Reviewer 1:

The paper by Abalos et al. compares stratospheric transport in three generations of climate models (CCMVal-2, CCMI-1, CCMI-2022) and evaluates different transport characterictics by comparison to satellite observations. The authors show that several long-standing model biases persist across the generations of models, even worsening in the most recent simulations for some transport features. In particular, stratospheric mean age of air is too low indicating too fast transport in the models, and this bias is largest for CCMI-2022. Also, the spring-time polar vortex break-up (final warming) date is delayed, strongest in most recent models, and this appears to be related to an overestimation of the ozone minimum in the simulations. Long-term trends show a robust acceleration of the stratospheric circulation for all model generations, with the trends before 2000 likely related to ozone depletion.

Overall, I find this a great model intercomparison paper which presents and evaluates different sophisticated and detailed transport diagnostics. The paper clearly falls within the scope of the journal and will be of much interest to a broad readership. Moreover, the paper is very well written, the results are concisely presented and the figures are clear and high-quality. In my opinion, particularly the identification of climate model biases and their evolution over different generations of models is urgently needed. And here the paper is doing a great job, pinpointing these biases in a very clear manner and discussing their causes and impacts, to the degree possible within such an extensive intercomparison. In summary, I do strongly recommend publication and have only a few specific and technical comments, which hopefully help to further improve the paper.

We are thankful to the reviewer for their positive comments and careful review of the paper. We provide responses to all individual comments below. In addition, we note a few changes to the paper in the revised version.

1) Two new coauthors have been included, responsible for providing data for one of the models and for the RID dataset. 2) The RID dataset has been updated (corrections to the uppermost levels ~0.01 hPa), and the figures changed very slightly (almost imperceptibly in most cases). 3) In Figures 18 and 19, the panels showing the CCMI-2022 trends over the ozone depletion and recovery periods are now represented for the residual timeseries, after removing the variability arising from ENSO, the QBO, the solar cycle and the aerosols with a standard multiple linear regression analysis. The regression has been applied to the period 1980-2018 and then the trends of the residual are computed for the two separate periods. The results are highly consistent and only slight quantitative differences are found in the 2000-2018 trends, which in any case remain much smaller than the 1980-1999 trends, and less robust across models. This has not been applied to the panels showing trends for the three generations, because the available period is too short to carry out the regression. 4) Figure 19 now includes the w* trends in the multi-reanalysis mean, with inter-reanalysis agreement indicated by stippling. We changed the stippling to indicate where all the models / reanalyses (instead of 70% as it was before) have a trend of the same sign. Even if the reanalysis trends are noisy, there broad conclusions obtained for the models (trends larger before than after 2000, and larger in the SH than in the NH) are supported by the multi-reanalysis mean trends.

Specific comments:

L128ff: From the following two paragraphs it seems important that different merged satellite data products are used for this study. If so, it would be helpful to describe here more clearly the merging techniques, and in particular differences between the data products.

We added more information on the differences and similarities between the two merged datasets and that we use both to demonstrate the degree of observational agreement in the context of comparisons with models.

L160: I'd mention already here "... version 3.5/3.6, as proposed by Saunders et al. (2025, 10.5194/acp-25-4185-2025)."

Added: " version 3.5/3.6 (February 2004–February 2021) and version 4.1/4.2 (February 2004–present), as proposed by Saunders et al. (2025)."

L288ff: The increase in age of air bias for the newest model generation is remarkable. Any ideas/hypotheses regarding potential causes? This could be briefly added here or in the discussion/conclusions section.

In the following sections there is discussion on possible causes. We moved to the end of this section the discussion on the families of models included in each generation and the possible impacts on the results. In the discussion, we added a mention to the Orbe et al. (2025): "Nevertheless, note that the age bias could have different origins in the various models. For instance, a recent study relates the young bias in the newest generation of the GEOSCCM model compared to its predecessors to a bias in tropical upwelling arising from differences in the transport scheme (Orbe et al. 2025)" .

Figures 3, 4, 9, 10, 18, 19, 20: The yellow lines for CCMVal-2 are very hard to see in some cases.

We have corrected this issue in Figs. 1, 3, 4, 6, 9, 10, 18, 19 and 20.

L324: I'm somewhat unsure about "The mass flux in our observational references...". Can reanalyses really be seen as observational references for upward mass flux? Since the residual circulation is not directly constrained by data assimilation, and substantial differences in its strength and structure are found among reanalyses (e.g. Abalos et al., 2015, 10.1002/2015JD023182; Fujiwara et al., 2024, 10.5194/acp-24-7873-2024), a brief discussion of these discrepancies could be appropriate here.

We have changed this to "observation-based references, the reanlyses, ...". A discussion on the caveats of residual circulation derived from reanalyses is included in section 2.2.2.

L338: "... the reduced spread is a common feature of the newest model generation found across metrics." Isn't Fig. 1 rather showing larger spread in age for the newest models? Please clarify what is meant here.

That is correct, we have changed this to " the reduced spread is a common feature of the newest model generation found across several metrics (but not in all)."

L353ff: I'm wondering about the comparison between mass flux results in Figs. 3 and 4. Why are relative differences between models different in the two diagnostics? For instance, in Fig. 4 the CCMI-2022 models show strongest upwelling throughout the profile while in Fig. 3 this is not the case in some layers (e.g. below 70hPa). It would be good to provide some further explanation to avoid confusion.

We have added a discussion on the discrepancies in the models (not involving the observational uncertainties: ", which is not observed in Fig. 3a. These differences between the overturning circulation results obtained with the two methods imply that they are not directly comparable. The inconsistencies are not due to the fact that different models provide residual circulation and mean

age output (Figs. 3S and 4S). Rather, they are likely due to the very different approaches, variables and calculations involved in each method. A comparison between residual circulation and age gradient mass flux was carried out in Linz et al. (2019), but was based on variability, not climatology." And then at the end of the Fig. 4 description we added: "Overall, we conclude that the overturning mass flux estimate from the age gradient should not be directly compared to residual circulation-based diagnostics."

L389: It seems to me that the upwelling at lowest stratospheric levels is for some models faster than in ERA5. So I don't fully see the "consistency" between upwelling and CPT differences mentioned here.

This sentence states: "Note that ERA5/5.1 has a colder CPT compared to the other reanalyses (by ∼1 K), consistent with the faster tropical upwelling seen in Fig. 3." So it is not comparing ERA5 to the models, but to the other reanalyses, which indeed have weaker upwelling at the lowest levels.

L439: Any idea why the inter-model spread in mixing efficiency changes so much between model generations?

We think this is related to an outlier in the mean age of air in the earlier generation (yellow dot in the right-upper corner of Fig. 2b). The spread is most different in CCMVal-2, due to the UMUCKA-UCAM, which has a very large mean age (see Fig. 2S), while the spread is more comparable in CCMI-1 (~0.4) and CCMI-2022 (~0.3).

L637: I find the reduced spread in age trends at lower levels for CCMI-2022 particularly interesting. Any idea why?

It is indeed interesting, and it is clearly seen in Fig. 38S. Unfortunately we do not have an explanation for this.

L650: The CCMI-2022 models here (Fig. 18 b/c) show a flip in the hemispheric difference of age change, with more negative trends before 2000 in the SH and afterwards in the NH (c.f. Strahan et al., 2020, 10.1029/2020GL088567; Ploeger and Garny, 2022, 10.5194/acp-22-5559-2022). Some related discussion of the robustness of hemispheric differences in age trends in different model generations could be interesting.

Note that we have now regressed out the most important variability modes (ENSO, QBO, solar, aerosols) before computing the trends for the pre and post-2000 periods for CCMI-2022 (the only one extending long enough after the year 2000 to compute trends) in Fig. 18b and cz. In the new results, the pre-2000 trends remain practically identical, but the 2018 trends change slightly, although they remain much smaller than those for the earlier period. The mentioned papers find a short-term trend towards relatively older air in the NH compared to the SH, maximizing over the period 2004-2017. We do see a hint of that interhemispheric pattern, which has been shown in refC2 simulations in Ploeger and Garny (2022). We added a sentence on this: "In the post-2000 period the much smaller trends still display an asymmetry, being more negative in the SH than in the NH, consistent with previous studies Ploeger et al. (2022). " We also added, relative to the robustness of the interhemispheric difference across renalyses (for the 1980-1999 period): "An additional proof that the Antarctic ozone hole is behind the large trend before 2000 is that the negative AoA trends are larger in the SH hemisphere over the earlier period. This is consistently seen in the three generations (not shown)."

Figure 19 a-c: Are the reanalysis trends significantly different from zero? Given the usually strong inter-annual variability in reanalysis upwelling, I guess that this is not the case at all levels. Adding a significance indicator to the plot would be helpful for interpretation of differences.

Thank you for pointing this out. The statistical significance of the reanalyses trends is indicated by the closed circles, and indeed it is very limited. We forgot to add this to the caption, we now added: “The only statistically significant trends in the reanalyses are those marked with a closed circle, based on a Student's t test with a 95% confidence level.”

L703: Given the fact that the newest models (CCMI-2022) show only very weak correlation between age and mixing efficiency changes (r=0.25, Fig. 21b), are the results here really robust? A more detailed discussion could be helpful.

We agree and we note that the result for the changes is less robust than that for the climatology. We now added “Also, the correlation with mixing efficiency is low in CCMI-2022 (0.25)”, and changed “confirms” to “suggests” in “This suggests that the mean age changes are more strongly connected with changes in mixing or diffusion than with the changing strength of the residual circulation.”

Figure 23: For better comparability, I'd find it better to use a common reference period for all datasets used (e.g. 1995-2005).

We redid the ozone timeseries plots in the main text and Supplementary material using the 1990-2000 as a common reference period (CCMVal-2 onlyl goes until 2000 in many models).

Technical corrections:

L192: Reference not properly linked.
corrected
L284: gin situ --> in situ
corrected
L348: Maybe better "satellite-based mean age data"?
changed
Figure 7, caption: add "(black line)" after "comparison".
Added (actually gray contours)
L517: "...will be discussed..."
corrected
L547: "...can imply"
corrected
Figure 16: An entry for ACE is missing in the legend.
added
L679: Is it really meant that differences between profiles in Fig. 20b and Fig. 19b at upper levels are less than 0.5 percent, or is it rather meant that trend values differ by less than 0.5 percent per decade?
The latter is correct. We have changed the wording to clarify this. Now it reads: “trend values differ less than 0.5%/decade”.
L690: Check the wording following "... and there is evidence ..."
corrected
L696: "contributions" to what?
The contribution of changes in mixing efficiency to the future changes in AoA. We have rephrased the sentence to: “Eichinger et al. (2019) show that the contribution of changes in mixing efficiency to the future changes in AoA in CCMI-1 refC2 simulations present a large spread across models,

varying between 10 and 30%.”
L709: Bracket after "Figure 22"
removed
L722: Check wording in "...and this it is likely ..."
corrected

Reviewer 2

Review of “Evaluation of stratospheric transport in three generations of Chemistry-Climate Models” by M. Abalos et al.
This paper summarizes and discusses the results of a very comprehensive comparative evaluation of stratospheric transport in three generations of chemistry climate models that participated in successive model intercomparison initiatives over the last 15 years, including the latest one, CCMI-2022. The authors analyze several direct and indirect metrics of transport, including but not limited to the advective component of the BDC, several mixing diagnostics, wave activity, and polar vortex breakup dates. They analyze the seasonality as well as long-term trends of these metrics against satellite and reanalysis-derived data. One disappointing conclusion is that well-known biases continue to haunt generations of CCMs, and in some cases, such as the age of air, the latest models perform worse than their predecessors. This is an important result pointing to something that we as a community need to tackle.
I don’t often say this in a referee report but this is an excellent and important paper. It is very well written, comprehensive, and logically organized. On the first read on several occasions I found myself thinking “this result merits more discussion” only to find more discussion in the next paragraph. As a reviewer it is my job to find problems with the paper but I came up with almost nothing;) I only have a few very minor suggestions, mostly editorial. I recommend that this work be promptly published after a few technical corrections.
Congratulations to the Authors!
Kris Wargan

Thank you for this very positive review and for the constructive comments! We provide responses to all individual comments below. In addition, we note a few changes to the paper in the revised version.

1) Two new coauthors have been included, responsible for providing data for one of the models and for the RID dataset. 2) The RID dataset has been updated (corrections to the uppermost levels ~0.01 hPa), and the figures changed very slightly (almost imperceptibly in most cases). 3) In Figures 18 and 19, the panels showing the CCMI-2022 trends over the ozone depletion and recovery periods are now represented for the residual timeseries, after removing the variability arising from ENSO, the QBO, the solar cycle and the aerosols with a standard multiple linear regression analysis. The regression has been applied to the period 1980-2018 and then the trends of the residual are computed for the two separate periods. The results are highly consistent and only slight quantitative differences are found in the 2000-2018 trends, which in any case remain much smaller than the 1980-1999 trends, and less robust across models. This has not been applied to the panels showing trends for the three generations, because the available period is too short to carry out the regression. 4) Figure 19 now includes the w* trends in the multi-reanalysis mean, with inter-reanalysis agreement indicated by stippling. We changed the stippling to indicate where all the models / reanalyses (instead of 70% as it was before) have a trend of the same sign. Even if the reanalysis trends are noisy, there broad conclusions obtained for the models (trends larger before than after 2000, and larger in the SH than in the NH) are supported by the multi-reanalysis mean trends.

Specific comments
LL27-28 This is something that people often say but it is not quite correct. As Brasseur and Solomon (2005) explain in Section 5.2.3:

Thank you for pointing this out, we understand that it is not quite correct to state it as it was. We have modified the sentence to avoid this inaccuracy. Now it reads: “Stratospheric transport is also responsible for redistributing ozone from its source region to the high latitude lower stratosphere.”

72. I suggest moving the parenthetical clause to right after “CCMs”, i.e. “…the last three generations of CCMs (CCMVal-2, CCMI-1, CCMI-2022) in representing…”
We agree, changed.

L121 and L124: Which is it, 68.1 hPa or 100 hPa? According to the MLS Data Quality Document it is 100 hPa in version 5.
Indeed, it is 100 hPa according to the Data quality document (Livesey et al. 2022), we now cite this document. Thank you for spotting this.

L192. Missing citation?
Corrected

L194. CO2 -> CO_2.
Corrected

L262. Does this mean that you do subsample the model output for comparisons with ACE-FTS?
No, we only subsampled one of the models (WACCM) as a test, as shown in the supplementary material (Fig. 11S). Based on the tests that we did, we decided to use monthly mean model output so we did not subsampled it along the trajectories. We have changed the order of the sentences and added a new sentence at the end to clarify what we do: “Thus, based on these tests we decided to use the monthly mean output to reduce the computation burden.”

L267. I suggest stating here what metric is used for wave activity (45º–75º eddy heat flux at 100 hPa).
Added

L280-281. Is it worth noting that this shift toward younger AoA was also reported and investigated in detail in the latest versions of the GEOS model (Orbe et al., 2025)?
Yes, indeed, thank you for this reference, we now mention it when discussing the evolution in the different models in Section 3.1: “Note that the shift toward a younger mean age in GEOSCCM models was recently highlighted and investigated in Orbe et al. (2025).” and in the the discussion: “Nevertheless, note that the age bias could have different origins in the various models. For instance, a recent study relates the young bias in the newest generation of the GEOSCCM model compared to its predecessors to a bias in tropical upwelling arising from differences in the transport scheme Orbe et al. (2025).” We also mention it as an example of the individual model studies that are recommended in the discussion.

L284. “gin” -> “in” ;)
Changed, thanks!

LL337–338. Not across all metrics, right? Figure 1 and discussion shows that CCMI-2022 has the largest spread. Same is seen in Fig. 4.
That is true, we changed this sentence to recognize this.

LL341-365 and Fig. 4. I understand that one advantage of deriving the overturning mass flux in this way is that this diagnostic can be computed from satellite data. However, as you say, it comes with some serious caveats. It’s not clear to me how useful this discussion

is in the context of this study. I suggest either dropping it or making a clearer case for keeping it in.
We hesitated about keeping it because of the caveats. However, as it is a relatively novel metric, we decided to keep it to examine how it compares with other metrics. We added: “The newest generation shows larger values throughout the altitude range, which is not observed in Fig. 4a. These differences between the overturning circulation results are not due to the fact that different models provide residual circulation and mean age output (Figs. 3S and 4S). Rather, they are likely due to the very different approaches, variables and calculations involved in each method. A comparison between residual circulation and age gradient mass flux was carried out in Linz et al. (2019), but was based on variability, not climatology.” And at the end of the discussion on that figure: “Overall, our analyses show that the overturning mass flux estimate from the age gradient should not be directly compared to residual circulation-based diagnostics.”

L383. Why not make 5f a separate figure (a new Figure 10)?
This would make sense because it is a mixing diagnostic. However we prefer to not increase the number of figures, which is already quite large. Since it is a diagnostic based on the tape recorder, which is shown in Fig. 5, we decided to keep it there, although it is discussed forther down.

LL481-484. Nice! I really appreciate your considering these possible sources of the discrepancy.
Thank you.

LL495-497. Since you’re not defining EHF and EP flux here I suggest citing some generic literature or, alternatively, adding the definition of eddy heat flux ($\overline{V'T'}$) to Table 4, although that would require defining the primes and overbar.
We added a reference to Andrews et al. (1987).

LL525-526. I don’t understand how weak downwelling contributes to the young bias in AoA. As it’s written my first reaction is that it should be the opposite: slow BDC in the polar regions would lead to older air.
Thank you for pointing this out. This can be understood by interpreting mean age as a tracer, as done in several previous studies (e.g. Ploeger et al. 2015, doi:10.1002/2014GL062927). In the polar regions the downwelling brings old air downward, so a weak polar downwelling implies relatively younger age a lower levels, compared to the situation with strong downwelling. However, we realize that this sounds counterintuitive because a faster circulation generally implies a younger mean age. We believe that the contradiction arises from the alternative local (tracer-like) versus integrated (along the trajectory) views of mean age. In any case, in our case the mean age young bias is seen across al latitudes, not just in the SH polar region. Moreover, other analyses suggest that mixing plays a more important role than advection in the mean age biases (although this may differ from model to model). Thus, we have decided to remove this sentence to avoid confusion.

LL536-538 What are “common models”?
This was mentioned at the end of the sentence, we now moved it to right after the mention to “common models”: “, which are the 8 model families introduced in Tables 1-3”

Fig. 12. I suggest adding one sentence in the caption explaining that in the NH/SH large positive/negative EHF means more wave activity.
We added: “Note that positive fluxes in the NH and negative fluxes in the SH both indicate poleward eddy heat transport and upward wave propagation.”

L684. “temperature … is cooling”. That’s a pet-peeve of mine: temperature can’t be cool or warm. It can be low or high. I suggest rephrasing.
That is correct, we changed it to “the cold point tropopause is cooling”

L709. Typo: drop “]”.
Removed

L750. Again, some metrics show the largest spread in CCMI-2022.
Agreed, we removed this sentence.

L815-816. I suggest adding two citations here: APARC 2025 for Hunga and Solomon et al., (2023) for wildfires. The first one, released in December, can serve as a canonical reference for the impacts of the eruption, the latter explicitly talks about the increasing importance of smoke injection from wildfires under climate change.
Thank you, added.

Tables 1–3. Would it be possible to include the number of simulations for each model?
We prefer not to include this to keep the tables simpler, as this is information is not crucial for our study which is mainly focused on the climatology.

Figures. Sometimes it’s difficult to see the yellow line (CCMVal2 MMMed) against the shading. See Figs 3, 4, 6, 9, 10, 18a, 20, 19a-c. Other similar figures look fine, e.g., Fig. 11. Appendix.
We have corrected this issue in Figs. 1, 3, 4, 6, 9, 10, 18, 19 and 20.

I don’t think it matters at this stage but Tables A1 and 2 are before the list of references and Table A3 is below the references section.
We hope this will be fixed in the typesetting phase by the journal.