Supplementary Information to "Exploiting large ensembles for a better yet simpler climate model evaluation"

Laura Suarez-Gutierrez[1], Sebastian Milinski[1], and Nicola Maher[1]

1. Max-Planck-Institut für Meteorologie, Hamburg, Germany.

S.1 Time series and rank frequency analysis for a common period

To avoid the confounding effects of different simulation lengths across the ensembles on the rank frequencies, as well as the relatively larger weight of the climatological reference period for shorter simulations, we repeat the rank frequency analysis for a period common to all models: 1950–2014 (Fig. S.1 and S.2). This change in our analysis period does not lead to substantial changes in our conclusions for most models, except for CSIRO-MK3.6 and MIROC6. For these two models, excluding the 19th and early 20th centuries leads to an improved performance, due to their overestimation of the observed forced warming during this time. Although this also leads to a reduction in the relative frequencies of low and zero ranks, these frequencies are still outside of the adequate performance range for both models.

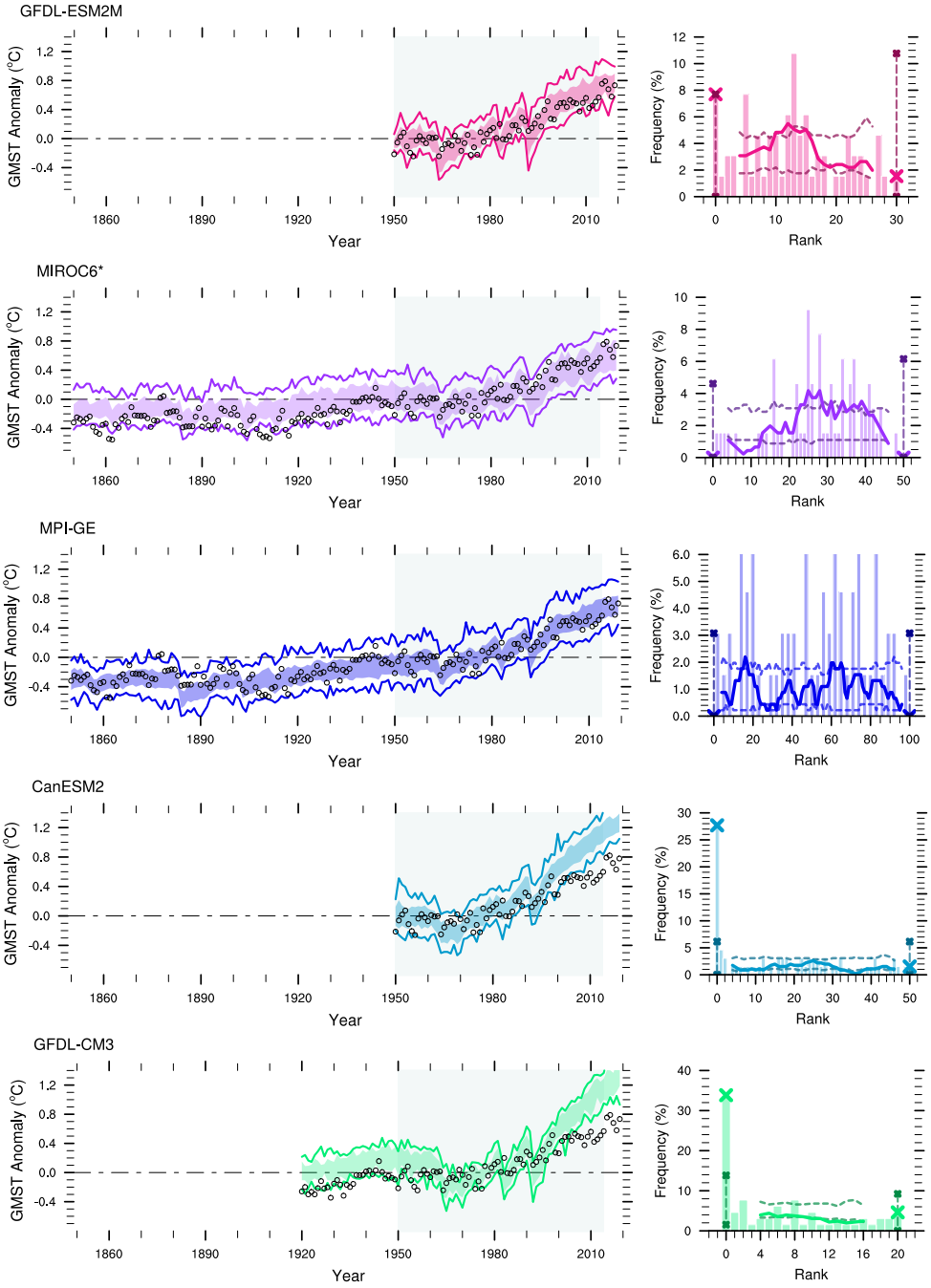

Figure S.1: Time series and rank histograms of annual GMST anomalies for the common period of 1950–2014. Time series of annual GMST anomalies simulated by each SMILE (colored) and GMST HadCRUT4 observed anomalies (black circles), as in Fig. 1 and 2. Blue shading highlights common period for rank histogram analysis. Rank histograms represent the frequency of each place that HadCRUT4 GMST observations (light colors) would take in a list of ensemble members ordered by ascending GMST values (right column), compared to perfect model tests based on each SMILE (dark colors), as for Fig. 1 and 2 but for the period of 1950–2014.

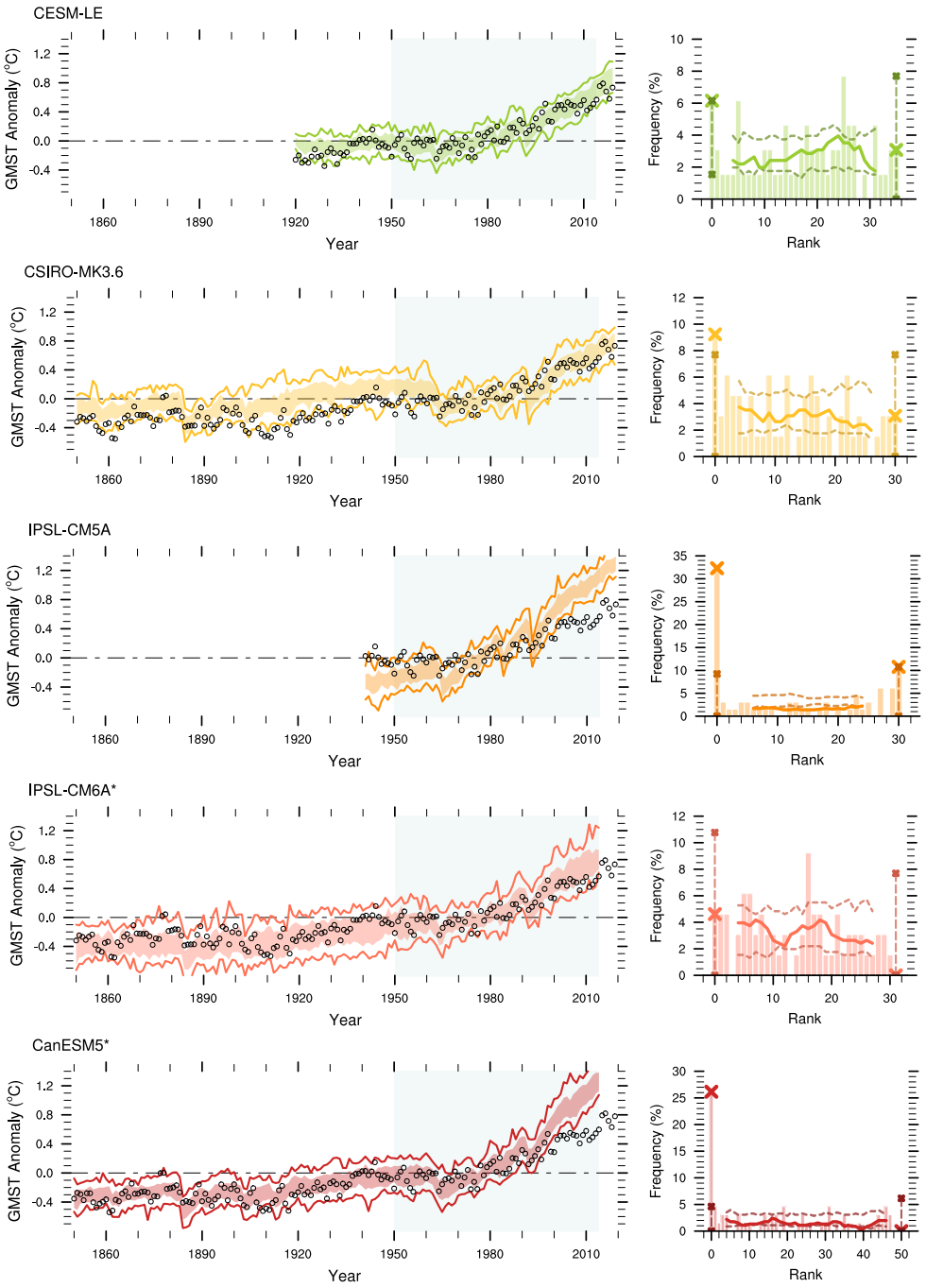

Figure S.2: Time series and rank histograms of annual GMST anomalies for the common period of 1950–2014. Time series of annual GMST anomalies simulated by each SMILE (colored) and GMST HadCRUT4 observed anomalies (black circles), as in Fig. 1 and 2. Blue shading highlights common period for rank histogram analysis. Rank histograms represent the frequency of each place that HadCRUT4 GMST observations (light colors) would take in a list of ensemble members ordered by ascending GMST values (right column), compared to perfect model tests based on each SMILE (dark colors), as for Fig. 1 and 2 but for the period of 1950–2014.

S.2 Sensitivity analysis of spatial evaluation

In this section we elaborate on the how our spatial evaluation analysis depends on ensemble size and on the selected threshold for identifying variability biases by replicating the analysis in figures Fig. 5 in the main article for slightly more permissive thresholds (Fig. S.3) and comparable ensemble sizes of 30 members (Fig. S.4). We find that, as expected, more permissive bias thresholds improve the performance metrics and area of no substantial biases for all ensembles (Fig. S.3), particularly for those with the largest range of variability such as GFDL-ESM2M, but do not alter our conclusions in terms of model performance to a substantial degree.

The evaluation under comprable ensemble sizes shows that reducing the ensemble size leads to a weakened performance and general decrease in the area where no biases occur to a substantial degree (Fig. S.4). This decrease is larger for the largest ensemble, MPI-GE, which exhibits large areas where observations occur both above and below the ensemble limit, particularly in the Southern Hemisphere. Under comparable ensemble sizes, the 30-member GFDL-ESM2M ensemble surpasses MPI-GE in fraction of unbiased area, and offers the largest area of adequate representation of surface temperatures. GFDL-ESM2M, which exhibits mostly biases of overestimated variability over land areas (Fig. S.4), is one the SMILEs with the largest annual surface temperature variability range in our study (Fig. S.6). Thus it is possible that the large internal variability in the ensemble is sufficient to conceal other biases in this threshold spatial evaluation (Fig. 1f); and such potential biases may be best identified or ruled out with a rank frequency analysis for the specific regions of interest.

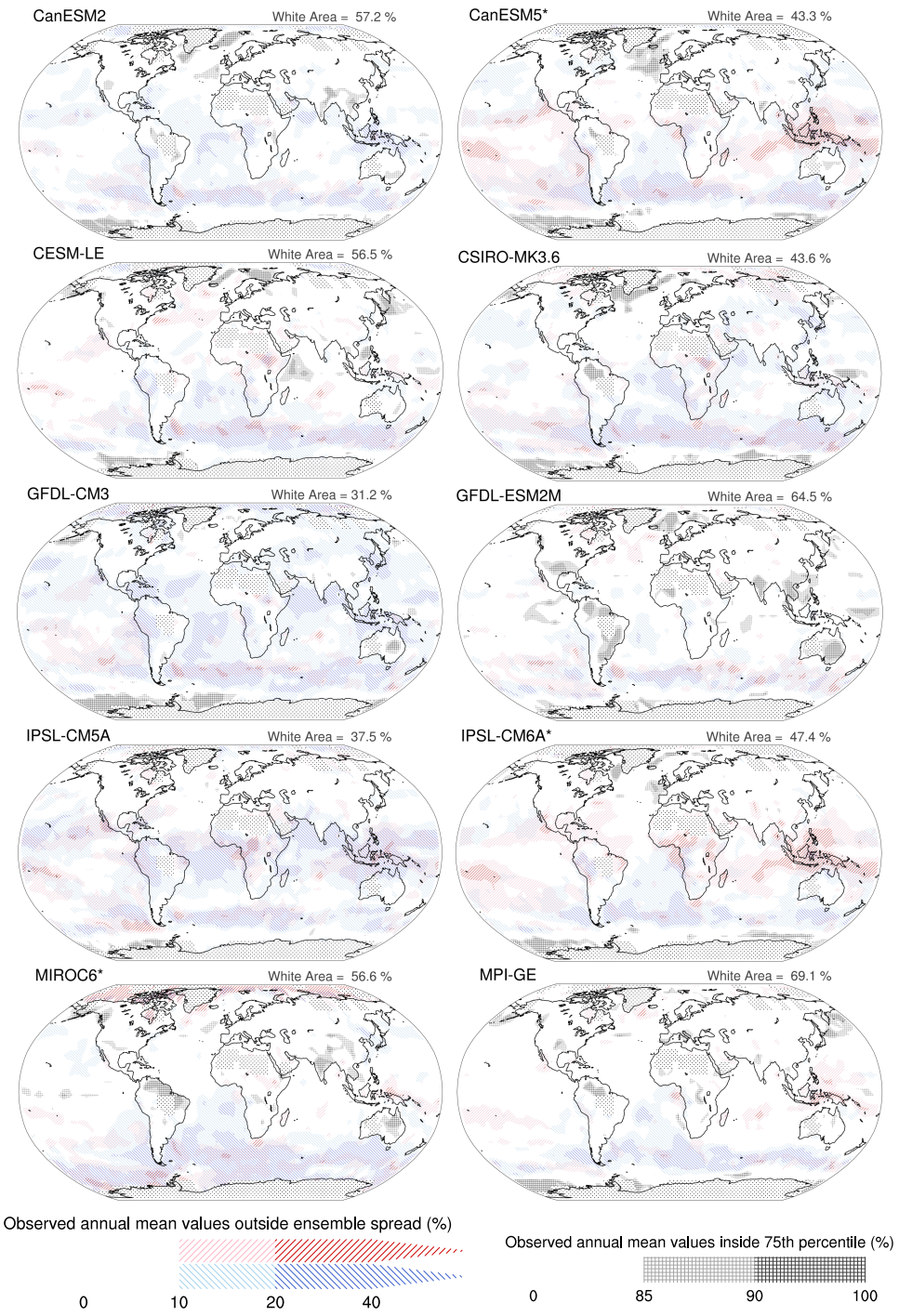

Figure S.3: Evaluation of internal variability and forced response in annual surface temperatures for more permissive bias threshold. As in Fig. 5, red shading represents where observed anomalies are larger than the ensemble maximum; while blue shading represents where observed anomalies are smaller than the ensemble minimum, both for more than 10% (light color) or 20% (dark color) of the time. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensembles (12.5th to 87.5th percentiles) now for more than 85% (light color) or 90% (dark color) of the time. White Area represents the percentage of total area included in the analysis that exhibits no substantial biases for each SMILE. Dotted areas represent where observations are available for less than 10 years, and therefore excluded from our analysis.

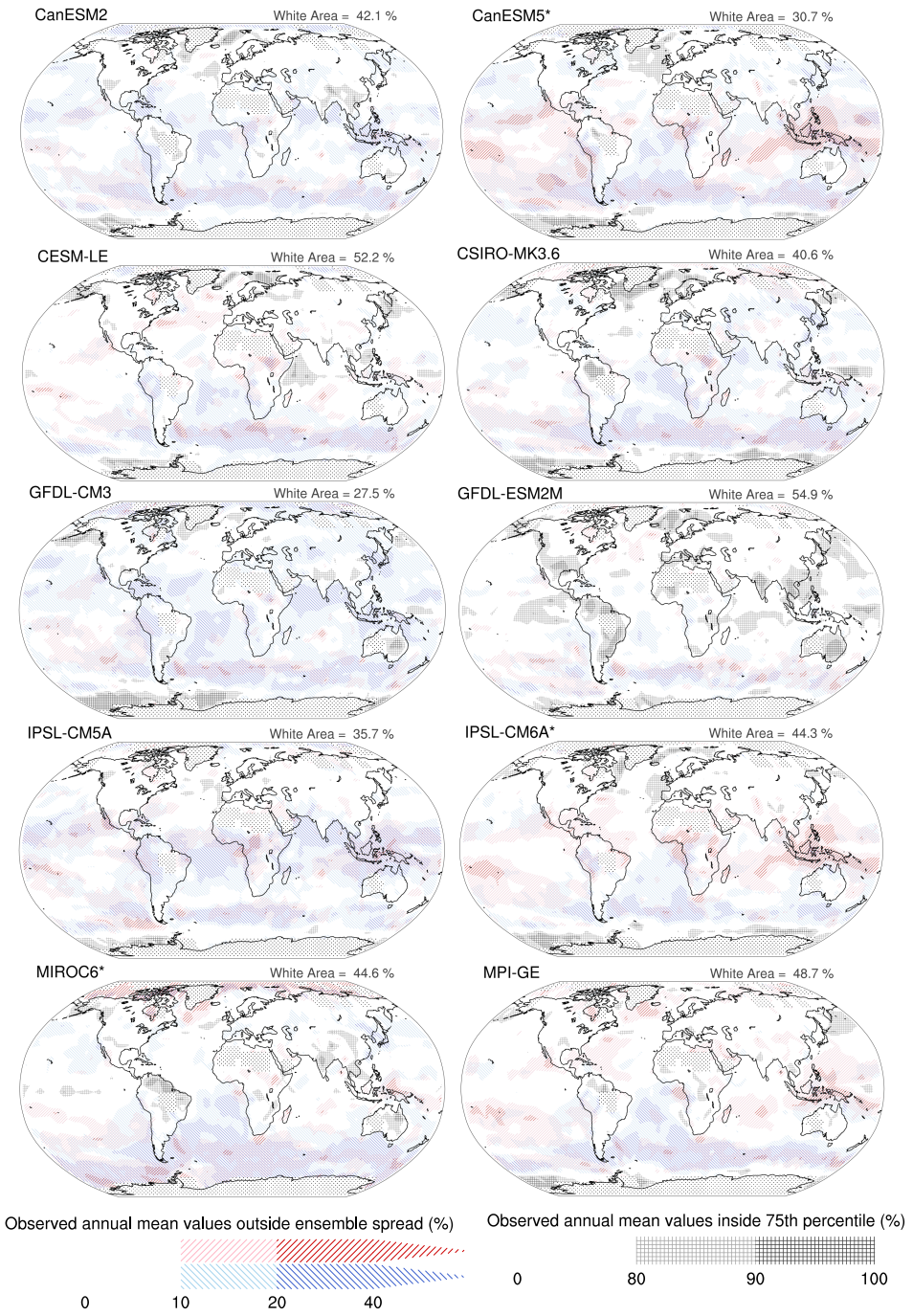

Figure S.4: Evaluation of internal variability and forced response in annual surface temperatures for comparable ensemble sizes. As in Fig. 5., red shading represents where observed anomalies are larger than the ensemble maximum; while blue shading represents where observed anomalies are smaller than the ensemble minimum, both for more than 10% (light color) or 20% (dark color) of the time. Gray hatching represents where observations cluster within the 75th percentile bounds of the ensembles (12.5th to 87.5th percentiles) more than 80% (light color) or 90% (dark color) of the time. White Area represents the percentage of total area included in the analysis that exhibits no substantial biases for each SMILE. Dotted areas represent where observations are available for less than 10 years, and therefore excluded. For SMILEs with more than 30 members, only the first 30 members are considered; for GFDL-CM3, only the 20 available members are considered.

S.3 Comparison of internal variability

We offer a multi-model comparison of the well-defined internal variability based on ten SMILE experiments. We measure the simulated internal variability as the ensemble spread between the 2.5th and 97.5th percentiles. We find that most of the ensembles present remarkably similar patterns and magnitude of internal variability in surface temperatures (Fig. S.5), especially once the different number of ensemble members is taken into account (Fig. S.6). The largest variability in annual mean surface temperatures, of more than 5°C in amplitude, occurs over the polar regions and sea-ice edges. This occurs similarly in all models with the exception of MIROC6, which exhibits less than 1°C of amplitude in variability of Arctic surface temperatures. The lowest variability across most models occurs in the mid-latitude and Southern Hemisphere oceans, with values between 2.5°C and 1°C. The ensembles simulate larger internal variability over land than over the oceans, and larger internal variability in the Northern Hemisphere than in the Southern Hemisphere. IPSL-CM5A exhibits the highest variability, followed by GFDL-CM3 and GFDL-ESM2M; whereas MIROC6 exhibits the lowest variability, especially in mid to high-latitude regions.

Compared to the 30-member variability spread, the range of internal variability increases when including more ensemble members, particularly for CanESM2 and CanESM5 (Fig. S.5). This indicates that the spread of these ensembles is not saturated at 30 members even for relatively low variability magnitudes such as annual mean surface temperatures. However, we do not find a consistent relationship between higher variability ranges and larger ensemble sizes for all models; and MPI-GE, the ensemble with the highest number of members, does not present the highest variability even when all available ensemble members are included.

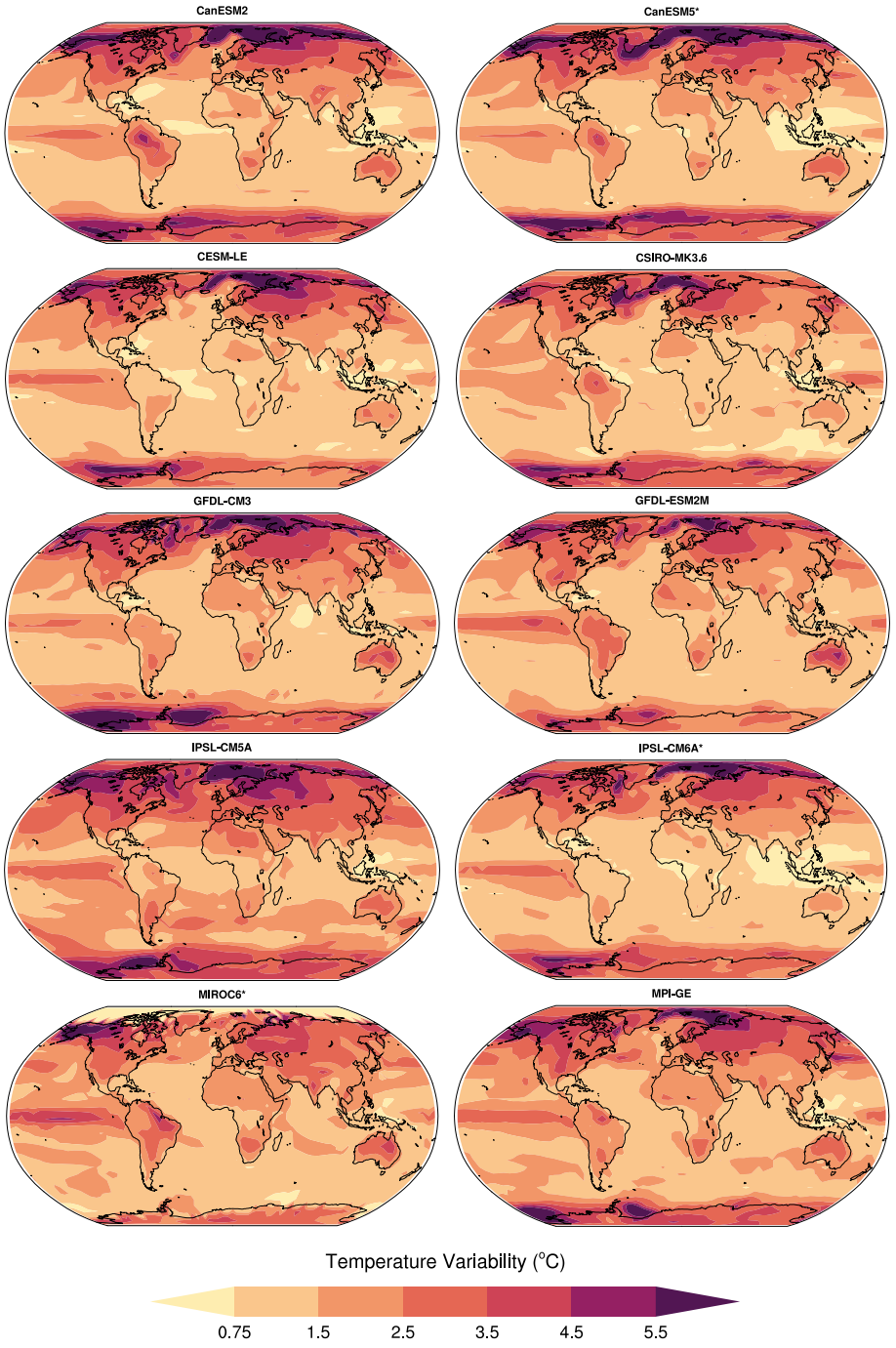

Figure S.5: Variability in annual surface temperatures. Full ensemble spread for annual surface temperature anomalies simulated by different SMILEs averaged for the period of 1950–2014 measured as the 2.5th to 97.5th percentiles. Simulated data are regridded to match the observational grid.

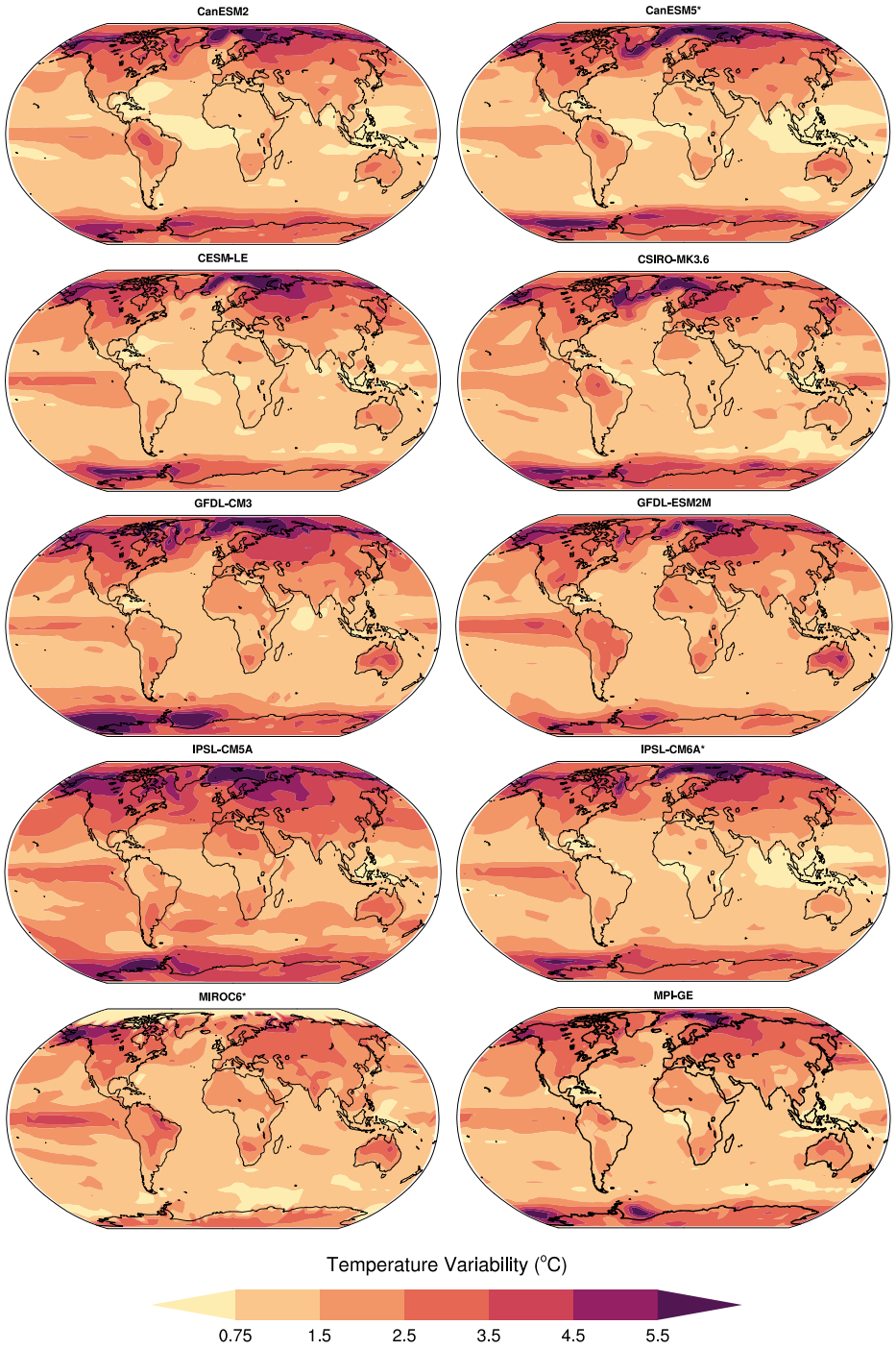

Figure S.6: Variability in annual surface temperatures for comparable ensemble sizes. 30-member ensemble spread for annual surface temperature anomalies simulated by different SMILEs, averaged for the period 1950–2014, and measured as the 2.5th to 97.5th percentile spread. For SMILEs with more than 30 members, only the first 30 members are considered; for GFDL-CM3, only the 20 available members have been considered. Model output data are regridded to match the observational grid.

S.4 Comparison of mean changes in recent decades

In this section we compare the observed mean annually averaged surface temperature anomalies in recent decades (Fig. S.7) with the simulated mean surface temperature anomalies in recent decades in each ensemble (Fig. S.8), and the bias in each model as the difference between the two means (Fig. S.9). Most models agree with the stronger observed mean warming over land, but some overestimate recent warming over the oceans, in particular CanESM2, CanESM5, GFDL-CM3 and IPSL-CM5A (Fig. S.8). In addition, most models overestimate recent mean increase in surface temperatures over the Arctic; while MIROC6 exhibits an underestimation of the mean temperature increase in this region (Fig. S.9). Over the Southern Ocean, most models substantially overestimate mean surface warming in recent decades, with mean temperature increases more than 1°C higher than observed for CESM-LE, CanESM2 and CanESM5.

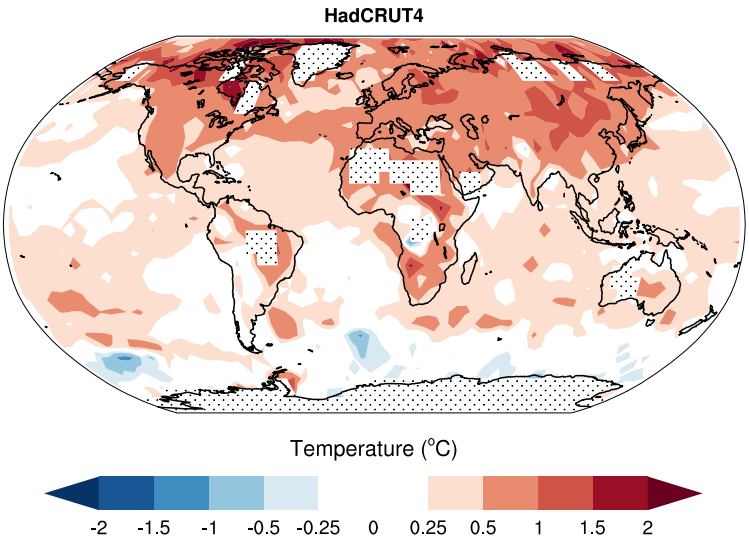

Figure S.7: Observed mean surface temperatures in recent decades. Mean annual surface temperature HadCRUT4 (**morice12**) observed anomalies anomalies relative to the period 1961–1990, and averaged for the period of 1990–2014. Dotted areas represent regions where observations are not available.

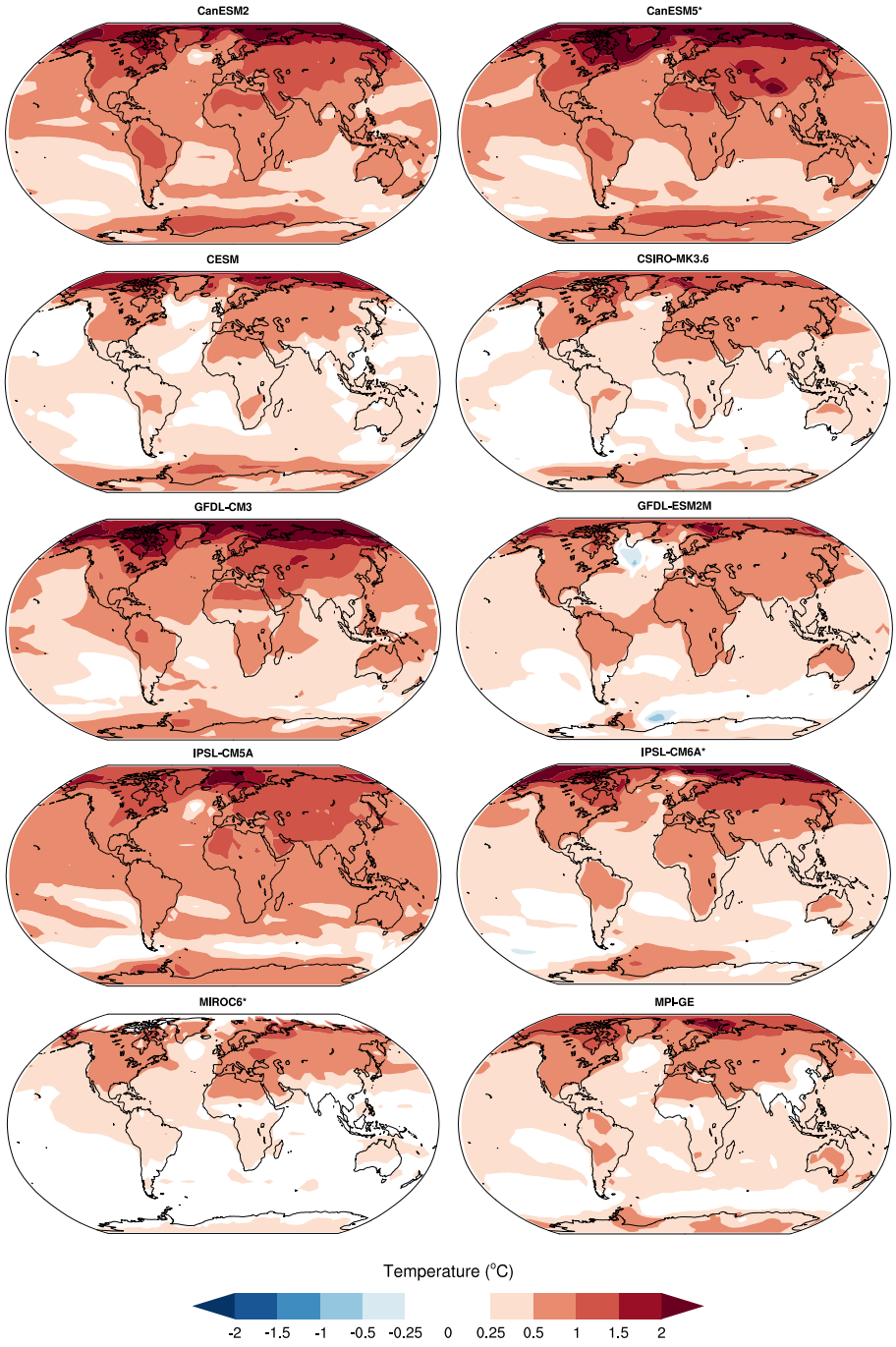

Figure S.8: Simulated mean surface temperatures in recent decades. Ensemble mean annual surface temperature anomalies simulated by different SMILEs averaged for the period of 1990–2014. Anomalies are relative to the period 1961–1990. Simulated data are regridded to match the observational grid (approximately 5°).

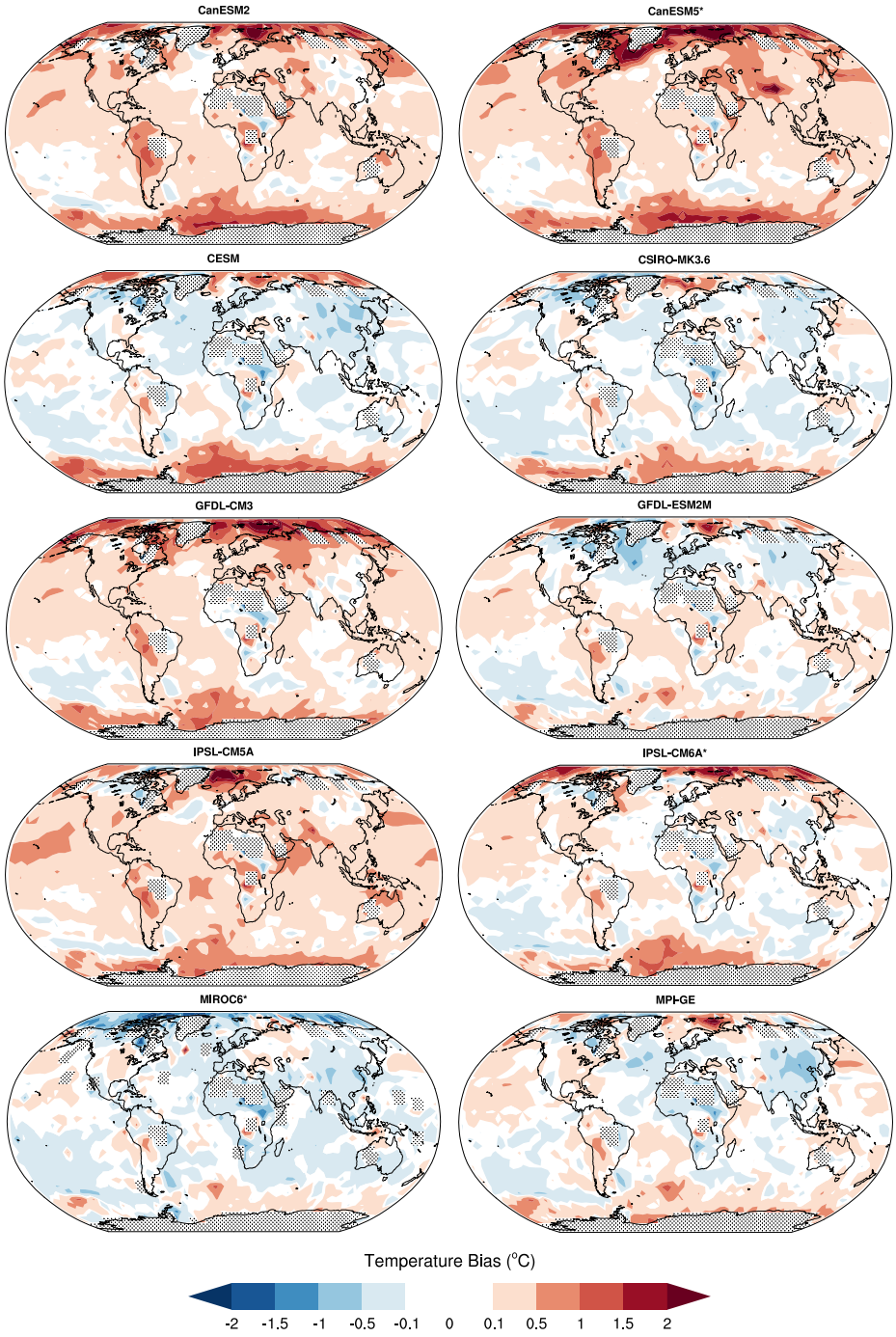

Figure S.9: Bias in recent mean surface temperatures. Difference of ensemble mean of the simulated annual surface temperature anomalies minus the average of observed annual surface temperature HadCRUT4 observed anomalies. This difference is averaged for the period of 1990–2014. Dotted areas represent regions where observations are not available. Anomalies are relative to the period 1961–1990. Simulated data are regridded to match the observational grid.