Response to Reviewer 2

We would like to thank the reviewer for their helpful comments. Our responses are printed in blue, while the reviewer's questions are in black. In addition to addressing the individual comments, we will review the manuscript for clarity and flow.

The paper is clear and well-written (although some parts have extensive descriptions of the atmospheric circulation that could be shortened in my opinion). I was already a reviewer in the first part of this article, and I have nothing new to add to the methods, which are essentially similar methods. I have some technical comments below, including some statistical significance computations that should be done differently in my opinion, but apart from those, I would be happy to recommend the paper after some revisions.

Response: Thank you for your continued review and positive assessment of our paper. We appreciate your suggestion to shorten the atmospheric circulation descriptions and will streamline these sections accordingly in the revised version.

Regarding the statistical significance computations, we agree that this is an important point and have implemented the bootstrapping method with a False Discovery Rate (FDR) test as suggested. The results are shown in Figure R2, and we will revise the manuscript accordingly.

The only main limitation that I see is that I find the paper very descriptive without testing any physical theory. In other words, it could have been interesting to explore how the changes you see fit with some physical expectations for how those mechanisms are supposed to evolve. It is mentioned several times that Arctic amplification, and the associated change in the temperature gradient and general circulation, is expected to decrease the importance of advection, and I think this kind of reasoning could be interesting to investigate further. I do not think this is a reason to reject the paper, but it could really add something on top of those descriptive mechanisms.

Response: Thank you for raising this important point regarding the connection between our results and the core concept of Arctic amplification. We have now incorporated a more detailed analysis to address this directly, and we will add these results to the supplementary materials.

Our trajectory analysis reveals a clear fingerprint of Arctic amplification: air originating from the Arctic experiences consistently larger future warming than air from lower latitudes (Figure 5, main paper). This reflects a weakening of the meridional temperature gradient. Furthermore, as shown in Figure R6, air masses on both day *t-1* and day *t* are substantially modified by the warming climate. The ratio of local temperature change to global warming exceeds 1 on all days (−72 h to 0 h), and is above 1.4 for both day *t-1* for warming and day *t* for cooling events, satisfying the quantitative definition of Arctic amplification. This confirms that air masses participating in these extreme DTDT events carry the signature of Arctic amplification, as they originate from regions warming at a rate significantly faster than the global average.

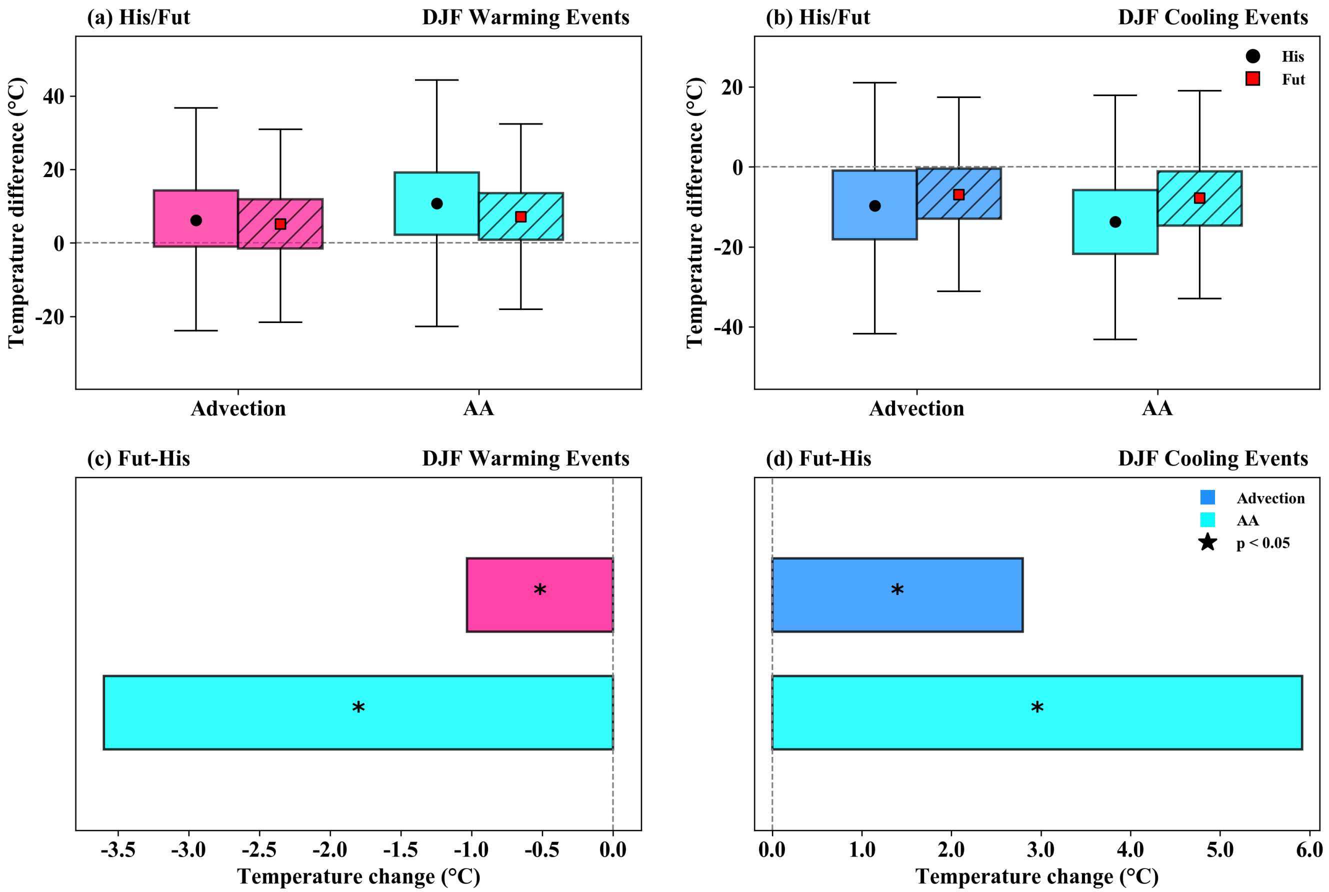

Figure R1. (a,b) Box plots comparing the contribution of advection and the estimated effect of Arctic Amplification (AA) for (a) warming events and (b) cooling events in North America, on a 3-day time scale. For each process, the left box (solid edge) represents the historical climate, and the right box (hatched) represents the future climate. The box spans the 25th to 75th percentiles; the black dot marks the historical mean; the red square marks the future mean; whiskers extend to 1.5 times the interquartile range. The effect of AA is estimated by comparing the seasonal-mean projected temperature change at the origins (at -72 h) of the historical backward trajectories between day t and day t-1. (c,d) Stacked horizontal bars showing the mean change (future minus historical) for advection and AA for (c) warming and (d) cooling events. The * indicates that the difference is statistically significant at the 95 % confidence level based on a t-test.

We isolated the Arctic Amplification (AA) contribution to extreme temperature events by comparing the seasonal-mean projected temperature change at the origins (at -72 h) of the historical backward trajectories between day *t* and day *t-1* in Figure R1. In this way, only the warming at the origins (which differs between *t* and *t-1* due to the typically different latitudes of air parcels at -72h) is considered under the assumption that the circulation is unchanged. The magnitudes of the AA effects are more pronounced than those from the changes in advection determined from equation 2. Specifically, for warming events, the change in advection (−1.1 °C) is smaller in magnitude than the change in AA (−3.5 °C), while for cooling events, the change in advection (+2.8 °C) is also smaller than the change in AA (+5.9 °C). The effect of AA explains the weakening of the advective contribution to both warming and cooling events, but the magnitude of the AA effect is mitigated by circulation changes. Also, the event-to-event variability (boxes and whiskers in Figure R1) is very similar between advection and estimated AA contributions.

We will add the Figure to the supplementary material and discuss this result in the main paper.

Major comments

1. A. Figure 1 and corresponding: the way the stipplings are computed does not look like a proper statistical test to me. a. If I understood correctly, for the present, the authors flag as “significant” the grid points where 80% of members are within +/- 10% of the EAR5-derived respective quantities. I do not think this is correct: first the +/-10% for ERA5 is an ad-hoc measure of the uncertainty. Second, I do not see why 80% of the members should be a correct measure of a significant difference. I suggest to do an actual statistical test with a standard reference level of 95% significance for example. In essence you want to know whether the climate of CESM is compatible with the value for ERA5: that is what you need to test for. The climate of CESM is defined by all the members being put together: what you need to test is whether the sigma_DTDT, sigma_T and r_1,T of this climate are compatible with the same values for ERA5 (which also has an uncertainty). You could for example, employ a bootstrapping approach on the ERA5 data (the values for the model are likely very well estimated given the amount of members you have) and check whether the distribution of values you obtain are compatible with the one from the model at the 95% level.

   Response: Thank you for your valuable feedback on our significance testing methodology. Following your suggestion, we have revised our approach (Figure R2). Given the present-day climate, we use bootstrapping (n=1000) on ERA5 data to construct an observational uncertainty distribution, which we compare with the CESM multi-member distribution at the 95% significance level. As noted in Comment 3, we apply a false discovery rate (FDR) correction to the bootstrapping results. We found that this FDR correction reduces the significant area by only 1–8% globally at the 1% level (Figure R2a-c). We will update the methodology and results accordingly.

   We will explain this result in the revised manuscript as follows: “The hatching in Figure R2a-c indicates that the CESM ensemble mean is systematically different from the ERA5 estimates in most regions, with only a few exceptions. Across approximately 40–70% of global land areas, CESM-LE exhibits systematic biases in daily temperatures and related metrics relative to ERA5. While the model reproduces the large-scale spatial patterns of the metrics (see Part I), notable discrepancies remain in the magnitude of anomalies (CESM-ERA5, Figure R3). These biases include an overestimation of DTDT extremes, particularly in the mid-to-high latitudes such as over North America and large parts of Asia. In contrast, underestimations occur at lower latitudes (e.g., over the Sahara Desert and Amazon) during DJF and persist across most regions during JJA.

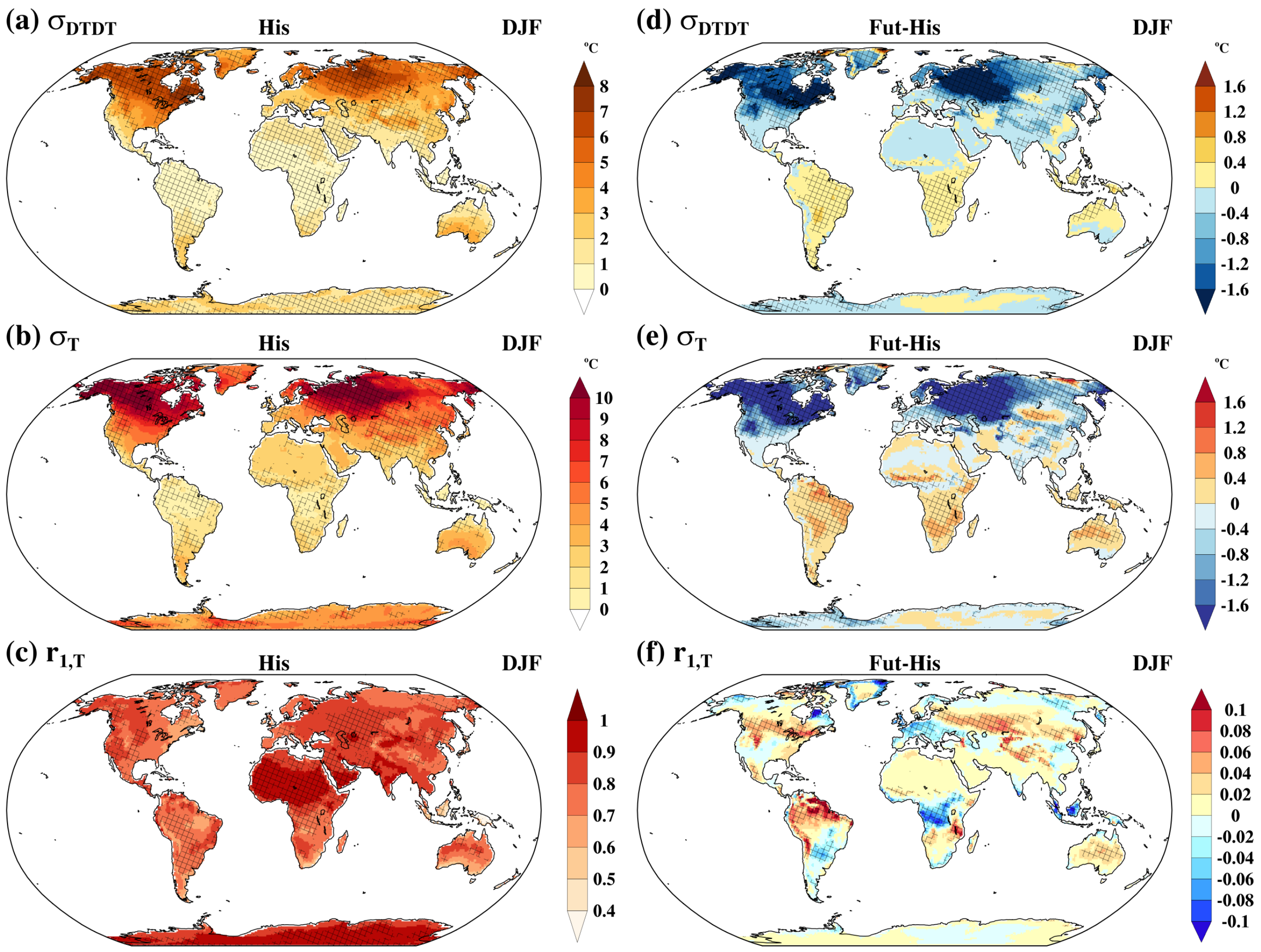

Figure R2. The ensemble means of (a, d) standard deviation of DTDT variations (σ_{DTDT}, °C), (b, e) standard deviation of daily mean temperature (σ_T, °C), and (c, f) lag-1 autocorrelation of daily mean temperature ($r_{1,T}$) in December-February (DJF) in the historical climate (a-c) and projected change (d-f). In the panels a–c, cross-hatching marks grid points where the CESM-LE ensemble mean differs significantly from the ERA5-derived metric, with statistical significance determined through bootstrap resampling. In panels d–f, cross-hatching denotes grid points where the future minus historical difference is significantly different from zero, assessed via a two-sample bootstrap test.

1. B. Same for the future: why don't you simply test whether there is a significant difference in sigma_DTDT, sigma_T and r_1,T between the two climates by putting all members together in each period?

   Response: Thank you for this helpful suggestion. For the future projections, we have adopted the same approach: test whether there is a significant difference in temperature metrics between the historical and future climates by pooling all ensemble members from each period (Figure R2d-f). We have applied bootstrapping with FDR correction at the 1% and will update the result accordingly.

2. For all your significance maps, you need to take into account correlations in statistical testing and employ a false discovery rate, see Wilks (2016).

   Response: Thank you for your suggestion. Following your suggestion, we applied the bootstrapping test with false discovery rate (FDR) correction (Wilks, 2016) to all climatological maps, while for the other composite maps, we applied a t-test at the 95% significance level. After applying the FDR correction, the area with significant differences is reduced only by 1–8%.

3. Several times, the authors argue that the model is doing a reasonable job in reproducing the statistics of ERA5. I am not sure this is so much the case, as exemplified by Figure 1, for example, where the stipplings do not really cover most of the regions (modulo my main comment 1). I think you should emphasise the differences more, including quantifying them when possible. One point, for example, is that the model seems to have a diabatic contribution larger than ERA5, which is something also found recently by Röthlisberger et al. (2025) in a different context: it seems to me that the model may be right for the wrong reasons.

   Response: Thank you for the suggestion. We agree that our initial framing was overly generous and have revised the manuscript to place greater emphasis on the quantitative differences between the model and ERA5. We have compared our results with previous studies by Röthlisberger et al. (2025) and will incorporate recent findings into the Discussion.

   Approximately 40–70% of global land regions exhibit systematic biases in daily temperature and related metrics simulated by CESM-LE compared with ERA5 (Figures R2–R3). While the model captures large-scale spatial patterns (cf. Part I), there are notable discrepancies in the magnitude of anomalies (Figure R3) and in the contributions of underlying physical processes. These biases include an overestimation of DTDT extremes, most pronounced in the mid-to-high latitudes (e.g., over North America and large parts of Asia), and an underestimation at lower latitudes during DJF and across most regions during JJA. The overestimation of extremes is linked to advection (primarily in the extratropics) and diabatic processes, with the relative influence of these processes varying across events and regions. This aligns with recent studies indicating an overestimation of daily temperature extremes due to advection and amplified diabatic heating from sensible heat fluxes (Röthlisberger et al., 2025). In contrast, the underestimation in the tropics is primarily associated with biases in the adiabatic contribution and, consequently, in vertical motion, likely stemming from inadequate representation of convective processes and turbulent fluxes in the models (Bao & Stevens, 2021; Bergman & Sardeshmukh, 2004; Stohl, 1998). Similarly, during JJA, the widespread underestimation across most regions results mainly from a combination of underestimated advective and adiabatic processes and also overestimated diabatic processes. These seasonal biases have important implications for targeted model development and improvement.

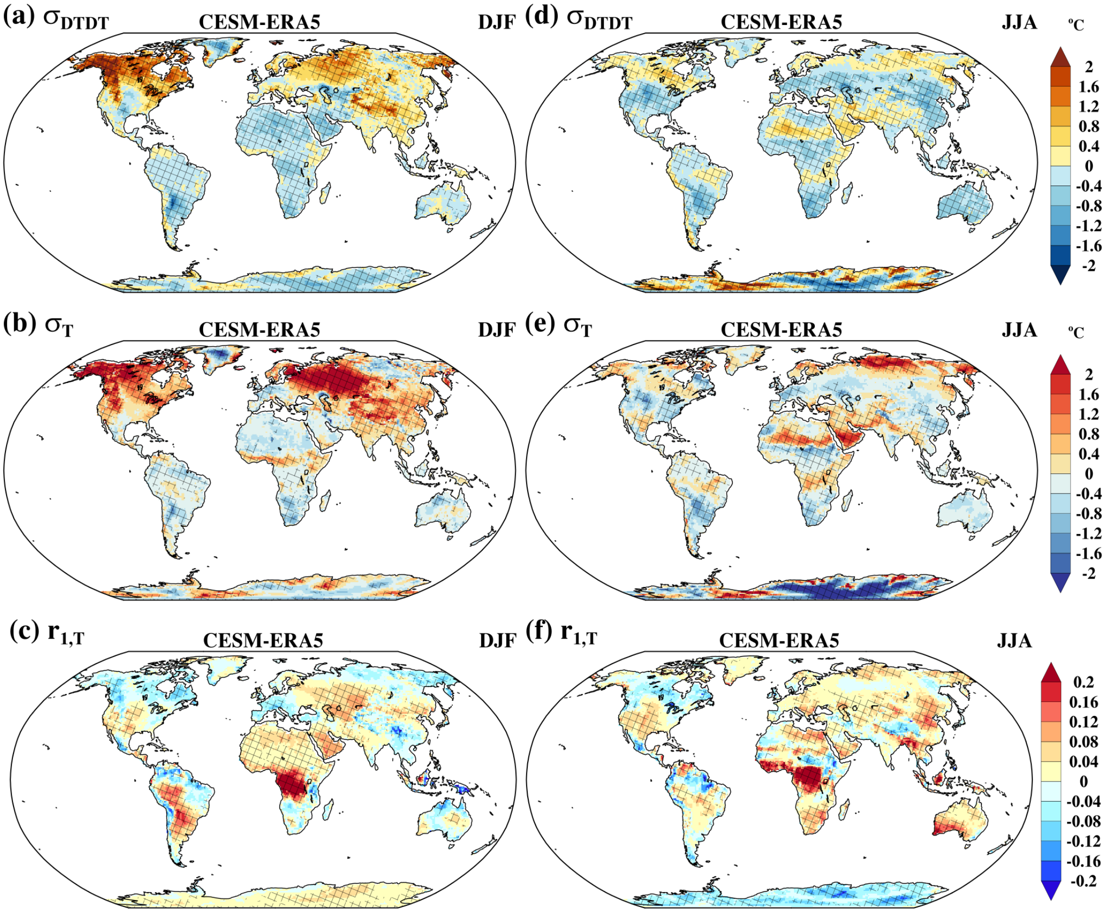

Figure R3. The absolute differences between the CESM-LE and ERA5 of (a, d) standard deviation of DTDT variations (σ_{DTDT}, °C), (b, e) standard deviation of daily mean temperature (σ_T, °C), and (c, f) lag-1 autocorrelation of daily mean temperature ($r_{1,T}$) in December-February (DJF, a-c) and June-August (JJA, d-f). In the panels a–d, cross-hatching marks grid points where the CESM-LE ensemble mean differs significantly from the ERA5-derived metric, with statistical significance determined through bootstrap resampling.

Minor comments

1. Please precise which version of CESM you are using.

   Response: We have used 1st version of CESM-LE and will correct this in the revised manuscript.

2. Figure 1: Because of the strong meridional differences, you could plot the changes in the second column in percentage rather than absolute values.

   Response: Thank you for the suggestion. We have plotted the projected percentage change in Figure R4. However, to maintain consistency with the trajectory analysis for which absolute changes are much easier to interpret, we keep the absolute differences in the main paper and add this Figure to the supplement.

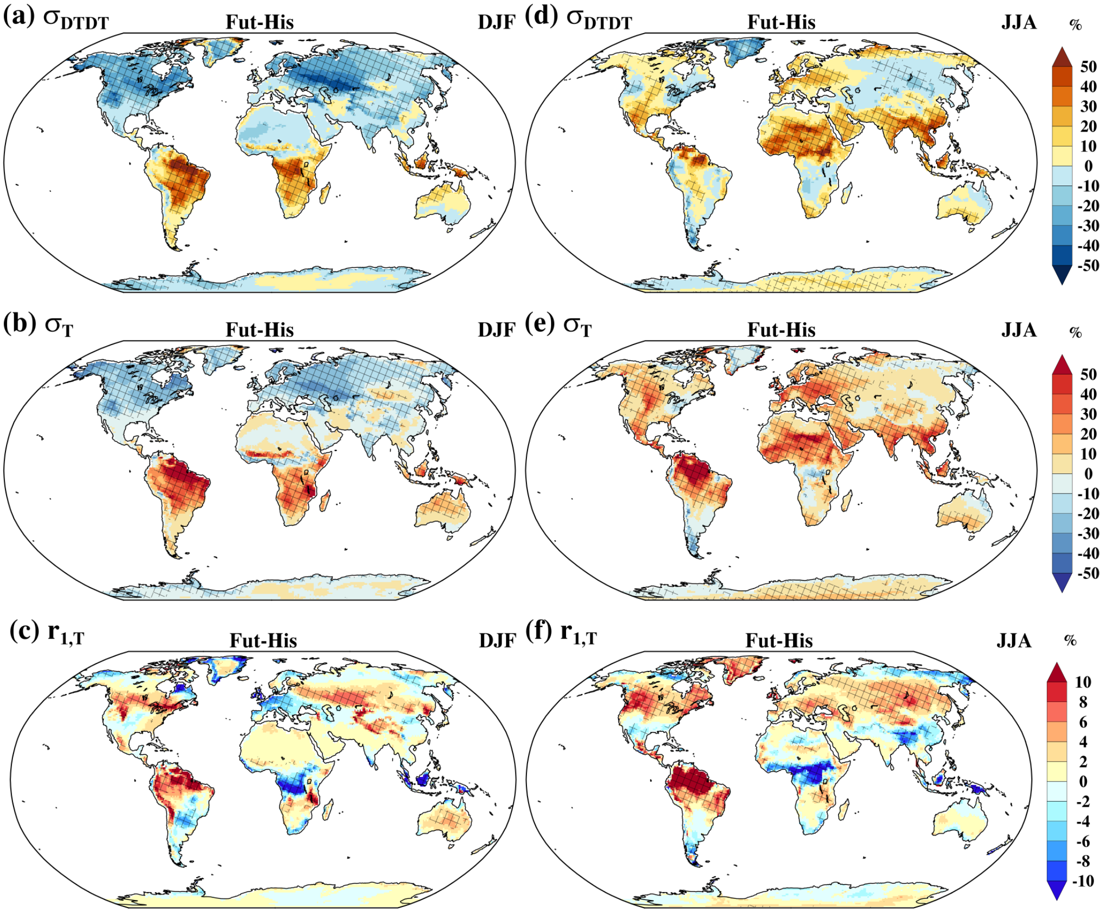

Figure R4. The projected percentage changes of (a, d) standard deviation of DTDT variations (σ_{DTDT}, %), (b, e) standard deviation of daily mean temperature (σ_T, %), and (c, f) lag-1 autocorrelation of daily mean temperature ($r_{1,T}$, %) in December-February (DJF, a-c) and June-August (JJA, d-f. In panels a–f, cross-hatching denotes grid points where the future-minus historical difference is significantly different from zero, assessed via a two-sample bootstrap test.

3. Figure 3: The sigma_DTDT should be delta_T?

   Response: We will change this.

4. Figure 4: The stipplings are barely visible.

   Response: Thank you for the suggestion. We have now changed the stippling to cross-hatching, which is more visible, and reduced the spacing between the panels, as shown in Figure R5 below.

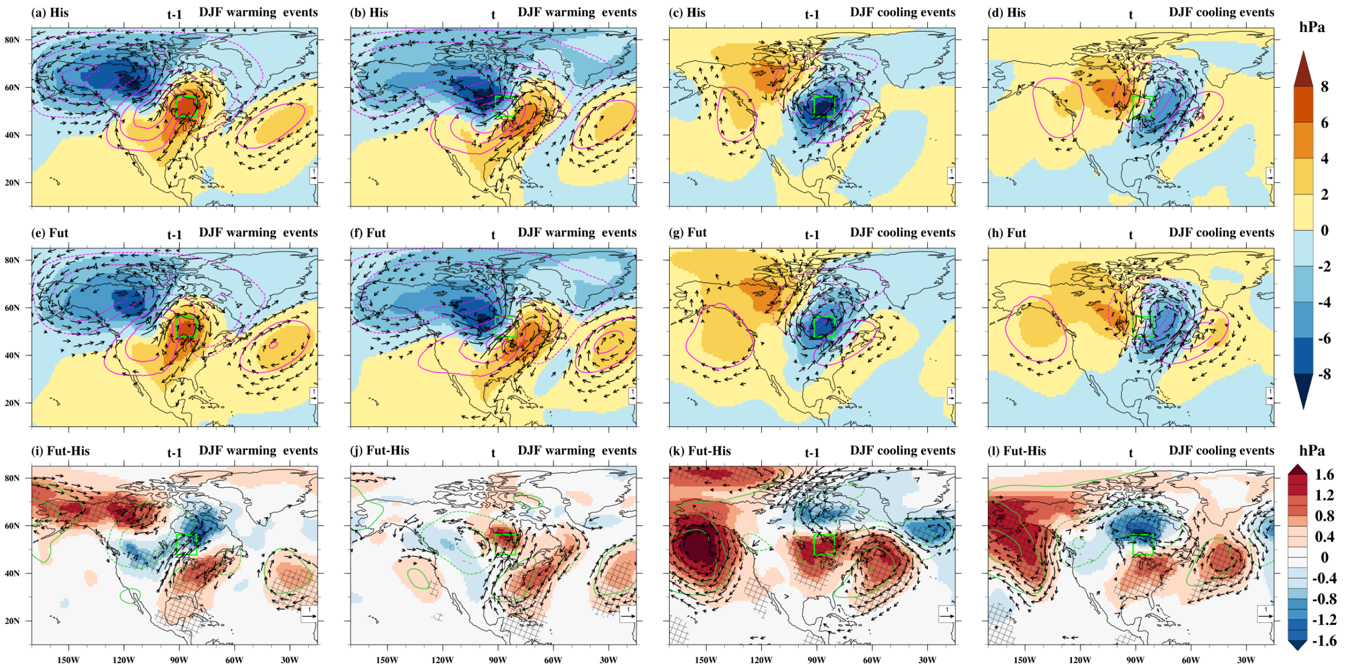
(a) His t-1 DJF warming events
(b) His t DJF warming events
(c) His t-1 DJF cooling events
(d) His t DJF cooling events
(e) Fut t-1 DJF warming events
(f) Fut t DJF warming events
(g) Fut t-1 DJF cooling events
(h) Fut t DJF cooling events
(i) Fut-His t-1 DJF warming events
(j) Fut-His t DJF warming events
(k) Fut-His t-1 DJF cooling events
(l) Fut-His t DJF cooling events
80N
60N
40N
20N
150W
120W
90W
60W
30W
hPa
8
6
4
2
0
-2
-4
-6
-8
hPa
1.6
1.2
0.8
0.4
0
-0.4
-0.8
-1.2
-1.6

Figure R5. Composite of sea level pressure anomalies (hPa, color shading), wind anomalies at 850 hPa ($\mathrm{m\,s^{-1}}$, vectors), and geopotential height anomalies at 500 hPa (gpm, magenta and darkgreen contours) relative to the seasonal mean on the (a, e, i, c, g, k) previous day (t-1) and (b, f, j, d, h, l) the event day (t) of the warming (a-b, e-f and i-j) and cooling (c-d, g-h and k-l) events during December-February (DJF) in (a-d) historical climate (His), (e-h) future climate (Fut), and (i-l) projected future changes (Fut-His) at a selected grid box in North America (green box). Note that, in (a-h), wind vector anomalies ≥ 2 $\mathrm{m\,s^{-1}}$ and in (i-l), wind vector difference anomalies ≥ 0.5 $\mathrm{m\,s^{-1}}$ are plotted. The dotted and bold contours indicate negative and positive geopotential height anomalies, respectively. Additionally, the cross-hatching area indicates where the ensemble mean of sea level pressure differences exceeds the 95% confidence threshold based on a t-test.

5. Figure 5: I would suggest scaling the temperature and pressure differences by a global/regional warming level to see what is changing beyond the expected local warming.

   Response: Thank you for the excellent suggestion. We have scaled the projected changes in temperature, pressure, and potential temperature with annual global temperature change, as shown in Figure R6. This updated analysis also illustrates how circulation patterns and temperatures are influenced by warming. The ratio of local temperature change to global warming exceeds 1 on all days (−72 h to 0 h) and is above 1.4 for both day *t-1* warming and day *t* cooling events, satisfying the quantitative definition of Arctic amplification. This confirms that air masses participating in these extreme DTDT events carry the signature of the Arctic amplification signal, as they originate from regions warming at a rate significantly faster than the global average.

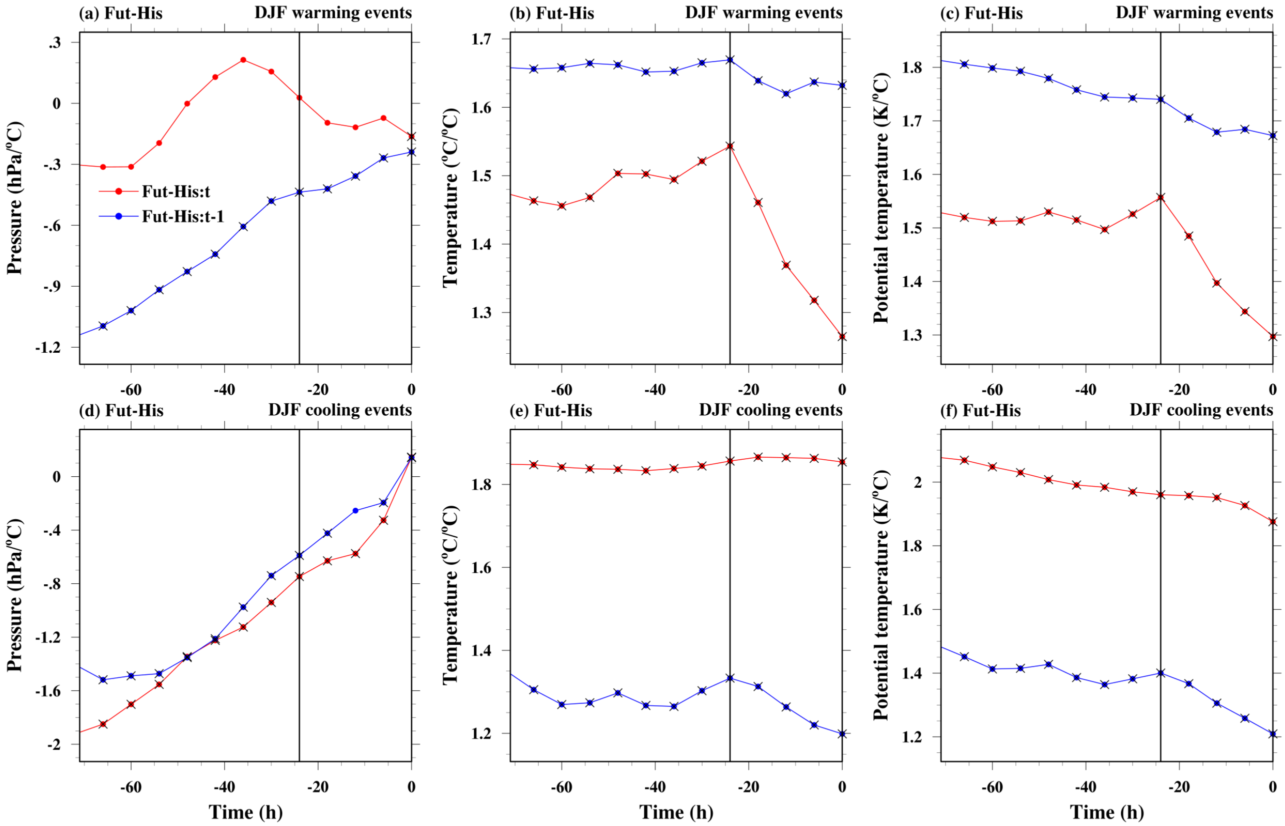

Figure R6: The mean Lagrangian evolution of distinct physical parameters: (a, d) pressure (Pres, hPa), (b, e) temperature (Tem, °C), and (c, f) potential temperature (PT, K) is shown along the air mass trajectories initialised on the previous (*t-1*) and event (*t*) days for projected changes in extremes scaled by annual global surface temperature (°C). Additionally, Bold circles with crosses mark time steps where the projected changes are statistically significant at the 95 % confidence level based on a t-test.

6. Figure 6 and similar: maybe you could add the boxplots for the future on panels a and b, also to compare the spread in each period (I do not expect the spread to be small, thus the changes you observe are probably much smaller in intensity compared to the spread between events in each period).

   Response: We thank the reviewer for this constructive suggestion. We will update Figure 6 (and other relevant figures) to include boxplots for both the historical and future periods in panels a and b. The spread among events within each period is indeed substantial. For the grid point over North America shown in Figure R7, the future spread is slightly smaller than (for DTDT, advection, and diabatic heating) or similar to (for adiabatic processes) the event-to-event variability in the historical climate, suggesting that this spread may also change in a warmer climate.

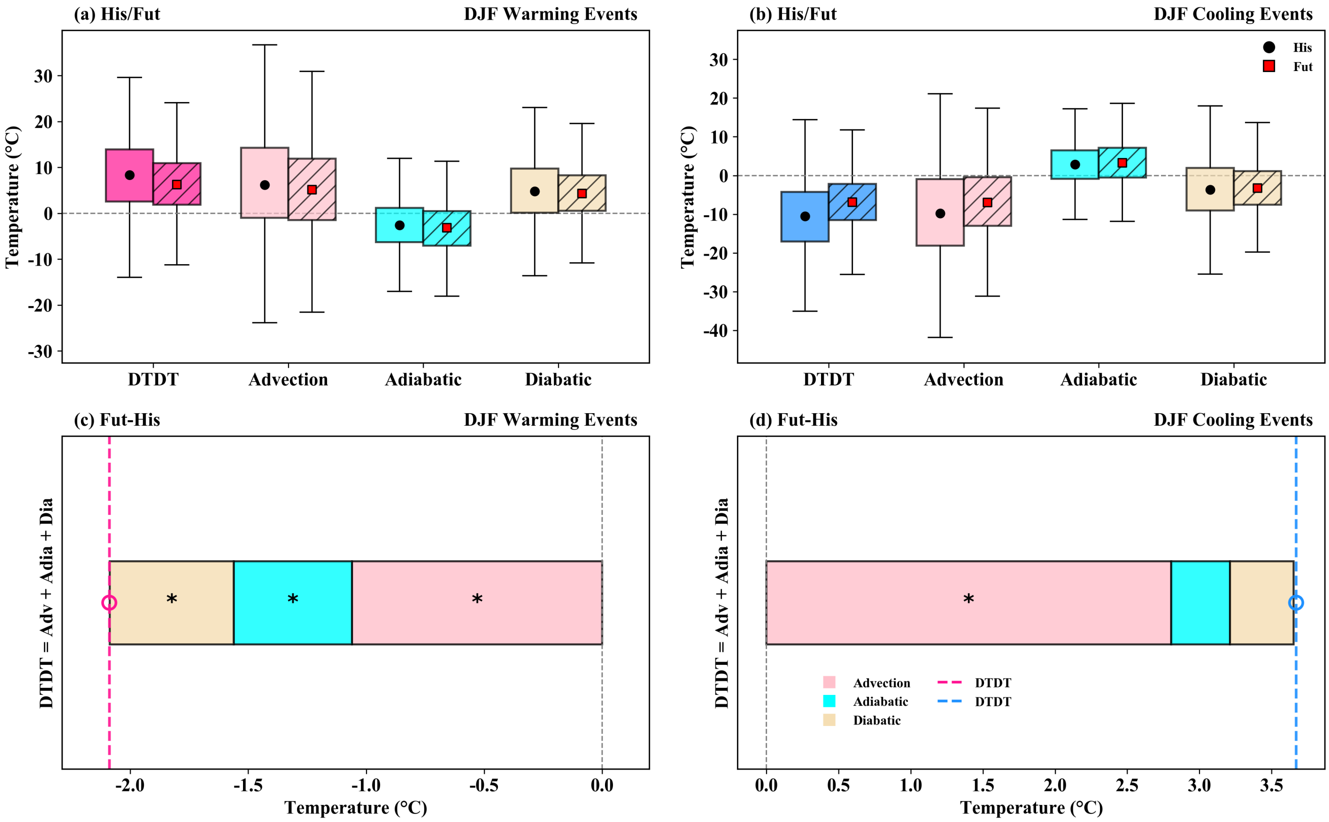

Figure R7. The contribution of the different physical processes (advection, adiabatic and diabatic temperature change) over North America during December-February (DJF) to genesis of DTDT (a, c) warming and (b, d) cooling events during historical/future climate (a-b, box plots) and projected future change (c-d, stacked plots) according to Eq. (2), which refers to a 3d-time scale. The box spans the 25th and 75th percentiles of the trajectory data; the black dot inside the box gives the mean of the related quantities in the historical climate, and the whiskers indicate 1.5 times the interquartile range in panels (a) and (b). The dotted lines in the stacked plots in panels (c) and (d) show the mean future change for DTDT warming and cooling events, respectively, and coloured bars indicate the contributions of the individual processes. Circle and * symbols mark future change distributions for which the ensemble mean differences exceed the 95% confidence threshold based on a t-test.

7. Figure 7: Why did you decide to change the position of the box for looking at extreme DTDT changes compared to Figure 4?

   Response: The grid points were chosen based on regions exhibiting the most significant changes (Figure R4). For DJF, the largest variations occurred over North America and Western North America, and the mechanisms behind the extremes are similar. For JJA, a grid point in western North America was selected, as the largest significant decrease was located near the coast (Figure R4d) and followed a different mechanism than at the northern grid point.

References

Bao, J., & Stevens, B. (2021). The Elements of the Thermodynamic Structure of the Tropical Atmosphere. *Journal of the Meteorological Society of Japan. Ser. II*, *99*(6), 1483-1499. https://doi.org/10.2151/jmsj.2021-072

Bergman, J. W., & Sardeshmukh, P. D. (2004). Dynamic Stabilization of Atmospheric Single Column Models. *Journal of Climate*, *17*(5), 1004-1021. https://doi.org/https://doi.org/10.1175/1520-0442(2004)017

Röthlisberger, M., Sprenger, M., Beyerle, U., Fischer, E. M., & Wernli, H. (2025). Advective, adiabatic and diabatic contributions to heat extremes simulated with the Community Earth System Model version 2. *EGUsphere*, *2025*, 1-32. https://doi.org/10.5194/egusphere-2025-5146

Stohl, A. (1998). Computation, accuracy and applications of trajectories—A review and bibliography. *Atmospheric Environment*, *32*(6), 947-966. https://doi.org/https://doi.org/10.1016/S1352-2310(97)00457-3

Wilks, D. S. (2016). “The Stippling Shows Statistically Significant Grid Points”: How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It. *Bulletin of the American Meteorological Society*, *97*(12), 2263-2273. https://doi.org/https://doi.org/10.1175/BAMS-D-15-00267.1