We thank both referees for a careful reading of the paper and many suggestions that have improved it substantially.
In the following we provide responses to all points raised by Referee 2.
Please, also note that the paper has been edited throughout for clarity and rigor, and we will provide a version with tracked changes.

In the following, we highlight our responses in blue, with italics used when we cite word by word from the paper (apologies for the occasional latex syntax left throughout these verbatim sections).

Thank you,

Claudia Tebaldi and co-authors.

Referee 2

The paper 'Extreme Metrics and Large Ensembles' considers the problem of ensemble size applied to 6 widely used extreme metrics. It asks whether we can estimate the ensemble size needed using a smaller 5-member ensemble, then validates the answer using two large ensembles. This manuscript provides additional information to the field and presents novel results. However, there are a few issues that need to be addressed prior to publication.

Thank you for an overall positive and constructive review and the many concrete suggestions.

Major comments:

1.Milinski et al 2020 discuss the problem of bootstrapping and how the errors increase as you approach the ensemble size. This is discussed in the manuscript on lines 108 onwards, however it is then largely ignored for the rest of the manuscript. This is a major issue as some of the results approaching the size of the ensemble may be buased due to this problem. Additionally this manuscript determines that the ensemble size needed is 20-25members, which is about hald the ensemble size and where this problem starts affecting the results. This needs to be addressed before this manuscript is published.

We agree wholeheartedly, and that is why we use the bootstrap approach only in the opening of our work, exactly to show how it may be biased, and to quantify such effect. The remaining of our analysis is based on the use of the formula for the determination of the ensemble size given a certain tolerance for error, so we never use the bootstrap again after showing and quantifying in our tables exactly what the reviewer, and the Milinski et al. analysis, pointed out.

2.All Figures are too small, on a printout it is impossible to see what the Figures are. On the screen one needs to magnify the pdf to be able to see anything.

We have redrawn all figures (also added several in the appendix) and we have improved legibility throughout.

3. Details of the statistics used and results are missing, sometimes the text is vague or non-speciifc. See comments below for details.

We apologize for the oversight when failing to describe our statistics. We hope to have addressed all shortcomings, please see below.

In the following we answer to specific comments, please note however that we have extensively edited the paper, in order to make the discussion clearer and more detailed and add caveats throughout according to Reviewer 1's and 2's suggestions.

Minor issues:

The title could be more descriptive as currently one has no idea the paper is about ensemble size from the title.

Thank you for this good observations. We have changed the title to: Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates.

line 20 'like' should be replaced with 'such as'

Corrected

line 40 I think your citation of CanESM2 large ensemble is wrong: https://open.canada.ca/data/en/dataset/aa7b6823-fd1e-49ff-a6fb-68076a4a477c

Thank you, we now use Kirchmeier-Young et al. 2017 and Kushner et al. 2018 as the webpage suggests.

lines 59-62 there are only citations for CESM not CanESM here

We added Arora et al., 2011 for CanESM2 which is used as a reference for CanESM2 in Kushner et al. 2018 and Kirchmeier-Young et al. 2017

line 63 - I don't think this is the correct initialisation procedure for CanESM2 see https://open.canada.ca/data/en/dataset/aa7b6823-fd1e-49ff-a6fb-68076a4a477c

Thank you, we have added that 5 different historical simulations were used for CanESM2 ensemble initialization.

line 65 - you could cite Marotzke, 2019 here: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wcc.563

Thank you, we do now.

Line 70 - would it make sense to show the results for the larger ensemble in the main paper?

We apologize for pushing back on this, which would mean a complete overhaul of the paper. In addition, by using the CESM ensemble we are choosing to show the results for a model whose resolution is closer to state-of the art ESMs at this point in time. This we hope would make the results more relatable, also considering the nature of CESM as a community model.

Equations on page 4 - you need to define what each term in the equation is

We have added mention of mean, scale and shape and their meaning, now, also in response to a similar note by Reviewer 1, as

"[...] *represent the location, scale and shape parameter respectively, responsible for the mean, variability and tail behavior of the random quantity z."*

line 126 - please define X

We have rewritten this sentence as in:

"On the basis of the GEV parameters estimated for a range of ensemble sizes n up to the full size available we compute return levels for several return periods X, X=2,5,10,20,50,100 and their uncertainty and assess when the estimates of the central value converge and what the trade-off is between sample size and width of the confidence interval."

Section 3.2 what tests do you use to detect changes in variance?

We use F-tests and we now have added that information which we forgot to specify because of an oversight.

Figure 1 - titles on the subplots could be more descriptive

We have changed them accordingly.

Figure 2 – ofthe should be of the

Corrected, thank you

Figure 2 - I don't fully understand what the diagonal line is. Is it the actual time evolution of the expected error?

It was the linear trend of the time evolution of sigma/sqrt(n). We have decided to erase this line, also according to Reviewer 1's criticism, since it seemed to create only confusion rather than add information.

[TODO]

Line 211 - do you have a hypothesis why the error exceeds the expected error in these regions?

We show in our table that the fraction of the global (or ocean, or land) areas that show an error exceeding the 95% bound is in fact consistent with 5%, so we deem this behavior consistent with the distribution of the mean component that we estimate through our sigma/sqrt(n) computation. However we have added a short point that Reviewer 1 suggested about the possible effect of autocorrelation in the samples used to estimate sigma:

"The prevalence of red areas over the oceans could be due to an underestimation of σ_i^c linked to the use of the 5-year windows and the autocorrelation possibly introduced, consistently with ocean quantities having more memory than land quantities, but we do not explore that further here."

Line 215 - be specific, for which variabiles are you talking about?

We added "*for all metrics considered in our analysis*"

Figure 5 - how do you calculate 95% confidence

We have now specified that we use the standard maximum likelihood approach to the computation of the CIs.

Line 252 - more detail please

We have added an explicit reference to the confidence intervals that are computable by the GEV approach, that is what we meant here. The sentence now reads:

"As for the results of the empirical counting approach, i.e., the blue point estimates, we can assess that in the majority of cases, but not across the board when we look closely to all the plots in the appendix, they do not deviate significantly from the central estimates based on fitting the GEV using the same sample size. However, while the latter can provide a measure of uncertainty through confidence intervals, the estimates based on counting events do not come with uncertainty bounds."

Line 275 - this is really an odd sentance cna you rephrase?

We have rephrased as in

"For all times considered, 5 members are sufficient to estimate an ensemble variance indistinguishable, statistically, from that which would be estimated using the full ensemble at most grid-points over the Earth's surface, as the light blue color indicates. For some sparse locations, however, 10 members are needed to achieve the same type of accuracy."

Line 281 - more detail please

The sentence now reads:

"We note here that the patterns shown in some of these figures have indeed the characteristics of noise. To minimize that possibility we have applied a threshold for the significance of the p-values from the F-test obtained through the method that controls the False Discovery Rate (Ventura et al., 2004). The method has been shown to control for the false identification of significant differences "by chance" due to repeating statistical tests hundreds or thousands of times, as in our situation. The same method has been proved effective, in particular for multiple testing over spatial fields, despite the presence of spatial correlation (Ventura et al., 2004; Wilks, 2016). We fix the false discovery rate to 5%. "

Figure 9 - could you zoom in on the Arctic as this is the only place there are colours, it is otherwise small, and seems to have limited information in the plot

We have changed the plots to show the Northern Hemisphere only for hot extremes and the Southern only for cold extremes (in the appendix too), thank you for the suggestion.

Line 294 - how do you detect changes?

We use an F-test comparing variance computed at the different times during the simulation. The F-test is now mentioned explicitly, apologies for the oversight.

Line 300 - missing specific details here

Line 307 - it is not clear how you calculate significance

In response to both these comments we have added a sentence that should clarify our analysis of significant changes in variance:

"Here as in the previous analysis a significant change is detected when the F-test for the ratio of the two variances that are being compared across time has a p-value smaller than the threshold determined by applying the false discovery rate method, and fixing the false discovery rate to 5%."

Section 4.4 - this is short and has limited detail in it. S/N really is dependent on the quantity due to the size of each term. I feel like very little is actually said in this section. Maybe think about what the point you want to make here is?

We have attempted to make this section more representative of what we consider some interesting differences in metrics which we had definitely overlooked in the previous version. We decided to focus on the mid-century results as the end-of century would be too boring, and we focused on the requirements for S_N=2. We hope the reviewer will find the new section improved and worth keeping.

5. Conclusions - perhaps mention recent regional large ensembles here such as: https://www.climex-project.org/

Thank you, done

Line 358 - Marotzke et al, 2019 and Hawkins et al, 2016 https://link.springer.com/article/10.1007/s00382-015-2806-8

Thank you, added.