

General comments

The authors aim to investigate the response of stratospheric transport to climate change by developing an integrated diagnostic framework that combines CMIP6 mean AoA, TEM diagnostics, zonal-wind tendency terms, and Aura MLS water-vapour observations.

However, in its present form, the manuscript contains several methodological and interpretative issues that affect its central conclusions. In particular, I have concerns regarding the interpretation of the idealised 1pctCO₂ experiment; the construction of the multi-model means is not sufficiently documented; period-mean climatologies are repeatedly interpreted as statistically significant trends; and the MLS correlation analysis does not provide an independent observational constraint on greenhouse-gas-driven circulation changes. Furthermore, the assessment of statistical significance does not appear to account adequately for temporal autocorrelation, smoothing, lag selection, or multiple testing.

I therefore consider that the main claims concerning enhanced lower-stratospheric “transport efficiency” and its greenhouse-gas-driven mechanism are not yet sufficiently supported by the evidence presented. Substantial reanalysis, clearer quantification of uncertainties, and a more cautious interpretation of the results are required before the manuscript can be considered for publication.

Specific comments

1. Interpretation and temporal coverage of the 1pctCO₂ experiment

Regarding the description on page 6, Table 3 lists the simulation period as 1850–1999, whereas Fig. 6 and the associated discussion refer to the period 1975–2014. The authors should clarify how the years 2000–2014, which fall outside the stated simulation period, could be included in the analysis as currently described.

More fundamentally, 1pctCO₂ is an idealised experiment in which atmospheric CO₂ concentrations increase by 1% yr⁻¹ from a pre-industrial control state, while other forcing agents do not follow their historical evolution. Consequently, the time coordinate of this experiment should not be interpreted as actual historical calendar time. The 1pctCO₂ results should instead be presented in terms of experiment year or CO₂ level, for example years 1–50, 51–100, and 101–140, or conditions corresponding approximately to 1 × CO₂, 2 × CO₂, and 4 × CO₂.

2. Period means should not be interpreted as trends

Regarding the descriptions on pages 10–12, Figs. 1–4 mainly present climatological means for selected periods. Nevertheless, the manuscript repeatedly refers to these results as “trends”. Comparisons among several period-mean contour plots are not sufficient to establish temporal trends or statistical significance.

The statement that the turnaround latitude remains unchanged also requires quantitative support. The turnaround latitude should be objectively diagnosed for each model and period, and its changes and associated uncertainties should be reported.

3. Definition of “transport efficiency”

Section 2.3 states that transport efficiency is not measured directly but is inferred collectively from AoA, TEM circulation diagnostics, wave forcing, and MLS lag correlations. As currently formulated, “transport efficiency” therefore appears to be an interpretative concept rather than a

clearly defined diagnostic quantity or quantitative metric.

The authors should either provide a clear, quantitative, and testable definition of transport efficiency or moderate the associated claims throughout the manuscript.

4. Relationship to the method of Flury et al. (2013)

The MLS analysis differs substantially from the method used by Flury et al. (2013). That study correlated one-year H₂O time series between selected, physically adjacent pressure levels and between the Equator and approximately 40° latitude at 100 hPa. In contrast, the present manuscript averages data over broad 30° latitude bands and broad isentropic layers, uses an approximately 18-year anomaly record, and selects correlation peaks only after imposing minimum lags of 45 or 90 days.

These methodological modifications may be reasonable, but they require independent justification and validation. In particular, selecting the first or largest correlation peak only after imposing a prescribed minimum lag creates a risk that the inferred result depends strongly on the selected threshold. Short-lag maxima should not simply be classified as spurious because they are inconsistent with an expected transport timescale.

At a minimum, the authors should:

- show the unconstrained lag-correlation functions and all relevant correlation peaks;
- specify the complete lag-search interval and peak-selection algorithm;
- provide uncertainty intervals for the selected lags;
- test the sensitivity of the results to different smoothing windows and latitude-band definitions;
 - compare results based on broad-layer averages with those obtained at individual isentropic levels;
- and
- validate the inferred lags using an independent tracer, model experiment, trajectory calculation, or age-spectrum diagnostic.

5. Statistical significance of the lag correlations

Table 5 reports $p < 0.001$ even for correlation coefficients as small as 0.14–0.18. The manuscript appears to apply a conventional Pearson-correlation significance test. Such a test assumes that the samples are approximately independent. This assumption is clearly violated by the daily stratospheric tracer records, the application of a 31-point smoothing filter, overlapping lagged samples, and strong seasonal and interannual persistence.

The effective number of independent observations is therefore likely to be substantially smaller than the nominal sample size. The statistical significance should be reassessed using a method that accounts for temporal autocorrelation, such as an effective-sample-size correction, a moving-block bootstrap, or autocorrelation-preserving surrogate time series. The effects of searching across multiple lags and selecting the maximum correlation should also be taken into account.

6. Figures and support for the conclusions

Several figures contain very small axis labels, tick labels, colour bars, and panel annotations, making quantitative evaluation difficult. The period-mean figures should be supplemented or replaced by change maps. Consistent colour scales should be used, and stippling or hatching should be added to indicate statistical significance and inter-model agreement.

The abstract and conclusions should also be revised. Statements such as “robust multi-model

evidence”, “direct observational evidence”, “observational constraint”, and “greenhouse-driven changes are expressed primarily” are currently stronger than warranted by the analyses presented.

Technical corrections

7. Use the standard experiment name “1pctCO2” consistently. The term “1ptCO2” in Table 3 is incorrect.
8. The caption of Fig. 6 gives the period as 1850–1990, whereas the corresponding period is defined elsewhere as 1850–1900.
9. Revise the legend and caption of Fig. 6. The EPFD, NOGW, and OGW terms represent zonal-momentum tendencies rather than “total upwelling”.
10. Carr et al. (1995) and Clark et al. (1998) refer to the earlier UARS MLS instrument. The Aura MLS Level-3 product should be supported by appropriate references to the Aura MLS product description and data-quality documentation.
11. Replace “pre-industry” with “pre-industrial” throughout the manuscript.