Author's response to RC2

We thank Reviewer 2 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 — This paper aims to dissect the Arctic warming simulated in the CMIP6 models by comparing them to observations. The analysis is centered on the geophysical variables related to the lapse rate feedback, which, as argued by a number of studies, is of critical importance for the Arctic warming amplification. To the extent this argument is valid, the comparisons in this paper are well motivated. A novel aspect of this paper is that it includes comparisons to several different kinds of data, some of which, such as the newly acquired Mosaic campaign data, provides fresh perspectives for model validation. However, although each comparison included here potentially provides a useful line of evidence for discriminating the models, unfortunately few results appear conclusive in the end. This calls into question whether one had better aim to identify and focus on what can be more conclusively stated about the models and/or nature, as opposed to a somewhat nonselective listing of results.

Reply: We acknowledge the major criticism that is being raised here, and we would have also hoped for a more clear story in parts. However, we want to justify presenting each of the results in this study: Firstly, to some extent and especially for the model comparison to the observations at higher time resolution, we do not have clear emerging relationships between simulated AA/ALRF and present-day climate aspects that is used to constrain the mediating processes. This is simply due to the fact that only few simulations exist, even in the historical simulations, to derive a relationship as done for e.g., an emergent constraint. However, especially these local processes are crucial in better understanding and constraining both ALRF, and AA as a whole. Our idea was to chose those models at the edge of the AA/ALRF distribution as a compromise. If the observational constraint then fits either one of these categories, we have a clear signal. If the OBS are in the range of inter-model mean, then the attribution to either weak or strong-AA/ALRF is less straightforward. This might appear as somewhat inconclusive, but it is still a result, i.e., an attribution to the inter-model mean. That is why we want to show all results, also to cover each process that is believed to have a mediating impact (inversion, sea ice retreat, transport, ...) and reduce gaps in the interpretation. We prefer not to select some that match one line of evidence and omit others. The collaborative project continues and in the next project phase one aim is to reconcile the differences in conclusions.

We adapted several major changes to the manuscript to clarify our intention: At the end of each section, we now comment more carefully on the significance in model differences, and the attribution of several observations to either one of the emerging subsets, just as the synergy that gradually appears trough the result sections. These inter-mediate results are later brought into context: The conclusion focuses only on results that show a clear signal in both model subset differences and their constraint through observations, and further brings attention to the synergy that emerges between inter-mediate results.

Reviewer Point P 0.2 — Moreover, the use of some data and analysis methods are not sufficiently explained (see comments below), raising questions about their properness. For these reasons, I think the paper would need a major revision before being considered for publication.

Reply: We thank the reviewer for their specific suggestions and performed the major revisions.

Reviewer Point P 0.3 — Figure 1. Can you also provide the observations for a comparison in these diagnostics?

Reply: It is a very good suggestion to add the observed estimate for AA 1985–2014 with respect to 1951–1980. We added the average from several observational estimates to Figure 1 of the manuscript. We present the OBS estimate of AA in Figure 1 of the manuscript as average from several observational data set (GISTEMP, Berkeley Earth, HadCRUT5, NOAA's MLOST, and ERA5; Rantanen et al., 2022). An overview of the observed AA as time series is shown in Fig. R 1. We added comments on the OBS estimate in Section 2 L151 ff. In addition, the inclusion of OBS in the introducing plot allows us to interpret the simulated model range with respect to observations. It ensures that our classification of either weak or strong-AA model subset actually shows AA values below or above the OBS, respectively. We include this interpretation in Section 3.1. L420 ff, and additionally expand the elaboration on statistical significance in the discussion. This concerns primarily the previously mentioned model-to-OBS comparisons at 6-hourly time resolution (MOSAiC, NSA, dropsondes), which is limited by the availability of models. Simply categorising the model range by taking the top-3 lowest and highest AA models might not do justice to the classification as either weak or low AA simulations at 6-hourly resolution, since the entire model spread (all models in Table 1 of the manuscript) is larger. However, by adding the observations we show that the discrimination is still valid, since the sub-set average of AA for CMIP6/w, and CMIP6/s lies below, and above the OBS estimate, respectively (for any time-resolution group). This gives further justification to our approach.

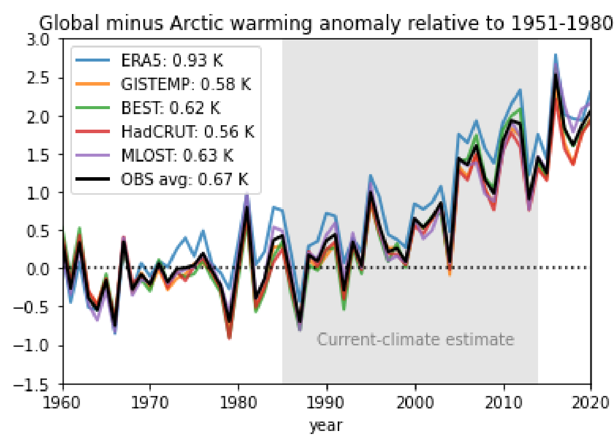

Figure R 1: Time series of AA: Difference in annual mean temperature anomalies in the Arctic with respect to global average as derived from the various observational datasets. Temperature anomalies have been calculated relative to the 30-year period of 1951–1980.

Reviewer Point P 0.4 — L172 "consistency": can you provide any reference to this belief? Note that it is quite known that there are noticeable differences between different kernels, especially in the Arctic. In either case, it would be move convincing to provide an error bar based on results computed from more than one kernel.

Reply: We agree with the reviewer that the formulation "consistency" is too unspecific in this context. We show in Fig R 2 the same scatter plot as in Figure 1 b of the manuscript, but with ALRF values derived from different kernels (the inter-model distribution of AA is not effected by the choice of kernel). There are indeed differences in the quantification of the ALRF across the kernels, and slight variations of the inter-model correlation between AA and ALRF. We acknowledge the criticism being raised from the reviewer and now show the scatter plot in Fig. 1 b, but with model-specific ALRF values derived as average from the output of all kernels. To avoid making Fig. 1 even more busy, we account for the inter-kernel spread across ALRF by adding the standard deviation in Table 1. We further added a comment on method Section 2.1 L192 ff. Albeit there are difference in the relationship of the inter-model spread in AA and ALRF across CMIP6 models, we emphasise that our results are not sensitive to the choice of kernels. The classification as either weak or strong-AA models remains unaffected, and the AA-ALRF relationship increases even for other kernels than the previously chosen HadGEM3 kernel. Thereby, the attribution of weak/strong-AA models to equally weak/strong-ALRF models is still valid. The newly added comment proves that point, and is important for the credibility of our results.

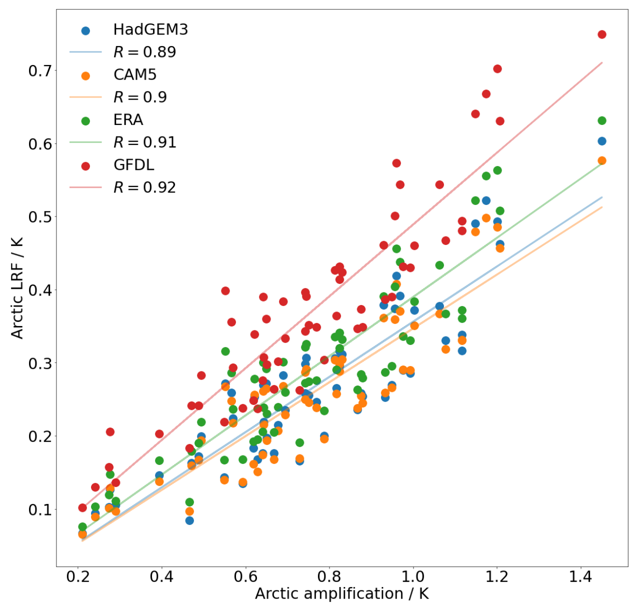

Figure R 2: As Fig. 1 b of the manuscript, but with different kernels to derive the ALRF.

Reviewer Point P 0.5 — L205, 251 use of years of 2010–2014. Can you justify the use of these model years to match the observation? It's understood coupled model years are nominal but what guarantees a comparison done here, between a single realization of nature of limited length and multiple model years, is proper? Very handwavy to "assume" they're "roughly the same".

Reply: We thank the reviewer for pointing to our insufficient elaboration on time comparison here. This comment is most valid for the limited data comparison at 6-hourly time resolution and concerns mostly the evaluation of model data based on very recent MOSAiC data (2019–2020). We have been considered using data from CMIP6 scenarios to expand the simulation period to the years following the historical simulations. However, this was again limited by the availability of data: Only three models in Table 1 of the manuscript provide the required diagnostics for simulation scenarios ongoing from 2014, which was not an option. We were still highly interested in comparing climate models against the valuable data conducted during the MOSAiC expedition. It is true that the time shift between 2000–2014 and 2019–2020 raises questions about the validity of the comparison. To prove that it is still valid to treat this periods as part of the same climate state, we show for the three models with scenario output the time series comparison between 2000–2014 and 2019–2020 in Fig. R 3. We use scenario outputs from the highest emission scenario SSP585 as boundary of the range of scenarios for 2019–2020. Even for this highest scenario, the 2019–2020 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 2.2 We added a comment in L233 ff.

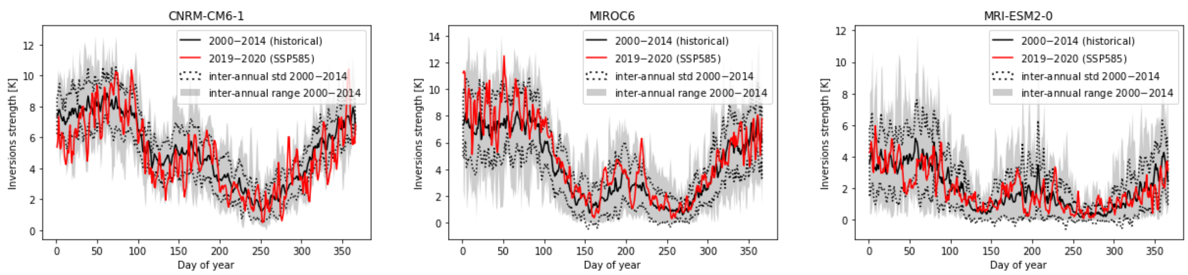

Figure R 3: Comparing time series for surface-based temperature inversion δT for for MOSAiC conduction time (2019–2020; SSP585 scenario in CMIP6), and for historical data 2000–2014, which we compare to the MOSAiC radiosonde data in Section 3.2 of the manuscript. Those models that facilitate the comparison are CNRM-CM6-1, MIROC6, and MRI-ESM2-0.

Reviewer Point P 0.6 — L261 The identification of different "regimes" looks an interesting approach to me. However, I found the description of the method too brief here. I'd suggest showing the relevant results such as the EOFs, as well as the associated PCs and eigenvalues. I think this method, like the other data and methods in this paper, is worth more careful/critical reasoning and more thorough discussion.

Reply: For this part of the study, we have used the concept of atmospheric circulation regimes (e.g., Hannachi et al., 2017) to characterise the large-scale circulation in terms of a few preferred states. This concept provides a framework for understanding low-frequency variability due to transitions between different regimes. In addition, Palmer (1993, 1999) introduced a dynamical paradigm for climate change which suggests, that a weak external forcing does not change the structure and number of atmospheric regimes, but instead changes the frequency of occurrence of the regimes. Since then, many studies have analysed the atmospheric circulation within this concept (see extended review by Hannach et al., 2017). To follow the reviewer's advise, we extended the description of the method for the determination of the regimes in section 2.5: L300 ff.

To characterize the reduced state space, we show here the spatial structure of the five leading EOFs over the North-Atlantic-Eurasian region (Fig. R 4a, 57.5% explained variance) and over the North-Pacific region (Fig. R 4 b, 54.5% explained variance) based on ERA5 daily mean SLP anomaly fields for the extended winter season (DJFM). The leading EOFs resemble the well-known teleconnection

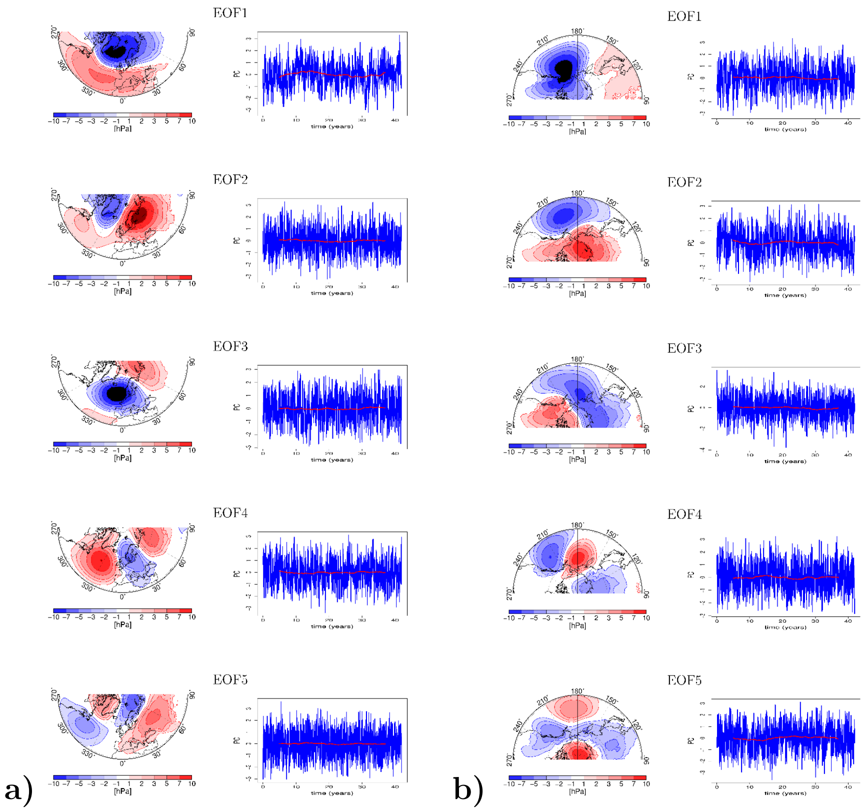

Figure R 4: **a)** Left from top to bottom: Five leading EOFs over the North-Atlantic-Eurasian region for DJFM, based on ERA5 daily mean SLP anomalies for DJFM, explaining 17.5 %,14.3 %,11.5 %, 8.9 %, 5.4 %. **b)** Left from top to bottom: Five leading EOFs over the North-Pacific region for DJFM, based on ERA5 daily mean SLP anomalies for DJFM, explaining 16.3 %, 13.7 %, 11.0 %, 7.2 %, 6.6 % of the total variance respectively. Right from top to bottom: Corresponding time-series of principal components (normalized), the red line represents running mean values over 10 years.

patterns such as the North-Atlantic Oscillation (North-Atlantic EOF1), Scandinavia pattern (North-Atlantic EOF2), East Atlantic pattern (North-Atlantic EOF3), Pacific/North American pattern (North-Pacific EOF1), West Pacific pattern (North-Pacific EOF2).

Reviewer Point P 0.7 — L290 What's the basis of using this proxy as a quantitative measure of the energy transport? How can the TOA-only perspective differentiate atmospheric vs. oceanic transports? How is equilibrium verified, so that horizontal transport can be inferred from vertical energy flux?

Reply: We thank the reviewer for commenting on the derivation of the transport term and acknowledge that the method has not been explained sufficiently at this point. In addition, we understand that it is useful to exclude the ocean signal while looking at atmospheric processes. Therefore, we now present the transport term as contribution from the atmosphere, and as divergence term wihtin an energy budget framework: Following previous works of e.g., (Nakamura and Oort, 1988; Trenberth, 1997; Serreze et al., 2007) we can consider the energy budget of an atmospheric column that extends from the surface to the TOA. For each column, the tendency in energy storage within an atmospheric column E_a can be estimated as

$$\frac{\partial E_\mathrm{a}}{\partial t} = R_\mathrm{a} + Q_\mathrm{H} - \nabla \cdot \vec{F}_\mathrm{a}, \tag{1}$$

with net atmospheric radiation budget R_a, the sum of turbulent heat fluxes at the surface Q_H, and the convergence of the horizontal atmospheric energy transport $-\nabla \cdot \vec{F}_\mathrm{a}$. In the long-term and large-scale energy budget, we can further neglect the storage tendency under assumption of steady state (Serreze et al., 2007; Linke and Quaas, 2022). We use this simplified energy budget framework to estimate the horizontal convergence of energy transport indirectly, i.e., residual of the budget equation. According to the reviewers comment, we expanded the method description in Section 2.6, specifically clarifying that we exclusively consider the atmospheric transport (convergence), and further commenting on the equilibrium criteria. Both results and discussions are updated accordingly.

Reviewer Point P 0.8 — L305 and Figure 9, concerning the use of satellite OLR records, it should be noted that various issues had been documented on how wrong it could be to take a non-SI-traceable radiation record as the ground-truth of "observed" long-term trends. For example:
Trishchenko et al. https://doi.org/10.1029/2002JD002353
Wong et al. https://doi.org/10.1175/JCLI3838.1
The OLR trending itself would be worth a full section if not a paper by itself. Before its correctness is established, it is very questionable to use this result as a model discrimination metric.

Reply: It is an important point being raised how to define the ground-truth of our data set. We want to comment on the credibility of the AVHRR climate data record: It is well known that suboptimal radiometric calibration of the AVHRR thermal channels might lead to inconsistencies, then source of discrepancies in Arctic cloud detection and radiation fluxes at the surface (Zygmuntowska et al., 2012). Prior to the production of the satellite record, every PM sensor was cross-calibrated with well-behaved sensors. The SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY) served as spectral reference for the visible wavelengths and the Infrared Atmospheric Sounding Interferometer (IASI) for the thermal channels (Stengel et al., 2020). This resulted in an improvement of the retrieved cloud parameters (Sus et al., 2018; McGarragh et al., 2018) in terms of precision, accuracy and stability (Stengel et al., 2017). Specifically to the Arctic, the AVHRR cloud record has not shown any scale-dependent bias upon validation with coincident measurements at four high-latitude ground sites (Vinjamuri et al., 2023). The accuracy of the cloud record is a key factor because the cloud properties are input for the derivation of the broadband fluxes (Henderson et al., 2013). The resulting accuracy in AVHRR-derived OLR amounts to $\pm 3\,\mathrm{W\,m^{-2}}$ against observations of the Geostationary Earth Radiation Budget (GERB) radiometer on board the Meteosat Second Generation (MSG-2) satellite (Christensen et al., 2016). This value is within GERB's calibration limits for radiation at TOA (Clerbaux et al., 2009). In relative terms, the average long-term bias of AVHRR-derived outgoing LW fluxes against CERES amounts to -2.7% (Stengel et al., 2020). In addition, using the same algorithm for the broadband fluxes, but applied at CloudSat, CALIPSO, and MODIS measurements instead, Kay and L'Ecuyer (2013) quantify an average bias against the Clouds and the Earth's Radiant Energy System (CERES) Energy Balanced and Filled (EBAF) record (Kato et al., 2018; Loeb et al., 2018) of the order of 4-5 $\mathrm{W\,m^{-2}}$. Consequently, the present AVHRR record has been used for the analysis of Arctic cloud radiative forcing (Lelli et al., 2023) and feedback (Philipp et al., 2020).

However, we acknowledge the comment of the reviewer and add additional data sources as reference to compare against CMIP6 models. The additional data records are the OLR flux from NOAA/NCEI from the High Resolution Infrared Radiation Sounder (HIRS) instruments on board the NOAA and MetOp satellites, and ERA5 reanalyses. Additionally, we adapt the data record in Figure 9 of the manuscript to display the anomaly of the OLR with respect to the first 15 years of the AVHRR record (1983–1997). We apply this change since the main focus is the trend in OLR during recent decades, which is more straightforward in the new plot version. We further consider CERES satellite data in the absolute time series (not shown), which fits well the records derived for HIRS, AVHRR, and ERA5. Only in the anomaly plot as now presented in the manuscript, CERES does not appear due to insufficient time coverage (start 2000).

Reviewer Point P 0.9 — L380 "significant". Although significant differences are stated here and at multiple other places (in this (Figure 4) and other figures), looking through these results, I am not convinced there is indeed any strong difference between the compared groups, either between "w" vs "s" or between them and the observation (Mosaic). If the discriminations are based on such weak evidence, I am not sure the observation used here provides any useful constraint as wished by the authors, or any model evaluation result can be considered conclusive. Please critically review and reason about this and other conclusions.

Reply: We thank the reviewer for pointing that out. In the specific case, and also in the two following analyses including 6-hourly data, we partly omitted the word "significance" since it implies the usage of relevant statistical tests. We want to elaborate a bit more on the raised criticism of weak evidence: It is true, primarily in the specifically mentioned case of comparing inversion data from models and observations, we cannot rely on strong inter-model relationships to constrain the mediating process of atmospheric stability. The motivation of constraining these highly-defining process to some extent with recent sate-of-the-art climate models is however a tempting option. Connecting to our first reply of this review, we rather show the full story rather than selecting some that match one line of evidence. However, in reflection of all results presented in the conclusions, we focus on what is, from our view, important and conclusive. This concerns two steps of the method: First, identifying model difference where they are clear and consistent across the results, second, attributing co-located observations to either one of the emerging categories (our proposed constraint).

We acknowledge the reviewers comment and elaborate more on data credibility. For the mentioned case that concerns the discussed season of ONDJFM: We argue that the model discrimination (albeit limited by data availability) is supported by there being no overlap in mean inversion strength across the models: During ON for CMIP6/w/s: 4.6–5.8 K / 1.8–3.5 K, and during DJFM for CMIP6/w/s 7.6–10.6 K / 5.8–6.9 K, respectively. A paragraph is added in Section 3.1 L459. This attribution to a specific model subset, not only in the subset average, but also for individual models, is true also for Section 3.2 and 3.3, that use the same models. We further performed a two-sample Kolmogorov-Smirnov test to compare the similarity of CMIP6/w and CMIP6/s distributions (addressed in the same paragraph in Section 3.1). Thereby, we conclude that the first point of model categorisation is fulfilled. Further, primarily during MOSAiC winter, the inversion distribution is most attributable to the range of CMIP6/w models, which is why in the conclusion we highlight the outcome, supported by the two following sections. The shared outcome of Section 3.1–3.3 is now more critically discussed in the discussion.

We further more critically review each individual Section regarding model subset discrimination, and constraints by observation, and further focus more specifically on the synergy of all Sections in the conclusion.

References

Christensen, M., Poulsen, C., McGarragh, G., and Grainger, R. (2016). Algorithm Theoretical Basis Document (ATBD) of the Community Code for CLimate (CC4CL) Broadband Radiative Flux Retrieval (CC4CL-TOAFLUX) module - Cloud_CCI Working Group. Technical report, European Space Agency. Last access July 2019.

Clerbaux, N., Russell, J., Dewitte, S., Bertrand, C., Caprion, D., De Paepe, B., Gonzalez Sotelino, L., Ipe, A., Bantges, R., and Brindley, H. (2009). Comparison of GERB instantaneous radiance and flux products with CERES Edition-2 data. *Remote Sensing of Environment*, 113(1):102–114.

Hannachi, A., Straus, D. M., Franzke, C. L., Corti, S., and Woollings, T. (2017). Low-frequency nonlinearity and regime behavior in the northern hemisphere extratropical atmosphere. *Reviews of Geophysics*, 55(1):199–234.

Henderson, D. S., L'Ecuyer, T., Stephens, G., Partain, P., and Sekiguchi, M. (2013). A multisensor perspective on the radiative impacts of clouds and aerosols. *Journal of Applied Meteorology and Climatology*, 52(4):853 – 871.

Kato, S., Rose, F. G., Rutan, D. A., Thorsen, T. J., Loeb, N. G., Doelling, D. R., Huang, X., Smith, W. L., Su, W., and Ham, S.-H. (2018). Surface irradiances of edition 4.0 clouds and the earth's radiant energy system (ceres) energy balanced and filled (ebaf) data product. *Journal of Climate*, 31(11):4501 – 4527.

Kay, J. E. and L'Ecuyer, T. (2013). Observational constraints on Arctic Ocean clouds and radiative fluxes during the early 21st century. *Journal of Geophysical Research: Atmospheres*, 118(13):7219–7236.

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).

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., Liang, L., Mitrescu, C., Rose, F. G., and Kato, S. (2018). Clouds and the earth's radiant energy system (ceres) energy balanced and filled (ebaf) top-of-atmosphere (toa) edition-4.0 data product. *Journal of Climate*, 31(2):895 – 918.

McGarragh, G. R., Poulsen, C. A., Thomas, G. E., Povey, A. C., Sus, O., Stapelberg, S., Schlundt, C., Proud, S., Christensen, M. W., Stengel, M., Hollmann, R., and Grainger, R. G. (2018). The community cloud retrieval for climate (cc4cl) – part 2: The optimal estimation approach. *Atmospheric Measurement Techniques*, 11(6):3397–3431.

Nakamura, N. and Oort, A. H. (1988). Atmospheric heat budgets of the polar regions. *Journal of Geophysical Research: Atmospheres*, 93(D8):9510–9524.

Palmer, T. N. (1993). Extended-range atmospheric prediction and the lorenz model. *Bulletin of the American Meteorological Society*, 74(1):49–66.

Palmer, T. N. (1999). A nonlinear dynamical perspective on climate prediction. *Journal of Climate*, 12(2):575–591.

Philipp, D., Stengel, M., and Ahrens, B. (2020). Analyzing the Arctic Feedback Mechanism between Sea Ice and Low-Level Clouds Using 34 Years of Satellite Observation. *Journal of Climate*, 33(17):7479 – 7501.

Rantanen, M., Karpechko, A. Y., Lipponen, A., Nordling, K., Hyvärinen, O., Ruosteenoja, K., Vihma, T., and Laaksonen, A. (2022). The arctic has warmed nearly four times faster than the globe since 1979. *Communications Earth & Environment*, 3(1):168.

Serreze, M. C., Barrett, A. P., Slater, A. G., Steele, M., Zhang, J., and Trenberth, K. E. (2007). The large-scale energy budget of the arctic. *Journal of Geophysical Research: Atmospheres*, 112(D11).

Stengel, M., Stapelberg, S., Sus, O., Finkensieper, S., Würzler, B., Philipp, D., Hollmann, R., Poulsen, C., Christensen, M., and McGarragh, G. (2020). Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties. *Earth System Science Data*, 12(1):41–60.

Stengel, M., Stapelberg, S., Sus, O., Schlundt, C., Poulsen, C., Thomas, G., Christensen, M., Carbajal Henken, C., Preusker, R., Fischer, J., Devasthale, A., Willén, U., Karlsson, K.-G., McGarragh, G. R., Proud, S., Povey, A. C., Grainger, R. G., Meirink, J. F., Feofilov, A., Bennartz, R., Bojanowski, J. S., and Hollmann, R. (2017). Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the Cloud_cci project. *Earth System Science Data*, 9(2):881–904.

Sus, O., Stengel, M., Stapelberg, S., McGarragh, G., Poulsen, C., Povey, A. C., Schlundt, C., Thomas, G., Christensen, M., Proud, S., Jerg, M., Grainger, R., and Hollmann, R. (2018). The community cloud retrieval for climate (cc4cl) – part 1: A framework applied to multiple satellite imaging sensors. *Atmospheric Measurement Techniques*, 11(6):3373–3396.

Trenberth, K. E. (1997). Using Atmospheric Budgets as a Constraint on Surface Fluxes. *J. Climate*, 10(11):2796 – 2809.

Vinjamuri, K. S., Vountas, M., Lelli, L., Stengel, M., Shupe, M. D., Ebell, K., and Burrows, J. P. (2023). Validation of the cloud_cci cloud products in the arctic [preprint]. *Atmospheric Measurement Techniques Discussions*, pages 1–24.

Zygmuntowska, M., Mauritsen, T., Quaas, J., and Kaleschke, L. (2012). Arctic Clouds and Surface Radiation - a critical comparison of satellite retrievals and the ERA-Interim reanalysis. *Atmospheric Chemistry and Physics*, 12(14):6667–6677.