How many ensemble members do we need?

Analyzing two IC large ensembles through an 'extreme' lens.

Extreme metrics from large ensembles: investigating the effects of ensemble size on their estimates
Claudia Tebaldi, Kalyn Dorheim, Michael Wehner, Ruby Leung
ESD 2021 https://doi.org/10.5194/esd-12-1427-2021

Motivation

We want to ESTIMATE A PRIORI* WHAT ENSEMBLE SIZE provides a *good*** estimate of both the FORCED COMPONENT and the size of INTERNAL VARIABILITY for several EXTREME METRICS.

* The perspective is that of a modeling center that has a small (5 member) ensemble ON THE BASIS OF WHICH TO DECIDE THE SIZE OF A LARGE ENSEMBLE EXPERIMENT.

** What *good* means depends on the application. We use Root Mean Square Errors (RMSEs) or other error metrics to exemplify the procedure and show results.

Models/Ensembles in the CLIVAR SMILEs repository

https://www.cesm.ucar.edu/projects/community-projects/MMLEA/

- **CESM1 LENS: 40 ensemble members**, 1950-2100 under RCP8.5 ~1degree grid (T96)
- **CanESM2: 50 ensemble members**, 1950-2100 under RCP8.5, ~2degrees grid (T42)

Quantities

- **Six metrics of annual extremes based on daily TASMIN, TASMAX and PR:**

  - ✓ Hottest day of the year: maximum TASMAX over the year -> **TXx**
  - ✓ Hottest night of the year: maximum TASMIN over the year -> **TNx**
  - ✓ Coolest day of the year: minimum TASMAX over the year -> **TXn**
  - ✓ Coolest night of the year: minimum TASMIN over the year -> **TNn**

  - ✓ Wettest day of the year: maximum daily PR amount over the year -> **Rx1Day**
  - ✓ Wettest pentad of the year: maximum mean daily PR amount over the wettest 5 consecutive days of the year -> **Rx5Day**

Spatial scales

- We consider the characterization of forced component and variability at a range of scales, **from global average to grid-point**.

Our Method for the Forced Component

Usually we estimate the forced component by the ensemble mean.

Determine sample size (i.e., ensemble size) needed to control the variance of the estimator (sample mean):

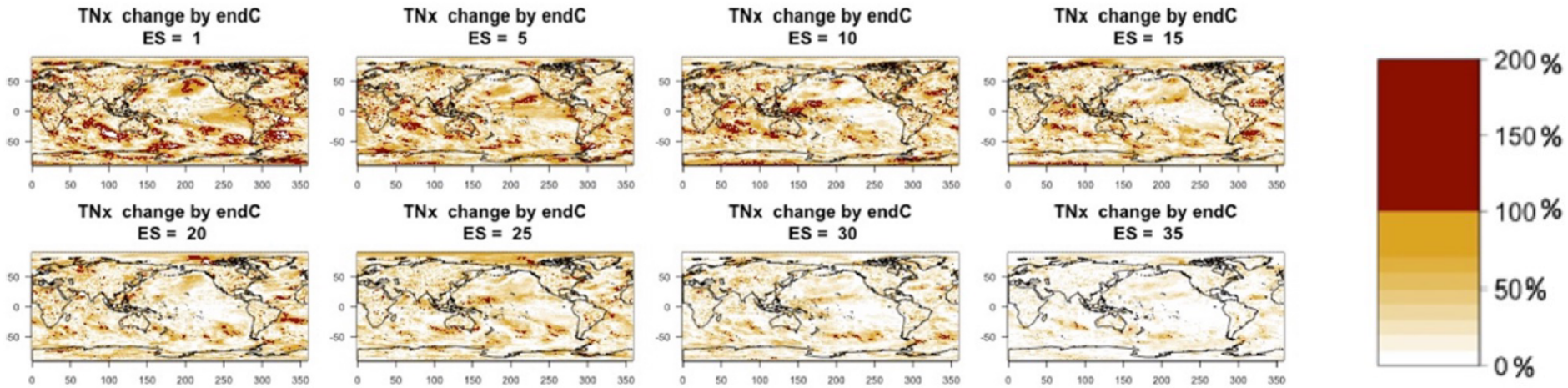
N such that

$$V(\bar{X}_N)=\sigma^2/N < K$$

Where we estimate σ using only 5 members (and note that we can actually get confidence intervals on σ and therefore on N).

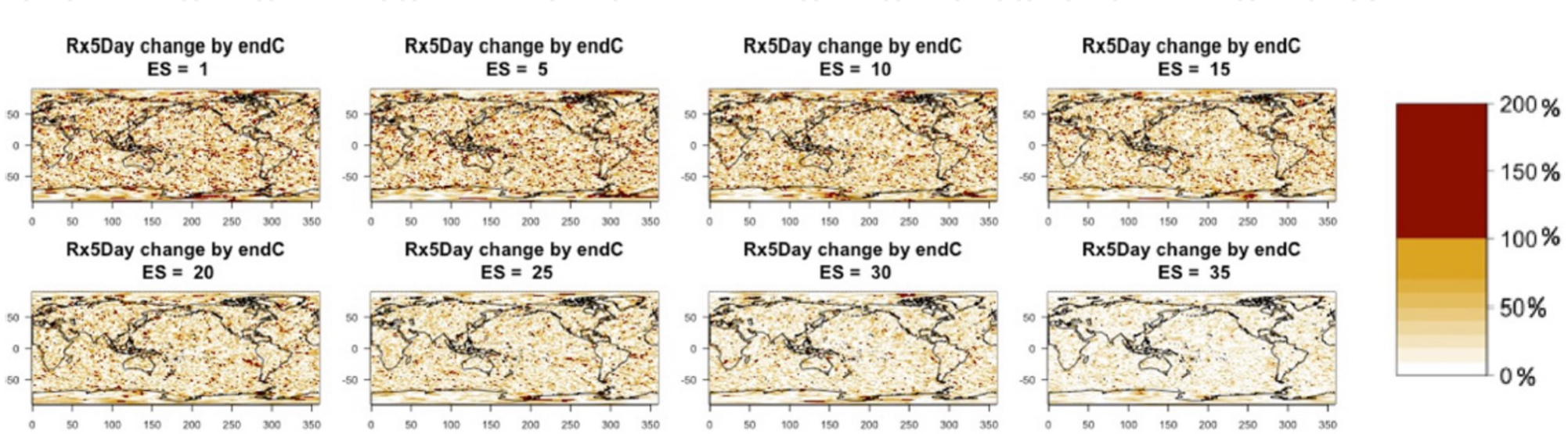
We use the σ computed from the full ensemble as the truth.

Comparison of estimated vs. true error at grid-point scales: Warmest Night



If the color pixels are other than red the error has been correctly estimated (in fact overestimated in most cases) by using the formula and estimating σ using 5 ensemble members.

Comparison of estimated vs. true error at grid-point scales: Wettest Pentad



If the color pixels are other than red the error has been correctly estimated (in fact overestimated in most cases) by using the formula and estimating σ using 5 ensemble members.

Forced Component Results:

We can use a small ensemble of five members to estimate the population variance and plug it into the formula for the variance/standard error of the sample mean as a function of sample size.

Imposing a ceiling for this error allows us then to determine how large an ensemble should be, in order to approximate the forced component to the desired level of accuracy.

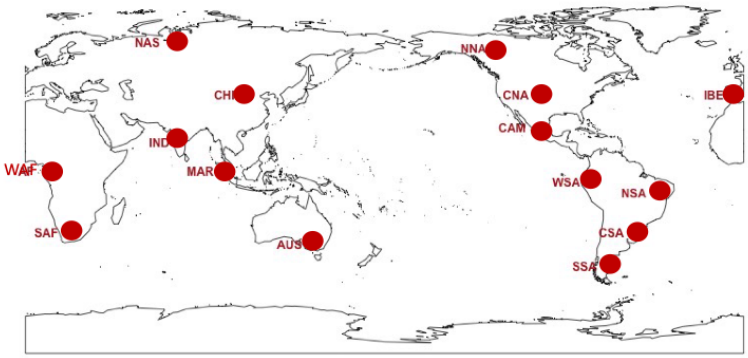
This holds true across the range of spatial scales afforded by these models, from global means all the way to grid-point values.

GEV analysis

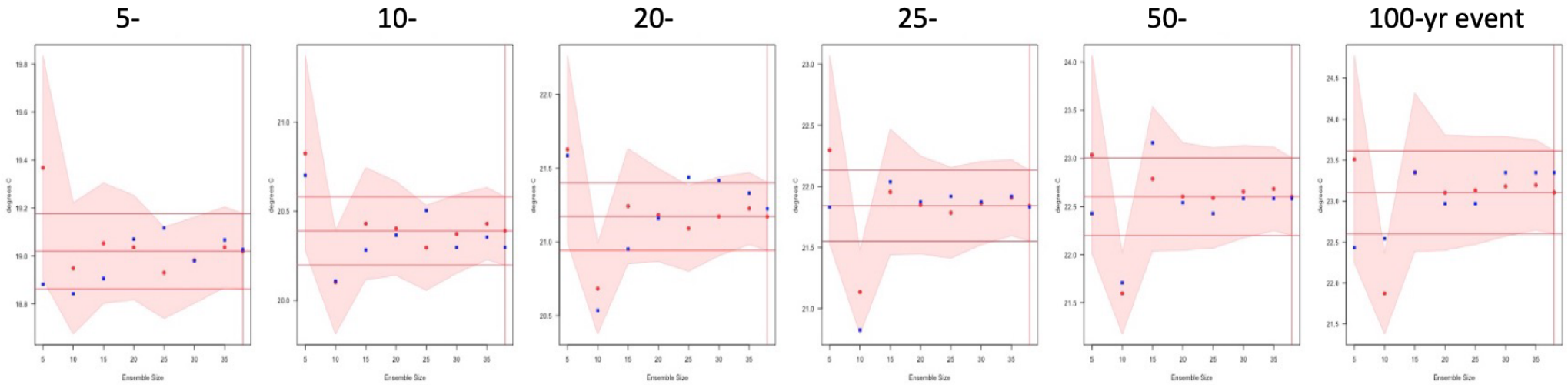
- We also apply Generalized Extreme Value distribution fitting to increasing sample sizes, at the grid-point level and consider how the central estimate and the confidence interval “stabilize” and how simply counting extreme events in the ensemble compares.

- We do this at a number of sites

  representing different climates

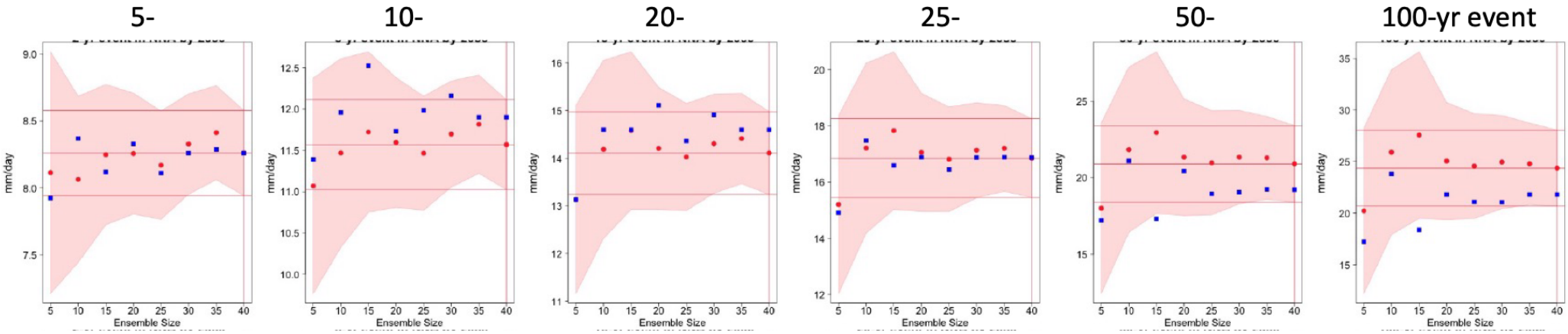


TNx – Warmest Night for Northern North America (2050)



X-axis: Ensemble Size
Y-axis: Event magnitude (Degrees C)
Red Dots/Pink Envelope: GEV analysis
Blue Dots: Empirical estimates

Rx5Day – Wettest Pentad for Northern North America (2050)



X-axis: Ensemble Size
Y-axis: Event magnitude (mm/day)
Red Dots/Pink Envelope: GEV analysis
Blue Dots: Empirical estimates

Results from GEV fitting

- It appears as if after 20 or 25 ensemble members the estimation "stabilizes" both as central estimate and as confidence interval;

- The empirical estimates have the drawback of not providing straightforwardly uncertainty bounds, still they often fall within the confidence intervals when the ensemble size reaches 20 or 25.

Internal Variability analysis

- We switch gears and ask how many ensemble members are needed to fully characterize the size and behavior of internal variability.
- Here we define IV as the standard deviation of the ensemble, i.e. the inter-members variability.
- Analysis conducted at the grid-point level.
- Again, we take the estimates of variability from the full ensemble as our truth.

Between 5 and 10 ensemble members seem sufficient to estimate the size of the intra-ensemble (inter-member) variability at the grid-point scale, for a specific time during the simulation, if a 5-year window is used to augment the sample size

For the problem of detecting a change in variance, it appears that the ensemble size needed is larger, especially for precipitation metrics, but still 20-25 members seem sufficient.

Conclusions

It is possible to **estimate a-priori with good accuracy what ensemble size is needed** in order to estimate signal and noise characteristics of several temperature and precipitation **extreme metrics**.

The analysis considered two different models, two resolutions, a large range of radiative forcing/warming levels, and several metrics of extreme Tmin, Tmax, Precipitation, producing **fairly robust outcomes**.

We find that an ensemble size of **20-25 members** meets our various estimation goals, both for identifying the forced response and for estimating internal variability around it, all the way to the grid-point scale.

The right answer to the question of how many ensemble members are needed depends on what the use of the ensemble is going to be, so it may always turn out to be an **ill-posed question**.

Identifying the size of the ensemble needed would be extremely valuable for designing experiments that **explore other sources of uncertainty besides initial conditions**.