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

A. Bodas-Salcedo[1], J. M. Gregory[1,2], D. M. H. Sexton[1], and C. P. Morice[2]

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.

METHODS

We decompose modelled and observational historical time series:

- $H_O(t) = S(t) + U_O(t) + E_O(t)$. $H_O(t)$: timeseries of the observed historical record anomalies; $S(t)$: forced signal; $U_O(t)$: unforced variability; $E_O(t)$: total observational error
- $H_M(t) = S(t) + D_M(t) + U_M(t)$. $D_M(t)$: discrepancy term or error in the forced response; $U_M(t)$: model's unforced variability.

We hypothesize that the model's forced response is realistic: $D_M(t)=0$, leading to our test for model ensemble means, denoted by overbars:

$$\overline{H_M}(t) - H_O(t) = \overline{U_M}(t) - U_O(t) - E_O(t) \qquad (1)$$

We define N(*T, y; RHS*), 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 construct empirical quantile functions of N(*T, y; RHS*) for a wide range of (*T, y*) pairs using *piControl* simulations (Figure 1). We test the *historical* simulations by comparing N(*T, y; LHS*) against the value of the 95th percentile of the empirical quantile function for (T, y). If N(*T, y; LHS*) > N(*T, y; RHS, p=0.95*) the model ensemble fails the test. A second test is carried out where the empirical quantile functions are calculated with a variance-scaled version of the RHS of Eq (1). Points in the (T, y) grid where the model fails the test (null hypothesis rejected) are shown with dots in Figures 2a and 2b.

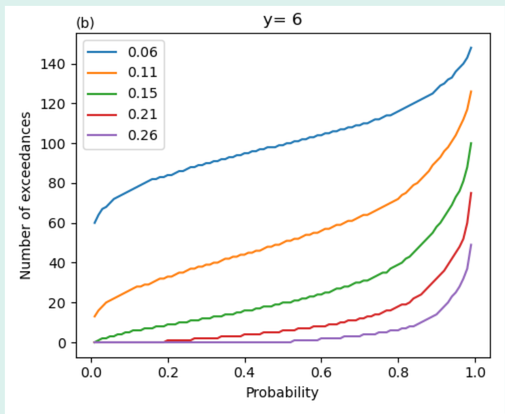


Figure 1. Examples of empirical quantile distribution functions of N(*T, y*). The averaging window used is 6 years as shown in the title. Each curve shows a different threshold *T*, as shown in the legend (in K). This example is for the global-mean index for a model ensemble with 3 members.

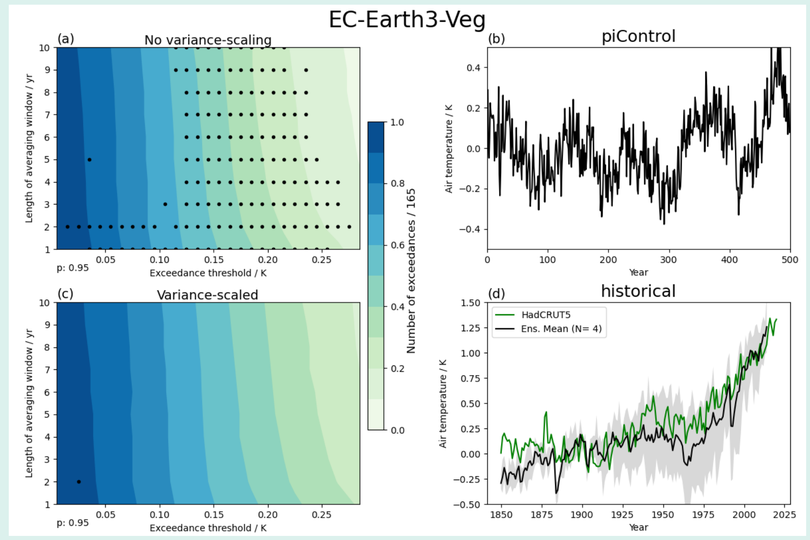
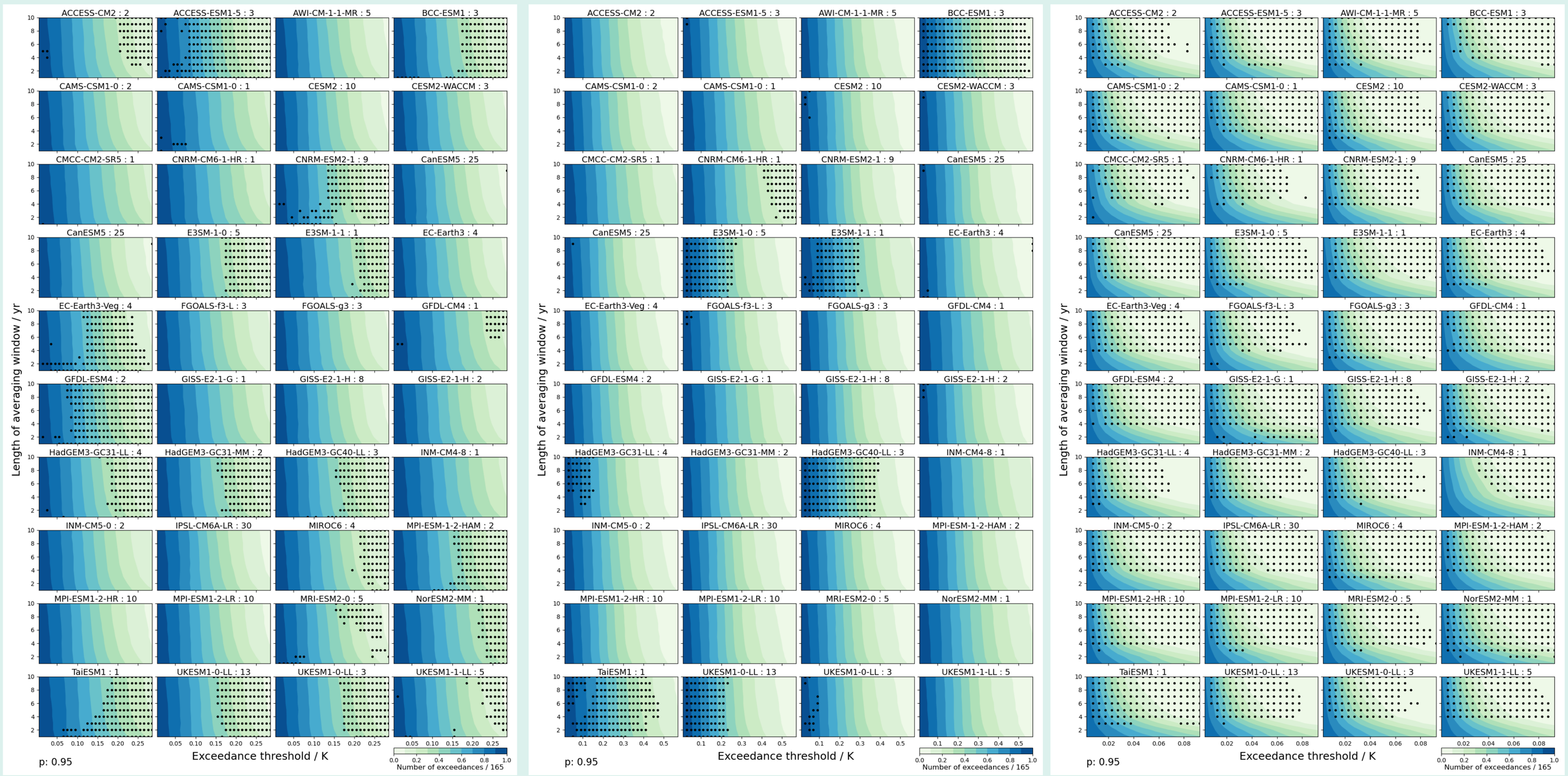


Figure 2. GMSAT tests for EC-Earth3-Veg. (a) and (b) show the results of the tests without and with variance-scaling, respectively (see text for details). The last 500 years of the *piControl* simulation of the model tested are shown in (b). Panel (d) shows the annual-mean historical anomalies of the 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.

RESULTS

Method applied to 40 *historical* ensembles (38 CMIP6 models plus HadGEM3-GC4.0 and UKESM1.1) for 3 indices: GMSAT, NH-SH difference and SST#.

GMSAT | **Hemispheric difference** | **SST#**



[1]**Met Office** Hadley Centre, FitzRoy Road, Exeter, Devon, EX1 3PB United Kingdom
Tel: +44 1392 885680 Fax: +44 1392 885681
Email: alejandro.bodas@metoffice.gov.uk

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

BIBLIOGRAPHY
Bodas-Salcedo., A., et al.: Assessment of large-scale indices of surface temperature during the historical period in the CMIP6 ensemble, J. Climate, submitted.

Fueglistaler, S., and Silvers, L.G.: The peculiar trajectory of global warming. Journal of Geophysical Research: Atmospheres, 126, 2021, e2020JD033629.