Replies to reviewer 1

for the manuscript submitted by Chabrillat et al. to EGUsphere (doi:10.5194/egusphere-2025-1327)

We thank the reviewer for their constructive and helpful comments. Please find below the comments in **bold**, our replies in *italic*, and the manuscript modifications in red.

Minor comments

1.) Length of paper: Overall, the paper is a bit long. I understand that it is important to present and discuss the model performance for several different species. However, making the text more concise and reducing the number of figures could improve readability, potentially increasing both the paper's readership and its impact. I don't have special recommendations here but would just encourage the authors to think about such potential improvements.

We appreciate the reviewer's careful and thorough reading of the manuscript, which is evident from the constructive comments provided. We have carefully considered a major revision to shorten the text in response to this suggestion. However, we anticipate that most readers will access the paper via a web browser and navigate directly to sections of interest using the automatically generated Table of Contents.

The manuscript was deliberately structured to provide a comprehensive and balanced evaluation, which both reviewers have welcomed. Although lengthy, this evaluation will be useful to assess the strengths and weaknesses of the stratospheric composition products provided by IFS-COMPO to CAMS. A substantial reduction in length, whether through text cuts or relocating figures to the Supplementary Material, would risk compromising this balance or result in only marginal gains in overall length. For these reasons, we respectfully propose to retain the current structure and level of detail.

2.) I'm wondering why the model moist biases in the extratropical lowermost stratosphere (below about 100hPa) are not discussed at all (around L542). These are the largest biases in the profiles shown in Fig. 5, and are similar to known moist biases in climate models (e.g. Charlesworth et al., 2023, https://doi.org/10.1038/s41467-023-39559-2), and in IFS have recently been shown to contribute to UTLS cold biases (Bland et al., 2024, https://doi.org/10.1002/qj.4873). I'd find it good to discuss these issues briefly here.

The following paragraph was added at the end of section 5.2:

> Most climate models suffer from large moist biases in the extratropical lowermost stratosphere, i.e. below about 100hPa, likely due to difficulties modelling transport of water vapor near the tropopause with a strong gradient (Charlesworth et al., 2023). This issue also impacts humidity in the lowermost stratosphere of IFS, contributing to a cold bias in the NWP-oriented configuration (Bland et al., 2024) and explaining the large overestimation shown by Fig. 5 in the mid-latitudes below 100hPa.

3.) Scorecard grading: I really like the summary of results in the scorecard in Sect. 7. But I'd suggest to be somewhat more careful with giving particularly high scores here, given the remaining biases in parts of the profiles. Such high scores could be misleading if quick readers don't look into specific details in the related subsections. A few examples where I'm sceptical about the choice of score are:

- **Fig. 19, U.S.: CH4, H2O, O3, ... show significant biases above about 10hPa (Figs. 4, 5, 9), so that I'm unsure whether "good performance" is suitable here.**
- **Fig. 19, Tropical M.S./O3: Also in the tropical profile (Fig. 9) the bias increases above 10hPa, such that I wouldn't rate the performance "very good".**
- **Fig. 19, Mid-lat. M.S./H2O: For H2O the mid-latitude correlation in Fig. 5 is very low, so that also here I'm wondering about the "good performance".**
- **Abstract, L30: "very good performance for O3, HC4, N2O and H2O..." perhaps too strong given the remaining biases in parts of the profiles.**
- **Conclusions, L1019: "very good performance for CH4, N2O and H2O" I find too positive.**

Related to these comments, it's not obvious to me that ACE-FTS is the better reference dataset for stratospheric water vapor (as chosen in Fig. 19 grading). MLS also provides a very good stratospheric water vapor product, and compared to MLS the IFS biases are generally larger.

We attempt to assess the initial performance of new forecast products (here stratospheric species beyond ozone). This is less straightforward as providing a scorecard for relative improvements between two consecutive versions (e.g. Eskes et al., 2024). We aim for an objective attribution of scores by selecting numerical criteria on bias, standard deviations and correlations (table S1) but those criteria are subjective themselves. You are correct in pointing out that initial scores should not be too optimistic, as this would prevent highlighting future model improvements in the evaluation of future model cycles.

We have carefully re-examined our criteria while considering these examples. As indicated by the comment about stratospheric water vapor, the issue did not lie as much in the choice of criteria as in the choice of only one reference dataset to apply them. The second column of Fig. 19 indicated the dataset chosen for performance evaluation, but this was not mentioned in the text and the rationale for this choice was not explained. ACE-FTS was chosen in most cases because it agreed better with the model, leading to more optimistic scores than allowed by a visual inspection of figures 4 to 9.

The revised manuscript thus keeps the same scoring criteria but requires two datasets in agreement to attribute the "very good" scores. As indicated by the Table S1 and the new table S2 in the Supplement, this is achieved by computing a simple score for each dataset (using the same criteria as before) before computing their sum for the species available in both datasets. A "very good" performance assessment thus requires the availability of, and very good agreement with, both datasets. This is outlined in the Supplementary material (see revised table S1 and new Table S2) and explained in the text as follows:

> The regional scores are determined objectively from the absolute values of the Normalized Mean Bias (NMB), Standard Deviations of differences between model and observations (STD) and corresponding correlations, using criteria chosen to segregate between the four proposed scores while prioritizing bias performance (see Table S1 in the Supplement). These scores are computed separately for each reference dataset (second column: "A" for ACE-FTS; "M" for Aura-MLS) and added for the species where both datasets are available. The total score provides a combined performance assessment, requiring the availability and agreement with both ACE-FTS and Aura-MLS to allow "very good performance"

assessment (see Table S2 for details). The assessment for N_2O in the lower and middle stratosphere relied only on ACE-FTS because Aura-MLS shows suspiciously large disagreements at pressures larger than 10 hPa while difficulties were reported in the retrieval of Aura-MLS N2O v4 (see section 3.3).

The resulting scorecard (Figure 19) has been simplified and updated accordingly:

Species	Ref. data	Tropical L.S. (70-150hPa)	Tropical M.S. (6-50hPa)	Mid-lat. M.S. (10-100hPa)	U.S.	Polar M.S. (6-30hPa)	Winter-spring L.S. (30-100hPa) N.P		S.P.		Ref. data
CH_4	A	n	+	+	n	n	/		/		/
H_2O	A,M	--	+	n	n	+	++		+	my	BRAM3
HCl	A,M	-	n	-	+	+	+	y	-	my	BRAM3
ClO	M	*0*	*0*	-	-	-	+	y	n	m	BRAM3
N_2O	A,M↑	-	+	+	-	n	+	m	+	my	BRAM3
HNO_3	A,M	*0*	-	-	--	-	+	y	n	my	BRAM3
N_2O_5	A	*0*	--	n	n	n	/		/		/
NOx/NO_2	A	-	n	-	-	n	+	m	+	m	MIPAS-REAN01
$ClONO_2$	A	*0*	n	n	/	n	/		/		/
BrO	H	/	/	/	/	+	/		/		/
$BrONO_2$		/	/	/	/	/	+	m	+	my	MIPAS
O_3	A,M	+	+	++	n	n	+	y	+	my	BRAM3

Note that all explanatory annotations were moved to the caption or to the main text due to a comment by the second reviewer.

As intended, the revised scorecard is less optimistic for the species which you commented about:

- ***CH_4** : the performance is downgraded from "good" to "neutral" in the tropical L.S. and from "very good" to "good" in the tropical and mid-latitudes M.S.*
- ***H_2O** : the performance is downgraded from "good" to "neutral" in the mid-latitudes M.S. and in the U.S.*
- ***HCl** : the performance is downgraded from "good" to "neutral" in the tropical M.S. and from "neutral" to "poor" in the mid-latitudes M.S.. Interestingly, the revised score is upgraded from "neutral" to "good" in the polar M.S. due to agreement between both datasets.*
- ***N_2O** : the performance is downgraded from "very good" to "good" in the extra-polar M.S. and from "neutral" to "poor" in the U.S.*
- ***HNO_3** : the performance is downgraded from "neutral" to "poor" in the mid-latitudes and polar M.S.*
- ***O_3** : the performance is downgraded from "very good" to "good" in the tropical M.S. and from "good" to "neutral" in the U.S. and polar M.S.*

The conclusions and abstract were updated accordingly.

Specific comments:

L138: How is the volcanic injection of sulphate species treated in the model? Would be good to mention here or point to the relevant place in the paper.

This helpful comment led to further clarification of the development history of IFS-COMPO, as the Pinatubo simulation experiment (rd.i9vv in Table 2) required a minor improvement of volcanic injection which was developed after the release of Cy49R1. Strictly speaking, the evaluation of this experiment thus concerns Cy49R2. This led to a minor modification of the title of the paper and the title of section 2.3, where "IFS-COMPO Cy49R1" is replaced by "IFS-COMPO Cy49". The treatment of volcanic injection is now described, first in section 2.1 (for Cy48R1):

> The volcanic injection in Cy48R1 concerns only SO_2 and is carried out only over a single grid cell, i.e. each volcano is treated as a point source. The injection specifics (amount injected, latitude/longitude of the volcano, times of beginning and end of injection, minimum and maximum injection altitude) are prescribed in a model namelist. The model determines the matching grid cell and model levels and distributes the injected amount equally between the model levels.

…and at the end of section 2.3 (for Cy49R2):

> Volcanic injection was further refined in IFS-COMPO Cy49R2, allowing injection over areas that comprise multiple grid cells and enabling the injection of water vapour alongside volcanic sulphur dioxide. These enhancements support the modelling of the impacts of the Hunga (2022) and Pinatubo (1991) eruptions, respectively.

Throughout the text, the cycle numbers were corrected from Cy49R1 to Cy49R2 where necessary or simplified to "Cy49" where appropriate.

The injection data for the Pinatubo simulation experiment was described in section 4.3 and has been clarified:

> For the Pinatubo eruption, a total of 14 Tg of SO_2 was injected on 15th of June 1991 between 18 and 24 km altitude (Sukhodolov et al., 2018). To better take into account the explosive nature of the eruption and local dynamical processes not described by the model, the injection was distributed over a 300 × 300 km area centered on the Pinatubo. The additional impact of the Cerro Hudson eruption is captured by also injecting 2.3 Tg of SO_2 on 15 August 1991, over a 300 × 300 km area centered on the Cerro Hudson.

L212 (Fig. 1): I don't understand the distinction between SO2 in CB05 and BASCOE. Please clarify in caption or text.

This comment also led to fruitful discussions between the co-authors, resulting in a major revision of Fig.1. While some fields are duplicated between the NWP core of IFS and its COMPO extension (e.g. GO3 and O3; q and H2O), there is no such duplication between the CB05 and BASCOE modules. These modules compute increments for the same fields, but differently depending on the location of each gridpoint in the troposphere or in the stratosphere. It was thus misleading to distinguish between SO2 in CB05 and SO2 in BASCOE. Fig.1 was revised to distinguish the conversion processes activated in the troposphere (module CB05) from those activated in the stratosphere (module BASCOE). This revision also led to a correction of the production process for sulfate in the troposphere, which is converted from SO3 rather than SO2.

Here is the revised Fig.1 and its caption:

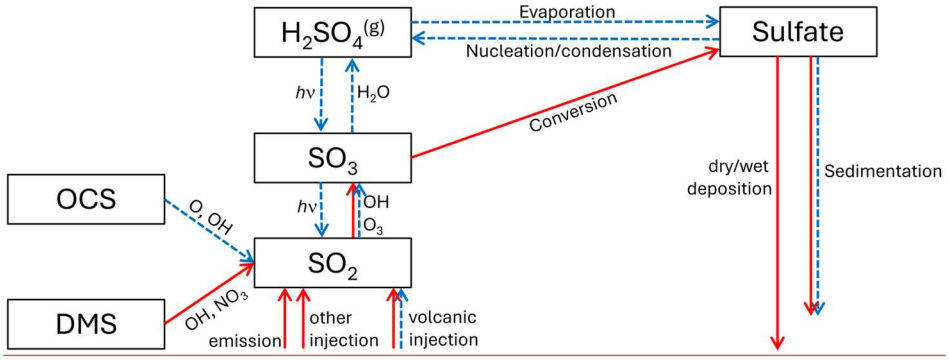

Figure 1. Architecture of the stratospheric sulphur cycle of IFS-COMPO as implemented in Cy49R2. Processes may be activated in the troposphere (red arrows), in the stratosphere (dashed blue arrows) or in the whole column (double arrows).1

L225: The variable "c" in Eq. 11 needs to be explained.

c is an adjustable parameter and its adjustment is explained on line 235-237. This has been clarified at L.225.

L403: I agree that the highest mean age values in the stratosphere are below 10 years. However, age spectrum tails extend well beyond. Hence, the statement "oldest air encountered in the stratosphere" is not correct and should be changed.

Done. We simply clarified by using the correct name i.e. "mean Age of Air":

> IFS-COMPO was thus spinned up during 10 years before this test case, i.e. for a longer time than the largest mean Age of Air encountered in the stratosphere (Chabrillat et al., 2018).

L507ff: How is the upper boundary condition treated? Can't this also be a source of bias in the upper stratosphere? Please add some explanation and discussion here.

The upper boundary condition is simply "null flux" for all tracers. This is justified by the location of the uppermost model at 0.01 hPa, i.e. in the mesosphere and far above the highest layer of interest and evaluated in the paper (upper stratosphere, 1-10 hPa). This has been clarified in the text:

> In the upper stratosphere, i.e. at pressures lower than 10 hPa, the N_2O biases between IFS-COMPO and ACE-FTS increase quickly to reach or exceed 50% at the upper limit of our evaluation (1 hPa pressure level). (…) Since the uppermost model level is at 0.01 hPa pressure, i.e. in the mesosphere and approximately 30 km above the upper limit of our evaluation, the upper boundary condition is not expected to play a role in this disagreement. This suggests that a common process…

L524: Adding age of air tracers to IFS would indeed be very interesting for future work.

Indeed: this led us to repeat this suggestion in the first bullet of the conclusions.

L890: What means "By elimination..." here?

This discussion paragraph was not clear and it did not contribute much to the evaluation. It was thus deleted from the revised version.

Technical corrections:

L72: ... CAMS was upgraded .... → *done*

L101: Lagrangian → *done*

L142: blank between "aerosols as" → *done*

L197: Cy48R1 - there are also other places where the "R" is lower-case (e.g. L201, L375, etc). Please check the entire manuscript again. → *done*

L307: one "solar" too much. → *done*

L403: "oldest age" sounds awkward, better "oldest air" or "largest/highest age values", etc.
→ *done:* for a longer time than the largest mean Age of Air

L415: Would change "Let us compare..." to "In the following, we compare ...", or similar.
→ *This short "linking" sentence sounded awkward and was not necessary. It has been removed.*

L431: Would change "It will be interesting to see how..." to "It is a particularly interesting question how ...", or similar. → *done:* It is an interesting question how…

L664: blank missing "inthe". → *done*

L691: number missing: "~150 and ~hPa". → *done: this sentence has been corrected and clarified by following increasing pressures (rather than altitudes). It now reads:*

> The vertical profiles of simulated extinctions match relatively well the retrievals, especially the constant or slow increase of retrieved extinction with increasing pressures from ~5 hPa to ~30 hPa and the stronger increase from ~30 hPa to ~150 hPa.

L748: "in the two control runs". → *done*

L804: blank missing "afterwardan", and better "afterwards ..." → *done*

L884: blank missing "quitesimilar" → *done*

L894: Just simplify to: "To conclude, IFS-COMPO ..." → *done*

L939: "stratospheric" → *done*

L962: blank missing "theagreement" → *done*

L983: blank missing "thisunderestimation" → *done*

L987: "Northern" → *done*

L1044: "spring" → *done*

Figures 4, 5, 6, 7, 9: Is the legend labelling of red solid and dashed lines correct? I guess the solid line should be Cy48... (not Cy49...), as for the blue lines?
→ *Indeed, the red solid lines are for Cy48R1 and not Cy49R1. This is now corrected.*
Thanks for spotting this!