Crossbreeding CMIP6 Earth System Models
With an Emulator for Regionally Optimized Land Temperature Projections

Lea Beusch1, Lukas Gudmundsson1, and Sonia I. Seneviratne1

1InstituteforAtmosphericandClimateScience,ETHZurich,Zurich,Switzerland

Abstract  The newest generation of the Coupled Model Intercomparison Project (CMIP6) exhibits a larger spread in temperature projections at the end of the 21st century than the previous generation. Here, a modular Earth System Model emulator is used to evaluate the realism of the warming signal in CMIP6 models on both global and regional scales, by comparing their global trends and regional response parameters to observations. Subsequently, the emulator is employed to derive large “crossbred” multimodel initial-condition ensembles of regionally optimized land temperature projections by combining observationally constrained global mean temperature trend trajectories with observationally constrained local parameters. In the optimized ensembles, the warmest temperature projections are generally reduced and for the coolest projections both higher and lower values are found, depending on the region. The median shows less changes in large parts of the globe. These regional differences highlight the importance of a geographically explicit evaluation of Earth System Model projections.

Plain Language Summary  The newest generation of state-of-the-art Earth System Models (ESMs) exhibits more uncertainty in temperature projections at the end of the 21st century than the previous generation. Here we use an ESM emulator, that is, a statistical tool which mimics the behavior of ESMs, to evaluate this newest generation of ESMs with respect to observations and to generate new ensembles combining the best performing features of these ESMs at both global and regional scales. The new ensembles consist of a large number of regionally optimized land temperature projections. At the end of the 21st century, the warmest regional temperature projections are generally lower in the optimized ensembles and the coolest projections are either reduced or enhanced, depending on the region. The temperature projections in the middle of the distribution are less affected in large parts of the globe. These regional differences emphasize the importance of also considering ESM performance in individual regions, not just the global mean, when evaluating climate change projections with current ESMs.

1. Introduction

The multimodel Earth System Model (ESM) ensembles of the Coupled Model Intercomparison Project (CMIP Eyring et al., 2016; Meehl et al., 2007; Taylor et al., 2012) play a crucial role in informing global and regional assessments of future climate change. Currently, ESM runs of the sixth CMIP phase (CMIP6 Eyring et al., 2016) are being produced and will serve as basis for the sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC). The runs available so far exhibit a large spread in land temperature projections at the end of the 21st century for a high-end emission scenario (SSP5-8.5 Riahi et al., 2017) both at global and regional scales (Figure 1). In fact, their spread even exceeds the full multimodel ensemble spread of the experiments of the fifth CMIP phase (CMIP5 Taylor et al., 2012) under a similar emission scenario (RCP8.5 Riahi et al., 2011).

In this study, we evaluate the realism of the past warming signal in individual CMIP6 models by comparing them to observations on global and regional scales. We subsequently derive large ensembles of regionally optimized and observationally constrained land temperature projections using a modular ESM emulator.
2. Data

Runs from 27 CMIP6 ESMs that span the historical time period (1870–2014) and the SSP5-8.5 emission scenario (2015–2100) are considered (listed in supporting information Table S1 and processed according to Brunner et al., 2020). To account for the largest ESM interdependencies (Abramowitz et al., 2019; Knutti, 2010) in terms of underlying modeling core, ESMs are grouped into “ESM families” according to their name. The 27 CMIP6 models considered in this study can be assigned to 19 ESM families (Table S1). Additionally, we use CMIP5 projections from 40 climate models over the same 1870–2100 period based on simulations for the historical period (1870–2005) and the RCP8.5 emission scenario (2006–2100) (Table S2). Furthermore, two spatially infilled, gridded observational products are employed to evaluate the realism of the ESM simulations while accounting for observational uncertainty (Cowtan and Way (COWTAN) and Berkeley Earth Surface Temperature (BEST) Cowtan & Way, 2014; Rohde et al., 2013). Observations are used from 1950–2018. Both observations and ESM runs were interpolated to a common $2.5^\circ \times 2.5^\circ$ grid. Yearly temperature anomalies with respect to a baseline of 1951–1980 are worked with and referred to as “temperature” for simplicity.

BEST and COWTAN consist of blended temperatures, meaning they report 2-m air temperature anomalies above land and sea ice and sea surface temperature (SST) anomalies above open ocean. SSTs warm systematically more slowly than air temperatures (Cowtan et al., 2015), and thus, blended temperatures are also employed for the ESMs to allow for a like-with-like comparison. To obtain blended temperatures from ESMs, 2-m air temperature anomalies, SST anomalies, sea ice concentrations, and land-sea mask fields are required. Note that we follow the BEST approach for blending ESM temperatures, which differs slightly from the COWTAN approach in the treatment of sea ice concentrations (Cowtan et al., 2015). In BEST, sea ice concentrations vary over time, whereas in COWTAN, they are fixed at climatological values.

For regional analyses, 26 SREX land regions (Seneviratne et al., 2012) and the Global Land region, which consist of all land grid points excluding Antarctica, are employed. A map with the SREX regions and their abbreviations is provided in the supporting information (Figure S1). To account for the fact that grid cell area depends on latitude and that coastal grid cells contain only a fraction of land cover, regional averages consist of area- and land-cover-weighted means.

Throughout the manuscript, a variety of variables, subscripts, and superscripts are employed. An overview is provided in supporting information Table S3.

3. ESM Emulator

MESMER, the Modular Earth System Model Emulator with spatially Resolved output employed in this study (Beusch et al., 2020), produces large, ESM-specific ensembles of yearly land temperature field time series which closely resemble initial-condition ensembles (e.g., Deser et al., 2012; Maher et al., 2019) at grid point to regional scales. Individual realizations of MESMER mimic single members of a true ESM initial-condition ensemble at a negligible computational cost and are referred to as emulations (Beusch et al., 2020).
Geophysical Research Letters

MESMER is based on the assumption that ESM-specific local land temperatures \( T_{m,s,t} \) for each climate model \( m \), at each grid point \( s \), and at each time step \( t \) can be expressed as a linear function \( f_{m,s} \) of the current global mean temperature \( T_{glob,m,t} \) and a spatiotemporally correlated local residual variability term \( \eta_{m,s,t} \) such that

\[
T_{m,s,t} = f_{m,s}(T_{glob,m,t}) + \eta_{m,s,t} = \beta_{m,s} \cdot T_{glob,m,t} + \eta_{m,s,t}.
\] (1)

Thereby, the global mean temperature trend, \( T_{glob,m,t} \), accounts for external forcings such as greenhouse gas emissions and volcanic eruptions, and \( \beta_{m} \) is its associated regression coefficient derived by ordinary least squares (OLS). \( T_{glob,m,t} \) is estimated by locally weighted scatterplot smoothing (LOWESS) of \( T_{glob,m,t} \) and a subsequent linear regression of the residuals from the smooth LOWESS trend to stratospheric aerosol optical depth. The global mean temperature variability, \( \gamma_{glob,var,m,t} = T_{glob,m,t} - T_{glob,trend,m,t} \), accounts for temporally correlated, internally generated variability and is described as an autoregressive (AR) process. \( \beta_{m,s} \) is used to regress this global variability to the local scale and \( \beta_{m,s} \) denotes the intercept term. Finally, \( \eta_{m,s,t} \) is represented by an AR process with spatially correlated innovations. A detailed description of the individual emulator components can be found in Beusch et al. (2020).

MESMER is trained for each of the 27 CMIP6 ESMs separately using data from 1870–2100 and employing all available initial-condition ensemble members. With the thereby derived ESM-specific emulator parameters, 1,000 emulations are generated for each ESM according to

\[
T_{e,m,s,t} = \gamma_{m,s} \cdot T_{glob,m,t} + \gamma_{var,m,s} \cdot \gamma_{glob,var,m,s,t} + \gamma_{int,m,s} + \eta_{m,s,t}.
\] (2)

whereby the subscript \( e_m \) refers to the ESM-specific emulations. The resulting "superensemble" contains 27,000 emulations and approximates a multimodel initial-condition ensemble. In the following, we refer to this ensemble as the "raw superensemble".

Please note that in Beusch et al. (2020), \( T_{m,s,t} \) referred to 2-m air temperature, whereas in this study, it refers to blended temperature. This does not affect land temperatures but leads to different \( T_{glob,m,t} \) and coastal grid cell values, as SSTs warm more slowly than 2-m air temperatures. Since the two \( T_{glob,m,t} \) quantities are directly linearly related (see supporting information Figure S2), MESMER can be trained with either one, leading to slightly different \( \beta_{m,s} \) values without impacting the results otherwise.

4. Emulator-Based ESM Evaluation

In general, local climate change can be characterized by global mean temperature increase (IPCC, 2013) alongside a linear scaling of local temperatures with global mean temperature (Seneviratne et al., 2016). Hence, we evaluate ESM performance by comparing the ESM runs to observations for these two features. Because natural climate variability affects both features, a direct comparison of observations, that is, the single realization of the real-world climate, and individual ESM runs, that is, single realizations of the ESM climate, is not possible. However, the ESM-specific raw emulations can be used to enable the comparison between ESMs and observations, since the emulations aim to approximate the true distribution of these two features the ESM would generate if it were run many times (Beusch et al., 2020). If the raw emulations perfectly approximated their ESM, the features of the ESM runs would fall within the 2.5th to 97.5th percentiles of the features of the raw emulations in 95% of all cases. In this study, the ESM is regarded as consistent with its raw emulations for a given feature, if the feature of at least one ESM run falls within these percentiles. If the feature of at least one observational product falls within these percentiles too, the ESM is additionally considered consistent with observations for this feature.

4.1. Global Mean Temperature Trend

During the last decades, observed global mean temperature \( T_{glob} \) has been linearly increasing with time (Cowtan & Way, 2014; Rohde et al., 2013). Here, we evaluate whether individual ESMs are able to reproduce the observed trend. For this purpose, we compute linear OLS trends in time for each observational product \( \gamma_{o,t} \), for each ESM run \( \gamma_{m,t} \), and for each raw emulation \( \gamma_{e,t} \) individually. We consider trends of varying length starting in 1975, 1980, 1985, 1990, 1995, and 2000 with the final year always being 2018.

Summarized results are provided in Figure 2, and the results for each time period are shown in supporting information Figure S3. Five ESMs from five ESM families are not consistent with the observed trend in any time period. All of them have been warming more rapidly than observations and continue to warm rapidly
Figure 2. Visualization of ESM performance for global trend and regional response to the global trend across multiple evaluation time periods. The color shows in how many time periods the ESM and its raw emulations are consistent with observations. The best performing models (blue) are furthermore grouped into ESMs, which are independent by name and consistent with their emulations (dark blue) and additional ESMs (light blue). (a) The ESM $T_{\text{glob}, \text{trend}}$ and the observed $T_{\text{glob}, \text{t}}$ time series. (b) ESM regionally averaged local response $\beta_{\text{reg}, \text{trend}} \cdot T_{\text{glob}, \text{trend}} + \beta_{\text{reg}, \text{int}}$ plotted against $T_{\text{glob}, \text{t}}$ for Global Land, East North America, and East Africa. The overlaid scatterplot shows observed $T_{\text{reg}, \text{t}}$ versus $T_{\text{glob}, \text{t}}$. (c) Tabular form of ESM performance for global trend and regional response in each region. White circles indicate regions where ESM runs and emulations are consistent with observations but the ESM runs are not consistent with the emulations. The associated time periods are counted as consistent. White crosses indicate regions where the ESM runs are consistent with observations but the emulations are neither consistent with the ESM runs nor the observations. The associated time periods are counted as inconsistent due to the inability of the raw emulations to represent either the ESM or the observations. A light faced symbol is employed if the result occurs in at least one evaluation time period and a bold faced symbol if it occurs in each one. The number of initial-condition ensemble members available for each ESM is provided in brackets behind the model name.

in the future. This finding is consistent with recent papers on individual high-end warming models (Sellner et al., 2019; Swart et al., 2019) and the CMIP6 ensemble (Forster et al., 2019; Tokarska et al., 2020). Eight models from six ESM families are consistent with observations in each time period (Figure 2). Nevertheless substantial intermodel uncertainty remains within these best performing models with their $T_{\text{glob}, \text{trend}}$ diverging by 2°C at the end of the 21st century (Figure 2a). Interestingly, their trajectories cluster into two distinct groups. We hypothesize that the models within each group share similar global feedbacks and heat uptake
efficiencies. This is also consistent with results of Tokarska et al. (2020) who show that the same subsets of models additionally share similarities in terms of transient climate response and equilibrium climate sensitivity.

4.2. Regionally Averaged Local Response

To evaluate the local grid point-level response to global mean temperature trend in ESMs, the emulator framework is applied again but this time different data are used to calibrate MESMER. Specifically, while MESMER has so far been trained for each ESM individually on the full ESM time period (1870–2100) by employing all initial-condition ensemble members (section 3), it is now trained on a single run and a set of shorter time periods instead. This setup allows for MESMER to also be calibrated with observations. The employed periods are 1950–2018, 1960–2018, and 1970–2018. The chosen periods result from a trade-off between maximizing the number of years available for training and maximizing the quality and quantity of the raw data in the observational products. For each period, MESMER is trained on each observational product, ESM run, and raw emulation individually according to equation (1), where the ESM subscript \( m \) is replaced with \( os \) for observations trained on a short time period \( st \), \( m_s \) for models, and \( e_{m,s} \) for raw emulations. Of the thereby derived parameters, the grid point-level regression coefficient \( \beta_{m,s,t}^{reg,trend} \) is of interest with \( x_{st} \) serving as a placeholder for \( o_s \), \( m_s \), or \( e_{m,s} \), depending on the data source.

The local response performance is evaluated on a regional level by averaging the grid point-level parameters \( (\beta_{m,s,t}^{reg,trend}) \) to regional parameters \( (\beta_{reg,trend}) \) before comparing them. Summarized results are shown in Figure 2, and the results for each individual time period are depicted in supporting information Figures S4–S6. If the ESM \( \beta_{m,s,t}^{reg,trend} \) are not consistent with the \( \beta_{reg,trend}^{st,trend} \) values of the raw emulations, the raw emulations do not account for all ESM processes relevant for explaining the behavior of the ESM during the short time periods used to calibrate MESMER here. In such regions, a second test is carried out to determine whether the ESM is nevertheless consistent with observations. Here, the ESM and the observations are considered as consistent if either \( \beta_{m,s,t}^{reg,trend} \) of an ESM run lies between the observational estimates \( \beta_{o_s,t}^{reg,trend} \) or vice versa. If the ESM and the observations are consistent, it is furthermore tested if the emulations and the observations are consistent (Figure 2). This test had to be carried out at least once in 41 of the total 729 model-region combinations, and thus, the result only marginally affects the outcome of the overall analysis.

In most regions, many of the 27 ESMs are consistent with observations in all evaluation time periods (Figure 2). Only in sixth out of 27 regions less than 50% of the ESMs are consistent with observations in all time periods. In these regions, the number of ESMs that are not consistent with observations during any time period lies between five from five different ESM families (Sahara, SAH) and twelve from ten different ESM families (North Asia, NAS). Overall, the performance of the ESMs appears higher on regional scales than on the global scale. Note that this could reflect the fact that the larger interannual variability of the climate signal at regional scales leads to a lower signal-to-noise ratio and thus a wider range of values, which are compatible with observations.

5. Crossbreeding to Avoid Loss of Information in Observationally Constrained ESM Ensembles

Since ESM biases remain largely stationary across different climate states (Krinner & Flanner, 2018), ESM evaluation results from the observational time period can be used to select a set of well-performing ESMs for future projections. However, our analysis revealed that there is no direct relation between ESM performance for global trend (here, \( T_{glob,trend}^{m_i,t} \)) and regional response (here, \( \beta_{m,i}^{reg,trend} \)) within individual ESMs, although the very best ESMs in terms of global trend frequently also perform well at regional scales (Figure 2).

The most naive approach to obtain projections that are consistent with observations would be to only consider models, which perform well with respect to all metrics. In that case, future projections would be based solely on MIROC6 and MPI-ESM1-2-LR, the only ESMs passing each test employed in this study (Figure 2). Naturally, this would constitute a large loss of information and a severe underestimation of intermodel uncertainty since many other ESMs also perform well in global trend and/or regionally averaged local response.

Alternatively, region-specific projections could be carried out by considering all ESMs which perform well in terms of both \( T_{glob,trend}^{m_i,t} \) and \( \beta_{m,i}^{reg,trend} \) in the region at hand (Figure 2). This would nevertheless disregard...
valuable ESM information, since ESMs which do not perform well in $T_{\text{glob}, \text{trend}}_{m.i}$ could not be employed for any regional projections.

Here, we propose a novel “crossbreeding” emulation approach to make full use of all ESMs that are consistent with observations at different scales. For this purpose, the equation to obtain raw emulations is reconsidered (equation (2)). To generate “crossbred” emulations, the global mean temperature trend of one ESM $j$, $T_{\text{glob}, \text{trend}}_{m.i}$, is combined with the local features of another ESM $i$ such that

$$T_{m.i,s,t} = \alpha_{\text{trend}}_{m.i} \cdot T_{\text{glob}, \text{trend}}_{m.i,t} + \beta_{\text{var}}_{m.i} \cdot T_{\text{var}, m.i,t} + \beta_{\text{int}}_{m.i} \cdot T_{\text{int}, m.i,t} + \eta_{m.i,s,t}.$$  

(3)

In each region, index $i$ runs across all ESMs whose local features are selected in this region and index $j$ across all ESMs whose global mean trend is selected. New crossbred emulations are generated by repeating each combination of $i$ and $j$ an equal number of times until the total number of crossbred emulations is as close to the total number of raw emulations (27,000) as achievable without exceeding it.

In the following, the results for two selection criteria are presented. The first criterion is referred to as “conservative removal” and excludes as few ESMs as possible by selecting each ESM which is consistent with observations in at least one of the considered time periods (Figure 2, gray ESMs are excluded). The second criterion is referred to as “best estimate” and solely selects ESMs, which are consistent with their emulations and observations in each of the considered time periods and which are additionally part of different ESM families (Figure 2, dark blue ESMs are selected). Thereby, the ESM providing the most initial-condition ensemble members of each family is selected such that future projections are based on the most robustly calibrated emulators. The resulting regionally optimized, observationally constrained, emulated ensembles are henceforth referred to as “crossbred superensembles”.

6. Temperature Projections: From the Original CMIP6 Multimodel Ensemble to Crossbred Superensembles

Regionally averaged time series highlight the added value of the raw superensemble compared to the actual CMIP6 ESM projections since the emulated ensemble samples the temperature projections phase space more thoroughly with 1,000 computationally cheap emulations available for each ESM (Figure 3). A caveat of this raw superensemble, however, is that its projections are based on all 27 available CMIP6 ESMs without taking ESM consistency with observations into account.

To improve the consistency of emulated projections with observations, crossbred superensembles, such as the conservative removal and the best estimate ensemble, can be used (Figure 3). Comparing the superensembles highlights that the impact of the observationally guided crossbreeding differs between regions, quantiles, and ESM selection criteria. Only the 95th percentile is reduced in each region and for both selection criteria.

Because the superensembles contain several thousand emulations, as opposed to the actual CMIP6 ensemble which contains only a few runs per ESM, quantiles can be reliably estimated at each grid point for any given year (Figure 4). Overall, temperature projection uncertainty is large in both the raw and the observationally constrained crossbred superensembles in the year 2100. The conservative removal superensemble shows a consistently reduced 95th and 50th percentiles compared to the raw superensemble, with the magnitude of the reduction being largest in the 95th percentile, especially in the Amazon region and eastern Asia. The best estimate superensemble's 95th percentile, on the other hand, shows a weaker reduction in many regions and even an increase in temperature projections in parts of the world. Its 50th percentile is generally similar to the raw superensemble. The spatial patterns of the changes in the 5th percentile of the two crossbred superensembles tend to be coherent but the best estimate superensemble is warmer.

7. Discussion and Conclusions

This study proposes a novel emulation approach to derive large ensembles of regionally optimized land temperature projections by combining global- and local-scale features of different ESMs through “crossbreeding”. For this purpose, global- and local-scale features of individual ESMs are evaluated with respect to observations to select ESM features, which are consistent with observations. Since local-scale ESM performance depends on the region at hand, the evaluation is carried out for different regions individually.
Figure 3. Regionally averaged temperature time series of the 27 CMIP6 ESMs, the raw superensemble, the conservative removal crossbred superensemble, and the best estimate crossbred superensemble for Global Land, East North America, and East Africa. Additionally, the difference between each crossbred superensemble and the raw superensemble is shown for different quantiles.
Figure 4. Maps of different temperature projection quantiles (5%, 50%, and 95%) in the raw, the conservative removal, and the best estimate superensembles for the year 2100. Additionally, the differences between the crossbred and the raw superensembles’ quantiles are given (\((\text{Crossbred } T - \text{Raw } T) \cdot 100\))\%. Note that the crossbred superensembles are optimized for each SREX region individually, which leads to discontinuities in the projections at the regions’ borders if different models are selected for the local response in adjacent regions. For land areas not covered by any SREX region, the Global Land optimization results are used.

In each region, crossbred emulations are created by drawing multiple realizations from each combination of the selected global mean temperature trends and the local features. Both a stringent and a soft ESM selection criterion are applied.

The advantages of the resulting crossbred superensembles are manifold. First, by retaining features from the full suite of ESMs which are consistent with observations at global and regional scales, they waste no valuable ESM information. Second, they provide a more complete sampling of the observationally constrained inter-ESM temperature projection uncertainty phase space as they couple global and local features of different ESMs. Third, due to their computational efficiency, it is feasible to generate superensembles containing thousands of emulations, and thus, quantiles can be reliably estimated at the grid cell level for any given year.
Nevertheless, some important caveats remain. The MESMER emulator is a simple statistical tool, which is subject to strong assumptions such as the linearity of the dependence of the local response on global mean temperature, the additivity of the local response and the variability around it, and the stationarity in time of the emulator parameters. Additionally, the only external forcing for local temperature in MESMER is global mean temperature. Other factors, which could be locally relevant too, such as human-induced aerosol emissions and land use forcing, are not considered. For a more complete discussion of the underlying assumptions of MESMER and their justification the reader is referred to Beusch et al. (2020). The caveats of the employed ESM selection approach consist of multiple testing, limited emulations’ sample size to estimate parameter distributions, and favoring ESMs with larger variability since they are more likely to reproduce the observed warming signal by chance. Lastly, also the treatment of inter-ESM dependencies (Abramowitz et al., 2019; Knutti, 2010) could be improved. Thus, while this study conceptually illustrates the possibility of crossbreeding ESM features with an emulator, further research is needed to fully exploit the capabilities of the resulting crossbred superensembles, for example, by improving the capacities of the considered emulator or by advancing model selection methodologies.

When applying the crossbreeding approach to CMIP6, the impacts are spatially diverse. Generally, the warmest temperature projections at the end of the 21st century are reduced whereas the median is less affected. For the coolest projections, complex patterns are observed with both increased and decreased projections depending on the region. Naturally, also the ESM selection criterion affects the projection patterns. Additionally, it should be noted that with CMIP6 still growing, these results cannot be regarded as final. Instead, they represent a snapshot based on the currently available CMIP6 climate model projections (status of April 2020).

In conclusion, this study introduces a novel emulation approach to recalibrate raw climate model projections based on historical performance in a modular way that retains the full suite of ESM features which are consistent with observations at global and regional scales. The thereby resulting superensembles of regionally optimized land temperature projections contain several thousand members and approximate recombined multimodel initial-condition ensembles at a negligible computational cost.

Acknowledgments
We acknowledge partial support from the H2020 CRESCENDO project (Grant Agreement 641816) and the FP7 Ideas ERC DROUGHT-HEAT project (Grant Agreement 617518). We furthermore acknowledge the World Climate Research Programme’s Working Group on Coupled Modeling and promotes CMIP. We thank the climate modeling groups (listed in supporting information Tables S1 and S2) for producing and making available their model output and the Earth System Grid Federation (ESGF) for archiving the data and providing access. We thank Urs Beyerle, Ian Sedláˇcek, Ruth Lorenz, and Lukas Brunner for retrieving and preprocessing the CMIP data and Iselin Medhaug, Reto Knutti, and Mathias Hauser for their scientific inputs. Lastly, we thank the two anonymous reviewers whose constructive feedback helped to substantially improve this study. The following data are available online: CMIP5 and CMIP6 data (https://esg-node.llnl.gov/projects/esgf-llnl/), the COWTAN data set (https://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html), BEST (https://berkeleyearth.org/data/), and the stratospheric aerosol optical depth data (https://data.giss.nasa.gov/modelforce/strataer/).

References
Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., et al. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. Earth System Dynamics, 10, 91–105. https://doi.org/10.5194/esd-10-91-2019
Beusch, L., Gudmundsson, L., & Seneviratne, S. I. (2020). Emulating Earth System Model temperatures with MESMER: From global mean temperature trajectories to grid-point level realizations on land. Earth System Dynamics, 11, 139–159. https://doi.org/10.5194/esd-11-139-2020
Brunner, L., Hauser, M., Lorenz, R., & Beyerle, U. (2020). The ETH Zurich CMIP6 next generation archive: Technical documentation. https://doi.org/10.5281/zenodo.3734128
Cowtan, K., Hausfather, Z., Hawkins, E., Jacobs, P., Mann, M. E., Miller, S. K., et al. (2015). Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. Geophysical Research Letters, 42, 6526–6534. https://doi.org/10.1002/2015GL064888
Cowtan, K., & Way, R. G. (2014). Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Quarterly Journal of the Royal Meteorological Society, 140, 1935–1944. https://doi.org/10.1002/qj.2297
Deser, C., Phillips, A., Boudette, V., & Teng, H. (2012). Uncertainty in climate change projections: The role of internal variability. Climate Dynamics, 38(3-4), 527–546. https://doi.org/10.1007/s00382-010-0977-x
Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development, 9(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016
Forster, P. M., Maycock, A. C., McKenna, C. M., & Smith, C. J. (2019). Latest climate models confirm need for urgent mitigation. Nature Climate Change Comment. https://doi.org/10.1038/s41558-019-0660-0
IPCC (2013). The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. https://doi.org/10.1017/CBO9781107015324.005
Knutti, R. (2010). The end of model democracy? Climate Change, 102, 395–404. https://doi.org/10.1007/s10584-010-9802-2
Krininger, G., & Flanner, M. G. (2018). Striking stationarity of large-scale climate model bias patterns under strong climate change. Proceedings of the National Academy of Sciences of the United States of America, 115, 9462–9466. https://doi.org/10.1073/pnas.1807912115
Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Kornblueh, L., Takano, Y., et al. (2019). The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. Journal of Advances in Modeling Earth Systems, 11, 1–21. https://doi.org/10.1029/2019MS001639
Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., et al. (2007). The WCRP CMIP3 multi-model dataset: A new era in climate change research. Bulletin of the American Meteorological Society, 88(9), 1383–1394. https://doi.org/10.1175/BAMS-88-9-1383
Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., et al. (2011). RCP 8.5—A scenario of comparatively high greenhouse gas emissions. Climatic Change, 109, 33–57. https://doi.org/10.1007/s10584-011-0449-y
Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O’Neill, B. C., Fujimori, S., et al. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. Global Environmental Change, 42, 153–168. https://doi.org/10.1016/j.gloenvcha.2016.05.009

Rohde, R., Muller, R., Jacobson, R., Perlmuter, S., & Mosher, S. (2013). Berkeley Earth temperature averaging process. Geoinformatics & Geostatistics: An Overview, 1, 1–13. https://doi.org/10.4172/2327-4581.1000103

Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., et al. (2019). UKESM1: Description and evaluation of the U.K. Earth system model. Journal of Advances in Modeling Earth Systems, 11, 4513–4558. https://doi.org/10.1029/2019MS001739

Seneviratne, S. I., Donat, M. G., Pitman, A. J., Knutti, R., & Wilby, R. L. (2016). Allowable CO2 emissions based on regional and impact-related climate targets. Nature, 529, 477–483. https://doi.org/10.1038/nature16542

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C., Kanae, S., Kossin, J., et al. (2012). Changes in climate extremes and their impacts on the natural physical environment. Managing the risk of extreme events and disasters to advance climate change adaptation. A special report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), pp. 109–230.

Swart, N. C., Cole, Jason N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019). The Canadian Earth System Model Version 5 (CanESM5.0.3). Geoscientific Model Development, 12, 4823–4873. https://doi.org/10.5194/gmd-12-4823-2019

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). 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

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