budgets, lifetimes, and other variables, are usually presented in this way. The values we are showing are averages over all the years of the simulations. We are not discussing variability of the different variables here, but go into the details of the variability of some of the variables in Section 4 and 5, as well as for AOD. We are very aware of the fact that many variables, for example ozone, are not distributed normally depending on regions and altitudes considered, as for example pointed out in Tilmes et al., 2012. This is why this table is not meant to serve as a full evaluation tool, but to give an overview of the global diagnostics, whereas detailed investigations are presented further along in the paper. We are not discussing the impact of different horizontal resolutions, however, in the new version of the paper, we acknowledge that some shortcoming in the model are likely a result of the coarse resolution (see the additional comparison to surface observations). Further case studies would be beneficial to investigate the behavior of different variables in more depth and specific aspects of the model, but this is beyond the scope of this paper.

New Figures:

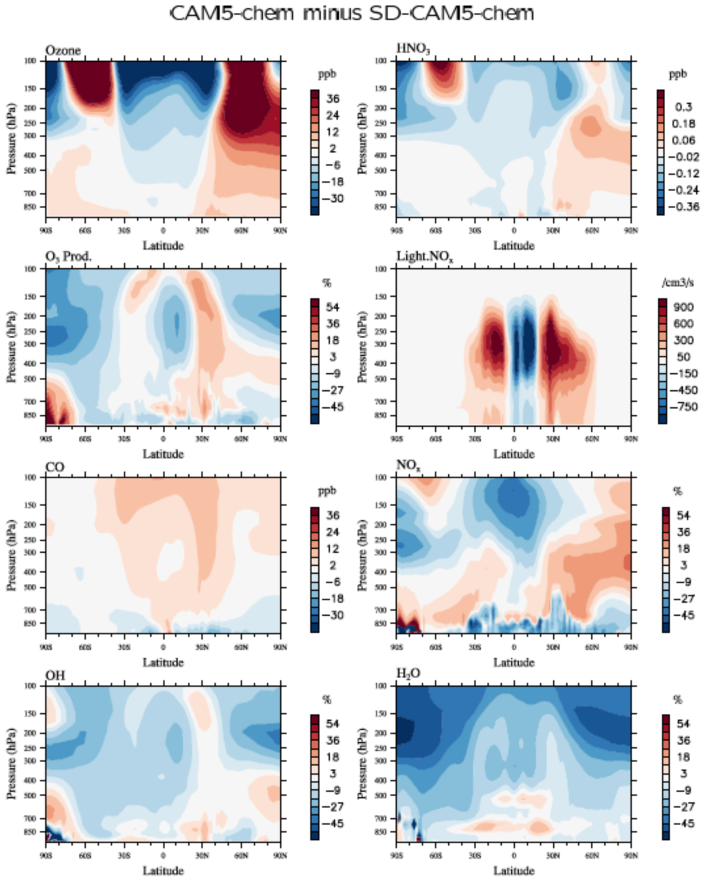

Figure 3: Comparison of ozone, nitric acid, ozone production, lightning NOx, carbon monoxide, NOx, hydroxyl radical, and water vapor, between CAM5-chem and SD-CAM5-chem.

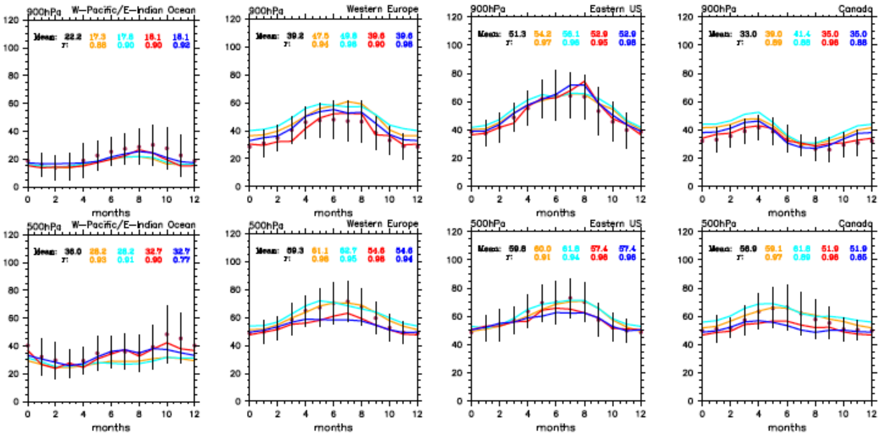

Figure 13:

Seasonal cycle comparison between observations using a present-day ozonesonde climatology between 1995--2011 (black) and model results: CAM5-chem (cyan) and CAM4-chem (orange), SD-CAM5-chem (blue) and SD-CAM4-chem (red). Model results are interpolated to the same locations as sampled by the observations and for different pressure levels, 900hPa (top panel) and 500hPa (bottom panel) for selected regions. The standard deviation of ozonesonde observations is shown as error bars and the mean and correlation of the seasonal cycle between observations and model results are printed on the top of each figure.

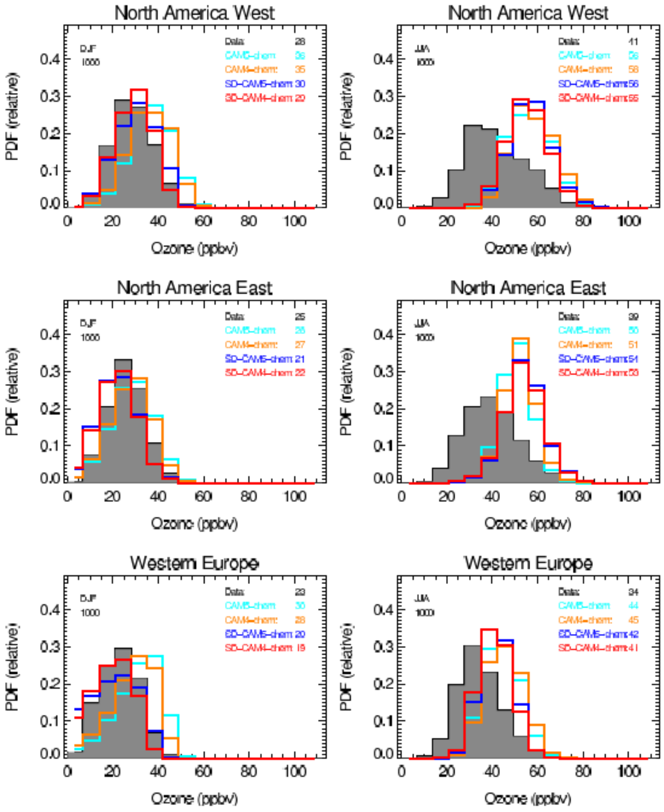

Figure 14: Probability distribution function (PDF) of the regionally-aggregated ozone distribution for Western North America, Eastern North America, and Western Europe from surface ozone observations (grey shaded area) in comparison to regionally-aggregated ozone distributions from

model results interpolated to the location of the ozone stations (different colors), for winter (left) and summer (right).

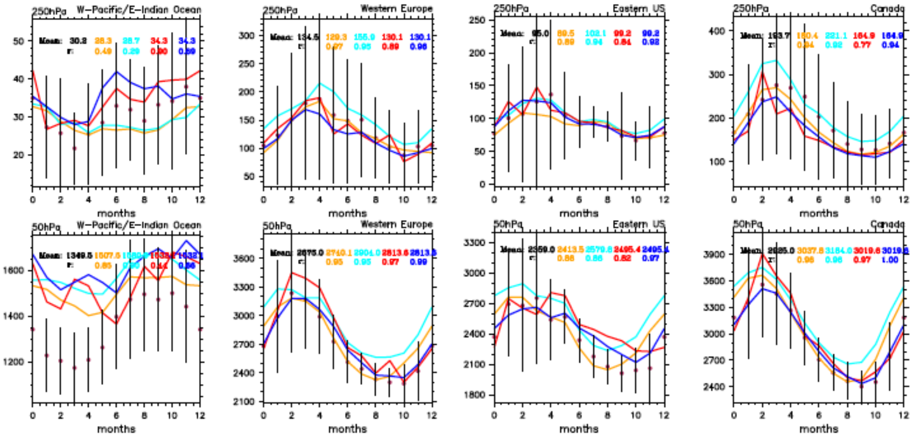

Figure 16:

As Fig.13, but for different pressure levels, 250hPa (top panel) and 50hPa (bottom panel).