Assessment of large-scale indices of surface temperature during the historical period in the CMIP6 ensembles

Article

Accepted Version

Bodas-Salcedo, A., Gregory, J. M. ORCID: https://orcid.org/0000-0003-1296-8644, Sexton, D. M. H. and Morice, C. P. (2023) Assessment of large-scale indices of surface temperature during the historical period in the CMIP6 ensembles. Journal of Climate, 36 (7). pp. 2055-2072. ISSN 1520-0442 doi: 10.1175/JCLI-D-22-0398.1 Available at https://centaur.reading.ac.uk/109130/

To link to this article DOI: http://dx.doi.org/10.1175/JCLI-D-22-0398.1

Publisher: American Meteorological Society

Assessment of large-scale indices of surface temperature during the historical period in the CMIP6 ensemble

A. Bodas-Salcedo,[a] J. M. Gregory,[a,b] D. M. H. Sexton,[a] C. P. Morice[a]

[a] *Met Office Hadley Centre, Exeter, United Kingdom*

[b] *National Centre for Atmospheric Science, University of Reading, Reading, United Kingdom*

Corresponding author: Alejandro Bodas-Salcedo, alejandro.bodas@metoffice.gov.uk

ABSTRACT

We develop a statistical method to assess CMIP6 simulations of large-scale surface temperature change during the historical period (1850-2014), considering all timescales, allowing for the different unforced variability of each model and the observations, observational uncertainty, and applicable to ensembles of any size. The generality of this method, and the fact that it incorporates information about the unforced variability, makes it a useful model assessment tool. We apply this method to the historical simulations of the CMIP6 multi-model ensemble. We use three indices which measure different aspects of large-scale surface-air temperature change: global-mean, hemispheric gradient, and a recently-developed index that captures the sea-surface temperature (SST) pattern in the tropics ($SST^{\#}$; Fueglistaler and Silvers, 2021). We use the following observations: HadCRUT5 for the first two indices, and AMIPII and ERSSTv5 for $SST^{\#}$. In each case, we test the hypothesis that the model's forced response is compatible with the observations, accounting for unforced variability in both models and observations as well as measurement uncertainty. This hypothesis is accepted more often (75% of the models) for the hemispheric gradient than for the global mean, for which half of the models fail the test. The tropical SST pattern is poorly simulated in all models. Given that the tropical SST pattern can strongly modulate the relationship between energy imbalance and global-mean surface temperature anomalies on annual to decadal time scales (short-term feedback parameter), we suggest this should be a focus area for future improvements due to its potential implications for the global-mean temperature evolution in decadal time scales.

1. Introduction

The historical record of near-surface air temperature (SAT) is widely used as a performance metric for climate models (e.g. Braganza et al., 2003; Reichler and Kim, 2008). The time series of annual-mean anomalies is a benchmark against which models are tested, and it has been used to assess the credibility of a model's ability to provide information on future changes (e.g. Brunner et al., 2020). Recent research suggests that the later part of the historical period (1980 onwards) contains information about the sensitivity of the Earth's climate to external forcing (Flynn and Mauritsen, 2020; Dittus et al., 2020), although this relationship may not be as strong as suggested due to common model biases in the simulation of historical SST patterns (Andrews and Webb, 2018; Ceppi and Gregory, 2017), the sensitivity to biomass aerosols (Fasullo et al., 2022), or a nonnegligible contribution of internal variability on multi-decadal trends (McKinnon and Deser, 2018). The tropical SST patterns are strongly connected to regional precipitation anomalies, of relevance for the accurate drought-inducing teleconnections (e.g. Annamalai et al., 2013; Zinke et al., 2021). Also, the radiative forcing over the historical record is uncertain, mainly due to the role of aerosols (e.g. Smith et al., 2021), with important implications for the historical warming shown by models (e.g. Wang et al., 2021; Zhang et al., 2021). Potentially, all this information can be used to improve the model's response to external forcing subject to the constraints of process observations. However, there is no common approach on how to incorporate the historical record into model development.

For example, several modeling centres have directly "calibrated" or "tuned" historical simulations (i.e. adjusted them to improve realism of climate change simulation) during the developments of the models used for the Climate Model Intercomparison Project phase 6 (CMIP6; Eyring et al., 2016). During the development of the Energy Exascale Earth System Model version 1 (E3SMv1), a historical simulation was performed with a near-final version of the model, but no action was taken to change the historical performance in the final version (Golaz et al., 2019). Boucher et al. (2020) describe the developments and performance of the IPSL-CM6A-LR model. Although historical simulations were not used as part of the development, the r1i1p1f1 simulation was selected qualitatively among the first ~12 available historical members, based on a few key observables of the historical period. The historical warming of the MPI-ESM1.2-LR model was tuned by reducing its climate sensitivity during its development (Mauritsen et al., 2019).

The use of historical runs (or any coupled run with transient forcing) for tuning is not part of the Met Office Unified Model (UM) development protocol. The Hadley Centre models submitted to CMIP6 were not tuned to the historical record, although several model improvements were added to ensure that the total present-day radiative forcing was positive (Mulcahy et al., 2019). This approach was revised in the 2020 UM Users Workshop, where it was agreed that one of the key model errors was the simulation of the historical record. As a result, a Prioritised Evaluation Group (PEG) was created with the objective of improving the simulation of the historical global-mean surface temperature record. Also, in a recent review of the UM's Global Configuration (GC) development protocol, it was agreed that a small ensemble of historical simulations will be run during the final stage of the development cycle, opening the option to implement model changes that target the performance of the simulation of the historical record before the final configuration is delivered to the users. In this paper we present the first step towards incorporating historical information into the UM's development process. We develop a statistical method to test whether simulations of large-scale surface temperature change are realistic during the historical period (1850-2014). The method is applied to annual-mean time series of three surface temperature indices: global-mean, hemispheric gradient, and a recently-developed index that captures the sea-surface temperature (SST) pattern in the tropics ($SST^{\#}$; Fueglistaler and Silvers, 2021). We test the historical simulations of the CMIP6 ensemble and post-CMIP6 versions of the HadGEM3 and UKESM models. We use the term 'realistic' in a relative manner: a model that performs well against the tests described here can do so due to compensating errors (e.g. between forcings and feedbacks). Consequently, those models that we label as realistic in the present study could nonetheless be rejected once other metrics with additional observational evidence or process understanding are considered. This shortcoming is not specific to this methodology, and the method we propose here should be used along a wide range of diagnostics to provide a detailed assessment.The structure of the paper is as follows. Section 2 describes the observational and model data. The statistical methodology is detailed in Section 3, and Section 4 presents the results of the method applied to the CMIP6 historical ensemble. Finally, Section 5 discusses the results and conclusions.

2. Model data and observations

We use near-surface air temperature (CMIP variable *tas*) data from the *piControl* and *historical* experiments of the CMIP6 archive (Table 1), which are atmosphere-ocean coupled simulations. The *piControl* are unforced simulations with forcing agents set at pre-industrial levels (year 1850). After a spin up period, the CMIP6 protocol requests a minimum of 500 simulation years, but not all models fulfil this criterion. We explain how we deal with different lengths of the *piControl* time series in the next section.

The CMIP 6 protocol (Eyring et al., 2016) recommended that the *historical* experiments are run with the current best estimates of the time-evolving datasets of forcing agents: atmospheric composition, solar irradiance, natural and anthropogenic aerosols, and land-use change, but not all institutions followed the protocol. They branch from the *piControl* simulation, running from 1850 to 2014 (165 years). The CMIP6 protocol recommends running at least 3 *historical* simulations, branching from different points in the *piControl* simulations. We use 40 *piControl* simulations from the CMIP6 ensemble, plus simulations from GC4.0-LL and UKESM1.1-LL (Mulcahy et al., submitted), models developed after CMIP6.

We use three different observational datasets of surface temperature: the Met Office Hadley Centre/Climatic Research Unit global surface temperature data set version 5 (HadCRUT5.0.1.0; Morice et al., 2021), the Program for Climate Model Diagnosis and Intercomparison (PCMDI) SST reconstruction (Hurrell et al., 2008; Taylor et al., 2000), and the Extended Reconstructed Sea Surface Temperatures Version 5 (ERSSTv5; Huang et al., 2017). The baseline period used for all historical datasets is 1880-1919.

HadCRUT5 provides temperature anomalies on a lat-lon rectangular grid. Two variants of the same dataset are provided: a non-infilled version, with data in gridboxes where measurements are available; a more spatially complete version. For global and regional time series, the HadCRUT5 analysis error model contains two terms (Morice et al., 2020): the analysis error (ε_a), and the coverage error (ε_c). The analysis error combines the errors from the Gaussian process used in the statistical infilling and the instrumental errors. The analysis grids are not generally globally complete, particularly in the early observed record. Regions are omitted where there are insufficient data available to form reliable grid cell estimates. The coverage error represents the uncertainty in spatial averages arising from these unrepresented regions. The analysis error is represented by the 200 realizations of the historical record, whereas the coverage error is reported as a time series of standard deviations. We use the more spatially complete version, also termed as “HadCRUT5 analysis”. The HadCRUT5 analysis data set uses a statistical method to extend temperature anomaly estimates into regions for which the underlying measurements are informative. This makes it more suitable for comparisons of large-scale regional average diagnostics against spatially complete model data, although variability in "infilled" regions will be lower than where observed measurement data is present (Jones, 2016). We use the HadCRUT5 analysis as a reference dataset for two of the indices: global-mean, and hemispheric gradient. We use the global means calculated by averaging the hemispheric means, as recommended by Morice et al. (2021).

The $SST^{\#}$ index is defined as the difference between the average of the warmest 30% SSTs (actual values, not anomalies) and the domain average. The domain used for this particular metric is the Tropics, from 30^{o}S to 30^{o}N. This index represents the difference in SSTs between the convective regions and the tropical average, and it explains the anomalies in low cloud cover (and cloud radiative feedbacks) over the historical record due to changes in SST patterns (Fueglistaler and Silvers, 2021). The index is calculated using monthly-mean SSTs, and then annual averages are calculated. The same process is followed for both models and observations. Since this index cannot be calculated from local anomalies, a dataset that provides absolute temperature estimates is required. The PCMDI dataset provides monthly mean sea surface temperature and sea ice concentration data from 1870 to the present on a regular lat-lon grid. These data are designed to be used as boundary conditions for atmosphere-only simulations. They use the AMIP-II mid-month calculation (Taylor et al., 2000), which ensures that the monthly mean of the time-interpolated data is identical to the input monthly mean. Following the convention in other studies, we refer to this dataset as PCMDI/AMIPII. $SST^{\#}$ is subject to a large observational uncertainty (Fueglistaler and Silvers, 2021), attributed to the different methodologies used to provide information where observations are not available. Given that the PCMDI/AMIPII dataset doesn't provide a comprehensive error characterization, we use the ERSST5 to test the robustness of our results to the observational uncertainty in $SST^{\#}$. We have chosen the PCMDI/AMIPII and ERSST5 datasets because they fall at opposite ends of the spectrum of $SST^{\#}$ anomalies provided by observational datasets, spanning the range of structural uncertainties in the observational reconstructions of $SST^{\#}$.There is evidence of differences between near-surface atmosphere temperature and surface temperature diagnostics (e.g. Richardson et al., 2016). The Intergovernmental Panel on Climate Change Assessment Report version 6 (IPCC AR6; Gulev et al., 2021) quantifies the global-mean uncertainty of long-term trends by at most 10% in either direction, with low confidence in the sign of any difference in long-term trends. Jones (2020) supports the use of global near-surface air temperature model diagnostics with blended datasets of observed temperature changes.

3. Methodology

Let $H_o(t)$ be the timeseries of the observed historical record anomalies of any given surface temperature index. We decompose it as $H_O(t) = S(t) + U_O(t) + E_O(t)$, where $S(t)$ represents the forced signal, $U_O(t)$ is the unforced variability, and $E_o(t)$ is the total observational error. Similarly, for a given model we decompose any historical simulation of the same index as $H_M(t) = S(t) + D_M(t) + U_M(t)$. $D_M(t)$ represents a discrepancy term or error in the forced response, and $U_M(t)$ is the model's unforced variability.

If we hypothesize that the model's forced response is realistic (i.e. $D_M(t)=0$), then $H_M(t) - H_O(t) = U_M(t) - U_O(t) - E_O(t)$. We can test this hypothesis by comparing $H_M(t)-H_O(t)$ with the expected distribution of $U_M(t) - U_O(t) - E_O(t)$. In general, we have more than one realization of a model's *historical* experiment, each of them with a different realization of the model unforced variability. Since we only have a single sample of the real world's unforced variability, tests on individual ensemble members are not independent. We avoid this problem by formulating the test for ensemble means noting that S(t) (and DM(t)) are the same for each ensemble member: $\overline{H_M}(t) - H_O(t) = \overline{U_M}(t) - U_O(t) - E_O(t)$. The overbars represent the ensemble mean. With this formulation, the observations are used only once for each model ensemble with the contribution of their internal variability remaining constant with ensemble size (unlike the contribution of the model internal variability which reduces with ensemble size).

The problem is now reduced to the characterization of the distribution of the right-hand side of the equation. Ideally, U_O should be characterized from a long time series of the real system under no external forcing. Paleoclimatic proxy reconstructions are available only for restricted regions, and therefore not representative of the large spatial scales of interest for this study, as well as having larger errors. They have the additional complication that the external forcing is not zero during the paleoclimate record. Therefore, we instead assume that unforced simulations of the multi-model ensemble provide us with a reasonable estimate of the real world's unforced variability, an approach that has been used in other studies (e.g. Gillet et al., 2002). Hence, we characterize $\overline{U_M}$ and U_O using *piControl* simulations.

The sub-sections below describe the next steps in the methodology: calculation of the observational error term; estimation of the distribution of $\overline{U_M}(t) - U_O(t) - E_O(t)$ using *piControl* simulations; definition of the metric and calculation of its control distribution; testing the historical ensembles; interpreting the tests.

a. Calculation of the observational error

For the HadCRUT5 observations, we combine analysis and coverage errors into a single error term (E_o) as follows. We add samples of a normally-distributed random variable of zero mean and variance $Var(\varepsilon_c(t))$ to the residuals of the 200 realizations of the HadCRUT5 analysis. The total error inherits the autocorrelation characteristics of the analysis error, which is correlated in time. E_o is then modelled by drawing random samples from this 200-member ensemble of realizations. The time-dependence of E_o for the global-mean is shown in Figure 1. The black lines show the 95% confidence interval (comparable to the orange range in Figure 2 of Morice et al., (2020)). In general, the observational error decreases with time, apart from periods of international conflicts. The time-dependence of E_o for the hemispheric difference is very similar to that of the global-mean, but larger in magnitude.

For the SST# index, we don't include an error term due to lack of error information in the observational datasets. However, we repeat the analysis with two different observational datasets to test the robustness of the results.

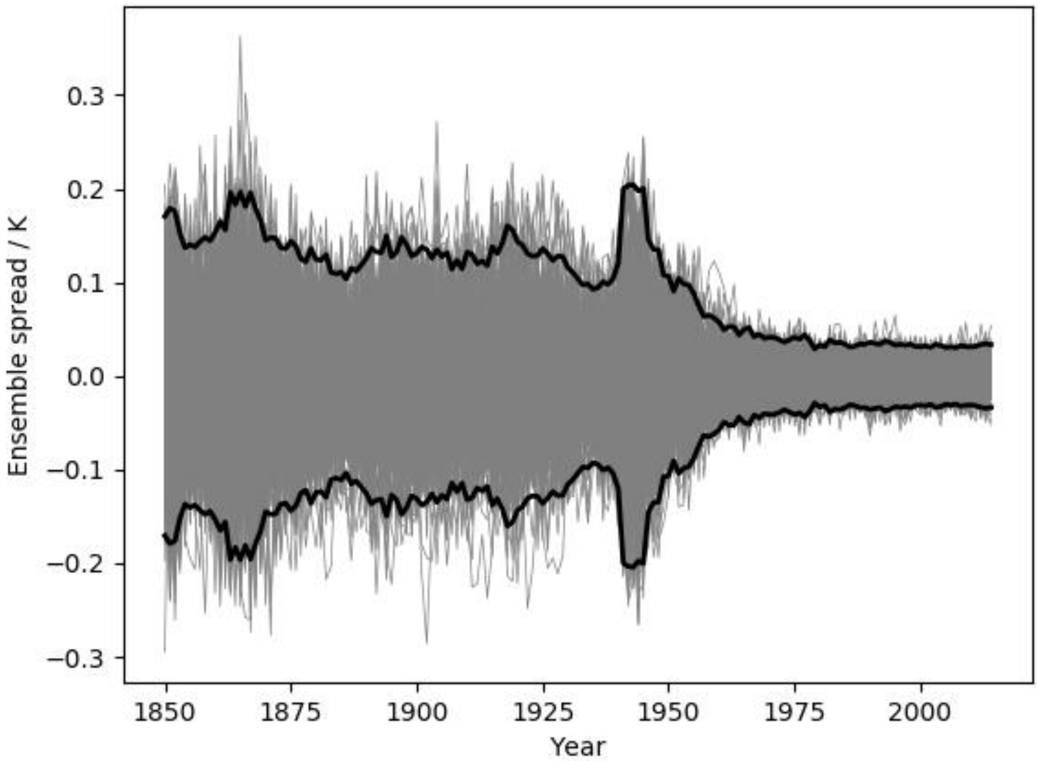

Figure 1. Total observational error (E_o) of the global-mean metric. The grey lines show the residuals of individual realizations of the HadCRUT5 global-mean analysis, including a randomly-generated contribution that accounts for the coverage error. The black lines are the bounds of the 95% confidence interval.

b. Construction of the unforced distribution of differences

Here we are concerned with the generation of random samples of $\overline{U_M}(t) - U_O(t) - E_O(t)$ using *piControl* simulations. Although the *piControl* simulations are started after a spin-up that is discarded, they are not in complete equilibrium (Eyring et al., 2016). For each model's control timeseries, we construct a linearly-detrended time series (X(t)) using the entire length of each control simulation. This increases the likelihood of adding noise to the detrended data (Sen Gupta et al., 2013; Jones et al., 2013), but some models show significant unforced variability on centennial timescales, which would be spuriously reduced by detrending shorter segments (Parsons et al., 2020).

We split the detrended control time series X(t) into non-overlapping segments 165 yr long, equal to the length of the CMIP6 historical simulations. The *piControl* simulations differ in length between models, so to give (nearly) equal weight to each model we use up to 3 segments of each piControl simulation. We also decide to retain models with shorter control time series. With these constraints, we use 41 *piControl* simulations, 32 of them with 3 segments, 5 with 2 segments, and 4 with only one segment. This gives 110 segments of piControl simulations of equal length. Then, we subtract the time average of the segment, so that the mean value of each segment is zero by construction. We call $U_{piControl}(t)$ to these detrended, 165 yr long, zero-average piControl samples of the unforced variability, which we use to generate samples of $\overline{U_M}(t;N_m) - U_O(t) - E_O(t)$. We sample both $U_M(t{:}\ N_m)$ and $U_o(t)$ from the ensemble of 110 $U_{piControl}(t)$ segments. For instance, for a *historical* ensemble with 10 members, we randomly draw 11 $U_{piControl}(t)$ segments, and average 10 of them to calculate $\overline{U_M}(t;N_m)$, and use the other one as $U_O(t)$. The $U_{piControl}(t)$ samples are drawn from the pool of *piControl* segments of all models, not only of the model whose *historical* ensemble is being tested. For GMSAT and hemispheric difference, $E_o(t)$ is randomly sampled from the ensemble of 200 realizations of the total HadCRUT5 total error as explained above, and the three timeseries are combined. We repeat this process 10000 times for each *historical* ensemble. For constructing the distribution of unforced differences, the only information extracted from the *historical* ensemble is its size N_m.

Other approaches for estimating internal variability exist, and a recent study by Olonscheck and Notz (2017) provide a brief description of the two main avenues and their caveats. We have used a method that is based on *piControl* simulations, which may be unsuitable if the unforced variability is state-dependent. However, Olonscheck and Notz (2017) show that the variability remains largely unchanged for historical simulations, even for those variables like sea ice area that show large changes in simulations of future warming. Therefore, we assume that the variability remains unchanged for the temperature indices used here and the amount of climate change in the historical period.

c. Definition of the metric: number of exceedances

Our interest is to characterize the quality of a historical ensemble of simulations against observations. As a metric of quality, in the next section (3d) we compute $\overline{H_M}(t) - H_O(t)$, for each model, and count the number of times that a running mean of the absolute value of this quantity exceeds a given value.

The samples of $\overline{U_M}(t) - U_O(t) - E_O(t)$ generated in the previous section (3b) serve as the basis to construct unforced distributions of this metric.

We define E(T, y, N_m), as the number of exceedances above a threshold T (in K) of a filtered time series of absolute values of $\left|\overline{U_M}(t) - U_O(t) - E_O(t)\right|$. The filter applied is a running mean with a window length of y years. We define a 2-dimensional rectangular grid in T and y, ranging between 0 and 0.3K, and between 1 and 10 years, respectively. We then calculate 10000 values of E for each combination (T, y). We use an absolute threshold in Kelvin, but the method could be easily reformulated in terms of a threshold defined in units of standard deviations of the unforced variability.

Figure 2 presents a an example of this process for the global-mean surface air temperature (GMSAT), leading to the calculation of one sample of E(0.1, 10, 5). The blue line shows one sample of $\overline{U_M}(t) - U_O(t) - E_O(t)$. The red line is the smoothed time series of the absolute value of the blue time series, using a y=10 yr running mean. The green line represents the temperature threshold T=0.1 K. The value of E(0.1, 10, 5) is the number of points from the red line that lie above the green line.

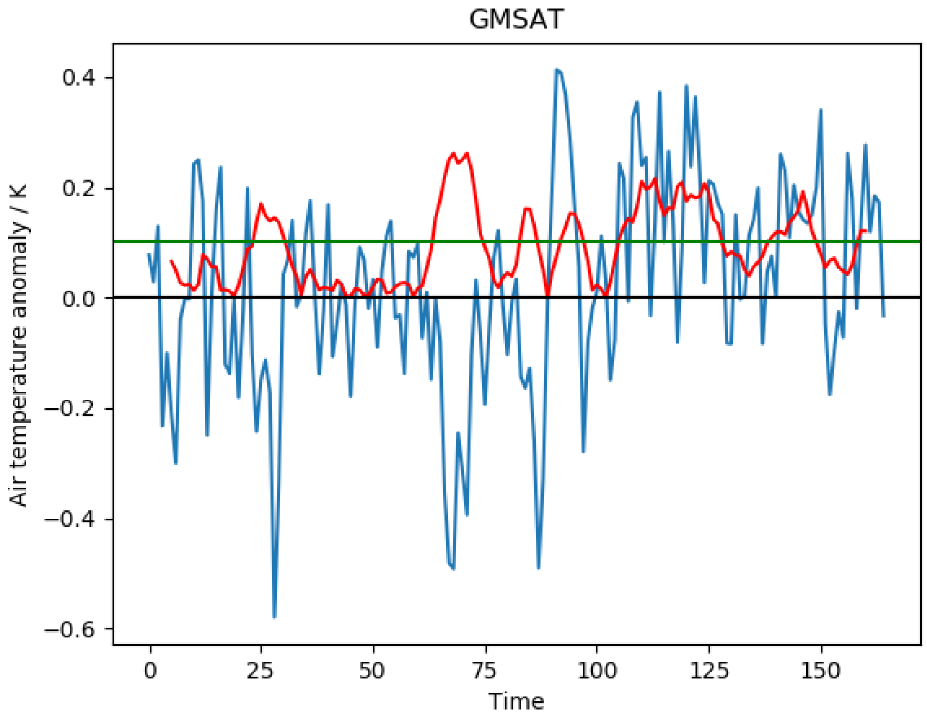

Figure 2. Graphical example of the calculation of the number of exceedances for a given pair of segments of the *piControl* simulations. This example is for GMSAT, but the method is the same for all indices. The blue line shows the difference between the two *piControl* segments that provides a sample of U_M-U_O. The red line is the absolute value of the 10-year running mean of the blue line. The green line represents the exceedance threshold, 0.1 K in this example. The number of exceedances is the number of red points above the green line.

We construct a second metric following the same steps, but using the variance-scaled samples $\sigma_M/\sigma\,(\overline{U_M}(t) - U_O(t)) - E_O(t)$, where σ_M is the model's standard deviation of the linearly-detrended *piControl* anomalies, and σ is the multi-model mean standard deviation of all the linearly-detrended *piControl* anomalies. This provides a variance-scaled set of samples of control distributions of exceedances that accounts for differences in the variance of the unforced variability across different models. We label this second metric as $E_s(T, y, N_m)$.

From these sets of samples of $E(T, y, N_m)$ and $E_s(T, y, N_m)$, we construct empirical quantile distribution functions $Q_Z(p; T, y, N_m)$, which give the number of exceedances for a given cumulative probability p. Z is a generic discrete random variable name that refers to either E or E_s. For simplicity, from now on we omit the dependency with the ensemble size N_m.

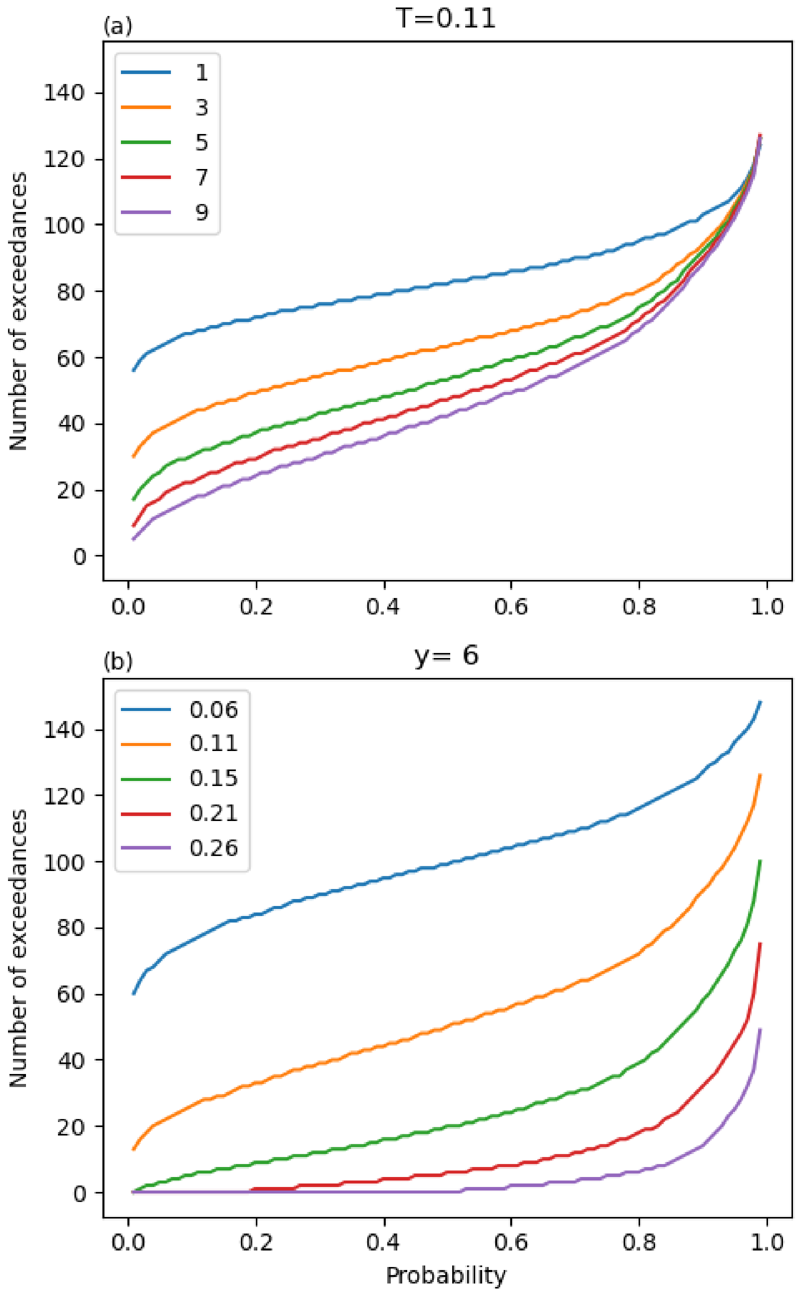

Figure 3. Examples of empirical quantile distribution functions $Q_Z(p; T, y, N_m)$ for an ensemble size N_m=3: (a) exceedance threshold set to 0.11K, length of the averaging window as shown in the legend (years); (b) length of averaging window of 6 years, exceedance threshold as shown in the legend (in K).

In summary, for each *historical* ensemble, we have calculated two (one with variance scaling and one without) empirical quantile functions in each point of the *(T, y)* grid. Figure 3 shows examples of Q_E for a *historical* ensemble of 3 members. For a given *T* and *y*, the probability is p that the number of exceedances (occurring during a 165-year historical integration) will be less than $Q_Z(p; T, y)$. There is zero chance that the number of exceedances will be less than zero, a small chance that it will be less than a small number, and we are certain that it will be less than a sufficiently large number (at most 165). Thus, Q_Z increases with p (Figures 3a and 3b). For any given p, the expected number of exceedances Q_Z is smaller for a longer meaning period y (Figure 3a) or a higher threshold T (Figure 3b).

d. Testing ensembles of historical simulations

We test each *historical* ensemble by comparing the number of exceedances of the difference between the ensemble mean and the observations against the expected number of differences given by the control distribution. First, we calculate $\overline{H_M}(t) - H_O(t)$, which we use as input to calculate the number of exceedances for each point in the (T, y) grid, $E_h(T, y)$, where the subscript h denotes that this is calculated from a *historical* ensemble, and $H_O(t)$ is the HadCRUT5 analysis ensemble mean. The linear drift of the *piControl* is subtracted from the *historical* time series. We then perform two one-tailed tests, each with a significance level α. This is done by comparing $E_h(T,y)$ against the empirical quantile function $Q_Z(p; T,y)$, separately for Z=E and Z=E_s. In each case, when either $E_h(T,y)>Q_Z(1-\alpha; T,y)$ or $E_h(T,y)<Q_Z(\alpha; T,y)$, the historical ensemble is flagged as incompatible in that point of the (T,y) grid. That is, we reject the null hypothesis that the difference between the historical simulation and observations is consistent with unforced variability if the number of times Z that the difference between them exceeds the threshold T in y-year means is either much larger than expected (upper-tail test), or much smaller than expected (lower-tail test).

Figure 4 shows an example for the upper tail test applied to the entire (T, y) grid, using a significance level α=0.05. For illustrative purposes, it is helpful to choose a model like EC-Earth3-Veg with large multidecadal unforced variability (Parsons et al., 2020). The filled contours in Figures 4a and 4c show $Q_Z(p=0.95; T, y)$ for $Z=E$ in and $Z=E_s$, respectively. The shape of Q_Z is very similar for all models and ensemble sizes. As shown also in Figure 3, Q_Z gets smaller as T gets larger for a given y (less likely to exceed a higher threshold), and smaller as y gets larger for a given T (less likely to for a longer time mean to exceed a threshold), although the dependency on y is much weaker.

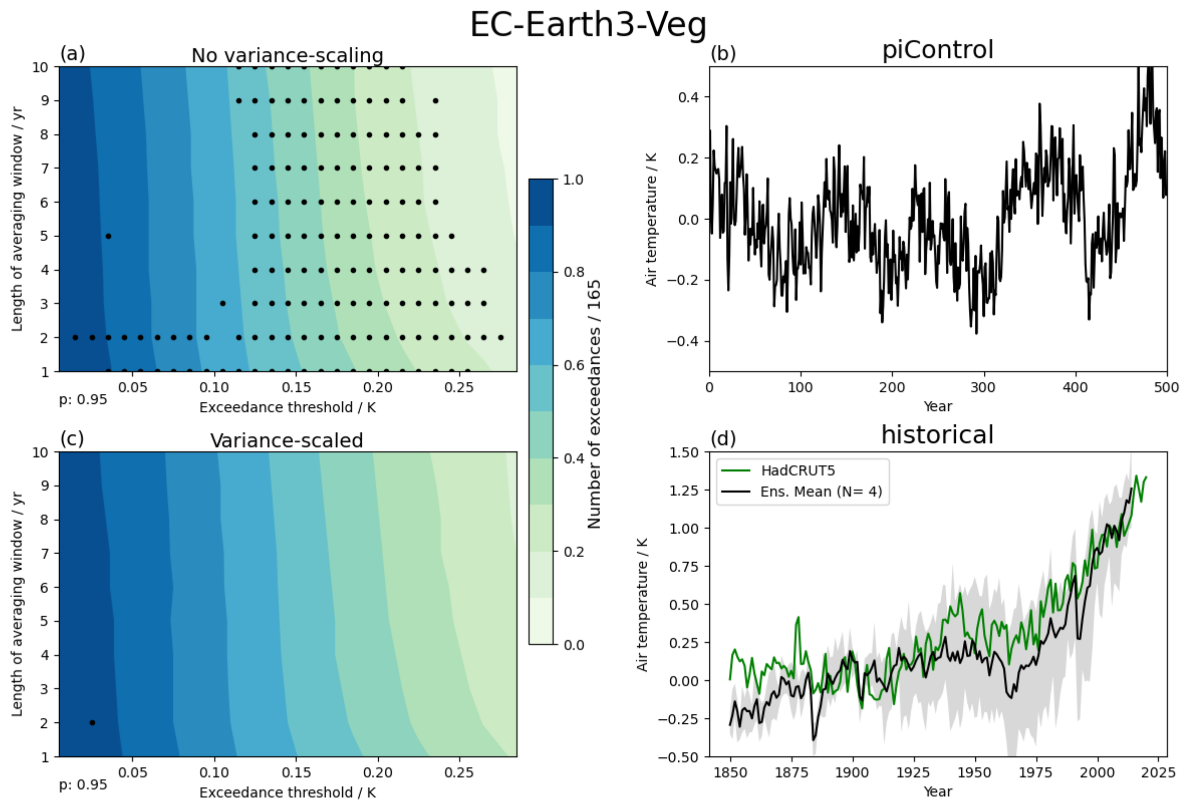

Figure 4. Tests of the *historical* ensemble of EC-Earth3-Veg. Tests without and with variance-scaling are shown in (a) and (b), respectively. The filled contours show $Q_Z(p=0.95;\ T, y)$ for (a) $Z=E$, and (c) $Z=E_s$. These surfaces show the expected number of exceedances normalized by 165 (maximum number of exceedances) for the 95th percentile (p=0.95, as noted in the bottom-left corner) of the *piControl* distributions in each point of the (T,y) grid. Observational uncertainty is included when available. The dots show the points in the (T, y) grid where the *historical* ensemble fails the test, i.e. $(E_h(T,y))>Q_Z(p=0.95;\ T, y)$. The last 500 years of the *piControl* simulation of the model tested are shown in (b). Panel (d) shows the annual-mean historical anomalies of the temperature index being tested: model's ensemble mean (black) and range (grey), and the observed anomalies (green). The historical anomalies in (d) are calculated with respect to the 1880-1919 time-average. The legend in (d) shows the number of *historical* realizations used in the calculation of the ensemble mean.

The dotted regions in the (T, y) grid mark where the null hypothesis is rejected ($E_h>Q_Z$). In this example, the test without variance scaling (Figure 4a) shows many rejections, whereas the variance-scaled test (Figure 4c) shows none. This contrast implies that the unforced variance of EC-Earth3-Veg is larger than the multi-model mean variance. The large variance increases the number of exceedances in the test without variance scaling, whereas variance scaling raises the control surface $Q_Z(p=0.95;\ T, y)$, making it easier for the model to pass the test. This scaling is trying to penalize those models that pass the non-scaled test due to a very small unforced variability compared to the multi-model mean variance, which we assume to be the best estimate of the unforced variability.

How much of the (T, y) space is needed to fail the statistical test for the model as a whole to be deemed "incompatible"? We have divided the (T, y) grid into 29×10 points, so we would expect a good model to fail at 290×10×0.05≈15 points in the (T, y) grid just by chance if there was no correlation in the number of exceedances between (T, y) neighbors. Because the time-scale and the threshold are correlated, if incompatibility occurs it is likely to cover patches of adjacent points in the (T, y) grid. Given that our main aim is to apply this method to intercompare models, we do not define a single, strict threshold for labelling a model as incompatible with the observations. Instead, we use the following guidance: models with less than 10 failures (dots) pass the test; models that fail between 10 and 20 times are considered marginal; model with more than 20 failures are labelled as incompatible.

The lower tail test ($E_h(T,y)<Q_Z(p=0.05; T,y)$) can be presented in a similar way, but only one of the models tested fails this test (FGOALS-g3, and only marginally). A model fails this test if its historical simulation deviates less than expected from reality, which can happen only if it has both a realistic forced response and unrealistically small unforced variability. It could be that the lower-tail test rarely fails because models in general do not have a realistic forced response. For the remainder of the paper we discuss the results of the upper-tail test only.

We have tested the sensitivity of the results to the order of the polynomial used for the detrending of the *piControl* time series. The results are largely insensitive to the use of quadratic instead of linear detrending, so we conclude that our method is robust with respect to the detrending method. If this test is applied to metrics that require non-linear detrending we would recommend the use of more flexible methods with better properties (e.g. splines).

4. Results and discussion

In this section we present results for three temperature indices: global mean, hemispheric difference, and $SST^{\#}$. These three metrics capture important complementary information about key aspects of temperature change over the historical record. The global mean has been widely used as the most fundamental metric of climate change. The hemispheric difference captures the influence of anthropogenic aerosols during the historical period, as emissions are dominated by sources in the Northern Hemisphere, and it is reasonably independent of the global mean (Braganza et al., 2003). The changes in tropical SST pattern control the sign and strength of low cloud feedbacks in response to CO_2 forcing (e.g. Miller, 1997; Gregory and Andrews, 2016), making it an important metric of the historical record.

a. Global mean

Figure 5 shows the tests without variance scaling. Out of the 40 models analyzed, 20 of them can be labelled as incompatible with the observed record, according to this test. These are models that show large, dotted areas. The other 20 models do not fail the test at all or only in a few instances. Models tend to fail the test for large exceedance thresholds *T*, with little dependence in the length of the averaging window *y*, i.e. they tend to fail along entire 'columns' in the contour plot.

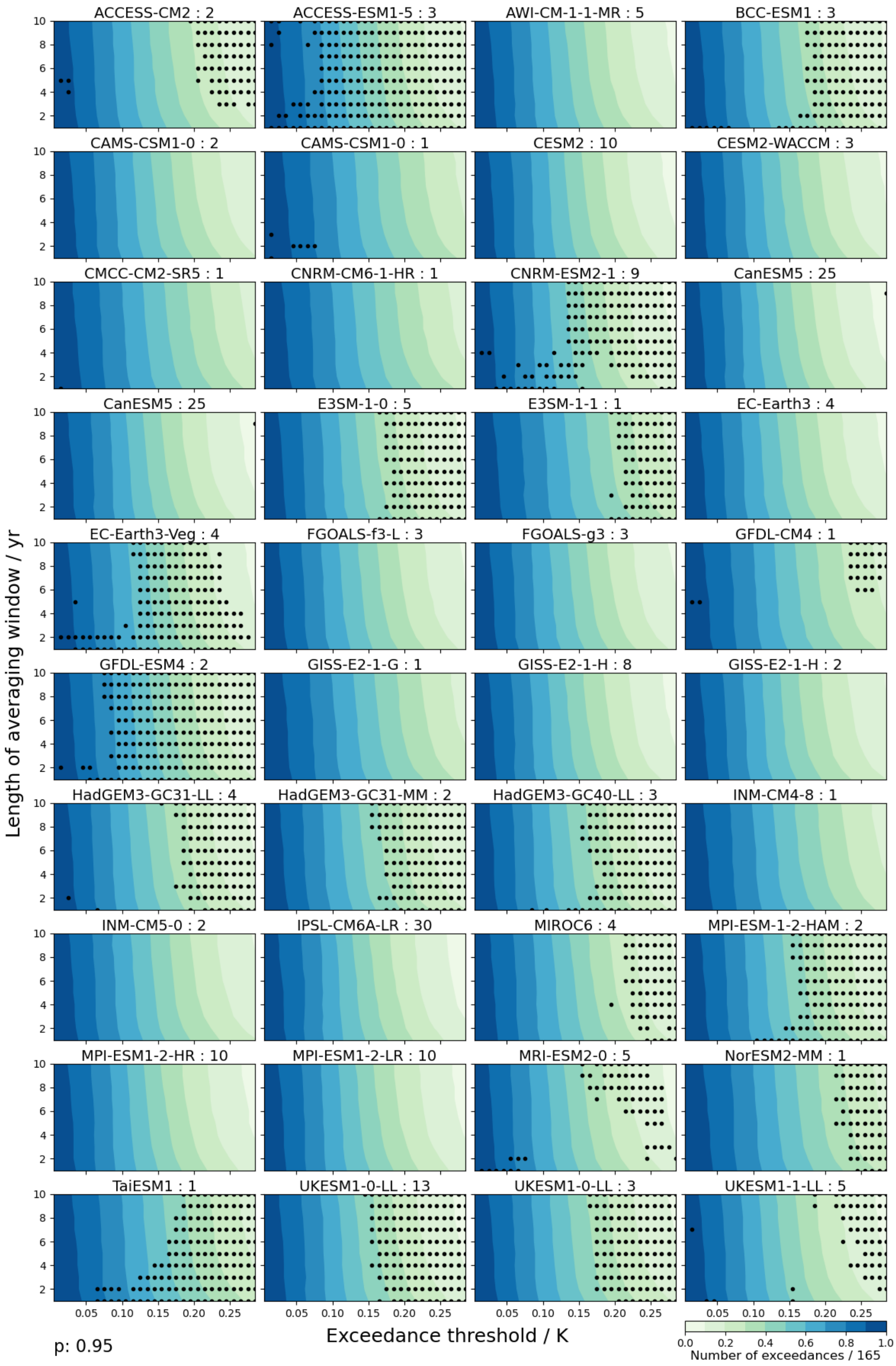

Figure 5. Multi-model summary of the test without variance scaling applied to the global-mean surface air temperature index. The number of exceedances is normalized by 165, the maximum number of exceedances given by the length of the historical record.

When the variance-scaled test is applied (Figure 6), 22 models are labelled as incompatible with the observed record, and 18 models pass the test. No models are in the marginal category. The variance-scaled test rejects 5 additional models, and labels as compatible 3 models that were rejected by the test without variance scaling. This is because these models have a *piControl* variance that is very different to the multi-model mean variance.

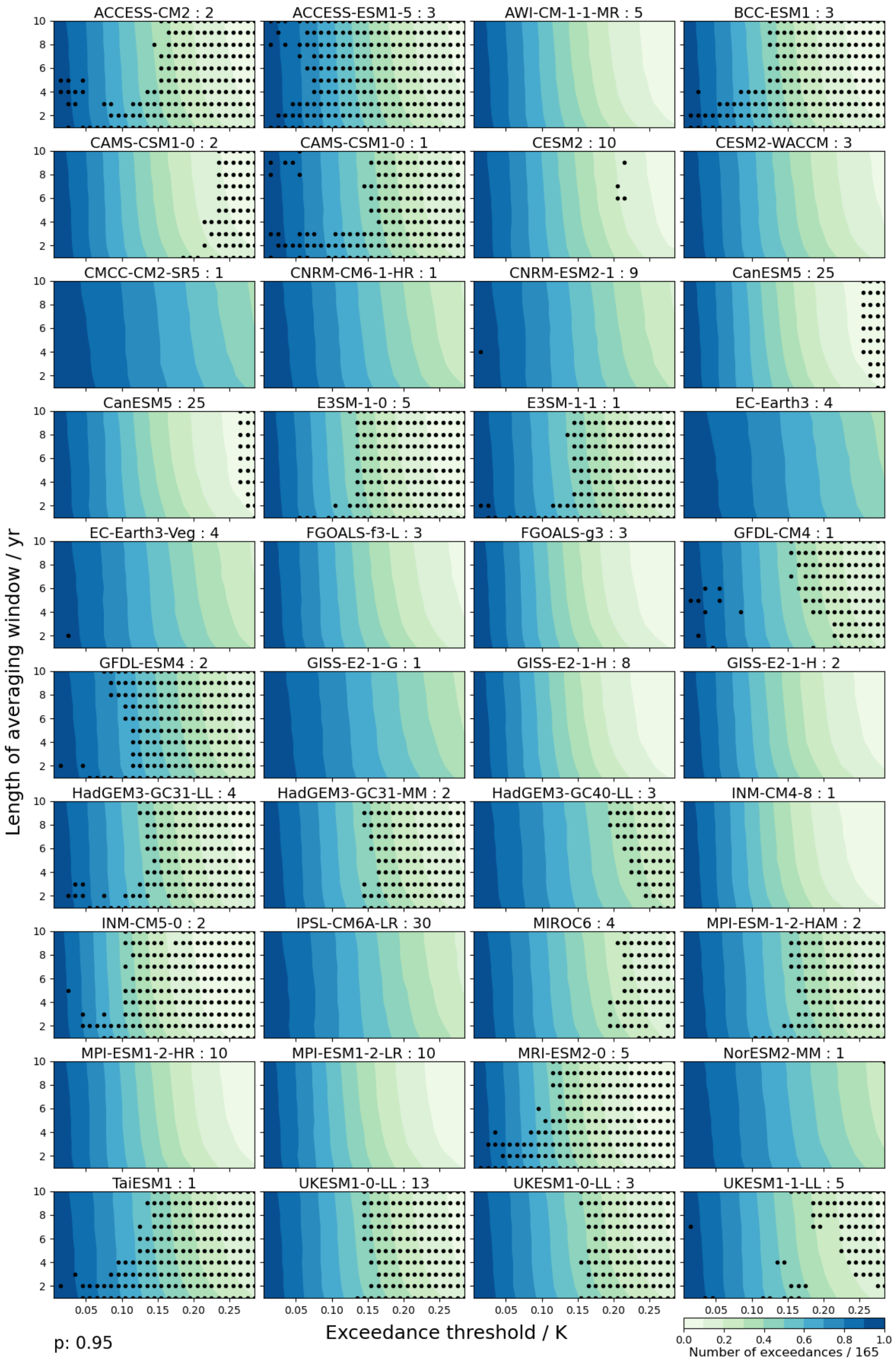

Figure 6. Multi-model summary of the test with variance scaling applied to the annual global-mean surface air temperature index. The number of exceedances is normalized by 165, the maximum number of exceedances given by the length of the historical record.

We have presented an example of a model with large unforced variability in Figure 4. Figure 7 shows an example for a model with a small unforced variability: MRI-ESM2-0. The control surface of the number of exceedances is lowered by the variance scaling, making it easier for the model to fail the test. Since we are not making any assumption about the quality of *piControl* simulations of individual models, the variance scaling method is an attempt to enable a fair comparison, when using other models with different unforced variability.

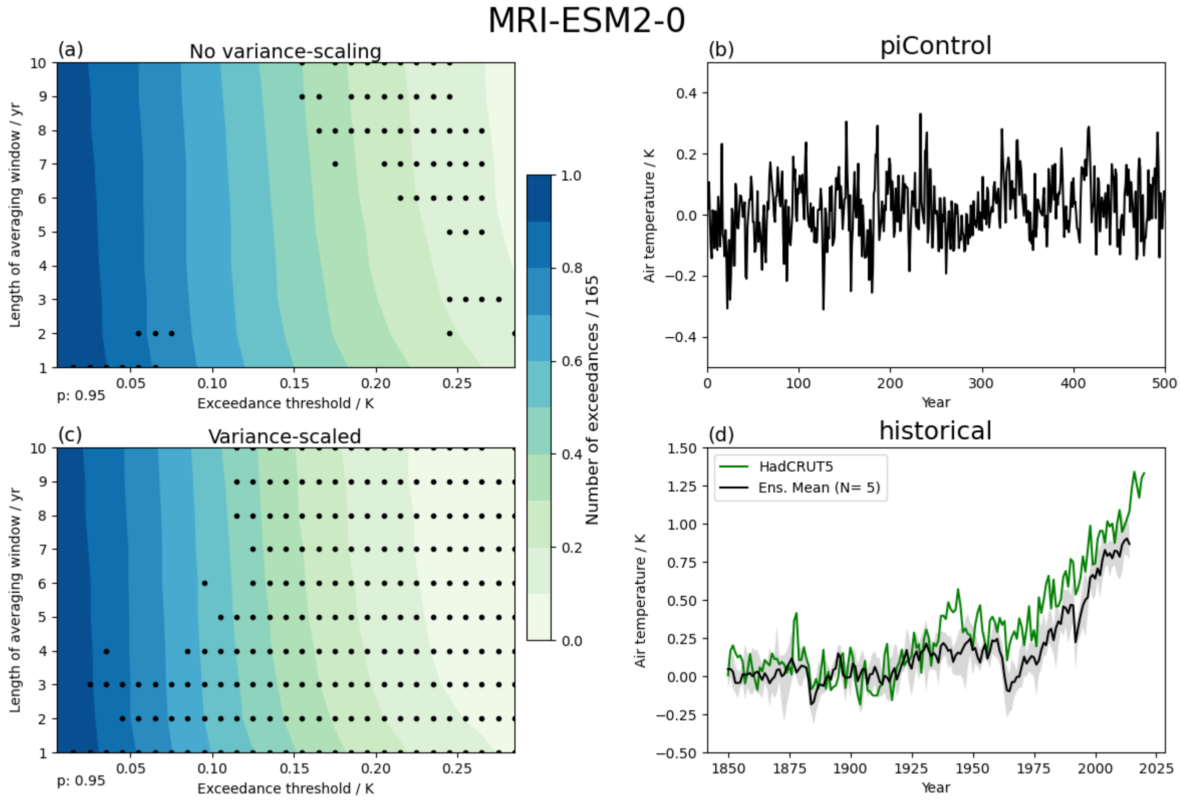

Figure 7. Same as Figure 4, but for model MRI-ESM2-0.

These two examples show how each model's characteristics of its unforced variability are incorporated into the test. This is particularly helpful when the ensemble size of historical simulations is small, which makes difficult the assessment of the impact of the unforced variability by visual inspection. It must be emphasized that we treat all *piControl* simulations as equally plausible, but the method could be refined by bringing in external information to better characterize the unforced variability of the real system. We expand on this below when we discuss the caveats of the methodology.

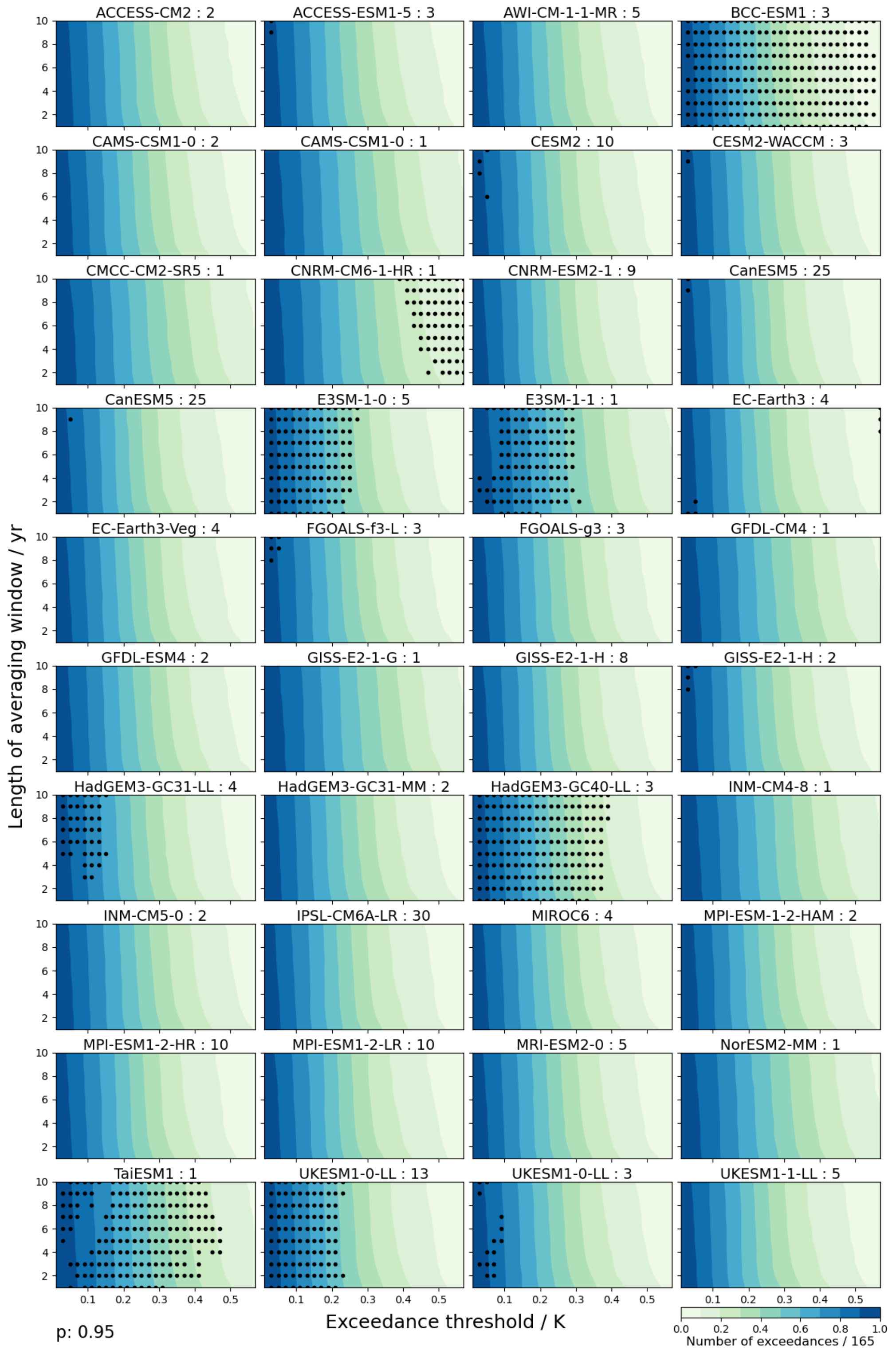

Figure 8. Same as Figure 5, but for the hemispheric gradient surface air temperature index.

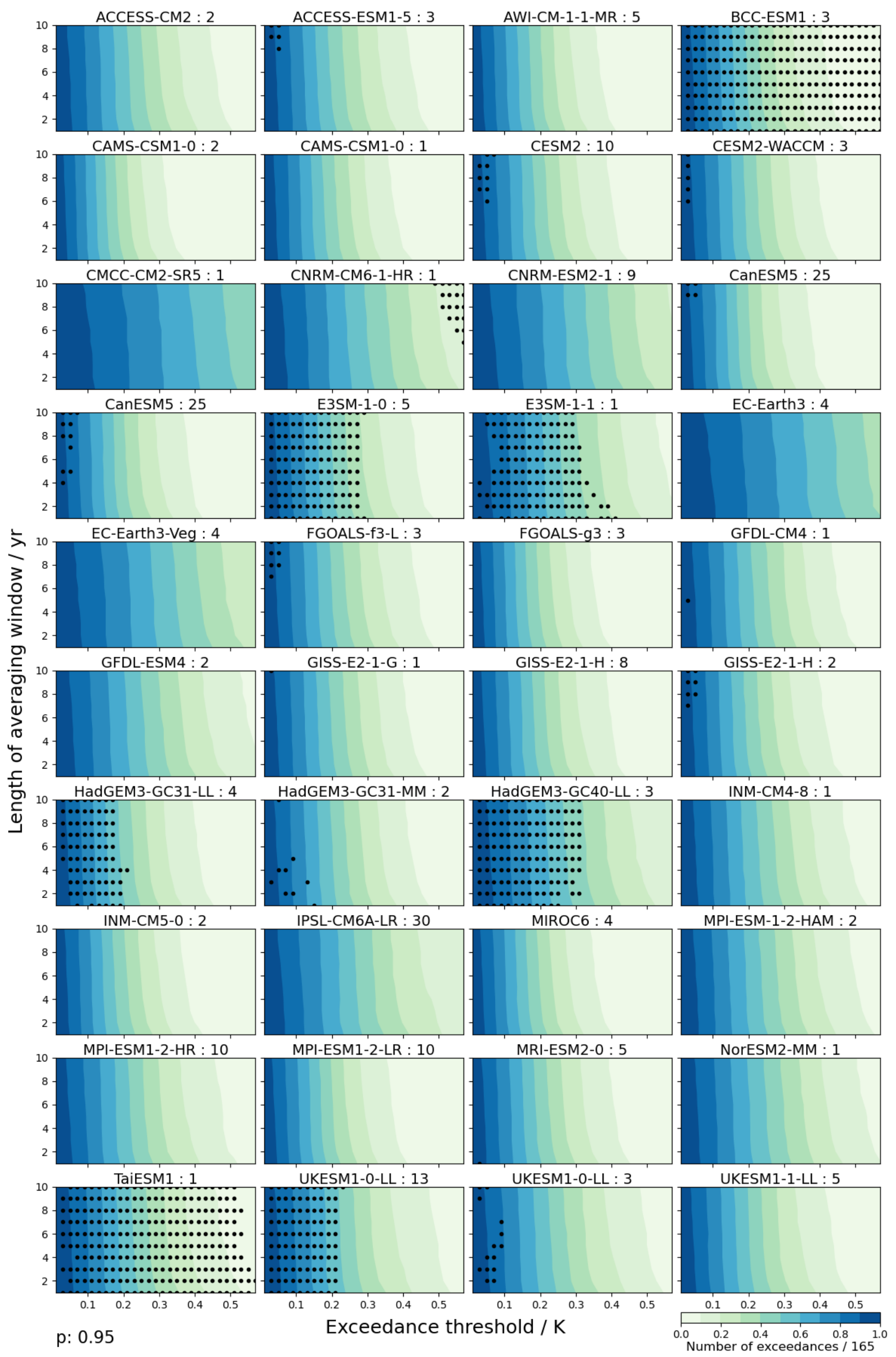

Figure 9. Same as Figure 6, but for the hemispheric gradient surface air temperature index.

b. Hemispheric gradient

Figure 8 shows the tests without variance scaling for the hemispheric gradient index. Out of the 40 models analyzed, 8 are labelled as incompatible with the observed record, 1 is marginal, and 31 pass the test. If variance scaling is used (Figure 9), the results are very similar, with 7 models rejected, 3 marginal, and 29 passing the test. As with the global-mean, failures tend to happen along ‘columns’, i.e. for all averaging window lengths. It is interesting to note that, contrary to the global-mean, the hemispheric gradient shows more failures for small exceedance thresholds.

In CESM2 there is a strong sensitivity of the hemispheric gradient to the variability in biomass emissions from 40ºN to 70ºN, which leads to spurious warming in the late historical period (Fasullo et al., 2022). However, this model passes the global and hemispheric tests, which may suggest the presence of compensating biases. This highlights the importance of having a large battery of diagnostics capable of assessing model performance from different angles.

c. $SST^{\#}$

Figures 10 and 11 show the multi-model ensemble results for $SST^{\#}$, without and with scaling of the unforced variance, respectively. The test without variance scaling rejects all CMIP6 models. Only one model is not rejected, namely GISS-E2-1-G, when variance scaling is used. The GISS models are examples of models with large unforced variability (Figure 12). Unlike in previous examples with large unforced variability on long time scales (Figure 4), the unforced variability of the GISS models is dominated by high-frequency (annual) variability. Given that the observational record does not show such a large high-frequency variability, we conclude that the test without variance scaling is probably a better assessment of the performance of the GISS models. This conclusion is also supported by Orbe et al. (2020) who show that GISS-E2-1-G is an outlier in the simulation of ENSO.

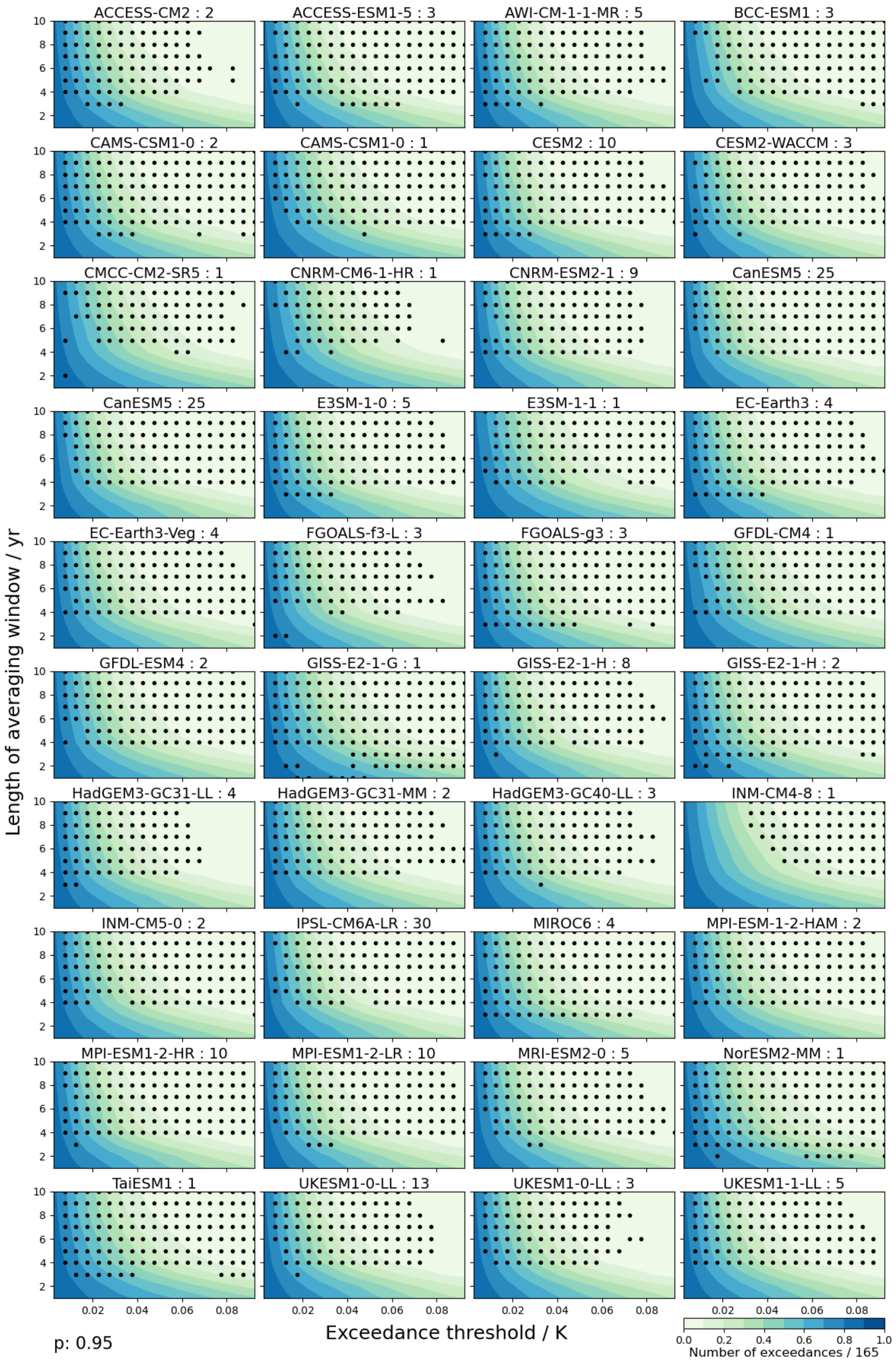

Figure 10. Same as Figure 5, but for the SST# index. The observational SST# index is calculated using the PCMDI/AMIPII dataset.

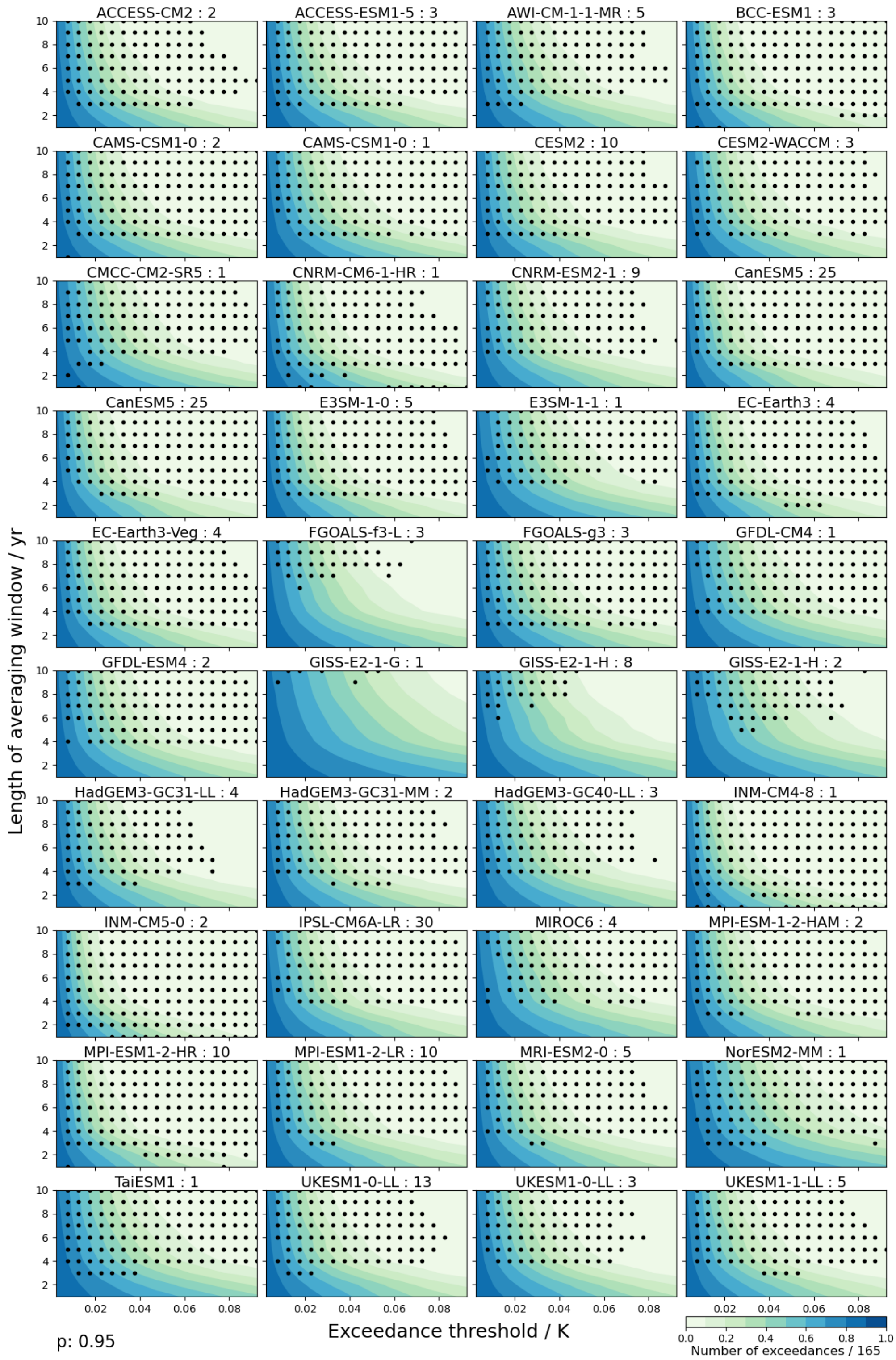

Figure 11. Same as Figure 6, but for the SST# index. The observational SST# index is calculated using the PCMDI/AMIPII dataset.

SST# is subject to a large observational uncertainty (Fueglistaler and Silvers, 2021). The observations show very good agreement during the satellite era (1979 onwards), where the spatial coverage is very dense, but they show large discrepancies before satellite data were available. The differences are attributed to the different methodologies used to provide information where observations are not available. Given that the PCMDI/AMIPII dataset doesn't provide a comprehensive error characterization, we have repeated the tests using the ERSST5 dataset to test the robustness of our conclusions. We have chosen the PCMDI/AMIPII and ERSST5 datasets because they fall at opposite ends of the spectrum of $SST^{\#}$ anomalies provided by observational datasets, giving us information about structural uncertainties in the observational reconstructions of $SST^{\#}$. The results with ERSST5 (not shown) are similar to the comparisons against PCMDI/AMIPII, all the CMIP6 models are rejected by both tests, with and without variance scaling. This confirms that the results are robust with respect to observational uncertainty in $SST^{\#}$.

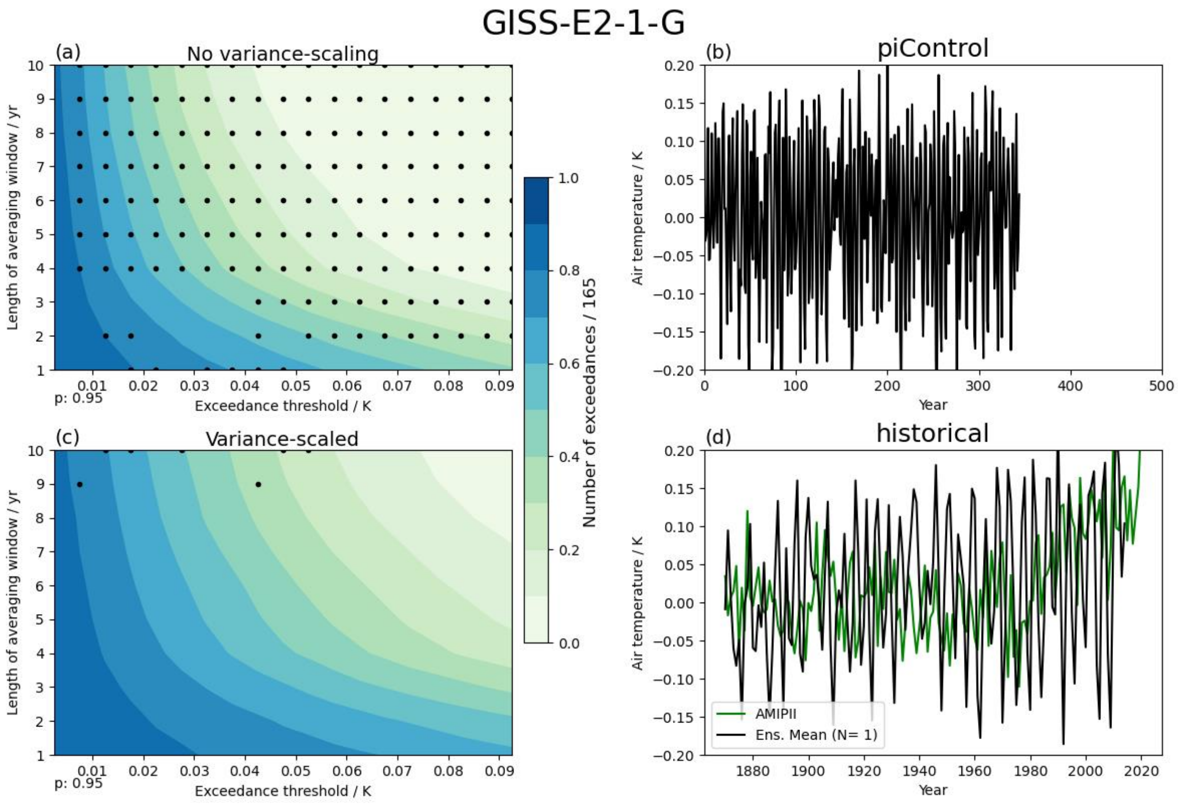

Figure 12.Same as Figure 4, but for the SST# index of model GISS-E2-1-G. The green line in (d) shows the PCMDI/AMIPII observational estimate..

The fact that the entire CMIP6 ensemble performs poorly in the $SST^{\#}$ index is consistent with previous studies showing that models in general do not reproduce the Pacific SST trends of recent decades (Seager et al., 2019; Gregory et al., 2020; Wills et al., 2022), and it has potential implications beyond the models' performance over the historical period.

Unlike for the two other indices, there is no consensus either that SST$^{\#}$ should contain a forced signal or that it is part of the unforced variability of the climate system. Some recent studies suggest that tropical Pacific SST patterns observed during the recent decades could arise from internal climate variability (e.g. Olonscheck et al, 2020; Watanabe et al. 2021). Other studies suggest that the SST patterns are consistent with a forced response to greenhouse forcing (Seager et al., 2019) that can be explained with simple models (Clement et al., 1996), or with a potential role for volcanic or anthropogenic aerosols in setting the recent patterns (Gregory et al., 2020; Heede and Fedorov, 2021; Dittus et al., 2021). If the observed evolution of SST$^{\#}$ is not forced, no model ensemble-mean can be expected to agree with the observations. In that case, if a model fails the test, it means that its simulation of SST$^{\#}$ variability has the wrong magnitude. On the other hand, if SST$^{\#}$ is forced, the rejection of the test means that the model doesn't replicate the forced response. In this case, if a large number of models fail the test it could imply a common bias in the forced response. In either case, a rejection of the test indicates some aspects of the model performanceare wrong somehow. Additional process-level analysis and physical hypothesis-testing is required to improve our understanding of the causes behind the model errors.

d. Caveats and interpretation of the tests

The results above show how the methodology presented here can be used to assess historical simulations during the model development process. We have applied it to surface temperature indices, but it can be applied to any variable for which observational estimates over the historical period exist. However, the methodology presents some interpretation challenges and caveats. How do we interpret a rejection of the null hypothesis that the model's forced response is realistic? Can we definitively conclude that there is a problem with the model's forced signal? There is a chance that the null hypothesis is wrongly rejected although true; that is a Type I error, whose probability is the chosen significance level. If we reject the null hypothesis, we must have an alternative hypothesis. Potential alternatives are: there is a problem with the model's forced signal; our model-based unforced variability is biased; the forcing is wrong. We do not have a statistical means to estimate the probability of these systematic errors.

It is also worth mentioning that agreement between the observations and simulations might be due to compensating errors. Potential problems that could contribute to compensating errors concern the following: aerosol radiative forcing and aerosol-cloud interactions (e.g. Paulot et al., 2018; Rieger et al., 2020; Wang et al., 2021; Fasullo et al., 2022); tropical SST patterns and their role on global radiative feedbacks (Ceppi and Gregory, 2017; Andrews and Webb, 2018). The unforced distributions used to define the exceedance quantile functions are constructed from *piControl* simulations. This assumes that the multi-model ensemble provides us with a good representation of the unforced variability, which is not necessarily true. As we have shown above when discussing the results of the variance-scaled results, there exist large discrepancies in the representation of unforced variability between models (Parsons et al., 2020), which raises questions about the ability of at least some models to provide a good estimate of unforced variability. If the unforced variability estimated from the multi-model ensemble is biased, then our method will be biased. One avenue that could be explored for improving this would be to incorporate information from proxy temperature reconstructions into a correction of the unforced variability. However, the use of proxy reconstructions is not free from problems. The reconstructions are for restricted regions where there are proxies (e.g. PAGES 2k Consortium, 2013), and much of their variability is forced by volcanoes and solar variability (PAGES 2k Consortium, 2019). In any case, a failure of this type would imply that the models *piControl* simulations are wrong (rather than the forced signal necessarily), so the test would still be highlighting a problem.

Our test with scaled variance is an initial attempt to identify outliers, but more sophisticated methods could be used. Perhaps a better estimate of the unforced variability could be achieved by restricting the set of models used to form the distributions of internal variability. This selection could be based on how models represent observational estimates of the spectra of some modes of variability (Fasullo et al., 2020). For $SST^{\#}$, basing this selection on some metric of ENSO could be particularly useful. Screening out models would reduce the number of *piControl* simulations, so this would have an impact the robustness of the unforced distributions.

A second caveat is the differing sizes of the historical ensembles. Out of the 40 models analyzed here, only 4 have *historical* ensembles with more than 10 members, and 31 models have 5 or fewer historical simulations. Large ensembles will provide more robust tests. A model with a small ensemble will provide a less precise estimate of the ensemble mean, making the result of the test more likely to be different from the result that would be obtained with a large ensemble. This is a general problem with statistical hypothesis testing, and it should be incorporated into the subjective interpretation of the tests. We propose some guidance based on the dependence of the variance of the control distribution with the size of the ensemble. As explained above, the control distribution is constructed from samples of $\overline{U_M}(t; N_m) - U_O(t) - E_O(t)$. The observational error is typically small compared to the unforced variability, so we can approximate dependence of the variance as $(1 + 1/N_m)*\sigma$, where σ is the variance of $U_M(t)$ and $U_o(t)$. As N_m becomes larger, the total variance decreases from 2 (in units of σ) to its asymptotic value of 1, with the rate of change being larger for small N_m. For instance, an ensemble of 10 members will reduce the variance to within 10% of its asymptotic value, which will significantly increase the robustness of the test.

We do not account for the uncertainty in radiative forcing, which could lead to overtuning if the only objective is to match the warming over the historical period (e.g. Hourdin et al., 2017). However, we are not advocating making development choices only based on the approach presented here. A wide range of other metrics, including process-based metrics need to be considered. The use of a much wider basket of metrics should reduce the risk of overtuning.

A final caveat is that the variance scaling can't account for differences in models' *piControl* variability on different timescales, so while the overall variability of two models can be scaled to be similar the interannual/multidecadal variability could be still very different. We have subjectively accounted for this in the discussion of the $SST^{\#}$ results for GISS-E2-1-G, whose variability is dominated by large interannual variability, which can be confidently assessed with observations of the historical period. However, this is not the case for variability at much longer timescales, for which the observational record provides much limited information. A possible approach to look at in the future is to account for this by applying different variance scaling factors for each p.

5. Conclusions

The historical record of surface temperature is an important metric that climate models should be able to reproduce. However, it is not consistently used by modelling centres during model development for two main and quite distinct reasons: first, coupled simulations are expensive to run, especially because the historical simulation must be preceded by a spin-up simulation long enough to eliminate drift; second, the observed historical record of surface temperature is reserved as an out-of-sample validation. It is generally argued that the warming during the historical record and emergent properties like equilibrium climate sensitivity should be used as an a posteriori evaluation and not as a target for model development, although there is not complete consensus among the modelling community on this topic (Hourdin et al., 2017). Bock et al. (2020) highlight the risk of tuning models to reproduce a set of metrics ignoring deficiencies elsewhere. However, this risk is not specific to metrics based on historical warming. Within the context of emergent constraints, Eyring et al. (2019) advocate the use of variability metrics or trends during model development.

We develop a statistical method to test whether simulations of large-scale surface temperature change are consistent with the observed warming of the historical period (1850-2014). The method uses information on a range of time scales. It incorporates information about unforced variability, and it is designed to test an entire ensemble of simulations of any size. The method is applied to annual-mean time series of three surface temperature indices: global-mean, hemispheric gradient, and a recently-developed index that captures the sea-surface temperature (SST) pattern in the tropics ($SST^{\#}$; Fueglistaler and Silvers, 2021). We test the historical simulations of the CMIP6 ensemble and post-CMIP6 versions of the HadGEM3 and UKESM models.

Around half the models fail the test for the global-mean time series, approximately a fifth of the models fail when the hemispheric temperature gradient is analyzed, and all models fail the $SST^{\#}$ test. We note the importance of the characteristics of the models' unforced variability (Parsons et al., 2020). Assessment of the quality of the historical simulations by visual comparison of the time series of a few ensemble members against the observations can be misleading, being reliable only for models with a large number of historical realisations. The method presented here complements other statistical approaches that have previously compared historical model simulations to observations (e.g. Sanderson et al., 2015; Brunner et al., 2020). Given that most modeling centres only run a small number of historical simulations, a method like the one presented here that accounts for the unforced variability is desirable, especially if the aim is to use it during the model development process, where large ensembles are not affordable.

We show that the method presented here can be used as a tool to assess historical simulations during the development process. The method is easy to apply and summarises a large amount of information in two plots, with and without variance-scaling. It accounts for the unforced variability of the model tested, and it can be applied to an ensemble of historical simulations of arbitrary size. We also plan to make this methodology available to the community by implementing it in ESMValTool (Eyring et al., 2020).

There are several avenues that could be explored to develop this method further. One potential improvement could be to incorporate information from proxy reconstructions to improve the estimate of the unforced variability, currently based on control model simulations. However, this may prove difficult given that many proxies do not resolve annual variability, and because of the non-stationarity of the magnitude of internal variability. Perhaps a better estimate of the unforced variability could be achieved by restricting the model set used to form the distributions of internal variability based on how models represent observational estimates of annual to decadal modes of variability (Fasullo et al., 2020).

A second area for further developments could be to apply a scaling factor, as it is done in optimal fingerprinting (e.g. Allen and Tett, 1999). Some of the models that are rejected by our current methodology could pass the test if they are appropriately scaled. The interpretation of the test results with the scaled time series is not straight forward, but it may be useful to know that a model that is rejected could be made realistic by a scaling factor.

Acknowledgments.

This work was supported by the Met Office Hadley Centre Climate Programme funded by BEIS. We thank an anonymous reviewer, J. T. Fasullo and R. C. J. Wills for their constructive comments that helped improve the original manuscript. We thank Gareth Jones for his contribution to the methodology and useful comments on an early draft version of the manuscript. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF.

Data Availability Statement.

HadCRUT.5.0.1.0 data were obtained from http://www.metoffice.gov.uk/hadobs/hadcrut5 on 15/02/2021 and are © British Crown Copyright, Met Office 2021, provided under an Open Government License, http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/. PCMDI AMIP SSTs were obtained from the ESGF archive, variable tosbcs from input4MIPs, version v20220201. NOAA_ERSST_V5 data provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html (Huang et al., 2017).

REFERENCES

Allen, M., and Tett, S., Checking for model consistency in optimal fingerprinting. Climate Dynamics 15, 419–434 (1999). https://doi.org/10.1007/s003820050291.

Andrews, T., & Webb, M. J. (2018). The Dependence of Global Cloud and Lapse Rate Feedbacks on the Spatial Structure of Tropical Pacific Warming, Journal of Climate, 31(2), 641-654.

Annamalai, H., Hafner, J., Sooraj, K. P., and Pillai, P. (2013). Global Warming Shifts the Monsoon Circulation, Drying South Asia, Journal of Climate, 26(9), 2701-2718.

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., et al. (2020). Quantifying progress across different CMIP phases with the ESMValTool. Journal of Geophysical Research: Atmospheres, 125, e2019JD032321. https://doi.org/10.1029/2019JD032321

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., & Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. Journal of Advances in Modeling Earth Systems, 12, e2019MS002010. https://doi.org/10.1029/2019MS002010.

Braganza, K., Karoly, D., Hirst, A. et al. Simple indices of global climate variability and change: Part I – variability and correlation structure. Climate Dynamics 20, 491–502 (2003). https://doi.org/10.1007/s00382-002-0286-0.

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth Syst. Dynam., 11, 995–1012, https://doi.org/10.5194/esd-11-995-2020, 2020.

Ceppi, P., and Gregory, J. M., 2017: Relationship of tropospheric stability to climate sensitivity and Earth's observed radiation budget, Proc. Nat. Acad. Sci., 114(50), 13126-13131, 10.1073/pnas.1714308114.

Dittus, A. J., Hawkins, E., Wilcox, L. J., Sutton, R. T., Smith, C. J., Andrews, M. B., & Forster, P. M. (2020). Sensitivity of historical climate simulations to uncertain aerosol forcing. Geophysical Research Letters, 47, e2019GL085806. https://doi.org/10.1029/2019GL085806.

Dittus, A. J., Hawkins, E., Robson, J. I., Smith, D. M., & Wilcox, L. J. (2021). Drivers of recent North Pacific Decadal Variability: The role of aerosol forcing. Earth's Future, 9, e2021EF002249. https://doi.org/10.1029/2021EF002249.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Eyring, V., Cox, P.M., Flato, G.M. et al. Taking climate model evaluation to the next level. Nature Clim Change 9, 102–110 (2019). https://doi.org/10.1038/s41558-018-0355-y.

Eyring, V., et al., Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geosci. Model Dev., 13, 3383–3438, https://doi.org/10.5194/gmd-13-3383-2020, 2020.

Fasullo, J. T., Phillips, A. S., & Deser, C. (2020). Evaluation of Leading Modes of Climate Variability in the CMIP Archives, *Journal of Climate*, *33*(13), 5527-5545.

Fasullo, J. T., Lamarque, J.-F., Hannay, C., Rosenbloom, N., Tilmes, S., DeRepentigny, P., et al. (2022). Spurious late historical-era warming in CESM2 driven by prescribed biomass burning emissions. Geophysical Research Letters, 49, e2021GL097420. https://doi.org/10.1029/2021GL097420

Flynn, C. M., and Mauritsen, T.: On the climate sensitivity and historical warming evolution in recent coupled model ensembles, Atmos. Chem. Phys., 20, 7829–7842, https://doi.org/10.5194/acp-20-7829-2020, 2020.

Fueglistaler, S., and Silvers, L.G., 2021: The peculiar trajectory of global warming. Journal of Geophysical Research: Atmospheres, 126, e2020JD033629. https://doi.org/10.1029/2020JD033629.

Gillett, N. P., Zwiers, F. W., Weaver, A. J., Hegerl, G. C., Allen, M. R., and Stott, P. A., Detecting anthropogenic influence with a multi-model ensemble, Geophys. Res. Lett., 29(20), 1970, doi:10.1029/2002GL015836, 2002.

Gregory, J. M., Andrews, T., Ceppi, P., Mauritsen, T., and M. J. Webb, 2020: How accurately can the climate sensitivity to CO2 be estimated from historical climate change?. Clim Dyn 54, 129–157. https://doi.org/10.1007/s00382-019-04991-y.

Gregory, J. M. & Andrews, T., 2016: Variation in climate sensitivity and feedback parameters during the historical period. Geophys. Res. Lett. 43, 3911–3920.

Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., et al. (2019). The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. Journal of Advances in Modeling Earth Systems, 11, 2089– 2129. https://doi.org/10.1029/2018MS001603.

Gulev, S.K., P.W. Thorne, J. Ahn, F.J. Dentener, C.M. Domingues, S. Gerland, D. Gong, D.S. Kaufman, H.C. Nnamchi, J. Quaas, J.A. Rivera, S. Sathyendranath, S.L. Smith, B. Trewin, K. von Schuckmann, and R.S. Vose, 2021: Changing State of the Climate System. In Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [MassonDelmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press. In Press.

Heede, U.K., Fedorov, A.V. Eastern equatorial Pacific warming delayed by aerosols and thermostat response to CO2 increase. Nat. Clim. Chang. 11, 696–703 (2021). https://doi.org/10.1038/s41558-021-01101-x.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson, D. (2017). The Art and Science of Climate Model Tuning, Bulletin of the American Meteorological Society, 98(3), 589-602. https://journals.ametsoc.org/view/journals/bams/98/3/bams-d-15-00135.1.xml

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H. (2017). Extended Reconstructed Sea

Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons, Journal of Climate, 30(20), 8179-8205. https://journals.ametsoc.org/view/journals/clim/30/20/jcli-d-16-0836.1.xml.

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., and Zhang, H. (2017): NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5 (2021-08-07). NOAA National Centers for Environmental Information. doi:10.7289/V5T72FNM. Accessed 2021-09-01.

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A new sea surface temperature and sea ice boundary dataset for the community atmosphere model. Journal of Climate, 21(19), 5145–5153. https://doi.org/10.1175/2008JCLI2292.1

Jones, P. The reliability of global and hemispheric surface temperature records. Adv. Atmos. Sci. 33, 269–282 (2016). https://doi.org/10.1007/s00376-015-5194-4

Jones, G. S. (2020). "Apples and Oranges": On comparing simulated historic near-surface temperature changes with observations, Quarterly Journal of the Royal Meteorological Society, 146, 733, 3747-3771. https://doi.org/10.1002/qj.3871.

Jones, G. S., Stott, P. A., and Christidis, N. (2013), Attribution of observed historical near–surface temperature variations to anthropogenic and natural causes using CMIP5 simulations, J. Geophys. Res. Atmos., 118, 4001– 4024, doi:10.1002/jgrd.50239.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and its response to increasing CO2. Journal of Advances in Modeling Earth Systems, 11, 998–1038. https://doi.org/10.1029/2018MS001400.

McKinnon, K. A., and Deser, C. (2018). Internal Variability and Regional Climate Trends in an Observational Large Ensemble, *Journal of Climate*, *31*(17), 6783-6802.

Miller, R. L., 1997: Tropical Thermostats and Low Cloud Cover, Journal of Climate, 10(3), 409-440.

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al. (2021). An updated assessment of near-surface temperature change from 1850: the HadCRUT5 data set. Journal of Geophysical Research: Atmospheres, 126, e2019JD032361. https://doi.org/10.1029/2019JD032361.

Mulcahy, J.P., C. Jones, S. Rumbold, T. Kuhlbrodt, A. J. Dittus, E. W. Blockley, A. Yool, J. Walton, C. Hardacre, T. Andrews, A. Bodas-Salcedo, M. Stringer, L. de Mora, P. Harris, R. Hill, D. Kelley, E. Robertson, and Y. Tang. UKESM1.1: Development and evaluation of an updated configuration of the UK Earth System Model. Submitted to Geoscientific Model Development.

Olonscheck, D., and Notz, D. (2017). Consistently Estimating Internal Climate Variability from Climate Model Simulations, *Journal of Climate*, *30*(23), 9555-9573.

Olonscheck, D., Rugenstein, M., & Marotzke, J. (2020). Broad consistency between observed and simulated trends in sea surface temperature patterns. Geophysical Research Letters, 47, e2019GL086773. https://doi.org/10.1029/2019GL086773.

Orbe, C., Van Roekel, L., Adames, Á. F., Dezfuli, A., Fasullo, J., Gleckler, P. J., Lee, J., Li, W., Nazarenko, L., Schmidt, G. A., Sperber, K. R., & Zhao, M. (2020). Representation of Modes of Variability in Six U.S. Climate Models, *Journal of Climate*, *33*(17), 7591-7617.Paulot, F., Paynter, D., Ginoux, P., Naik, V., and Horowitz, L. W.: Changes in the aerosol direct radiative forcing from 2001 to 2015: observational constraints and regional mechanisms, Atmos. Chem. Phys., 18, 13265–13281, https://doi.org/10.5194/acp-18-13265-2018, 2018.

Parsons, L. A., Brennan, M. K., Wills, R. C. J., and Proistosescu, C. (2020). Magnitudes and spatial patterns of interdecadal temperature variability in CMIP6. Geophysical Research Letters, 47, e2019GL086588. https://doi.org/10.1029/2019GL086588.

PAGES 2k Consortium. Continental-scale temperature variability during the past two millennia. Nature Geosci 6, 339–346 (2013). https://doi.org/10.1038/ngeo1797.

PAGES 2k Consortium. Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era. Nat. Geosci. 12, 643–649 (2019). https://doi.org/10.1038/s41561-019-0400-0.

Reichler, T., and Kim, J. (2008). How Well Do Coupled Models Simulate Today's Climate?, Bulletin of the American Meteorological Society, 89(3), 303-312.

Richardson, M., Cowtan, K., Hawkins, E. et al. Reconciled climate response estimates from climate models and the energy budget of Earth. Nature Clim Change 6, 931–935 (2016). https://doi.org/10.1038/nclimate3066.

Rieger, L. A., Cole, J. N. S., Fyfe, J. C., Po-Chedley, S., Cameron-Smith, P. J., Durack, P. J., Gillett, N. P., and Tang, Q.: Quantifying CanESM5 and EAMv1 sensitivities to Mt. Pinatubo volcanic forcing for the CMIP6 historical experiment, Geosci. Model Dev., 13, 4831–4843, https://doi.org/10.5194/gmd-13-4831-2020, 2020.

Sanderson, B. M., Knutti, R., and Caldwell, P. (2015). A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble, Journal of Climate, 28(13), 5171-5194. https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00362.1.xml

Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., et al. (2021). Energy budget constraints on the time history of aerosol forcing and climate sensitivity. Journal of Geophysical Research: Atmospheres, 126, e2020JD033622. https://doi.org/10.1029/2020JD033622.

Taylor, K.E., D. Williamson and F. Zwiers, 2000: The sea surface temperature and sea ice concentration boundary conditions for AMIP II simulations. PCMDI Report 60, Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, 25 pp. Available online: https://pcmdi.llnl.gov/report/pdf/60.pdf

Gupta, A. S., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate Drift in the CMIP5 Models, Journal of Climate, 26(21), 8597-8615. https://journals.ametsoc.org/view/journals/clim/26/21/jcli-d-12-00521.1.xml.

Taylor, K. E., D. Williamson, and F. Zwiers, 2000: The sea surface temperature and sea-ice concentration boundary conditions of AMIP II simulations. PCMDI Rep. 60, 20 pp.

Wang, C., Soden, B. J., Yang, W., and Vecchi, G. A. (2021). Compensation between cloud feedback and aerosol-cloud interaction in CMIP6 models. Geophysical Research Letters, 48, e2020GL091024. https://doi.org/10.1029/2020GL091024.

Watanabe, M., Dufresne, JL., Kosaka, Y. et al. Enhanced warming constrained by past trends in equatorial Pacific sea surface temperature gradient. Nat. Clim. Chang. 11, 33–37 (2021). https://doi.org/10.1038/s41558-020-00933-3.

Wills, R. C. J., Dong, Y., Proistosecu, C., Armour, K. C., and Battisti, D. S. (2022). Systematic climate model biases in the large-scale patterns of recent sea-surface temperature and sea-level pressure change. Geophysical Research Letters, 49, e2022GL100011. https://doi.org/10.1029/2022GL100011.

Zhang, J., Furtado, K., Turnock, S. T., Mulcahy, J. P., Wilcox, L. J., Booth, B. B., Sexton, D., Wu, T., Zhang, F., and Liu, Q.: The role of anthropogenic aerosols in the anomalous cooling from 1960 to 1990 in the CMIP6 Earth system models, Atmos. Chem. Phys., 21, 18609–18627, https://doi.org/10.5194/acp-21-18609-2021, 2021.

Zhou, C., Zelinka, M., and S. Klein, 2016: Impact of decadal cloud variations on the earth's energy budget. Nature Geoscience, 9, 871–874. https://doi.org/10.1038/ngeo2828.

Zinke, J., Browning, S.A., Hoell, A. et al. The West Pacific Gradient tracks ENSO and zonal Pacific sea surface temperature gradient during the last Millennium. Sci Rep 11, 20395 (2021). https://doi.org/10.1038/s41598-021-99738-3.