Reply to the review from Referee 2

We are thankful to this referee for the review and the associated suggestions, listed in italics below. We provide our detailed responses (regular font) and plans; our revised manuscript will be available in a fairly short time. For added information, we have provided the revised Abstract in our reply here, although most of those (highlighted) changes were done as a response to comments from the other Referee.

It would seem from the referee comments that there are no demonstrable big issues with the science (or math), besides some requested clarifications, and we are pleased that Referee 2 found our manuscript to contain "a lot of valuable and detailed information" [and Referee 1 also found "some interesting results"].

The manuscript aims to evaluate the stratospheric composition of the free-running and specified-dynamics version of CESM1 (WACCM). The evaluations are based on comparisons to satellite measurements including single-instrument and merged data records. The model diagnostics include zonal monthly mean comparisons, seasonal and semi-annual cycles as well as long-term trends. All evaluations are described in detail and valuable information on various aspects of the model performance is provided. Overall, the manuscript is of great interest for scientist directly working with WACCM or potentially with other earth-system models. Therefore, such a detailed manuscript would seem much more appropriate in a journal focused on geoscientific model development/validation and I would urge the authors to submit it to a journal focused on this topic.

Reply: We do not agree that the fairly comprehensive analyses presented here are of more limited interest to modeling groups only, because there are also inferences from data sets that have not been presented before (in particular trend analyses from MLS data alone). We have also added a few clarifications to better explain certain aspects of these trends and comparisons (largely in response to the other Referee).

Regarding the Journal issue, we feel quite strongly that such a paper is (or can certainly be) in the ACP domain, given that the model description is really a small part of this manuscript (WACCM having been used and described previously, including in GMD, *Morgenstern et al.*, gmd-10-639-2017), and that there are some scientific results discussed here (to be further clarified, as mentioned in our replies and in the upcoming revision), even if some of this confirms past/recent work, but from our own model/data comparisons. There is some "grey region" between ACP and

GMD papers, with the latter being more geared towards model description and development (if one looks through many of those articles), although there are some model evaluation papers there as well. To be more specific, we include Table 1 at the end of this reply, and this provides a summary of all the papers that are part of the current CCMI special issue, which is what we are submitting to here; this special issue encompasses several journals (including ACP and GMD). As one can see from Table 1, the more recent papers have nearly all been part of ACP, after some initial work with much more of a model description focus. Some of the articles in ACP could compare broadly to the work we are trying to present, with a combination of model and data (and comparisons). We also feel that there are detailed aspects of the MLS data sets described in our work (regarding absolute error bars and trend uncertainties, including some drift issues) that would be of much interest to the stratospheric component of the ACP readership. Without attempting to be more comprehensive, we can state that we did consider the Journal topic seriously, which also led to some delays. We also consulted with the ACP editors on this topic, and we are pleased that they agree with our views; this topic is also something that editors consider as part of the pre-review process. It is also true that going through another 4 months of review with a completely new set of reviewers and editors is a considerable burden not just on the authors, with further time delays, but also on the reviewer community (especially for longer papers). We are thus thankful for the support we obtained towards finalizing this process for ACP, and we feel that we can now focus our efforts to that end; we would very much welcome reviewer support on this aspect as well.

Major comments

1) The paper delivers a lot of valuable and detailed information, however, is overall very long. In particular, the number of figures could be reduced from 32 to around 20. To give one example, Figure 2 is only discussed very briefly in the text in order to illustrate mean biases and annual cycle differences shown elsewhere and could be removed.

Reply: We really prefer to keep Fig. 2 in the main text as this does show much better than Fig. 1 how the models and the data differ (in certain regions) in terms of not only the mean differences, but also the annual cycle; these differences are now also better quantified in the upcoming revised version (in answer to a comment from reviewer 1).

We are planning to cut down on the length of this manuscript, mainly by relegating some of the less critical Figures to the Supplement. Although this does not necessarily translate into a very large cut in terms of text length, we consider this work to be a fairly comprehensive analysis, which therefore leads to a longer paper; there have definitely been some longer (atmospheric) papers in the literature, and specifically in ACP. Turning this into two separate papers mainly for the sake of overall length seems too artificial, and this would be quite an elaborate proposition, with the need for some duplication regarding both the data sets and the models; as an aside, this would actually lead to more reviewing work for the community. We hope to have shown that detailed analyses are necessary to enable identification of both good agreement (a result in itself) or significant differences between model runs and the data sets, but also for some of the more subtle differences, and furthermore, that an understanding and discussion of error bars and potential data issues is important. We will also strive to reduce the amount of text in the revised manuscript, especially where some less critical aspects can be discussed more succinctly, or taken out altogether. In particular, we plan to shorten Section 5.1.1 (pages 11-13) to a text length that roughly matches (rather than exceeds) the text length of Section 5.1.2 (on variability issues); the cuts to Sect. 5.1.1 will be of order 30% (or more).

In terms of reducing the amount of Figures and related changes, our specific plans are to remove Figs. 13, 14, 15, and 17 from the main text (and relegate these to the Supplement, with a slightly shortened discussion), since these mainly reinforce the expectation (already noted for O_3 and H_2O) of better model/data fits from SD-WACCM, as one might expect from a model with better dynamical constraints than the FR-WACCM version. Such an expectation does not hold for the variability diagnostics, so these are really best left in the main text, although we will plan to displace Figure 22 (on the N_2O and HNO_3 variability comparisons), and move it to the Supplement. Moreover, we feel that Figure 31 on lat/p contours of short-term trend for various species can be moved to the Supplement, as it is less critical, and given past (and ongoing) work on this topic. While Figure 32 is interesting to us, it is more of a side note on lower stratospheric tropical cohesiveness for various species exhibiting similar dynamical variability, so we decided that the text and Figure in this case can be eliminated altogether without much of an impact on this paper.

In summary, the total number of Figures in the main text will be trimmed down by almost a quarter, with a more manageable total of 25 Figures; writing up a multi-year effort of (part-time) work on detailed model/data comparisons is bound to lead to a longer manuscript than several shorter analyses; to our knowledge, fits, correlations, variability, and trend comparisons are rarely investigated to this extent in model/data comparisons, even for a single model (or two flavors of one model). This, with some reductions (and clarifications) in the text (including the Abstract and Conclusions section), will at least show our good faith effort towards the referee comments. Recommending a goal of exactly 20 Figs. is rather arbitrary, but our point here is that we have considered these requests with some care, and that we are being responsive.

2) Differences are often only listed and not explored more in detail. To give one example, model HCl shows systematic differences in the lower stratosphere (evaluation based on Fig. 4) and a discussion relating those differences to shortcomings in the model transport or model chemistry would be interesting. Given the length of the manuscript, one could focus on the gases for which the detected differences are discussed in terms of model behavior (e.g., HNO_3). Differences for other gases can be mentioned in the manuscript with the according figures being moved to the supplement.

Reply: Yes, we pursued this type of reorganization, as explained above, with what we would consider a reasonable amount of delegating of material to the Supplement. We find some value in the remaining Figures, and feel that using a somewhat arbitrary number (such as 20) is not justified for a paper that covers a fair amount of ground and wishes to confront the models with a multi-species approach, in order to check for potential areas of weakness. Just stating good agreement and putting almost every Figure in the Supplement could work also, in principle, but that would be the other extreme, with a nearly complete lack of visual confirmation, which we think is important to preserve as well. Also, while we are striving to cut down on the length here, there are other long papers in the literature (but we will most likely avoid this sort of length in the future).

3) In section 3, existing evaluations of WACCM and the WACCM composition in particular should be discussed. Such references come up in the latter part of the manuscript. If they are given combined in this section, it will easier for the reader to identify what the current challenges are and what is new in this manuscript.

Reply: Yes, this section and/or the Introduction will be modified in the revised version (without adding too much length) to take this into account; in particular, we will add some motivation for the comparisons done here for FR-WACCM versus SD-WACCM (and observations).

Minor comments:

1) Consider changing the title to 'Evaluation of CESM1 (WACCM) free-running and specified-dynamics stratospheric composition simulations using global multi-species satellite data records.

Reply: Given that water vapor is considered all the way through the mesosphere, we prefer to stick to our original title, but we did consider this suggestion.

2) Page 5, line 31 – Page 6, line 2: This text could be moved to the discussion of the MLS data record in section 2.1.

Reply: While this could be done in principle, we feel that the species-specific discussions of error bars and validation work is really best kept as part of the discussions for each species, and that the flow is less awkward this way; we have thus not tried to reorganize these portions of text.

3) Page 7, line 24: Do you mean all earth system model or just WACCM with the term 'general model underestimation'?

Reply: We mean just the WACCM models here. This is clarified in the revised version by stating "model underestimation by both WACCM versions." However, it is implicit that other models without the proper (more complicated) chemical processes and energetic particle pathways will also underestimate HNO_3 in the same fashion.

4) Page 9, line 7 -10: Here, and also in other places, the sentence is too long for easy understandability. Consider splitting into two sentences at the semicolon.

Reply: Yes, we will start a new sentence instead of using a semi-colon, if/as that may help. We will also consider some other places for such an issue.

5) Page 12, line 5-8: The statement is made for the upper mesosphere. But isn't it also true for the stratosphere?

Reply: The statement (regarding worse diagnostic values) is somewhat true for the upper stratosphere as well, but we are mainly referring to SD-WACCM here; nevertheless, we have modified the revised text to state that the (SD-WACCM) diagnostics "are of poorest quality in the mesosphere" (etc…).

6) Page 13, line 7: MIPAS has been used earlier in the manuscript.

Reply: Yes, thank you; this is readily fixed by defining MIPAS earlier on in the text (in the 2nd part of section 4).

Revised Abstract:

We evaluate the recently delivered Community Earth System Model version 1 (CESM1) Whole Atmosphere Community Climate Model (WACCM) using satellite-derived global composition datasets, focusing on the stratosphere. The simulations include free-running (FR-WACCM) and specified-dynamics (SD-WACCM) versions of the model. Model evaluations are made using global monthly zonal mean time series obtained by the Aura Microwave Limb Sounder (MLS), as well as longer-term global data records compiled by the Global Ozone Chemistry and Related Trace gas Data Records for the Stratosphere (GOZCARDS) project. A recent update (version 2.20) to the original GOZCARDS merged ozone (O_3) data set is used.

We discuss upper atmospheric climatology and zonal mean variability using O_3, hydrogen chloride (HCl), nitrous oxide (N_2O), nitric acid (HNO_3), and water vapor (H_2O) data. There are a few significant model/data mean biases, such as for lower stratospheric O_3, for which the models at mid- to high latitudes overestimate the mean MLS values by as much as 50% and the seasonal amplitudes by ~60%; such differences require further investigations, but would appear to implicate a transport-related issue in the models. Another clear difference occurs for HNO_3 during recurring winter periods of strong HNO_3 enhancements at high latitudes; the strong model underestimate in this case (by a factor of about 2 to 6) stems from the (known) omission of ion chemistry relating to particle precipitation effects, in the global models used here. In the lower stratosphere at high southern latitudes, the variations in polar winter/spring composition observed by MLS are generally well matched by SD-WACCM, the main exception being for the early winter rate of decrease in HCl, which is too slow in the model. In general, we find that the latitude/pressure distributions of annual and semi-annual oscillation amplitudes derived from the MLS data are properly captured by the corresponding model values. ~~Nevertheless, detailed aspects of the interactions between the quasi-biennial, annual, and semi-annual ozone variations in the upper stratosphere are not as well represented by FR-WACCM as by SD-WACCM.~~

One of the model evaluation diagnostics we use represents the closeness of fit between the model/data anomaly time series, and we also consider the correlation coefficients. Not surprisingly, SD-WACCM, which is driven by realistic dynamics, generally matches observed deseasonalized anomalies better than FR-WACCM does. We use the root mean square variability as a more valuable way to estimate differences between the two models and the observations. We find, most notably, that FR-WACCM underestimates the observed interannual variability for H_2O by ~30%, typically, and by as much as a factor of two in some regions; this has some implications for estimates of the time needed to detect small trends, based on model predictions.

We have derived trends using a multivariate linear regression (MLR) model, and there is a robust signal in both MLS observations and WACCM of an upper stratospheric O_3 increase from 2005 to 2014 by ~0.2-0.4%/yr (± 0.2%/yr, 2σ), depending on which broad latitude bin (tropics or mid-latitudes) is considered. In the lower stratosphere, some decreases are indicated for 1998-2014 (based on merged GOZCARDS O_3), but we find near-zero or positive trends when using MLS O_3 data alone for 2005-2014. The SD-WACCM results track these observed tendencies, although there is little statistical significance in either result; however, the patterns of O_3 trends versus latitude and pressure are remarkably similar between SD-WACCM and MLS results. For H_2O, the most statistically significant trend result for 2005-2014 is an upper stratospheric increase, peaking at slightly more than 0.5%/yr in the lower mesosphere, in fairly close agreement with SD-WACCM trends, but with smaller values in FR-WACCM. As shown before by others, there are multiple factors that can influence low-frequency variability in H_2O; indeed, these recent short-term trends go beyond what one would expect from changes associated with a slow, secular increase in

methane. For HCl, while the lower stratospheric vertical gradients of MLS trends are duplicated to some extent by SD-WACCM, the model trends (decreases) are always on the low side of the data trends. There is also little model-based indication (in SD-WACCM) of a significantly positive HCl trend derived from the MLS tropical series at 68 hPa. These differences deserve further study. For N_2O, the MLS-derived trends (for 2005-2012) point to negative trends (of up to about -1%/yr) in the NH mid-latitudes and positive trends (of up to about +3%/yr) in the SH mid-latitudes, in good agreement with the asymmetry that exists in SD-WACCM trend results. The small observed positive N_2O trends of ~0.2%/yr in the 100 to 30 hPa tropical region are also consistent with model results (SD-WACCM in particular), which in turn are very close to the known rate of increase in tropospheric N_2O. In the case of HNO_3, MLS-derived lower stratospheric trend differences (for 2005-2014) between hemispheres are opposite in sign to those from N_2O and in reasonable agreement with both WACCM results.

The data sets and tools discussed here for the evaluation of the models could be expanded to additional comparisons of species not included here, as well as to model intercomparisons using a variety of CCMs, in order to search for systematic differences versus observations or between models, keeping in mind the range of model parameterizations and approaches.

Table 1. Pubs. in CCMI special issue (mostly ACP papers recently, with a variety of topics/thrusts).

Reference	Title	Type of study (model vs data, etc…)	Some novel aspects of atm. science?	Mostly model description or model analyses? > not much data
Jockel, P. et al. (2016), 10.5194/ **gmd**-9-1153-2016 **GMD**	Earth System Chemistry integrated Modelling (ESCiMo) with the Modular Earth Submodel System v-2.5	*One model* with different scenarios	Not really	Yes, model sensitivity (scenario) runs
Tilmes, S. et al. (2016), 10.5194/ **gmd**-9-1853-2016 **GMD**	Representation of the CESM1 CAM4-chem within the CCMI	*One model* (different scenarios) & some data	Not really	Model evaluation studies
Strode, S. A. et al. (2016), 10.5194/ **acp**-16-7285-2016	Interpreting space-based trends in CO with multiple models.	Model and data	Yes, in terms of model/data differences.	A combination of models and data
Morgenstern, O. et al. (2017), 10.5194/ **gmd**-10-639-2017 **GMD**	Review of the global models used within phase 1 of CCMI.	Descriptions of various CCMI models	Not directly	Model descriptions only
Fernandez, R. P. et al. (2017), 10.5194/**acp**-17-1673-2017	Impact of biogenic VSL bromine on the Antarctic O_3 hole during the 21st century.	*One model* and data - with model predictions	Not directly, but based on model predictions	Yes, mostly model predictions
Smalley, K. M. et al. (2017), 10.5194/ **acp**-17-8031-2017	Contribution of different processes to changes in tropical LS H_2O in CCMs.	Models and some data	Yes, based on model behaviors & inferences	Yes, mostly model analyses
Hardiman, S. C. et al. (2017), 10.5194/ **gmd**-10-1209-2017 **GMD**	The Met Office HadGEM3-ES CCM: evaluation of strat. dynamics, impact on O_3	*One model:* different simulations (FR vs SD)	Not directly	Yes, mostly model analyses and evaluations
Lin, M. et al. (2017), 10.5194/ **acp**-17-2943-2017	US surface O_3 trends & extremes (1980- 2014): quantifying the roles of rising Asian emissions, domestic controls, wildfires, and climate.	*One model* with data comparisons	Yes, based on one model's behavior & inferences	Mostly model inferences (with some data comparisons)
Maycock, A. C. et al. (2018), 10.5194/ **acp**-18-11323-2018	The representation of solar cycle signals in strat O_3- Part-2: Analysis of global models.	Mostly multi-model results	Not directly, mostly model dependence on inputs	Yes, mostly a model sensitivity study

Reference	Title	Type of study (model vs data, etc…)	Some novel aspects of atm. science?	Mostly model description or model analyses? >not much data
Morgenstern, O. et al. (2018), 10.5194/ **acp**-18-1091-2018	O_3 sensitivity to varying greenhouse gases and ozone-depleting substances in CCMI-1 simulations.	Multi-model description & consistency of responses to forcings	Not directly	Yes, a model sensitivity study
Revell, L. E. et al. (2018), 10.5194/**acp**-17-13139-2017	Impacts of Mt. Pinatubo volcanic aerosol on the tropical stratosphere in CCM simulations using CCMI & CMIP6 stratos. Aerosol data	*One model*. Sensitivity of T and O_3 response to volcanic aerosol data	Not directly	Yes, mostly a model sensitivity study
Hou, P. et al. (2018) **acp**-18-8173-2018	Sensitivity of atmos. aerosol scavenging to precip. intensity and frequency in context of climate change	Some data but mostly a prediction sensitivity study	Yes, but based on prediction sensitivities	Yes, mostly a model sensitivity study (with different met. fields)
Phalitnonkiat, P. et al. (2018), 10.5194/ **acp**-18-11927-2018	Extremal dependence between T and O_3 over the continental US.	Some data but mostly multi-model prediction	Yes, but based on model predictions	Yes, mostly a model sensitivity study
Orbe, C. et al. (2018), 10.5194/ **acp**-18-7217-2018	Large-scale tropospheric transport in the CCMI simulations.	Multi-model diffs.: AOA, transport.	Not directly	Yes, mostly a model sensitivity study
Wu, X. et al. (2018), 10.5194/ **acp**-18-7439-2018	Spatial and temporal variability of interhemispheric transport times.	*One model:* Variability of idealized tracers	To some extent, based on model sensitivity	Yes, mostly a model sensitivity study (of variability)
Dietmuller, S. et al. (2018), 10.5194/ **acp**-18-6699-2018	Quantifying the effect of mixing on the mean age of air in CCMVal-2 and CCMI-1 models.	Multi-model look: factors influencing AOA	Not directly	Yes, mostly a model sensitivity study
Dhomse, S. S. et al. (2018), 10.5194/ **acp**-18-8409-2018	Estimates of ozone return dates from CCMI simulations.	Multi-model estimates: O_3 return dates	Yes, but based on predictions	Yes, mostly a model sensitivity study
Ayarzaguena, B. et al. (2018), 10.5194/ **acp**-18-11277-2018	No robust evidence of future changes in major stratospheric sudden warmings: a multi-model CCMI assessment	Multi-model study of major strat. sudden warmings	Yes, based on model predictions	Yes, mostly a model sensitivity study

Reference	**Title**	**Type of study (model vs data, etc…)**	**Some novel aspects of atm. science?**	**Mostly model description or model analyses? >not much data**
Lamy, K. et al. (2018), ACPD, 10.5194/**acp**-2018-525	UV radiation modelling using output from the CCMI	Multi-model UVI versus climo UVI data	Yes, based on model results	A combination of models and data
Revell, L. E. et al. (2018) **acp**-2018-615	Tropospheric ozone in CCMI models and Gaussian emulation to understand biases in the SOCOLv3 CCM.	Multi-model comparison of tropos. ozone vs data	Mostly geared towards model refinements	A combination of models and data