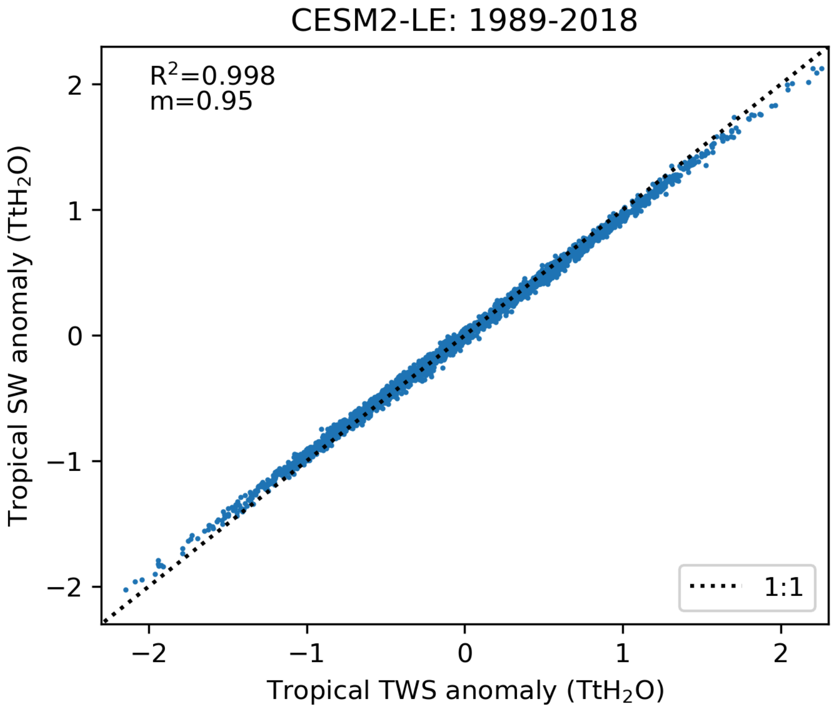
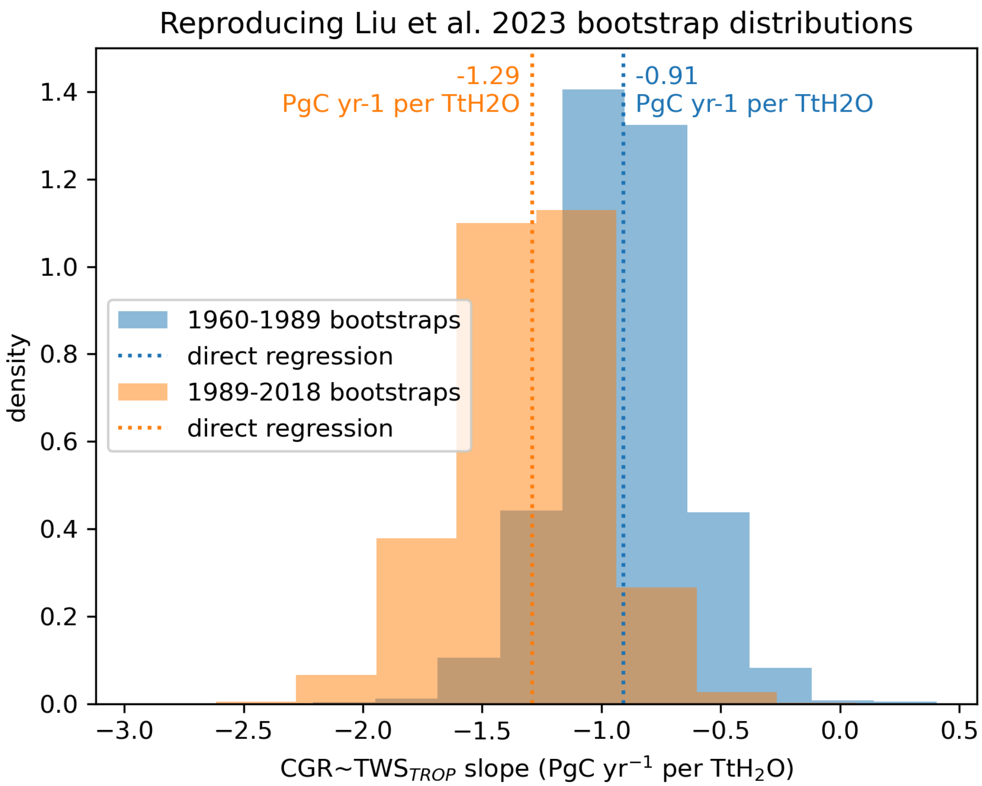
Extended Data for:
Terrestrial carbon-water relations driven by internal climate variability

D Kennedy, AT Trugman, DM Lawrence, SC Swenson and IR Simpson

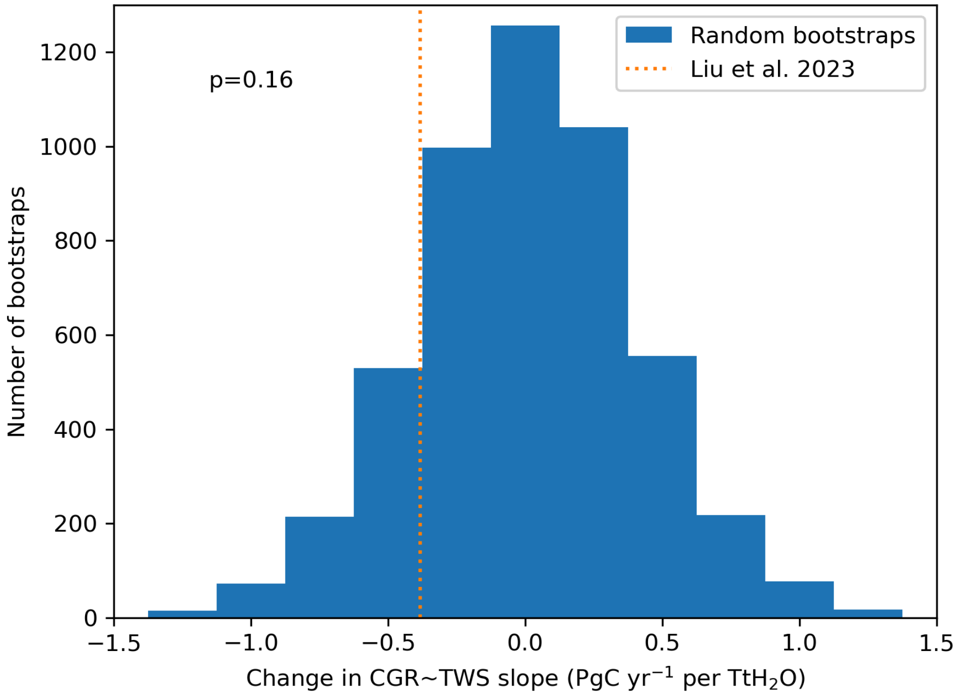
Contains:
Extended Data Figures 1-13



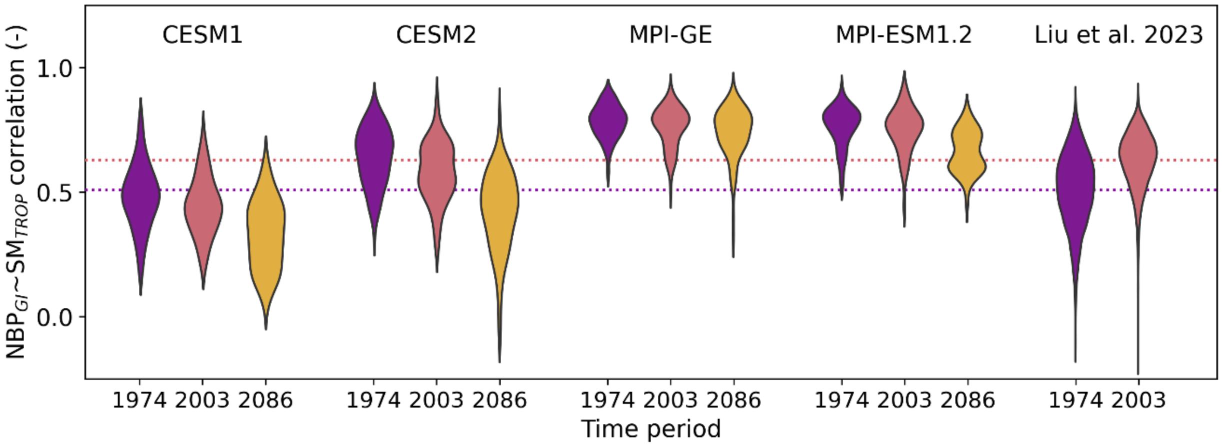
Extended Data Figure 1: Total column soil moisture anomalies are both highly correlated with TWS anomalies and very nearly 1:1, examined for CESM2 where both SM and TWS are available as model output. Blue dots represent individual years between 1989-2018 across all ensemble members in the CESM2-LE. The dotted black line is the 1:1 line.



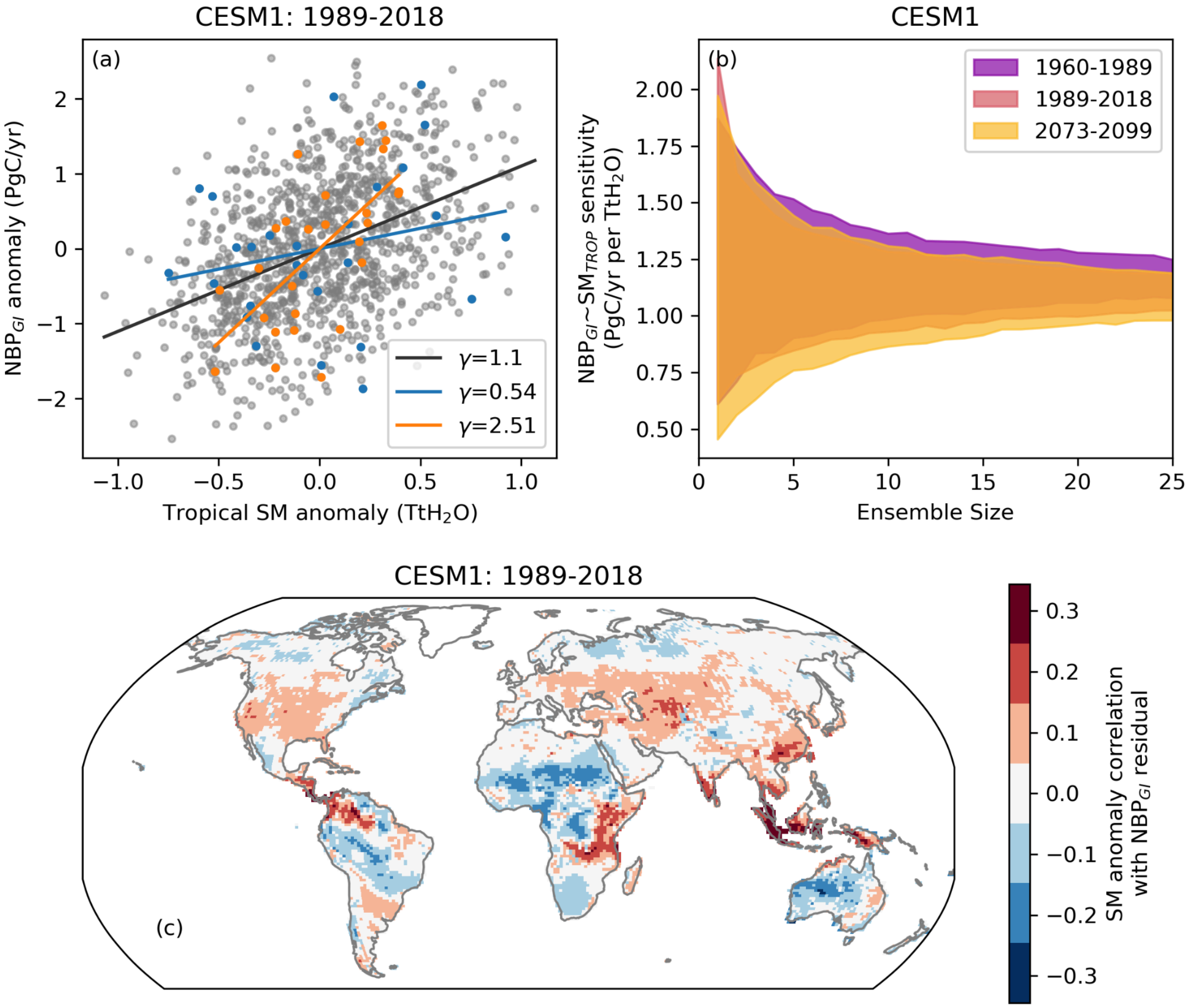
Extended Data Figure 2: Bootstrap distributions reproduced from Liu *et al.* (2023). We used data the authors of Liu *et al.* (2023) provided to us via email, but our calculations yielded a slightly different slopes compared to Liu *et al.* (2023) Figure 1d, which were reported as −0.95 (compared to -0.91) PgC yr^{-1} per Tt H_2O to −1.26 (compared to -1.29) PgC yr^{-1} per Tt H_2O. However, the match is reasonably close (see Methods).



Extended Data Figure 3: Based on 5000 time series generated from bootstrapping the Liu *et al.* (2023) CGR and TWS observations, we found that the probability of a random decrease in slope larger than observed (-0.38 PgC yr^{-1} per TtH2O) was 16%. This conflicts with a one-sided Wilcoxon signed rank test (also with 5000 bootstraps), where p was equal to 0 to machine precision.

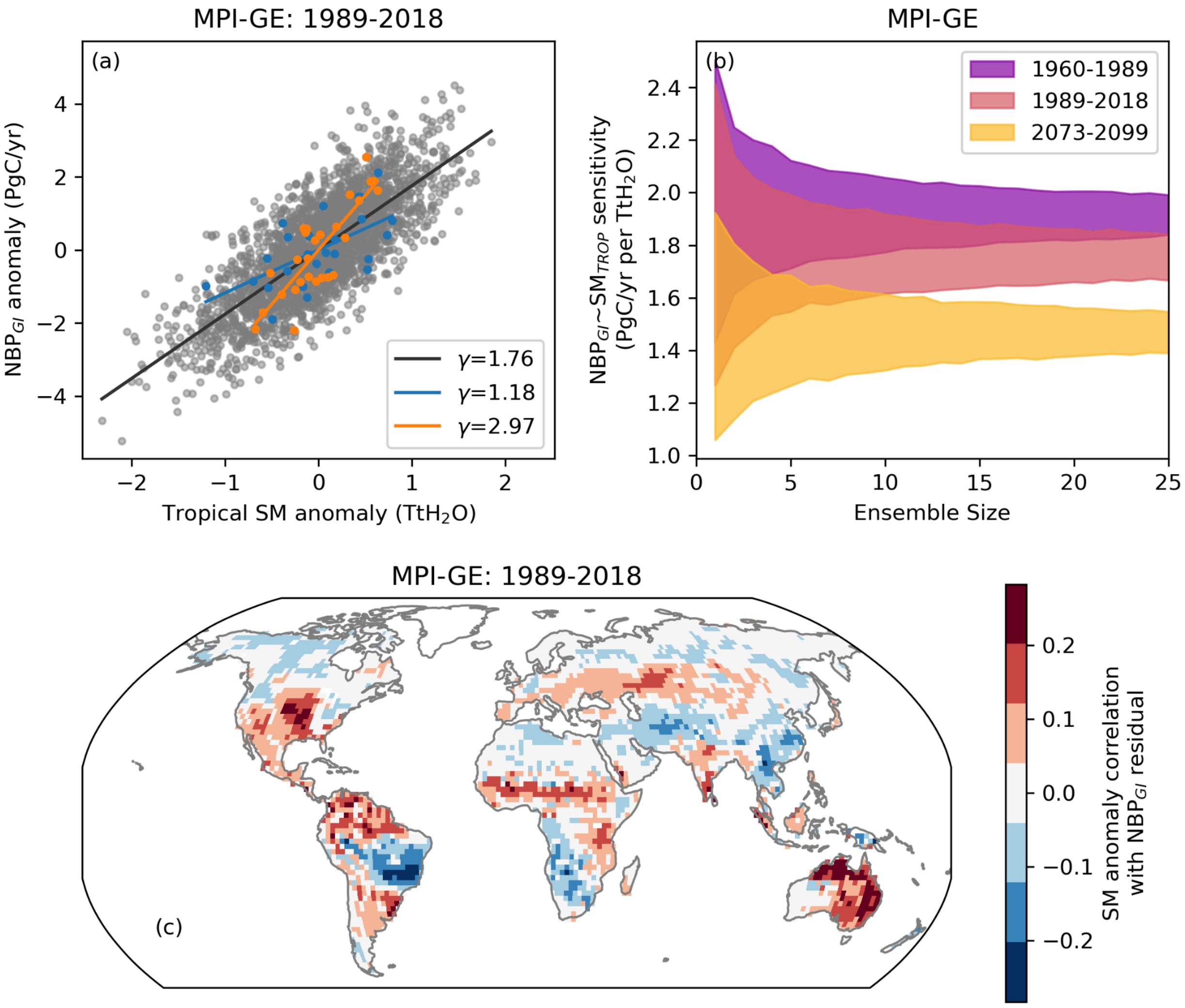


Extended Data Figure 4: Correlation coefficients of the NBP$_{GI}$ - tropical SM relationship across the four LEs and from the Liu *et al.* (2023) data. Liu et al. (2023) utilizes CGR, which is comparable to NBP$_{GI}$, albeit opposite in sign. As such, we reversed the sign for the Liu *et al.* (2023) correlations. For the large ensembles, the violins represent the distributions of the slopes of the various ensemble members. For Liu et al. (2023), the violin represents the distribution of 5000 bootstraps of the observational time series. Dotted lines plot the correlation coefficients of the Liu *et al.* (2023) data directly, without bootstrapping.



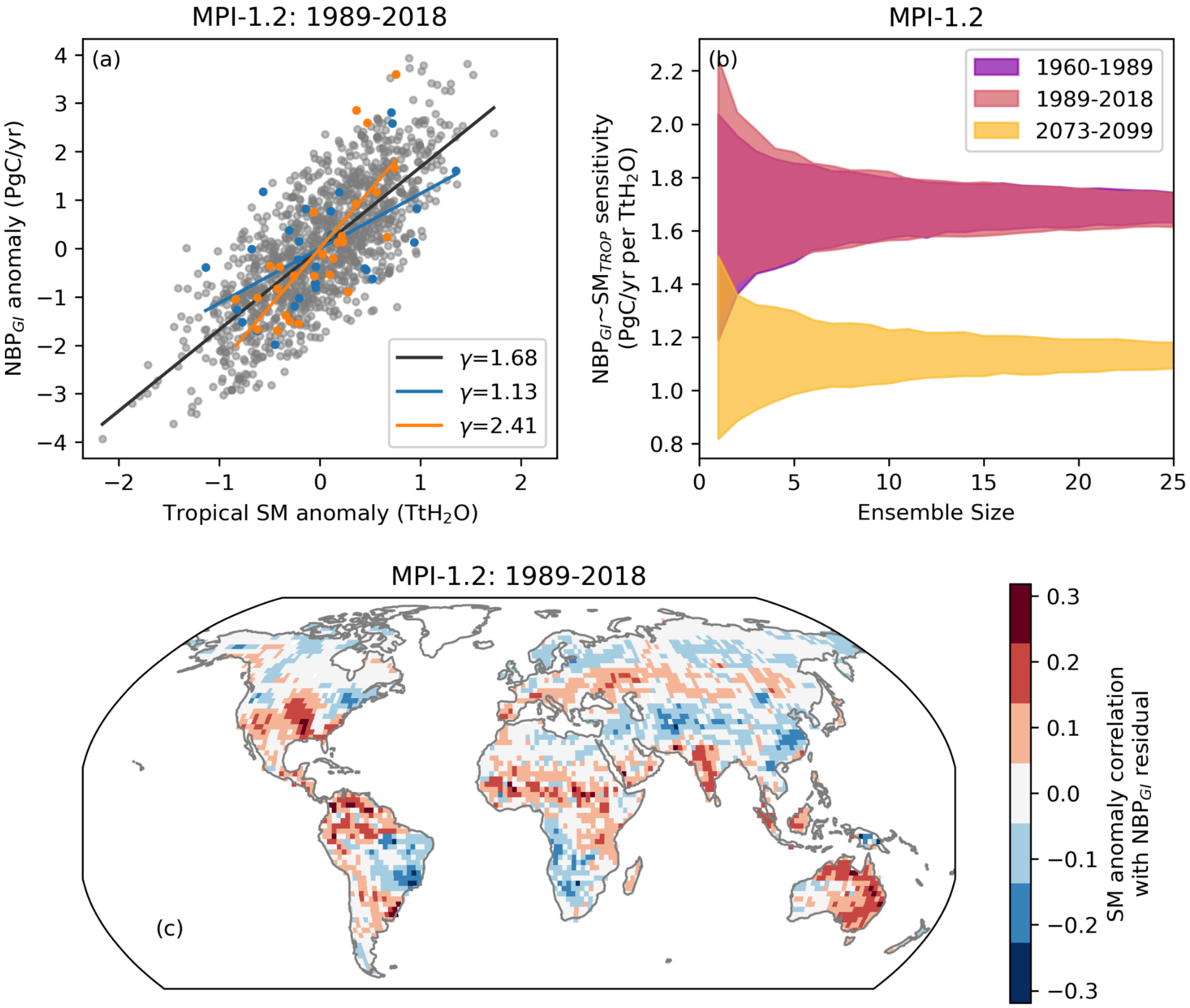
Extended Data Figure 5

Reproduction of main text Figure 2, for CESM1.



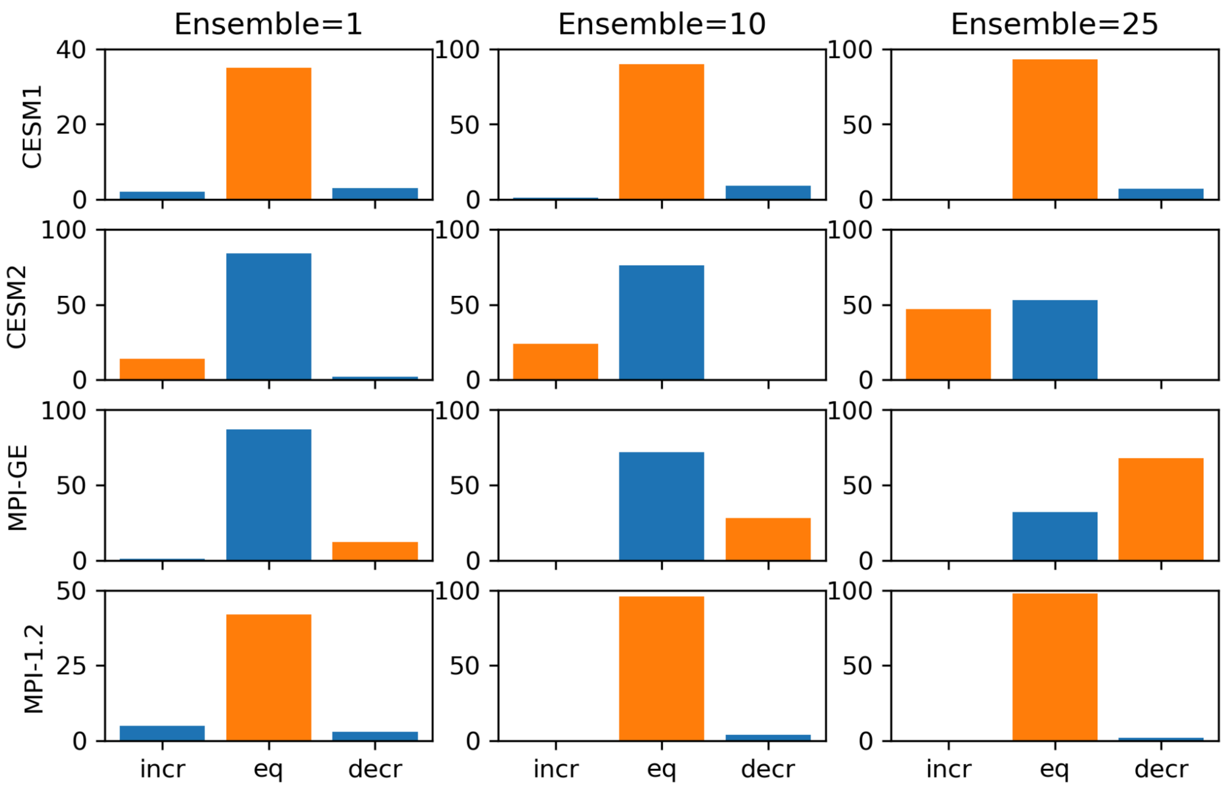
Extended Data Figure 6

Reproduction of main text Figure 2, for MPI-GE.

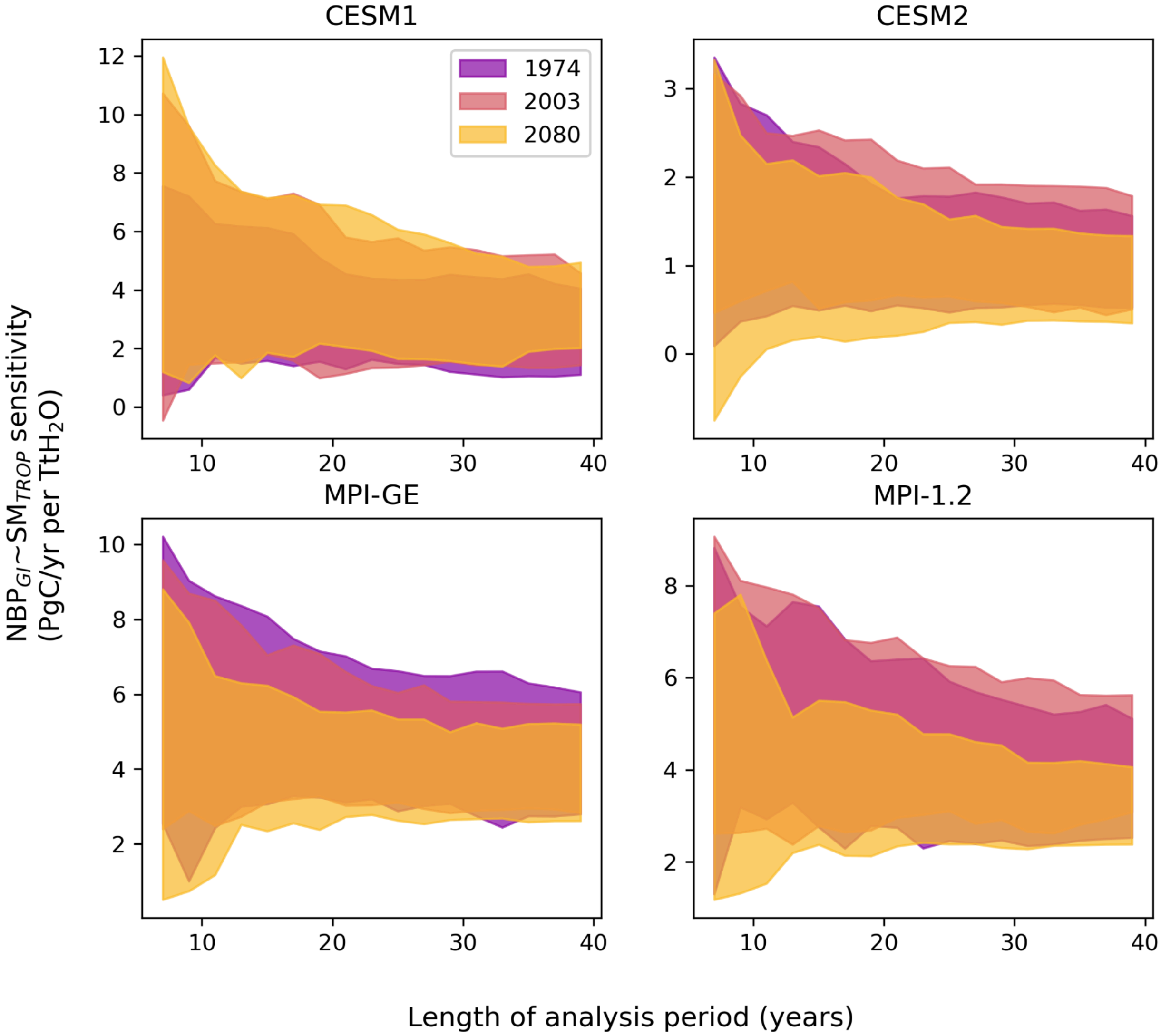


Extended Data Figure 7

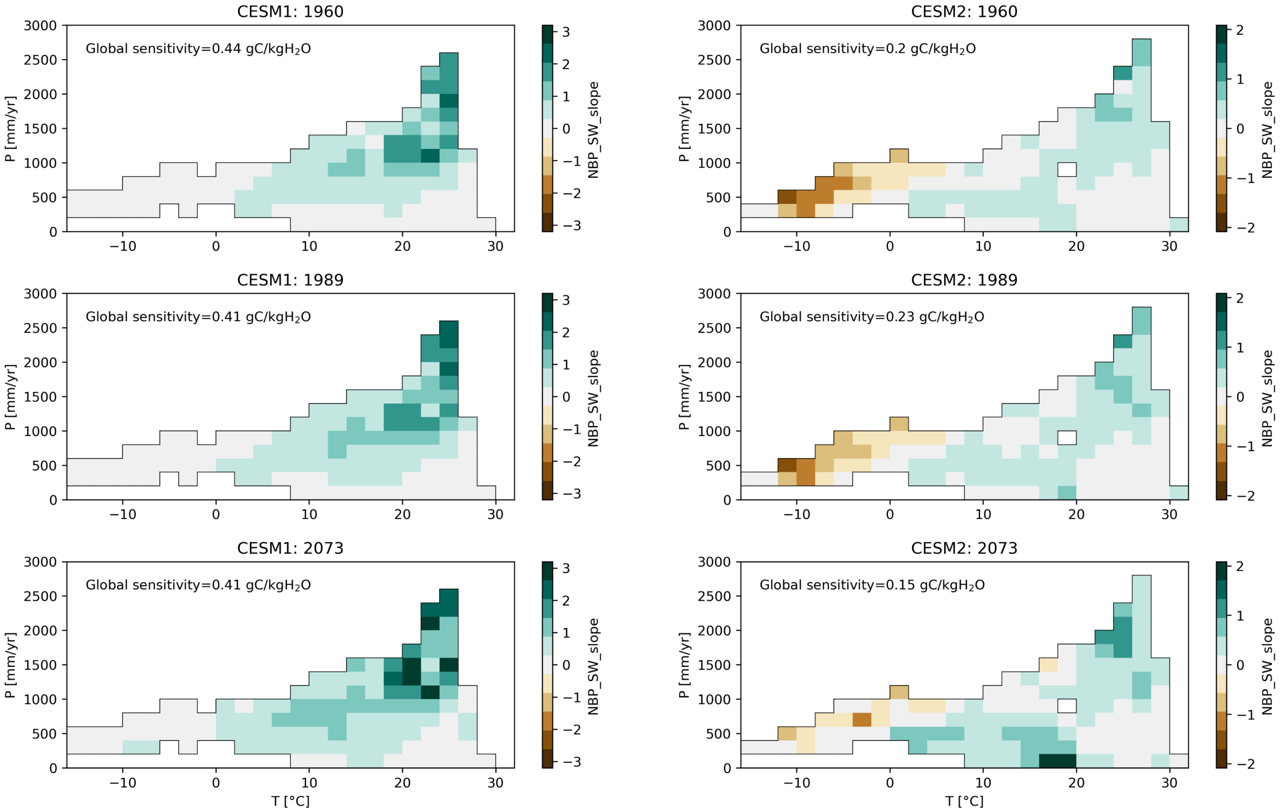
Reproduction of main text Figure 2, for MPI-1.2.



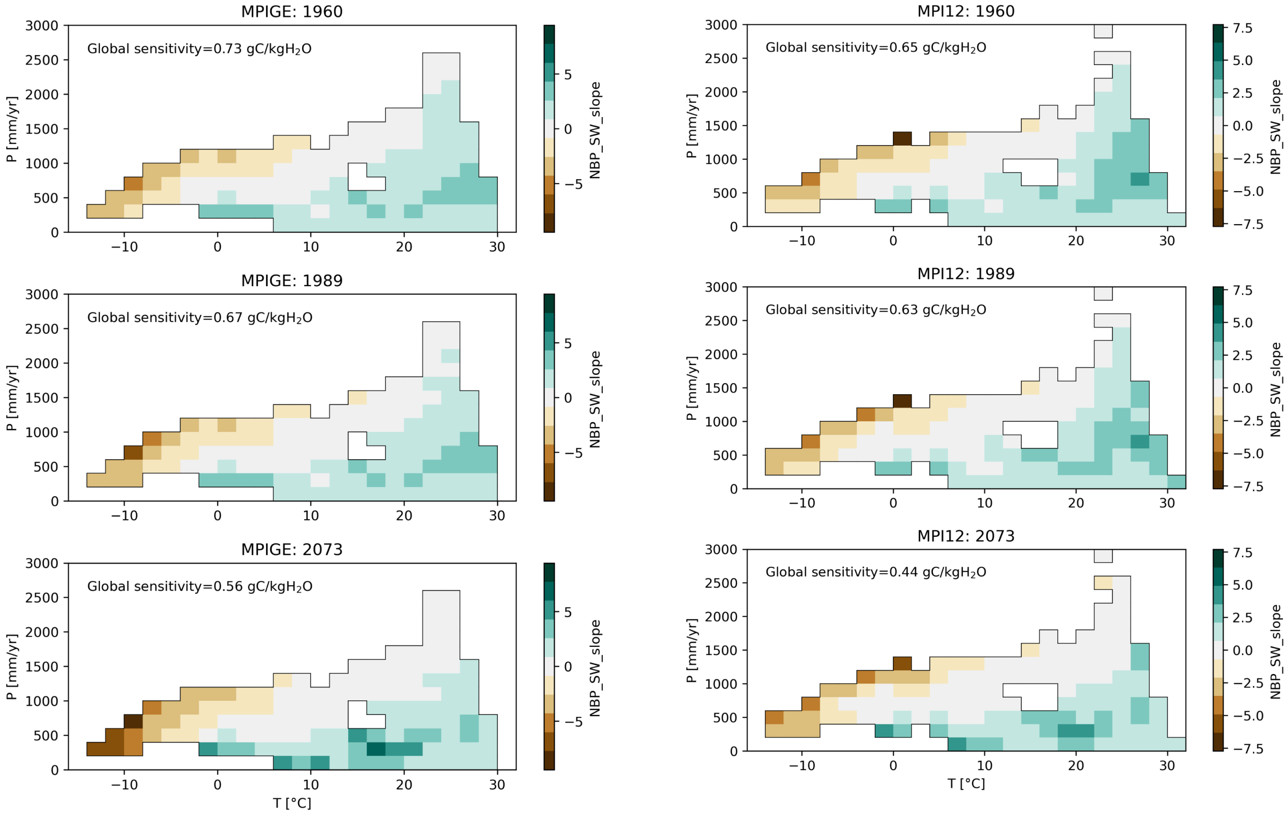
Extended Data Figure 8: The number of ESM ensemble realizations required to confidently distinguish between statistically different slopes varies depending on ESM. Shown are the frequency of NBP_{GI}-SM_{TROP} slope changes for various sub-ensemble sizes and ESMs, comparing the periods 1960-1989 vs. 1989-2018. In the case of a sub-ensemble of size 1, we asked whether the sensitivity of each ensemble member increased, decreased, or was statistically equivalent at the 5% confidence level, using our bootstrap delta test described in the Methods. For the sub-ensembles of 10 and 25, we generated 100 random sub-ensembles of the given size, and asked if the sensitivity in that subsetted dataset appears to have increased (incr), decreased (decr), or stayed the same (eq). Blue bars indicate when the sub-ensemble signal is contrary to that from the full ensemble signal in Main Text Figure 1 and orange indicate when it is consistent. As ensemble size increases, the proportion contained within the orange bars increases.



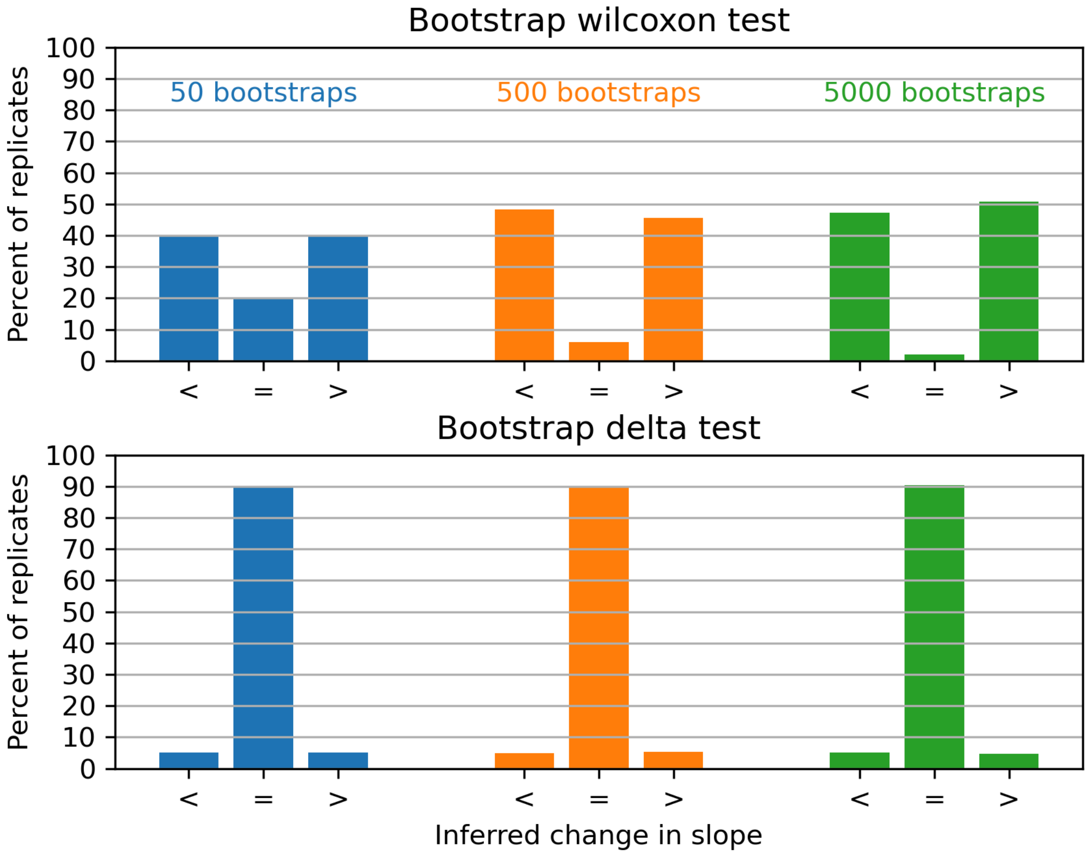
Extended Data Figure 9: The influence of analysis period length on the range of sensitivity slopes across the four large ensembles. Shown are the 5th-95th percentile range of sensitivity slopes based on different length time windows, centered around the years 1974, 2003, and 2080.



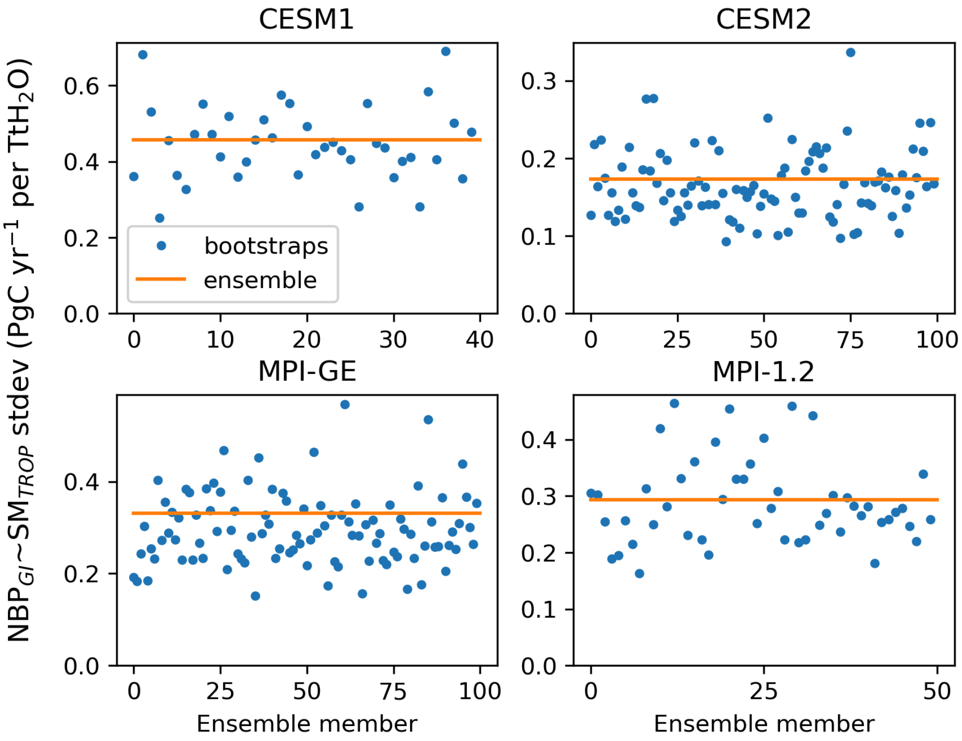
Extended Data Figure 10: Whittaker diagrams (with precipitation P on the y-axis and temperature T on the x-axis) for climatic variations in the NBP – TWS slope (PgC/yr per TtH$_2$O) for the CESM1 LE (left column) and the CESM2 LE (right column) for time periods 1960-1989 (top row), 1989-2018 (middle row), and 2073-2100 (bottom row).



Extended Data Figure 11: Whittaker diagrams (with precipitation P on the y-axis and temperature T on the x-axis) for climatic variations in the NBP – TWS slope (PgC/yr per TtH_2O) for the MPI GE (left column) and MPI-1.2 LE (right column) for time periods 1960-1989 (top row), 1989-2018 (middle row), and 2073-2100 (bottom row).



Extended Data Figure 12: Synthetic data test of statistical significance illustrates artificially inflated statistical confidence of inappropriately applied Wilcoxon sign ranked test. (upper) Based on 1000 randomly generated time series of synthetic data where we explicitly impose no change in slope, a bootstrap Wilcoxon signed rank test frequently identifies a significant change in slope (at the 5% level). The number of false positives likewise increases with the number of bootstraps. (lower) We propose an alternative test (see Methods) that better comports with expectations, with approximately 5% false positive rate on either side, independent of the number of bootstraps.



Extended Data Figure 13: The standard deviation of NBP$_{GI}$~SM$_{TROP}$ slopes (1960-1989) calculated for a set of 500 bootstraps for each ensemble member (blue dots) plotted in the context of the standard deviation across ensemble members without bootstrapping (orange line). Bootstrapping appears to be a reasonable proxy for internal variability. In general the standard deviation across a single ensemble member's bootstraps underpredicts the standard deviation across ensemble members, but only slightly (-2.0, -4.3, -10.5, -1.6% on average for CESM1, CESM2, MPI-GE, and MPI-1.2, respectively).