General comments

This paper addresses temperature and precipitation extremes in the Arctic in CESM experiments with prescibed SST and sea ice. Two versions of CESM, one with uniform resolution and another with a refined grid at high latitudes are considered. First, comparisons with reanalysis and regional model datasets are presented for near-present climate simulations, and second, future changes in extremes are evaluated for two storylines in a high-emission scenario.

The results show that in warmer climate conditions, high temperature extremes and precipitation extremes will become stronger and/or more frequent, while cold/dry extremes will generally become weaker and/or less frequent in the Arctic region. Aside from these somewhat expected findigs, there are some differences between the storylines and some (mainly subtler) differences between the experiments performed with the two different grids.

This paper represents an interesting application of the storyline approach to the future Arctic climate, also documenting how the results depend on the model grid. Overall, the paper is thorough and well-written. The technical quality of figures is good, but statistical testing is mostly missing.

Major comments

1. The evaluation of statistical significance of the findings is largely lacking in this work. In most of the figures, it is not considered at all, and while it is included in some (e.g., Figs. 3 and 13), the method for calculating the statistical significance is not explained. For example, does the statistical testing account for the multiplicity problem? That is, when tests are performed for a large number of grid points, some of them would likely show nominally significant differences even if the results were generated randomly. See, for example:

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. Bull. Amer. Meteor. Soc., 97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Testing the statistical significance is especially relevant because the climate change signal is defined based on relatively short runs (10 years) and because extremes are more subject to the effects of internal variability than mean values.

While, for example, the general increase of warm extremes and wet extremes at most Arctic locations might be robust, it is less obvious how robust are the differences between the changes in extremes for the two storylines, let alone the corresponding differences between POLARRES and and NE30 runs. Indeed, it is obvious that some of the figures contain mostly noise (Fig. S13f,h being a particularly "good" example).

I recommend that the statistical significance of the results should be evaluated more systematically, and the reporting should be mainly focused on those aspects found robust. In the case of areal-mean results, such as those Fig. 10 and 12, uncertainty could be illustrated by showing confidence intervals. Just be careful with their interpretation:

Lanzante, J. R., 2005: A Cautionary Note on the Use of Error Bars. J. Climate, 18, 3699–3703, https://doi.org/10.1175/JCLI3499.1.

2. The manuscript is not fully clear about how regridding of the results from different model runs and reanalysis to a common grid is made. For example, is it made using conservative interpolation (which conserves area means) or bilinear interpolation between the nearest grid points, or some other approach? Also, it is not stated clearly, whether regridding is applied to the original temperature and precipitation values before calculating the extremes, or whether the extremes are first calculated at the original model resolution and then regridded to the common grid for plotting.

These choices are especially important for precipitation extremes. In particular, I would expect that if (1) bilinear (rather than conservative) interpolation is used and/or (2) the extremes are calculated before regridding, there would be a systematic increase in high precipitation rates (e.g., P99) with improving resolution, simply because daily precipitation values show substantial small-scale variations. On the other hand, if extremes are calculated after a conservative regridding is applied, it is not obvious this will happen (or at least, there would be no trivial reason for that).

There is not necessarily a single "right" solution to how this matter should be handled, but at any rate, you should be clear about how you do it, and justify your choice. See also minor comments 11 and 12.

Minor comments

1. lines 13 and 102: replace the latter "strong/weak" with "weak/strong"? The idea of ST1 and ST2 is to contrast a case with strong land warming and weak SST warming with a case with weak land warming and strong SST warming.

2. line 16: It could be mentioned that the better performance of the 1° grid in simulating temperature extremes is related to a larger negative temperature bias in the VR grid. (Incidentally, I think this might be a matter of model tuning rather than any fundamental issue associated with higher resolution).

3. line 124: You can delete "cloud-aerosol" before "radiation scheme".

4. Some parts of the model and experimental description appear unnecessarily detailed in the context of this work. In particular, the description of CLM5 on lines 132–143 and land surface treatment on lines 216-223.

5. lines 176–177: at least in CICE4, increasing the parameter r_{snw} actually decreases the snow grain radius over sea ice, and therefore, increases snow albedo. Please check this.

6. lines 182–184: Are microphysical substep and microphysical timestep the same thing or not?

7. lines 198–203: Please use a notation that it consistent with Levine et al. (2024). Storyline ST1 corresponds to storyline D (not B2) in Levine et al., denoted as ArcAmp+/BKWarm- (not PolAmpl+BKSSTWarm-). Similarly, ST2 corresponds to storyline A (not B1) denoted as ArcAmp-/BKWarm+ in Levine et al.

8. line 213: It would be useful to mention the global-mean near-surface temperature change from the present-day (2005–2014) to the future period (2090–2099) for ST1 and ST2.

9. Sect. 2.3. Mention that in regions with sparse observations, such as large parts of the Arctic, the reanalysis fields are strongly influenced by the underlying forecast model.

10. line 307–309: If/when CSDI is calculated based on CESM's own climatology, there is no reason why a cold bias would lead to a positive CSDI bias. So the negative CSDI bias is note particularly conunter-intuitive.

11. lines 348–357: If the precipitation PDFs are based on daily precipitation rates

regridded to the NE30 grid (as stated), then what explains the systematic change in the PDFs (more frequent very high values) with improving horizontal resolution?

12. As a follow-up comment, if Fig. 7 is based on (e.g.) bilinearly interpolated precipitation data, it would be interesting to see how the PDFs behave if conservative interpolation is used instead.

13. lines 368–369 (also 552–553). Please check if the cited references support your statement regarding extra-tropical storm systems. Based on a cursory reading, they seem to focus on low-latitude systems.

14. lines 458–461: It is not clear what these "average" absolute and relative increases represent. The ranges quoted do not cover the actual range of values in Fig. 12. A simple approach would be to take averages over 60-90°N.

15. line 547–549: A possible physical interpretation of this is that the regions with high present-day CDD are very high-latitude regions with cold winters, in which daily precipitation above 1 mm is relatively rare (but will increase with warming), while the regions with lower present-day CDD are in the southern parts of Arctic and might experience longer dry periods in summer in a warmer climate.

16. line 566: "this narrows the uncertainty range". This requires a more careful reasoning. The definition of storylines in Levine et al. (2024) is based on Arctic land and Barents-Kara Sea warming *normalized* by global-mean warming. Therefore, while the storyline approach is invaluable for impact studies, it is not obvious to me how it reduces the uncertainty related to the overall Arctic warming.

17. Figures 3, 5, 6, 8, S1, S2, S3, S5 and S6: Given that even the renalysis are affected strongly by the underlying model, I suggest to replace root-mean-square error (RMSE) with root-mean-square differene (RMSD), and "Bias" with average difference (AVG or AVGD).

18. Figs. 3 and 6: I suggest to add one more panel showing the difference POLARRES−NE30.

19. Fig. 4: Are these PDFs based on daily-mean data? Please mention that in the caption.

20. Figs. 13 and S15: "... vectors represent the significance". Does this mean that vectors are drawn only where the change is significant (in the third column, they seem to be drawn everywhere!)?

21. Fig. S4: Explain the meaning of stippling.

Language and technical corrections

1. line 307. Replace “decrease in the number” with “smaller number”.

2. line 318: Replace “(TXx)” with “and TXx”.

3. lines 352–353: “A possible explanation for the lower extreme precipitation ... is that extreme precipitation rates are underestimated”. This could be shortened: “It is possible that extreme precipitation rates are underestimated in ERA5 and JRA-3Q”.

4. line 380: “although CESM tends to be slightly drier”. You presumably mean “although the dry bias for CESM appears slightly larger”? The CESM results remain unaltered, only the reference changes . . .

5. lines 414–415 and 430–432. The use of parentheses to shorten sentences often makes the text more difficult to read. It should be avoided especially in cases in which parentheses are also used in their conventional purpose. See

https://eos.org/opinions/parentheses-are-are-not-for-references-and-clarification-saving-space.

6. line 419: “response on temperature” should be “effect on temperature”?

7. line 472: replace “wetter and experience” with “wetter but experience”?

8. line 496: “Fig. 13b-c,e-f”. Double-check that these are the correct figure and panel numbers.

9. line 519: replace “increased cloud cover” with “overestimated cloud cover”.

10. caption of Fig. 1: SIC and OCN are marked with colours rather than diagonal and crossed pattern.

11. Fig. S7: “Same as Figs. S8 and S9”. It is not reader-friendly to refer to later figures in the caption. Move Fig. S7 after S8 and S9?