Supplement: Reduced global warming from CMIP6 projections when weighting models by performance and independence

Lukas Brunner[1], Angeline G. Pendergrass[2,1*], Flavio Lehner[1*], Anna L. Merrifield[1], Ruth Lorenz[1], and Reto Knutti[1]

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[2]National Center for Atmospheric Research, Boulder, CO, USA
[*]Now at: Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

S1 Summary

This supplementary material includes additional method details, as well as tables and figures supporting the findings presented in the main paper.

- **Section S2**: Additional information for the calculation of performance diagnostics
- **Section S3**: Additional information for the performance shape parameter (σ_D) calibration
- **Section S4**: Additional information for the independence shape parameter (σ_S) calibration
- **Section S5**: Additional information for the hierarchical clustering
- **Section S6**: Additional tables
  - **Table S1**: Table of performance shape parameter (σ_D) values as calculated by the weighting method.
  - **Table S2**: Table of weights, TCR, and warming per model.
  - **Table S3**: Table of temperature and TCR statistics for the unweighted and weighted distributions.
  - **Table S4 [external .csv file]:** List of all CMIP6 models and ensemble members used in this study as well as their institutions and DOIs.
  - **Table S5 [external .csv file]:** List of all CMIP5 models used in the study.
  - **Table S6 [external .csv file]:** List of all CMIP6 files used in the study including version date and tracking ID for tractability. Model issues are constantly updated and reported on the ES-DOC Errata page (https://errata.es-doc.org/static/pid.html). They can be accessed by searching for the tracking ID. For multiple version dates with the same tracking ID (in cases where more than one file exists for a given setting) the most recent version date is relevant.

- **Section S7**: Additional figures
  - **Figure S1**: Schematic of the performance shape parameter calibration.
  - **Figure S2**: Extended figure 2 showing all CMIP5 models.
  - **Figure S3**: Extended figure 3b showing all four combinations of scenarios and time periods.
  - **Figure S4**: Extended figure 8a showing distributions from a bootstrap approach.

S2 Additional information for the calculation of performance diagnostics

For a variable X_l^t which depends on a rolling horizontal index $l = l$ (lat, lon) and a time index t the time aggregations are calculated as follows. Climatology:

$$X_l^{\text{CLIM}} = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \left(X_l^t \right), \tag{1}$$

Anomaly:

$$X_l^{\text{ANOM}} = X_l^{\text{CLIM}} - \sum_l \left(w_l X_l^{\text{CLIM}} \right), \tag{2}$$

with $\sum_l w_l = 1$ being the area weights for each grid cell. Trend:

$$X_l^{\text{TREND}} = \text{TREND}_{t=t_1}^{t_2} \left(X_l^t \right), \tag{3}$$

with the TREND operator extracting the linear trend between t_1 and t_2 using ordinary least squares. Standard deviation:

$$X_l^{\text{STD}} = \text{STDDEV}_{t=t_1}^{t_2} \left(X_l^t - t * X_l^{\text{TREND}} \right), \tag{4}$$

with the STDDEV operator calculating the temporal standard deviation $(1/(N-1)\sum_l^N (x_l - \bar{x})^2)^{1/2}$ from (temporally) de-trended data. A diagnostic is then calculated as the area weighted root-mean-squared error between a model and the observations:

$$d = \sqrt{\sum_l w_l \left(X_l^{\text{AGG, Model}} - X_l^{\text{AGG, Obs}} \right)^2}, \tag{5}$$

with AGG denoting one of the time aggregations (CLIM, ANOM, TREND, or STD). So far we have used a notation which skipped some dependencies of d (and X) for simplicity. For the next steps we will generalise it to include the dependence on

the model index i and the initial-condition member index k, hence $d = d_i^k$. Like stated in equation (2) in the main paper, a mean diagnostic per model i is then given by:

$$d'_i = \frac{\sum_k^K d_i^k}{K_i} \tag{6}$$

Finally, consider multiple diagnostics indicated by the index a, which denotes the combination of variable X and time aggregation AGG (e.g., tasCLIM), hence $d'_i = d'^a_i$. The generalised distance D_i is then given as the weighted mean of the diagnostics, where each diagnostic is normalised by its median over all models:

$$D_i = \sum_a \frac{w_a d_i^a}{\text{MEDIAN}_i\,(d_i^a)}, \tag{7}$$

with $\sum_a w_a = 1$ being the weights for each diagnostic (see, e.g., figure 1 in the main paper).

S3 Additional information for the performance shape parameter (σ_D) calibration

The performance shape parameter σ_D is a constant that translates the observation-model distances into model weights (via equation (1) in the main paper). While different approaches exist to estimate this parameter, we here use a target specific calibration. This means that we use model information from the target period (which in our case is in the future) during the estimation process in order to avoid overconfident weighted projections for the selected target. Therefore, models can receive different weights for different targets (such as mid-century temperature change under SSP1-2.6 change versus end-of-century temperature change under SSP5-8.5) even though the same diagnostics are used in the historical period. This reflects the different levels of confidence based on the properties of the target we are interested in. Crucially, however, the rank of the models (i.e., the order from best to worst model in the ensemble) is the same in every case and only the "strength" of the weighting differs.

A schematic of the performance shape parameter (σ_D) calibration is shown in figure S1. A range of different sigma values are tested iteratively (ranging from $20\,\%$ to $200\,\%$ of the median of the generalised model-observation distance D_i) and the smallest value (i.e., strongest weighting) for which $80\,\%$ of perfect models fall within the 10-90 percentile range of the weighted target distribution is selected (Knutti et al., 2017). The σ_D values for all combinations of diagnostics and targets investigated in the main paper are summarised in table S1.

S4 Additional information for the independence shape parameter (σ_S) calibration

The independence shape parameter σ_S is a constant that translates the model-model distances into weights (via equation (1)). Similar to σ_D different approaches exist to determine an ideal value for σ_S (see, e.g., Lorenz et al., 2018; Brunner et al., 2019; Merrifield et al., 2020). Pragmatically speaking, the aim is to make sure that initial-condition ensemble members of a model are

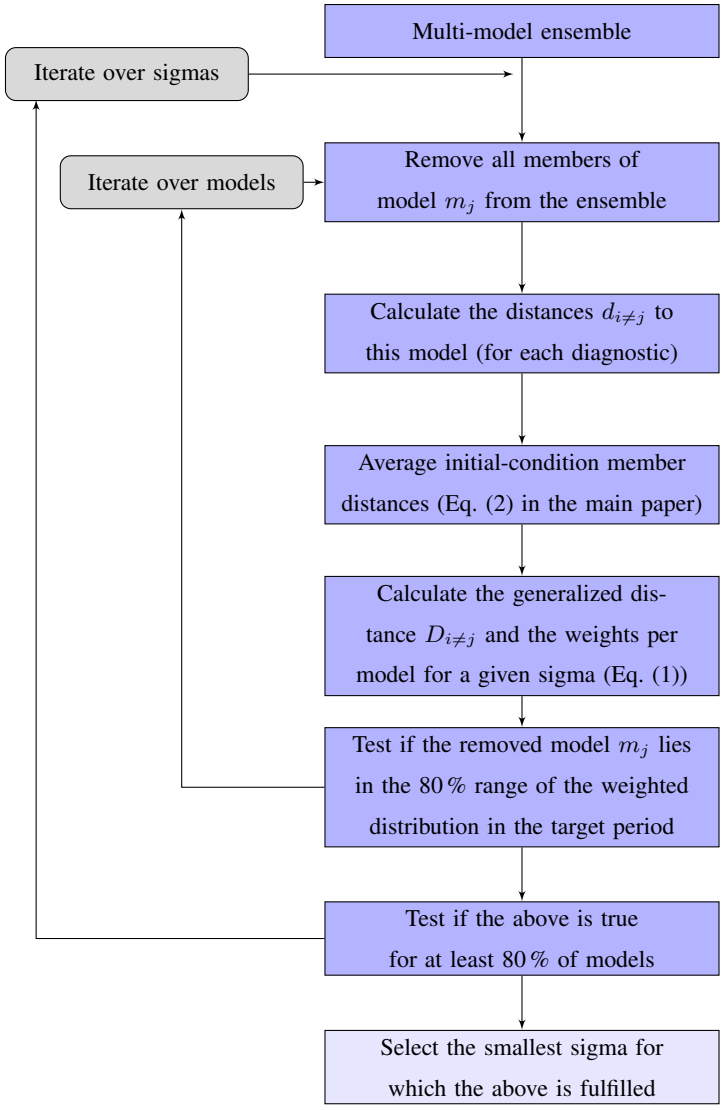

Figure S1. Schematic of the performance shape parameter calibration

Table S1. Model performance shape parameter σ_D for different target periods (sub-tables), SSPs (rows), and trend importance (columns) as well as the respective mean values. The mean value of $50\,\%$ highlighted in bold font is used throughout the manuscript.

2041-2060	$0\,\%$	$33\,\%$	$50\,\%$	$66\,\%$	$100\,\%$	Mean
SSP126	0.64	0.60	0.58	0.63	0.93	0.68
SSP585	0.47	0.37	0.35	0.31	0.29	0.36
Mean	0.55	0.48	0.46	0.47	0.61	0.52

2081-2000	$0\,\%$	$33\,\%$	$50\,\%$	$66\,\%$	$100\,\%$	Mean
SSP126	0.55	0.44	0.39	0.42	0.32	0.42
SSP585	0.47	0.37	0.39	0.67	1.20	0.62
Mean	0.51	0.40	0.39	0.55	0.76	0.52

Mean	$0\,\%$	$33\,\%$	$50\,\%$	$66\,\%$	$100\,\%$	Mean
SSP126	0.60	0.52	0.48	0.52	0.62	0.55
SSP585	0.47	0.37	0.37	0.49	0.74	0.49
Mean	0.53	0.44	**0.43**	0.51	0.68	0.52

recognised as copies (see figure 6 and corresponding discussion in the main paper), partly dependent models receive reduced weighting based on their similarity to other models in the ensemble and independent models are identified as such. To estimate σ_S we here follow the approach detailed in section 3 of the appendix of Brunner et al. (2019).

The resulting value we find is $\sigma_S = 0.54$. To put this in context we briefly look into the composition of the multi-model ensemble used: it consists of 33 different models with up to 50 realisations and a total of 129 runs. The median of the generalised distance between two initial-condition ensemble members of the same model (which differ only due to internal variability) is about 0.12. The median of the generalised distance between two models (including models from the same institutions) is about 1.09. Looking at only two initial-condition ensemble members of the same model ($M = 1, 2$), which we here take to have the typical distance (0.12), the pure independence weighting becomes (derived from equation (1) in the main paper):

$$w_i^{\text{ind}} = \frac{1}{1 + \sum_{j \neq i}^{M} e^{-\left(\frac{S_{ij}}{\sigma_S}\right)^2}} = \frac{1}{1 + e^{-\left(\frac{0.12}{0.54}\right)^2}} = \frac{1}{1 + 0.952} = 0.512, \tag{8}$$

which is close to $\frac{1}{2}$ which we would expect for the idealised case. The independence weight for two different models taken to have the typical distance (1.09), in turn, becomes

$$w_i^{\text{ind}} = \frac{1}{1 + e^{-\left(\frac{1.09}{0.54}\right)^2}} = \frac{1}{1 + 0.017} = 0.983, \tag{9}$$

which would identify them as mostly independent. As mentioned in the main paper it is important to remind ourselves that the definition of independence used here does not hold in a purely statistical sense. It rather aims at reducing obvious inter-dependencies between models based on their output while assuming that the majority of models (after averaging initial-condition members) is mostly independent.

S5 Additional information for the hierarchical clustering

Here a short description of the hierarchical clustering used for creating the CMIP6 "family tree" in figure 5 of the main manuscript is given. We use an implementation from the Python SciPy package (https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html), which is based on work by Müllner (2011).

Consider an example distance matrix of four models A, B, C, and D with distances: A-B: 1, A-C: 3, A-D: 6, B-C: 2, B-D: 5, and C-D: 6. The first cluster is formed by the two models with the smallest distance A and B. Since we use the "average" method the distance of this cluster to the remaining models is the average of this elements: AB-C: 2.5 (mean of A-C and B-C) and AB-D: 5.5. The next cluster is formed by the now two closest "clusters" AB-C. This process is repeated until all models are connected. For figure 5 in the main paper the models then are sorted by decreasing number of branches from top to bottom. This sorting does not change the results and is only done for visual reasons; the order of models in the initial clusters is arbitrary.

S6 Additional tables

Table S2. List of CMIP6 models used including their weight, Transient Climate Response (TCR), and warming relative to the 1995-2014 baseline. The colours are locked to the values. Weights are coloured relative to equal weighting (which is about 0.03): x0.5 to x1.5 (white), up to x2 (lightest red), x2.5, x3, x3.5, and above (darkest red); equivalent for models with less than equal weight. TCR is coloured equivalent to figure 4 in the main paper and the values are taken from Tokarska et al. (2020), updated for more models.

			2041-2060		2081-2100	
Model	Weight	TCR	SSP1-2.6	SSP5-8.5	SSP1-2.6	SSP5-8.5
ACCESS-CM2	0.0499	2.11 °C	1.62 °C	2.08 °C	1.89 °C	4.85 °C
ACCESS-ESM1-5	0.0358	1.95 °C	1.15 °C	1.80 °C	1.34 °C	3.98 °C
AWI-CM-1-1-MR	0.0436	2.07 °C	0.92 °C	1.46 °C	0.92 °C	3.62 °C
BCC-CSM2-MR	0.0354	1.5 °C	0.98 °C	1.69 °C	0.89 °C	3.31 °C
CAMS-CSM1-0	0.0507	1.75 °C	0.60 °C	1.03 °C	0.68 °C	2.51 °C
CanESM5-CanOE	0.0019	2.64 °C	1.54 °C	2.55 °C	1.62 °C	5.82 °C
CanESM5	0.0013	2.66 °C	1.50 °C	2.51 °C	1.59 °C	5.79 °C
CESM2-WACCM	0.0106	1.98 °C	1.28 °C	1.93 °C	1.50 °C	4.78 °C
CESM2	0.0140	2.06 °C	1.21 °C	1.98 °C	1.43 °C	4.74 °C
CNRM-CM6-1-HR	0.0218	2.47 °C	1.46 °C	1.94 °C	1.71 °C	4.76 °C
CNRM-CM6-1	0.0170	2.13 °C	1.12 °C	1.74 °C	1.39 °C	4.87 °C
CNRM-ESM2-1	0.0192	1.92 °C	1.14 °C	1.76 °C	1.47 °C	4.46 °C
EC-Earth3-Veg	0.0092	2.61 °C	1.08 °C	1.80 °C	1.30 °C	4.40 °C
EC-Earth3	0.0079	2.49 °C	1.08 °C	1.70 °C	1.26 °C	4.43 °C
FGOALS-f3-L	0.0630	2.06 °C	0.88 °C	1.52 °C	0.88 °C	3.57 °C
FGOALS-g3	0.0069	1.57 °C	0.44 °C	1.26 °C	0.48 °C	2.76 °C
FIO-ESM-2-0	0.0643	2.24 °C	1.01 °C	1.69 °C	1.03 °C	4.32 °C
GFDL-ESM4	0.1287	1.61 °C	0.78 °C	1.29 °C	0.79 °C	3.11 °C
GISS-E2-1-G	0.0862	1.8 °C	1.16 °C	1.64 °C	1.22 °C	3.40 °C
HadGEM3-GC31-LL	0.0011	2.51 °C	1.52 °C	2.43 °C	2.00 °C	5.46 °C
INM-CM4-8	0.0142	1.32 °C	0.65 °C	1.34 °C	0.61 °C	2.90 °C
INM-CM5-0	0.0430	1.39 °C	0.75 °C	1.38 °C	0.68 °C	2.81 °C
IPSL-CM6A-LR	0.0224	2.31 °C	1.21 °C	1.96 °C	1.31 °C	4.97 °C
KACE-1-0-G	0.0347	2.19 °C	1.61 °C	2.26 °C	1.81 °C	4.62 °C
MCM-UA-1-0	0.0328	1.94 °C	0.86 °C	1.58 °C	0.93 °C	3.63 °C
MIROC6	0.0378	1.55 °C	0.81 °C	1.28 °C	0.81 °C	3.17 °C
MIROC-ES2L	0.0014	1.55 °C	1.02 °C	1.56 °C	0.97 °C	3.38 °C
MPI-ESM1-2-HR	0.0524	1.65 °C	0.66 °C	1.16 °C	0.67 °C	3.02 °C
MPI-ESM1-2-LR	0.0401	1.84 °C	0.64 °C	1.19 °C	0.60 °C	3.09 °C
MRI-ESM2-0	0.0189	1.65 °C	1.08 °C	1.77 °C	1.03 °C	3.68 °C
NESM3	0.0072	2.79 °C	1.07 °C	1.93 °C	1.03 °C	4.17 °C
NorESM2-MM	0.0223	1.34 °C	0.84 °C	1.40 °C	0.87 °C	3.32 °C
UKESM1-0-LL	0.0045	2.75 °C	1.77 °C	2.62 °C	2.08 °C	5.86 °C

Table S3. Overview of statistics from figure 8.

SSP1-2.6 2041-2060	Mean	Median	66 % range	90 % range
Unweighted	1.07	1.08	0.75 - 1.50	0.61 - 1.61
Weighted	0.96	0.92	0.67 - 1.18	0.62 - 1.61
Change	-0.11	-0.16	-32.00	-0.99

SSP5-8.5 2041-2060	Mean	Median	66 % range	90 % range
Unweighted	1.73	1.70	1.29 - 2.08	1.17 - 2.55
Weighted	1.54	1.55	1.20 - 1.80	1.08 - 2.13
Change	-0.19	-0.15	-24.05	-23.91

SSP1-2.6 2081-2100	Mean	Median	66 % range	90 % range
Unweighted	1.17	1.03	0.68 - 1.62	0.60 - 1.98
Weighted	1.01	0.92	0.68 - 1.33	0.60 - 1.84
Change	-0.16	-0.11	-30.85	-10.14

SSP5-8.5 2081-2100	Mean	Median	66 % range	90 % range
Unweighted	4.05	3.98	3.09 - 4.87	2.76 - 5.82
Weighted	3.63	3.50	3.05 - 4.41	2.73 - 4.85
Change	-0.42	-0.48	-23.60	-30.16

TCR	Mean	Median	66 % range	90 % range
Unweighted	2.01	1.98	1.55 - 2.51	1.35 - 2.74
Weighted	1.90	1.89	1.60 - 2.21	1.37 - 2.48
Change	-0.11	-0.09	-36.46	-20.14

S7 Additional figures

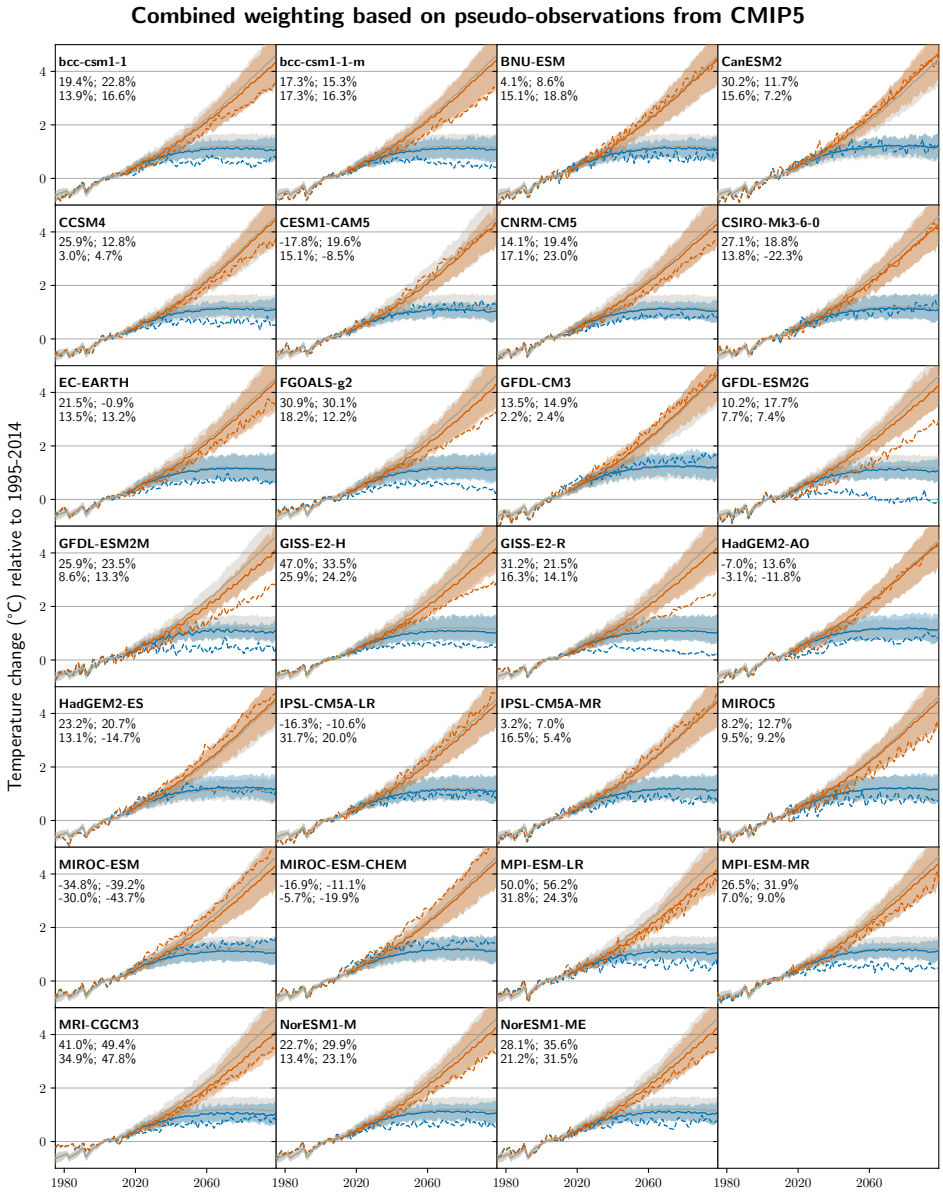

Figure S2. Similar to figure 2 but for all different pseudo-observations as given in the left top corner of each subplot. The values below each model name give the change in skill (CRPSS) for (top row) SSP5-8.5 as well as (bottom row) SSP1-2.6 in the mid- and end-of-century time periods.

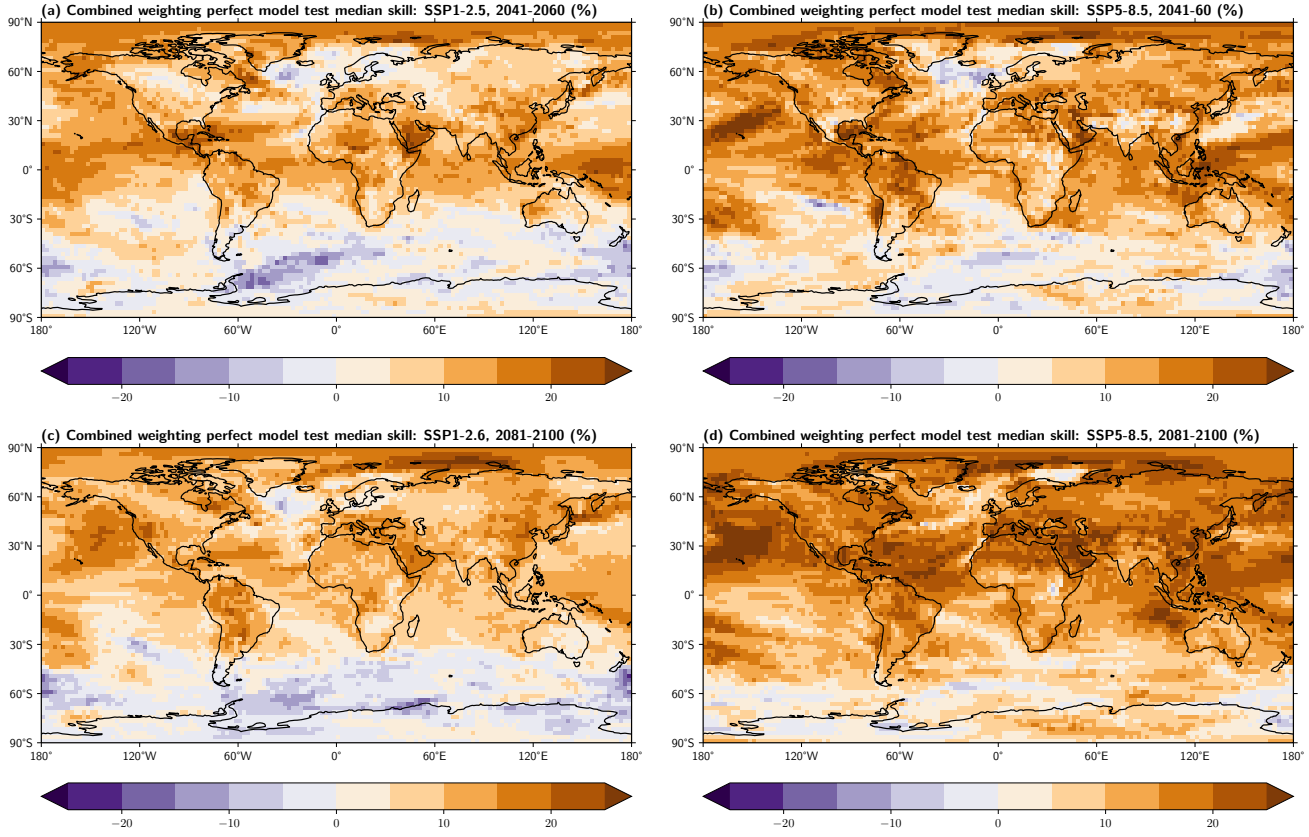

Figure S3. Same as figure 3b but for all four combinations of SSPs and time periods.

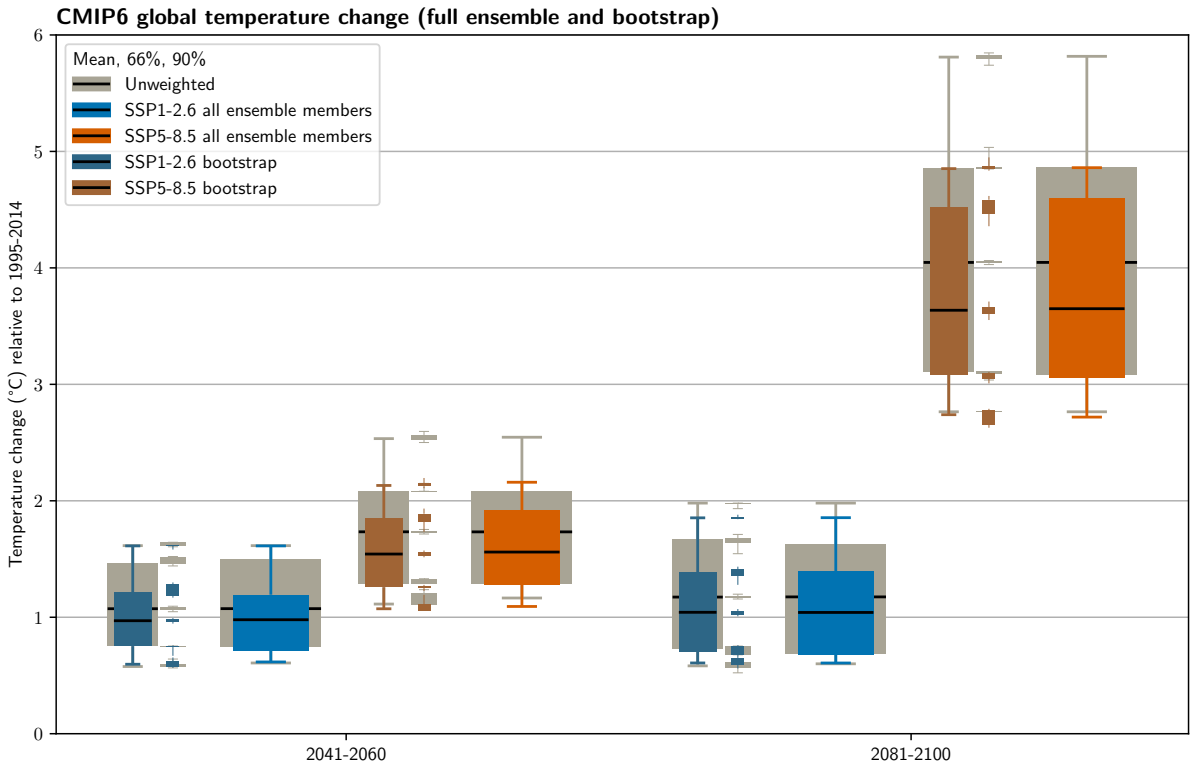

Figure S4. Unweighted (gray) and weighted (colors) temperature change for both periods and scenarios. The wide boxes show the same distributions as in figure 8a in the main paper based on all ensemble members. The larger narrow boxes show the median over all 100 bootstrap members. The tiny boxes show the uncertainty for each percentile.

References

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, Environmental Research Letters, 14, 124 010, https://doi.org/10.1088/1748-9326/ab492f, http://dx.doi.org/10.1038/ngeo3017, 2019.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V.: A climate model projection weighting scheme accounting for performance and interdependence, Geophysical Research Letters, 44, 1909–1918, https://doi.org/10.1002/2016GL072012, http://doi.wiley.com/10.1002/2016GL072012, 2017.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America, Journal of Geophysical Research: Atmospheres, 123, 4509–4526, https://doi.org/10.1029/2017JD027992, http://doi.wiley.com/10.1029/2017JD027992, 2018.

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., and Knutti, R.: An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles, Earth System Dynamics, 11, 807–834, https://doi.org/10.5194/esd-11-807-2020, https://esd.copernicus.org/articles/11/807/2020/, 2020.

Müllner, D.: Modern hierarchical, agglomerative clustering algorithms, pp. 1–29, http://arxiv.org/abs/1109.2378, 2011.

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., and Knutti, R.: Past warming trend constrains future warming in CMIP6 models, Science Advances, 6, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549, https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aaz9549, 2020.