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

Article

Accepted Version

**CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

1 **Assessment of large-scale indices of surface temperature during the**

2 **historical period in the CMIP6 ensemble**

3

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

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

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

7

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

9

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

2

42  internal variability on multi-decadal trends (McKinnon and Deser, 2018). The tropical SST

43  patterns are strongly connected to regional precipitation anomalies, of relevance for the

44  accurate drought-inducing teleconnections (e.g. Annamalai et al., 2013; Zinke et al., 2021).

45  Also, the radiative forcing over the historical record is uncertain, mainly due to the role of

46  aerosols (e.g. Smith et al., 2021), with important implications for the historical warming

47  shown by models (e.g. Wang et al., 2021; Zhang et al., 2021). Potentially, all this information

48  can be used to improve the model's response to external forcing subject to the constraints of

49  process observations. However, there is no common approach on how to incorporate the

50  historical record into model development.

51      For example, several modeling centres have directly "calibrated" or "tuned" historical

52  simulations (i.e. adjusted them to improve realism of climate change simulation) during the

53  developments of the models used for the Climate Model Intercomparison Project phase 6

54  (CMIP6; Eyring et al., 2016). During the development of the Energy Exascale Earth System

55  Model version 1 (E3SMv1), a historical simulation was performed with a near-final version

56  of the model, but no action was taken to change the historical performance in the final

57  version (Golaz et al., 2019). Boucher et al. (2020) describe the developments and

58  performance of the IPSL-CM6A-LR model. Although historical simulations were not used as

59  part of the development, the r1i1p1f1 simulation was selected qualitatively among the first

60  ~12 available historical members, based on a few key observables of the historical period.

61  The historical warming of the MPI-ESM1.2-LR model was tuned by reducing its climate

62  sensitivity during its development (Mauritsen et al., 2019).

63      The use of historical runs (or any coupled run with transient forcing) for tuning is not part

64  of the Met Office Unified Model (UM) development protocol. The Hadley Centre models

65  submitted to CMIP6 were not tuned to the historical record, although several model

66  improvements were added to ensure that the total present-day radiative forcing was positive

67  (Mulcahy et al., 2019). This approach was revised in the 2020 UM Users Workshop, where it

68  was agreed that one of the key model errors was the simulation of the historical record. As a

69  result, a Prioritised Evaluation Group (PEG) was created with the objective of improving the

70  simulation of the historical global-mean surface temperature record. Also, in a recent review

71  of the UM's Global Configuration (GC) development protocol, it was agreed that a small

72  ensemble of historical simulations will be run during the final stage of the development cycle,

73  opening the option to implement model changes that target the performance of the simulation

74  of the historical record before the final configuration is delivered to the users. In this paper

75  we present the first step towards incorporating historical information into the UM's

76  development process. We develop a statistical method to test whether simulations of large-

77  scale surface temperature change are realistic during the historical period (1850-2014). The

78  method is applied to annual-mean time series of three surface temperature indices: global-

79  mean, hemispheric gradient, and a recently-developed index that captures the sea-surface

80  temperature (SST) pattern in the tropics (SST[#]; Fueglistaler and Silvers, 2021). We test the

81  historical simulations of the CMIP6 ensemble and post-CMIP6 versions of the HadGEM3

82  and UKESM models. We use the term 'realistic' in a relative manner: a model that performs

83  well against the tests described here can do so due to compensating errors (e.g. between

84  forcings and feedbacks). Consequently, those models that we label as realistic in the present

85  study could nonetheless be rejected once other metrics with additional observational evidence

86  or process understanding are considered. This shortcoming is not specific to this

87  methodology, and the method we propose here should be used along a wide range of

88  diagnostics to provide a detailed assessment.The structure of the paper is as follows. Section

89  2 describes the observational and model data. The statistical methodology is detailed in

90  Section 3, and Section 4 presents the results of the method applied to the CMIP6 historical

91  ensemble. Finally, Section 5 discusses the results and conclusions.

## 2. Model data and observations

93  We use near-surface air temperature (CMIP variable *tas*) data from the *piControl* and

94  *historical* experiments of the CMIP6 archive (Table 1), which are atmosphere-ocean coupled

95  simulations. The *piControl* are unforced simulations with forcing agents set at pre-industrial

96  levels (year 1850). After a spin up period, the CMIP6 protocol requests a minimum of 500

97  simulation years, but not all models fulfil this criterion. We explain how we deal with

98  different lengths of the *piControl* time series in the next section.

99  The CMIP 6 protocol (Eyring et al., 2016) recommended that the *historical* experiments

100 are run with the current best estimates of the time-evolving datasets of forcing agents:

101 atmospheric composition, solar irradiance, natural and anthropogenic aerosols, and land-use

102 change, but not all institutions followed the protocol. They branch from the *piControl*

103 simulation, running from 1850 to 2014 (165 years). The CMIP6 protocol recommends

104 running at least 3 *historical* simulations, branching from different points in the *piControl*

105 simulations. We use 40 *piControl* simulations from the CMIP6 ensemble, plus simulations

4

106 from GC4.0-LL and UKESM1.1-LL (Mulcahy et al., submitted), models developed after
107 CMIP6.

108     We use three different observational datasets of surface temperature: the Met Office
109 Hadley Centre/Climatic Research Unit global surface temperature data set version 5
110 (HadCRUT5.0.1.0; Morice et al., 2021), the Program for Climate Model Diagnosis and
111 Intercomparison (PCMDI) SST reconstruction (Hurrell et al., 2008; Taylor et al., 2000), and
112 the Extended Reconstructed Sea Surface Temperatures Version 5 (ERSSTv5; Huang et al.,
113 2017). The baseline period used for all historical datasets is 1880-1919.

114     HadCRUT5 provides temperature anomalies on a lat-lon rectangular grid. Two variants of
115 the same dataset are provided: a non-infilled version, with data in gridboxes where
116 measurements are available; a more spatially complete version. For global and regional time
117 series, the HadCRUT5 analysis error model contains two terms (Morice et al., 2020): the
118 analysis error ($\varepsilon_a$), and the coverage error ($\varepsilon_c$). The analysis error combines the errors from
119 the Gaussian process used in the statistical infilling and the instrumental errors. The analysis
120 grids are not generally globally complete, particularly in the early observed record. Regions
121 are omitted where there are insufficient data available to form reliable grid cell estimates. The
122 coverage error represents the uncertainty in spatial averages arising from these unrepresented
123 regions. The analysis error is represented by the 200 realizations of the historical record,
124 whereas the coverage error is reported as a time series of standard deviations. We use the
125 more spatially complete version, also termed as "HadCRUT5 analysis". The HadCRUT5
126 analysis data set uses a statistical method to extend temperature anomaly estimates into
127 regions for which the underlying measurements are informative. This makes it more suitable
128 for comparisons of large-scale regional average diagnostics against spatially complete model
129 data, although variability in "infilled" regions will be lower than where observed
130 measurement data is present (Jones, 2016). We use the HadCRUT5 analysis as a reference
131 dataset for two of the indices: global-mean, and hemispheric gradient. We use the global
132 means calculated by averaging the hemispheric means, as recommended by Morice et al.
133 (2021).

134     The SST[#] index is defined as the difference between the average of the warmest 30%
135 SSTs (actual values, not anomalies) and the domain average. The domain used for this
136 particular metric is the Tropics, from 30ºS to 30ºN. This index represents the difference in
137 SSTs between the convective regions and the tropical average, and it explains the anomalies

5

138 in low cloud cover (and cloud radiative feedbacks) over the historical record due to changes

139 in SST patterns (Fueglistaler and Silvers, 2021). The index is calculated using monthly-mean

140 SSTs, and then annual averages are calculated. The same process is followed for both models

141 and observations. Since this index cannot be calculated from local anomalies, a dataset that

142 provides absolute temperature estimates is required. The PCMDI dataset provides monthly

143 mean sea surface temperature and sea ice concentration data from 1870 to the present on a

144 regular lat-lon grid. These data are designed to be used as boundary conditions for

145 atmosphere-only simulations. They use the AMIP-II mid-month calculation (Taylor et al.,

146 2000), which ensures that the monthly mean of the time-interpolated data is identical to the

147 input monthly mean. Following the convention in other studies, we refer to this dataset as

148 PCMDI/AMIPII. $SST^{\#}$ is subject to a large observational uncertainty (Fueglistaler and

149 Silvers, 2021), attributed to the different methodologies used to provide information where

150 observations are not available. Given that the PCMDI/AMIPII dataset doesn't provide a

151 comprehensive error characterization, we use the ERSST5 to test the robustness of our results

152 to the observational uncertainty in $SST^{\#}$. We have chosen the PCMDI/AMIPII and ERSST5

153 datasets because they fall at opposite ends of the spectrum of $SST^{\#}$ anomalies provided by

154 observational datasets, spanning the range of structural uncertainties in the observational

155 reconstructions of $SST^{\#}$.There is evidence of differences between near-surface atmosphere

156 temperature and surface temperature diagnostics (e.g. Richardson et al., 2016). The

157 Intergovernmental Panel on Climate Change Assessment Report version 6 (IPCC AR6;

158 Gulev et al., 2021) quantifies the global-mean uncertainty of long-term trends by at most 10%

159 in either direction, with low confidence in the sign of any difference in long-term trends.

160 Jones (2020) supports the use of global near-surface air temperature model diagnostics with

161 blended datasets of observed temperature changes.

## 3. Methodology

163     Let $H_o(t)$ be the timeseries of the observed historical record anomalies of any given

164 surface temperature index. We decompose it as $H_O(t) = S(t) + U_O(t) + E_o(t)$, where $S(t)$

165 represents the forced signal, $U_O(t)$ is the unforced variability, and $E_o(t)$ is the total

166 observational error. Similarly, for a given model we decompose any historical simulation of

167 the same index as $H_M(t) = S(t) + D_M(t) + U_M(t)$. $D_M(t)$ represents a discrepancy term or error

168 in the forced response, and $U_M(t)$ is the model's unforced variability.

169  If we hypothesize that the model's forced response is realistic (i.e. $D_M(t)=0$), then $H_M(t) -$

170 $H_O(t) = U_M(t) - U_O(t) - E_O(t)$. We can test this hypothesis by comparing $H_M(t)$-$H_O(t)$ with the

171 expected distribution of $U_M(t) - U_O(t) - E_O(t)$. In general, we have more than one realization

172 of a model's *historical* experiment, each of them with a different realization of the model

173 unforced variability. Since we only have a single sample of the real world's unforced

174 variability, tests on individual ensemble members are not independent. We avoid this

175 problem by formulating the test for ensemble means noting that S(t) (and DM(t)) are the

176 same for each ensemble member: $\overline{H_M}(t) - H_O(t) = \overline{U_M}(t) - U_O(t) - E_O(t)$. The overbars

177 represent the ensemble mean. With this formulation, the observations are used only once for

178 each model ensemble with the contribution of their internal variability remaining constant

179 with ensemble size (unlike the contribution of the model internal variability which reduces

180 with ensemble size).

181  The problem is now reduced to the characterization of the distribution of the right-hand

182 side of the equation. Ideally, $U_O$ should be characterized from a long time series of the real

183 system under no external forcing. Paleoclimatic proxy reconstructions are available only for

184 restricted regions, and therefore not representative of the large spatial scales of interest for

185 this study, as well as having larger errors. They have the additional complication that the

186 external forcing is not zero during the paleoclimate record. Therefore, we instead assume that

187 unforced simulations of the multi-model ensemble provide us with a reasonable estimate of

188 the real world's unforced variability, an approach that has been used in other studies (e.g.

189 Gillet et al., 2002). Hence, we characterize $\overline{U_M}$ and $U_O$ using *piControl* simulations.

190  The sub-sections below describe the next steps in the methodology: calculation of the

191 observational error term; estimation of the distribution of $\overline{U_M}(t) - U_O(t) - E_O(t)$ using

192 *piControl* simulations; definition of the metric and calculation of its control distribution;

193 testing the historical ensembles; interpreting the tests.

194 *a. Calculation of the observational error*

195 For the HadCRUT5 observations, we combine analysis and coverage errors into a single error

196 term ($E_o$) as follows. We add samples of a normally-distributed random variable of zero mean

197 and variance $Var(\varepsilon_c(t))$ to the residuals of the 200 realizations of the HadCRUT5 analysis.

198 The total error inherits the autocorrelation characteristics of the analysis error, which is

199 correlated in time. $E_o$ is then modelled by drawing random samples from this 200-member

7

200   ensemble of realizations. The time-dependence of $E_o$ for the global-mean is shown in Figure

201   1. The black lines show the 95% confidence interval (comparable to the orange range in

202   Figure 2 of Morice et al., (2020)). In general, the observational error decreases with time,

203   apart from periods of international conflicts. The time-dependence of $E_o$ for the hemispheric

204   difference is very similar to that of the global-mean, but larger in magnitude.

205   For the SST[#] index, we don't include an error term due to lack of error information in the

206   observational datasets. However, we repeat the analysis with two different observational

207   datasets to test the robustness of the results.
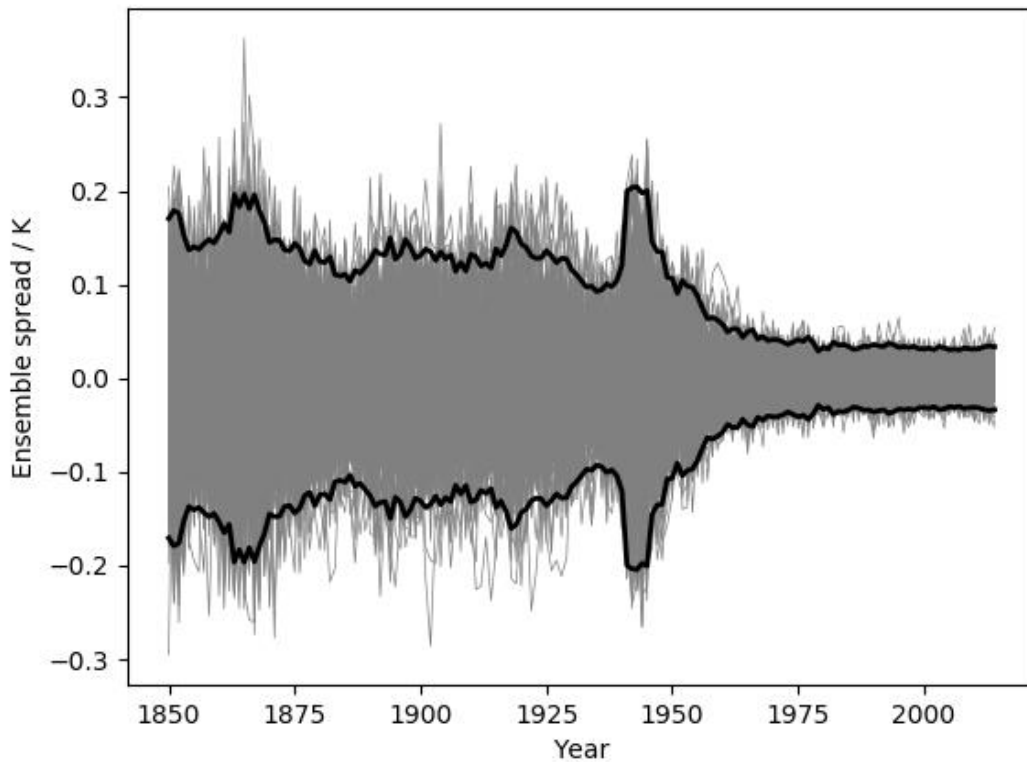


208

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

213   *b. Construction of the unforced distribution of differences*

214       Here we are concerned with the generation of random samples of $\overline{U_M}(t) - U_O(t) -$

215   $E_O(t)$ using *piControl* simulations. Although the *piControl* simulations are started after a

216   spin-up that is discarded, they are not in complete equilibrium (Eyring et al., 2016). For each

217   model's control timeseries, we construct a linearly-detrended time series (X(t)) using the

8

218    entire length of each control simulation. This increases the likelihood of adding noise to the

219    detrended data (Sen Gupta et al., 2013; Jones et al., 2013), but some models show significant

220    unforced variability on centennial timescales, which would be spuriously reduced by

221    detrending shorter segments (Parsons et al., 2020).

222        We split the detrended control time series X(t) into non-overlapping segments 165 yr

223    long, equal to the length of the CMIP6 historical simulations. The *piControl* simulations

224    differ in length between models, so to give (nearly) equal weight to each model we use up to

225    3 segments of each piControl simulation. We also decide to retain models with shorter

226    control time series. With these constraints, we use 41 *piControl* simulations, 32 of them with

227    3 segments, 5 with 2 segments, and 4 with only one segment. This gives 110 segments of

228    piControl simulations of equal length. Then, we subtract the time average of the segment, so

229    that the mean value of each segment is zero by construction. We call $U_{piControl}(t)$ to these

230    detrended, 165 yr long, zero-average piControl samples of the unforced variability, which we

231    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

232    $U_o(t)$ from the ensemble of 110 $U_{piControl}(t)$ segments. For instance, for a *historical* ensemble

233    with 10 members, we randomly draw 11 $U_{piControl}(t)$ segments, and average 10 of them to

234    calculate $\overline{U_M}(t; N_m)$, and use the other one as $U_O(t)$. The $U_{piControl}(t)$ samples are drawn from

235    the pool of *piControl* segments of all models, not only of the model whose *historical*

236    ensemble is being tested. For GMSAT and hemispheric difference, $E_o(t)$ is randomly sampled

237    from the ensemble of 200 realizations of the total HadCRUT5 total error as explained above,

238    and the three timeseries are combined. We repeat this process 10000 times for each *historical*

239    ensemble. For constructing the distribution of unforced differences, the only information

240    extracted from the *historical* ensemble is its size $N_m$.

241        Other approaches for estimating internal variability exist, and a recent study by

242    Olonscheck and Notz (2017) provide a brief description of the two main avenues and their

243    caveats. We have used a method that is based on *piControl* simulations, which may be

244    unsuitable if the unforced variability is state-dependent. However, Olonscheck and Notz

245    (2017) show that the variability remains largely unchanged for historical simulations, even

246    for those variables like sea ice area that show large changes in simulations of future warming.

247    Therefore, we assume that the variability remains unchanged for the temperature indices used

248    here and the amount of climate change in the historical period.

249    *c. Definition of the metric: number of exceedances*

250    Our interest is to characterize the quality of a historical ensemble of simulations against

251    observations. As a metric of quality, in the next section (3d) we compute $\overline{H_M}(t) - H_O(t)$, for

252    each model, and count the number of times that a running mean of the absolute value of this

253    quantity exceeds a given value.

254    The samples of $\overline{U_M}(t) - U_O(t) - E_O(t)$ generated in the previous section (3b) serve as

255    the basis to construct unforced distributions of this metric.

256    We define E($T, y, N_m$), as the number of exceedances above a threshold $T$ (in K) of a

257    filtered time series of absolute values of $\left|\overline{U_M}(t) - U_O(t) - E_O(t)\right|$. The filter applied is a

258    running mean with a window length of $y$ years. We define a 2-dimensional rectangular grid in

259    T and $y$, ranging between 0 and 0.3K, and between 1 and 10 years, respectively. We then

260    calculate 10000 values of E for each combination ($T, y$). We use an absolute threshold in

261    Kelvin, but the method could be easily reformulated in terms of a threshold defined in units

262    of standard deviations of the unforced variability.

263    Figure 2 presents a an example of this process for the global-mean surface air temperature

264    (GMSAT), leading to the calculation of one sample of E(0.1, 10, 5). The blue line shows one

265    sample of $\overline{U_M}(t) - U_O(t) - E_O(t)$. The red line is the smoothed time series of the absolute

266    value of the blue time series, using a y=10 yr running mean. The green line represents the

267    temperature threshold T=0.1 K. The value of E(0.1, 10, 5) is the number of points from the

268    red line that lie above the green line.
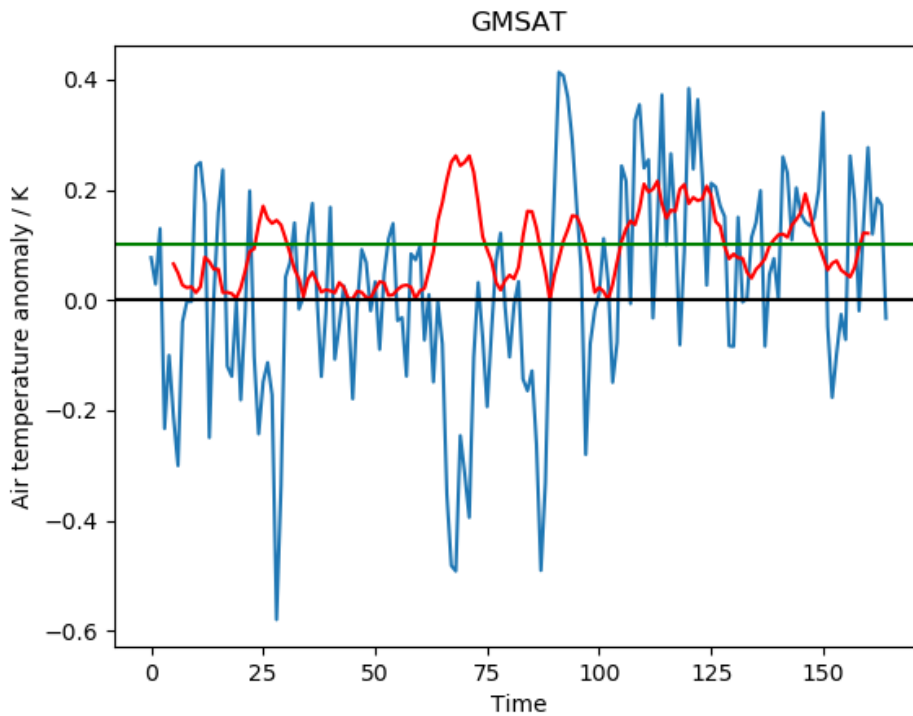
GMSAT

269

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

276

277    We construct a second metric following the same steps, but using the variance-scaled

278    samples $\sigma_M/\sigma\ (\overline{U_M}(t) - U_O(t)) - E_O(t)$, where $\sigma_M$ is the model's standard deviation of the

279    linearly-detrended *piControl* anomalies, and $\sigma$ is the multi-model mean standard deviation of

280    all the linearly-detrended *piControl* anomalies. This provides a variance-scaled set of samples

281    of control distributions of exceedances that accounts for differences in the variance of the

282    unforced variability across different models. We label this second metric as $E_s(T, y, N_m)$.

283    From these sets of samples of $E(T, y, N_m)$ and $E_s(T, y, N_m)$, we construct empirical

284    quantile distribution functions $Q_Z(p; T, y, N_m)$, which give the number of exceedances for a

285    given cumulative probability $p$. $Z$ is a generic discrete random variable name that refers to

286    either $E$ or $E_s$. For simplicity, from now on we omit the dependency with the ensemble size

287    $N_m$.

File generated with AMS Word template 2.0

288
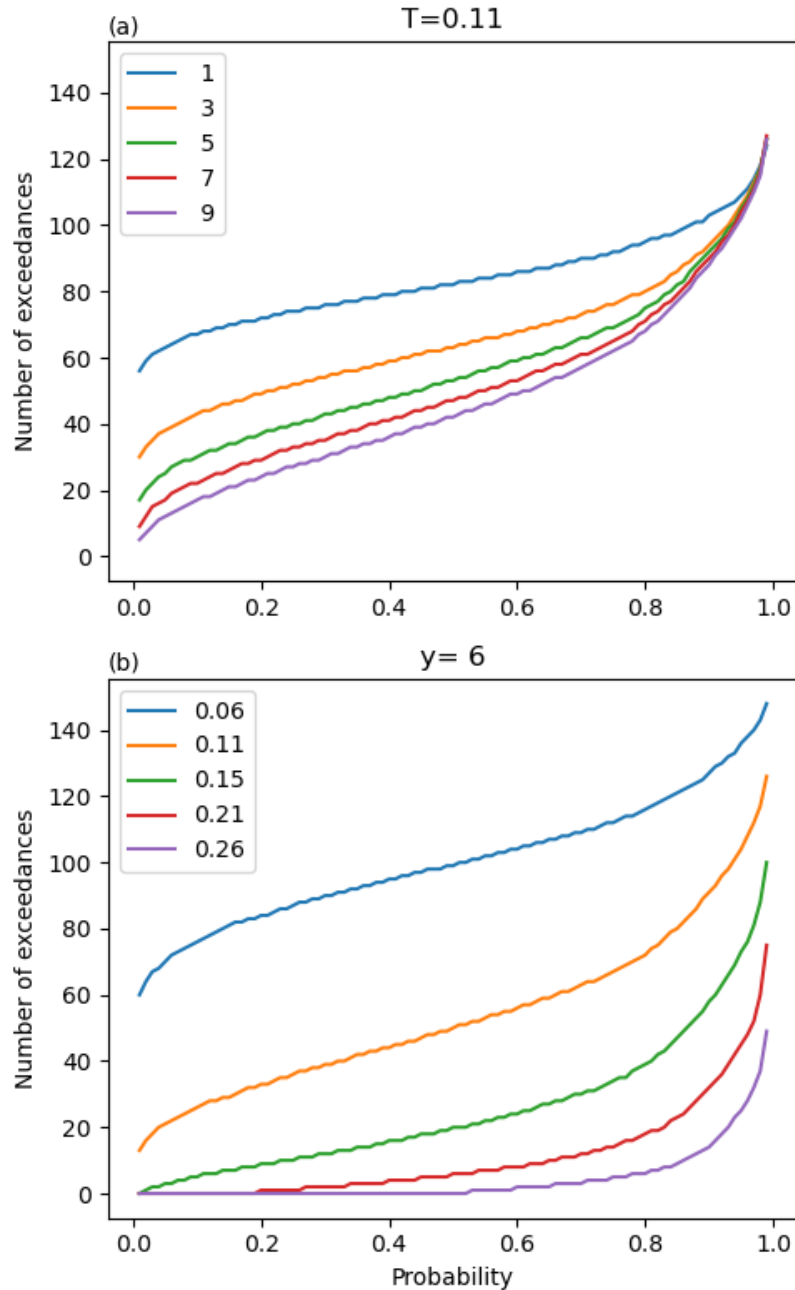


Figure 3. Examples of empirical quantile distribution functions $Q_Z(p; T, y, N_m)$ for an ensemble size $N_m=3$: (a) exceedance threshold set to 0.11K, length of the averaging window as shown in the legend (years); (b) length of averaging window of 6 years, exceedance threshold as shown in the legend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

294

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

298 probability is $p$ that the number of exceedances (occurring during a 165-year historical

299 integration) will be less than $Q_Z(p; T, y)$. There is zero chance that the number of

300 exceedances will be less than zero, a small chance that it will be less than a small number,

301 and we are certain that it will be less than a sufficiently large number (at most 165). Thus, $Q_Z$

302 increases with $p$ (Figures 3a and 3b). For any given $p$, the expected number of exceedances

303 $Q_Z$ is smaller for a longer meaning period $y$ (Figure 3a) or a higher threshold T (Figure 3b).

304 *d. Testing ensembles of historical simulations*

305 We test each *historical* ensemble by comparing the number of exceedances of the

306 difference between the ensemble mean and the observations against the expected number of

307 differences given by the control distribution. First, we calculate $\overline{H_M}(t) - H_O(t)$, which we

308 use as input to calculate the number of exceedances for each point in the *(T, y)* grid, $E_h(T, y)$,

309 where the subscript $h$ denotes that this is calculated from a *historical* ensemble, and $H_O(t)$ is

310 the HadCRUT5 analysis ensemble mean. The linear drift of the *piControl* is subtracted from

311 the *historical* time series. We then perform two one-tailed tests, each with a significance level

312 α. This is done by comparing $E_h(T,y)$ against the empirical quantile function $Q_Z(p; T,y)$,

313 separately for Z=E and Z=E$_s$. In each case, when either $E_h(T,y) > Q_Z(1-\alpha; T,y)$ or

314 $E_h(T,y) < Q_Z(\alpha; T,y)$, the historical ensemble is flagged as incompatible in that point of the

315 *(T,y)* grid. That is, we reject the null hypothesis that the difference between the historical

316 simulation and observations is consistent with unforced variability if the number of times Z

317 that the difference between them exceeds the threshold $T$ in $y$-year means is either much

318 larger than expected (upper-tail test), or much smaller than expected (lower-tail test).

319 Figure 4 shows an example for the upper tail test applied to the entire *(T, y)* grid, using a

320 significance level α=0.05. For illustrative purposes, it is helpful to choose a model like EC-

321 Earth3-Veg with large multidecadal unforced variability (Parsons et al., 2020). The filled

322 contours in Figures 4a and 4c show $Q_Z(p=0.95; T, y)$ for Z=E in and Z=E$_s$, respectively. The

323 shape of $Q_Z$ is very similar for all models and ensemble sizes. As shown also in Figure 3, $Q_Z$

324 gets smaller as $T$ gets larger for a given $y$ (less likely to exceed a higher threshold), and

325 smaller as $y$ gets larger for a given $T$ (less likely to for a longer time mean to exceed a

326 threshold), although the dependency on $y$ is much weaker.
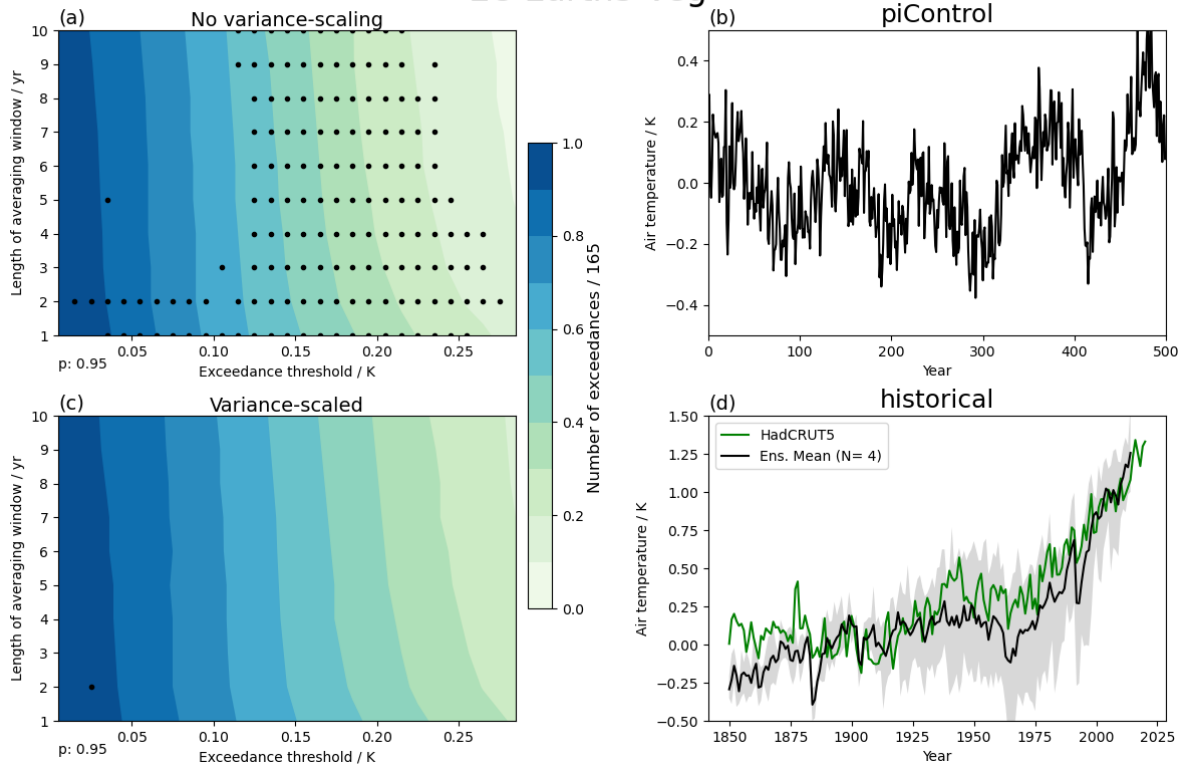
Figure 4. Tests of the *historical* ensemble of EC-Earth3-Veg. Tests without and with variance-scaling are shown in (a) and (b), respectively. The filled contours show $Q_Z(p=0.95; T, y)$ for (a) $Z=E$, and (c) $Z=E_s$. These surfaces show the expected number of exceedances normalized by 165 (maximum number of exceedances) for the 95$^{th}$ 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.

14

349      How much of the *(T, y)* space is needed to fail the statistical test for the model as a whole

350    to be deemed "incompatible"? We have divided the *(T, y)* grid into 29×10 points, so we

351    would expect a good model to fail at 290×10×0.05≈15 points in the *(T, y)* grid just by chance

352    if there was no correlation in the number of exceedances between *(T, y)* neighbors. Because

353    the time-scale and the threshold are correlated, if incompatibility occurs it is likely to cover

354    patches of adjacent points in the *(T, y)* grid. Given that our main aim is to apply this method

355    to intercompare models, we do not define a single, strict threshold for labelling a model as

356    incompatible with the observations. Instead, we use the following guidance: models with less

357    than 10 failures (dots) pass the test; models that fail between 10 and 20 times are considered

358    marginal; model with more than 20 failures are labelled as incompatible.

359      The lower tail test $(E_h(T,y) < Q_Z(p=0.05; T,y))$ can be presented in a similar way, but only

360    one of the models tested fails this test (FGOALS-g3, and only marginally). A model fails this

361    test if its historical simulation deviates less than expected from reality, which can happen

362    only if it has both a realistic forced response and unrealistically small unforced variability. It

363    could be that the lower-tail test rarely fails because models in general do not have a realistic

364    forced response. For the remainder of the paper we discuss the results of the upper-tail test

365    only.

366      We have tested the sensitivity of the results to the order of the polynomial used for the

367    detrending of the *piControl* time series. The results are largely insensitive to the use of

368    quadratic instead of linear detrending, so we conclude that our method is robust with respect

369    to the detrending method. If this test is applied to metrics that require non-linear detrending

370    we would recommend the use of more flexible methods with better properties (e.g. splines).

## 371  **4. Results and discussion**

372      In this section we present results for three temperature indices: global mean, hemispheric

373    difference, and SST[#]. These three metrics capture important complementary information

374    about key aspects of temperature change over the historical record. The global mean has been

375    widely used as the most fundamental metric of climate change. The hemispheric difference

376    captures the influence of anthropogenic aerosols during the historical period, as emissions are

377    dominated by sources in the Northern Hemisphere, and it is reasonably independent of the

378    global mean (Braganza et al., 2003). The changes in tropical SST pattern control the sign and

379  strength of low cloud feedbacks in response to $CO_2$ forcing (e.g. Miller, 1997; Gregory and

380  Andrews, 2016), making it an important metric of the historical record.

381  *a. Global mean*

382  Figure 5 shows the tests without variance scaling. Out of the 40 models analyzed, 20 of

383  them can be labelled as incompatible with the observed record, according to this test. These

384  are models that show large, dotted areas. The other 20 models do not fail the test at all or only

385  in a few instances. Models tend to fail the test for large exceedance thresholds $T$, with little

386  dependence in the length of the averaging window $y$, i.e. they tend to fail along entire

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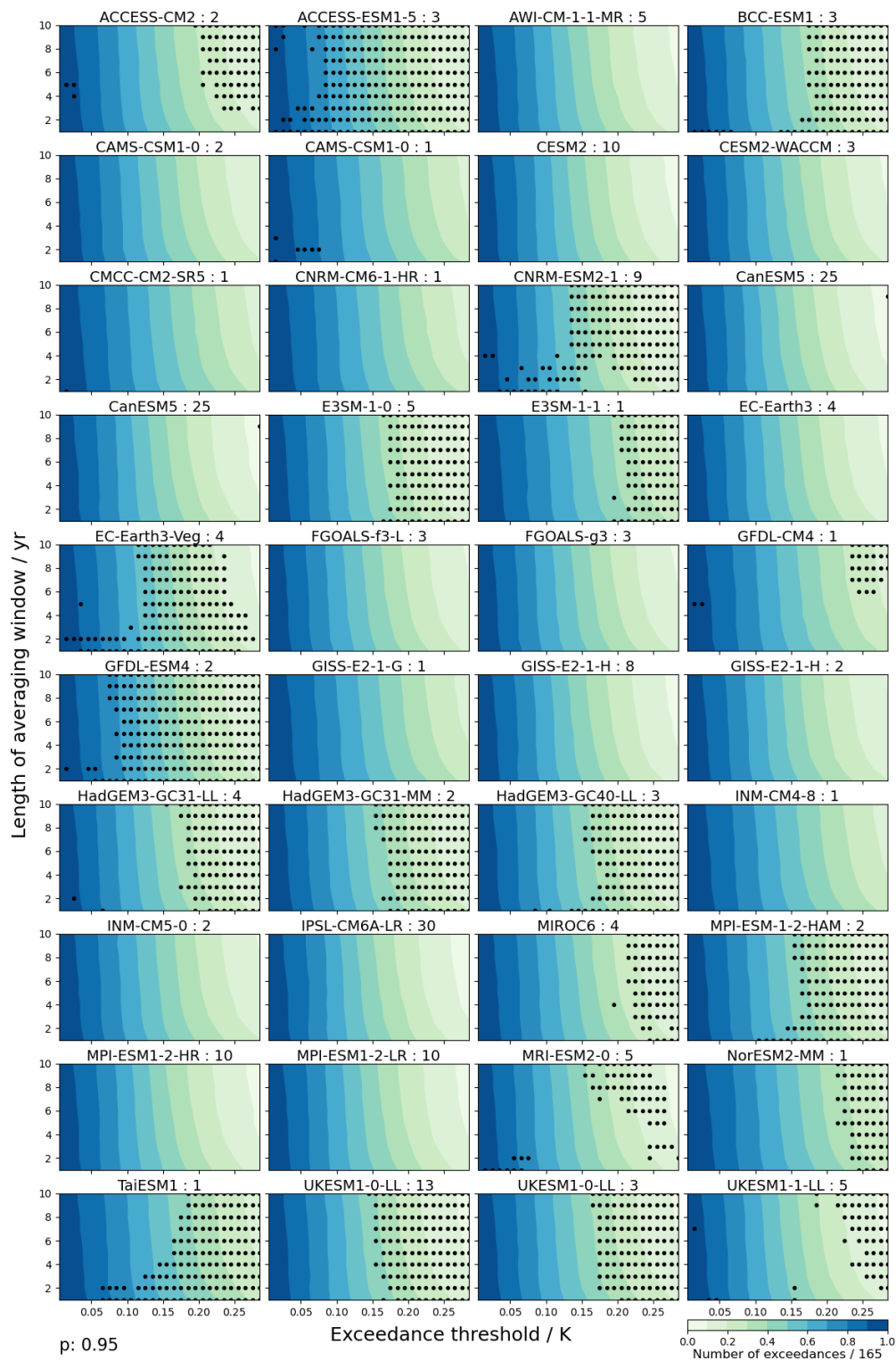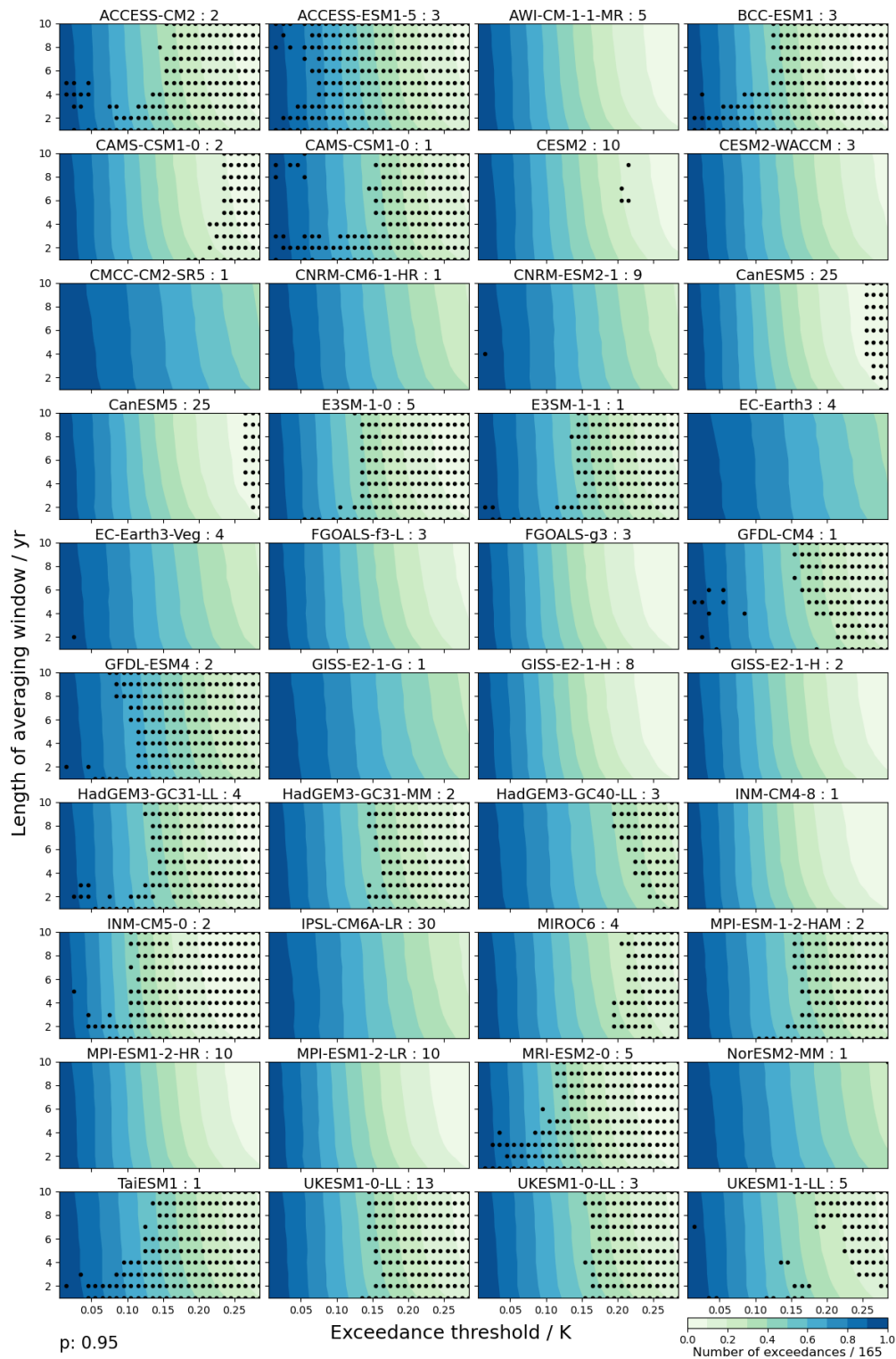
Figure 5. Multi-model summary of the test without variance scaling applied to the global-mean surface air temperature index. The number of exceedances is normalized by 165, the maximum number of exceedances given by the length of the historical record.

388

389
390
391

392

393    When the variance-scaled test is applied (Figure 6), 22 models are labelled as

394    incompatible with the observed record, and 18 models pass the test. No models are in the

395    marginal category. The variance-scaled test rejects 5 additional models, and labels as

396    compatible 3 models that were rejected by the test without variance scaling. This is because

397    these models have a *piControl* variance that is very different to the multi-model mean

398    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.

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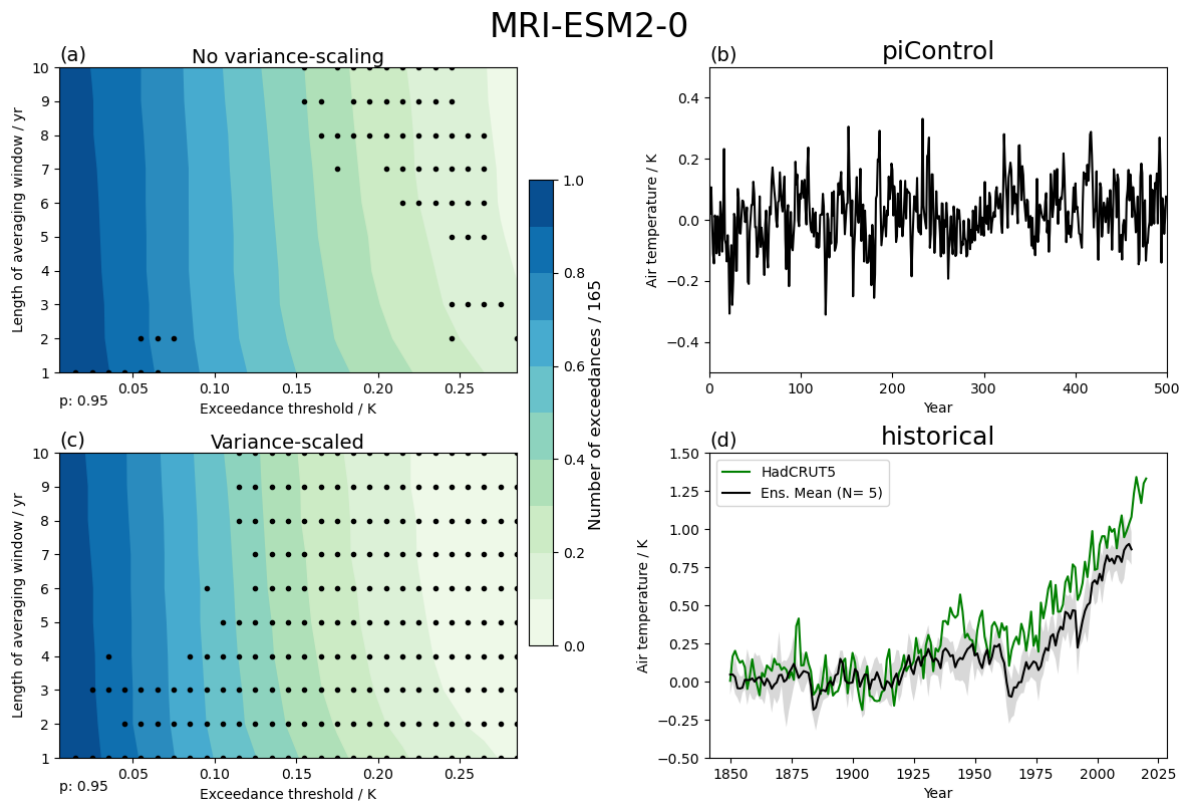


410

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

412

413     These two examples show how each model's characteristics of its unforced variability are
414     incorporated into the test. This is particularly helpful when the ensemble size of historical
415     simulations is small, which makes difficult the assessment of the impact of the unforced
416     variability by visual inspection. It must be emphasized that we treat all *piControl* simulations
417     as equally plausible, but the method could be refined by bringing in external information to
418     better characterize the unforced variability of the real system. We expand on this below when
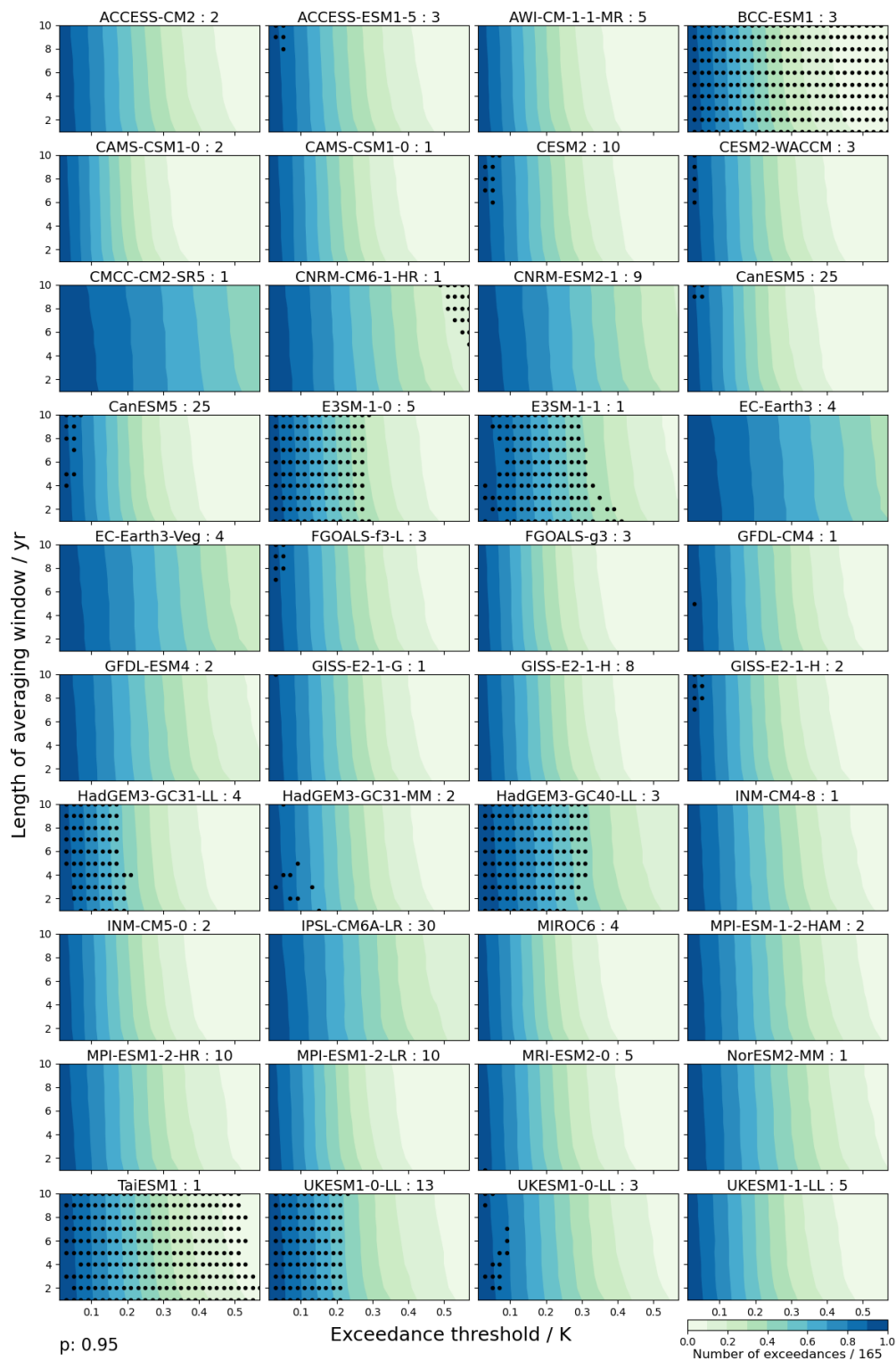419     we discuss the caveats of the methodology.
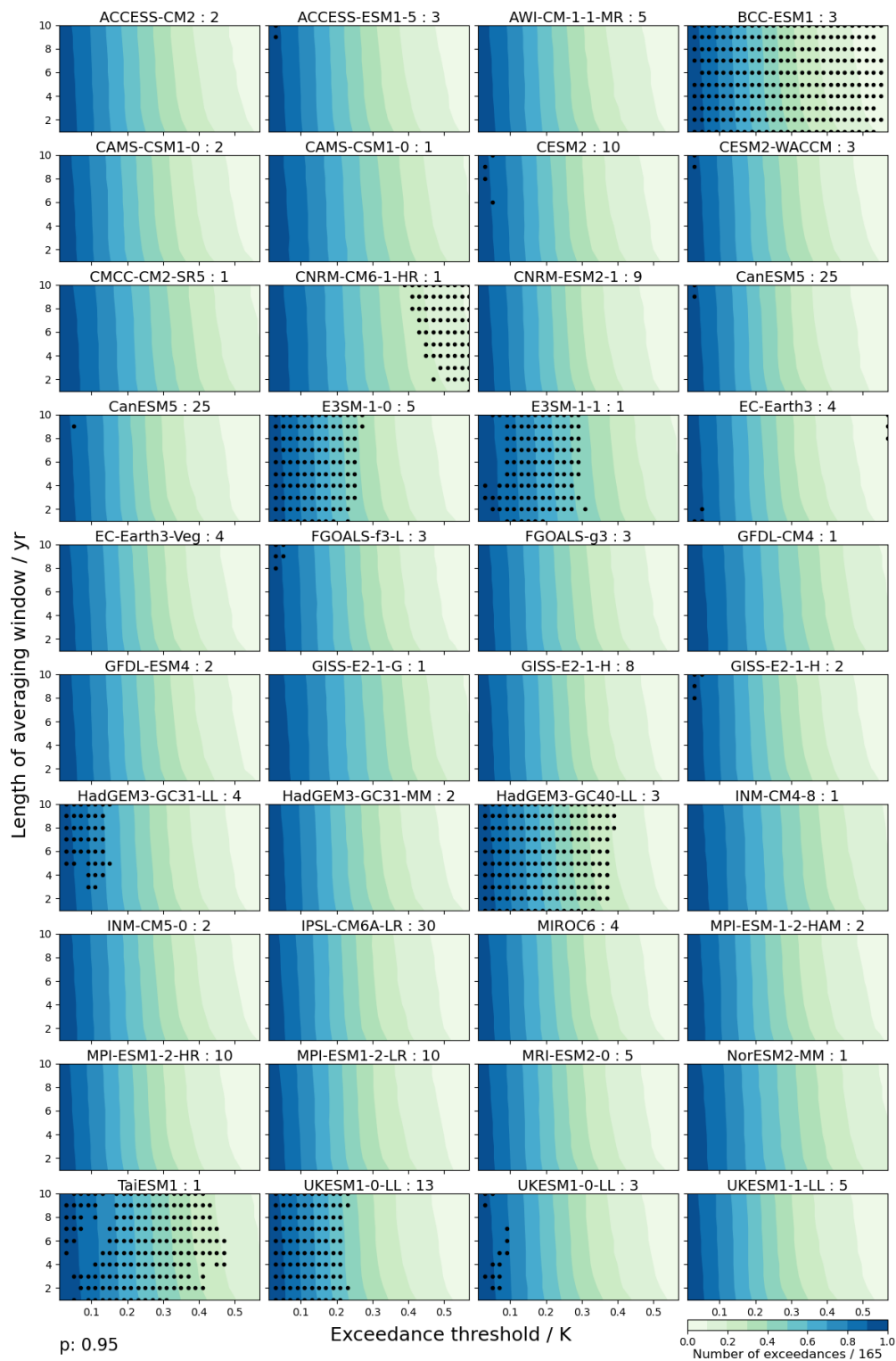
Figure 8. Same as Figure 5, but for the hemispheric gradient surface air temperature index.

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

*b. Hemispheric gradient*

22

428    Figure 8 shows the tests without variance scaling for the hemispheric gradient index. Out

429    of the 40 models analyzed, 8 are labelled as incompatible with the observed record, 1 is

430    marginal, and 31 pass the test. If variance scaling is used (Figure 9), the results are very

431    similar, with 7 models rejected, 3 marginal, and 29 passing the test. As with the global-mean,

432    failures tend to happen along 'columns', i.e. for all averaging window lengths. It is interesting

433    to note that, contrary to the global-mean, the hemispheric gradient shows more failures for

434    small exceedance thresholds.

435    In CESM2 there is a strong sensitivity of the hemispheric gradient to the variability in

436    biomass emissions from 40ºN to 70ºN, which leads to spurious warming in the late historical

437    period (Fasullo et al., 2022). However, this model passes the global and hemispheric tests,

438    which may suggest the presence of compensating biases. This highlights the importance of

439    having a large battery of diagnostics capable of assessing model performance from different

440    angles.

441    *c. SST[#]*

442    Figures 10 and 11 show the multi-model ensemble results for SST[#], without and with

443    scaling of the unforced variance, respectively. The test without variance scaling rejects all

444    CMIP6 models. Only one model is not rejected, namely GISS-E2-1-G, when variance scaling

445    is used. The GISS models are examples of models with large unforced variability (Figure 12).

446    Unlike in previous examples with large unforced variability on long time scales (Figure 4),

447    the unforced variability of the GISS models is dominated by high-frequency (annual)

448    variability. Given that the observational record does not show such a large high-frequency

449    variability, we conclude that the test without variance scaling is probably a better assessment

450    of the performance of the GISS models. This conclusion is also supported by Orbe et al.

451    (2020) who show that GISS-E2-1-G is an outlier in the simulation of ENSO.
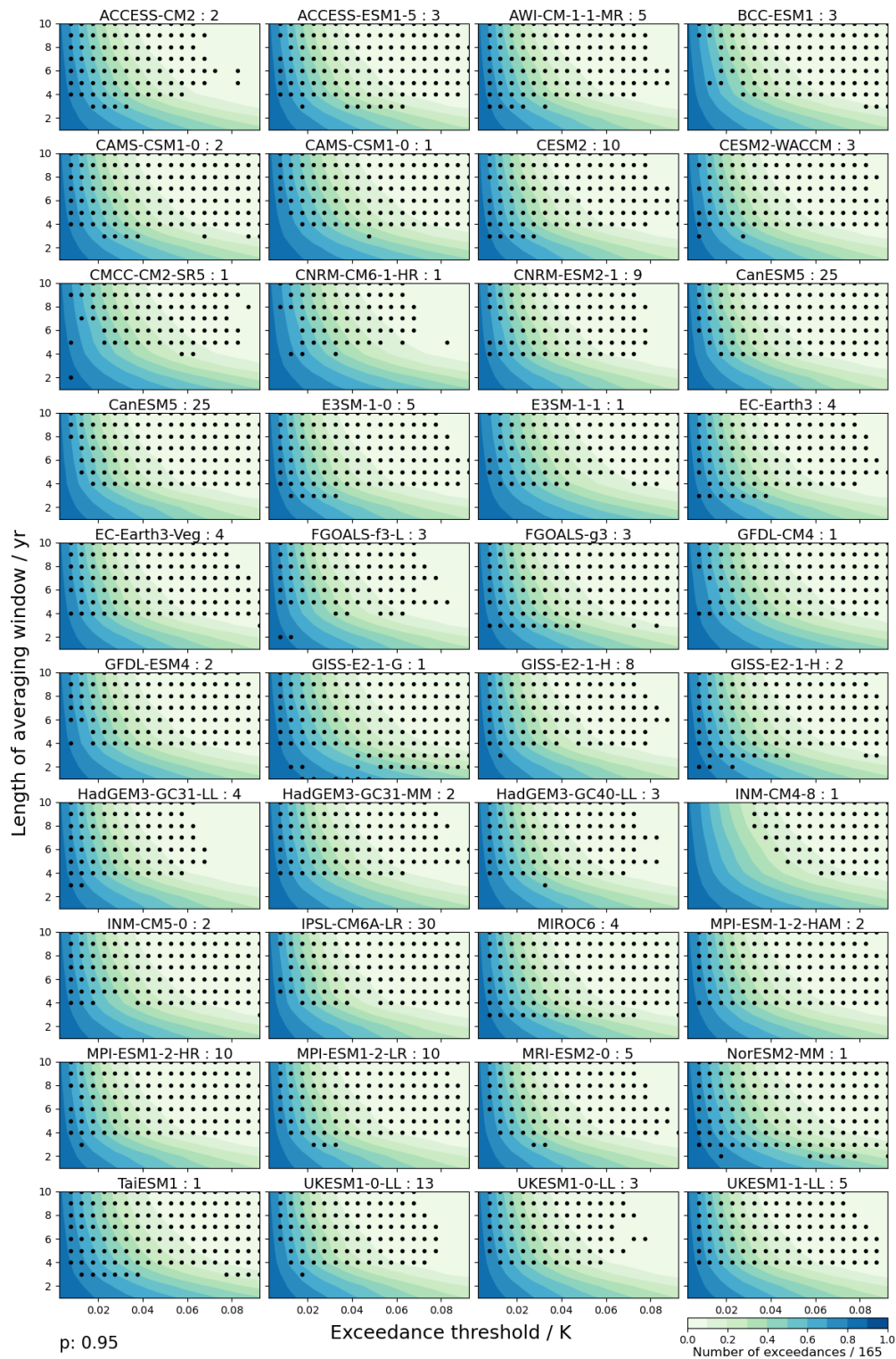
Figure 10. Same as Figure 5, but for the SST$^{\#}$ index. The observational SST$^{\#}$ index is calculated using the PCMDI/AMIPII dataset.
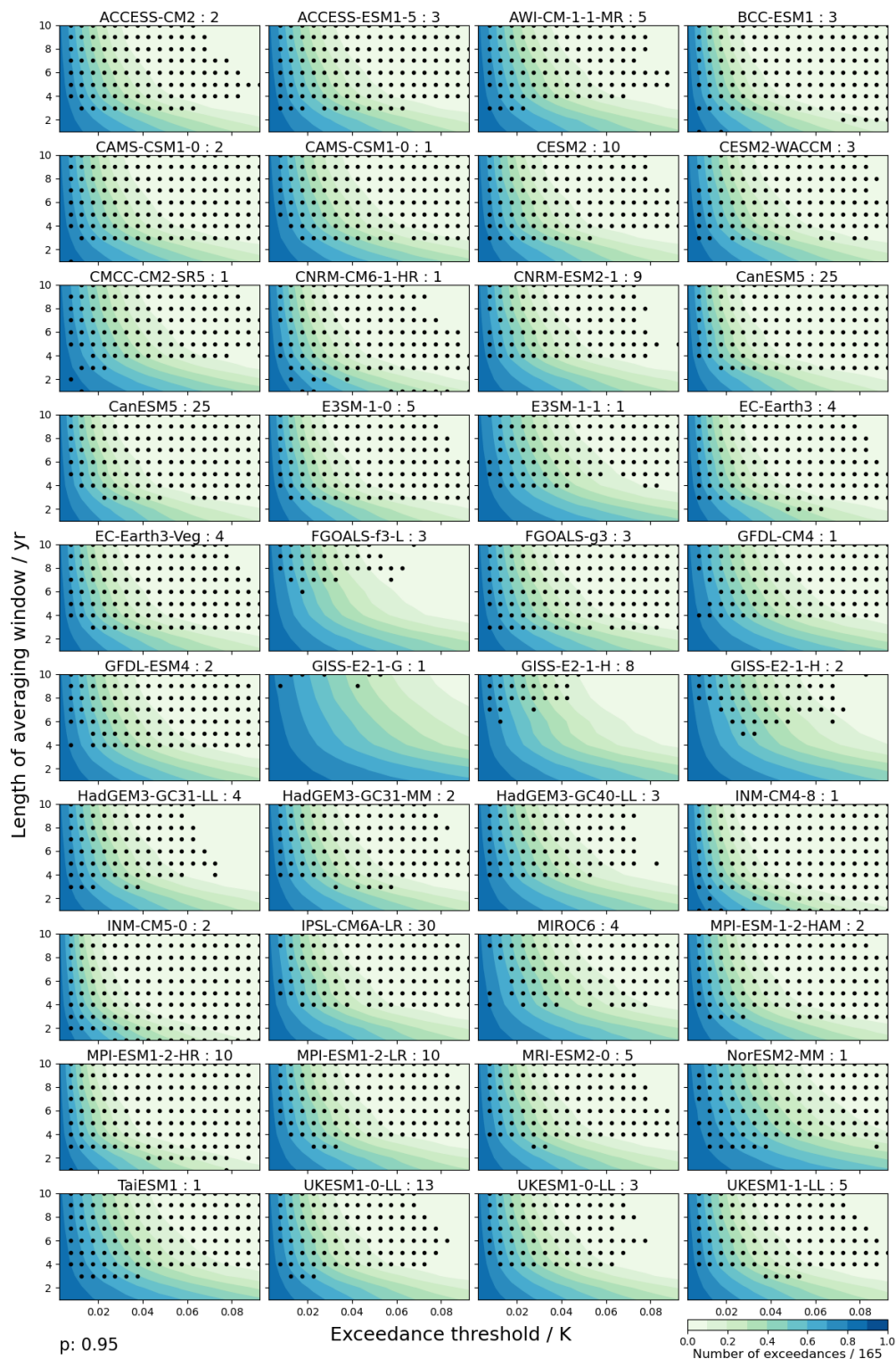
Figure 11. Same as Figure 6, but for the SST[#] index. The observational SST[#] index is calculated using the PCMDI/AMIPII dataset.

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

File generated with AMS Word template 2.0

462  spatial coverage is very dense, but they show large discrepancies before satellite data were

463  available. The differences are attributed to the different methodologies used to provide

464  information where observations are not available. Given that the PCMDI/AMIPII dataset

465  doesn't provide a comprehensive error characterization, we have repeated the tests using the

466  ERSST5 dataset to test the robustness of our conclusions. We have chosen the

467  PCMDI/AMIPII and ERSST5 datasets because they fall at opposite ends of the spectrum of

468  SST# anomalies provided by observational datasets, giving us information about structural

469  uncertainties in the observational reconstructions of SST#. The results with ERSST5 (not

470  shown) are similar to the comparisons against PCMDI/AMIPII, all the CMIP6 models are

471  rejected by both tests, with and without variance scaling. This confirms that the results are

472  robust with respect to observational uncertainty in SST#.
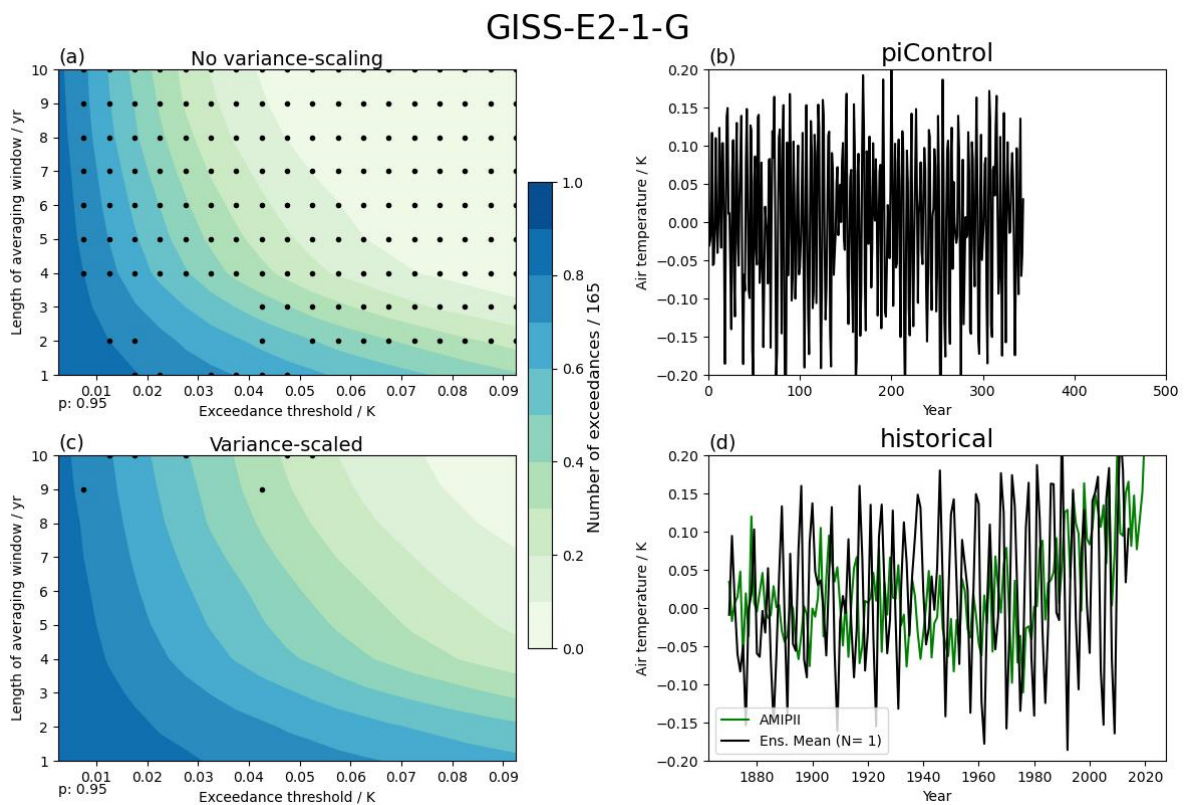


473

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

476

477  The fact that the entire CMIP6 ensemble performs poorly in the SST# index is consistent

478  with previous studies showing that models in general do not reproduce the Pacific SST trends

479  of recent decades (Seager et al., 2019; Gregory et al., 2020; Wills et al., 2022), and it has

480  potential implications beyond the models' performance over the historical period.

26

File generated with AMS Word template 2.0

481　　　　Unlike for the two other indices, there is no consensus either that SST[#] should contain a

482　　　forced signal or that it is part of the unforced variability of the climate system. Some recent

483　　　studies suggest that tropical Pacific SST patterns observed during the recent decades could

484　　　arise from internal climate variability (e.g. Olonscheck et al, 2020; Watanabe et al. 2021).

485　　　Other studies suggest that the SST patterns are consistent with a forced response to

486　　　greenhouse forcing (Seager et al., 2019) that can be explained with simple models (Clement

487　　　et al., 1996), or with a potential role for volcanic or anthropogenic aerosols in setting the

488　　　recent patterns (Gregory et al., 2020; Heede and Fedorov, 2021; Dittus et al., 2021). If the

489　　　observed evolution of SST[#] is not forced, no model ensemble-mean can be expected to agree

490　　　with the observations. In that case, if a model fails the test, it means that its simulation of

491　　　SST[#] variability has the wrong magnitude. On the other hand, if SST[#] is forced, the rejection

492　　　of the test means that the model doesn't replicate the forced response. In this case, if a large

493　　　number of models fail the test it could imply a common bias in the forced response. In either

494　　　case, a rejection of the test indicates some aspects of the model performanceare wrong

495　　　somehow. Additional process-level analysis and physical hypothesis-testing is required to

496　　　improve our understanding of the causes behind the model errors.

497　　　*d. Caveats and interpretation of the tests*

498　　　　The results above show how the methodology presented here can be used to assess

499　　　historical simulations during the model development process. We have applied it to surface

500　　　temperature indices, but it can be applied to any variable for which observational estimates

501　　　over the historical period exist. However, the methodology presents some interpretation

502　　　challenges and caveats. How do we interpret a rejection of the null hypothesis that the

503　　　model's forced response is realistic? Can we definitively conclude that there is a problem

504　　　with the model's forced signal? There is a chance that the null hypothesis is wrongly rejected

505　　　although true; that is a Type I error, whose probability is the chosen significance level. If we

506　　　reject the null hypothesis, we must have an alternative hypothesis. Potential alternatives are:

507　　　there is a problem with the model's forced signal; our model-based unforced variability is

508　　　biased; the forcing is wrong. We do not have a statistical means to estimate the probability of

509　　　these systematic errors.

510　　　　It is also worth mentioning that agreement between the observations and simulations

511　　　might be due to compensating errors. Potential problems that could contribute to

512　　　compensating errors concern the following: aerosol radiative forcing and aerosol-cloud

File generated with AMS Word template 2.0

513 interactions (e.g. Paulot et al., 2018; Rieger et al., 2020; Wang et al., 2021; Fasullo et al.,

514 2022); tropical SST patterns and their role on global radiative feedbacks (Ceppi and Gregory,

515 2017; Andrews and Webb, 2018). The unforced distributions used to define the exceedance

516 quantile functions are constructed from *piControl* simulations. This assumes that the multi-

517 model ensemble provides us with a good representation of the unforced variability, which is

518 not necessarily true. As we have shown above when discussing the results of the variance-

519 scaled results, there exist large discrepancies in the representation of unforced variability

520 between models (Parsons et al., 2020), which raises questions about the ability of at least

521 some models to provide a good estimate of unforced variability. If the unforced variability

522 estimated from the multi-model ensemble is biased, then our method will be biased. One

523 avenue that could be explored for improving this would be to incorporate information from

524 proxy temperature reconstructions into a correction of the unforced variability. However, the

525 use of proxy reconstructions is not free from problems. The reconstructions are for restricted

526 regions where there are proxies (e.g. PAGES 2k Consortium, 2013), and much of their

527 variability is forced by volcanoes and solar variability (PAGES 2k Consortium, 2019). In any

528 case, a failure of this type would imply that the models *piControl* simulations are wrong

529 (rather than the forced signal necessarily), so the test would still be highlighting a problem.

530     Our test with scaled variance is an initial attempt to identify outliers, but more

531 sophisticated methods could be used. Perhaps a better estimate of the unforced variability

532 could be achieved by restricting the set of models used to form the distributions of internal

533 variability. This selection could be based on how models represent observational estimates of

534 the spectra of some modes of variability (Fasullo et al., 2020). For SST[#], basing this selection

535 on some metric of ENSO could be particularly useful. Screening out models would reduce

536 the number of *piControl* simulations, so this would have an impact the robustness of the

537 unforced distributions.

538     A second caveat is the differing sizes of the historical ensembles. Out of the 40 models

539 analyzed here, only 4 have *historical* ensembles with more than 10 members, and 31 models

540 have 5 or fewer historical simulations. Large ensembles will provide more robust tests. A

541 model with a small ensemble will provide a less precise estimate of the ensemble mean,

542 making the result of the test more likely to be different from the result that would be obtained

543 with a large ensemble. This is a general problem with statistical hypothesis testing, and it

544 should be incorporated into the subjective interpretation of the tests. We propose some

545 guidance based on the dependence of the variance of the control distribution with the size of

546 the ensemble. As explained above, the control distribution is constructed from samples of

547 $\overline{U_M}(t; N_m) - U_O(t) - E_O(t)$. The observational error is typically small compared to the

548 unforced variability, so we can approximate dependence of the variance as $(1 + 1/N_m)*\sigma$,

549 where $\sigma$ is the variance of $U_M(t)$ and $U_o(t)$. As $N_m$ becomes larger, the total variance

550 decreases from 2 (in units of $\sigma$) to its asymptotic value of 1, with the rate of change being

551 larger for small $N_m$. For instance, an ensemble of 10 members will reduce the variance to

552 within 10% of its asymptotic value, which will significantly increase the robustness of the

553 test.

554    We do not account for the uncertainty in radiative forcing, which could lead to overtuning

555 if the only objective is to match the warming over the historical period (e.g. Hourdin et al.,

556 2017). However, we are not advocating making development choices only based on the

557 approach presented here. A wide range of other metrics, including process-based metrics

558 need to be considered. The use of a much wider basket of metrics should reduce the risk of

559 overtuning.

560    A final caveat is that the variance scaling can't account for differences in models'

561 *piControl* variability on different timescales, so while the overall variability of two models

562 can be scaled to be similar the interannual/multidecadal variability could be still very

563 different. We have subjectively accounted for this in the discussion of the SST[#] results for

564 GISS-E2-1-G, whose variability is dominated by large interannual variability, which can be

565 confidently assessed with observations of the historical period. However, this is not the case

566 for variability at much longer timescales, for which the observational record provides much

567 limited information. A possible approach to look at in the future is to account for this by

568 applying different variance scaling factors for each p.

## 5. Conclusions

570    The historical record of surface temperature is an important metric that climate models

571 should be able to reproduce. However, it is not consistently used by modelling centres during

572 model development for two main and quite distinct reasons: first, coupled simulations are

573 expensive to run, especially because the historical simulation must be preceded by a spin-up

574 simulation long enough to eliminate drift; second, the observed historical record of surface

575 temperature is reserved as an out-of-sample validation. It is generally argued that the

576   warming during the historical record and emergent properties like equilibrium climate

577   sensitivity should be used as an a posteriori evaluation and not as a target for model

578   development, although there is not complete consensus among the modelling community on

579   this topic (Hourdin et al., 2017). Bock et al. (2020) highlight the risk of tuning models to

580   reproduce a set of metrics ignoring deficiencies elsewhere. However, this risk is not specific

581   to metrics based on historical warming. Within the context of emergent constraints, Eyring et

582   al. (2019) advocate the use of variability metrics or trends during model development.

583   We develop a statistical method to test whether simulations of large-scale surface

584   temperature change are consistent with the observed warming of the historical period (1850-

585   2014). The method uses information on a range of time scales. It incorporates information

586   about unforced variability, and it is designed to test an entire ensemble of simulations of any

587   size. The method is applied to annual-mean time series of three surface temperature indices:

588   global-mean, hemispheric gradient, and a recently-developed index that captures the sea-

589   surface temperature (SST) pattern in the tropics ($SST^{\#}$; Fueglistaler and Silvers, 2021). We

590   test the historical simulations of the CMIP6 ensemble and post-CMIP6 versions of the

591   HadGEM3 and UKESM models.

592   Around half the models fail the test for the global-mean time series, approximately a fifth

593   of the models fail when the hemispheric temperature gradient is analyzed, and all models fail

594   the $SST^{\#}$ test. We note the importance of the characteristics of the models' unforced

595   variability (Parsons et al., 2020). Assessment of the quality of the historical simulations by

596   visual comparison of the time series of a few ensemble members against the observations can

597   be misleading, being reliable only for models with a large number of historical realisations.

598   The method presented here complements other statistical approaches that have previously

599   compared historical model simulations to observations (e.g. Sanderson et al., 2015; Brunner

600   et al., 2020). Given that most modeling centres only run a small number of historical

601   simulations, a method like the one presented here that accounts for the unforced variability is

602   desirable, especially if the aim is to use it during the model development process, where large

603   ensembles are not affordable.

604   We show that the method presented here can be used as a tool to assess historical

605   simulations during the development process. The method is easy to apply and summarises a

606   large amount of information in two plots, with and without variance-scaling. It accounts for

607   the unforced variability of the model tested, and it can be applied to an ensemble of historical

30

608    simulations of arbitrary size. We also plan to make this methodology available to the

609    community by implementing it in ESMValTool (Eyring et al., 2020).

610        There are several avenues that could be explored to develop this method further. One

611    potential improvement could be to incorporate information from proxy reconstructions to

612    improve the estimate of the unforced variability, currently based on control model

613    simulations. However, this may prove difficult given that many proxies do not resolve annual

614    variability, and because of the non-stationarity of the magnitude of internal variability.

615    Perhaps a better estimate of the unforced variability could be achieved by restricting the

616    model set used to form the distributions of internal variability based on how models represent

617    observational estimates of annual to decadal modes of variability (Fasullo et al., 2020).

618        A second area for further developments could be to apply a scaling factor, as it is done in

619    optimal fingerprinting (e.g. Allen and Tett, 1999). Some of the models that are rejected by

620    our current methodology could pass the test if they are appropriately scaled. The

621    interpretation of the test results with the scaled time series is not straight forward, but it may

622    be useful to know that a model that is rejected could be made realistic by a scaling factor.

623

634    *Data Availability Statement.*

635        HadCRUT.5.0.1.0 data were obtained from http://www.metoffice.gov.uk/hadobs/hadcrut5

636    on 15/02/2021 and are © British Crown Copyright, Met Office 2021, provided under an Open

637    Government License, http://www.nationalarchives.gov.uk/doc/open-government-

638    licence/version/3/. PCMDI AMIP SSTs were obtained from the ESGF archive, variable

31

File generated with AMS Word template 2.0

639    tosbcs from input4MIPs, version v20220201. NOAA_ERSST_V5 data provided by the

640    NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, from their Web site at

641    https://psl.noaa.gov/data/gridded/data.noaa.ersst.v5.html (Huang et al., 2017).

642

643                                  REFERENCES

644    Allen, M., and Tett, S., Checking for model consistency in optimal fingerprinting. Climate

645        Dynamics 15, 419–434 (1999). https://doi.org/10.1007/s003820050291.

646    Andrews, T., & Webb, M. J. (2018). The Dependence of Global Cloud and Lapse Rate

647        Feedbacks on the Spatial Structure of Tropical Pacific Warming, Journal of Climate,

648        31(2), 641-654.

649    Annamalai, H., Hafner, J., Sooraj, K. P., and Pillai, P. (2013). Global Warming Shifts the

650        Monsoon Circulation, Drying South Asia, Journal of Climate, 26(9), 2701-2718.

651    Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., et al. (2020).

652        Quantifying progress across different CMIP phases with the ESMValTool. Journal of

653        Geophysical Research: Atmospheres, 125, e2019JD032321.

654        https://doi.org/10.1029/2019JD032321

655    Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., & Bastrikov, V., et al.

656        (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. Journal of

657        Advances in Modeling Earth Systems, 12, e2019MS002010.

658        https://doi.org/10.1029/2019MS002010.

659    Braganza, K., Karoly, D., Hirst, A. et al. Simple indices of global climate variability and

660        change: Part I – variability and correlation structure. Climate Dynamics 20, 491–502

661        (2003). https://doi.org/10.1007/s00382-002-0286-0.

662    Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R.:

663        Reduced global warming from CMIP6 projections when weighting models by

664        performance and independence, Earth Syst. Dynam., 11, 995–1012,

665        https://doi.org/10.5194/esd-11-995-2020, 2020.

666    Ceppi, P., and Gregory, J. M., 2017: Relationship of tropospheric stability to climate

667        sensitivity and Earth's observed radiation budget, Proc. Nat. Acad. Sci., 114(50), 13126-

668        13131, 10.1073/pnas.1714308114.

File generated with AMS Word template 2.0

669    Dittus, A. J., Hawkins, E., Wilcox, L. J., Sutton, R. T., Smith, C. J., Andrews, M. B., &

670        Forster, P. M. (2020). Sensitivity of historical climate simulations to uncertain aerosol

671        forcing. Geophysical Research Letters, 47, e2019GL085806.

672        https://doi.org/10.1029/2019GL085806.

673    Dittus, A. J., Hawkins, E., Robson, J. I., Smith, D. M., & Wilcox, L. J. (2021). Drivers of

674        recent North Pacific Decadal Variability: The role of aerosol forcing. Earth's Future, 9,

675        e2021EF002249. https://doi.org/10.1029/2021EF002249.

676    Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K.

677        E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6)

678        experimental design and organization, Geosci. Model Dev., 9, 1937–1958,

679        https://doi.org/10.5194/gmd-9-1937-2016, 2016.

680    Eyring, V., Cox, P.M., Flato, G.M. et al. Taking climate model evaluation to the next level.

681        Nature Clim Change 9, 102–110 (2019). https://doi.org/10.1038/s41558-018-0355-y.

682    Eyring, V., et al., Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set

683        of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth

684        system models in CMIP, Geosci. Model Dev., 13, 3383–3438,

685        https://doi.org/10.5194/gmd-13-3383-2020, 2020.

686    Fasullo, J. T., Phillips, A. S., & Deser, C. (2020). Evaluation of Leading Modes of Climate

687        Variability in the CMIP Archives, *Journal of Climate*, *33*(13), 5527-5545.

688    Fasullo, J. T., Lamarque, J.-F., Hannay, C., Rosenbloom, N., Tilmes, S., DeRepentigny, P., et

689        al. (2022). Spurious late historical-era warming in CESM2 driven by prescribed biomass

690        burning emissions. Geophysical Research Letters, 49, e2021GL097420.

691        https://doi.org/10.1029/2021GL097420

692    Flynn, C. M., and Mauritsen, T.: On the climate sensitivity and historical warming evolution

693        in recent coupled model ensembles, Atmos. Chem. Phys., 20, 7829–7842,

694        https://doi.org/10.5194/acp-20-7829-2020, 2020.

695    Fueglistaler, S., and Silvers, L.G., 2021: The peculiar trajectory of global warming. Journal

696        of Geophysical Research: Atmospheres, 126, e2020JD033629.

697        https://doi.org/10.1029/2020JD033629.

698    Gillett, N. P., Zwiers, F. W., Weaver, A. J., Hegerl, G. C., Allen, M. R., and Stott, P. A.,

699        Detecting anthropogenic influence with a multi-model ensemble, Geophys. Res. Lett., 29(

700        20), 1970, doi:10.1029/2002GL015836, 2002.

701    Gregory, J. M., Andrews, T., Ceppi, P., Mauritsen, T., and M. J. Webb, 2020: How

702        accurately can the climate sensitivity to CO2 be estimated from historical climate

703        change?. Clim Dyn 54, 129–157. https://doi.org/10.1007/s00382-019-04991-y.

704    Gregory, J. M. & Andrews, T., 2016: Variation in climate sensitivity and feedback

705        parameters during the historical period. Geophys. Res. Lett. 43, 3911–3920.

706    Golaz, J.-C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., et

707        al. (2019). The DOE E3SM coupled model version 1: Overview and evaluation at

708        standard resolution. Journal of Advances in Modeling Earth Systems, 11, 2089– 2129.

709        https://doi.org/10.1029/2018MS001603.

710    Gulev, S.K., P.W. Thorne, J. Ahn, F.J. Dentener, C.M. Domingues, S. Gerland, D. Gong,

711        D.S. Kaufman, H.C. Nnamchi, J. Quaas, J.A. Rivera, S. Sathyendranath, S.L. Smith, B.

712        Trewin, K. von Schuckmann, and R.S. Vose, 2021: Changing State of the Climate

713        System. In Climate Change 2021: The Physical Science Basis. Contribution of Working

714        Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate

715        Change [MassonDelmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N.

716        Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R.

717        Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)].

718        Cambridge University Press. In Press.

719    Heede, U.K., Fedorov, A.V. Eastern equatorial Pacific warming delayed by aerosols and

720        thermostat response to CO2 increase. Nat. Clim. Chang. 11, 696–703 (2021).

721        https://doi.org/10.1038/s41558-021-01101-x.

722    Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J., Balaji, V., Duan, Q., Folini, D., Ji, D.,

723        Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., and Williamson,

724        D. (2017). The Art and Science of Climate Model Tuning, Bulletin of the American

725        Meteorological Society, 98(3), 589-602.

726        https://journals.ametsoc.org/view/journals/bams/98/3/bams-d-15-00135.1.xml

727    Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne,

728        M. J., Smith, T. M., Vose, R. S., and Zhang, H. (2017). Extended Reconstructed Sea

File generated with AMS Word template 2.0

729      Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and

730      Intercomparisons, Journal of Climate, 30(20), 8179-8205.

731      https://journals.ametsoc.org/view/journals/clim/30/20/jcli-d-16-0836.1.xml.

732    Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne,

733      M. J., Smith, T. M., Vose, R. S., and Zhang, H. (2017): NOAA Extended Reconstructed

734      Sea Surface Temperature (ERSST), Version 5 (2021-08-07). NOAA National Centers for

735      Environmental Information. doi:10.7289/V5T72FNM.  Accessed 2021-09-01.

736    Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A new sea surface

737      temperature and sea ice boundary dataset for the community atmosphere model. Journal

738      of Climate, 21(19), 5145–5153. https://doi.org/10.1175/2008JCLI2292.1

739    Jones, P. The reliability of global and hemispheric surface temperature records. Adv. Atmos.

740      Sci. 33, 269–282 (2016). https://doi.org/10.1007/s00376-015-5194-4

741    Jones, G. S. (2020). "Apples and Oranges": On comparing simulated historic near-surface

742      temperature changes with observations, Quarterly Journal of the Royal Meteorological

743      Society, 146, 733, 3747-3771. https://doi.org/10.1002/qj.3871.

744    Jones, G. S., Stott, P. A., and Christidis, N. (2013), Attribution of observed historical near–

745      surface temperature variations to anthropogenic and natural causes using CMIP5

746      simulations, J. Geophys. Res. Atmos., 118, 4001– 4024, doi:10.1002/jgrd.50239.

747    Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019).

748      Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and its

749      response to increasing CO2. Journal of Advances in Modeling Earth Systems, 11, 998–

750      1038. https://doi.org/10.1029/2018MS001400.

751    McKinnon, K. A., and Deser, C. (2018). Internal Variability and Regional Climate Trends in

752      an Observational Large Ensemble, *Journal of Climate*, *31*(17), 6783-6802.

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

754      409-440.

755    Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al.

756      (2021). An updated assessment of near-surface temperature change from 1850: the

757      HadCRUT5 data set. Journal of Geophysical Research: Atmospheres, 126,

758      e2019JD032361. https://doi.org/10.1029/2019JD032361.

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

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

766 Olonscheck, D., Rugenstein, M., & Marotzke, J. (2020). Broad consistency between observed
767     and simulated trends in sea surface temperature patterns. Geophysical Research Letters,
768     47, e2019GL086773. https://doi.org/10.1029/2019GL086773.

769 Orbe, C., Van Roekel, L., Adames, Á. F., Dezfuli, A., Fasullo, J., Gleckler, P. J., Lee, J., Li,
770     W., Nazarenko, L., Schmidt, G. A., Sperber, K. R., & Zhao, M. (2020). Representation of
771     Modes of Variability in Six U.S. Climate Models, *Journal of Climate*, *33*(17), 7591-
772     7617.Paulot, F., Paynter, D., Ginoux, P., Naik, V., and Horowitz, L. W.: Changes in the
773     aerosol direct radiative forcing from 2001 to 2015: observational constraints and regional
774     mechanisms, Atmos. Chem. Phys., 18, 13265–13281, https://doi.org/10.5194/acp-18-
775     13265-2018, 2018.

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

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

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

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

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

File generated with AMS Word template 2.0

789 Rieger, L. A., Cole, J. N. S., Fyfe, J. C., Po-Chedley, S., Cameron-Smith, P. J., Durack, P. J.,

790      Gillett, N. P., and Tang, Q.: Quantifying CanESM5 and EAMv1 sensitivities to Mt.

791      Pinatubo volcanic forcing for the CMIP6 historical experiment, Geosci. Model Dev., 13,

792      4831–4843, https://doi.org/10.5194/gmd-13-4831-2020, 2020.

793 Sanderson, B. M., Knutti, R., and Caldwell, P. (2015). A Representative Democracy to

794      Reduce Interdependency in a Multimodel Ensemble, Journal of Climate, 28(13), 5171-

795      5194. https://journals.ametsoc.org/view/journals/clim/28/13/jcli-d-14-00362.1.xml

796 Smith, C. J., Harris, G. R., Palmer, M. D., Bellouin, N., Collins, W., Myhre, G., et al. (2021).

797      Energy budget constraints on the time history of aerosol forcing and climate sensitivity.

798      Journal of Geophysical Research: Atmospheres, 126, e2020JD033622.

799      https://doi.org/10.1029/2020JD033622.

800 Taylor, K.E., D. Williamson and F. Zwiers, 2000: The sea surface temperature and sea ice

801      concentration boundary conditions for AMIP II simulations. PCMDI Report 60, Program

802      for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National

803      Laboratory, 25 pp. Available online: https://pcmdi.llnl.gov/report/pdf/60.pdf

804 Gupta, A. S., Jourdain, N. C., Brown, J. N., & Monselesan, D. (2013). Climate Drift in the

805      CMIP5 Models, Journal of Climate, 26(21), 8597-8615.

806      https://journals.ametsoc.org/view/journals/clim/26/21/jcli-d-12-00521.1.xml.

807 Taylor, K. E., D. Williamson, and F. Zwiers, 2000: The sea surface temperature and sea-ice

808      concentration boundary conditions of AMIP II simulations. PCMDI Rep. 60, 20 pp.

809 Wang, C., Soden, B. J., Yang, W., and Vecchi, G. A. (2021). Compensation between cloud

810      feedback and aerosol-cloud interaction in CMIP6 models. Geophysical Research Letters,

811      48, e2020GL091024. https://doi.org/10.1029/2020GL091024.

812 Watanabe, M., Dufresne, JL., Kosaka, Y. et al. Enhanced warming constrained by past trends

813      in equatorial Pacific sea surface temperature gradient. Nat. Clim. Chang. 11, 33–37

814      (2021). https://doi.org/10.1038/s41558-020-00933-3.

815 Wills, R. C. J., Dong, Y., Proistosecu, C., Armour, K. C., and Battisti, D. S. (2022).

816      Systematic climate model biases in the large-scale patterns of recent sea-surface

817      temperature and sea-level pressure change. Geophysical Research Letters, 49,

818      e2022GL100011. https://doi.org/10.1029/2022GL100011.

819 Zhang, J., Furtado, K., Turnock, S. T., Mulcahy, J. P., Wilcox, L. J., Booth, B. B., Sexton, D.,
820     Wu, T., Zhang, F., and Liu, Q.: The role of anthropogenic aerosols in the anomalous
821     cooling from 1960 to 1990 in the CMIP6 Earth system models, Atmos. Chem. Phys., 21,
822     18609–18627, https://doi.org/10.5194/acp-21-18609-2021, 2021.

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

825 Zinke, J., Browning, S.A., Hoell, A. et al. The West Pacific Gradient tracks ENSO and zonal
826     Pacific sea surface temperature gradient during the last Millennium. Sci Rep 11, 20395
827     (2021). https://doi.org/10.1038/s41598-021-99738-3.