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 about the calculation of performance diagnostics
- **Section S3**: Additional information about the performance shape parameter (σ_D) calibration
- **Section S4**: Additional information about the calculation of independence diagnostics
- **Section S5**: Additional information about the independence shape parameter (σ_S) calibration
- **Section S6**: Additional information about the hierarchical clustering
- **Section S7**: Additional tables and figures
  - **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.
  - **Figure S1**: Schematic of the performance shape parameter calibration.
  - **Figure S2 & S3**: Model-model distance matrices.
  - **Figure S4**: Similar to figure 5 but using a different metric.
  - **Figure S5**: Extended figure 2 showing all CMIP5 models.
  - **Figure S6**: Extended figure 3b showing all four combinations of scenarios and time periods.
  - **Figure S7**: Extended figure 8a showing distributions from a bootstrap approach.

S2 Additional information about 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:

$$X_l^{\text{CLIM}} = \frac{1}{t_2 - t_1} \sum_{t=t_1}^{t_2} \left(X_l^t \right), \tag{S1}$$

Anomaly:

$$X_l^{\text{ANOM}} = X_l^{\text{CLIM}} - \sum_l \left(w_l X_l^{\text{CLIM}} \right), \tag{S2}$$

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

$$X_l^{\text{TREND}} = \text{TREND}_{t=t_1}^{t_2} \left(X_l^t \right), \tag{S3}$$

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

$$X_l^{\text{STD}} = \text{STDDEV}_{t=t_1}^{t_2} \left(X_l^t - t * X_l^{\text{TREND}} \right), \tag{S4}$$

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:

$$d = \sqrt{\sum_l w_l \left(X_l^{\text{AGG, Model}} - X_l^{\text{AGG, Obs}} \right)^2}, \tag{S5}$$

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:

$$d_i' = \frac{\sum_k^K d_i^k}{K_i} \tag{S6}$$

Finally, consider multiple diagnostics indicated by the index a, which denotes the combination of variable X and time aggregation AGG (e.g., tasCLIM), hence $d_i' = d_i'^a$. The generalised distance D_i is then given as the weighted mean of

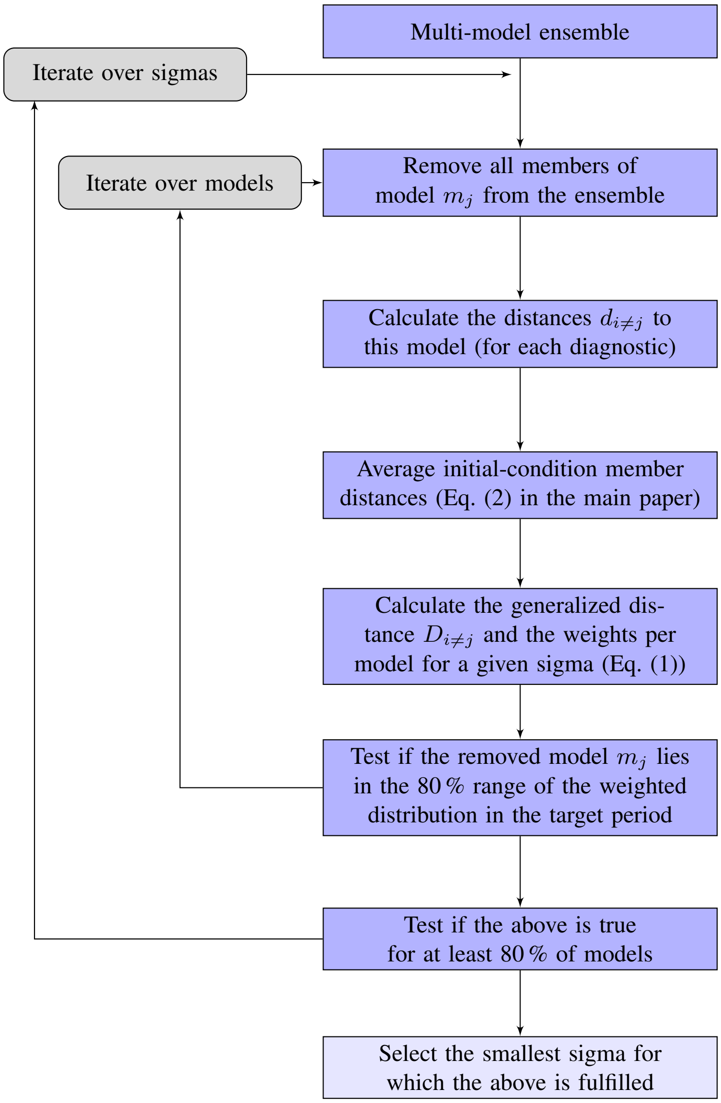

Multi-model ensemble
Iterate over sigmas
Iterate over models
Remove all members of model m_j from the ensemble
Calculate the distances $d_{i\neq j}$ to this model (for each diagnostic)
Average initial-condition member distances (Eq. (2) in the main paper)
Calculate the generalized distance $D_{i\neq j}$ and the weights per model for a given sigma (Eq. (1))
Test if the removed model m_j lies in the 80 % range of the weighted distribution in the target period
Test if the above is true for at least 80 % of models
Select the smallest sigma for which the above is fulfilled

Figure S1. Schematic of the performance shape parameter calibration

the diagnostics, where each diagnostic is normalised by its median over all models:

$$D_i = \sum_a \frac{w_a d_i'^a}{\mathrm{MEDIAN}_i\left(d_i^a\right)}, \tag{S7}$$

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

S3 Additional information about 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. Using only historical information might lead to overconfident results as a more skillful representation of the base state does not necessarily translate into a more skillful representation of the future as it would a priori assume that chosen diagnostics are relevant for the projections (Sanderson et al., 2015a). This means, in turn, that 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.

Finally, we here summarize considerations regarding the use of some future model information in the calibration of σ_D and the subsequent skill test using CMIP5 (figure 3 in the main paper), which also draws on the future (which might arise questions of circularity):

- Firstly, it is important to remind ourselves that the main information in the weights is always based on the comparison between models and the observations in the historical period. In particular, the ranking of the models (from best to worst) is determined solely by historical information.

- Secondly, the parameter estimation does not aim at maximizing (mean) skill but rather ensures that the results are not overconfident. To illustrate this point, consider an example case of very badly chosen diagnostics without any relationship to the target: in such a case, any separation into better or worse models in the target period based on the diagnostics is overconfident as it is based on pure chance. The parameter calibration, therefore, leads to a very large σ_D value and, hence, to an approximation of equal weighting. A subsequent test of the weighting skill in the target period can then reveal the actual increase in skill (or the lack thereof) given a set of diagnostics, which was not the optimization target of the performance shape parameter calibration.

- Thirdly, the pseudo-observations used for the skill calculation of the full weighting (figure 3 in the main paper) are drawn from the CMIP5 ensemble and have not been used in the parameter calibration. However, several models in CMIP6 are related to their CMIP5 predecessors and are, therefore, not fully independent. Nonetheless, we argue here that the degree of dependence is very limited due to several reasons:

    - Models evolve with time and there are about eight years of additional model development as well as additional observations to tune to between the CMIP5 and CMIP6 model generations. In particular, several CMIP6 models have been found to lead to considerably stronger warming than their predecessors and even then all the previous-generation CMIP5 models in general.

    - In the future period CMIP5 and CMIP6 are driven by different emission pathways (RCPs and SSPs, respectively) which lead to somewhat different radiative forcings (Forster et al., 2020).

    - For each CMIP5 pseudo-observation we exclude the directly related CMIP6 models from the calculation as listed in table S5 to further increase the independence.

S4 Additional information about the calculation of independence diagnostics

One crucial consideration when dealing with multi-model ensembles, such as CMIP6 or its predecessors, is the issue of model independence. Already for CMIP3 Jun et al. (2008) point out, that, for example, models from the same institution have highly correlated biases. They estimate the effective number of independent climate models to be considerably lower than the about 25 models included in total. In a 2010 editorial Knutti (2010) picks up this topic, asking if this should mean the "end of model democracy". Subsequently, many different approaches have been developed and tested to account for model dependence in multi-model ensembles. Some of them use background knowledge about the models' origin (e.g., Leduc et al., 2016) or components (e.g., Boé, 2018), but in most cases the models' output is used to infer their degree of inter-dependence. While all of these methods share a similar goal, their approaches on how to identify and quantify the degree of dependence arising from shared parametrisations, code, or full components is somewhat divergent. This is perhaps not surprising given that the notion of independence is often not cleanly defined for climate models (Annan and Hargreaves, 2017).

Existing methods have used regional (e.g., Steinschneider et al., 2015; Knutti et al., 2017; Lorenz et al., 2018; Brunner et al., 2019; Amos et al., 2020; Merrifield et al., 2020) or global (e.g., Masson and Knutti, 2011; Bishop and Abramowitz, 2013; Knutti et al., 2013; Sanderson et al., 2015a; Merrifield et al., 2020) model information for the estimation of model dependence. While the regional approach might be able to identify dependencies which are more specific for a given target region (e.g., two models having the same sea-ice component might lead to more dependence in projections at high latitudes than for the tropics), it also seams reasonable to interpret model dependence as a property of the multi-model ensemble in use, which does not change based on the target region so that both avenues have their justifications. Similarly, methods differ in the variables they use to derive model dependence, ranging from only a single variable (e.g., Masson and Knutti, 2011; Bishop and Abramowitz, 2013) to a basket of different variables (e.g.,

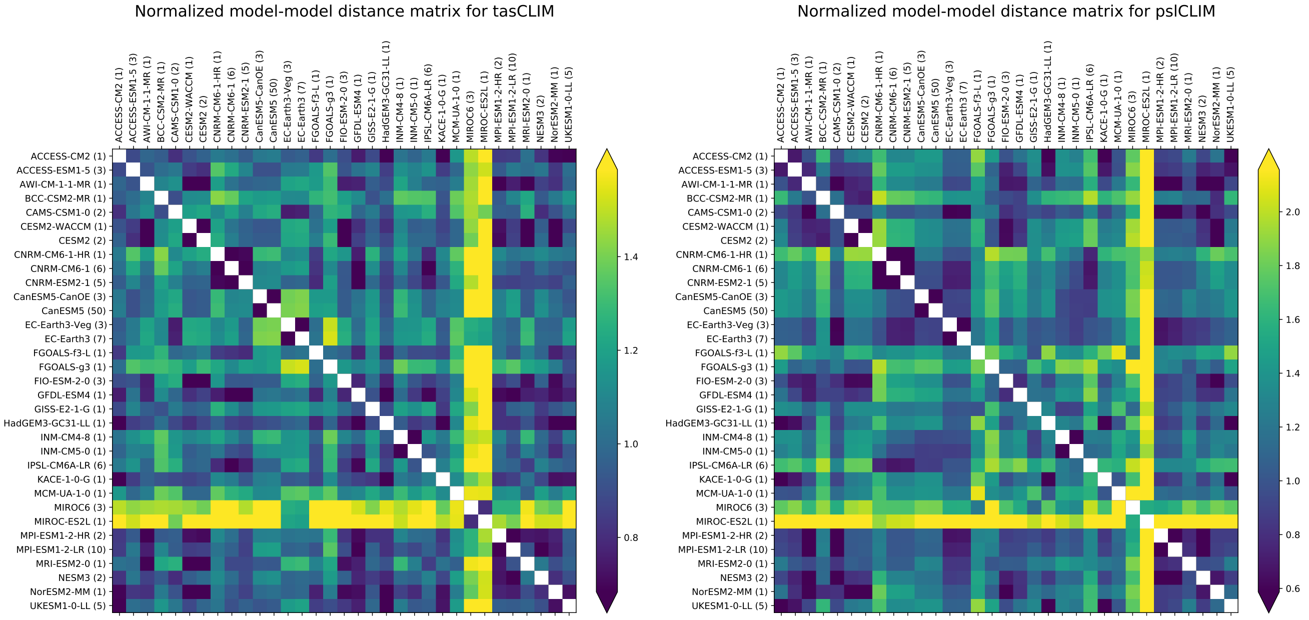


Figure S2. Model-model distance matrix s_{ij} normalized by its median for (a) surface air temperature and (b) sea level pressure climatologies between 1980-2014. Note the different colorbar-ranges in the panels.

Sanderson et al., 2015b; Knutti et al., 2017; Amos et al., 2020). Also, the way these model output fields are interpreted to translate them into a measure of model dependence differs between methods. Two approaches that are frequently applied are based either on the model correlation (e.g., Bishop and Abramowitz, 2013; Steinschneider et al., 2015) or the euclidean model distance (e.g., Sanderson et al., 2015a; Knutti et al., 2017).

Here we use the 35-year climatology of global, horizontally resolved fields of surface air temperature (tasCLIM) and sea level pressure (pslCLIM) as basis for our independence weighting. As described in the main paper, the generalized distance is calculated as difference between each model pair, which is equivalent to the difference of the model errors e:

$$e_{li} - e_{lj} = (X_{li}^{\text{AGG, Model}} - X_{l}^{\text{AGG, Obs}}) - (X_{lj}^{\text{AGG, Model}} - X_{l}^{\text{AGG, Obs}}) = X_{li}^{\text{AGG, Model}} - X_{lj}^{\text{AGG, Model}}, \tag{S8}$$

where the rolling index l runs over all longitudes and latitudes and the indices $i \neq j$ mark the different models. The model-model distance matrix is then calculated equivalent to (S5):

$$s_{ij} = \sqrt{\sum_{l} w_l \left(X_{li}^{\text{AGG, Model}} - X_{lj}^{\text{AGG, Model}} \right)^2} \tag{S9}$$

The two resulting distance matrices for tasCLIM and pslCLIM used in this study are shown in figure S2. The resulting generalised model-model distance matrix S_{ij} (calculated as the mean over both normalized s_{ij} equivalent to (S6) and (S7)) is shown in figure S3. Already from this visualisation, models from the same institution can be identified to be close (e.g., the three CNRM models), while other models (e.g., both MIROC models) are found to be quite far away from most other models in the ensemble. This method has the advantage of not needing any observations, compared to, for example, the approach of using the models' spatial error correlation distances similar to Bishop and Abramowitz (2013):

$$s_{ij}^{\text{CORR}} = 1 - \text{CORR}_l\,(X_{li}^{\text{ANOM, Model}} - X_{l}^{\text{ANOM, Obs}}, X_{lj}^{\text{ANOM, Model}} - X_{l}^{\text{ANOM, Obs}}) \tag{S10}$$

This means that the independence weights could, in theory, be based on variables for which no (or spare) observations are available as, for example, the pre-industrial control runs which often provide considerably more data than the historical runs. To test the robustness of our approach we apply the family tree clustering (figure 5 in the main manuscript) also based on the correlation distances described in (S11) (figure S4). The resulting patterns are quite similar between both approaches, with most known model families again falling into the same cluster also when using the error correlation distance as metric. In the details there are also some differences such as the two MIROC models being considerably

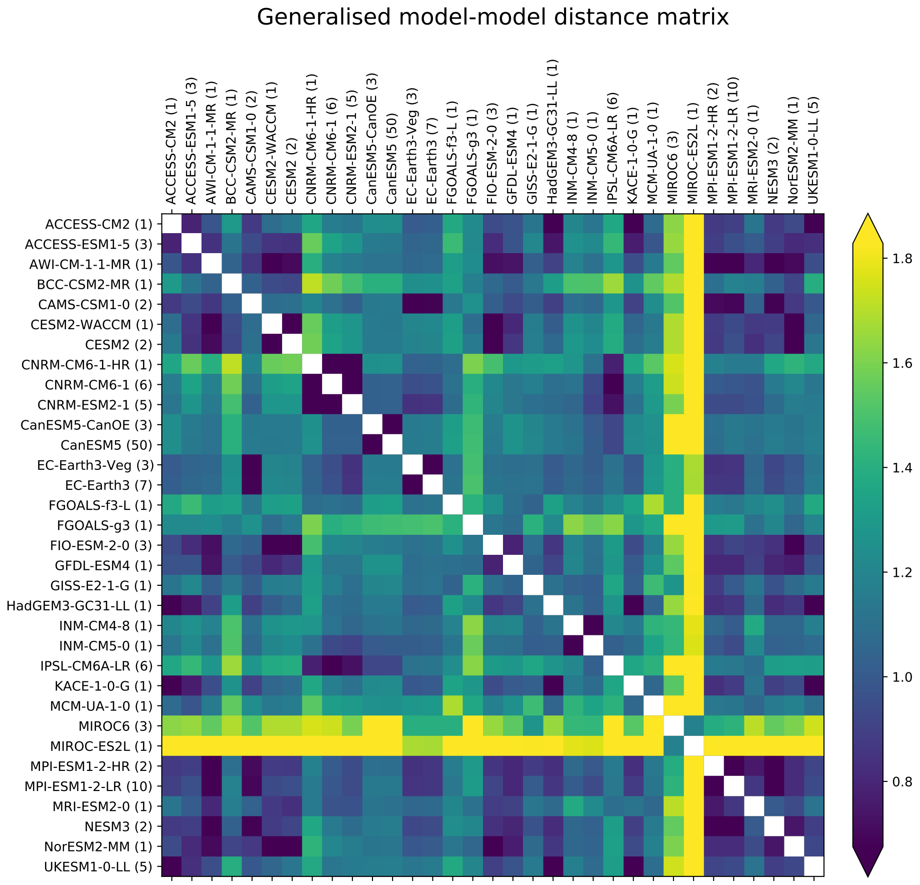


Figure S3. Generalised model-model distance matrix S_{ij}, calculated by averaging the two matrices shown in figure S2.

closer related when basing their independence on the error correlation distance. In general, however, we do not expect any major differences in the weighting using either metric, in particular since the weighting (in our case) is dominated to a large degree by several models receiving quite low performance weights. Indeed, calculating the weighted distributions (equivalent to figure 8 in the main manuscript) based on error correlation distances, reveals only minor differences (not shown). A more detailed analysis and comparison of the differences between the different approaches to model independence constitutes an interesting topic for further research but is outside the scope of this study.

S5 Additional information about 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 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):

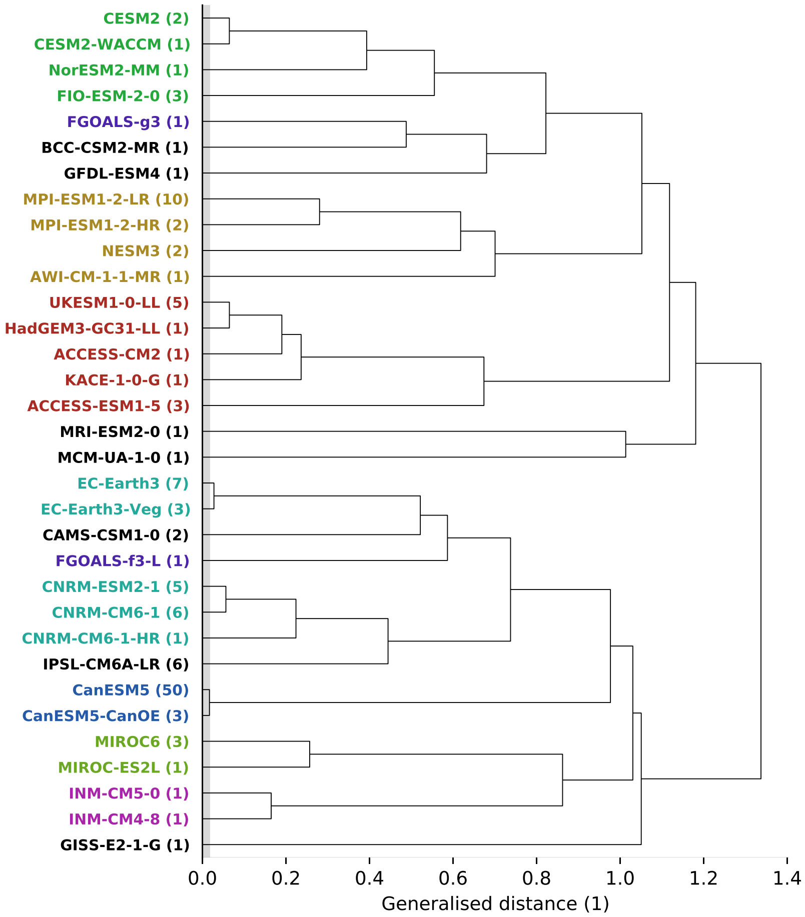


Figure S4. Similar to figure 5 in the main paper but based on error correlation distances.

$$w_i^{\text{ind}} = \frac{1}{1+\sum_{j\neq i}^{M} e^{-\left(\frac{S_{ij}}{\sigma_S}\right)^2}} = \frac{1}{1+e^{-\left(\frac{0.12}{0.54}\right)^2}} = \frac{1}{1+0.952} = 0.512, \tag{S11}$$

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

$$w_i^{\text{ind}} = \frac{1}{1+e^{-\left(\frac{1.09}{0.54}\right)^2}} = \frac{1}{1+0.017} = 0.983, \tag{S12}$$

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.

S6 Additional information about 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 generalised model-model distance matrix (figure S3) is used as basis. In the resulting tree models 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. Internal variability is estimated using the distance between initial-condition ensemble members of the same model. For each model with more than one member we calculate the mean distance between the members. The estimate of internal variability is then calculated as median over all these mean distances.

S7 Additional tables and figures

See next pages.

References

Amos, M., Young, P. J., Hosking, J. S., Lamarque, J.-F., Abraham, N. L., Akiyoshi, H., Archibald, A. T., Bekki, S., Deushi, M., Jöckel, P., Kinnison, D., Kirner, O., Kunze, M., Marchand, M., Plummer, D. A., Saint-Martin, D., Sudo, K., Tilmes, S., and Yamashita, Y.: Projecting ozone hole recovery using an ensemble of chemistry-climate models weighted by model performance and independence, Atmospheric Chemistry and Physics Discussions, 2020, 1–26, https://doi.org/10.5194/acp-2020-86, https://www.atmos-chem-phys-discuss.net/acp-2020-86/, 2020.

Annan, J. D. and Hargreaves, J. C.: On the meaning of independence in climate science, Earth System Dynamics, 8, 211–224, https://doi.org/10.5194/esd-8-211-2017, 2017.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Climate Dynamics, 41, 885–900, https://doi.org/10.1007/s00382-012-1610-y, 2013.

Boé, J.: Interdependency in Multimodel Climate Projections: Component Replication and Result Similarity, Geophysical Research Letters, 45, 2771–2779, https://doi.org/10.1002/2017GL076829, http://doi.wiley.com/10.1002/2017GL076829, 2018.

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.

Forster, P. M., Maycock, A. C., McKenna, C. M., and Smith, C. J.: Latest climate models confirm need for urgent mitigation, Nature Climate Change, 10, 7–10, https://doi.org/10.1038/s41558-019-0660-0, 2020.

Jun, M., Knutti, R., and Nychka, D. W.: Spatial analysis to quantify numerical model bias and dependence: How many climate models are there?, J. Am. Stat. Assoc., 103, 934–947, https://doi.org/10.1198/016214507000001265, 2008.

Knutti, R.: The end of model democracy?, Clim. Change, 102, 395–404, https://doi.org/10.1007/s10584-010-9800-2, 2010.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophysical Research Letters, 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.

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.

Leduc, M., Laprise, R., de Elía, R., and Šeparović, L.: Is institutional democracy a good proxy for model independence?, J. Clim., 29, 8301–8316, https://doi.org/10.1175/JCLI-D-15-0761.1, 2016.

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.

Masson, D. and Knutti, R.: Climate model genealogy, Geophysical Research Letters, 38, 1–4, https://doi.org/10.1029/2011GL046864, 2011.

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.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A representative democracy to reduce interdependency in a multimodel ensemble, Journal of Climate, 28, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing interdependency in a multimodel ensemble by interpolation of model properties, Journal of Climate, 28, 5150–5170, https://doi.org/10.1175/JCLI-D-14-00361.1, 2015b.

Steinschneider, S., McCrary, R., Mearns, L. O., and Brown, C.: The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning, Geophys. Res. Lett., 42, 5014–5022, https://doi.org/10.1002/2015GL064529, 2015.

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.

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

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 but in blue 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 °C	1.08 °C	0.75 °C to 1.50 °C	0.61 °C to 1.61 °C
Weighted	0.98 °C	0.91 °C	0.71 °C to 1.19 °C	0.62 °C to 1.61 °C
Change	−0.09 °C	−0.17 °C	−36.00 %	−0.99 %

SSP5-8.5 2041-2060	Mean	Median	66 % range	90 % range
Unweighted	1.73 °C	1.70 °C	1.29 °C to 2.08 °C	1.17 °C to 2.55 °C
Weighted	1.56 °C	1.56 °C	1.28 °C to 1.91 °C	1.09 °C to 2.16 °C
Change	−0.17 °C	−0.14 °C	−18.99 %	−22.46 %

SSP1-2.6 2081-2100	Mean	Median	66 % range	90 % range
Unweighted	1.17 °C	1.03 °C	0.68 °C to 1.62 °C	0.60 °C to 1.98 °C
Weighted	1.04 °C	0.91 °C	0.68 °C to 1.40 °C	0.61 °C to 1.85 °C
Change	−0.13 °C	−0.12 °C	−24.47 %	−9.42 %

SSP5-8.5 2081-2100	Mean	Median	66 % range	90 % range
Unweighted	4.05 °C	3.98 °C	3.09 °C to 4.87 °C	2.76 °C to 5.82 °C
Weighted	3.65 °C	3.46 °C	3.06 °C to 4.59 °C	2.72 °C to 4.86 °C
Change	−0.40 °C	−0.52 °C	−13.48 %	−29.84 %

TCR	Mean	Median	66 % range	90 % range
Unweighted	2.01 °C	1.98 °C	1.55 °C to 2.51 °C	1.35 °C to 2.74 °C
Weighted	1.87 °C	1.83 °C	1.58 °C to 2.17 °C	1.38 °C to 2.43 °C
Change	−0.14 °C	−0.15 °C	−37.50 %	−24.46 %

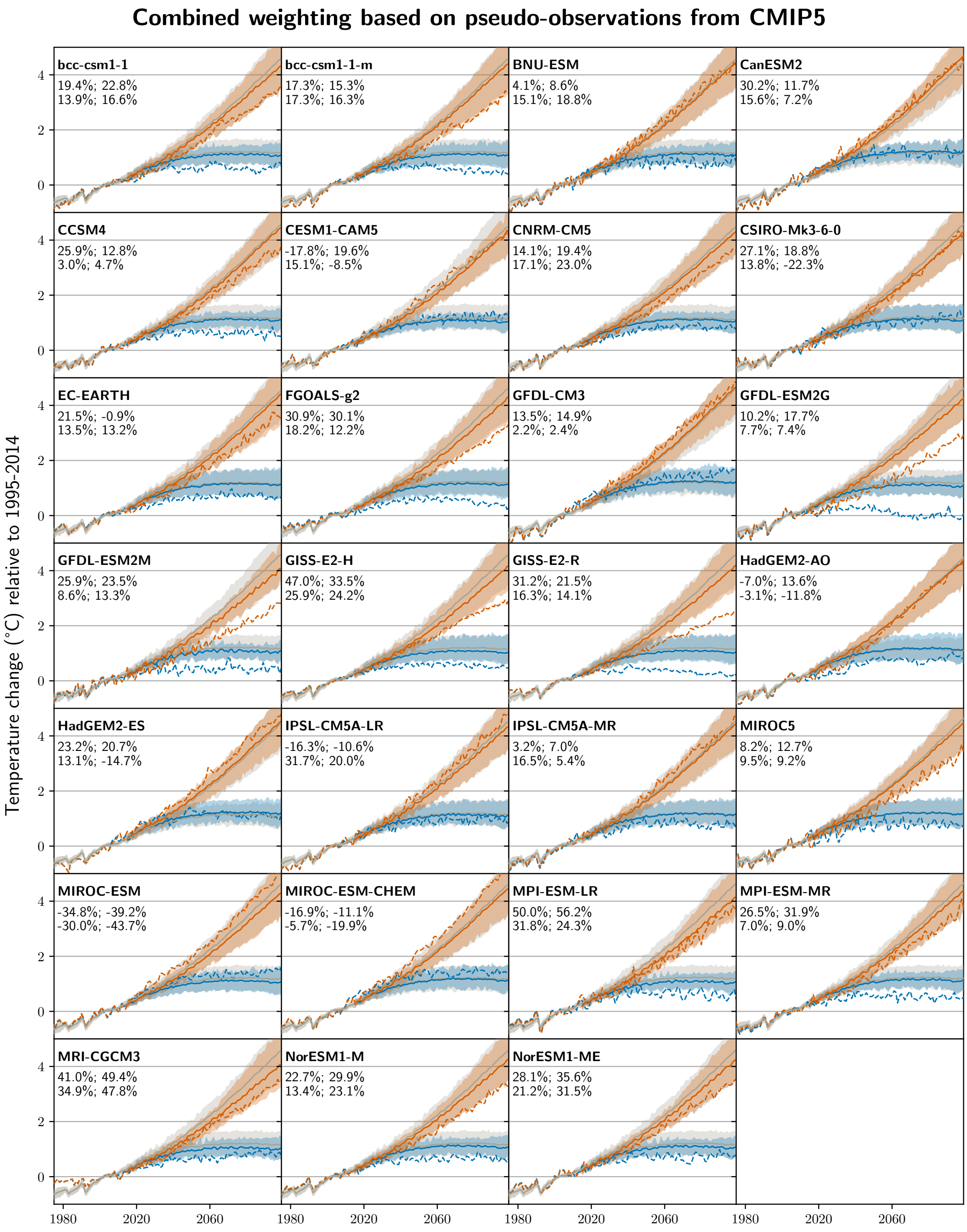
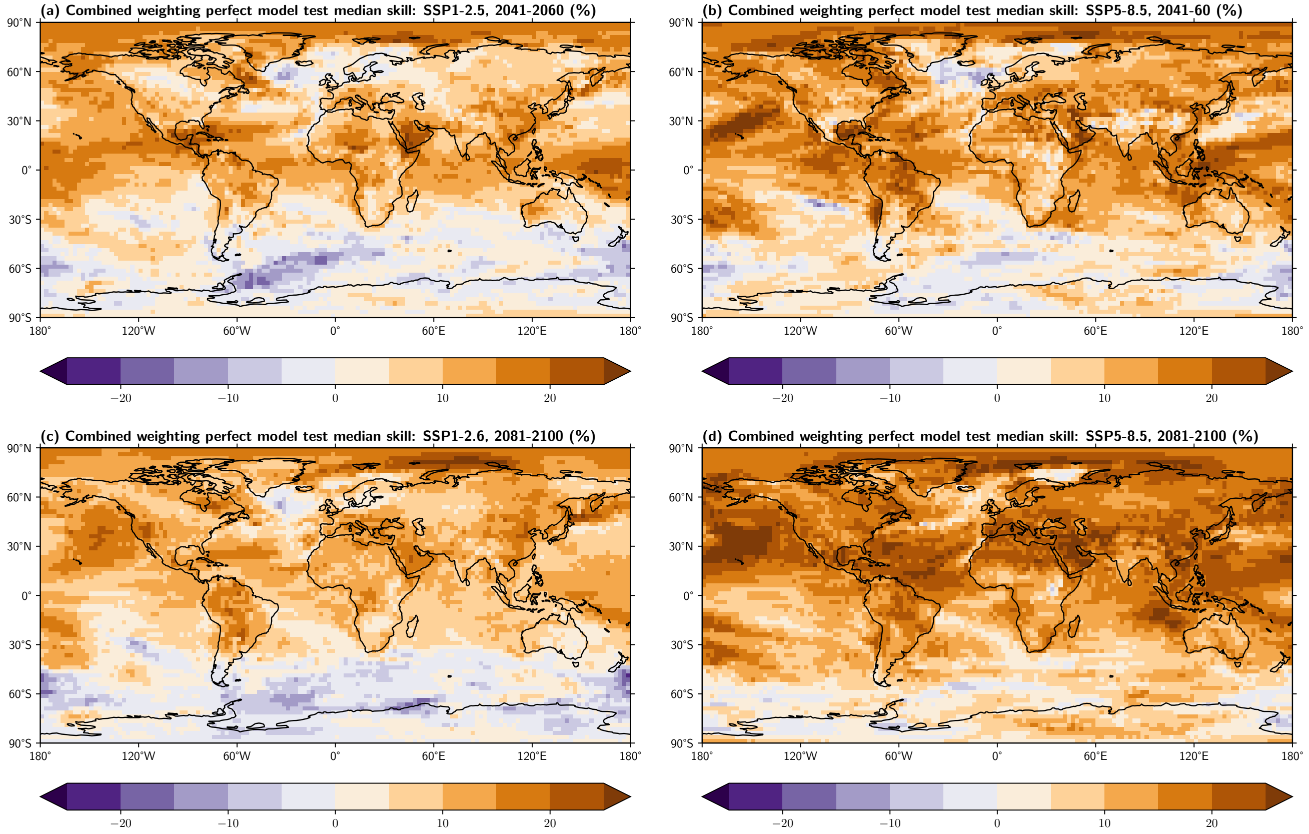


Figure S5. Similar to figure 2 but for all different pseudo-observations as given in the left top corner of each subplot. The values below each model name give the change in skill (CRPSS) for (top row) SSP5-8.5 as well as (bottom row) SSP1-2.6 in the mid- and end-of-century time periods.



(a) Combined weighting perfect model test median skill: SSP1-2.5, 2041-2060 (%)

(b) Combined weighting perfect model test median skill: SSP5-8.5, 2041-60 (%)

(c) Combined weighting perfect model test median skill: SSP1-2.6, 2081-2100 (%)

(d) Combined weighting perfect model test median skill: SSP5-8.5, 2081-2100 (%)

Figure S6. Same as figure 3b but for all four combinations of SSPs and time periods.

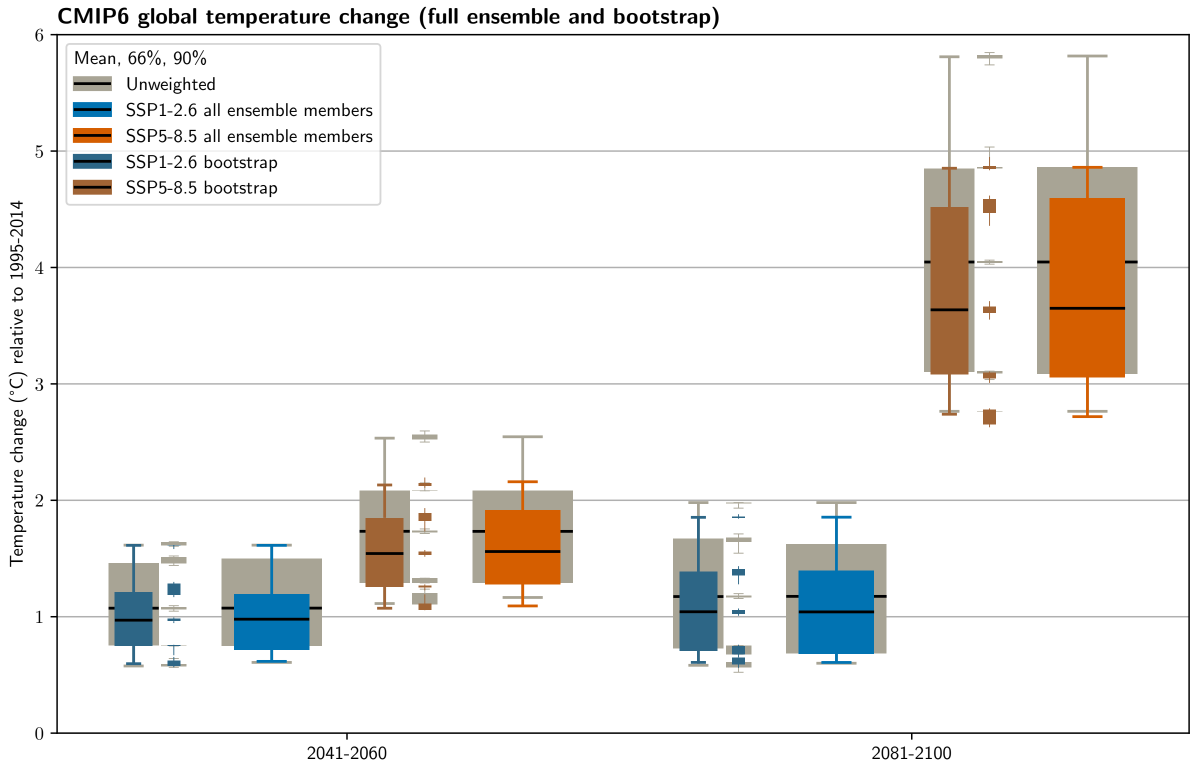


Figure S7. Unweighted (gray) and weighted (colors) temperature change for both periods and scenarios. The wide boxes show the same distributions as in figure 8a in the main paper based on all ensemble members. The larger narrow boxes show the median over all 100 bootstrap members. The tiny boxes show the uncertainty for each percentile.