Dear Christian,

We thank the reviewers for the care they took in assessing our first revision and suggesting further improvements to the solidity of our results and discussion, which we have tried to implement. Here below please find a point-by-point response to the new round of reviews. We are also submitting a version of the manuscript with track changes.

Please note however that we are not sure how to interpret Reviewer 2's discomfort with our results pointing at 20-25 members as a sufficient size for most applications we have considered. His/her point would be justified, we think, only if we were using the bootstrap as the basis for our results. As we also explained in the first round of review, since we had the impression that the reviewer thought our study was based mainly on bootstrapping, we instead only use the bootstrap as a preliminary and limited comparison to our proposed approach, which does not rely on resampling. We have made this argument in our response, but if we are mistaken in interpreting what the reviewer meant we are ready to work on it further. In that case however we need help in understanding the criticism.

Meanwhile, we hope what we have done for this new version is satisfactory. Thank you and the reviewers for your fair handling of our work.

Claudia and co-authors.

PS We left the table colored but they do not need to be, since we were told that cannot be done. We have underlined the values that we want to highlight, and refer to them in the caption as underlined, not colored differently.

Reviewer #1

The authors have incorporated or responded satisfactorily to all my comments on the initial submission. There remain two essential issues to rectify before the paper is ready to publish (corresponding comments are marked with an asterisk):

1) The authors conclusions regarding the number of ensembles needed to detect variance changes are too strong with regard to precipitation.

2) The method for estimating the empirical CDF, from which the blue dots in figure 5&6 are obtained, should be discussed. (I missed this the first round, apologies).

In the interest of time, I think the editor can determine whether the authors sufficiently address the points above, so I do not need to see the second revision, though I am happy to review it should the editor request.
We thank the reviewer for pointing out these additional needs, we have tried to satisfy them carefully. Please see below for details.

Detailed comments:
*Line 15: This implies variance change can be detected with 5-10 members. The results do not support this for the case of Rx5Day (see comments below re lines 278, 343.)
Thank you for pointing this out, we do agree and apologize for this oversight. Also, we agree that histograms could bring out these results more clearly. Therefore, we have reworded the discussion of the results regarding the detection of the variance changes, starting from the abstract, and added histogram plots mirroring the layout of Figure 9 and 10 to the supplementary material.
The abstract now has this additional sentence: "However, the detection of changes in the variance when comparing different times along the simulation requires larger sizes, up to 15 or 20 especially for the precipitation-based metrics." Please also see below for additional modifications regarding this issue.

Line 127: I would change 'responsible for' to 'related to' since the location parameter is not exactly the mean of the tail and variability and tail behavior and not well defined.
Agreed, and changed accordingly.

Line 133: The URL runs off the page.
We rely on the typesetting to fix this type of issue, but we have for now moved the url in a footnote.

Line 232 consistently —> consistent
Corrected, thank you.

Line 276-278 While this is true, the actual size of the confidence intervals will not be known in your scenario of 5 initial runs, as a larger ensemble would be needed.
For this set of results we are using increasing sample sizes, not relying on any estimate derived from the first 5, so the assessment that we are talking about would be achieved by considering the size of the confidence intervals as the ensemble size is increased gradually. Also, in theory, even the width of the confidence interval for the estimate obtained by 5 ensemble members could be considered narrow enough for some type of application not requiring highly precise estimates.

*Line 278 The 'counting' approach should be inaccurate for long return periods using a small number of ensembles. For N-yr events where N is larger than n*11, the CDF

values need to be extrapolated. Consider removing the blue dots (and associated discussion in the text) or only using a single estimate from the full ensemble. If left in, provide a description of how the CDF was estimated.
We consider more appropriate to compare the results obtained by the GEV analysis to that obtained by computing the empirical CDF on the basis of the same number of data points, but we agree with the reviewer that for the only case of the first blue point (n=5, based on 55 data points) in the 100-yr return level plots the empirical result is based on interpolating the value of the CDF between the 54/55~=98th percentile and the 100th, rather than a straightforward count. We have specified that in the caption by adding the sentence "Note that in the 100-yr return level plots the first such dot is obtained by interpolation of the last two values of the CDF, since the sample size is less than 100 (see text).". We have also now added a description of the computation of the empirical CDF in the text of the Methods section. That section of the text now reads: "Lastly, we can use a simple counting approach, based on computing the empirical cumulative distribution function from the same sample, to determine those same X-year events. I.e., after computing the empirical CDF we choose the value that leaves to its right no more than $p*n*11$ data points, where p is the tail probability corresponding to the $1/p$-year return period as defined above."

Line 299 Consider rephrasing "bound to be an upper bound".
☺ _yes, we now say "as that answer serves as an upper bound"._

Line 341 Correct 'a spatially noisier pictures'.
Done, thank you.

*Line 343 The results do not support this conclusion, in my opinion. There seem to be quite a number of locations where n is larger (often much larger) than 5. To quantify this, show the histograms corresponding to the panels in Figures 9 & 10 in the supplementary material.
*Line 398 - 400. Related to the above comment, this statement is far too strong. There are a lot of point where n>5 in the top right panel of Figure 9, corresponding a change in Rx5Day between 1950 and 2100 (the largest time difference). My interpretation is that for rain metrics, you need a minimum of ten ensemble members to be reasonably confident you can detect variance changes in most locations where they occur (around 90%) or so. But this is hard to estimate from the maps, which is why I suggest showing the histograms in the supplement.
We agree with these comments, apologize for the oversight, and we have added such figures in the supplementary material, together with rewording the discussion of these results.
The relevant section now reads: "In the case of TNx, a metric based on daily minimum temperature, the changes are confined to the Arctic region and in most cases the ensemble size required is again 5, with only one instance where the changes between mid-century and end-of the century require consistently a larger ensemble size over an appreciable extent (as

many as 15 members over the region). When we conduct the same analysis on the precipitation metric, shown in Figure~\ref{VarianceChanges_Rx5Day}, we are presented with a spatially noisier picture, with changes in variance scattered throughout the Earth's surface, especially over the oceans. In the case of this precipitation metric the ensemble size required is in many regions as large as 15 or 20 members. These results are made clearer by Figures~\ref{VarianceChanges_TNx_histograms}and~\ref{VarianceChanges_Rx5Day_histograms} in the Appendix, where the grid-boxes where significant changes are present are gathered into histograms (weighted according to the Earth's fraction that the grid-boxes represent) that show the ensemble size required along the x-axis.
We highlight in those figures the fact that for the temperature-based metric only three histograms, corresponding to three specific time-comparisons, gather grid-boxes covering more than 5\% of the Earth's surface, while the coverage is more extensive than 5\% for all time comparisons for the precipitation metric.
These results are representative of the remaining metrics and the alternative model as Figures~\ref{VarianceDiff_TNn} through~\ref{VarianceChanges_Rx5Day_CanESM} in the Appendix document."

Figure 3, 4 9 & 10 are sideways.
We did this by purpose as they are wider than they are tall, to facilitate reading, but we hope to rely on the journal technical help for an optimal setting.

Reviewer #2

Second review of Extreme Metrics from Large Ensembles: Investigating the Effects of Ensemble Size on their Estimates

The authors have largely responded to and made changes to my original comments, I recommend minor revisions.

My main concern is still the claim that 20-25 members is enough given that this is ½ the ensemble size. The argument on lines 10, 391 and 395 seems to be on contradiction with your argument on line 185 and Milinski et al.

We wonder if the reviewer still thinks that we are basing our results on a bootstrap approach, which we are not. All our results are based on estimating quantities/variability/standard errors on the basis of formulas in which we plug in an estimate derived from a 5-member ensemble, or by showing how results change for increasing sample sizes, but without involving bootstrapping (aside from the initial exercise to set the stage). Therefore, none of our result should be undermined by the fact that 20/25 is half the ensemble size, while we agree with the reviewer that if we were using the bootstrap there would be a problem, also in line with Milinski et al.'s argument. If we are *misinterpreting the reviewer's concerns, then we apologize and remain eager to clarify what needs to be clarified.*

Minor comments:

46 – you should cite both Hawkins & Sutton papers:
https://link.springer.com/article/10.1007/s00382-010-0810-6
Done, thank you.

69 – this is not completely true, this is partially addressed in the CanESM2 setup
Noted. We have rephrased as in "We note that sources of variability from different ocean states, particularly at depth, are not systematically sampled by these type of ensembles, albeit they are partially addressed by the design of the CanESM2 that uses de-facto different ocean states."

125 – I am still unclear on what each term in the equation is
This was a standard way to introduce a GEV distribution, so we are not sure how to change it to make it more readable. We have tried to reword the definition slightly. We are not pasting the new wording here due to the mathematical notation that would go awry.

133 – issue with formatting
We have moved the url to a footnote.

251 – should this paragraph be in the methods?
Good point, we have moved it there and reworded a bit the entire section about the GEV methodology in the Methods section.

369 – why do we expect this?
Every quantity emerging at the grid-point of a GCM represents an average quantity, and the average is representative of the size of the grid-box. Therefore, we expect quantities that are representative of larger averages to be less affected by noise. The same way as if we were averaging across grid-boxes, the larger the region considered, the stronger the signal should be, all else being equal.

391 – should read ' we compared;
We changed "We use" to "we compared" hoping this is what the reviewer meant.

417 – should be e.g. Knutti as there are many studies on this topic
Good point, corrected.

Papers that also look at ensemble size that might be worth citing in the introduction. The authors can decide if these are useful.
Thank you for this list, and for taking the time to include the urls!

Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Clim. Dynam., 38, 527–546, 2012
We already cite this first one.

Olonscheck, D. and Notz, D.: Consistently Estimating Internal Climate Variability from Climate Model Simulations, J. Climate, 30, 9555–9573, https://doi.org/10.1175/JCLI-D-16-0428.1, 2017.
This paper does not use initial condition ensembles.

Li, H. and Ilyina, T.: Current and Future Decadal Trends in the Oceanic Carbon Uptake Are Dominated by Internal Variability, Geophys. Res. Lett., 45, 916–925, 2018
Included, thank you.

Steinman, B. A., Frankcombe, L. M., Mann, M. E., Miller, S. K., and England, M. H.: Response to Comment on “Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures”, Science, 350, 1326, https://doi.org/10.1126/science.aac5208, 2015
Included, thank you.

Pausata, F. S. R., Grini, A., Caballero, R., Hannachi, A., and Seland, Ø.: High-latitude volcanic eruptions in the Norwegian Earth System Model: the effect of different initial conditions and of the ensemble size, Tellus B, 11, 2050–2069, 2015
Included, thank you.

Bittner, M., Schmidt, H., Timmreck, C., and Sienz, F.: Using a large ensemble of simulations to assess the Northern Hemisphere stratospheric dynamical response to tropical volcanic eruptions and its uncertainty, Geophys. Res. Lett., 43, 9324–9332, 2016
Included, thank you.

Daron, J. D. and Stainforth, D. A.: On predicting climate under climate change, Environ. Res. Lett., 8, 034021, https://doi.org/10.1088/1748-9326/8/3/034021, 2013
We left this out as it uses simpler models.

Drótos, G., Bódai, T., and Tél, T.: On the importance of the convergence to climate attractors, Eur. Phys. J. Spec. Top., 226, 2031–2038, https://doi.org/10.1140/epjst/e2017-

70045-7, 2017
We left this out due to the theoretical nature of the discussion.

Maher, N., Matei, D., Milinski, S., and Marotzke, J.: ENSO change in climate projections: forced response or internal variability?, Geophys. Res. Lett., 45, 11390–11398, 2018
Included, thank you.

Table 1 – did you try bootstrapping with replacement? Is the colour or the underline important as you refer to both?
We use the bootstrap, without replacement, to confirm the Milinski et al. argument, and to position our alternative method with respect to it. We do not use the bootstrap for anything further, and not for any of our results. We therefore did not try anything but the simple bootstrap approach without replacement, as we do not consider the bootstrap a focus of our study. Sampling with replacement would also introduce unrealistic replicates of the values, making the estimation of the error even more biased low.
We did not mean for "highlighting" to refer to the colors, but rather in the general sense. We changed that to "pointing out" to avoid confusion. We expect the colors to disappear in the journal version, as we were told the journal does not print colored tables.

Figure 3/4 – add % on colourbar
Added.

Figures 5/6 are still really small
Apologies but we cannot figure out a way to expand the size without cutting out the legend. We will work with the journal to try and maximize the quality of the graphics.

Figure 9 does not show much, did you consider zooming in on just the Arctic?
Apologies but we cannot get a projection of the Arctic looking satisfactory with our R graphic package.