Editor's comment

Thank you for submitting a revised version of the manuscript to ACP. I have received two evaluation reports from the original referees. While both referees agree that most of the previous comments are addressed and the manuscript is clearly improved, there are remaining concerns. A major one shared by them both is the potential uncertainty in conclusions imposed by internal variability. The observations and CMIP6 model results are from different time periods, during which the modes of internal variability can be different. Also, how well can the CMIP6/w and CMIP6/s groups of models reproduce the observed internal variability in the analyzed historical period? How will the claimed model biases change if the effect of internal variability is removed from the CMIP6 models? The referees also raised a couple of other major issues. Please refer to their reports. These will need to be addressed before I can make a recommendation for the publication of your manuscript in ACP.

Reply: We thank the editor and reviewers for their constructive comments which have helped us to improve the manuscript. In the revised manuscript we now put a new emphasis on the role of internal variability. It has been criticised before that by using the mean of realizations from each participating CMIP6 model alone might not justify 1) interpreting the differences between CMIP6/w and CMIP6/s subsets (weak and strong AA/ALRF models in the historical period), and 2) constraining the climate relevant parameters by observations. We want to give a general remark on these points, before addressing the specific comments of the reviewers below:

1) By taking the average of model realizations over the past decades, we average out the effect of internal variability, and isolate the response to external forcing. As such, the differences between CMIP6/w and CMIP6/s can be attributed to external forcing. The observations, however, represent a single climate trajectory and thus combine both the effect of internal variability and response to external forcing. We revised the method section that elaborates on the CMIP6 simulations accordingly (newly added Section 2.9 in the manuscript). When comparing the observations to the model subsets (CMIP6/w and CMIP6/s) it is thus important to discuss if the attribution to either one (in our work, based on their respective distributions) can be justified if accounting for internal variability. This leads to the second point:

2) We now discuss our results concerning thermodynamic structure of the boundary layer (i.e., inversion), energy transport, and TOA energy budget (OLR) also in context of internal variability. Specifically, we examine whether the differences between observations, and CMIP6/w / CMIP6/s models, can be explained by internal variability within each subset. In particular, we compute the difference in parameters (OBS minus CMIP/w / CMIP6/s) and compare that difference to the respective range of model realizations which is attributable to internal variability. This range is calculated by subtracting the ensemble mean from each realization (to remove the forced response), and then calculating the central 95 % range of internal variability per model subset. If the OBS-model difference lies within (without) that range, it can (cannot) with confidence be explained by internal variability, which justifies verifying (falsifying) the specific subset based on the OBS. We specify these points within the reviewer's comments and revised the manuscript accordingly.

Author's response to RC1

Major

Reviewer Point P 0.1 — The authors stick with the use of short time series as climatological averages, disregarding internal variability. Especially for the MOSAiC winter, that has been shown to have a particular large-scale circulation with less meridional advection of warm air then other recent winter, I do not think this is an appropriate choice.

Reply: We thank the reviewer for bringing up internal variability and agree that it should be accounted for. Regarding MOSAiC, the observations during winter are roughly consistent with the ensemble mean of CMIP6/w. This leads to the conclusion that the ensemble mean of CMIP6/w models (response to external forcing) more realistically represents the OBS. We acknowledge, however, the reviewer's concern that the observations might be a low-probability trajectory of the climate system, and therefore need to be put into context of the envelope of model realizations.

We show in Fig. R 1 (also added to the manuscript as Fig. B1) the averages of inversion during DJFM, observed and simulated (ensemble averages), corresponding to Fig. 4 of the manuscript. We also indicate the residuals after subtracting the CMIP6/w and CMIP6/s data from the observations. The error bars account for internal variability of the respective model subset. They are computed by subtracting the subset ensemble mean from each realization, and then calculating the central 95 % range (e.g. England et al., 2021).

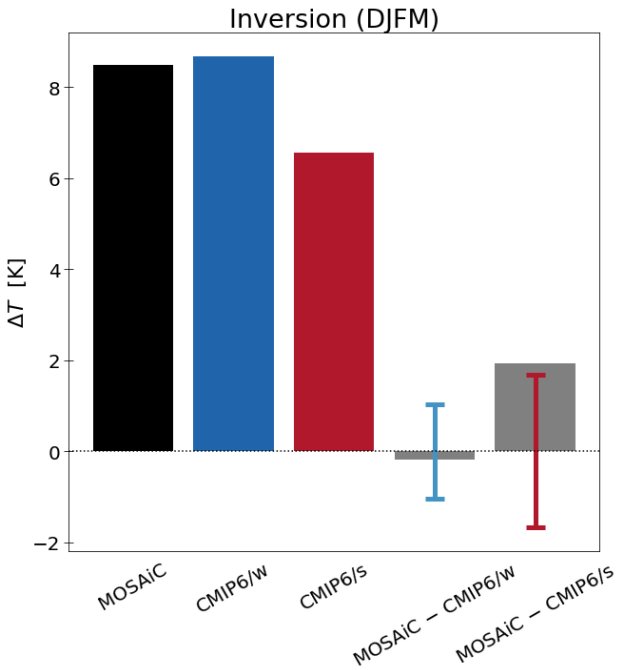


Figure R 1: Averaged inversion during DJFM for MOSAiC, CMIP6/w and CMIP/s (ensemble means, respectively), corresponding to Fig. 4 of the manuscript. Gray bars give the residuals after subtracting the externally forced simulation from the observed inversion. The error bars indicate the 95 % range of simulated internal variability of both CMIP/w (blue), and CMIP6/s (red) models, respectively.

The difference between observations and CMIP6/w is small compared to the one for CMIP6/s (this result is already discussed in the manuscript and has led us to the conclusion that CMIP6/w models more realistically represent the inversion). However, individual CMIP6/s realizations might still be consistent with the observed inversion. Fig. R 1 shows that this is not the case: the MOSAiC − CMIP6/s difference cannot be explained with confidence by internal variability of the CMIP6/s ensemble, as is the case for CMIP6/w. This justifies our main conclusion that CMIP6/s models systematically underestimate the inversion. We added these results to the manuscript in L507ff (a similar analysis is also done for atmospheric energy transport and OLR at TOA in Fig. B1 of the manuscript).

The second point of the reviewer concerns the usage of MOSAiC data in general, as it might represent anomalous inversion conditions. It is true that during MOSAiC the Polarstern experienced certain anomalous events, e.g., extreme cases of warm, moist air transported from the northern North Atlantic or northwestern Siberia during late fall until early spring. Rinke et al. (2021) compared the near-surface meteorological conditions during MOSAiC to the context of the recent climatology (characterised by co-located ERA5 reanalyses with hourly resolution 1979–2020). They show that for the full time series, the near-surface meteorological variables were mostly within the record, even during storms and moisture intrusion events. We want to emphasise in particular that this is true for the near-surface air temperature. In order to respond in depth to the reviewer comment, we examined whether this statement also is true for the inversion strength. The result is shown in Fig. R 2, namely a comparison between MOSAiC inversion time series and co-located ERA5 data as statistics of the 30 years preceding MOSAiC (1991–2020). Note that the MOSAiC inversion appears generally smaller than in the manuscript, since we had to interpolate the radiosonde data to common pressure levels of ERA5 (CMIP6 models have pressure data, ERA5 not).

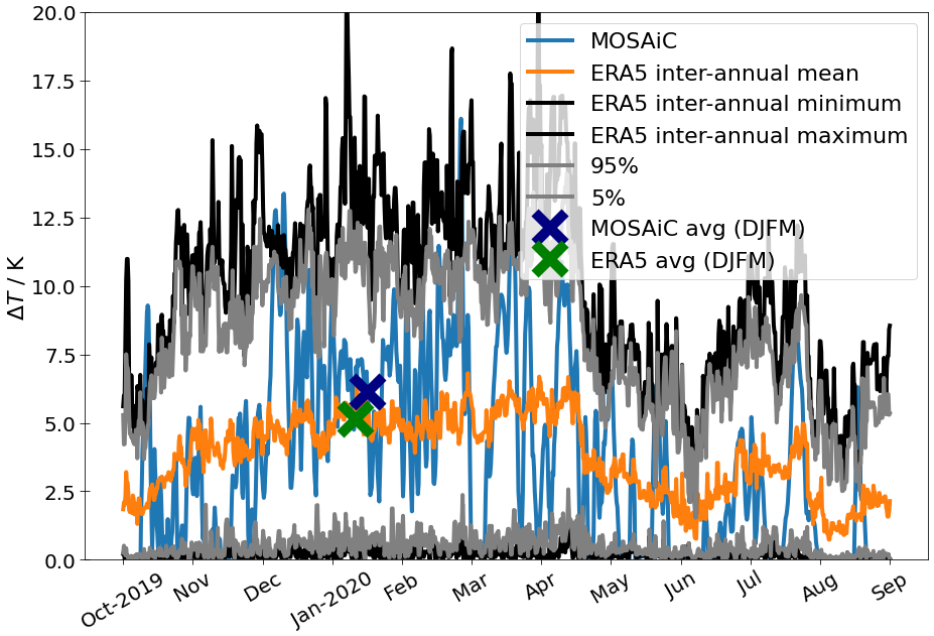


Figure R 2: Inversion strengths ΔT obtained from radio soundings and concurrent 2-m temperature measurements from the nearby ice camp during MOSAiC, and for ERA5. ERA5 inversion data is computed as the difference between the maximum temperature below the 250 hPa isobar, and the surface (as for CMIP6 and MOSAiC). MOSAiC data is interpolated to ERA5 pressure levels. The blue line shows MOSAiC, and the orange line ERA5 data (inter-annual average 1991–2020). Grey lines give the 5th and 95th percentiles, black lines the minimum–maximum range from 1991–2020 data from ERA5, respectively. Blue and green crosses give the DJFM average value for MOSAiC and ERA5, respectively.

In conclusion, the inversion strength observed during MOSAIC was not unusual. Extending the results of Rinke et al. (2021), it is evident that the MOSAIC inversion lies within the climatological range. In particular, the seasonal average during DJFM is close to the average values from the past years, which justifies the comparison between climate models and MOSAiC data. Another line of evidence is that the average winter-time inversion during MOSAiC is fundamentally similar to the average winter inversion during the SHEBA campaign (approx. 8 K in the averaged DJF temperature profile; Stramler et al., 2011). In addition (albeit not relevant for DJFM), recent work of Svensson et al. (2023) shows that for the MOSAiC April (where most of the warm air intrusion events were recorded), the 2-m temperature from observations and ERA5 are in large agreement for most of the month. We added a comment to the manuscript (L484ff).

Reviewer Point P 0.2 — It is unclear to me what the profiles over sea ice ($> 15\,\%$) and their trends actually show. My understanding of the definition is that they would include grid points passing from 100 to 30 % sea-ice cover between the reference and recent period, and thus an important amount of sea-ice retreat, which the authors attempt to exclude from this analysis.

Reply: The reviewer is exactly right: where the sea ice concentration is 15 % or higher, the area is considered ice-covered; where sea ice concentration is below 15 %, the area is considered ice-free.

With this definition we follow the recommendation of the NSIDC (https://nsidc.org/data/soac/sea-ice-concentration), and it allows to classify sea ice conditions and their changes. We define sea ice as areas with SIC of >15 % in both reference and warmer climate, open ocean with SIC of <15 % in both reference and warmer climate, and sea-ice retreat as SIC of >15 % in reference climate and<15 % in warmer climate, respectively (e.g. Lauer et al., 2020; Boeke et al., 2021; Linke and Quaas, 2022). What we are actually interested in (Fig. 2 of the manuscript) is the effect of both, surface type (i.e. difference sea ice vs. ocean profile) and cloudiness (overcast and non-overcast).

By comparing, e.g., the two black lines with squared markers in each panel of Fig. 2, we account for different cloud conditions over an equal surface type (sea ice). This allows to isolate the cloud effect at least partly, and its changes in panel c. By comparing, e.g., the solid black and red line (square and triangle markers, respectively), we compare profiles over sea ice and ocean, respectively, at equal cloud conditions (overcast). This aims to isolate the effect of the surface type on the temperature profile, and its changes in panel c. We do not account explicitly for the profile over sea ice retreat, but it is implied by the surface-type difference, and the way it changes.
In order to address the reviewer's comment, we now adapted the caption of Fig. 2 in the revised manuscript in order to clearly explain the above statements, and further expanded the text (L80ff).

Reviewer Point P 0.3 — The authors conclude that "local processes mediating the lower thermodynamic structure of the atmosphere are more realistically depicted in climate models with weak simulated ALRF/AA in the past". However, in my view the authors have not actually studied the representation of these processes in any subset of models, let alone to the extent to generalize such conclusions. The differences in inversion strengths shown in the manuscript may well be due to different combinations of compensating or non-compensating biases in the underlying processes (mixed-phase clouds, turbulence, sea-ice concentration and thickness, heat conduction through snow and ice).

Reply: The reviewer has a good point that this formulation was misleading. What we intended to do is summarise the results of Section 2.2–2.4 (inversion and profiling), and 2.5–2.6. (energy transport). The former conclusions cover the vertical thermodynamic structure of the lower troposphere, whereas the latter can also impact the free troposphere. We agree that "processes mediating" these features is only partly correct, since there are other processes that can impact them, as rightly pointed out. In response to the reviewer's comment, we now adapt the formulation to clarify that we constrain the climate-relevant parameters that relate to the processes, without excluding other impacts: It is now e.g., "Local, near-surface features like temperature inversion", or "the vertical temperature structure of the Arctic boundary layer". We also added subsection 3.2 and 3.3 to more clearly sort local vs. remote features that are linked to AA/ALRF.

Author's response to RC1

Minor

Reviewer Point P 0.4 — The authors addressed most of my comments, although they didn't reduce the comparisons to fewer but more robust metrics. For that reason, I am hesitant to recommend "accept". For example, a major, remaining concern is the use of OLR data. As recognised by

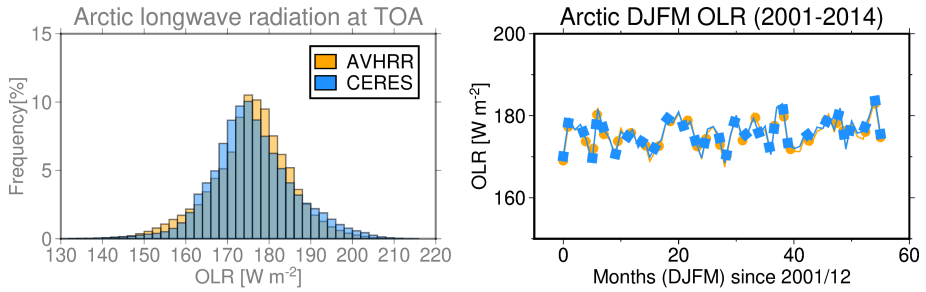


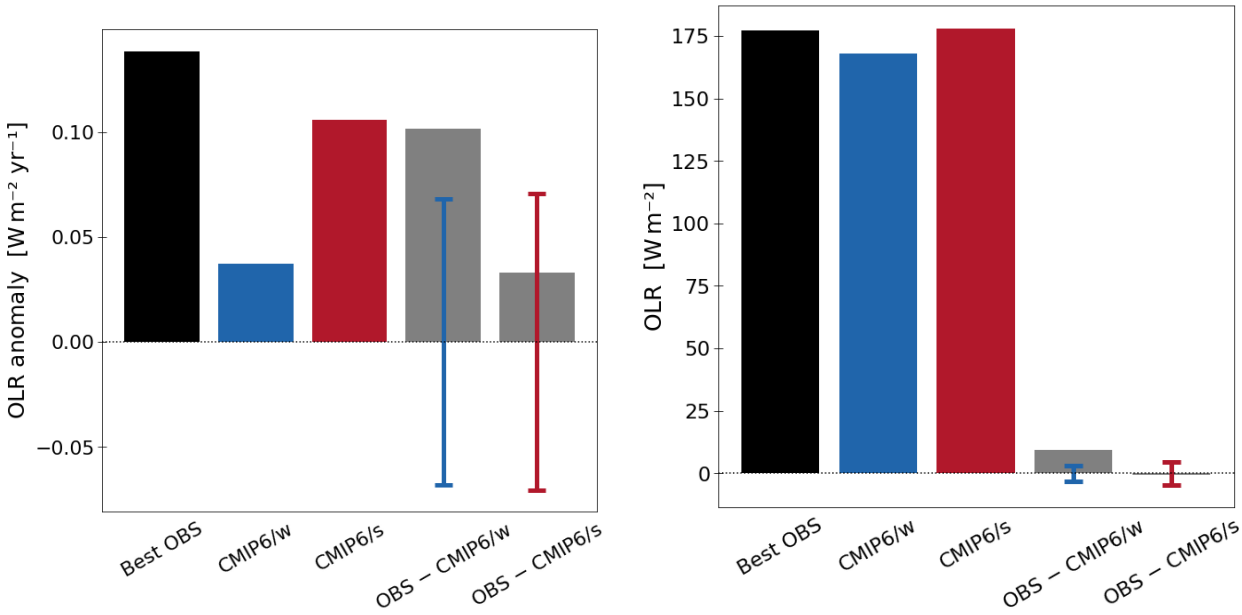
Figure R 3: Comparison between OLR data from CERES and AVHRR. Left: Distribution of OLR at TOA. Right: Overlapping time series during 2001–2014. All values are derived as Pan-Arctic and seasonal averages during DJFM.

the authors in their reply, there are significant uncertainties in the observation datasets, obscuring the determination of a trend signal from it. This uncertainty, together with the uncertainty from matching the time periods (against internal variability in the models), makes it very questionable to use the data to discriminate the linear trends in the GCMs. I'd suggest again the authors think more critically about their use of the different data.

Reply: We thank the reviewer for their comments regarding credibility of observational data, and the role of internal variability. To address the first point regarding credibility of the OBS we added a second satellite from NOAA/NCEI HIRS, just as ERA5 reanalyses data to the previously used AVHRR record in response to the reviewer's concern in the first review. All three datasets support our conclusions. The combined observational estimate is now derived as the average of these three data records (BEST COMB; added to the article). We further added uncertainty ranges to the trends, which are computed as standard deviation of trends following Lelli et al. (2023). In addition, we compared the AVHRR record with the current standard and that is CERES EBAF 4.2 first edition, published on January 27, 2023 (`https://ceres.larc.nasa.gov/documents/DQ_summaries/CERES_EBAF_Ed4.2_DQS.pdf`). CERES data has not been used in the manuscript due to insufficient time coverage, but it is widely used for data evaluation. For the available overlap years, for latitudes north of 66° N and during boreal winter (DJFM), Fig. R 3 shows the OLR distribution (left), and the time series of the two records (right), respectively. Although some small differences can be detected in the distributions (mainly due to surface characterisation, e.g. emissivity), the consistency of the two time series is further confirmation of the robustness of the records, and the soundness of the derived trend data.

The second point addresses the valid concern regarding the role of internal variability. So far, all model-to-OBS/reanalysis comparisons rely on ensemble averages in the climate model data, i.e., internal variability has been averaged out. The observations, however, comprise only one possible climate trajectory reflecting both the response to external forcing as well as internal climate variability, and therefore need to be put into context of the envelope of model realizations. We revised the manuscript and now discuss our main results (temperature inversion, energy transport, and OLR at TOA) also in the context of internal climate variability (see new method Section 2.9 in the manuscript). To address the specific comment of the reviewer, we show in Fig. R 4 the averages of the OLR anomaly trend during DJFM, observed and simulated (ensemble averages), corresponding to Fig. 9 of the manuscript. We also indicate the residuals after subtracting the CMIP6/w and CMIP6/s data from the BEST COMB (best combined estimate from satellite observations and ERA5). The error bars account for internal variability of the respective model subset. They are computed by subtracting the subset ensemble mean

from each realization, and then calculating the central 95 % ranges (England et al., 2021).



(a) OLR trend anomaly w.r.t. 1983–1994 (DJFM)

(b) OLR average 2000–2014 (DJFM)

Figure R 4: a) Averaged OLR anomaly trend during DJFM for BEST COMB (average from NOAA/NCEI HIRS, ERA5, and AVHRR), CMIP6/w and CMIP/s (ensemble means, respectively), corresponding to Fig. 9 of the manuscript. Gray bars give the residuals after subtracting the externally forced simulation (ensemble means) from the BEST COMB. The error bars indicate the 95 % ranges which could be explained by internal variability per subset. b) Same as a), but for climatological OLR averages 2000–2014.

Fig. R 4a indicates that the difference between BEST COMB and CMIP6/s is smaller compared to CMIP6/w (this result is already discussed in the manuscript, and it has led us to the conclusion that the mean of CMIP6/s simulations more realistically represents the observed OLR trends, albeit still underestimating them). However, also the CMIP6/w ensemble members might still be consistent with the observed OLR trends. Fig. R 4 now allows to conclude that this is unlikely. The fact that the BEST-COMB − CMIP6/s difference is within the range of internal variability simulated by CMIP6/s, but the BEST-COMB − CMIP6/w difference cannot be fully explained by the range of internal variability simulated by CMIP6/w, justifies our previous conclusion. We further show the absolute values of OLR at the end of the historical period (2000–2014), with a similar result: The difference BEST-COMB − CMIP6/s is small compared to OBS − CMIP6/w, and the differences are covered by the range of simulated internal variability for CMIP6/s, but not for CMIP6/w. Fig. R 4a is now added to the manuscript (as Fig. B1), together with a similar analysis of MOSAiC inversion and Pan-Arctic energy transport, to support the main conclusions of our study with regards to the role of internal variability.

References

Boeke, R. C., Taylor, P. C., and Sejas, S. A. (2021). On the nature of the arctic's positive lapse-rate feedback. *Geophysical Research Letters*, 48(1):e2020GL091109.

England, M. R., Eisenman, I., Lutsko, N. J., and Wagner, T. J. (2021). The recent emergence of arctic amplification. *Geophysical Research Letters*, 48(15):e2021GL094086.

Lauer, M., Block, K., Salzmann, M., and Quaas, J. (2020). Co2-forced changes of arctic temperature lapse-rates in cmip5 models. *Met. Z.*, 29(1):79–93.

Lelli, L., Vountas, M., Khosravi, N., and Burrows, J. P. (2023). Satellite remote sensing of regional and seasonal arctic cooling showing a multi-decadal trend towards brighter and more liquid clouds. *Atmospheric Chemistry and Physics*, 23(4):2579–2611.

Linke, O. and Quaas, J. (2022). The impact of co_2-driven climate change on the arctic atmospheric energy budget in cmip6 climate model simulations. *Tellus A: Dynamic Meteorology and Oceanography*, 74(2022).

Rinke, A., Cassano, J. J., Cassano, E. N., Jaiser, R., and Handorf, D. (2021). Meteorological conditions during the mosaic expedition: Normal or anomalous? *Elem Sci Anth*, 9(1):00023.

Stramler, K., Genio, A. D. D., and Rossow, W. B. (2011). Synoptically driven arctic winter states. *Journal of Climate*, 24(6):1747 – 1762.

Svensson, G., Murto, S., Shupe, M., Pithan, F., Magnusson, L., Day, J., Doyle, J., Renfrew, I., Spengler, T., and Vihma, T. (2023). Warm air intrusions reaching the mosaic expedition in april 2020—the yopp targeted observing period (top). *Elem Sci Anth*, 11.