Author's response to RC1

The authors combine CMIP6 model output with reanalysis data, observations and LES model results to investigate the inter-model spread in Arctic amplification (AA) and the Arctic lapse-rate feedback (ALRF). When sorting models into models with stronger and weaker AA and ALRF, strong AA/LRF models better match reanalysis trends in heat advection, whereas weak AA/ALRF better match observed present-day inversion strength. The presented data and work is interesting and relevant to important research questions, but I have a few major concerns on how the model-observation analysis is carried out.

Reply: We thank Reviewer 1 for very constructive and helpful comments on our manuscript. We have addressed the concerns which have helped us to improve our manuscript. Our responses are listed below.

Major

Reviewer Point P 0.1 — The authors do not investigate the role of internal variability for model results. Investigating only one ensemble member per model without regard for the ensemble spread might not do justice to models – even a clear mismatch with observations does not rule out that the model in question is consistent with the observed trend or phenomenon (see e.g. Notz, 2015)

Reply: The reviewer raises an important point. Firstly, we haven't been clear enough in elaborating on the use of different ensemble members within CMIP6. We use the entire data set (all available ensemble members), but as ensemble means over all realisations per model – this way, each model carries equal weight in the CMIP6 distribution, and we exclude the chance of accidentally choosing a model realisation that deviates substantially from the entire population The reviewer still is right, we do not account for internal variability by using ensemble means, we merely exclude the chance of catching an outlier among the realisations. We want to state that while internal variability is an important and very interesting point, unfortunately there are only few models with enough members to engage in a deeper study: Only four models have more than 30 realisations which could be considered enough for such an analysis. Six other models on the other hand only have only one realisation, more than half only 2–3. Since we noticed that this topic has not been addressed properly in our manuscript, we added a paragraph in the methods section (L134 ff) and thank the reviewer for pointing it out.

Reviewer Point P 0.2 — Important conclusions rely on small subsets of the analysed models, comparing only the top and bottom three models in terms of AA/ALRF. For the weak AA group, these are clear outliers in the CMIP ensemble, and two of the three are different versions of the same model. Would the results remain the same (just with weaker signals) if models 4-8/24-28 were used instead?

Reply: It is an important question that the reviewer asks here. Firstly, it is exactly that, we chose CMIP6 models at the respective edges of the range of simulated past AA. This is to ensure a clear signal in the comparison and to allow for an attribution to either weak and strong-AA models (we added a comment in the method section L146 and following). Since we don't take the classic approach of an emergent constraint where statistically strong relationships across model simulations of past/future

and the observable current climate are used, we instead agreed on a number of models that represent an either weak or strong-AA cluster (split by the observed value of AA, also added now in Fig. 1 and explained in the text L151 and L420 and following). Unfortunately, some comparisons required a high temporal resolution of the model output (L140 ff in the manuscript). The model-data comparison at 6-hourly time resolution in particular included only 12 models with all required diagnostics in total (Section 3.2-3.4 concerning stability and vertical temperature structures). This has lead us to the compromise of choosing 3 models at the respective edge of AA distribution (model 5, 6, 10 for weak, and 25, 28, 29 for strong), as the inclusion of more models would rather represent the inter-model mean.

However, we strongly agree with the point being raised. To demonstrate that the comparison is still valid, we added model 11 to the weak-AA, and model 21 to the strong-AA ensemble to extend the individual range. This has no effect on the key messages of Section 3.2-3.4, the results remain the same (e.g. shown for MOSAiC in Fig. R 1). We added a paragraph on the sensitivity of model choice on the results in the discussion part L734 ff. In addition, the model-to-data comparison is explained more thoroughly in Section 3.2, adding a supporting Figure to the Appendix, and further commenting on statistical representatives of the results that rely on sparse high-resolution model diagnostics.

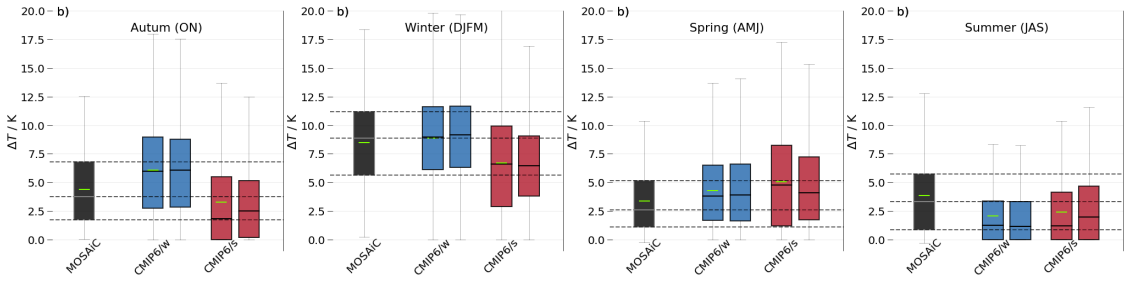

Figure R 1: As Fig. 4 b in the manuscript, with the addition of model 11 and 21 to CMIP6/w and CMIP6/s subsets, respectively. In each panel, the left box plots show the original subsets with three models, and the right box plots show the subsets with 4 models, respectively.

Reviewer Point P 0.3 — The definition of AA as a difference dT_{Arctic}-dT_{global} rather than a ratio dT_{Arctic}/dT_{global} is surprising to me. Wouldn't one expect most mechanisms driving AA to act in a multiplicative rather than additive way? Similarly, the choice of the reference period is unclear to me. If no observations from the reference period are used, why not choose an earlier reference period (PI or at least 1850-1880 historical) to maximize the signal?

Reply: We thank the reviewer for bringing up these points, which we have discussed in the preparation of the study also internally. To address the first point of defining AA: There are different metrics which can be used to describe the difference in temperature change between the northern high latitudes and the global (or mid-latitude / tropical) mean to quantify AA. There are several studies that apply different metrics, e. g., the difference between present and base climate (like us; e.g., Francis and Vavrus, 2015, the ratio, or ratio between linear trends (Johannessen et al., 2016; Kobashi et al., 2013). Indeed the ratio is an established metric, as the reviewer suggests, but there is no fundamental information it carries that the difference would not carry. The reason for choosing the difference is a practical one. When using the ratio of anomalies (e.g., here for the temperature) the denominator may approach small numbers down to zero. In the period of interest, for some model realisations, it turned out that global

warming is rather close to 0 (e.g. model 13 realisation r20i1p1f2 with 0.11 K global warming), so the ratio estimator may be arbitrarily inflating the model spread. A consequence in our study is that when using the ratio metric, the correlation between ALRF and AA degrades to $r = 0.66$, instead of 0.86 as in Figure 1. We consider the LRF a stable metric to quantify AA as it has essentially the same physical basis: The feedback contributes to slight cooling on global average in the time period of interest, but strong warming in the Arctic, both of which is a result of the effect of strong vs. limited mixing abilities in the tropics vs. Arctic on the vertical redistribution of the warming. In the Arctic, this imposes the key feature of bottom-heavy warming, which is AA. Thereby, we chose the difference definition: first, it reduces the problem with small global-warming in some model runs, and second, we can make use of the stronger ALRF-AA relationship by classifying strong/weak AA models also as strong/weak ALRF models, by extension. We now added this explanation in Section 2.1. (L185 ff).

To address the second point of time framing: In our first analyses we did consider the entire time series of historical simulations. However, there are two main periods which are identified to have AA, and both occur in the 20th century: in the 1920–1940s, and at the end of the 20th century continuing into the 21st century (Davy et al., 2018 and references therein). We added this important information at the beginning of the introduction, and in the methods L189 ff. In addition, in Section 3.5, we actually address changes in the reference period (relative frequency of circulation regimes) in ERA5 data. This type of comparison is only feasible with reanalysis data, which starts from 1950. This has led us to adapt the reference time period. Another important point is that large changes in the global surface temperature (simulated and observed) have started to occur since the second half of the 20th century. This leads to the result that excluding the first century of historical simulations imposes no large impact on the order of models 1-31 which are sorted by the degree of AA. Short message: It does not matter for the outcome of this study if 1850–2014 or 1950–2014 is used, but it allows for the inclusion of Section 3.5, and it addresses the second (and stronger) period of identified AA.

Minor

Reviewer Point P 0.4 — l 22 ff and elsewhere in the manuscript: Now that the work is done, I feel that the manuscript would be stronger by focusing on what has been achieved rather than what the authors want to achieve.

Reply: The last paragraph was omitted and L11-12 added "...to provide different perspectives on AA and the Arctic LRF." instead.

Reviewer Point P 0.5 — l. 65 ff: The impact of clouds on the vertical temperature profile has not been introduced at this point in the manuscript.

Reply: That is true. We added a comment in L68 ff.

Reviewer Point P 0.6 — l 205: showing that 2019/2020 is equivalent to 2000-2014 using scenario output would be stronger than just assuming it – strong changes have happened in the Arctic in the early 21st century.

Reply: We thank the reviewer for pointing to our insufficient elaboration on time comparison here. The comment is most valid for comparing the very recent and highly valuable MOSAiC data during 2019/2020 with the last years of historical simulations (2000–2014). Our choice of model data here

is, again, somewhat a result of data availability, which is unfortunately limited in the 6-hourly time resolution: Only three models from Table 1 of the manuscript provide the 6-hourly time resolution also in the scenario simulations ongoing from 2014. However, we acknowledge that this alone cannot justify a comparison here. We argue that our comparison is valid, and show a comparing time series between scenario data (SSP585 as upper boundary of the range of scenarios) for 2019–2020 (MOSAiC time frame) and historical data 2000–2014 for those models that provide both. Fig. R 2 shows that the SSP585 time series lies within the inter-annual range of the 2000–2014 period, and for most of the year, within the range of inter-annual standard deviation. Even though we cannot show this comparison for each model used in our study, we argue that the correspondence between 2000–2014 and 2019–2020 time series from the highest emission scenario justifies our comparison in Section 3.2. We added a comment in L233 ff.

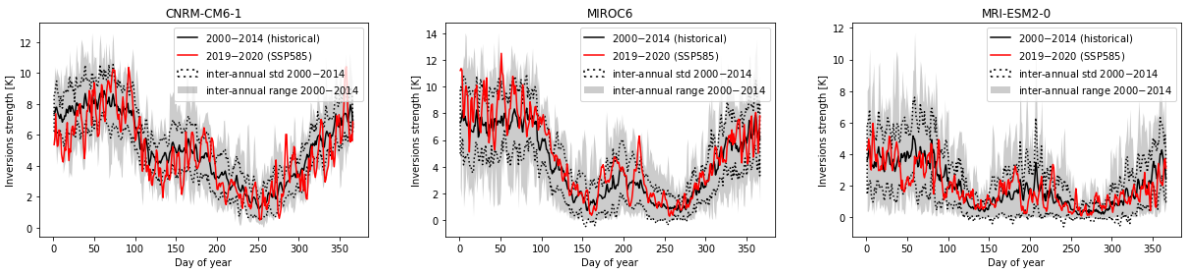

Figure R 2: Comparing time series for surface-based temperature inversion ΔT for MOSAiC conduction time (2019–2020; SSP585 scenario in CMIP6), and for historical data 2000–2014, which we compare to the MOSAiC radiosonde data in Section 3.2 of the manuscript. Those models that facilitate the comparison are CNRM-CM6-1, MIROC6, and MRI-ESM2-0.

Reviewer Point P 0.7 — For the comparison with radiosondes, I would recommend coarsening the radiosonde profiles to the vertical resolution of the models at least as a sensitivity test (same for NSA).

Reply: Unfortunately, the suggested sensitivity test is complicated here, since the model diagnostics are given on model levels. This would require interpolating the models profiles to common pressure levels in order to coarsen the radiosonde profiles to a common vertical resolution. Our approach is to keep each model on its instantaneous vertical resolution and derive the inversion as described in 2.2 and 2.3 of the manuscript. We now specify this approach in L227.

Reviewer Point P 0.8 — Section 2.4: Comparing March/April measurements with DJFM model data – did you check that model data looks similar for March as for the entire winter season?

Reply: There might have been a misunderstanding due to an imprecise formulation on our side. We compare the flight campaign data exclusively during March with model data during the same month (not DJFM). We re-formulated the sentence "The measurements presented here were performed during March to ensure similar thermodynamic conditions compared to the extended winter season, DJFM." to: "Since the measurements presented here are available only for March, we restrict the model-observation-comparison to this month." (L269 ff).

Reviewer Point P 0.9 — Do we expect the 1993 campaign to show the same climate state as the 2019 campaign?

Reply: The reviewer is right, between 1993 and 2019, the climate state is different. However, the comparing time period is 2000–2014, and the year-average of the aircraft campaigns lies within the range of model data: $\mathrm{avg}(1993, 2013, 2019) \approx 2008$. We were still interested in the comparison without the 1993 campaign, but the results are similar. Only over ice, the inversion is slightly elevated and weaker (by around 1 K), which does not affect the conclusions, however. The warming effect by transforming from sea ice to ocean is less, compared to the combination of all campaigns, which brings the observations even closer to the CMIP6/w model ensemble. However, we prefer keeping as many data as possible for the observational constraint: When including too little data, it becomes more illusive to which extend our results are mediated by climate change or ambient meteorology. We thereby included the REFLEX data to achieve a wider range of conditions.

Reviewer Point P 0.10 — l 385: do all models have similar inversion strengths in the reference period?

Reply: We thank the reviewer for bringing more attention to this comparison. The models within each subset do not have exactly the same strength, but both model groups show no overlap (weak-AA models 7.55–10.62 K, and strong-AA models 5.75–6.91 K during DJFM on average). Thereby, the subsets are clearly distinguishable, and the MOSAiC inversion average of 8.49 K lies in the range of CMIP6/w models. We added this important comment in L459 ff, and further elaborate on the statistical representatives of the comparison, primarily during the season of highest interest which is DJFM.

Reviewer Point P 0.11 — l 407: what is the time frame covered by the Kahl (1990) study? Do we expect it to be representative of 2020 conditions?

Reply: Agreed, the mentioned study should not be used in this argumentation here, especially since we expect the inversion strength to decrease with time, which would explain stronger inversions observed in the study of 1990. We drop the reference and adapted the text accordingly.

Reviewer Point P 0.12 — l 487: what significance level? How did you do the bootstrap analysis?

Reply: We now explain the bootstrap analysis more clearly in Section 2.5

Reviewer Point P 0.13 — Fig. 10 and related analysis: This shows data year-round, is there a relevant seasonal cycle?

Reply: There are mild seasonal variations, however the two-state feature is evident throughout the year. A cloudless atmosphere is thereby in approximate RAE, and cloudiness adds a heat source to the boundary layer. This features confirms the results of Figure 2 of the manuscript, and is further in line with previous findings (e.g., Pithan et al., 2014). An explicit evaluation of the seasonality was not pursued here, as this plot is mostly an outlook and frame to the introducing Figure 2 (GCM results also confirmed by LES simulations).

Reviewer Point P 0.14 — l. 564: Cronin and Jansen (2016) would be a good reference here.

Reply: Added here, and also in L602.

Reviewer Point P 0.15 — l. 585–590: I think this is an important result deserving a stronger emphasis in the paper, since entrainment has not received a lot of attention in this context so far.

Reply: The reviewer is right that this is a very interesting result. However, the results are meant to give a final view and supplement to the introducing Figure 2 of the manuscript, rather than following the model-to-OBS/reanalysis framework as the other sections. Therefore, we do not want to overemphasise the point here. However, the implication is clear: entrainment warming due to the presence of clouds is a considerably large heat source for the surface, and the presence of clouds might therefore reduce the change in lapse rate in the lower boundary layer (as already suggested from climate models in Fig. 2). This is an important result for understanding the LRF, and leaves room for deeper studies, not only due to the under-representation of the role of clouds when studying the LRF. We motivate the importance of clear vs. cloudy states in the discussion, but do not further dig into the results here, since this point deserves a dedicated study on its own and would inflate our study at the moment, (rather, dedicated studies are underway).

Reviewer Point P 0.16 — l. 592 "we compile a sizeable amount of observations" Here and elsewhere in the paper: There is nothing to be said against impressing the reader with the large array of observations you bring to the task in addition to CMIP and LES data, but in my view this works better if you leave being impressed to the reader.

Reply: Changed in "We have presented data from several Arctic-based observations and reanalyses in conjunction with co-located CMIP6 model simulations to constrain various processes that mediate both Arctic amplification and the Arctic LRF." Also, the abstract is adapted according to the suggestion.

Reviewer Point P 0.17 — l. 687: I think a crucial point here is that CMIP6/s models generate less warming for a given amount of sea-ice retreat. If this is correct, it should be stated more explicitly.

Reply: It is actually the opposite: Weak-AA models have a stronger present-day inversion (over both sea ice and open ocean), and when transforming from sea ice to open ocean, the expected warming of the lower boundary layer is less compared to CMIP6/s. We now state this more clearly in point 1 of the conclusions.

References

Davy, R., Chen, L., and Hanna, E. (2018). Arctic amplification metrics. *International Journal of Climatology*, 38(12):4384–4394.

Francis, J. A. and Vavrus, S. J. (2015). Evidence for a wavier jet stream in response to rapid arctic warming. *Environmental Research Letters*, 10(1):014005.

Johannessen, O. M., Kuzmina, S. I., Bobylev, L. P., and Miles, M. W. (2016). Surface air temperature variability and trends in the arctic: new amplification assessment and regionalisation. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1):28234.

Kobashi, T., Shindell, D., Kodera, K., Box, J., Nakaegawa, T., and Kawamura, K. (2013). On the origin of multi-decadal to centennial greenland temperature anomalies over the past 800 yr. *Climate of the Past*, 9(2):583–596.

Notz, D. (2015). How well must climate models agree with observations? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2052):20140164.

Pithan, F., Medeiros, B., and Mauritsen, T. (2014). Mixed-phase clouds cause climate model biases in arctic wintertime temperature inversions. *Climate dynamics*, 43:289–303.