



# **Extreme Metrics and Large Ensembles**

Claudia Tebaldi<sup>1</sup>, Kalyn Dorheim<sup>1</sup>, Michael Wehner<sup>2</sup>, and Ruby Leung<sup>3</sup> <sup>1</sup>Joint Global Change Research Institute, Pacific Northwest National Laboratory, College Park, MD, USA <sup>2</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA <sup>3</sup>Pacific Northwest National Laboratory, Richland, WA, USA Correspondence: Claudia Tebaldi (claudia.tebaldi@pnnl.gov)

Abstract. We consider the problem of estimating the ensemble sizes required to characterize the forced component and the internal variability of a range of extreme metrics. While we exploit existing large ensembles contributed to the CLIVAR Large Ensemble Project, our perspective is that of a modeling center wanting to estimate a-priori such sizes on the basis of an existing small ensemble (we use five members here). We therefore ask if such small-size ensemble is sufficient to estimate the population

- variance in a way accurate enough to apply a well established formula that quantifies the expected error as a function of n (the 5 ensemble size). We find that indeed we can anticipate errors in the estimation of the forced component for temperature and precipitation extreme metrics as a function of n by applying the population variance derived by five members in the formula. For a range of spatial and temporal scales, forcing levels (we use RCP8.5 simulations), and both models considered here as our proof of concept, CESM1-CAM5 and CanESM2, it appears that an ensemble size of 20 or 25 members can provide estimates of
- 10 the forced component for the extreme metrics considered that remain within small absolute and percentage errors. Additional members beyond 20 or 25 add only marginal precision to the estimate, which remains true when extreme value analysis is used. We then ask about the ensemble size required to estimate the ensemble variance (a measure of internal variability) along the length of the simulation, and – importantly – about the ensemble size required to detect significant changes in such variance along the simulation with increased external forcings. When an F-test is applied to the ratio of the variances in question, one
- 15 estimated on the basis of only 5 or 10 ensemble members, one estimated using the full ensemble (up to 50 members in our study) we do not obtain significant results even when the analysis is conducted at the grid-point scale. While we recognize that there will always exist applications and metric definitions requiring larger statistical power and therefore ensemble sizes, our results suggest that for a wide range of analysis targets and scales an effective estimate of both forced component and internal variability can be achieved with sizes below 30 members. This invites consideration of the possibility of exploring additional sources of uncertainty, like physics parameter settings, when designing ensemble simulations.
- 20

25

Keywords. Large Ensembles, Extreme Metrics, Internal Variability, Forced Signal Detection, ETCCDI

# 1 Introduction

Recently much attention and resources have been dedicated to running and analyzing large ensembles of climate model simulations under perturbed initial conditions (e.g., Deser et al., 2012, 2020; Lehner et al., 2020; Maher et al., 2021a, b). Both detecting the forced component in externally forced experiments, and quantifying the role of internal variability are being





facilitated by the availability of these large ensembles. Many variables and metrics of model output have been analyzed, with large ensembles allowing precise estimates of their current and future statistics. Large ensembles are also being used to answer methodological questions, particularly about the precision these experiments can confer to the estimate of those variables and metrics, and how that varies with increasing ensemble sizes (e.g., Milinski et al., 2020). Recent efforts by multiple modeling centers to coordinate these experiments so that they can be comparable (by being run under the same scenarios of future).

30

ing centers to coordinate these experiments so that they can be comparable (by being run under the same scenarios of future greenhouse gas emissions) allow answering those questions robustly, accounting for the size and behavior over time of internal variability, which is known to be a model-specific characteristic (Deser et al., 2020).

In this methodological study we adopt the point of view of a modeling center deciding on the size of a large ensemble, when interested in estimating current and future behavior of several metrics of extremes. Such decision needs to be reached on the

- 35 basis of a limited number of initial condition members, which the center would run as a standard experiment. We start from a size of five, which is a fairly common, middle-of-the-road, choice for future projection experiments, and use the statistics we derive on the basis of such small ensemble to estimate the optimal size of an ensemble, according to standards of performance that we will specify. We test our estimate of the optimal size by using 'the truth' that we find in two large ensembles available through the CLIVAR SMILES initiative (Lehner et al., 2020), the CESM1-CAM5 LENS (of 40 ensemble members, (Kay et al.,
- 40 2015)) and the CanESM2 ensemble (of 50 members, (Arora et al., 2011)), both run over the historical period and RCP8.5 under the CMIP5 protocol (Riahi et al., 2011; Taylor et al., 2012).

Our metrics of interest are indices describing the tail behavior of daily temperature and precipitation. We conduct the analysis in parallel for extremes of temperature and precipitation because we expect our answers to be dependent on the signal-to-noise ratio affecting these two variables, which we know to be different in both space and time (Hawkins and Sutton, 2009; Lehner

45 et al., 2020). The consideration of two models, two atmospheric quantities and several extreme metrics for each help our conclusions to be robust and – we hope – applicable beyond the specifics of our study.

We consider the goal of identifying the forced change over the course of the  $21^{st}$  century in the extremes behavior. We seek an answer in terms of the ensemble size for which we expect the estimate of the forced component to approximate the truth within a given tolerance, or for which our estimate does not change significantly with additional ensemble members.

- 50 We also consider the complementary problem of identifying the ensemble size that fully characterizes the variability around the forced component. After all, considering future changes in extremes usually has salience for impact risk analysis, and any risk-oriented framework will be better served by characterizing both the expected outcomes (i.e. the central estimates) and the uncertainties surrounding them. Both types of questions can be formulated at a wide range of geographic scales, as the information that climate model experiments provide is used for evaluation of hazards at local scales, for assessment of risk and
- 55 adaptation options, all the way to global-scales of aggregation, usually most relevant for mitigation policies. The time horizon of interest may vary as well. Therefore we present results from grid-point scales all the way to global scale, and for mid-century and late-century projections, specific years or decades along the simulations, or whole century-long trajectories.





### 2 Models, Experiments and Metrics

The CESM1-CAM5 LENS (CESM ensemble from now on) has been the object of significant interest and many published 60 studies, as the more than 1,300 citations of Kay et al. (2015) testify to, and, if in lesser measure, so has been the CanESM2 ensemble (CanESM ensemble from now on). The CESM model has a resolution of about 1 degree in the longitude-latitude dimensions (Hurrell et al., 2013), while CanESM has a coarser resolution of about 2 degrees. Both have been run by perturbing the atmospheric state at a certain date of the historical simulation by applying "errors" of the order of magnitude of machine precision. These perturbations have been found to generate alternative system trajectories that spread out losing memory in

the atmosphere of the respective initial conditions within a few days of simulation time. We note that sources of variability 65 from the ocean, particularly at depth, are not well represented in these experiments. For CESM, 38 or 40 ensemble members (depending on the variable considered) are available, covering the period between 1920 and 2100, while CanESM only starts from 1950 but has 48 or 50 ensemble members. In the following we won't distinguish precisely between the full size or the full size minus two, as the results are not influenced by this small difference. Both models were run under historical and RCP8.5

external forcing, the latter applied starting at 2006. In our analysis we focus first on results from the CESM ensemble, and use 70 CanESM to confirm the robustness of our results. For consistency, we use the period 1950-2100 for both ensembles.

We use daily output of minimum and maximum temperature at the surface (TASMIN and TASMAX) and average precipitation (PR) and compute a number of extreme metrics, all of them part of the ETCCDI suite of indices (Alexander, 2016). All the metrics result in annual statistics descriptive of daily output. They are:

- 75 - **TXx**: highest value over the year of daily maximum temperature (interpretable as the warmest day of the year);
  - TXn: lowest value over the year of daily maximum temperature (interpretable as the coldest day of the year);
  - TNx: highest value over the year of daily minimum temperature (interpretable as the warmest night of the year);
  - **TNn**: lowest value over the year of daily minimum temperature (interpretable as the coldest night of the year);
  - **Rx1Day**: precipitation amount falling on the wettest day of the year;
- 80 - **Rx5Day**: average daily amount of precipitation during the wettest five consecutive days (i.e., the wettest pentad) of the year.

We choose these indices as they reflect diverse aspects of daily extremes, but also because of a technical matter: their definitions all result in the identification of what statistical theory of extreme values calls "block maxima" or "block minima" (here the block is composed of the 365 days of the year). The same theory establishes that quantities so defined lend themselves

85

to be fitted by the Generalized Extreme Value distribution (GEVs) (Coles, 2001). GEV fitting allows us to apply the power of inferential statistics, through which we can estimate return levels for any given period (e.g., the 20-, 50- or 100-year events), and their confidence intervals. We will be looking at how these statistics - i.e., tail inference by a statistical approach that was intended specifically for data poor problems - change with the number of data points at our disposal, varying with ensemble





size, and asking if the statistical approach buys any power with respect to the simple "counting" of events across the ensemble 90 realizations.

If a random variable z (say the temperature of the hottest day of the year, TXx) is distributed according to a GEV distribution, its distribution function has the form:

$$G(z) = \exp\left\{-\left[1-\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},\label{eq:G}$$

defined on  $\{z: 1 + \xi(z-\mu)/\sigma > 0\}$ , where  $-\infty < \mu < \infty$ ,  $\sigma > 0$  and  $-\infty < \xi < \infty$ .

We estimate the parameters of the GEVs by maximum likelihood. If p (say p = 0.01) is the tail probability to the right of 95 level z under the GEV probability density function,  $z_p$  is said to be the return level associated to the 1/p-year return period (100-yr return period in this example), and is given by:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - \{ -\log(1-p) \}^{-\xi} ], & \text{for } \xi \neq 0 \\ \mu - \sigma \log\{ -\log(1-p) \}, & \text{for } \xi = 0. \end{cases}$$

Thus  $z_p$  in our example represents the temperature in the hottest day of the year expected to occur only once every 100 year (in a stationary climate) or with 0.01 probability every year (a definition more appropriate in the case of a transient climate). We use the R package extRemes available from https://CRAN.R-project.org/package=extRemes to fit GEVs and determine return levels and confidence intervals.

100

### 3 Methods

# 3.1 Identifying the forced component

Milinski et al. (2020), use the ensemble mean computed on the basis of the full ensemble as a proxy for the true forced signal, and analyze how its approximation gains in precision by using an increasingly larger ensemble size. By a bootstrap approach, subsets of the full ensemble of a given size n are sampled (without replacement) multiple times (in our analysis we use 100 times), their mean (for the metric of interest) is computed, and the multiple replications of this mean are used to compute the Root Mean Square Error (RMSEs) with respect to the full ensemble mean, assumed to be the true forced signal. Note that this bootstrap approach at estimating errors is expected to become less and less accurate as n increases, as was also noted in Milinski et al. (2020). For n approaching the size of the full ensemble, the repeated sampling from a

- 110 finite population introduces increasingly stronger dependencies among the samples, which share larger and larger numbers of members, therefore underestimating RMSE(n). More problematically, this approach would not be possible if we did not have a full-ensemble to exploit, and if our model was thought of having different characteristics in variability than the models for which large ensembles are available. As a more realistic approach, therefore, we assume that only 5 ensemble members are available, and test how our estimate of the forced signal and the expected RMSE(n) (as *n* increases) may differ. We base our
- 115 expectation on the statistics we can gather from the 5 members and we compare them to the "truth" that the availability of a





large ensemble provides. It is a well known result of descriptive statistics that the standard error of the sample mean around the true mean decreases as a function of n as in  $\sigma/\sqrt{(n)}$ . Here  $\sigma$  is the true standard deviation of the population of ensemble members and we are left with estimating it on the basis of five of them. (This result has in fact been applied in Wehner (2000) to estimate the sampling size in ensemble simulation well before the advent of Large Ensembles.) We will compare the estimated RMSE(n) for  $n \ge 5$  with the actual departure of the mean computed by averaging an *n*-size ensemble from the "true" mean computed on the basis of all available members.

Since we are considering extreme metrics that can be modelled by a GEV, we also derive return levels for given multi-year periods (e.g., 10-, 50-, 100-year events) Because of the availability of multiple ensemble members we can choose a narrow window along the simulations (we choose 11 years, short enough to satisfy the requirement of stationarity that the GEV fit postulates) centered around several dates along the  $20^{\text{th}}$  and  $21^{\text{st}}$  century, i.e., 1953, 2000, 2050, 2097 (the first and last chosen to allow extracting a symmetric window at the beginning and end of the simulations). On the basis of the GEV parameters we compute X = 1/p-year events (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. Lastly, we can use a simple counting approach to determine those same X-year events from all available ensemble members, and compare those estimates

130 to the ones derived by the GEV. The comparison will test if fitting the GEV allows any saving (in terms of sample size) to

120

### 3.2 Characterizing internal variability

Recognizing the importance of characterizing variability besides the signal of change, we ask how many ensemble members are required to fully characterize the size of internal variability and its possible changes over the course of the simulation due to increasing anthropogenic forcing. Process based studies are suited to tackle the question of how and why changes in internal variability manifest themselves in transient scenarios (Huntingford et al., 2013), while here we simply describe the behavior of a straightforward metric, the within-ensemble standard deviation. We look at this quantity at the grid-point scale and we investigate how many ensemble members are needed to robustly characterize the full ensemble behavior, which here again

we assume to be representative of the truth, i.e., the true variability of the system. This translates into two separate questions.

achieve an accurate estimate of the same event obtained on the basis of the full ensemble.

140 First, for a number of dates along the simulation spanning the 20<sup>th</sup> and 21<sup>st</sup> centuries, we ask how many ensemble members are needed to estimate an ensemble variance that is not statistically significantly different from that estimated on the basis of the full ensemble. Second, we first detect changes in variance between all possible pairs of these dates on the basis of the full ensemble, and we then ask how many ensemble members are needed to detect the same changes.

## 4 Results

145 In the following presentation of our main findings we choose two representative metrics, TNx (warmest night of the year) and Rx5Day (average rainfall amount during the 5 wettest days of the year) using the 40-member ensemble from CESM1-CAM5. In the supplementary material we include the same type of results for the additional metrics considered and the CanESM 50-





member ensemble. We will discuss if and when the results presented in this section differ from those in the supplementary material.

# 150 4.1 Identifying the forced component

We start from time series of annual values of globally averaged TNx and Rx5Day (Figure 1, top panels). We compute them for each ensemble member separately, and average them over n ensemble members as the ensemble size n increases, applying the bootstrap approach and computing RMSE(n) (see Section 3.1) at every year along the simulation.



**Figure 1.** Time series for TNx (warmest night of the year, left) and Rx5Day (average daily amount during the 5 consecutive wettest days of the year, right) showing how the estimate of the forced component of their global mean trajectories over the period 1950-2100 changes when averaging an ensemble of increasing size. Top row shows the entire time series. Middle row zooms into the relatively flatter period of 1950-2000, so that the y-axis range allows a clearer assessment of the relative size of the uncertainty ranges for different sizes of the ensembles. Bottom row plots the bootstrapped RMSE for every year and each ensemble size.

155

As Figure 1 indicates, for both quantities the marginal effect of increasing the ensemble size by 5 members is not constant but rather decreases as the ensemble size increases. This is qualitatively visible in the evolution of the ranges in the panels of the first two rows, and is measured along the y-axis of the plots along the bottom row, where RMSE(n) for increasing *n* is shown (each *n* corresponding to a different color).





	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.10	0.10 (0.08,0.13)	0.09 (0.07,0.17)	0.14	0.14 (0.11,0.18)	0.12 (0.09,0.17)
n=5	0.04	0.05 (0.04,0.06)	0.04 (0.03,0.06)	0.05	0.06 (0.05,0.08)	0.05 (0.04,0.08)
n=10	0.03	0.03 (0.03,0.04)	0.03 (0.02,0.04)	0.04	0.04 (0.04,0.06)	0.04 (0.03,0.05)
n=15	0.02	0.03 (0.02,0.03)	0.02 (0.02,0.03)	0.03	0.04 (0.03,0.05)	0.03 (0.02,0.04)
n=20	0.02	0.02 (0.02,0.03)	0.02 (0.02,0.03)	0.02	0.03 (0.03,0.04)	0.03 (0.02,0.04)
n=25	0.01	0.02 (0.02,0.03)	0.02 (0.01,0.02)	0.02	0.03 (0.02,0.04)	0.02 (0.02,0.03)
n=30	0.01	0.02 (0.02,0.02)	0.02 (0.01,0.02)	0.01	0.03 (0.02,0.03)	0.02 (0.02,0.03)
n=35	0	0.02 (0.01,0.02)	0.02 (0.01,0.02)	0.01	0.02 (0.02,0.03)	0.02 (0.02,0.03)
	2050 (B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050 (B) 0.11	2050 (F) 0.12 (0.09,0.15)	2050 (F-5) 0.15 (0.12,0.21)	2097 (B) 0.15	2097 (F) 0.15 (0.12,0.19)	2097 (F-5) 0.13 (0.10,0.17)
n=1 n=5	2050 (B) 0.11 0.05	2050 (F) 0.12 (0.09,0.15) 0.05 (0.04,0.07)	2050 (F-5) 0.15 (0.12,0.21) 0.07 (0.05,0.09)	2097 (B) 0.15 0.06	2097 (F) 0.15 (0.12,0.19) 0.07 (0.05,0.09)	2097 (F-5) 0.13 (0.10,0.17) 0.06 (0.04,0.08)
n=1 n=5 n=10	2050 (B) 0.11 0.05 0.03	2050 (F) 0.12 (0.09,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05)	2050 (F-5) 0.15 (0.12,0.21) 0.07 (0.05,0.09) 0.05 (0.04,0.07)	2097 (B) 0.15 0.06 0.04	2097 (F) 0.15 (0.12,0.19) 0.07 (0.05,0.09) 0.05 (0.04,0.06)	2097 (F-5) 0.13 (0.10,0.17) 0.06 (0.04,0.08) 0.04 (0.03,0.06)
n=1 n=5 n=10 n=15	2050 (B) 0.11 0.05 0.03 0.02	2050 (F) 0.12 (0.09,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.02,0.04)	2050 (F-5) 0.15 (0.12,0.21) 0.07 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.05)	2097 (B) 0.15 0.06 0.04 0.03	2097 (F) 0.15 (0.12,0.19) 0.07 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05)	2097 (F-5) 0.13 (0.10,0.17) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20	2050 (B) 0.11 0.05 0.03 0.02 0.02	2050 (F) 0.12 (0.09,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.03)	2050 (F-5) 0.15 (0.12,0.21) 0.07 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.05)	2097 (B) 0.15 0.06 0.04 0.03 0.03	2097 (F) 0.15 (0.12,0.19) 0.07 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04)	2097 (F-5) 0.13 (0.10,0.17) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25	2050 (B) 0.11 0.05 0.03 0.02 0.02 0.01	2050 (F) 0.12 (0.09,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.15 (0.12,0.21) 0.07 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.05) 0.03 (0.02,0.04)	2097 (B) 0.15 0.06 0.04 0.03 0.03 0.02	2097 (F) 0.15 (0.12,0.19) 0.07 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04)	2097 (F-5) 0.13 (0.10,0.17) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.03)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050 (B) 0.11 0.05 0.03 0.02 0.02 0.01 0.01	2050 (F) 0.12 (0.09,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.15 (0.12,0.21) 0.07 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)	2097 (B) 0.15 0.06 0.04 0.03 0.03 0.02 0.01	2097 (F) 0.15 (0.12,0.19) 0.07 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04) 0.03 (0.02,0.04)	2097 (F-5) 0.13 (0.10,0.17) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03)

**Table 1.** Global mean of TNx as simulated by the CESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 35, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma_t/\sqrt{n}$  where  $\sigma_t$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma_t$  estimated by a subset of 5 ensemble members, and 5 years around the year t considered (columns labelled by "(F-5)"). Results are shown for four individual years (t) along the simulation (column-wise), since  $\sigma_t$  varies along its length.

160

This behavior is to be expected, as we know the RMSE of a mean behaves in inverse proportion to the square root of the size of the sample from which the mean is computed, but the actual behavior shown in the plots and the table could be misleading, as the variability of the largest means (largest in sample size *n*) could be underestimated by our bootstrap (see Section 3.1). More importantly, this assessment would not be possible if all we had was a 5-member ensemble for our model. We can therefore compute the formula for the standard error of a mean,  $\sigma/\sqrt{n}$  (see Section 3.1), using the full ensemble to estimate  $\sigma$ , which we assume to be the true standard deviation of the ensemble. We then repeat the estimation by substituting an estimate for  $\sigma$ derived using only 5 ensemble members. Table 1 shows RMSEs for the same increasing values of *n*. Each pair of columns compares side by side the bootstrap (B) and the formula (F) results, the latter also reporting the 95% confidence intervals due to having to estimate  $\sigma_t$ . Also shown is the result of applying the formula by estimating  $\sigma_t$  on the basis of a small ensemble (5

165





members) but, importantly for the accuracy of our results, increasing the sample size by using a window of 5 years around the individual dates. From the table entries we can assess that the bootstrap estimation is the most optimistic about the size of the RMSE for the estimation of the forced signal of these two quantities once the ensemble size exceeds about 15 - 20 (out of 40 available). For the larger sizes the RMSE estimated by the bootstrap falls in all cases outside of the confidence interval under the (F) column. However, the estimates of RMSE associated with an ensemble size of 10 or 15 already quantifies a high degree of accuracy for the approximation of the ensemble mean of the full 40-member ensemble: those RMSEs for TNx are on the order of 0.02°C - 0.04°C.

Table 2 reports the same analysis results for the precipitation metric, Rx5Day. The same general message can be drawn,
with narrower estimates by the bootstrap approach for ensemble sizes starting at around 20 or 25 members. Even in this case however the estimates for the RMSE is on the order of 0.1 – 0.2mm/day for Rx5Day once the ensemble size exceeds 10.

	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.40	0.40 (0.33,0.51)	0.33 (0.26,0.46)	0.49	0.50 (0.41,0.64)	0.45 (0.35,0.63)
n=5	0.16	0.18 (0.15,0.23)	0.15 (0.12,0.21)	0.15	0.22 (0.18,0.29)	0.20 (0.16,0.28)
n=10	0.11	0.13 (0.10,0.16)	0.10 (0.08,0.15)	0.11	0.16 (0.13,0.20)	0.14 (0.11,0.20)
n=15	0.08	0.10 (0.08,0.13)	0.09 (0.07,0.12)	0.09	0.13 (0.11,0.17)	0.12 (0.09,0.16)
n=20	0.07	0.09 (0.07,0.11)	0.07 (0.06,0.10)	0.07	0.11 (0.09,0.14)	0.10 (0.08,0.14)
n=25	0.05	0.08 (0.07,0.10)	0.07 (0.05,0.09)	0.06	0.10 (0.08,0.13)	0.09 (0.07,0.13)
n=30	0.03	0.07 (0.06,0.09)	0.06 (0.05,0.08)	0.05	0.09 (0.07,0.12)	0.08 (0.06,0.11)
n=35	0.02	0.07 (0.06,0.09)	0.06 (0.04,0.08)	0.03	0.08 (0.07,0.11)	0.08 (0.06,0.11)
	2050 (B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050 (B) 0.54	2050 (F) 0.55 (0.45,0.71)	2050 (F-5) 0.52 (0.40,0.72)	2097 (B) 0.76	2097 (F) 0.77 (0.63,0.98)	2097 (F-5) 0.65 (0.51,0.90)
n=1 n=5	2050 (B) 0.54 0.23	2050 (F) 0.55 (0.45,0.71) 0.25 (0.20,0.32)	2050 (F-5) 0.52 (0.40,0.72) 0.23 (0.18,0.32)	2097 (B) 0.76 0.33	2097 (F) 0.77 (0.63,0.98) 0.34 (0.28,0.44)	2097 (F-5) 0.65 (0.51,0.90) 0.29 (0.23,0.40)
n=1 n=5 n=10	2050 (B) 0.54 0.23 0.17	2050 (F) 0.55 (0.45,0.71) 0.25 (0.20,0.32) 0.17 (0.14,0.22)	2050 (F-5) 0.52 (0.40,0.72) 0.23 (0.18,0.32) 0.16 (0.13,0.23)	2097 (B) 0.76 0.33 0.19	2097 (F) 0.77 (0.63,0.98) 0.34 (0.28,0.44) 0.24 (0.20,0.31)	2097 (F-5) 0.65 (0.51,0.90) 0.29 (0.23,0.40) 0.21 (0.16,0.29)
n=1 n=5 n=10 n=15	2050 (B) 0.54 0.23 0.17 0.14	2050 (F) 0.55 (0.45,0.71) 0.25 (0.20,0.32) 0.17 (0.14,0.22) 0.14 (0.12,0.18)	2050 (F-5) 0.52 (0.40,0.72) 0.23 (0.18,0.32) 0.16 (0.13,0.23) 0.13 (0.10,0.19)	2097 (B) 0.76 0.33 0.19 0.15	2097 (F) 0.77 (0.63,0.98) 0.34 (0.28,0.44) 0.24 (0.20,0.31) 0.20 (0.16,0.25)	2097 (F-5) 0.65 (0.51,0.90) 0.29 (0.23,0.40) 0.21 (0.16,0.29) 0.17 (0.13,0.23)
n=1 n=5 n=10 n=15 n=20	2050 (B) 0.54 0.23 0.17 0.14 0.09	2050 (F) 0.55 (0.45,0.71) 0.25 (0.20,0.32) 0.17 (0.14,0.22) 0.14 (0.12,0.18) 0.12 (0.10,0.16)	2050 (F-5) 0.52 (0.40,0.72) 0.23 (0.18,0.32) 0.16 (0.13,0.23) 0.13 (0.10,0.19) 0.12 (0.09,0.16)	2097 (B) 0.76 0.33 0.19 0.15 0.10	2097 (F) 0.77 (0.63,0.98) 0.34 (0.28,0.44) 0.24 (0.20,0.31) 0.20 (0.16,0.25) 0.17 (0.14,0.22)	2097 (F-5) 0.65 (0.51,0.90) 0.29 (0.23,0.40) 0.21 (0.16,0.29) 0.17 (0.13,0.23) 0.15 (0.11,0.20)
n=1 n=5 n=10 n=15 n=20 n=25	2050 (B) 0.54 0.23 0.17 0.14 0.09 0.08	2050 (F) 0.55 (0.45,0.71) 0.25 (0.20,0.32) 0.17 (0.14,0.22) 0.14 (0.12,0.18) 0.12 (0.10,0.16) 0.11 (0.09,0.14)	2050 (F-5) 0.52 (0.40,0.72) 0.23 (0.18,0.32) 0.16 (0.13,0.23) 0.13 (0.10,0.19) 0.12 (0.09,0.16) 0.10 (0.08,0.14)	2097 (B) 0.76 0.33 0.19 0.15 0.10 0.10	2097 (F) 0.77 (0.63,0.98) 0.34 (0.28,0.44) 0.24 (0.20,0.31) 0.20 (0.16,0.25) 0.17 (0.14,0.22) 0.15 (0.13,0.20)	2097 (F-5) 0.65 (0.51,0.90) 0.29 (0.23,0.40) 0.21 (0.16,0.29) 0.17 (0.13,0.23) 0.15 (0.11,0.20) 0.13 (0.10,0.18)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050 (B) 0.54 0.23 0.17 0.14 0.09 0.08 0.04	2050 (F) 0.55 (0.45,0.71) 0.25 (0.20,0.32) 0.17 (0.14,0.22) 0.14 (0.12,0.18) 0.12 (0.10,0.16) 0.11 (0.09,0.14) 0.10 (0.08,0.13)	2050 (F-5) 0.52 (0.40,0.72) 0.23 (0.18,0.32) 0.16 (0.13,0.23) 0.13 (0.10,0.19) 0.12 (0.09,0.16) 0.10 (0.08,0.14) 0.09 (0.07,0.13)	2097 (B) 0.76 0.33 0.19 0.15 0.10 0.10 0.07	2097 (F) 0.77 (0.63,0.98) 0.34 (0.28,0.44) 0.24 (0.20,0.31) 0.20 (0.16,0.25) 0.17 (0.14,0.22) 0.15 (0.13,0.20) 0.14 (0.11,0.18)	2097 (F-5) 0.65 (0.51,0.90) 0.29 (0.23,0.40) 0.21 (0.16,0.29) 0.17 (0.13,0.23) 0.15 (0.11,0.20) 0.13 (0.10,0.18) 0.12 (0.09,0.16)

Table 2. Same as Table 1, for Rx5Day.

For both metrics, estimating  $\sigma_t$  using only 5 ensemble members (and a window of 5 years around the year t of interest) delivers accurate estimates of the RMSE and its confidence interval as soon as the ensemble size considered exceeds 5 or 10. The lesson learned here is that

180 1. if global average quantities of these indices are concerned, and





2. if the formula for estimating the RMSE on the basis of a given sample size is adopted,

it is possible, on the basis of an existing 5-member ensemble to accurately estimate the required ensemble size to identify the forced component within a given tolerance for error. Of course the size of this tolerance would be ideally dictated by the use the analysis is put towards.

We note here that the calculation of the RMSE for increasing ensemble sizes is straightforward, once σ<sub>t</sub> is estimated. Even more straightforward is the calculation of the expected "gain" in narrowing the RMSE. A simple ratio calculation shows that for n spanning the range 5 to 45 (relevant sizes for our specific examples) the reduction in RMSE follows the sequence {100 \* 1/√n}<sub>n=1,5,...,45</sub>. If we take the RMSE affecting estimates based on only one member (in essence, our estimate of σ) as reference, we would expect an RMSE that is 45%, 32%, 22%, 17% or 15% of that for ensemble sizes of n = 5, 10, 20, 35, or 45
respectively.

190 respectively.

We assess how the results of the formula compare to the actual error by considering the difference between the smaller size ensemble means and the truth (the full ensemble mean), year by year and comparing that difference to twice the expected RMSE derived by the formula (akin to considering twice the standard deviation of a normally distributed quantity). Figure 2 for global averages of the two same quantities, shows the ratios of actual vs. expected error, indicating the 100% level by a

horizontal line for reference (and indicating by a diagonal line the behavior of  $\sigma_t$  over time, for reference, as summarized by fitting a linear regression in t). As can be gauged, the actual error is in most cases much smaller than the expected, especially for ensemble sizes greater than 5, and we see that only occasionally the actual error spikes above the expected (above 100%) for individual years, consistent with what would be expected of a normally distributed error compared to the  $2\sigma$  quantity.







**Figure 2.** In each plot, for each year, the height of the bar gives the error in the estimate of the forced component (defined as the mean of the entire ensemble) as a percentage of the expected error size, estimated by the formula  $\sigma_t/\sqrt{n}$  with *n* the ensemble size. Grey bars if  $\sigma_t$  is estimated using the whole ensemble, colored bars if  $\sigma_t$  is estimated using only 5 ensemble members (but using five years around each year). The diagonal line and the scale on the right axis give the actual size and behavior over time (as  $\sigma_t$  may vary with *t*) of the expected error. We choose to show the stylized behavior of the estimated error here, as the the least square linear regression fit of  $\sigma_t/\sqrt{n}$  onto *t*. Each plot corresponds to a different and increasing ensemble size: 1,5,10,15,20,25,30,35. The top two rows of plots are for TNx; the bottom two rows are for Rx5Day. CESM1-CAM5 results.

In the supplementary material we report the results of applying the same analysis to the rest of the indices, and for regional results. Even if we can't show all results, we tested country averages, zonal averages, land and ocean regions separately, confirming that the qualitative behavior we assess here is common to all these other instances.

Here we go on to show how the same type of analysis can be applied at the grid scale, and still deliver an accurate bound for the error in approximating the forced component. For the grid scale analysis, we define as the forced component anomalies by mid- and end-of-century (compared to a baseline) obtained as differences between five year averages: 2048-2052 and 2096-2100 vs 2000-2005. We use the full ensemble or only 5 members to estimate the ensemble standard deviation  $\sigma$  at each

205 2096-2100 vs 2000-2005. We use the full ensemble or only 5 members to estimate the ensemble standard deviation  $\sigma$  at each grid-point, and compare the actual error when approximating the "true" anomalies (i.e., those obtained on the basis of the full ensemble) by increasingly larger ensemble sizes to the expected error, calculated by the formula  $2\sigma/\sqrt{n}$ . In Figures 3 and 4 we show fields of the ratio of actual to expected error as the ensemble size increases. Red areas are ones where the





ratio exceeds 100%, i.e., where the estimated error failed to anticipate the actual error in the ensemble in estimating the true anomalies. As can be gauged even by eye, only small and sparse areas appear where the actual error exceeds the expected

- 210 anomalies. As can be gauged even by eye, only small and sparse areas appear where the actual error exceeds the expected error. Over the majority of the Earth's surface, particularly when errors are estimated for ensemble sizes of 20 or more, the expected error is a good measure of what the actual is, consistently providing a conservative estimate of it according to normal distribution theory. Tables B1 through B4 in the appendix report percentages of surface areas (distinguishing global, land-only or ocean-only aggregation) where the estimate fails to provide a ceiling for the actual error, i.e., where the values of the fields exceed 100%. As can be assessed, 20 ensemble members consistently keep such fraction at or under 5% for the CESM model
- ensemble, while the coarser resolution CanESM requires 25 ensemble members for that to be true.



**Figure 3.** Anomalies in TNx by mid century (top two rows) and end-of-century (bottom two rows). In each plot, for increasing ensemble sizes, the color at each grid-point indicates the fraction (as a percentage) of the estimated error that the real error represents. Values less than 100% indicate that the estimated error is an effective upper bound for the real error in estimating the anomaly at that location. The color scale highlights in dark red the values above 100%, whose total fraction is reported in Table B1.







**Figure 4.** Anomalies in Rx5Day by mid century (top two rows) and end-of-century (bottom two rows). In each plot, for increasing ensemble sizes, the color at each grid-point indicates the fraction (as a percentage) of the estimated error that the real error represents. Values less than 100% indicate that the estimated error is an effective upper bound for the real error in estimating the anomaly at that location. The color scale highlights in dark red the values above 100%, whose total fraction is reported in Table B3.

Overall, these results attest to the fact that we can use a small ensemble of 5 members to estimate the population standard deviation around the population mean, and apply the formula for the standard error of a mean as a function of sample size to decide how large an ensemble we need in order to approximate the forced component to a given degree of accuracy. This holds true across the range of spatial scales afforded by these models, starting from global means down to subcontinental regional averages all the way to grid-point values.

4.2 GEV results

220

As explained in Section 3, the extreme metrics we chose can be fit by a Generalized Extreme Value distribution, and return levels for arbitrary return periods derived, with their confidence interval. In this section we ask two questions.

- 1. How many ensemble members are needed for the estimates to stabilize and the size of the confidence interval not to change in a significant way? And
  - 2. Is there any gain in applying GEV fitting rather than simply "counting" rare events across the ensemble?





230

We perform the analysis for a set of individual locations (i.e., grid-points), as for most extreme quantities there would be little value in characterizing very rare events as means of large geographical regions. Figure C1 shows the 15 locations that we chose with the goal of testing a diverse set of climatic conditions. We choose three years along the simulations (2000, 2050 and 2095<sup>1</sup>) around which we extract an 11-yr window of data. This relatively short period allows us to assume quasi-stationarity eliminating the need for temporal covariates in the estimation of the GEV paramaters. We estimate return levels (event sizes,  $z_p$  in our notation) for a number of return periods, i.e., 2-,5-,10-,20-,50- and 100-years by concatenating the 11 year segments across the n ensemble members. Here we show results for our two metrics, choosing two different locations for each. These results are indicative of what happens across the rest of locations, and for the other metrics and the other model considered (see 235 supplementary material for a sampling of those).

Figure 5. Return Levels for TNx at (row-wise) 2000, 2050 and 2100 for (column-wise) 2-, 5-, 10-, 20-, 50-, 100-year return periods, based on estimating a GEV by using 11-vr windows of data around each date. In each plot, for increasing ensemble sizes along the x-axis (from 5 to the full ensemble, 40), the red dots indicate the central estimate, and the pink envelope represents the 95% confidence interval. The estimates based on the full ensemble, which we consider the truth, are also drawn across the plot for reference, as horizontal lines. The blue dots in each plot show the same quantities estimated simply by counting, i.e., computing the empirical cumulative distribution function of TNx on the basis of the  $n \times 11$  years where n is the ensemble size. The first three rows show results for a location in Australia while the following three rows show results for a location in Northern North America (see Figure C1).

<sup>&</sup>lt;sup>1</sup>As some of the experiments end at 2099 and we want an 11-year window around each date this latter date becomes 2094 in some cases).







**Figure 6.** Return Levels for Rx5Day at (row-wise) 2000, 2050 and 2100 for (column-wise) 2-, 5-, 10-, 20-, 50-, 100-year return periods, based on estimating a GEV by using 11-yr windows of data around each date. In each plot, for increasing ensemble sizes along the x-axis (from 5 to the full ensemble, 40), the red dots indicate the central estimate, and the pink envelope represents the 95% confidence interval. The estimates based on the full ensemble, which we consider the truth, are also drawn across the plot for reference, as horizontal lines. The blue dots in each plot show the same quantities estimated simply by counting, i.e., computing the empirical cumulative distribution function of Rx5Day on the basis of the  $n \times 11$  years where n is the ensemble size. The first three rows show results for a location in Northern Asia while the following three rows show results for a location in Southern Africa (see Figure C1).

Figures 5 and 6 compare for each return level (along the columns), and across the projection dates (along the rows), the behavior of the GEV central estimates (red dots) and 95% confidence intervals (pink envelope) based on an increasing ensemble size (along the x-axis) to the ones obtained by the full ensemble (considered to be the truth), which are drawn as a reference

- 240 across each plot as horizontal lines. Further, estimates of the central quantities based on computing an empirical cumulative distribution function from the data are added to each plot as blue dots for each of the ensemble sizes considered (in this case as well using 11-year windows for each ensemble member, so that the sample has the same size as that used for the GEV fitting of  $n \times 11$ ). The general message can be summarized by two observations. First, an ensemble size of 20 or 25 (corresponding to a sample size of 220 to 275 years) appears to be the lower bound at which the estimates stabilize in most
- 245 cases. Both the central estimates and the confidence intervals do not depart significantly from the point estimate and range of the truth (horizontal lines), and gradually converge to it reliably (barred the odd result that we cannot exclude given the large





number of trials we are performing). The additional precision gained by increasing the sample size from 25 becomes soon marginal. Second, the empirical estimates have the drawback of not providing straightforward uncertainty bounds, but seem not to deviate significantly and consistently from those obtained by the GEV, falling within those confidence intervals in most cases, again starting most reliably once the ensemble size exceeds 20 or 25. In summary, an ensemble size of this magnitude seems appropriate to compute, either empirically or by GEV fits, return levels for events as rare as having p = 0.01 to occur in a given year. The use of a GEV approach allows to characterize the uncertainty bounds straightforwardly as opposed to the empirical "counting" approach but the central estimates from the two approaches do not seem to differ significantly in most cases tested.

255 The same statistical precision may be realized with fewer ensemble members by relaxing the quasi-stationary assumption and extending the analysis period to contain a similar number of years. However, this then necessitates usage of temporal covariates adding another source of fitting uncertainty.

# 4.3 Characterizing internal variability

- After concerning ourselves with the characterization of the forced component we turn to the complementary problem of characterizing internal variability. Rather than aiming at eliminating the effects of internal variability as we have done so far in the estimation of a forced signal, we take here the opposite perspective, wanting to fully characterize that internal variability. After all, the real world realization won't be akin to the mean of the ensemble, but to one of its members, and we want to be sure to estimate the range of variations such members may display. Thus, we ask how large the ensemble needs to be to fully characterize the variations that the full-size ensemble produces, in the form of the ensemble variance; we also ask how large an ensemble is needed to detect changes in the size of internal variability with changing external forcing. Both these questions we
- tackle directly at the grid-point scale, as the answer to that problem is bound to be a conservative answer to any other problem that concerns the characterization of variability at a larger spatial scale. Figures 7 through 10 synthesize our findings for both these questions.







**Figure 7.** Estimating the ensemble variance for TNx: Each plot corresponds to a year along the simulation length (1950, 1975, 2000, 2025, 2050, 2075, 2100). The color indicates the number of ensemble members needed to estimate an ensemble variance at that location that is statistically indistinguishable from that computed on the basis of the full 40-member ensemble. The results of the first two columns use only the specific year across the members. The results of the third and fourth columns enrich the samples by using five years around the specific date.







**Figure 8.** Estimating the within ensemble variance for Rx5Day: Each plot corresponds to a year along the simulation length (1950, 1975, 2000, 2025, 2050, 2075, 2100). The color indicates the number of ensemble members needed to estimate an ensemble variance at that location that is statistically indistinguishable from that computed on the basis of the full 40-member ensemble. The results of the first two columns use only the specific year across the members. The results of the third and fourth columns enrich the samples by using five years around the specific date.





The two columns on the left-hand side of Figure 7 show for several years along the simulation how many ensemble members 270 are needed (denoted by the colors of the legends) in order to estimate an ensemble variance at each grid-point that is not statistically distinguishable from the same variance estimated by the full 40-member ensemble, for TNx. Note that we do this at various times along the length of the simulation (1950, 1975, 2000, 2025, 2050, 2075, 2100) because we account for the possibility that internal variability might change over its course with increasing external forcing, but for now we remain agnostic on this issue. For all dates, most of the area indicates that 5 members are sufficient, but a remaining noisy pattern shows that at some locations ten members are needed. The same type of Figure for the precipitation metric, Figure 8, confirms 275 that for the noisier quantity a larger extent of the Earth's surface needs ensemble sizes of ten or more to accurately estimate the behavior of the full ensemble. The two right-hand columns in Figure 7 show corresponding plots where now most of the areas only require 5-members. This is the result of "borrowing strength" in the estimation of the ensemble variance by using a 5-year window around the date as we have done for the analysis of  $\sigma$  in the previous sections. This solution addresses the problem of

estimating the variance for both the temperature and the precipitation metric, as Figure 8 confirms. The supplementary material 280 attests to this remaining true for the other model and the remaining metrics as well. We note here that the patterns shown in these two figures have the characteristics of noise, but we have in fact determined the significance of these areas by applying a testing method that controls the False Discovery Rate (Ventura et al., 2004) under multiple testing conditions.

Detecting changes in the size of the variance over time by comparing two dates over the simulation is a problem that we 285 expect to require more statistical power than the problem of characterizing the size of the variance at a given point, as the difference between stochastic quantities is affected by larger uncertainty than the quantities individually considered, unless those are strongly correlated. Figure 9 shows the ensemble size required to detect the same changes in the ensemble variance of TNx that the full ensemble of 40-members detects. Each plot is at the intersection of a column and a row corresponding to two of the dates considered in the previous analysis, indicating that the solution applies to detecting a change in variance 290 between those two dates.







**Figure 9.** Estimating changes in ensemble variance for TNx: each plot corresponds to a pair of years along the simulation (same set of years as depicted in Figures 7 and 8 above). Colored areas are regions where on the basis of the full 40-member ensemble a significant change in variance was detected. The colors indicate the size of the smaller ensemble needed to detect the same change. Here the sampling size is increased by using 5 years around each date.







**Figure 10.** Estimating changes in within-ensemble variance for Rx5Day: each plot corresponds to a pair of years along the simulation (same set of years as depicted in Figures 7 and 8 above). Colored areas are regions where on the basis of the full 40-member ensemble a significant change in variance was detected. The colors indicate the size of the smaller ensemble needed to detect the same change. Here the sampling size is increased by using 5 years around each date.





Blank areas are regions where the full ensemble has not detected any changes in the ensemble variance at that location when comparing the two dates. Colored areas are regions where such change has been detected by the full ensemble, and the color indicates what (smaller) ensemble size is sufficient to detect the same change. These results are obtained by increasing the sample size using 5 years around the dates, as in the right-hand columns of Figures 7 and 8. In the case of minimum 295 temperature, on which this metric is based, the changes are confined to the Arctic region and in most cases the ensemble size required is again 5, with only an instance where the changes between mid-century and end-of the century require consistently a larger ensemble size (as many as 15 members over the region). When we conduct the same analysis on the precipitation metric, shown in Figure 10, we are presented with a spatially noisier pictures, with changes in variance scattered throughout the Earth's surface, especially over the oceans. This can again is being accurately characterized by a 5-member ensemble almost everywhere, aside from some isolated patches.

300

305

We don't show it explicitly here, as it is not the focus of our analysis, but, for both model ensembles, when the change is significant, the ensemble variance increases over time for both precipitation metrics, indicating that the ensemble spread increases with the strength of external forcing (which happens over time under RCP8.5). This is expected as the variance of precipitation increases in step with its mean. For the temperature based metrics, the changes, when significant, are mostly towards an increase in variance (ensemble spread) with forcings for hot extremes (TNx and TXx, the hottest night and day of the year), for which the significant changes are mostly located in the Arctic region. The ensemble variance/spread decreases instead for cold extremes (TNn and TXn, the coldest night and day of the year), for which the significant changes are mostly located along the Antarctic continent edge.

#### 4.4 Signal-to-Noise considerations

- Another aspect that is implicitly relevant to the establishment of a required ensemble size, if the estimation is concerned with 310 emergence of the forced component, or, more in general, with 'detection and attribution'-type analysis is the Signal-to-Noise ratio of the quantity of interest. Assuming as we have done so far that the quantity of interest can be regarded as the mean  $\mu$ of a noisy population, the signal to noise ratio is defined as  $S_N = \mu/\sigma$  where  $\sigma$  is the standard deviation of the population. A critical threshold, say K, for  $S_N$  is usually set at K = 1 or 2, and it is immediate to derive the sample size required for such
- threshold to be hit, by computing the value of n that makes  $\mu/(\sigma/\sqrt{n}) \ge K$ , i.e.,  $n \ge K^2/S_N^2$ . Figure 11 shows two maps 315 of the spatially varying ensemble sizes required for the signal to noise ratio to exceed 1, when computing anomalies at midand end-of-century for the warmest night of the year, TNx. The anomalies are computed as five year mean differences, as in Section 4.1 under RCP8.5, and clearly by the end of the century the entire Earth's surface experiences the emergence of the signal, even by averaging 1 or 2 ensemble members. The map of changes by mid-century is more interesting, as evidently some
- 320 areas require more statistical power, i.e., a larger n, especially around the polar regions, as expected, where internal variability is significantly larger than in lower latitudes, translating into a smaller  $S_N$ .







TNx Anomalies at end of Century



**Figure 11.** Ensemble size n required for the signal to noise ratio of the grid-point scale TNx anomalies to exceed 1. Top map: anomalies defined as the mean of 2048-2052 minus the historical baseline taken as 2000-2005. Bottom map: anomalies as 2096-2100 minus 2000-2005 means. In both cases the future simulations follow RCP8.5.





#### Conclusions 5

325

In this study we have addressed the need of deciding a-priori the size of a large ensemble, using an existing five member ensemble as our guidance. Aware that the optimal size ultimately depends on the purpose the ensemble is used for, and in order to cover a wide range of possible uses, we chose metrics of temperature and precipitation extremes and we considered output at grid-point scale and at various scales of aggregation, up to global averages. We tackled the problem of characterizing forced changes along the length of a transient scenario simulation, and that of characterizing the system's internal variability and its possible changes. By using a high emission scenario like RCP8.5, but considering behaviors all along the length of the simulations, we are also implicitly addressing a wide range of signal-to-noise magnitudes. Using the availability of existing large ensembles with two different models, CESM1 and CanESM2, we could compare our estimates of the expected errors that 330 a given ensemble size would generate with actual errors, obtained using the full ensembles' estimates as our "truth".

First, we find that for the many uses that we explored, it is possible to put a ceiling on the expected error associated with a given ensemble size by exploiting a small ensemble of 5 members. We estimate the ensemble variance at a given simulation date (e.g., 2000, or 2050, or 2095), which is the basis for all our error computations, on the basis of five members, "borrowing

- strength" by using a window of five years around that date. The results we assess are consistent with assuming that the quantities 335 of interest are normally distributed with standard deviation  $\sigma/\sqrt{n}$ , where  $\sigma$  can be estimated on the basis of the 5 members available: the error estimates and therefore the optimal sizes computed on the basis of choosing a given tolerance for such errors provide a safe upper bound to the errors that would be committed for a given ensemble size n. This is true for all metrics considered, both models, and the full range of scales of aggregation. When we use such estimates (later verified by the
- availability of the actual large ensembles) there appears to be a sweet spot in the range of ensemble sizes that provides accurate 340 estimates for both forced changes and internal variability, consisting of 20 or 25 members. The larger of these sizes also appears approximately sufficient to conduct an estimation of rare events with as low as 0.01 probability of occurrence each year, by fitting a GEV and deriving return levels and their confidence intervals. In most cases (locations around the globe, times along the simulation, and metrics considered) enlarging the sample size beyond 25 members provides only marginal improvement in
- the confidence intervals, while the central estimate does not change significantly from the one established using 25 members, 345 and accurately approximating that obtained by the full ensemble.

In all cases considered a much smaller ensemble size of 5 to 10 members, if enriched by sampling along the time dimension (that is, using a 5-year window around the date of interest) is sufficient to characterize the ensemble variability, and its changes along the course of the simulations under increasing greenhouse-gases, when found significant using the full ensemble size.

- 350 Some caveats are in order. Obviously, the question of how many ensemble members are needed is fundamentally ill-posed, as the answer ultimately and always depends on the most exacting use to which the ensemble is put. One can always find a higher-frequency, smaller-scale metric, and a tighter error bound to satisfy, requiring a larger ensemble size than any previously identified. As tropical cyclone permitting and eventually convection permitting climate model simulations become available, these metrics will be more commonly analyzed. Even for a specific use, the answer depends on the characteristics of internal
- 355 variability, and the fact that for these two models 5 ensemble members are sufficient to obtain an accurate estimate of it is



variation.



promising, but not guarantee that 5 are sufficient for all models. In fact, this could also be invalidated by a different experimental exploration of internal variability: new work is exploring different types of initialization, involving ocean states, which could uncover a dimension of internal variability that has so far being under-appreciated. This would likely change our best estimates of internal variability, and with it possibly the ensemble sizes required to accurately estimate it.

- With this work, however we have shown a way to attack the problem "bottom up", starting from a smaller ensemble and building estimates of what would be required for a given problem. One can imagine a more sophisticated set-up where an ensemble can be recursively augmented (rather than assuming a fixed 5-member ensemble as we have done here) in order to approximate the full variability incrementally better. We have also shown that for a large range of questions the size needed is actually well below what we have come to associate with "Large Ensembles". There exist other important sources of uncertainties in climate modeling, one of which is beyond reach of any single modeling center, having to do with structural uncertainty Knutti et al. (2010). Adopting the perspective of an individual model, however, parameter settings have as important a role – at least – as initial conditions. Together with scenario uncertainty, all these dimensions compete over computational resources for their exploration. Our results may be of guidance in choosing how to allocate those resources among these alternative sources of
- 370 Code and data availability. The large ensembles output is available through the CLIVAR Large Ensemble Working Group webpage, in the archive maintained through the NCAR CESM community project cesm.ucar.edu/projects/community-projects/MMLEA/. R code for these analyses is available from the first author on reasonable request.





# Appendix A: RMSE estimation for more indices and based on the CanESM ensemble

	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.23	0.24 (0.19,0.30)	0.20 (0.16,0.28)	0.20	0.20 (0.17,0.26)	0.22 (0.17,0.31)
n=5	0.10	0.11 (0.09,0.14)	0.09 (0.07,0.13)	0.07	0.09 (0.07,0.12)	0.10 (0.08,0.14)
n=10	0.07	0.07 (0.06,0.10)	0.06 (0.05,0.09)	0.05	0.06 (0.05,0.08)	0.07 (0.05,0.10)
n=15	0.05	0.06 (0.05,0.08)	0.05 (0.04,0.07)	0.04	0.05 (0.04,0.07)	0.06 (0.04,0.08)
n=20	0.03	0.05 (0.04,0.07)	0.05 (0.04,0.06)	0.03	0.05 (0.04,0.06)	0.05 (0.04,0.07)
n=25	0.02	0.05 (0.04,0.06)	0.04 (0.03,0.06)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.06)
n=30	0.02	0.04 (0.04,0.06)	0.04 (0.03,0.05)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.06)
n=35	0.01	0.04 (0.03,0.05)	0.03 (0.03,0.05)	0.01	0.03 (0.03,0.04)	0.04 (0.03,0.05)
L						
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.20	2050 (F) 0.20 (0.16,0.26)	2050 (F-5) 0.21 (0.17,0.30)	2097 (B) 0.24	2097 (F) 0.24 (0.20,0.32)	2097 (F-5) 0.18 (0.14,0.25)
n=1 n=5	2050(B) 0.20 0.08	2050 (F) 0.20 (0.16,0.26) 0.09 (0.07,0.12)	2050 (F-5) 0.21 (0.17,0.30) 0.10 (0.08,0.13)	2097 (B) 0.24 0.12	2097 (F) 0.24 (0.20,0.32) 0.11 (0.09,0.14)	2097 (F-5) 0.18 (0.14,0.25) 0.08 (0.06,0.11)
n=1 n=5 n=10	2050(B) 0.20 0.08 0.05	2050 (F) 0.20 (0.16,0.26) 0.09 (0.07,0.12) 0.06 (0.05,0.08)	2050 (F-5) 0.21 (0.17,0.30) 0.10 (0.08,0.13) 0.07 (0.05,0.09)	2097 (B) 0.24 0.12 0.06	2097 (F) 0.24 (0.20,0.32) 0.11 (0.09,0.14) 0.08 (0.06,0.10)	2097 (F-5) 0.18 (0.14,0.25) 0.08 (0.06,0.11) 0.06 (0.04,0.08)
n=1 n=5 n=10 n=15	2050(B) 0.20 0.08 0.05 0.04	2050 (F) 0.20 (0.16,0.26) 0.09 (0.07,0.12) 0.06 (0.05,0.08) 0.05 (0.04,0.07)	2050 (F-5) 0.21 (0.17,0.30) 0.10 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08)	2097 (B) 0.24 0.12 0.06 0.06	2097 (F) 0.24 (0.20,0.32) 0.11 (0.09,0.14) 0.08 (0.06,0.10) 0.06 (0.05,0.08)	2097 (F-5) 0.18 (0.14,0.25) 0.08 (0.06,0.11) 0.06 (0.04,0.08) 0.05 (0.04,0.07)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.20 0.08 0.05 0.04 0.03	2050 (F) 0.20 (0.16,0.26) 0.09 (0.07,0.12) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06)	2050 (F-5) 0.21 (0.17,0.30) 0.10 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07)	2097 (B) 0.24 0.12 0.06 0.06 0.05	2097 (F) 0.24 (0.20,0.32) 0.11 (0.09,0.14) 0.08 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07)	2097 (F-5) 0.18 (0.14,0.25) 0.08 (0.06,0.11) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.20 0.08 0.05 0.04 0.03 0.02	2050 (F) 0.20 (0.16,0.26) 0.09 (0.07,0.12) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06) 0.04 (0.03,0.05)	2050 (F-5) 0.21 (0.17,0.30) 0.10 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06)	2097 (B) 0.24 0.12 0.06 0.06 0.05 0.03	2097 (F) 0.24 (0.20,0.32) 0.11 (0.09,0.14) 0.08 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06)	2097 (F-5) 0.18 (0.14,0.25) 0.08 (0.06,0.11) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.20 0.08 0.05 0.04 0.03 0.02 0.02	2050 (F) 0.20 (0.16,0.26) 0.09 (0.07,0.12) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.04 (0.03,0.05)	2050 (F-5) 0.21 (0.17,0.30) 0.10 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2097 (B) 0.24 0.12 0.06 0.06 0.05 0.03 0.02	2097 (F) 0.24 (0.20,0.32) 0.11 (0.09,0.14) 0.08 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.06)	2097 (F-5) 0.18 (0.14,0.25) 0.08 (0.06,0.11) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.03 (0.03,0.05)

**Table A1.** Global mean of TNn as simulated by the CESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 35, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.



Earth System	Ope
Dynamics	A (EGU
Discussions	ess

	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.11	0.11 (0.09,0.14)	0.10 (0.08,0.14)	0.14	0.14 (0.12,0.19)	0.13 (0.10,0.18)
n=5	0.05	0.05 (0.04,0.06)	0.04 (0.03,0.06)	0.05	0.06 (0.05,0.08)	0.06 (0.05,0.08)
n=10	0.03	0.04 (0.03,0.05)	0.03 (0.02,0.04)	0.04	0.05 (0.04,0.06)	0.04 (0.03,0.06)
n=15	0.02	0.03 (0.02,0.04)	0.03 (0.02,0.04)	0.03	0.04 (0.03,0.05)	0.03 (0.03,0.05)
n=20	0.02	0.02 (0.02,0.03)	0.02 (0.02,0.03)	0.02	0.03 (0.03,0.04)	0.03 (0.02,0.04)
n=25	0.01	0.02 (0.02,0.03)	0.02 (0.02,0.03)	0.02	0.03 (0.02,0.04)	0.03 (0.02,0.04)
n=30	0.01	0.02 (0.02,0.03)	0.02 (0.01,0.03)	0.01	0.03 (0.02,0.03)	0.02 (0.02,0.03)
n=35	0.01	0.02 (0.02,0.02)	0.02 (0.01,0.02)	0.01	0.02 (0.02,0.03)	0.02 (0.02,0.03)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.12	2050 (F) 0.12 (0.10,0.16)	2050 (F-5) 0.16 (0.12,0.22)	2097 (B) 0.15	2097 (F) 0.16 (0.13,0.20)	2097 (F-5) 0.14 (0.11,0.19)
n=1 n=5	2050(B) 0.12 0.06	2050 (F) 0.12 (0.10,0.16) 0.05 (0.04,0.07)	2050 (F-5) 0.16 (0.12,0.22) 0.07 (0.06,0.10)	2097 (B) 0.15 0.07	2097 (F) 0.16 (0.13,0.20) 0.07 (0.06,0.09)	2097 (F-5) 0.14 (0.11,0.19) 0.06 (0.05,0.09)
n=1 n=5 n=10	2050(B) 0.12 0.06 0.03	2050 (F) 0.12 (0.10,0.16) 0.05 (0.04,0.07) 0.04 (0.03,0.05)	2050 (F-5) 0.16 (0.12,0.22) 0.07 (0.06,0.10) 0.05 (0.04,0.07)	2097 (B) 0.15 0.07 0.04	2097 (F) 0.16 (0.13,0.20) 0.07 (0.06,0.09) 0.05 (0.04,0.06)	2097 (F-5) 0.14 (0.11,0.19) 0.06 (0.05,0.09) 0.04 (0.03,0.06)
n=1 n=5 n=10 n=15	2050(B) 0.12 0.06 0.03 0.02	2050 (F) 0.12 (0.10,0.16) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04)	2050 (F-5) 0.16 (0.12,0.22) 0.07 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06)	2097 (B) 0.15 0.07 0.04 0.03	2097 (F) 0.16 (0.13,0.20) 0.07 (0.06,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05)	2097 (F-5) 0.14 (0.11,0.19) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.12 0.06 0.03 0.02 0.02	2050 (F) 0.12 (0.10,0.16) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04)	2050 (F-5) 0.16 (0.12,0.22) 0.07 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2097 (B) 0.15 0.07 0.04 0.03 0.03	2097 (F) 0.16 (0.13,0.20) 0.07 (0.06,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.05)	2097 (F-5) 0.14 (0.11,0.19) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.12 0.06 0.03 0.02 0.02 0.01	2050 (F) 0.12 (0.10,0.16) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04) 0.02 (0.02,0.03)	2050 (F-5) 0.16 (0.12,0.22) 0.07 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04)	2097 (B) 0.15 0.07 0.04 0.03 0.03 0.02	2097 (F) 0.16 (0.13,0.20) 0.07 (0.06,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04)	2097 (F-5) 0.14 (0.11,0.19) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.12 0.06 0.03 0.02 0.02 0.01 0.01	2050 (F) 0.12 (0.10,0.16) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.16 (0.12,0.22) 0.07 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)	2097 (B) 0.15 0.07 0.04 0.03 0.03 0.02 0.01	2097 (F) 0.16 (0.13,0.20) 0.07 (0.06,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04)	2097 (F-5) 0.14 (0.11,0.19) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.03 (0.02,0.04)

**Table A2.** Global mean of TXx as simulated by the CESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 35, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.





	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.21	0.21 (0.17,0.27)	0.19 (0.15,0.27)	0.19	0.20 (0.16,0.25)	0.21 (0.16,0.29)
n=5	0.09	0.09 (0.08,0.12)	0.09 (0.07,0.12)	0.07	0.09 (0.07,0.11)	0.09 (0.07,0.13)
n=10	0.06	0.07 (0.05,0.09)	0.06 (0.05,0.08)	0.05	0.06 (0.05,0.08)	0.07 (0.05,0.09)
n=15	0.04	0.05 (0.04,0.07)	0.05 (0.04,0.07)	0.04	0.05 (0.04,0.07)	0.05 (0.04,0.08)
n=20	0.02	0.05 (0.04,0.06)	0.04 (0.03,0.06)	0.03	0.04 (0.04,0.06)	0.05 (0.04,0.07)
n=25	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.05)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.06)
n=30	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.05)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.05)
n=35	0.01	0.04 (0.03,0.05)	0.03 (0.03,0.05)	0.01	0.03 (0.03,0.04)	0.04 (0.03,0.05)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.19	2050 (F) 0.19 (0.15,0.27)	2050 (F-5) 0.19 (0.18,0.32)	2097 (B) 0.22	2097 (F) 0.22 (0.18,0.29)	2097 (F-5) 0.17 (0.13,0.23)
n=1 n=5	2050(B) 0.19 0.08	2050 (F) 0.19 (0.15,0.27) 0.09 (0.07,0.12)	2050 (F-5) 0.19 (0.18,0.32) 0.09 (0.08,0.14)	2097 (B) 0.22 0.11	2097 (F) 0.22 (0.18,0.29) 0.10 (0.08,0.13)	2097 (F-5) 0.17 (0.13,0.23) 0.08 (0.06,0.10)
n=1 n=5 n=10	2050(B) 0.19 0.08 0.04	2050 (F) 0.19 (0.15,0.27) 0.09 (0.07,0.12) 0.06 (0.05,0.09)	2050 (F-5) 0.19 (0.18,0.32) 0.09 (0.08,0.14) 0.06 (0.06,0.10)	2097 (B) 0.22 0.11 0.06	2097 (F) 0.22 (0.18,0.29) 0.10 (0.08,0.13) 0.07 (0.06,0.09)	2097 (F-5) 0.17 (0.13,0.23) 0.08 (0.06,0.10) 0.05 (0.04,0.07)
n=1 n=5 n=10 n=15	2050(B) 0.19 0.08 0.04 0.04	2050 (F) 0.19 (0.15,0.27) 0.09 (0.07,0.12) 0.06 (0.05,0.09) 0.05 (0.04,0.07)	2050 (F-5) 0.19 (0.18,0.32) 0.09 (0.08,0.14) 0.06 (0.06,0.10) 0.05 (0.05,0.08)	2097 (B) 0.22 0.11 0.06 0.05	2097 (F) 0.22 (0.18,0.29) 0.10 (0.08,0.13) 0.07 (0.06,0.09) 0.06 (0.05,0.07)	2097 (F-5) 0.17 (0.13,0.23) 0.08 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.19 0.08 0.04 0.04	2050 (F) 0.19 (0.15,0.27) 0.09 (0.07,0.12) 0.06 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.06)	2050 (F-5) 0.19 (0.18,0.32) 0.09 (0.08,0.14) 0.06 (0.06,0.10) 0.05 (0.05,0.08) 0.04 (0.04,0.07)	2097 (B) 0.22 0.11 0.06 0.05 0.04	2097 (F) 0.22 (0.18,0.29) 0.10 (0.08,0.13) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06)	2097 (F-5) 0.17 (0.13,0.23) 0.08 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.19 0.08 0.04 0.04 0.03 0.03	2050 (F) 0.19 (0.15,0.27) 0.09 (0.07,0.12) 0.06 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2050 (F-5) 0.19 (0.18,0.32) 0.09 (0.08,0.14) 0.06 (0.06,0.10) 0.05 (0.05,0.08) 0.04 (0.04,0.07) 0.04 (0.04,0.06)	2097 (B) 0.22 0.11 0.06 0.05 0.04 0.03	2097 (F) 0.22 (0.18,0.29) 0.10 (0.08,0.13) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.06)	2097 (F-5) 0.17 (0.13,0.23) 0.08 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.19 0.08 0.04 0.04 0.03 0.02 0.02	2050 (F) 0.19 (0.15,0.27) 0.09 (0.07,0.12) 0.06 (0.05,0.09) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2050 (F-5) 0.19 (0.18,0.32) 0.09 (0.08,0.14) 0.06 (0.06,0.10) 0.05 (0.05,0.08) 0.04 (0.04,0.07) 0.04 (0.04,0.06) 0.04 (0.03,0.06)	2097 (B) 0.22 0.11 0.06 0.05 0.04 0.03 0.02	2097 (F) 0.22 (0.18,0.29) 0.10 (0.08,0.13) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.06) 0.04 (0.03,0.05)	2097 (F-5) 0.17 (0.13,0.23) 0.08 (0.06,0.10) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.05) 0.03 (0.02,0.04)

**Table A3.** Global mean of TXn as simulated by the CESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 35, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.





	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	1.05	1.06 (0.87,1.36)	0.90 (0.70,1.25)	1.23	1.25 (1.02,1.60)	1.15 (0.89,1.59)
n=5	0.5	0.48 (0.39,0.61)	0.40 (0.31,0.56)	0.52	0.56 (0.46,0.72)	0.51 (0.40,0.71)
n=10	0.34	0.34 (0.28,0.43)	0.28 (0.22,0.39)	0.36	0.39 (0.32,0.51)	0.36 (0.28,0.50)
n=15	0.20	0.27 (0.22,0.35)	0.23 (0.18,0.32)	0.29	0.32 (0.26,0.41)	0.30 (0.23,0.41)
n=20	0.19	0.24 (0.19,0.31)	0.20 (0.16,0.28)	0.17	0.28 (0.23,0.36)	0.26 (0.20,0.36)
n=25	0.12	0.21 (0.17,0.27)	0.18 (0.14,0.25)	0.17	0.25 (0.20,0.32)	0.23 (0.18,0.32)
n=30	0.08	0.19 (0.16,0.25)	0.16 (0.13,0.23)	0.11	0.23 (0.19,0.29)	0.21 (0.16,0.29)
n=35	0.07	0.18 (0.15,0.23)	0.15 (0.12,0.21)	0.08	0.21 (0.17,0.27)	0.19 (0.15,0.27)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 1.77	2050 (F) 1.79 (1.47,2.30)	2050 (F-5) 1.69 (1.32,2.36)	2097 (B) 2.18	2097 (F) 2.20 (1.81,2.83)	2097 (F-5) 2.22 (1.73,3.08)
n=1 n=5	2050(B) 1.77 0.82	2050 (F) 1.79 (1.47,2.30) 0.80 (0.66,1.03)	2050 (F-5) 1.69 (1.32,2.36) 0.76 (0.59,1.05)	2097 (B) 2.18 0.80	2097 (F) 2.20 (1.81,2.83) 0.99 (0.81,1.27)	2097 (F-5) 2.22 (1.73,3.08) 0.99 (0.77,1.38)
n=1 n=5 n=10	2050(B) 1.77 0.82 0.42	2050 (F) 1.79 (1.47,2.30) 0.80 (0.66,1.03) 0.57 (0.46,0.73)	2050 (F-5) 1.69 (1.32,2.36) 0.76 (0.59,1.05) 0.54 (0.42,0.75)	2097 (B) 2.18 0.80 0.58	2097 (F) 2.20 (1.81,2.83) 0.99 (0.81,1.27) 0.70 (0.57,0.89)	2097 (F-5) 2.22 (1.73,3.08) 0.99 (0.77,1.38) 0.70 (0.55,0.98)
n=1 n=5 n=10 n=15	2050(B) 1.77 0.82 0.42 0.43	2050 (F) 1.79 (1.47,2.30) 0.80 (0.66,1.03) 0.57 (0.46,0.73) 0.46 (0.38,0.59)	2050 (F-5) 1.69 (1.32,2.36) 0.76 (0.59,1.05) 0.54 (0.42,0.75) 0.44 (0.34,0.61)	2097 (B) 2.18 0.80 0.58 0.38	2097 (F) 2.20 (1.81,2.83) 0.99 (0.81,1.27) 0.70 (0.57,0.89) 0.57 (0.47,0.73)	2097 (F-5) 2.22 (1.73,3.08) 0.99 (0.77,1.38) 0.70 (0.55,0.98) 0.57 (0.45,0.80)
n=1 n=5 n=10 n=15 n=20	2050(B) 1.77 0.82 0.42 0.43 0.29	2050 (F) 1.79 (1.47,2.30) 0.80 (0.66,1.03) 0.57 (0.46,0.73) 0.46 (0.38,0.59) 0.40 (0.33,0.51)	2050 (F-5) 1.69 (1.32,2.36) 0.76 (0.59,1.05) 0.54 (0.42,0.75) 0.44 (0.34,0.61) 0.38 (0.30,0.53)	2097 (B) 2.18 0.80 0.58 0.38 0.34	2097 (F) 2.20 (1.81,2.83) 0.99 (0.81,1.27) 0.70 (0.57,0.89) 0.57 (0.47,0.73) 0.49 (0.40,0.63)	2097 (F-5) 2.22 (1.73,3.08) 0.99 (0.77,1.38) 0.70 (0.55,0.98) 0.57 (0.45,0.80) 0.50 (0.39,0.69)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 1.77 0.82 0.42 0.43 0.29 0.18	2050 (F) 1.79 (1.47,2.30) 0.80 (0.66,1.03) 0.57 (0.46,0.73) 0.46 (0.38,0.59) 0.40 (0.33,0.51) 0.36 (0.29,0.46)	2050 (F-5) 1.69 (1.32,2.36) 0.76 (0.59,1.05) 0.54 (0.42,0.75) 0.44 (0.34,0.61) 0.38 (0.30,0.53) 0.34 (0.26,0.47)	2097 (B) 2.18 0.80 0.58 0.38 0.34 0.25	2097 (F) 2.20 (1.81,2.83) 0.99 (0.81,1.27) 0.70 (0.57,0.89) 0.57 (0.47,0.73) 0.49 (0.40,0.63) 0.44 (0.36,0.57)	2097 (F-5) 2.22 (1.73,3.08) 0.99 (0.77,1.38) 0.70 (0.55,0.98) 0.57 (0.45,0.80) 0.50 (0.39,0.69) 0.44 (0.35,0.62)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 1.77 0.82 0.42 0.43 0.29 0.18 0.16	2050 (F) 1.79 (1.47,2.30) 0.80 (0.66,1.03) 0.57 (0.46,0.73) 0.46 (0.38,0.59) 0.40 (0.33,0.51) 0.36 (0.29,0.46) 0.33 (0.27,0.42)	2050 (F-5) 1.69 (1.32,2.36) 0.76 (0.59,1.05) 0.54 (0.42,0.75) 0.44 (0.34,0.61) 0.38 (0.30,0.53) 0.34 (0.26,0.47) 0.31 (0.24,0.43)	2097 (B) 2.18 0.80 0.58 0.38 0.34 0.25 0.22	2097 (F) 2.20 (1.81,2.83) 0.99 (0.81,1.27) 0.70 (0.57,0.89) 0.57 (0.47,0.73) 0.49 (0.40,0.63) 0.44 (0.36,0.57) 0.40 (0.33,0.52)	2097 (F-5) 2.22 (1.73,3.08) 0.99 (0.77,1.38) 0.70 (0.55,0.98) 0.57 (0.45,0.80) 0.50 (0.39,0.69) 0.44 (0.35,0.62) 0.40 (0.32,0.56)

**Table A4.** Global mean of Rx1Day as simulated by the CESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 35, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.

Earth System Dynamics Discussions

https://doi.org/10.5194/esd-2021-53 Preprint. Discussion started: 6 July 2021 © Author(s) 2021. CC BY 4.0 License.



	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.14	0.14 (0.12,0.17)	0.12 (0.10,0.17)	0.15	0.16 (0.13,0.19)	0.12 (0.09,0.17)
n=5	0.06	0.06 (0.05,0.08)	0.05 (0.04,0.08)	0.07	0.07 (0.06,0.09)	0.05 (0.04,0.07)
n=10	0.04	0.04 (0.04,0.05)	0.04 (0.03,0.05)	0.04	0.05 (0.04,0.06)	0.04 (0.03,0.05)
n=15	0.04	0.04 (0.03,0.04)	0.03 (0.02,0.04)	0.03	0.04 (0.03,0.05)	0.03 (0.02,0.04)
n=20	0.03	0.03 (0.03,0.04)	0.03 (0.02,0.04)	0.03	0.03 (0.03,0.04)	0.03 (0.02,0.04)
n=25	0.02	0.03 (0.02,0.03)	0.02 (0.02,0.03)	0.02	0.03 (0.03,0.04)	0.02 (0.02,0.03)
n=30	0.01	0.03 (0.02,0.03)	0.02 (0.02,0.03)	0.02	0.03 (0.02,0.04)	0.02 (0.02,0.03)
n=35	0.01	0.02 (0.02,0.03)	0.02 (0.02,0.03)	0.01	0.03 (0.02,0.03)	0.02 (0.02,0.03)
n=40	0.01	0.02 (0.02,0.03)	0.02 (0.02,0.03)	0.01	0.02 (0.02,0.03)	0.02 (0.01,0.03)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.13	2050 (F) 0.14 (0.11,0.17)	2050 (F-5) 0.13 (0.10,0.18)	2097 (B) 0.12	2097 (F) 0.12 (0.10,0.15)	2097 (F-5) 0.14 (0.11,0.20)
n=1 n=5	2050(B) 0.13 0.06	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08)	2097 (B) 0.12 0.06	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09)
n=1 n=5 n=10	2050(B) 0.13 0.06 0.04	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08) 0.04 (0.03,0.06)	2097 (B) 0.12 0.06 0.04	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06)
n=1 n=5 n=10 n=15	2050(B) 0.13 0.06 0.04 0.03	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.04 (0.03,0.04)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05)	2097 (B) 0.12 0.06 0.04 0.03	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.13 0.06 0.04 0.03 0.03	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.04 (0.03,0.04) 0.03 (0.03,0.04)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04)	2097 (B) 0.12 0.06 0.04 0.03 0.02	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.13 0.06 0.04 0.03 0.03 0.02	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.04 (0.03,0.04) 0.03 (0.03,0.04) 0.03 (0.02,0.03)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)	2097 (B) 0.12 0.06 0.04 0.03 0.02 0.02	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.13 0.06 0.04 0.03 0.03 0.03 0.02 0.02	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.04 (0.03,0.04) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04) 0.02 (0.02,0.03)	2097 (B) 0.12 0.06 0.04 0.03 0.02 0.02 0.01	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25 n=30 n=35	2050(B) 0.13 0.06 0.04 0.03 0.03 0.02 0.02 0.01	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.04 (0.03,0.04) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.13 (0.10,0.18) 0.06 (0.04,0.08) 0.04 (0.03,0.06) 0.03 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2097 (B) 0.12 0.06 0.04 0.03 0.02 0.02 0.01 0.01	2097 (F) 0.12 (0.10,0.15) 0.05 (0.04,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.02 (0.02,0.03)

**Table A5.** Global mean of TNx as simulated by the CanESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 40, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.



Earth System	Ope		
Dynamics	n Acc	EGU	
Discussions	ess		

	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.29	0.29 (0.24,0.36)	0.35 (0.27,0.49)	0.26	0.27 (0.22,0.33)	0.26 (0.20,0.36)
n=5	0.1	0.13 (0.11,0.16)	0.16 (0.12,0.22)	0.11	0.12 (0.10,0.15)	0.12 (0.09,0.16)
n=10	0.09	0.09 (0.08,0.12)	0.11 (0.09,0.15)	0.07	0.08 (0.07,0.10)	0.08 (0.06,0.11)
n=15	0.06	0.08 (0.06,0.09)	0.09 (0.07,0.13)	0.05	0.07 (0.06,0.09)	0.07 (0.05,0.09)
n=20	0.05	0.07 (0.05,0.08)	0.08 (0.06,0.11)	0.04	0.06 (0.05,0.07)	0.06 (0.05,0.08)
n=25	0.04	0.06 (0.05,0.07)	0.07 (0.05,0.10)	0.04	0.05 (0.04,0.07)	0.05 (0.04,0.07)
n=30	0.03	0.05 (0.04,0.07)	0.06 (0.05,0.09)	0.03	0.05 (0.04,0.06)	0.05 (0.04,0.07)
n=35	0.03	0.05 (0.04,0.06)	0.06 (0.05,0.08)	0.02	0.05 (0.04,0.06)	0.04 (0.03,0.06)
n=40	0.02	0.05 (0.04,0.06)	0.06 (0.04,0.08)	0.02	0.04 (0.04,0.05)	0.04 (0.03,0.06)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	0.20					
	0.29	0.29 (0.25,0.37)	0.30 (0.23,0.41)	0.33	0.34 (0.28,0.42)	0.31 (0.24,0.44)
n=5	0.29	0.29 (0.25,0.37) 0.13 (0.11,0.16)	0.30 (0.23,0.41) 0.13 (0.10,0.18)	0.33 0.17	0.34 (0.28,0.42) 0.15 (0.13,0.19)	0.31 (0.24,0.44) 0.14 (0.11,0.19)
n=5 n=10	0.29	0.29 (0.25,0.37) 0.13 (0.11,0.16) 0.09 (0.08,0.12)	0.30 (0.23,0.41) 0.13 (0.10,0.18) 0.09 (0.07,0.13)	0.33 0.17 0.10	0.34 (0.28,0.42) 0.15 (0.13,0.19) 0.11 (0.09,0.13)	0.31 (0.24,0.44) 0.14 (0.11,0.19) 0.10 (0.08,0.14)
n=5 n=10 n=15	0.29 0.13 0.08 0.05	0.29 (0.25,0.37) 0.13 (0.11,0.16) 0.09 (0.08,0.12) 0.08 (0.06,0.09)	0.30 (0.23,0.41) 0.13 (0.10,0.18) 0.09 (0.07,0.13) 0.08 (0.06,0.11)	0.33 0.17 0.10 0.08	0.34 (0.28,0.42) 0.15 (0.13,0.19) 0.11 (0.09,0.13) 0.09 (0.07,0.11)	0.31 (0.24,0.44) 0.14 (0.11,0.19) 0.10 (0.08,0.14) 0.08 (0.06,0.11)
n=5 n=10 n=15 n=20	0.29 0.13 0.08 0.05 0.05	0.29 (0.25,0.37) 0.13 (0.11,0.16) 0.09 (0.08,0.12) 0.08 (0.06,0.09) 0.07 (0.05,0.08)	0.30 (0.23,0.41) 0.13 (0.10,0.18) 0.09 (0.07,0.13) 0.08 (0.06,0.11) 0.07 (0.05,0.09)	0.33 0.17 0.10 0.08 0.06	0.34 (0.28,0.42) 0.15 (0.13,0.19) 0.11 (0.09,0.13) 0.09 (0.07,0.11) 0.08 (0.06,0.09)	0.31 (0.24,0.44) 0.14 (0.11,0.19) 0.10 (0.08,0.14) 0.08 (0.06,0.11) 0.07 (0.05,0.10)
n=5 n=10 n=15 n=20 n=25	0.29 0.13 0.08 0.05 0.05 0.04	0.29 (0.25,0.37) 0.13 (0.11,0.16) 0.09 (0.08,0.12) 0.08 (0.06,0.09) 0.07 (0.05,0.08) 0.06 (0.05,0.07)	0.30 (0.23,0.41) 0.13 (0.10,0.18) 0.09 (0.07,0.13) 0.08 (0.06,0.11) 0.07 (0.05,0.09) 0.06 (0.05,0.08)	0.33 0.17 0.10 0.08 0.06 0.05	0.34 (0.28,0.42) 0.15 (0.13,0.19) 0.11 (0.09,0.13) 0.09 (0.07,0.11) 0.08 (0.06,0.09) 0.07 (0.06,0.08)	0.31 (0.24,0.44) 0.14 (0.11,0.19) 0.10 (0.08,0.14) 0.08 (0.06,0.11) 0.07 (0.05,0.10) 0.06 (0.05,0.09)
n=5 n=10 n=15 n=20 n=25 n=30	0.29 0.13 0.08 0.05 0.05 0.04 0.04	0.29 (0.25,0.37) 0.13 (0.11,0.16) 0.09 (0.08,0.12) 0.08 (0.06,0.09) 0.07 (0.05,0.08) 0.06 (0.05,0.07) 0.05 (0.04,0.07)	0.30 (0.23,0.41) 0.13 (0.10,0.18) 0.09 (0.07,0.13) 0.08 (0.06,0.11) 0.07 (0.05,0.09) 0.06 (0.05,0.08) 0.05 (0.04,0.07)	0.33 0.17 0.10 0.08 0.06 0.05 0.04	0.34 (0.28,0.42) 0.15 (0.13,0.19) 0.11 (0.09,0.13) 0.09 (0.07,0.11) 0.08 (0.06,0.09) 0.07 (0.06,0.08) 0.06 (0.05,0.08)	0.31 (0.24,0.44) 0.14 (0.11,0.19) 0.10 (0.08,0.14) 0.08 (0.06,0.11) 0.07 (0.05,0.10) 0.06 (0.05,0.09) 0.06 (0.04,0.08)
n=5 n=10 n=15 n=20 n=25 n=30 n=35	0.29 0.13 0.08 0.05 0.05 0.04 0.04 0.03	0.29 (0.25,0.37) 0.13 (0.11,0.16) 0.09 (0.08,0.12) 0.08 (0.06,0.09) 0.07 (0.05,0.08) 0.06 (0.05,0.07) 0.05 (0.04,0.07) 0.05 (0.04,0.06)	0.30 (0.23,0.41) 0.13 (0.10,0.18) 0.09 (0.07,0.13) 0.08 (0.06,0.11) 0.07 (0.05,0.09) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.07)	0.33 0.17 0.10 0.08 0.06 0.05 0.04 0.03	0.34 (0.28,0.42) 0.15 (0.13,0.19) 0.11 (0.09,0.13) 0.09 (0.07,0.11) 0.08 (0.06,0.09) 0.07 (0.06,0.08) 0.06 (0.05,0.08) 0.06 (0.05,0.07)	0.31 (0.24,0.44) 0.14 (0.11,0.19) 0.10 (0.08,0.14) 0.08 (0.06,0.11) 0.07 (0.05,0.10) 0.06 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07)

**Table A6.** Global mean of Rx5Day as simulated by the CanESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 40, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.

Earth System

https://doi.org/10.5194/esd-2021-53 Preprint. Discussion started: 6 July 2021 © Author(s) 2021. CC BY 4.0 License.



	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.20	0.20 (0.17,0.25)	0.20 (0.15,0.28)	0.20	0.21 (0.17,0.26)	0.22 (0.17,0.30)
n=5	0.09	0.09 (0.07,0.11)	0.09 (0.07,0.12)	0.10	0.09 (0.08,0.12)	0.10 (0.08,0.13)
n=10	0.06	0.06 (0.05,0.08)	0.06 (0.05,0.09)	0.05	0.07 (0.05,0.08)	0.07 (0.05,0.10)
n=15	0.05	0.05 (0.04,0.06)	0.05 (0.04,0.07)	0.04	0.05 (0.04,0.07)	0.06 (0.04,0.08)
n=20	0.04	0.04 (0.04,0.06)	0.04 (0.03,0.06)	0.04	0.05 (0.04,0.06)	0.05 (0.04,0.07)
n=25	0.03	0.04 (0.03,0.05)	0.04 (0.03,0.06)	0.03	0.04 (0.03,0.05)	0.04 (0.03,0.06)
n=30	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.05)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.06)
n=35	0.02	0.03 (0.03,0.04)	0.03 (0.03,0.05)	0.02	0.03 (0.03,0.04)	0.04 (0.03,0.05)
n=40	0.01	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.01	0.03 (0.03,0.04)	0.03 (0.03,0.05)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.22	2050 (F) 0.22 (0.19,0.28)	2050 (F-5) 0.22 (0.17,0.30)	2097 (B) 0.21	2097 (F) 0.22 (0.18,0.27)	2097 (F-5) 0.23 (0.19,0.32)
n=1 n=5	2050(B) 0.22 0.1	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13)	2097 (B) 0.21 0.08	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15)
n=1 n=5 n=10	2050(B) 0.22 0.1 0.06	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12) 0.07 (0.06,0.09)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13) 0.07 (0.05,0.09)	2097 (B) 0.21 0.08 0.07	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12) 0.07 (0.06,0.09)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15) 0.07 (0.06,0.10)
n=1 n=5 n=10 n=15	2050(B) 0.22 0.1 0.06 0.04	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08)	2097 (B) 0.21 0.08 0.07 0.04	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15) 0.07 (0.06,0.10) 0.06 (0.05,0.08)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.22 0.1 0.06 0.04 0.04	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07)	2097 (B) 0.21 0.08 0.07 0.04 0.03	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15) 0.07 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.22 0.1 0.06 0.04 0.04 0.02	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.06)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06)	2097 (B) 0.21 0.08 0.07 0.04 0.03 0.03	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.05)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15) 0.07 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.22 0.1 0.06 0.04 0.04 0.02 0.03	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.06) 0.04 (0.03,0.05)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2097 (B) 0.21 0.08 0.07 0.04 0.03 0.03 0.02	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.05) 0.04 (0.03,0.05)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15) 0.07 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06) 0.04 (0.03,0.06)
n=1 n=5 n=10 n=15 n=20 n=25 n=30 n=35	2050(B) 0.22 0.1 0.06 0.04 0.04 0.02 0.03 0.02	2050 (F) 0.22 (0.19,0.28) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.06) 0.04 (0.03,0.05) 0.04 (0.03,0.05)	2050 (F-5) 0.22 (0.17,0.30) 0.1 (0.08,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2097 (B) 0.21 0.08 0.07 0.04 0.03 0.03 0.02 0.02	2097 (F) 0.22 (0.18,0.27) 0.1 (0.08,0.12) 0.07 (0.06,0.09) 0.06 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.05) 0.04 (0.03,0.05)	2097 (F-5) 0.23 (0.19,0.32) 0.1 (0.08,0.15) 0.07 (0.06,0.10) 0.06 (0.05,0.08) 0.05 (0.04,0.07) 0.05 (0.04,0.06) 0.04 (0.03,0.06)

**Table A7.** Global mean of TNn as simulated by the CanESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 40, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.

Earth System Dynamics Discussions

https://doi.org/10.5194/esd-2021-53 Preprint. Discussion started: 6 July 2021 © Author(s) 2021. CC BY 4.0 License.



	10.50 (2)	10 <b>50</b> (D)	10.50 (7) 5)		• • • • • • •	••••• ••
	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.16	0.16 (0.13,0.20)	0.14 (0.11,0.19)	0.17	0.17 (0.14,0.22)	0.14 (0.11,0.19)
n=5	0.07	0.07 (0.06,0.09)	0.06 (0.05,0.09)	0.07	0.08 (0.06,0.10)	0.06 (0.05,0.08)
n=10	0.05	0.05 (0.04,0.06)	0.04 (0.03,0.06)	0.05	0.05 (0.05,0.07)	0.04 (0.03,0.06)
n=15	0.03	0.04 (0.03,0.05)	0.04 (0.03,0.05)	0.04	0.04 (0.04,0.06)	0.03 (0.03,0.05)
n=20	0.03	0.04 (0.03,0.04)	0.03 (0.02,0.04)	0.03	0.04 (0.03,0.05)	0.03 (0.02,0.04)
n=25	0.02	0.03 (0.03,0.04)	0.03 (0.02,0.04)	0.03	0.03 (0.03,0.04)	0.03 (0.02,0.04)
n=30	0.02	0.03 (0.02,0.04)	0.03 (0.02,0.04)	0.02	0.03 (0.03,0.04)	0.02 (0.02,0.03)
n=35	0.01	0.03 (0.02,0.03)	0.02 (0.02,0.03)	0.01	0.03 (0.02,0.04)	0.02 (0.02,0.03)
n=40	0.01	0.02 (0.02,0.03)	0.02 (0.02,0.03)	0.01	0.03 (0.02,0.03)	0.02 (0.02,0.03)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.13	2050 (F) 0.14 (0.11,0.17)	2050 (F-5) 0.14 (0.11,0.20)	2097 (B) 0.13	2097 (F) 0.13 (0.11,0.16)	2097 (F-5) 0.14 (0.11,0.20)
n=1 n=5	2050(B) 0.13 0.06	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09)	2097 (B) 0.13 0.06	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09)
n=1 n=5 n=10	2050(B) 0.13 0.06 0.03	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.04 (0.03,0.06)	2097 (B) 0.13 0.06 0.04	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07) 0.04 (0.03,0.05)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06)
n=1 n=5 n=10 n=15	2050(B) 0.13 0.06 0.03 0.03	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.03 (0.03,0.04)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2097 (B) 0.13 0.06 0.04 0.03	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.13 0.06 0.03 0.03 0.03	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.03 (0.03,0.04) 0.03 (0.03,0.04)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04)	2097 (B) 0.13 0.06 0.04 0.03 0.02	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.13 0.06 0.03 0.03 0.03 0.03 0.02	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.03 (0.03,0.04) 0.03 (0.03,0.04) 0.03 (0.02,0.03)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)	2097 (B) 0.13 0.06 0.04 0.03 0.02 0.02	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.03 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.13 0.06 0.03 0.03 0.03 0.02 0.01	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.03 (0.03,0.04) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04)	2097 (B) 0.13 0.06 0.04 0.03 0.02 0.02 0.01	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.03 (0.02,0.03) 0.02 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04)
n=1 n=5 n=10 n=15 n=20 n=25 n=30 n=35	2050(B) 0.13 0.06 0.03 0.03 0.03 0.02 0.01 0.01	2050 (F) 0.14 (0.11,0.17) 0.06 (0.05,0.08) 0.04 (0.04,0.05) 0.03 (0.03,0.04) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2050 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.02 (0.02,0.03)	2097 (B) 0.13 0.06 0.04 0.03 0.02 0.02 0.01 0.01	2097 (F) 0.13 (0.11,0.16) 0.06 (0.05,0.07) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.03) 0.03 (0.02,0.03) 0.02 (0.02,0.03) 0.02 (0.02,0.03)	2097 (F-5) 0.14 (0.11,0.20) 0.06 (0.05,0.09) 0.05 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04) 0.03 (0.02,0.04) 0.03 (0.02,0.04) 0.02 (0.02,0.03)

**Table A8.** Global mean of TXx as simulated by the CanESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 40, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.





	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.2	0.20 (0.17,0.25)	0.19 (0.15,0.27)	0.20	0.20 (0.17,0.25)	0.21 (0.17,0.30)
n=5	0.09	0.09 (0.07,0.11)	0.09 (0.07,0.12)	0.07	0.09 (0.07,0.11)	0.10 (0.07,0.13)
n=10	0.05	0.06 (0.05,0.08)	0.06 (0.05,0.08)	0.04	0.06 (0.05,0.08)	0.07 (0.05,0.09)
n=15	0.04	0.05 (0.04,0.06)	0.05 (0.04,0.07)	0.03	0.05 (0.04,0.06)	0.06 (0.04,0.08)
n=20	0.03	0.04 (0.04,0.06)	0.04 (0.03,0.06)	0.03	0.04 (0.04,0.06)	0.05 (0.04,0.07)
n=25	0.03	0.04 (0.03,0.05)	0.04 (0.03,0.05)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.06)
n=30	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.05)	0.02	0.04 (0.03,0.05)	0.04 (0.03,0.05)
n=35	0.02	0.03 (0.03,0.04)	0.03 (0.03,0.05)	0.02	0.03 (0.03,0.04)	0.04 (0.03,0.05)
n=40	0.01	0.03 (0.03,0.04)	0.03 (0.03,0.04)	0.01	0.03 (0.03,0.04)	0.03 (0.03,0.05)
	2050(D)	<b>2050</b> (T)				
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.21	0.21 (0.18,0.26)	2050 (F-5) 0.21 (0.17,0.30)	2097 (B) 0.20	2097 (F) 0.20 (0.17,0.25)	2097 (F-5) 0.21 (0.17,0.30)
n=1 n=5	0.21 0.08	0.21 (0.18,0.26) 0.09 (0.08,0.12)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13)	2097 (B) 0.20 0.1	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13)
n=1 n=5 n=10	0.21 0.08 0.05	0.21 (0.18,0.26) 0.09 (0.08,0.12) 0.07 (0.06,0.08)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09)	2097 (B) 0.20 0.1 0.06	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11) 0.06 (0.05,0.08)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09)
n=1 n=5 n=10 n=15	0.21 0.08 0.05 0.05	0.21 (0.18,0.26) 0.09 (0.08,0.12) 0.07 (0.06,0.08) 0.05 (0.05,0.07)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08)	2097 (B) 0.20 0.1 0.06 0.05	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11) 0.06 (0.05,0.08) 0.05 (0.04,0.06)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.05 (0.04,0.08)
n=1 n=5 n=10 n=15 n=20	2030(B) 0.21 0.08 0.05 0.05 0.04	0.21 (0.18,0.26) 0.09 (0.08,0.12) 0.07 (0.06,0.08) 0.05 (0.05,0.07) 0.05 (0.04,0.06)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07)	2097 (B) 0.20 0.1 0.06 0.05 0.03	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11) 0.06 (0.05,0.08) 0.05 (0.04,0.06) 0.04 (0.04,0.06)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.05 (0.04,0.08) 0.05 (0.04,0.07)
n=1 n=5 n=10 n=15 n=20 n=25	2030(B) 0.21 0.08 0.05 0.05 0.04 0.03	0.21 (0.18,0.26) 0.09 (0.08,0.12) 0.07 (0.06,0.08) 0.05 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.05)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06)	2097 (B) 0.20 0.1 0.06 0.05 0.03 0.02	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11) 0.06 (0.05,0.08) 0.05 (0.04,0.06) 0.04 (0.04,0.06) 0.04 (0.03,0.05)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.05 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2030(B) 0.21 0.08 0.05 0.05 0.04 0.03 0.02	2050 (F)         0.21 (0.18,0.26)         0.09 (0.08,0.12)         0.07 (0.06,0.08)         0.05 (0.05,0.07)         0.05 (0.04,0.06)         0.04 (0.04,0.05)         0.04 (0.03,0.05)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)	2097 (B) 0.20 0.11 0.06 0.05 0.03 0.02 0.02	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11) 0.06 (0.05,0.08) 0.05 (0.04,0.06) 0.04 (0.04,0.06) 0.04 (0.03,0.05) 0.04 (0.03,0.05)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.05 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05)
n=1 n=5 n=10 n=15 n=20 n=25 n=30 n=35	2030(B) 0.21 0.08 0.05 0.05 0.04 0.03 0.02 0.02	0.21 (0.18,0.26) 0.09 (0.08,0.12) 0.07 (0.06,0.08) 0.05 (0.05,0.07) 0.05 (0.04,0.06) 0.04 (0.04,0.05) 0.04 (0.03,0.05) 0.04 (0.03,0.04)	2050 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.06 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.04 (0.03,0.05)	2097 (B) 0.20 0.1 0.06 0.05 0.03 0.02 0.02 0.02	2097 (F) 0.20 (0.17,0.25) 0.09 (0.07,0.11) 0.06 (0.05,0.08) 0.05 (0.04,0.06) 0.04 (0.04,0.06) 0.04 (0.03,0.05) 0.03 (0.03,0.04)	2097 (F-5) 0.21 (0.17,0.30) 0.1 (0.07,0.13) 0.07 (0.05,0.09) 0.05 (0.04,0.08) 0.05 (0.04,0.07) 0.04 (0.03,0.06) 0.04 (0.03,0.05) 0.04 (0.03,0.05)

**Table A9.** Global mean of TXn as simulated by the CanESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 40, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.





	1953 (B)	1953 (F)	1953 (F-5)	2000 (B)	2000 (F)	2000 (F-5)
n=1	0.89	0.90 (0.75,1.12)	0.68 (0.53,0.94)	0.86	0.87 (0.73,1.08)	0.68 (0.53,0.95)
n=5	0.38	0.40 (0.33,0.50)	0.30 (0.24,0.42)	0.44	0.39 (0.32,0.48)	0.31 (0.24,0.43)
n=10	0.25	0.28 (0.24,0.35)	0.21 (0.17,0.30)	0.26	0.28 (0.23,0.34)	0.22 (0.17,0.30)
n=15	0.2	0.23 (0.19,0.29)	0.17 (0.14,0.24)	0.21	0.22 (0.19,0.28)	0.18 (0.14,0.25)
n=20	0.17	0.20 (0.17,0.25)	0.15 (0.12,0.21)	0.15	0.19 (0.16,0.24)	0.15 (0.12,0.21)
n=25	0.12	0.18 (0.15,0.22)	0.14 (0.11,0.19)	0.11	0.17 (0.15,0.22)	0.14 (0.11,0.19)
n=30	0.1	0.16 (0.14,0.20)	0.12 (0.10,0.17)	0.10	0.16 (0.13,0.20)	0.12 (0.10,0.17)
n=35	0.08	0.15 (0.13,0.19)	0.11 (0.09,0.16)	0.08	0.15 (0.12,0.18)	0.12 (0.09,0.16)
n=40	0.07	0.14 (0.12,0.18)	0.11 (0.08,0.15)	0.05	0.14 (0.11,0.17)	0.11 (0.11,0.17)
	2050(B)	2050 (F)	2050 (F-5)	2097 (B)	2097 (F)	2097 (F-5)
n=1	2050(B) 0.98	2050 (F) 0.99 (0.82,1.23)	2050 (F-5) 0.97 (0.76,1.35)	2097 (B) 0.97	2097 (F) 0.98 (0.82,1.22)	2097 (F-5) 0.87 (0.68,1.21)
n=1 n=5	2050(B) 0.98 0.4	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60)	2097 (B) 0.97 0.49	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54)
n=1 n=5 n=10	2050(B) 0.98 0.4 0.28	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55) 0.31 (0.26,0.39)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60) 0.31 (0.24,0.43)	2097 (B) 0.97 0.49 0.27	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55) 0.31 (0.26,0.39)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54) 0.27 (0.21,0.38)
n=1 n=5 n=10 n=15	2050(B) 0.98 0.4 0.28 0.2	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60) 0.31 (0.24,0.43) 0.25 (0.20,0.35)	2097 (B) 0.97 0.49 0.27 0.18	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54) 0.27 (0.21,0.38) 0.22 (0.17,0.31)
n=1 n=5 n=10 n=15 n=20	2050(B) 0.98 0.4 0.28 0.2 0.2	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.28)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60) 0.31 (0.24,0.43) 0.25 (0.20,0.35) 0.22 (0.17,0.30)	2097 (B) 0.97 0.49 0.27 0.18 0.13	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.27)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54) 0.27 (0.21,0.38) 0.22 (0.17,0.31) 0.19 (0.15,0.27)
n=1 n=5 n=10 n=15 n=20 n=25	2050(B) 0.98 0.4 0.28 0.2 0.2 0.2 0.15	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.28) 0.2 (0.16,0.25)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60) 0.31 (0.24,0.43) 0.25 (0.20,0.35) 0.22 (0.17,0.30) 0.19 (0.15,0.27)	2097 (B) 0.97 0.49 0.27 0.18 0.13 0.13	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.27) 0.2 (0.16,0.24)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54) 0.27 (0.21,0.38) 0.22 (0.17,0.31) 0.19 (0.15,0.27) 0.17 (0.14,0.24)
n=1 n=5 n=10 n=15 n=20 n=25 n=30	2050(B) 0.98 0.4 0.28 0.2 0.2 0.2 0.15 0.11	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.28) 0.2 (0.16,0.25) 0.18 (0.15,0.22)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60) 0.31 (0.24,0.43) 0.25 (0.20,0.35) 0.22 (0.17,0.30) 0.19 (0.15,0.27) 0.18 (0.14,0.25)	2097 (B) 0.97 0.49 0.27 0.18 0.13 0.13 0.13	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.27) 0.2 (0.16,0.24) 0.18 (0.15,0.22)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54) 0.27 (0.21,0.38) 0.22 (0.17,0.31) 0.19 (0.15,0.27) 0.17 (0.14,0.24) 0.16 (0.12,0.22)
n=1 n=5 n=10 n=15 n=20 n=25 n=30 n=35	2050(B) 0.98 0.4 0.28 0.2 0.2 0.15 0.11 0.1	2050 (F) 0.99 (0.82,1.23) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.28) 0.2 (0.16,0.25) 0.18 (0.15,0.22) 0.17 (0.14,0.21)	2050 (F-5) 0.97 (0.76,1.35) 0.43 (0.34,0.60) 0.31 (0.24,0.43) 0.25 (0.20,0.35) 0.22 (0.17,0.30) 0.19 (0.15,0.27) 0.18 (0.14,0.25) 0.16 (0.13,0.23)	2097 (B) 0.97 0.49 0.27 0.18 0.13 0.13 0.1 0.1	2097 (F) 0.98 (0.82,1.22) 0.44 (0.37,0.55) 0.31 (0.26,0.39) 0.25 (0.21,0.32) 0.22 (0.18,0.27) 0.2 (0.16,0.24) 0.18 (0.15,0.22) 0.17 (0.14,0.21)	2097 (F-5) 0.87 (0.68,1.21) 0.39 (0.30,0.54) 0.27 (0.21,0.38) 0.22 (0.17,0.31) 0.19 (0.15,0.27) 0.17 (0.14,0.24) 0.16 (0.12,0.22) 0.15 (0.11,0.20)

**Table A10.** Global mean of Rx1Day as simulated by the CanESM ensemble: Values of the RMSE in approximating the full ensemble mean by the individual runs (first row, n = 1), and by ensembles of increasingly larger sizes (from 5 to 40, along the remaining rows). The estimates obtained by the bootstrap approach (columns labeled by "(B)") are compared to the estimates obtained by the formula  $\sigma/\sqrt{n}$  where  $\sigma$  is estimated as the ensemble standard deviation using all ensemble members (columns labelled by "(F)", where also the 95% conf. int. is shown). We also compare estimates derived by plugging into the formula a value of  $\sigma$  estimated by a subset of 5 ensemble members, and 5 years around the year considered (columns labelled by "(F-5)"). Results are shown for four individual years along the simulation (column-wise), since  $\sigma$  varies along it.





Ens. size	1	5	10	15	20	25	30	35
TNx Global midC	4.74	8.27	5.52	4.81	2.65	2.91	0.87	0.07
TNx Global endC	3.07	6.49	6.34	3.57	3.73	2.38	1.02	0.08
TNx Land midC	3.8	11.05	7.14	4.3	3.11	2.95	0.96	0.14
TNx Land endC	3.88	9.01	6.35	3.34	3.06	1.99	0.79	0.06
TNx Ocean midC	5.13	7.13	4.85	5.01	2.46	2.89	0.83	0.05
TNx Ocean endC	2.73	5.45	6.33	3.67	4	2.55	1.11	0.08
TNn Global midC	11.98	10.18	8.66	5.18	3.68	2.67	1.21	0.2
TNn Global endC	11.4	8.11	8.42	6.03	3.58	3.57	0.53	0.11
TNn Land midC	13.03	10.05	8.98	5.41	2.91	4	0.73	0.18
TNn Land endC	11.99	9.92	10.02	7.48	4.83	6.45	0.93	0.21
TNn Ocean midC	11.55	10.23	8.53	5.08	3.99	2.12	1.41	0.21
TNn Ocean endC	11.15	7.36	7.77	5.44	3.07	2.39	0.36	0.08
TXx Global midC	14.99	7.78	8.65	4.9	5.88	3.58	1.06	0.2
TXx Global endC	12.28	8.98	7.92	5.3	4.98	1.72	1.11	0.29
TXx Land midC	13.48	8.4	6.68	5.27	4.56	2.18	1.49	0.14
TXx Land endC	14.08	11.64	7.76	4.84	4.7	2.19	1.37	0.29
TXx Ocean midC	15.6	7.52	9.46	4.75	6.42	4.16	0.88	0.23
TXx Ocean endC	11.54	7.89	7.98	5.49	5.1	1.53	1	0.3
TXn Global midC	11.1	10.84	7.59	5.14	5.63	2.68	0.92	0.38
TXn Global endC	10.33	11.1	7.27	4.34	3.66	2.02	0.66	0.34
TXn Land midC	12.16	9.22	9.3	4.84	3.52	2.39	0.85	0.26
TXn Land endC	11.41	12.73	8.94	5.75	4.3	3.12	0.79	0.4
TXn Ocean midC	10.67	11.51	6.89	5.26	6.5	2.79	0.94	0.44
TXn Ocean endC	9.89	10.44	6.58	3.76	3.4	1.57	0.61	0.32

# Appendix B: Summary of error ratio patterns as shown in Figures 3 and 4 for all metrics and models

**Table B1.** Percentage of the global, land or ocean surface where the actual errors exceed the errors estimated on the basis of the formula "a-priori" using 5 ensemble members to estimate  $\sigma$ . Results for all temperature extreme metrics, derived from the CESM ensemble whose full size is 40 members. Calculations apply cosine-of-latitude weighting. Results for TNx are summaries of the behavior shown in Figure 3, i.e., the fraction of surface represented by locations where the error ratio is larger than 100%. Numbers under small *n*'s are affected by noise, as we randomly choose *n* members from the full ensemble, only once. As can be gauged, the decreasing behavior of the fractions stabilizes for  $n \ge 15$ .





Ens. size	1	5	10	15	20	25	30	35	40	45
TNx Global midC	15.89	7.38	13.35	8.32	4.77	4.59	2.77	1.76	1.48	0.64
TNx Global endC	26.37	10.09	13.73	9.21	6.7	4.88	4.76	1.84	1.45	0.74
TNx Land midC	12.82	7.68	10.36	9.16	4.85	5.6	2.94	1.78	0.59	0.38
TNx Land endC	18.52	9.9	10.65	8.94	6.61	5.33	4.91	2.3	1.07	0.49
TNx Ocean midC	17.23	7.25	14.65	7.96	4.73	4.15	2.7	1.76	1.87	0.76
TNx Ocean endC	29.79	10.17	15.06	9.32	6.74	4.68	4.69	1.64	1.62	0.85
TNn Global midC	11.91	12.12	8.79	6.45	5.06	3.43	2.76	1.72	0.73	0.04
TNn Global endC	9.64	14.05	11.08	11.77	7.56	6.37	5.08	3.46	1.07	0.28
TNn Land midC	11.85	10.39	8.38	7.17	4.62	4.24	2.96	1.78	0.59	0.02
TNn Land endC	10.76	13.02	10.01	10.14	6.5	6.75	4.15	3.11	1.41	0.48
TNn Ocean midC	11.93	12.87	8.96	6.14	5.25	3.07	2.67	1.69	0.79	0.05
TNn Ocean endC	9.15	14.5	11.55	12.48	8.03	6.2	5.49	3.61	0.92	0.19
TXx Global midC	9.36	10.3	7.81	6.87	5.54	4.45	3.58	1.34	0.67	0.15
TXx Global endC	12.06	11.61	7.99	8.57	5.72	4.14	3.54	1.6	0.4	0.05
TXx Land midC	11.13	10.19	8.14	5.19	6.1	3.75	3.01	2.09	0.45	0.13
TXx Land endC	15.21	10.16	9.41	9.27	6.44	4.05	3.47	1.76	0.51	0
TXx Ocean midC	8.59	10.35	7.67	7.6	5.3	4.76	3.82	1.02	0.77	0.16
TXx Ocean endC	10.7	12.24	7.37	8.26	5.4	4.18	3.57	1.53	0.35	0.07
TXn Global midC	10.59	8.66	9.07	5.73	5.45	3.34	1.8	1.04	0.72	0.06
TXn Global endC	10.96	10.74	9.05	7.22	4.29	3.95	2.29	1.93	0.42	0.13
TXn Land midC	11.76	10.32	10.95	6.62	5.45	4.15	2.06	1.19	0.69	0
TXn Land endC	12.22	11.2	8.46	6.66	3.88	2.89	2.68	1.46	0.55	0.04
TXn Ocean midC	10.07	7.94	8.26	5.34	5.45	2.99	1.69	0.97	0.74	0.09
TXn Ocean endC	10.41	10.54	9.31	7.46	4.48	4.41	2.12	2.13	0.36	0.16

**Table B2.** Percentage of the global, land or ocean surface where the actual errors exceed the errors estimated on the basis of the formula "a-priori" using 5 ensemble members to estimate  $\sigma$ . Results for all temperature extreme metrics, derived from the CanESM ensemble whose full size is 50 members. Calculations apply cosine-of-latitude weighting. Numbers under small *n*'s are affected by noise, as we randomly choose *n* members from the full ensemble, only once. As can be gauged, the decreasing behavior of the fractions stabilizes for  $n \ge 15$ .





Ens. size	1	5	10	15	20	25	30	35
Rx5Day Global midC	4.89	9.78	8.05	6.33	5.03	3.26	1.79	0.49
Rx5Day Global endC	5.83	8.84	7.72	6.46	4.58	3.22	1.61	0.39
Rx5Day Land midC	4.81	9.97	8.08	6.26	5.7	3.21	1.58	0.48
Rx5Day Land endC	5.42	8.71	7.16	6.07	4.91	2.61	1.66	0.38
Rx5Day Ocean midC	4.92	9.71	8.04	6.36	4.76	3.28	1.87	0.5
Rx5Day Ocean endC	5.99	8.89	7.95	6.63	4.45	3.48	1.59	0.39
Rx1Day Global midC	12.83	10.71	8.58	6.77	5.33	3.36	2.03	0.49
Rx1Day Global endC	12.45	10.52	7.94	6.22	4.78	3.2	1.73	0.47
Rx1Day Land midC	11.65	11.5	7.72	6.34	5.27	3.14	1.67	0.4
Rx1Day Land endC	11.39	9.82	7.83	5.24	4.6	3.04	1.72	0.29
Rx1Day Ocean midC	13.32	10.39	8.93	6.95	5.35	3.46	2.18	0.53
Rx1Day Ocean endC	12.89	10.8	7.99	6.63	4.86	3.26	1.73	0.54

**Table B3.** Percentage of the global, land or ocean surface where the actual errors exceed the errors estimated on the basis of the formula "a-priori" using 5 ensemble members to estimate  $\sigma$ . Results for the two precipitation extreme metrics, derived from the CESM ensemble whose full size is 40 members. Calculations apply cosine-of-latitude weighting. Results for Rx5Day are summaries of the behavior shown in Figure 4, i.e., the fraction of surface represented by locations where the error ratio is larger than 100%. Numbers under small *n*'s are affected by noise, as we randomly choose *n* members from the full ensemble, only once. As can be gauged, the decreasing behavior of the fractions stabilizes for  $n \ge 15$ .





Ens. size	1	5	10	15	20	25	30	35	40	45
Rx5Day Global midC	12.54	10.76	7.72	7.31	6.97	5.03	4.1	2.09	0.91	0.32
Rx5Day Global endC	12.69	11.43	9.35	8.53	6.99	5.43	3.9	2.35	1.17	0.59
Rx5Day Land midC	13.03	11.12	7.6	7.53	6.63	4.84	4.68	2.23	1.13	0.4
Rx5Day Land endC	11.5	10.98	8.4	8.95	5.89	4.98	3.27	2.16	1.04	0.33
Rx5Day Ocean midC	12.33	10.59	7.77	7.21	7.11	5.12	3.85	2.02	0.82	0.29
Rx5Day Ocean endC	13.21	11.62	9.77	8.34	7.47	5.62	4.17	2.43	1.23	0.71
Rx1Day Global midC	12.52	11.77	8.69	9.11	6.74	5.69	4.09	2.56	1.33	0.28
Rx1Day Global endC	11.56	12.17	9.29	8.64	7.12	4.84	3.65	2.78	1.31	0.23
Rx1Day Land midC	11.43	11.78	8.94	8.58	5.58	5.89	3.71	2.33	1.33	0.15
Rx1Day Land endC	11.8	12.64	10.11	7.09	6.72	4.77	3.33	2.4	1.17	0.19
Rx1Day Ocean midC	13	11.77	8.58	9.35	7.24	5.6	4.25	2.67	1.33	0.34
Rx1Day Ocean endC	11.45	11.97	8.93	9.31	7.3	4.87	3.8	2.94	1.37	0.25

**Table B4.** Percentage of the global, land or ocean surface where the actual errors exceed the errors estimated on the basis of the formula "a-priori" using 5 ensemble members to estimate  $\sigma$ . Results for the two precipitation extreme metrics, derived from the CanESM ensemble whose full size is 50 members. Calculations apply cosine-of-latitude weighting. Numbers under small *n*'s are affected by noise, as we randomly choose *n* members from the full ensemble, only once. As can be gauged, the decreasing behavior of the fractions stabilizes for  $n \ge 15$ .





# 375 Appendix C: Additional Figures

# C1 Forced Component



Figure C1. The fifteen locations at which we extract model output to fit GEV distributions.







Figure C2. Like Figure 1 for the remaining metrics, derived from the CESM ensembles.



Figure C3. As Figure 2 for TNn and Rx1Day. CESM ensemble results.







Figure C4. As Figure 2 for TXn and TXx. CESM ensemble results.







Figure C5. As Figure 2 but using the CanESM ensemble.







Figure C6. As Figure C5 for TNn and Rx1Day and using the CanESM ensemble.







Figure C7. As Figure C5 for TXn and TXx and using the CanESM ensemble.







**Figure C8.** Return Levels for TNn at (row-wise) 2000, 2050 and 2100 for (column-wise) 2-, 5-, 10-, 20-, 50-, 100-year return periods, based on estimating a GEV by using 11-yr windows of data around each date. In each plot, for increasing ensemble sizes along the x-axis (from 5 to the full ensemble, 40), the red dots indicate the central estimate, and the pink envelope represents the 95% confidence interval. The estimates based on the full ensemble, which we consider the truth, are also drawn across the plot for reference, as horizontal lines. The blue dots in each plot show the same quantities estimated simply by counting, i.e., computing the empirical cumulative distribution function of TNn on the basis of the  $n \times 11$  years where n is the ensemble size. The first three rows show results for a location in Australia while the following three rows show results for a location in Northern North America (see Figure C1).







**Figure C9.** Return Levels for TXx at (row-wise) 2000, 2050 and 2100 for (column-wise) 2-, 5-, 10-, 20-, 50-, 100-year return periods, based on estimating a GEV by using 11-yr windows of data around each date. In each plot, for increasing ensemble sizes along the x-axis (from 5 to the full ensemble, 40), the red dots indicate the central estimate, and the pink envelope represents the 95% confidence interval. The estimates based on the full ensemble, which we consider the truth, are also drawn across the plot for reference, as horizontal lines. The blue dots in each plot show the same quantities estimated simply by counting, i.e., computing the empirical cumulative distribution function of TXx on the basis of the  $n \times 11$  years where n is the ensemble size. The first three rows show results for a location in Central South America while the following three rows show results for a location on the Iberian Peninsula (see Figure C1).







Figure C10. Return Levels for TXn at (row-wise) 2000, 2050 and 2100 for (column-wise) 2-, 5-, 10-, 20-, 50-, 100-year return periods, based on estimating a GEV by using 11-yr windows of data around each date. In each plot, for increasing ensemble sizes along the x-axis (from 5 to the full ensemble, 40), the red dots indicate the central estimate, and the pink envelope represents the 95% confidence interval. The estimates based on the full ensemble, which we consider the truth, are also drawn across the plot for reference, as horizontal lines. The blue dots in each plot show the same quantities estimated simply by counting, i.e., computing the empirical cumulative distribution function of TXn on the basis of the  $n \times 11$  years where n is the ensemble size. The first three rows show results for a location in Northern South America while the following three rows show results for a location on the Maritime Continent (see Figure C1).







**Figure C11.** Return Levels for Rx1Day at (row-wise) 2000, 2050 and 2100 for (column-wise) 2-, 5-, 10-, 20-, 50-, 100-year return periods, based on estimating a GEV by using 11-yr windows of data around each date. In each plot, for increasing ensemble sizes along the x-axis (from 5 to the full ensemble, 40), the red dots indicate the central estimate, and the pink envelope represents the 95% confidence interval. The estimates based on the full ensemble, which we consider the truth, are also drawn across the plot for reference, as horizontal lines. The blue dots in each plot show the same quantities estimated simply by counting, i.e., computing the empirical cumulative distribution function of Rx1Day on the basis of the  $n \times 11$  years where n is the ensemble size. The first three rows show results for a location in Southern South America while the following three rows show results for a location in Northern Asia (see Figure C1).





# C2 Variability



**Figure C12.** Estimating the ensemble variance for TNn: Each plot corresponds to a year along the simulation length (1950, 1975, 2000, 2025, 2050, 2075, 2100). The color indicates the number of ensemble members needed to estimate a variance at that location that is statistically indistinguishable from that computed on the basis of the full 40-member ensemble. The results of the first two columns use only the specific year across the members. The results of the third and fourth columns enrich the samples by using five years around the specific date.







**Figure C13.** Estimating the within ensemble variance for TXx: Each plot corresponds to a year along the simulation length (1950, 1975, 2000, 2025, 2050, 2075, 2100). The color indicates the number of ensemble members needed to estimate a variance at that location that is statistically indistinguishable from that computed on the basis of the full 40-member ensemble. The results of the first two columns use only the specific year across the members. The results of the third and fourth columns enrich the samples by using five years around the specific date.







**Figure C14.** Estimating the ensemble variance for TXn: Each plot corresponds to a year along the simulation length (1950, 1975, 2000, 2025, 2050, 2075, 2100). The color indicates the number of ensemble members needed to estimate a variance at that location that is statistically indistinguishable from that computed on the basis of the full 40-member ensemble. The results of the first two columns use only the specific year across the members. The results of the third and fourth columns enrich the samples by using five years around the specific date.







**Figure C15.** Estimating the within ensemble variance for Rx1Day: Each plot corresponds to a year along the simulation length (1950, 1975, 2000, 2025, 2050, 2075, 2100). The color indicates the number of ensemble members needed to estimate a variance at that location that is statistically indistinguishable from that computed on the basis of the full 40-member ensemble. The results of the first two columns use only the specific year across the members. The results of the third and fourth columns enrich the samples by using five years around the specific date.







**Figure C16.** Estimating changes in within-ensemble variance for TNn: each plot corresponds to a pair of years along the simulation (same set of years as depicted in Figures 7 and 8 above). Colored areas are regions where on the basis of the full 40-member ensemble a significant change in variance was detected. The colors indicate the size of the smaller ensemble needed to detect the same change. Here the sampling is increased by using 5 years around each date as part of the samples.







**Figure C17.** Estimating changes in within-ensemble variance for TXx: each plot corresponds to a pair of years along the simulation (same set of years as depicted in Figures 7 and 8 above). Colored areas are regions where on the basis of the full 40-member ensemble a significant change in variance was detected. The colors indicate the size of the smaller ensemble needed to detect the same change. Here the sampling is increased by using 5 years around each date as part of the samples.







**Figure C18.** Estimating changes in within-ensemble variance for TXn: each plot corresponds to a pair of years along the simulation (same set of years as depicted in Figures 7 and 8 above). Colored areas are regions where on the basis of the full 40-member ensemble a significant change in variance was detected. The colors indicate the size of the smaller ensemble needed to detect the same change. Here the sampling is increased by using 5 years around each date as part of the samples.







**Figure C19.** Estimating changes in within-ensemble variance for Rx1Day: each plot corresponds to a pair of years along the simulation (same set of years as depicted in Figures 7 and 8 above). Colored areas are regions where on the basis of the full 40-member ensemble a significant change in variance was detected. The colors indicate the size of the smaller ensemble needed to detect the same change. Here the sampling is increased by using 5 years around each date as part of the samples.



385



*Author contributions.* CT conceived the study, analyzed the data and wrote the paper. KD and MW provided data pre-processing and co-wrote the paper. RL advised and co-wrote the paper.

380 *Competing interests.* The authors declare no competing interests.

Acknowledgements. This study was supported by the Energy Exascale Earth System Model (E3SM) project funded by U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. MW was supported by the CASCADE project also funded by U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research. The Pacific Northwest National Laboratory is operated by Battelle for the US Department of Energy under Contract DE-AC05-76RLO1830. Lawrence Berkeley National Laboratory is operated by U.S. Department of Energy under Contract No. DE340AC02-05CH11231.





### References

- Alexander, L. V.: Global Observed Long-term Changes in Temperature and Precipitation Extremes: a Review of Progress and Limitations in IPCC Assessments and Beyond., Weather and Climate Extremes, 11, 4-16, 2016.
- Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W. G., and Merryfield, W. J.: Carbon
- emission limits required to satisfy future representative concentration pathways of greenhouse gases, Geophysical Research Letters, 38, 390 https://doi.org/https://doi.org/10.1029/2010GL046270, 2011.
  - Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer-Verlag, London, London, United Kingdom, 2001.
  - Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in Climate Change Projections: the Role of Internal Variability, Climate Dynamics, 38, https://doi.org/10.1007/s00382-010-0977-x, 2012.
- Deser, C., Lehner, F., Rodgers, K. B., Ault, T., Delworth, T. L., DiNezio, P. N., Fiore, A., Frankignoul, C., Fyfe, J. C., Horton, D. E., 395 Kay, J. E., Knutti, R., Lovenduski, N. S., Marotzke, J., McKinnon, K. A., Minobe, S., Randerson, J., Screen, J. A., Simpson, I. R., and Ting, M.: Insights from Earth system model initial-condition large ensembles and future prospects, Nature Climate Change, 10, 277–286, https://doi.org/10.1038/s41558-020-0731-2, 2020.
  - Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, Bulletin of the American Meteorological Society, 90, 1095-1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.
  - Huntingford, C., Jones, P. D., Livina, V. N., Lenton, T. M., and Cox, P. M.: No increase in global temperature variability despite changing regional patterns, Nature, 500, 327-330, https://doi.org/10.1038/nature12310, 2013.
    - Hurrell, J. W., Holland, M. M., and Gent, P. R.: The Community Earth System Model: A Framework for Collaborative Research, Bulletin of the American Meteorological Society, 94, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.
- 405 Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., G., S., and et al.: The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, Bulletin of the American Meteorological Society, 96, 1333-1349, https://doi.org/10.1175/BAMS-D-13-00255.1, 2015.
  - Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, Journal of Climate, 23, 2739-2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.
- 410 Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., and Hawkins, E.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, Earth System Dynamics, 11, 491-508, https://doi.org/10.5194/esd-11-491-2020, 2020.
  - Lehner, F., Deser, C., Maher, N., Marotzke, J., Fischer, E. M., Brunner, L., Knutti, R., and Hawkins, E.: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6, EARTH SYSTEM DYNAMICS, 11, 491-508, https://doi.org/10.5194/esd-11-491-2020, 2020.
- 415

400

- Maher, N., Milinski, S., and Ludwig, R.: Large ensemble climate model simulations: introduction, overview, and future prospects for utilising multiple types of large ensemble, Earth System Dynamics, 12, 401-418, https://doi.org/10.5194/esd-12-401-2021, 2021a.
- Maher, N., Power, S. B., and Marotzke, J.: More accurate quantification of model-to-model agreement in externally forced climatic responses over the coming century, Nature Communications, 12, 788, https://doi.org/10.1038/s41467-020-20635-w, 2021b.
- Milinski, S., Maher, N., and Olonscheck, D.: How large does a large ensemble need to be?, Earth System Dynamics, 11, 885-901, 420 https://doi.org/10.5194/esd-11-885-2020, 2020.



425



- Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5—A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33, https://doi.org/10.1007/s10584-011-0149-y, 2011.
- Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.
- Ventura, V., Paciorek, C., and Risbey, J. S.: Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data, Journal of Climate, 17, 4343–4356, https://doi.org/10.1175/3199.1, 2004.
- Wehner, M. F.: A method to aid in the determination of the sampling size of AGCM ensemble simulations, Climate Dynamics, 16, 321–331, https://doi.org/10.1007/s003820050331, 2000.