General comments

This study assesses the 11-year solar cycle signals in the middle atmosphere in multiple-model ensemble simulations. This study starts with initial solar cycle signals in short wave heating rate, ozone, and temperature anomalies and continues with an analysis of whether the top-down mechanism explaining the downward propagation of these initial solar signals can be found in the presented models. I find the study highly relevant for long-term discussion of indirect solar effects and recommend it for publication with minor comments listed below.

Specific comments

I am not convinced that three sets of historical-like simulations with 9, 6 and 10 ensemble members, respectively, can be called "large ensembles" (l2+l330) since we do not know how large the large ensemble needs to be (Milinski et al, 2020).

➔ Thanks for the reference. Indeed, as shown in the work of Drews et al. (2022), even a 10-member ensemble is still not large enough to quantify the solar signal in the zonal wind at stratopause (~1 hPa) or the surface. But it's good enough to quantify the solar signal in the temperature at the tropical stratopause (see Figure 14 in the Extended Data of Drews et al. (2022)). We deleted the word "large" to avoid misleading.

Reference:

Drews, A., Huo, W., Matthes, K., Kodera, K., and Kruschke, T.: The Sun's role in decadal climate predictability in the North Atlantic, Atmos. Chem. Phys., 22, 7893–7904, https://doi.org/10.5194/acp-22-7893-2022, 2022.

Can you specify the threshold in the abstract (l8)?

➔ Revised. **Please see lines 7-8**.

What does "partly confirmed" (l25) mean?

➔ It means that top-down propagation of the solar signal was found in subsequent studies, but with varying times of propagation. The texts are revised**, please see lines 26-29.**

The authors should elaborate more on the fact that the solar signal may not be stationary (Thejll et al, 2003) related to modulation by QBO and PDO (l30).

➔ Thanks for the suggestion, the descriptions are revised. **Please see lines 29-33.**

I would omit the "controversial" label (l39) even though these studies may reduce the confidence level of the solar-NAO connection as you state.

➔ Replaced the "controversial" with "diverse".

As shown in previous studies (e.g. Mitchell et al, 2014; Kuchar et al, 2015) the upper stratospheric equatorial temperature anomaly related to the solar cycle has been detected showed a statistically significant signal with structure and amplitude of 1–1.25 K. Temperature response in Fig. 2A maximizes at 0.6 K. I would say that models a bit underestimate the response even with comparison (l167) with Kunze et al (2020). These facts should be discussed and even analyzed more thoroughly in your models.

➔ Thanks for pointing this out. In general, I would say, yes, the climate models with interactive chemistry used in this study (i.e., FOCI and EMAC) might a bit underestimate the response compared to the model with prescribed ozone chemistry (i.e., MPI-ESM-HR). This is consistent with the results based on CMIP5 simulations (as shown in Figure 4 in the work of Mitchell et al (2015)) that the models with interactive ozone chemistry simulate a response of 0.5 K in the tropical stratopause.
Besides, we calculated the composite differences between solar maxima and minima of the annual tropical SWHR, temperature, and O3 volume mixing ratio anomalies in the FULL ensemble mean with respect to the FIX ensemble mean (Figure. 2A). Following the method used in the work of Drews et al. (2022), here three years — the year of the peak (valley) of each solar cycle and two years around it — are selected as solar maximum (minimum). This definition could avoid a problem from double "peaks" in one solar cycle and smooth out the high-frequency interannual variability (< 3 years). But it also might lead to an "underestimation" compared to other methods, like multiple linear regression used in the works of Mitchell et al (2015) and Spiegl et al. (2023). To check the influence of the method, we repeated the composite analysis only based on the peak and valley years (Table R1), and the results are shown in Figure. R1.

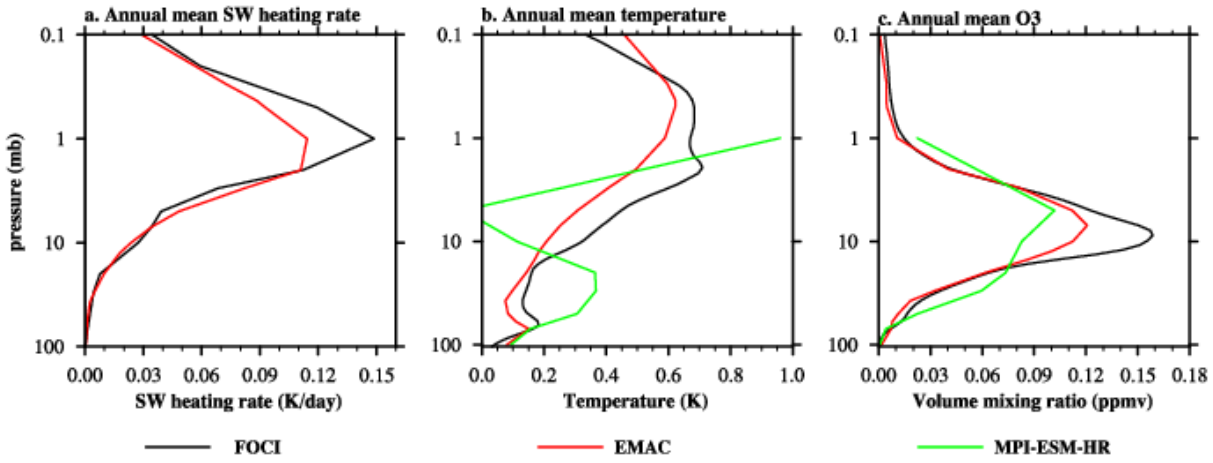

Figure. R1. Same as Figure. 2A, but composite based on the solar maximum and minimum years listed in Table R1 (below).

Table R1. The peak (maximum) and valley (minimum) years of solar cycles used for the "test" composite shown in Fig. R1

Solar peak	11, 21, 34, 44, 57, 68, 78, 88, 98, 108, 119, 131, 140, 151
Solar valley	6, 17, 28, 39, 52, 63, 73, 83, 94, 104, 114, 126, 136, 146

Comparison between Figure. R1 and Figure. A2, we found our method has very little impact on the responses in chemistry-climate models (FOCI and EMAC), but reduces the simulated temperature response in the MPI-ESM-HR model a lot (from 1.0 K to 0.6 K at the stratopause). Well, even the 1 K response of the tropical stratopause (Figure. R1) in the MPI-ESM-HR is still smaller than the result in the work of Spiegl et al. (2023) (about 1.2 K), which is estimated by a multiple linear regression method and CMIP5 historical simulations.

Compared to the reanalysis datasets (e.g., ERA-I, MERRA, and JRA-55 used in Mitchell et al (2015)), there is an "underestimation" of the initial solar signals in the upper and middle stratosphere in both the CMIP5 models (Figure 4 in Mitchell et al (2015)) and the models used in our study, especially for the models with interactive ozone chemistry. However, we should note that the reanalysis datasets do not cover the "weak" solar cycles (i.e., solar cycles before 1940) and only one member for each dataset. The SWHR is sensitive to the strength of the solar cycle (as demonstrated in Figure 1 of the work of Spiegl et al. (2023)) and hence a weaker solar signal is achieved when more "weak" solar cycles are included (e.g., back to 1850). In addition, the ensemble mean of transition simulations could smooth out some "coincide" between the solar signal and internal variability. In the study of Kunze et al (2020), the response of SWHR in tropical stratopause is about 0.2 Kd^{-1} based on sensitivity simulations forced by perpetual solar maximum conditions of the solar cycle 22 maximum and 0.15 Kd^{-1} in our transient simulations. Due to the difference in methods and experiments' design, we will not directly conclude that our models underestimate the solar signals but include a discussion on it. **Please see lines 167-182.**

References:

Mitchell, D., Misios, S., Gray, L., Tourpali, K., Matthes, K., Hood, L., Schmidt, H., Chiodo, G., Thiéblemont, R., Rozanov, E., Shindell, D., and Krivolutsky, A.: Solar signals in CMIP-5 simulations: the stratospheric pathway, Q.J.R. Meteorol. Soc., 141, 2390–2403,https://doi.org/https://doi.org/10.1002/qj.2530, 2015.

Spiegl, T. C., Langematz, U., Pohlmann, H., and Kröger, J.: A critical evaluation of decadal solar cycle imprints in the MiKlip historical ensemble simulations, Weather and Climate Dynamics, 4, 789–807, https://doi.org/https://doi.org/10.5194/wcd-4-789-2023, 2023.

Kunze, M., Kruschke, T., Langematz, U., Sinnhuber, M., Reddmann, T., and Matthes, K.: Quantifying uncertainties of climate signals in chemistry climate models related to the 11-year solar cycle – Part 1: Annual mean response in heating rates, temperature, and ozone, Atmos. Chem. Phys., 20, 6991–7019, https://doi.org/https://doi.org/10.5194/acp-20-6991-2020, 2020.

Based on Fig. 1.c (l159), the authors suggest that a nonlinear response can occur when the solar forcing is strong enough but I would soften these statements given the large spread and not enough samples for high sfu values.

➔ Thanks for the suggestion. The statements are revised. **Please see lines 162-166.**

I would omit the publications of Gray et al (2010) which provides a review of the Kodera and Kuroda mechanism and Mitchell et al (2015; CMIP5) which does not show any BDC response (l219) and only highlight the link between weaker BDC and lower-stratospheric temperature induced by the 11-year solar cycle.

➔ Revised.

How different (l240)?

➔ Spiegl et al. (2023) analyzed a set of historical simulations forced by CMIP5 external forcings (i.e., the CMIP5 protocol) and the solar forcing follows the reconstruction of lean (2000). The simulations were integrated from 1850 to 2005 and they focused on the period 1880 – 1999 in their study. We revised the texts, **please see lines 251-252.**

Would you find relevant to reproduce composite differences between Smax and Smin as in e.g. A8 for ERA5 and assess whether the response of temperature and zonal wind in a reanalyzed dataset also reveals a sensitivity to weak and strong solar epochs?

➔ It's hard to assess the sensitivity based on the ERA5 data (1950--present) because it only partly overlaps the strong epoch (i.e., 1932--2014). However, here we reproduced the composite analysis based on the ERA5 (Figure. R2) and discussed the comparison with the modeling results in this study. We should notice the lower top of the ERA5 data (up to 1 hPa) compared to the climate models used in this study. Similar to the modeling results (Figure 4 in manuscript), a weak warm response can be found in the tropical stratosphere in most winter months (except February) with ERA5 data (first row of Figure. R2). However, the stratospheric temperature response in EAR5 does not pass the significant test. Compared with using a short single member, (e.g., EAR5 here), the ensemble mean of climate modeling runs can extract the external forcing signal. The temperature response in FOCI during the strong epoch (second row of Figure A7 in the manuscript) is a bit larger than in the ERA5, which may be due to the warm bias in the tropical upper stratosphere of FOCI (as shown in Figures 10 and 12) leads to a higher sensitivity of the model to the solar forcing. As a result of the increased meridional temperature gradient, response in the zonal mean zonal wind anomalies can be found in the ERA5 (second row of Figure. R2), which is stronger than the FOCI modeling results (second row of Figure. A8) and interrupted by the strong internal variability in February.

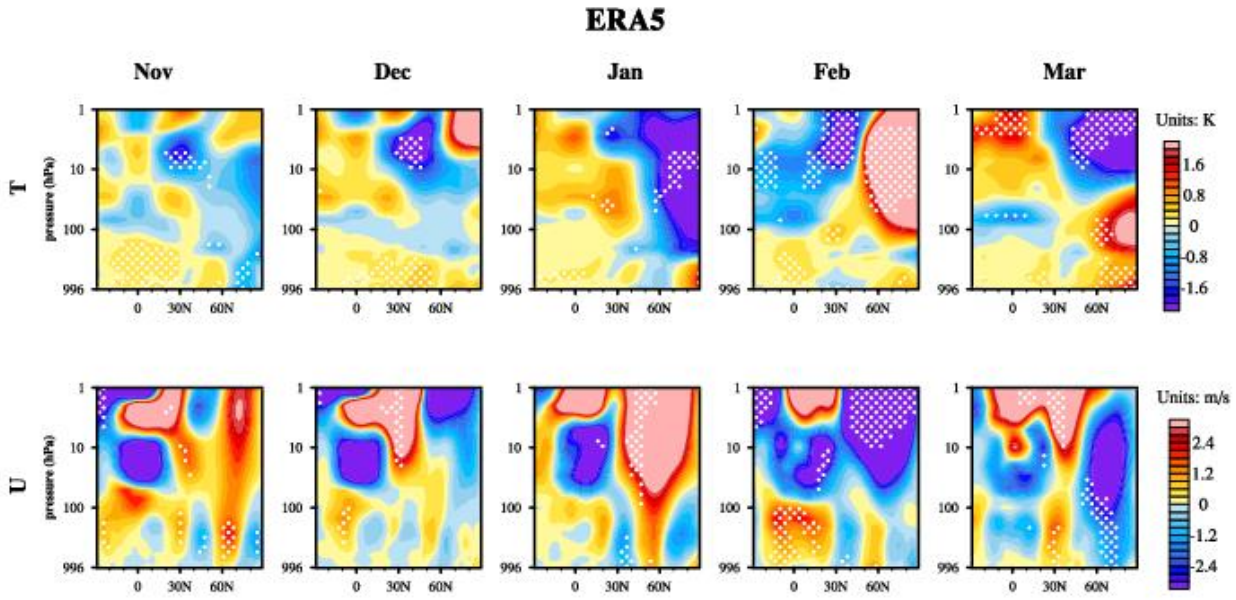

Figure. R2. Composite differences between solar maximum and minimum of the zonal-mean temperature anomalies (Units: K, first row) and zonal mean zonal wind anomalies (Units: m/s, second row) from ERA5 data (1950--2014).

Using vector figures instead of raster ones may help to improve the quality of your publication.

➔ Thanks for the suggestion. We submitted the high-resolution figures (in .pdf formats) to the system for publication.

Due to the extensiveness and unique methodology of the study, I think the whole community would appreciate an adoption of Open Science approaches to allow reproduce the extensive analysis in this study (e.g. Laken, 2016). In particular, I would recommend any kind of willingness of the authors to publish the code allowing to reproduce the figures in the paper. There are multiple ways how to proceed, either to allow the access upon request or via portals allowing to assign Digital Object Identifier (DOI) to the research outputs, e.g. ZENODO. I think it could enhance the quality and reliability of this publication.

➔ Thanks for the suggestion. All the codes involved in this study are achievable via the ZENODO link: https://doi.org/10.5281/zenodo.13358940. **We added the description in lines 407-409**.

I really appreciate the authors's willingness to use the robust bootstrap method to but why do you use only 1000 samples? Furthermore, this should be used to assess the significance level of the correlation coefficient to secure methodological consistency. Or was the temporal autocorrelation taken into account in your composites? Can you discuss how the inclusion of the effective sample size (see Section 5 in Bretherton et al, 1999) would influence the t-test results? Do your composite samples comply with the t-test assumptions?

➔ Following the method described by Diaconis and Efron (1983), we performed a 1000-fold bootstrapping test with replacement to estimate the statistical confidence level (90%) of the ensemble mean composites in this study. Here is a bit more explanation of this method and we took a composite of temperature anomalies as an example. (1) We calculated the ensemble mean temperature response to the solar cycle by the difference of ensemble mean temperature anomalies between solar maximum and minimum --- the true value for the bootstrap method. (2) Using all the original data (i.e., all years and all members) as a seeds pool, we calculated an averaged value of the all-years (i.e., 165 years in total) from the seeds pool --- the observed value for a random set. (3) We mimicked step (2) 1000 times by replacing the temperature anomaly randomly. If 90% of the amount of the observed values from step (3) were different from (and smaller than) the true value of step (1), we then marked the true value as a significant response. This method can be used to identify the significance level of the solar signal different from the background noise without assuming that the data have a normal distribution, especially for the cases where only several solar cycles were included (i.e., only a few data).

Of course, we could also mimic the process 10000 times more or increase the critical level (like 95%). We tested the 10000 times and only very tiny changes happened in the results, so we kept the 1000 times to save the computing resources. However, when the critical level is increased to 95%, the significance of the composite of temperature anomalies only reduces a bit but a large reduction in the composite of zonal mean zonal wind anomalies. To facilitate the comparison with our previous work (Drews, et al., 2022) which used the same method, we kept the 90% significance level in this study.

➔ The bootstrap t-test does improve the power of the t-test for a pair of non-normality datasets. In this study, we calculated the correlation coefficients in the 45-year running windows and demonstrated their dependencies on the solar cycle amplitudes. However, the 95% significant levels based on the bootstrap t-test in all the 45-year windows and for all the ensemble members mixed up in one busy figure (in a way as Figure 2), and most of them overlapped. It is hard to interpret and compare. Considering the effective degree of freedom in the 45-year running windows (method described below) are quite similar, we prefer to use a consistent t-value from the two-sides student's t-test to show the 95% significant level (as indicated by the black dash line in Figure 2) to facilitate the comparisons and reproduction. The effective degree of freedom in each 45-year window was calculated following the method used in the work of Pyper and Peterman (1998) and simplified as only the autocorrelation coefficients at lag 1 are considered. More details of this method are also described in the work of Huo et al. (2023).

We briefly described the method of calculating the effective degree of freedom in the method section, **please see lines 143-147.**

References:

Diaconis, P. and Efron, B.: Computer-Intensive Methods in Statistics, Scientific American, 248, 116–131, 1983.

Pyper, B. J. and Peterman, R. M.: Comparison of methods to account for autocorrelation in correlation analyses of fish data, Canadian Journal of Fisheries and Aquatic Sciences, 55, 2127–2140, https://doi.org/10.1139/f98-104, 1998.

Huo, W., Xiao, Z., and Zhao, L.: Phase-Locked Impact of the 11-Year Solar Cycle on Tropical Pacific Decadal Variability, Journal of Climate, 36, 421–439, https://doi.org/https://doi.org/10.1175/JCLI-D-21-0595.1, 2023.455

Please specify what CCR in your figures stands for

➔ "CCR." stands for "correlation coefficients in a running window". An explanation is added in the method section. **Please see line 133.**

l288 replace EAR with ERA5

➔ Revised.

l290 replace EAR with ERA5

➔ Revised.

References

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., & Bladé, I. (1999). The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field, Journal of Climate, 12(7), 1990-2009. Retrieved Jan 19, 2022, from https://journals.ametsoc.org/view/journals/clim/12/7/1520-0442_1999_012_1990_tenosd_2.0.co_2.xml

Kuchar, A., Sacha, P., Miksovsky, J., & Pišoft, P. (2015). The 11-year solar cycle in current reanalyses: a (non)linear attribution study of the middle atmosphere. Atmospheric Chemistry and Physics, 15(12), 6879–6895. https://doi.org/10.5194/acp-15-6879-2015

Kunze, M., Kruschke, T., Langematz, U., Sinnhuber, M., Reddmann, T., and Matthes, K.: Quantifying uncertainties of climate signals in chemistry climate models related to the 11-year solar cycle – Part 1: Annual mean response in heating rates, temperature, and ozone, Atmos. Chem. Phys., 20, 6991–7019, https://doi.org/https://doi.org/10.5194/acp-20-6991-2020, 2020.

Laken, B. A. (2016). Can Open Science save us from a solar-driven monsoon? Journal of Space Weather and Space Climate, 6, A11. http://doi.org/10.1051/swsc/2016005020.

Milinski, S., Maher, N., and Olonscheck, D.: How large does a large ensemble need to be?, Earth Syst. Dynam., 11, 885–901, https://doi.org/10.5194/esd-11-885-2020, 2020.

Mitchell, D.M., Gray, L.J., Fujiwara, M., Hibino, T., Anstey, J.A., Ebisuzaki, W., Harada, Y., Long, C., Misios, S., Stott, P.A. and Tan, D. (2015), Signatures of naturally induced variability in the atmosphere using multiple reanalysis datasets. Q.J.R. Meteorol. Soc., 141: 2011-2031. https://doi.org/10.1002/qj.2492

Thejll, P., Christiansen, B., and Gleisner, H.: On correlations between the North Atlantic Oscillation, geopotential heights, and geomagnetic activity, Geophys. Res. Lett., 30, https://doi.org/10.1029/2002GL016598, 2003