Earth System Dynamics Discussions

Earth Syst. Dynam. Discuss., referee comment RC1
https://doi.org/10.5194/esd-2021-53-RC1, 2021

Comment on esd-2021-53

Anonymous Referee #1

Referee comment on "Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates" by Claudia Tebaldi et al., Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2021-53-RC1, 2021

The authors address the problem of choosing the size of a climate model ensemble by investigating the number of members needed to obtain robust statistics for the warmest night (TNx) and wettest pentad (Rx5d) of the year. The main text presents results based on the CESM1-CAM5 38/40 member ensemble (CanESM2 48/50 member results are shown in the supplement). They investigate the accuracy, compared to the full ensemble, of variety of statistics obtained based on 1,5,10,…,35 ensemble members, namely: 1) the forced component (ensemble mean) and its uncertainty, 2) generalized extreme value (GEV) distribution fits, and 3) the internal variability (ensemble variance) at various time points, as well as detecting changes over time. They also briefly touch on the number of members needed to detect a signal to noise ratio greater than 1 for TNx and Rx5d by mid and end century. They argue that often, an ensemble of 20-25 members provides reliable statistical estimates. They also show that the full ensemble standard deviation (and error estimates that rely on it) can be accurately estimated using 5 ensemble members, supplemented by using 5 years along the time dimension of each member.

The study is of practical interest for optimizing computing resources and for potentially allowing broader investigations into structural/parameter uncertainty. While the general question of determining ensemble size has been addressed in previous studies, the analysis using yearly extremes is a new contribution as far as I'm aware. However, there are major issues that need to be addressed before the manuscript could be considered suitable to publish.

Major comments:

- There are several places where details of the methods are not provided (see specific comments below), particularly in Secion 4.3 where the statistical tests used to determine the accuracy of the ensemble estimates of variability and for detecting the

change in variance over time are not discussed. As such, I cannot yet comment on the validity of the methods for this section.

- The bootstrap estimates are all dubious for n > 20 due to oversampling. This is mentioned in the manuscript and was addressed in some detail in Milinski et al., 2020. They are therefore not suitable for comparison in the Tables, and misleading in Figures. I strongly suggest removing any bootstrap results for n>20. Alternatively, you can visually highlight the dubious numbers in the Tables and provide a caveat in the captions.

As a side note, I also have some doubts about number 20-25 being sufficient to provide reliable estimates, as that number constitutes a majority of the ensemble members and therefore may be biased low. I cannot point to any theory to confirm or deny my suspicion however. Repeating the analysis with an ensemble that has more members would increase the confidence in them, perhaps this can be done in another study.

3) When supplementing the estimate using 5 ensemble members with information along the time dimension, one runs into the problem of serial correlation, which can reducing the effective sample size. So 5 members x 5 years results is a sample size of 25 at most. Depending on the variable of interest it can be substantially less. I am fairly confident that this is the reason that more ensembles are needed to estimate TNx variance over the ocean for example (Fig. 3). A caveat regarding this issue should be included in the text.

Specific comments:

All the figures are too small.

Line 43: Surface temperature is one of the strongest climate signals generally and with regard to extreme metrics. So an ensemble number based on analyzing TS is functionally a minimum value. Precipitation will also be a strong signal, especially averaging over larger regions where the dynamic changes might cancel out, leaving only the increase due to thermodynamics. So changes in precipitation extremes are also not difficult to detect. You later address this by looking a specific times when the changes are small, which I think is worth mentioning here.

Line 65-66: This is an overly strong statement. Model variability on increasingly long time scales (and hence increasing ocean depths) is indeed not included by design. But how do you justify saying these GCMs do not represent ocean variability?

Line 68, elsewhere: Generally, you should avoid contractions in scientific papers

Line 93: Some brief info on the location, shape, and scale parameters would be helpful. If nothing else just name them here, which would help the reader's intuition.

Line 124: A symmetric sample around 1953, 2097 would result in 7 years, not 11.

Fig 1: The full ensemble line looks white, not blue as in the legend. Only plot up to n=25, to avoid issues with oversampling.

Line 168: Since the full ensemble statistics are being treated as the true statistics, the F values are the true RMSE values. So the results show that the bootstrap method is not "optimistic" but inaccurate, as mentioned in line 109 and Milinski, 2020.

Line 167-168: Need to consider autocorrelation. Are the F-5 values bootstrapped, or is only a single 5 member subset used?

Table 1: Show percentages for more precision.Remove or visual highlight the spurious bootstrap numbers. Comparing the n=5 results to the bootstrap result that are known to be spurious is not useful. Consider putting the "true" full ensemble number on the left.

Lines 180-184: It seems the more general result that 5 ensemble members are sufficient to obtain an accurate measure the true instantaneous (i.e. changing over time) model internal variability at global scales.

Line 186: It's not clear what you mean by "RMSE affecting estimates" or "our estimate of σ"; What about starting with: "Thus, compared to a single model run, we would expect …."

Line 191: Why twice the expected RMSE? The 2 sigma level is roughly the 95% confidence interval of the expected error, so you would expect to exceed that level 5% of the time. You cannot conclude from these results that "the actual error is in most cases much smaller than the expected…"

Line 205: "We use the full ensemble or only 5 members to estimate the ensemble standard deviation…" You specify two definitions here and then never clarify which one you use.

Figure 2: Why is 2-sigma the cutoff when the standard error is sigma/sqrt(n)?

The diagonal lines are spurious; they show a large widening of the variability that is absent in Fig 1 (no change in the width of the lines in a-d.) Table 1 also indicates that TNx variability increase by maybe 50%, not a factor of 3 as shown here. Also the variability may not change monotonically with time.

Line 210: "Small and sparse" over land perhaps, but certainly not over the ocean due to the longer timescales of variability which necessitate more samples.

Line 212: "consistently providing a conservative estimate of it according to normal distribution theory": 1) I don't think that conservative is the goal, but accuracy, and 2) this is dependent on using the 2-sigma threshold, which was not justified.

Fig 3: The 2 sigma level is not an "upper bound", but an estimate of the 95% level.

Fig 3&4: From Fig. 1, it seems safe to assume a normal distribution for the ensemble range of the global average of the block maxima (TNx, Rx5Day). But is that true at each grid point. Don't we expect them to be follow the GEV distribution?

Line 219; The goal is to find the number of ensembles, n, that can be used to accurately "decide how large an ensemble we need in order to approximate the forced component to a given degree of accuracy." As I read it, that amounts to finding the a threshold for n, such that the ensemble standard deviation sigma_n accurate measure the true sigma. So rather that comparing to a threshold of 2 sigma (the rationale for which is never given) you should compare the error with respect to the true sigma, including the confidence intervals of both the true value and the estimate. The two aspects being how accurate is the estimate, and how narrow is the range of estimate.

Line 220: You do not actual address regional means.

Line 232: Provide a bit more information or reference about temporal covariates for the reader. You can safely assume that the forced climate change signal is not causing temporal autocorrelation, but interannual to decadal internal variability can still induce serial correlation.

Fig 5: Is this based on a bootstrap, or a single selection of N ensemble members? From the NNA plots, I assume the latter, because the N=10 return levels are way off. If that is the case, were the same ensemble members used in all the plots? How are the confidence intervals for the full ensemble (the upper and lower horizontal lines) calculated?

Line 243: You're not bootstrapping here so oversampling is not an issue, but 20-25 is still a majority of the ensemble members, which may artificially improve the results. The exercise would need to be repeated with a larger ensemble to increase confidence in the result. A caveat along these lines should be included.

Line 271: What test is used to determine if they are distinguishable?

Line 283: The Ventura et al, 2004 method addressing spatial covariability, not temporal, correct?

Line 286: What test is used to detect variance changes?

Fig 9,10: Several color bars are mislabelled

Line 294: "minimum temperature" might be confusing; consider using "warmest night" as you do elsewhere

Line 315: Is sigma held constant for this calculation?

Line 326: Only grid points and global averages were shown, not a range of aggregation.

Line 336: Are the variables of interest here normally distributed? Shoudn't they obey an GEV distribution

Line 365: Missing parentheses