Supplementary Material

In Figure S1 we illustrate the dependence of the results on whether all CMIP3 models were used versus using only the models which passed two statistical tests of their ability to simulate tropical variability. The variability of modelled tropical sea surface temperatures (SST) compares well with NOAA/NASA (Advanced Very High Resolution Radiometer (AVHRR) Oceans Pathfinder SST data (Kilpatrick et al., 2001) over the time period 1989-2006. SST is used here rather than land surface temperature since tropical variability is largely dominated by the ocean and the associated ENSO phenomenon. The models that pass two statistical tests for their ability to simulate the tropical variability are identified here. The first test is described by Santer et al. (2009) and aims to estimate whether the interannual climate noise is realistically represented in the climate models. The second is the F-test which tests whether the natural variability of the models and the observations are similar in magnitude. Of the models which pass both tests, the first run available in the CMIP3 archive is used for the analysis. Figure S1a shows the variability of the models compared to the observations for the subset of models, and Figure S1b shows results for all models. Figures S1c and d compare the subset of models which passed the two statistical tests to the complete set of models (as in Figure 2 of the main text). The pattern of the results is nearly the same; only in a few countries slight differences are found. No clear criteria exist to test the quality of simulated ENSO variability. Different studies define different models as ‘best’(Leloup et al., 2008). Leloup et al. (2008) define four groups of model quality in terms of their capability to reproduce ENSO variability. When performing the same analysis as in Figure S1 using the models in the two best groups (Leloup et al., 2008) the outcome is the same as before. Furthermore, detection results are not influenced by model quality in case of water vapour (Santer et al., 2009). This further implies that systematic biases and different implemented forcings in the models used do not alter our findings.

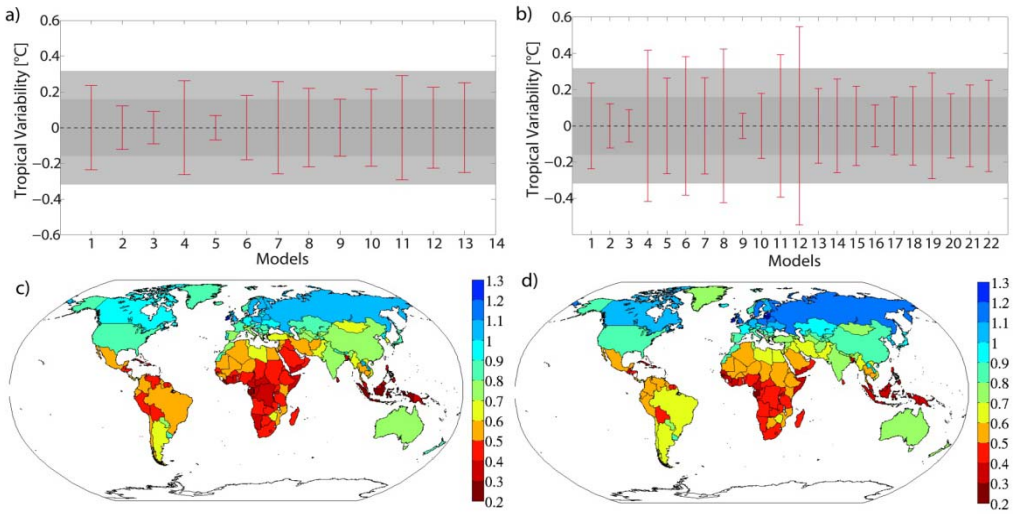

Figure S1 a) Model interannual variability (red) compared to observed variability (shaded in dark gray is the variability and in light gray twice the variability from the observed SST’s) for the models which passed both statistical

tests. b) the same as a), but for all models. c) the global temperature increase (°C) needed for a single location to undergo a statistically significant change in average summer seasonal surface temperature (TAS), aggregated on a country level (as in as Figure 2 in the main text) but for the subset of models only, d) as in panel c but for all models.

We next examine the dependence on the baseline period. When using the time period 1950-1979 as a baseline period (e.g. for comparison to an observational data set), the pattern of detected changes is similar. However, the year when perceptible changes can be detected is later. This implies that in most mid to high latitude regions, significant local warming would not be detected before about 2015 as shown in Figure S2 unless high quality century-scale records of observations are available.

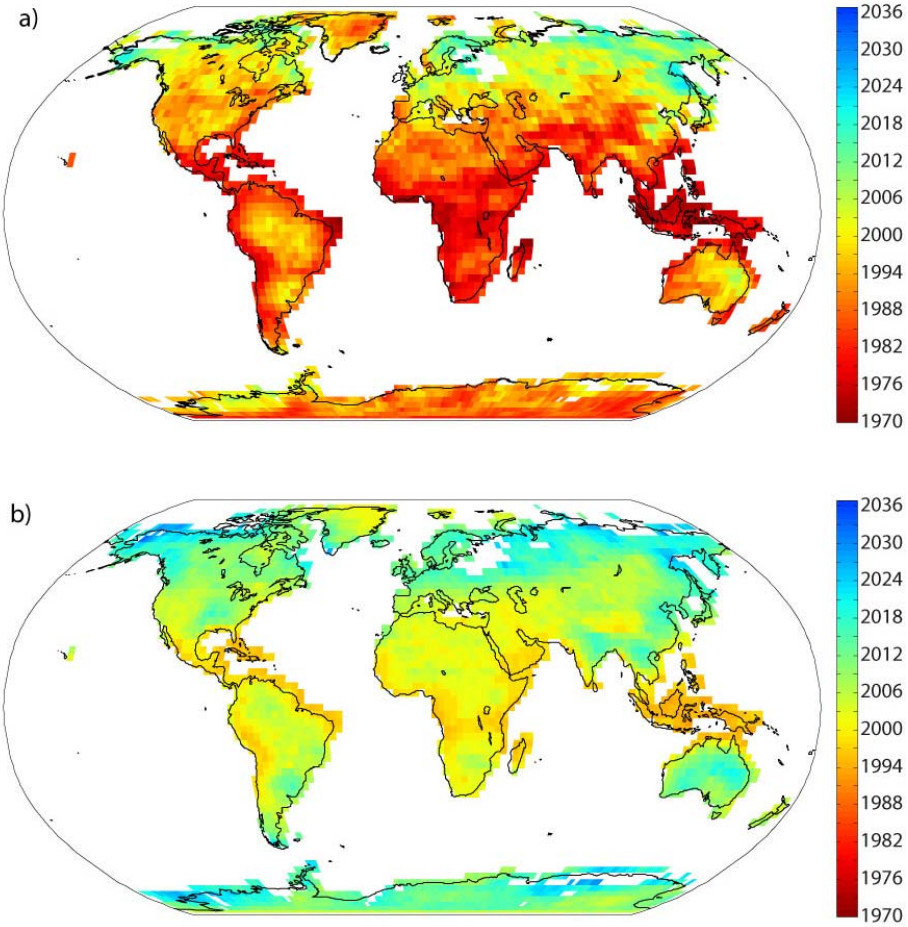

Figure.S2 a) Year of emergence on a grid scale level with 1900-1929 as a baseline period compared to using b) 1950-1979 as a baseline for surface summer temperature.

The method applied for the statistical test in order to calculate when the signal emerges from the noise is also probed here. In Figure S3, three different results are shown using

different statistical tests. Small differences can be found in the higher latitudes but for most low and mid-latitude regions, the results are not significantly altered.

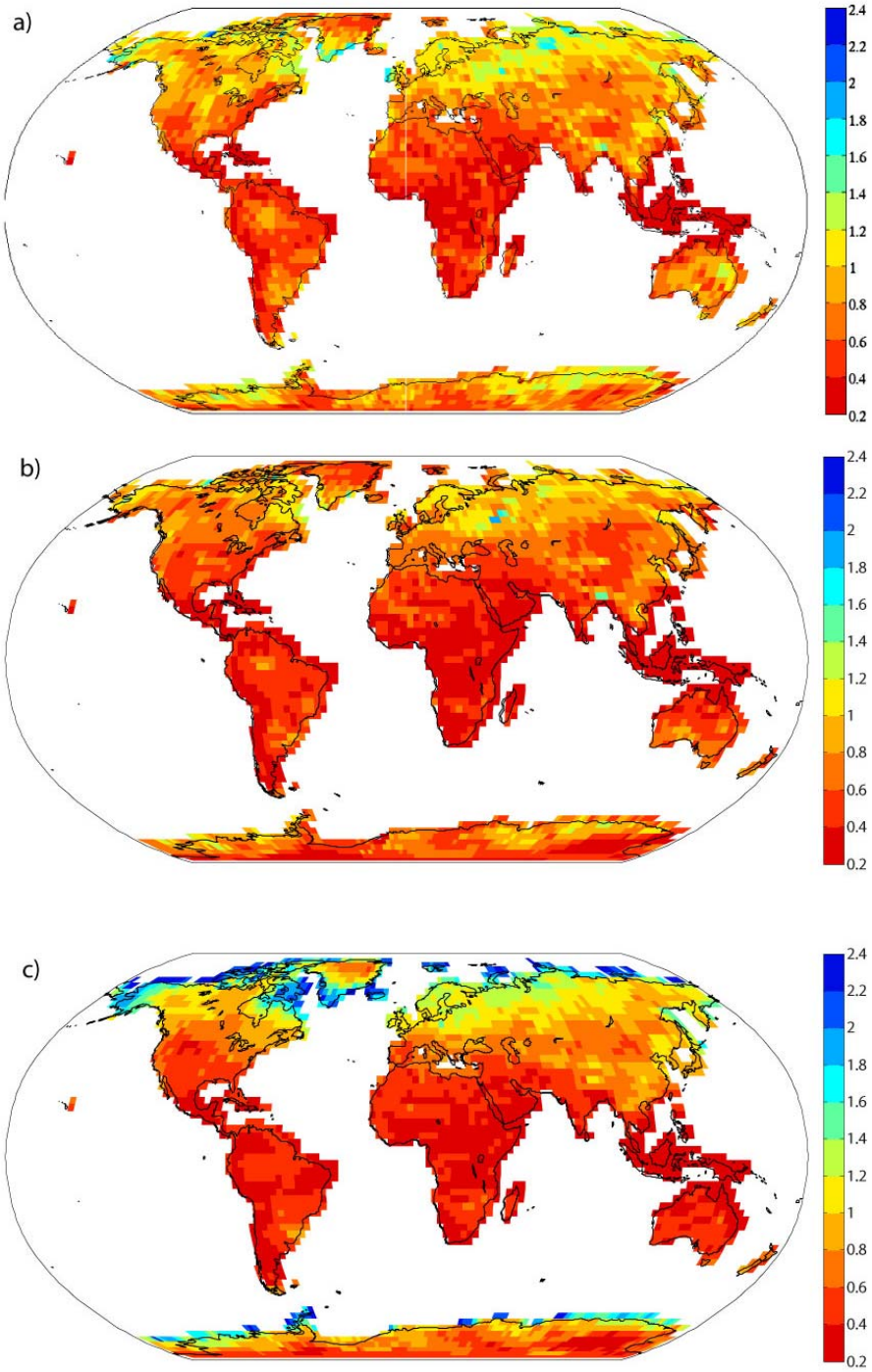

Figure S3 Results of three different statistical approaches to test how much global warming (°C) is needed for the local signal to emerge of the noise for the summer season. a) shows the results for the Kolmogorov-Smirnov test used in the main text, b) the students t-test and c) for the simple test.

Using annual mean instead of summer means leads to similar results in the tropics. However the high latitudes show an even lower warming to variability ratio as shown in Figure S4.

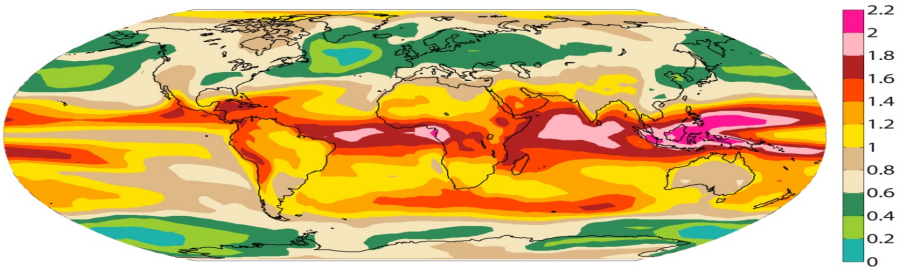

Figure S4 Warming to variability ratio of the annual mean trend to the annual mean variability.

We next raise the question how noise should be defined. One may argue that the 100-year trend from 1900-1999 should be compared to unforced trends on the same timescales using segments of the control model runs, which is a standard procedure in detection and attribution studies (Hegerl et al., 1997, Santer et al., 1995, Tett et al., 1999 and many more). This analysis was performed by fitting a linear trend from 1900-1999 to the transient runs and averaging them across models. The noise is estimated by performing the same trend fitting to non-overlapping segments of 100 years of the preindustrial control runs. The standard deviation of the sampling distribution of these unforced 100-year trends is the noise. Our analysis suggests that irrespective of how the variability is quantified, the resulting patterns are similar, as shown in Figure S5 . Nonetheless, there are some differences in the higher latitudes and the warming to variability ratio is higher in the tropics. Because only one realization of the observations exists, Figure 1 shows the results based on interannual variability which can directly be compared to the observed interannual variability. From an impacts perspective an important question is the date when we are leaving the range of interannual variations to which ecosystems have adapted, consistent with the analysis in Fig. 1. It is important to note that Figure 1 is only an illustration of the idea, but the main results (Fig. 2) are not dependent on Figure 1. The statistical test used to detect the emerging signal is simply testing whether a future time window is different from the base period and does not require an assumption of signal and noise. Furthermore, the analysis presented in Figure 2 is not affected by autocorrelation in summer temperatures. The autocorrelation is less than 0.1 in most areas and has no significant effect on the conclusions. Figure 2 is also robust to different rates in increasing global temperature across models. If some models warm at a slower rate than others, the global temperature increase needed for an equivalent signal to emerge will still be the same as for other models. However, if a model has a larger interannual variability, a greater global temperature increase would

be needed for that model compared to others. But most models simulate the observed low latitude variability reasonably well. However, the results shown in Figure 2 are sensitive to the length of the moving window. For example the increase in global temperature needed for a significant change is 0.1-0.2 °C greater when using a 50-years window. Yet the pattern remains unchanged and the conclusions are not altered in a significant way. A 30-years moving window was chosen as 30 years are long enough for short-term fluctuations to become irrelevant but yet it they are long enough to echo long term trend (http://www.aos.wisc.edu/~sco/normals.html).

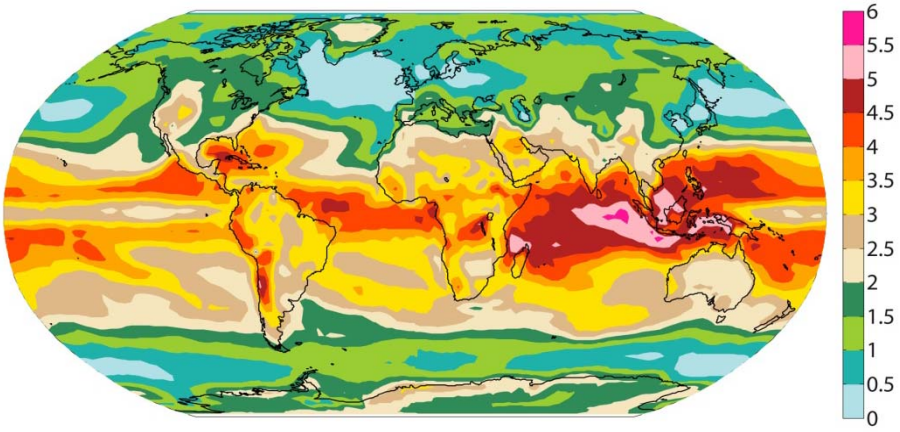

Figure S5 Warming to variability ratio of the summer mean trend (1900-1999) to the standard deviation of summer trends in 100 year segments in the control runs.

Concerning the observational data and its uncertainty, GISTEMP is compared with GHCN-gridded data (Peterson and Vose, 1997). Figure S6 shows the same as Figure 1 in the main text, but here the two observational datasets are shown. Please note that the signal is computed as a linear trend, rather than a temperature difference. The two datasets are not inconsistent although especially in the tropical region there is a considerable lack of data in case of GHCN which makes it difficult to judge. Furthermore, the results of GISTEMP for the period 1900-1999 show many similar features as for the period 1950-1999 which justifies the use of the longer time period for the model comparison.

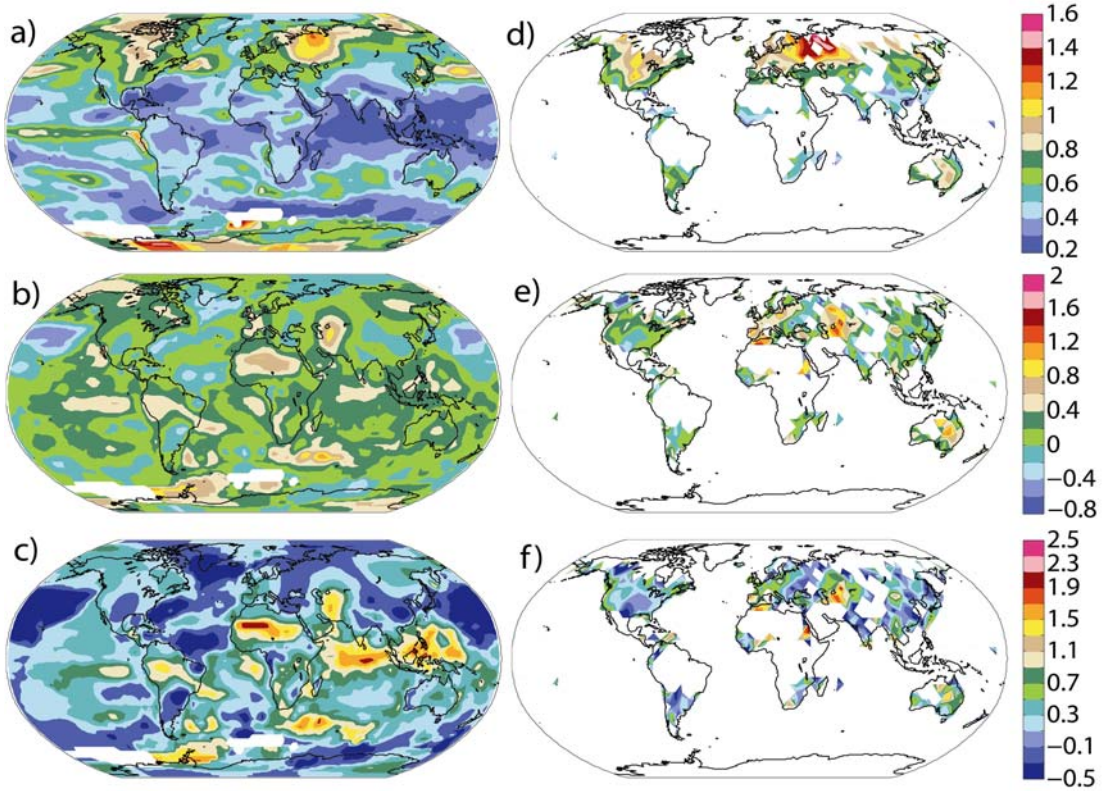

Figure S6 Comparison between GISTEMP (left) and GHCN-gridded (right) data for a/d) variability [°C], b/e) trend [°C] and c/f) warming to variability ratio during summer season for the period 1950-1999.

HEGERL, G. C., HASSELMANN, K., CUBASCH, U., MITCHELL, J. F. B., ROECKNER, E., VOSS, R. & WASZKEWITZ, J. (1997) Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dynamics,* 13**,** 613-634.

KILPATRICK, K. A., PODESTA, G. P. & EVANS, R. (2001) Overview of the NOAA/NASA advanced very high resolution radiometer Pathfinder algorithm for sea surface temperature and associated matchup database. *Journal of Geophysical Research-Oceans,* 106**,** 9179-9197.

LELOUP, J., LENGAIGNE, M. & BOULANGER, J. P. (2008) Twentieth century ENSO characteristics in the IPCC database. *Climate Dynamics,* 30**,** 277-291.

PETERSON, T. C. & VOSE, R. S. (1997) An overview of the global historical climatology network temperature database. *Bulletin of the American Meteorological Society,* 78**,** 2837-2849.

SANTER, B. D., TAYLOR, K. E., GLECKLER, P. J., BONFILS, C., BARNETT, T. P., PIERCE, D. W., WIGLEY, T. M. L., MEARS, C., WENTZ, F. J., BRUGGEMANN, W., GILLETT, N. P., KLEIN, S. A., SOLOMON, S., STOTT, P. A. & WEHNER, M. F. (2009) Incorporating model quality information in climate change detection and attribution studies.

Proceedings of the National Academy of Sciences of the United States of America, 106**,** 14778-14783.

SANTER, B. D., TAYLOR, K. E., WIGLEY, T. M. L., PENNER, J. E., JONES, P. D. & CUBASCH, U. (1995) Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dynamics,* 12**,** 77-100.

TETT, S. F. B., STOTT, P. A., ALLEN, M. R., INGRAM, W. J. & MITCHELL, J. F. B. (1999) Causes of twentieth-century temperature change near the Earth's surface. *Nature,* 399**,** 569-572.