

Journal of Climate

Multi-Model Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 20th Century Simulations --Manuscript Draft--

Manuscript Number:	JCLI-D-12-00567
Full Title:	Multi-Model Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 20th Century Simulations
Article Type:	Article
Corresponding Author:	Thomas Knutson Geophysical Fluid Dynamics Lab/NOAA Princeton, New Jersey UNITED STATES
Corresponding Author's Institution:	Geophysical Fluid Dynamics Lab/NOAA
First Author:	Thomas Knutson
Order of Authors:	Thomas Knutson Fanrong Zeng Andrew Wittenberg
Abstract:	<p>Regional surface temperature trends from the CMIP3 and CMIP5 20th century runs are compared with observations -- at spatial scales ranging from global averages to individual grid points -- using simulated intrinsic climate variability from pre-industrial control runs to assess whether observed trends are detectable and/or consistent with the models' historical run trends. The CMIP5 models are also used to detect anthropogenic components of the observed trends, by assessing alternative hypotheses based on scenarios driven with either anthropogenic plus natural forcings combined, or with natural forcings only. Modeled variability is assessed via inspection of control run time series, standard deviation maps, spectral analyses, and low-frequency variance consistency tests. The models are found to provide plausible representations of internal climate variability, though there is room for improvement. The influence of observational uncertainty on the trends is assessed, and found to be generally small compared to intrinsic climate variability.</p> <p>Observed temperature trends over 1901-2010 are found to contain detectable anthropogenic warming components over a large fraction (about 80%) of the analyzed global area. In several regions, the observed warming is significantly underestimated by the models, including parts of the southern Ocean, south Atlantic, far eastern Atlantic, and far west Pacific. Regions without detectable warming signals include the high latitude North Atlantic, the eastern U.S., and parts of the eastern Pacific. For 1981-2010, the observed warming trends over about 45% of the globe are found to contain a detectable anthropogenic warming; this includes much of the globe within about 40-45 degrees of the equator, except for the eastern Pacific.</p>

Responses to Reviewer Comments

Last modified: Mar. 1, 2013

Ref.: JCLI-D-12-00567

"Multi-Model Assessment of Regional Surface Temperature Trends: CMIP3 and CMIP5 20th Century Simulations" by Thomas Knutson; Fanrong Zeng; Andrew Wittenberg

Note to reviewers:

Although this was not requested by the reviewers, since the last submission we have implemented an improved way of comparing the low-frequency variance between model control runs and the observed record. This new approach allows us to then display and compare the models' control run low-frequency variance (standard deviation) with the (adjusted) observed estimate with less leakage of forced variance than was present previously. Here is how we have done this: At each gridpoint, we low-pass filter the observations using a decadal filter with half power point at 9 years. Rather than compare this variance directly to that of a model control run, we first attempt to estimate how much of an amplification of variance there is in the observed estimate owing to the presence of forced variability (in addition to internal, unforced variability). We correct or adjust for this amplification in two stages. For CMIP5 (or CMIP3, analyzed separately), we used the grand ensemble mean of the seven model All-Forcing runs as the estimate of the forced signal to remove from observations. This provides the first level of adjusted observations. But we know that due to errors in simulating the forced response and the limited number of ensemble members that are used to estimate the forced response, there is some forced variance that remains in the adjusted observed series. We try to estimate how much remains by using this same general procedure to filter the forced response from the model all-forcing runs and then calculate the residual variance that remains by direct comparison to the control run for the same model. We do this for each model in turn and generate and average "second-level" adjustment. This average adjustment is then applied to the standard deviation from the first-level adjusted observations to obtain a new observed internal variability standard deviation estimate that is more comparable with the model control runs. Given this new approach (derived with two separate levels of adjustment) we can now more defensibly compare the model control run and observed low-frequency variability. We stress that our procedure remains a very rough test of decadal variance consistency, especially in data poor regions such as the southern ocean. There will be inherent limits to what we can do because there is only so much observational data and only so many ensemble members supplied by the modeling centers. In terms of observational temporal coverage, in order for a comparison to be done between model and observations at a gridpoint, we require at least 50 points (out of 110) to be available in the 110-year annually resolved decadal filtered record. Forty percent temporal coverage is required for an annual mean to be considered valid, and the decadal filter does a modest degree of gap-infilling by computing in a seven-year wide sliding window, a filtered value if at least four of seven annual values are available.

The adjusted standard deviation of low-pass filtered observations (which we call "Obs. St. Dev.*") is the basis of the difference maps and spatial correlations in Figs. 2 and 3, and for the

revised variance consistency statistics quoted in the text (global mean), in the maps (Figs. 14-16) and in the summary time series (Fig. 17).

We have added text to explain this new procedure and accompanying in Section 3b..

Reviewer Comments: (Responses are in red.)

Reviewer #1:

The ms is much more readable, and the authors have done a remarkable job extending the ms and improving it - congratulations on an excellent job, I am happy with it, just a few questions / suggestions for the authors to address as they see fit:

On the low number of models that both have detectable changes and realistic variability: it would be important to clarify what faults there are in variability - is the variability too high or too low? While the first queries model quality but makes d+a results conservative, the latter is more problematic. if the authors can see a way to make this more clear, or ideally be specific on if the var is too high or low, would be great. If not at least flag please.

For the global mean variance comparisons, we now also state the number of models that are significantly too high and too low. For the maps, this is more difficult to convey without adding new panels to the figures. However, we've made some changes to at least partially address this issue, without adding new panels to the fraction of model plots. We've decided to change the format of Fig. 2 and 3 so that these now show differences in low frequency standard deviation between model control runs and observations (where an adjustment correction has been applied to the observed field to make it more comparable to control run variability in the models as explained above). These maps allow the reader to see, for a general region, which models have low-frequency variability that is less than or greater than observed (although doesn't indicate whether the difference is significant or not). The multi-model average standard deviation plot shows the difference vs. observations for the average low-frequency standard deviation of the models.

abstract: 'weighted against' is not quite the right way to phrase it - maybe say 'and reject an alternative explanation based on scenarios driven with natural forcings only'

We have adopted this change.

line 100/101 twice section 7, bit awkward.

Second occurrence should have been "8". This is now corrected.

l 332: describing a trend measured over just a few decades in K/100yrs is a bit misleading - maybe change notation to K/decade? thats much more common pracitce -again a few lines down.

We've made this change.

l 333: at what percentile of the multimodel range is a trend starting in the 1990s? or one starting in 2000? (just a curious question if easy to answr might be useful to have)

We have not yet computed this yet, but we plan to insert a sentence with this calculation at the galley proof stage, as it is not critical to the paper at this point.

l 337/339 I cant quite understand this - so the ensemble mean of each model is averaged to a multimodel mean? And then the noise is added on that - for a reasonably small ensemble, this would make the noise quite large as there will be still quite a bit in the ensemble mean (eg ensemble of 4 would give you total standard dev 10% too large). This is not a big deal, but if you have done it like that without correcting the variance down, would be good to caveat, if you have corrected (eg following von Storch and Zwiers) worth mentioning. Based on this noise magnitude without correcting you might find too many models to have too large variability!

First, we need to clarify how the pink and blue envelopes on Figs. 7, 8 are constructed. These envelopes for a given trend start date, are 5th to 95th percentile range of trends about a central value which is the grand ensemble mean trend of the All Forcing or Natural Forcing models. The ensemble mean of each model is averaged to obtain the grand multimodel mean as the reviewer notes. However, the spread is computed in a different way such that it includes uncertainty not only due to internal variability, but also uncertainty due to the different model responses to forcing. To compute the spread, we build a distribution of trends as follows. For example, for the case of All-Forcing runs, where we have seven models, we sample each of the seven models equally often in building up a large distribution of trends. When we sample a model, this means we combine that model's ensemble mean with some randomly sampled trend from that model's control run. Each time we do this we create one sample, and the process is repeated a large number of times (50 times for each model included, or 350 times for the seven model analysis), sampling from each of the seven models equally often. The "grand distribution" built up in this way then has spread due to the differences in model mean trends among the seven models and due to the control run variability.

Turning to the second part of the reviewer's question, we agree that the limited number of ensemble members for the individual models means that there is additional variance in the grand distribution due to our imperfect knowledge of each model's forced response. However, the net impact of this on the spread of the total distribution is a complicated function of several factors. These include the following four factors: 1) the number of ensemble members a particular model has (which we now show in Fig. 1; the larger the number of ensemble members the

smaller the overestimate of variance; 2) where the models with few ensemble members sit in the distribution (if they are close to the outer edge, the overestimate can be greater than if they are near the middle of the distribution); 3) what is the variance of the model with few ensemble members or that sits at the outer edge of the distribution; and 4) what is the relative size of the spread of the individual model ensemble responses vs. the internal variability of the models near the outer edge of the distribution.

We can also estimate an upper limit on the overestimate of the standard deviation, based on the number of ensemble members we use, as about 15-40% at most, with the worst case being for a single ensemble member, where the variance is as much as doubled, so the standard deviation is 40% overestimated. However, given the four factors mentioned above, the effect will typically be considerably smaller than this.

It is also worth noting that the effect of an overestimation of variance in our framework is to make trends too difficult to detect (compared to internal variability or to the internal variability plus natural forcing), but to also make it too easy for All-Forcing trends to be consistent with observations.

We could in principle attempt a simulation to essentially estimate confidence intervals on our confidence intervals, but these would be situation dependent and would vary for different locations around the globe, time period, etc. We have chosen to leave this for further studies and add a caveat summarizing the above discussion in the conclusion section along with other caveats of our analysis.

l 383 typo 'for'

Corrected.

l 398: the 'most' assessment is also based on using spatial patterns to distinguish between forcings and estimate ghg alone, maybe rephrase slightly

We have rephrased the second part of the paragraph to address this issue and compare a little further the nature of our approach compared with other complementary approaches.

l 507: not sure Hegerl 2007 is the best quote for this - Portman et al. 2009 might be more useful here Portmann R. W., S. Solomon and G.C. Hegerl (2009): Linkages between climate change, extreme temperature and precipitation across the United States . PNAS, 2009, www.pnas.org/cgi/doi/10.1073/pnas.0808533106 (no strong view just a suggestion)

This is a good suggestion and we've added this reference, which we had intended to add earlier but had forgotten to do so.

l 588: All these explanations are possible and useful, but its also worth noting that among many at least partly independent regions you would expect some to be high or low just by chance (although I doubt thats the case here - but worth listing) - same in line 984 and probably most important to mention there.

We have adopted these suggestions.

1748: might be worth mentioning that the hiatus in the mid20th was preceded by strong early 20th century warming

We have adopted this suggestion.

1877 and 881 see general comment above - are they low or high?

We now refer the reader back to revised Figs. 2 and 3, which allow one to assess whether models tend to be too high or too low in terms of simulated variability.

1917: isn't this at least partly because there is less detectable in the first place for the short interval?

Yes, so we have moved the comment on the variance test up to the end of the preceding paragraph so that it is not confused with the issue of relatively limited regions with detectable trends over the short period (1991-2010).

Additional Material for Reviewer Reference

[Click here to download Additional Material for Reviewer Reference: Knutson_JCLIM_regional_rev_Feb2013_w_track_changes.pdf](#)

1 Multi-Model Assessment of Regional Surface Temperature Trends:
2 CMIP3 and CMIP5 20th Century Simulations

3

4

5 Thomas R. Knutson, Fanrong Zeng, and Andrew T. Wittenberg

6

7 Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ 08542

8

9 Submitted to *J. of Climate*

10

11 Version: Revised March 1, 2013

12

13 Email contact: Tom.Knutson@noaa.gov

14

15

16

17

18 **Abstract.**

19

20 Regional surface temperature trends from the CMIP3 and CMIP5 20th century runs are compared
21 with observations -- at spatial scales ranging from global averages to individual grid points --
22 using simulated intrinsic climate variability from pre-industrial control runs to assess whether
23 observed trends are detectable and/or consistent with the models' historical run trends. The
24 CMIP5 models are also used to detect anthropogenic components of the observed trends, by
25 assessing alternative hypotheses based on scenarios driven with either anthropogenic plus natural
26 forcings combined, or with natural forcings only. Modeled variability is assessed via inspection
27 of control run time series, standard deviation maps, spectral analyses, and low-frequency
28 variance consistency tests. The models are found to provide plausible representations of internal
29 climate variability, though there is room for improvement. The influence of observational
30 uncertainty on the trends is assessed, and found to be generally small compared to intrinsic
31 climate variability.

32 Observed temperature trends over 1901-2010 are found to contain detectable anthropogenic
33 warming components over a large fraction (about 80%) of the analyzed global area. In several
34 regions, the observed warming is significantly underestimated by the models, including parts of
35 the southern Ocean, south Atlantic, far eastern Atlantic, and far west Pacific. Regions without
36 detectable warming signals include the high latitude North Atlantic, the eastern U.S., and parts of
37 the eastern Pacific. For 1981-2010, the observed warming trends over about 45% of the globe
38 are found to contain a detectable anthropogenic warming; this includes much of the globe within
39 about 40-45 degrees of the equator, except for the eastern Pacific.

40

41 **1. Introduction**

42 Are historical simulations of surface temperature trends, obtained using climate models with the
43 best available estimates of past climate forcings, consistent with observations? Where on the
44 globe can observed temperature trends be attributed to anthropogenic forcing? These questions
45 can be examined using a substantial number of different climate models and using different
46 analysis methods. Here we attempt to incorporate information from a relatively large sample of
47 climate models, from the Coupled Model Intercomparison Project 3 (CMIP3; Meehl et al. 2007)
48 and CMIP5 (Taylor et al. 2012), using various multi-model combination techniques. The general
49 approach is to compare the modeled and observed trends, in terms of both magnitude and
50 pattern, by considering trends at each grid point in the observational grid, as well as trends over
51 broader-scale regions.

52 The term “*detectable climate trend*” used here refers to a trend in the observations that is
53 inconsistent with (i.e., outside of the 5th to 95th percentile range of) simulated trends, either from
54 control runs (the internal or intrinsic climate variability background) or from a sample of natural-
55 forcing response and control run variability combined (the natural climate variability
56 background). (Control runs are long runs with pre-industrial forcings that do not change from
57 year to year.) We interpret a trend in observations as “*attributable (at least in part) to*
58 *anthropogenic forcing*” if it is both inconsistent with simulated natural climate variability
59 (detectable) and consistent with the All-Forcing runs that contain both anthropogenic forcing
60 agents (e.g., changes in greenhouse gases and aerosols) and natural forcings (e.g., changes in
61 solar insolation or volcanic aerosol loading). If an observed trend is detectable but inconsistent

62 with All-Forcing runs because it is larger than the simulated distribution of trends, we still
63 interpret the observed trend as attributable, at least in part, to anthropogenic forcing. While a
64 number of CMIP5 models have Natural-Forcing-Only runs available on-line, for the CMIP3
65 models, relatively few such runs are available. Therefore, for CMIP3, we adopt a simpler
66 approach of assessing whether observed trends are consistent with All-Forcing runs, but
67 inconsistent with internal variability alone. The simpler approach does not allow us to draw
68 conclusions about whether an observed trend is attributable to anthropogenic forcing or not.

69 The modeled internal climate variability from long control runs is used to determine whether
70 observed and simulated trends are consistent or inconsistent. In other words, we assess whether
71 observed and simulated forced trends are more extreme than those that might be expected from
72 random sampling of internal climate variability. This approach has been applied to earlier
73 models in a number of studies, beginning with the analyses of Stouffer et al. (1994; 2000).
74 Similarly, we use the available ensemble of simulated forced trends to assess whether observed
75 trends are compatible with the forcing-and-response hypotheses embodied by those forced
76 simulations.

77 Formal detection/attribution techniques often use a model-generated pattern from a single or set
78 of climate forcing experiments, and then regress this pattern against the observations to compute
79 a scaling amplitude (e.g., Hegerl et al. 1996; Hasselmann 1997; Allen and Tett 1999; Allen and
80 Stott 2003) . If the scaling is significantly different from zero, the forced signal is detected. If
81 the scaling does not significantly differ from unity, then the amplitude of the signal agrees with
82 observations, or is at least close enough to agree within an expected range based on internal
83 climate variability. Optimal detection techniques also filter the data during the analysis such that
84 the chance of detecting a specified signal, or “fingerprint”, is enhanced if the signal is present in

85 the data. An alternative approach that is less focused on model-defined patterns has been
86 proposed by Schneider and Held (2001). In contrast to the optimal detection/attribution
87 methods, we compare both the amplitude and pattern simulated by the models directly with the
88 observations, without rescaling of patterns or application of optimization filtering. Our analysis
89 is thus a consistency test for both the amplitude and pattern of the observed versus simulated
90 trends, building on earlier work along these lines by Knutson et al. 1999; Karoly and Wu 2005;
91 Knutson et al. 2006; and Wu and Karoly 2007 to test for detectable anthropogenic contributions.
92 Other variants and enhancements to this general type of analysis have recently been presented by
93 Sakaguchi et al. (2012). More discussion of various detection and attribution methods and their
94 use in general is contained in Hegerl et al. 2009.

95

96 In this report, the models, methods, and observed data are described in Section 2. We examine
97 the model control runs and their variability in Section 3. Global-mean time series from the
98 20C3M (approximately 1860-2010) historical runs are examined in Section 4. Section 5 contains
99 consistency tests for observed vs simulated trends, as discussed above, for temperatures averaged
100 over various defined regions of the globe. Maps based on results of consistency tests at the grid
101 point scale are presented in Section 6. A brief description of online supplemental material is
102 given in Section 7, and the discussion and conclusions are given in Section 8.

103

104 **2. Model and Observed Data Sources**

105

106 *a. Observed data*

107

108 The observed surface temperature dataset used in this study is the HadCRUT4 (Morice et al.
109 2012) which is available as a set of anomalies relative to the period 1961-1990. The dataset
110 contains some notable revisions, particularly to SSTs (HadSST3; Kennedy et al. 2011), relative
111 to previous versions, so it important to retest earlier conclusions regarding climate trends using
112 the revised data. The dataset also contains uncertainty information, in the form of 100 ensemble
113 members sampling the estimated observational uncertainty. Some of our tests examine the
114 sensitivity of trend results to this observational uncertainty.

115

116 To form a combined product of SST and land surface air temperature, Morice et al. (2012) adopt
117 the following procedure. If both land data and SST data are available in a particular grid box,
118 they are weighted according to the fraction of the grid box that is covered by land or ocean,
119 respectively. A minimum of 25% coverage is assumed, even if the fraction of the grid box
120 covered by land is less than 25%. In our study, we use this same general procedure, adapted to a
121 model's land-sea mask, to combine SST and land surface air temperature data sets from each
122 model that we analyze.

123

124 *b. CMIP3 and CMIP5 models*

125

126 Figure 1 displays the complete collection of control runs from both CMIP3 and CMIP5 used in
127 our analysis. The data were downloaded from the CMIP3 ([www-](http://www-pcmdi.gov/ipcc/about_ipcc.php)
128 pcmdi.gov/ipcc/about_ipcc.php) and CMIP5 (cmip-pcmdi.llnl.gov/cmip5) model archives. We
129 regridded (averaged) the model data from the 20C3M historical runs and control runs onto the
130 observational grid. In cases where we needed to use combined model land surface air
131 temperature and SST data to compare with observations, we used a procedure resembling that
132 used for the observations, but based on the model's own land-sea mask. For example, if any land
133 is present in a grid box, a minimum of 25% land coverage is assumed, even if the fraction of the
134 grid box covered by land is less than 25%. Our general approach in this study is to attempt to
135 mimic observations with the models, in terms of data coverage over time. To mimic the space-
136 time history of data gaps in the observations, we masked out (withheld from the analysis) model
137 data at times and locations where data were labeled missing in the observations. Finally, we
138 computed the model's climatology over the same years as for observations (1961-1990) and then
139 created anomalies from this climatology. For example, this same procedure was used for 150-yr
140 samples from the model control runs for analyses where we wanted to ensure that the control
141 runs had missing data characteristics that were similar to those of the observed data.

142

143 The historical forcings for the CMIP3 20C3M historical forcing runs are summarized in Rind et
144 al. (2009; Table 3.6). An important distinction among the models is the treatment of volcanic
145 forcing. Ten of the 24 CMIP3 models we examined include volcanic forcing, while 14 do not.
146 However, as discussed further below, for most of our assessments, we used a maximum of 19 of
147 the 24 CMIP3 models of which eight included volcanic forcing while 11 models (identified by
148 “*” after model name in Fig. 1 a,b) did not. We refer to these sets of models as the eight

149 “Volcanic” and 11 “Non-Volcanic” CMIP3 model subsets, respectively. All 23 of the CMIP5
150 models included in this study included volcanic forcing in their 20C3M runs. However, only
151 seven of the 23 CMIP5 models had Natural-Forcing-Only runs that extended to 2010 (see Fig.
152 1). These Natural-Forcing runs extending to 2010 were necessary for some of our detection and
153 attribution analyses concerning anthropogenic forcing, and those seven models form the CMIP5
154 seven-model subset referred to in subsequent section.

155

156 **3. Model Control Run Analysis**

157 *a. Global mean time series*

158 The global-mean surface air temperature series from the CMIP3 and CMIP5 model control runs
159 are shown in Fig. 1. Data are displayed with arbitrary vertical offsets for visual clarity. The
160 figure also shows the observed surface temperature anomalies from HadCRUT4. The curves
161 labeled “Residual” were obtained by subtracting the multi-model mean of the historical volcanic
162 forcing runs (either CMIP3 or CMIP5) from the full observed time series. These observed
163 residual series thus contain estimates of the internal variability of the climate system as derived
164 from the observations in combination with the climate models’ response to estimated historical
165 forcing. In section 3b we will further refine this estimate of observed internal variability.

166

167 The model control runs exhibit long-term drifts. The magnitudes of these drifts tend to be larger
168 in the CMIP3 control runs (Fig. 1a,b) than in the CMIP5 control runs (Fig. 1 c,d), although there
169 are exceptions. We assume that these drifts are due to the models not being in equilibrium with

170 the control run forcing, and we remove the drifts by a linear trend analysis (depicted by the
171 orange straight lines in Fig. 1). In some CMIP3 cases the drift initially proceeds at one rate, but
172 then the trend becomes smaller for the remainder of the run. We approximate the drift in these
173 cases by two separate linear trend segments, which are identified in the figure by the short
174 vertical orange line segments. These long-term drift trends are removed to produce the “drift-
175 corrected” series. The procedure for removing the trends involves calculating and removing the
176 linear trends (over the time periods shown in Fig. 1) at each model grid point separately. The
177 orange trend lines shown in Fig. 1 depict also the starting and ending years for the trends used
178 for each model.

179 Five of the 24 CMIP3 models, identified by “(-)” in Fig. 1, were not used, or practically not
180 used, beyond Fig. 1 in our analysis. For instance, the IAP_fgoals1.0.g model has a strong
181 discontinuity near year 200 of the control run. We judge this as likely an artifact due to some
182 problem with the model simulation, and we therefore chose to exclude this model from further
183 analysis. The Miroc_3.2_hires and INGV_echam4 model control runs are so short in length that
184 they are essentially unused in our analysis, since we require the control run record to be at least
185 three times as long as a trend that is being assessed. For two other models, we were not able to
186 successfully obtain sea surface temperature information from the CMIP3 archive, and so these
187 were excluded from further analysis.

188

189 While some of the trends in the CMIP3 and CMIP5 control runs (Fig. 1) approach the observed
190 ~150 yr trend in terms of general magnitude, these few cases are associated with either the long-
191 term drifts discussed above or with a few spurious discontinuity issues (e.g., IAP_fgoals1.0.g).

192 Controlling for these apparent problems, none of the control runs in the CMIP3 or CMIP5
193 samples exhibit a centennial scale trend as large as the trend in the observations. On the other
194 hand, the variability of observed residual series appears roughly similar in scale to that from
195 several of the control runs. Three of the CMIP3 control runs illustrated in Fig. 1 (GISS_aom,
196 GISS_model_e_h, and GISS_model_e_f) have much lower levels of global surface temperature
197 variability than in the observed residual series. For some sensitivity tests on the multi-model
198 assessments, we have excluded these three models to test for robustness.

199 *b. Geographical distribution of variability*

200

201 In this section, we describe a method for comparing the geographical distributions of observed
202 variability with model control run variability. The geographical distribution of an adjusted
203 standard deviation of low-pass-filtered (> 10 yr) surface temperature from observations (Obs.
204 St. Dev.*) is shown in Fig. 2 (middle column: b, e, h). These observed estimates contain
205 adjustments (described in detail below) that make them more suitable for comparison to the
206 variability in the model control run. This is necessary because the variability within the model
207 control runs is generated strictly internally within the models and does not contain contributions
208 from external climate forcings. In contrast, observed temperature will contain some mixture of
209 variability due to external climate forcing agents and internally generated processes in the
210 climate system. The models' average standard deviation fields, based on the full available time
211 series of surface air temperature from each control run, are shown in the left column (a, d, g).
212 Prior to computing the individual model standard deviations, the long-term drift has been
213 subtracted from each control run as discussed in Section 3a. The individual model standard
214 deviations are then averaged for the three model sets to form the fields in (a, d, g). Difference

215 maps, computed as the models' average low-frequency standard deviations minus Obs. St. Dev*,
216 are shown in the right column (c, f, i) of Fig. 2.

217 We now describe the process for computing the adjusted observed low-pass filtered standard
218 deviation (Obs. St. Dev.*; Fig. 2 b, e, h). At each gridpoint, we low-pass filter the observations
219 using a decadal filter with a half-power point at nine years. Rather than compare this variance
220 directly to variance from a model control run, we first attempt to estimate how much of an
221 amplification of variance there is in the observed estimate owing to the presence of forced
222 variability (in addition to internal, unforced variability). We then correct or adjust for this
223 amplification in two stages. For each of the three sets of models (CMIP3 eight-model set;
224 CMIP5 23-model set, and CMIP5 seven-model subset), analyzed separately, we use the grand
225 ensemble mean of the model All-Forcing runs (n=8, 23, or 7) as an estimate of the forced signal
226 to remove from observations. This provides the "first level" adjustment for the observations,
227 which is slightly different for each set of models. However, since the true forced response of a
228 given model is only approximately known, given the limited number of ensemble members that
229 are used to estimate this forced response, it follows that some residual forced variance will
230 remain in the observed series after this initial adjustment. We try to estimate how much variance
231 remains by using the same procedure that we used for observations, but applying it to each
232 individual All-Forcing run ensemble member. That is, for a given model, we consider each of its
233 All-Forcing ensemble member separately and remove the multi-model ensemble mean (as for
234 observations) to derive an internal variability estimate. We average this estimate across all of
235 that model's ensemble members to create an average standard deviation for that model, and then
236 average across all models to create a multi-model ensemble internal variability standard
237 deviation estimate. Next, we consider the model control runs, and compute the average standard

238 deviations for a sample of 50 randomly drawn 110-yr time series from each control run, average
239 those, and then average across all of the control runs to create an ensemble-average internal
240 variability estimate from the control runs. For each of the 110-yr segments, the control model
241 data is masked with the observed mask for the given grid point before being low-pass filtered.
242 By comparing the internal variability estimate derived from the All-Forcing runs with that from
243 the control runs for the same models, we derive the “second-level” adjustment. This average
244 adjustment is then applied to the standard deviation from the “first-level” adjusted observations
245 to obtain a new observed internal variability standard deviation estimate (Obs. St. Dev.*) that is
246 more suitable to compare with the model control runs. Given this method (which includes two
247 separate levels of adjustment) we can now more defensibly compare the model control run and
248 observed low-frequency variability.

249 We stress that our variance-comparison procedure described above is only a very rough test of
250 decadal variance consistency, and is not even attempted in data-poor regions such as the deep
251 Southern Ocean. There are inherent limitations to our estimates because there is only so much
252 observational data and only so many ensemble members supplied by the modeling centers. In
253 terms of observational temporal coverage, in order for a comparison to be done between model
254 and observations at a grid point, we require at least 50 points (out of 110) to be available in the
255 110-year annually resolved decadal filtered record. Forty percent temporal coverage is
256 required for an annual mean to be considered valid, and the decadal filter does a modest degree
257 of gap-infilling by computing a filtered value if at least four of seven annual values are available
258 within a seven-year wide sliding window.

259 The adjusted standard deviation of low-pass filtered observations (“Obs. St. Dev.*”) forms the
260 basis of the observed estimates and difference maps in Figs. 2 and 3, and of the variance
261 consistency tests that will be described later in this report.

262 The adjusted observed fields (“Obs. St. Dev.*”) suggest that the strongest low-frequency
263 internal surface temperature variability occurs over higher latitude land and oceanic regions of
264 the Northern Hemisphere. The modeled fields also show these features, though they are
265 somewhat stronger in the models than for the observed estimate. Thus, a feature that stands out
266 in the modeled minus observed (Obs. St. Dev.*) standard deviation field (Fig. 2 c, f, i) is the
267 tendency for model-simulated low-frequency internal variability to exceed the observed estimate
268 in high-latitude oceanic and continental regions of the Northern Hemisphere. Another feature is a
269 tendency for the modeled variability to be too small over much of the remaining ocean regions
270 and Southern Hemisphere as far south as about 40°S. Limited data coverage precludes an
271 assessment of low-frequency variability over of the Arctic Ocean, Antarctica, and the Southern
272 Ocean south of 40°S (gray regions on the maps).

273
274 The general features shown in the ensemble mean difference maps in Fig. 2 (c, f, i) are also
275 present to some degree for many of the individual models (Fig. 3). We also list in Fig. 3 the
276 spatial correlation coefficients between the individual model standard deviation fields (not
277 shown) and the observed field (Obs. St. Dev.*). These spatial correlations vary from about 0.5
278 to 0.7 for the models shown, indicating a relatively good agreement between individual models
279 and observations in the overall spatial structure of the variability. This gives us some confidence
280 in the models’ ability to simulate at least the broad-scale features of surface temperature low-
281 frequency variability.

282
283 There are a number of caveats to the comparison presented here. For example, uncertainties
284 remain in estimating the forced variability component from observations, which is used to create
285 the observed residual, and thus there are uncertainties in the observed internal variability
286 estimate used for comparison to the model control runs, as noted earlier. In addition, the
287 available observational records are relatively short compared with many of the model control
288 runs. As noted by Wittenberg (2009) and Vecchi and Wittenberg (2010), long-running control
289 runs suggest that internally generated SST variability, at least in the ENSO region, can vary
290 substantially between different 100-yr periods (approximately the length of record used here for
291 observations), which again emphasizes the caution that must be placed on comparisons of
292 modeled vs. observed internal variability based on records of relatively limited duration.

293

294

295 **4. Global mean surface temperature: Historical forcing runs**

296 *a. Time series of global mean surface temperature*

297 The global mean time series of surface temperature from the 20C3M historical runs are
298 compared with observations (black curves) in Fig. 4 in a form similar to that presented by Hegerl
299 et al. (2007). The historical dates of large volcanic eruptions are shown by vertical brown lines.
300 An analysis of the model time series for the CMIP3 and CMIP5 All-Forcing experiments is
301 presented in Figs. 4a-c, and for the available CMIP5 Natural-Forcing-Only experiments in Fig.
302 4d. The large shaded region on each plot shows the 5th to 95th percentile range of a single model
303 realization from the multi-model sample. The multi-model sample is formed by combining the
304 distributions of each of the models, with each model having an equal probability weight in the

305 multi-model distribution. The sub-distribution from each model is centered on that model's
306 ensemble mean with the distribution about that mean based on the control run for that model.
307 Thus the multi-model distribution incorporates the uncertainty due to differences between the
308 model ensemble means (i.e., forcing and response-to-forcing uncertainties) and uncertainties due
309 to internal variability for each model.

310 The analysis shows that for the All-Forcing runs (Fig. 4 a-c) most of the time the observed
311 annual means lie within the 5th to 95th percentile range of single model realizations, implying that
312 there is a consistency between the observed record and the multi-model ensemble of runs taken
313 as a whole. However, the range for the CMIP5 Natural-Forcing-Only simulations (Fig. 4d)
314 clearly separates from the observed time series after about 1960, indicating that Natural-Forcing-
315 Only runs are inconsistent with observations, particularly for the late 20th century global
316 warming.

317 The narrower shaded region between the two thick red lines (a-c) depicts the 5th to 95th percentile
318 range of the multi-model ensemble mean. This is fairly narrow, indicating that the multi-model
319 ensemble means of these particular sets of models are fairly well-constrained, with relatively
320 small uncertainty. The ensemble means of the CMIP3 and CMIP5 volcanic models (Fig. 4 a,c)
321 track the observations remarkably well although the apparent volcanically induced temporary
322 dips are not in full agreement with the observed behavior for those periods. For example, in Fig.
323 4a, and 4c, the multi-model responses to the Pinatubo and Krakatau eruptions appear to be larger
324 than in observations. These apparent discrepancies in the volcanic responses will require further
325 analysis (see e.g. Stenchikov et al. 2009) and are not a focus of the present study. For example,
326 one must carefully assess the role of internal climate variability in judging whether these
327 differences are significant or not.

328 The combined volcanic and non-volcanic CMIP3 ensemble (Fig. 4 (b)) shows a substantially
329 wider envelope of model behavior, as expected with the larger number of models and with the
330 wider discrepancy in forcing among these models. Since the “Non-Volcanic” runs have a
331 substantially less realistic representation of the forcing, we will generally emphasize the eight
332 CMIP3 models with “Volcanic” runs in panel (a) in our remaining forced model assessments for
333 the CMIP3 models in this study.

334 *b. Spectra of global mean surface temperature*

335 Figure 5 (a,b) shows the variance spectra of observed global mean temperature (black curves,
336 with a shaded range for the 90% confidence intervals) and of the individual CMIP3 and CMIP5
337 “Volcanic forcing” historical runs (red curves) from Fig. 4 (a, c), using data from the years 1880-
338 2010. The data were not detrended prior to computing the spectra. Before plotting, the raw
339 spectra were smoothed using a non-overlapping sliding boxcar window that groups the raw
340 spectra into groups of three calculable frequencies. The 90% confidence intervals on the
341 observed spectrum assume six degrees of freedom for each spectral estimate (group of three)
342 shown. The sum of the variance is plotted at the central frequency of the sliding boxcar window.
343 The enhanced power at low frequencies in (a,b) relative to (c,d) is associated with the strong
344 warming trend in both observations and the All-Forcing model runs. There is a strong tendency
345 for the model spectra to lie within the 90% confidence intervals of the observed spectra,
346 particularly at periods longer than 10 yr (frequency $< 0.1 \text{ yr}^{-1}$).

347 The spectra in Fig 5 (c) and (d) are based on residual time series from observations or model
348 historical runs, where the multi-model ensemble surface temperature time series from the
349 20C3M volcanically forced historical runs is first subtracted from the observed global mean

350 temperature series or from the individual model historical runs to form residual time series. As a
351 result of this filtering procedure, most of the long-term warming trend (e.g., Fig. 4 a, c) is
352 removed from the time series. The agreement between variance spectra of model and observed
353 residual time series in Fig. 5 (c,d) is not as good as for the original unfiltered spectra (Fig. 5 a,b),
354 particularly for the CMIP3.

355

356 Overall, the results of these comparisons suggest that the model simulations have a plausible
357 representation of variability of the climate system, in terms of the spatial pattern of variability
358 and the direct comparison of the time series of observed and historical run global mean surface
359 temperature. The spectral results suggest that the models, particularly the CMIP3, may have
360 some shortcomings in global low-frequency variability simulations, although there are
361 uncertainties in estimates of the internal climate variability as obtained by creating observed
362 residual time series. Overall, these findings encourage us to use the models to assess surface
363 temperature trends at the regional scale in the following sections, with the caveat that there is
364 likely room for improvement in the model simulations of internal variability. Further tests of
365 low-frequency variability are presented in Section 6.

366

367 **5. Trend assessments: global mean and regional time series**

368 *a. Methodology for the “sliding trend” analysis: CMIP5 models*

369 In this section we compare the observed and simulated historical (20C3M) temperature trends
370 obtained from global or regional averages, to assess whether a linear trend signal has emerged

371 from the “background noise” of internal or natural climate variability, as estimated by the
372 models. The primary focus is on the seven CMIP5 models that have Natural-Forcing-Only runs
373 extending to 2010. While we can extend All-Forcing runs to 2010, when necessary, using
374 RCP4.5 projections, this is not tractable for the Natural-Forcing-Only runs. We can use these
375 seven CMIP5 model runs together to assess whether the observed trends have emerged from the
376 background of natural variability and whether they contain an attributable anthropogenic
377 component. We also examine the full sample (23 models) of CMIP5 runs for our All-Forcing
378 run vs. control run analysis. For these 23 models and for the eight CMIP3 models that include
379 volcanic forcing (but for which we generally do not have Natural-Forcing-Only runs), we can
380 ask a more limited set of questions, namely whether the linear trend signal to 2010 in the
381 observations has emerged from the background of internal climate variability and whether the
382 All-Forcing run trends are consistent with the observed trends.

383 We assess the trends across a wide “sliding range” of start years beginning as early as 1861. All
384 trends in the analysis use 2010 as the end year. The general procedure we use is illustrated in
385 Fig. 7 (a) for global mean surface temperature. The black shaded curve in the figure shows the
386 value of the linear trend in observed global mean temperature for each beginning year from 1880
387 to 2000, in each case with the trend ending in the year 2010. The HadCRUT4 observed data set
388 contains an ensemble of 100 estimates, and these are used to create an ensemble of observed
389 trend estimates. The black shading depicts the 5th to 95th percentile range of this ensemble. The
390 first year plotted for global mean temperature was 1880 because the areal coverage and temporal
391 coverage requirements for a trend to 2010 were reached in that year. The observed temperature
392 trend to 2010 is about 0.5°C/100 yr (0.05 °C/decade) beginning early in the record (late 1800s)
393 and increases to about 2°C/100 yr (0.2°C/decade) by around 1980. The observed trend has

394 decreased for more recent start dates, falling below $1^{\circ}\text{C}/100\text{ yr}$ ($0.1^{\circ}\text{C}/\text{decade}$) for trends
395 beginning in the late 1990s.

396 The blue curve in Fig. 7a shows the “mean of ensemble mean trends” for the Natural-Forcing-
397 Only runs of the seven CMIP5 model subset (see caption). Each of the seven models is weighted
398 equally in the mean of ensemble means, even if a modeling center provided a greater or smaller
399 than average number of within-model ensemble members. The light blue shading in Fig. 7 (a)
400 shows the 5th to 95th percentile range of trend values for the Natural-Forcing-Only runs, which is
401 constructed using the long-term drift-adjusted control run variability (Fig. 1 c,d) from each
402 model. Under an assumption that internal variability in the control run is not substantially
403 different from that in the forced runs, we can use the long control run for each model to estimate
404 the component of inter-realization uncertainty that would be present in the forced trends; this is
405 helpful, since most centers did not provide enough ensemble members to precisely assess this
406 component of the uncertainty.

407 To prevent any one model from dominating the analysis, our approach also attempts to weight
408 the various models roughly equally. Thus even if one modeling center provided a much longer
409 control run than the others, each of these models would still get an equal weighting in
410 constructing a multi-model sample of internal climate variability. Control runs from each of the
411 seven CMIP5 models contribute equally to the multi-model sample from which the percentile
412 range is constructed, as long a particular model control run is “eligible” for use, meaning here
413 that the length of the usable part of the control run is at least three times the length of the
414 observed trend being examined.

415 Each randomly selected control run trend (from the seven models used) is combined with that
416 model's ensemble-mean Natural-Forcing-Only trend for that trend length, thus creating a
417 distribution of historical Natural-Forcing-Only trends that includes the uncertainty due to both
418 internal variability and the spread of forced responses across the seven models. The blue region
419 is the 5th to 95th percentile range of this distribution of trends, and thus relates to the uncertainty
420 of single ensemble members (which mimics the real world, itself a "single ensemble member").
421 Therefore, the distribution of trends used to construct the percentile range includes uncertainty
422 due to both the different natural forcings and responses of the individual models, and the
423 uncertainty due to the internal variability as simulated in the control runs. The random
424 resampling approach is necessary because the available control runs for the various models are of
425 different lengths and yet we purposely chose to give each available model an equal "vote" in
426 estimating internal variability. The samples are drawn from the control runs in the form of 150-
427 yr samples with randomly chosen start dates, with each sample masked with the observed mask
428 of missing data over the period 1861-2010 to create data sets with missing data characteristics
429 that are similar to those of the observations. The analysis in Fig. 7 (a) shows that observed
430 global temperature trends-to-2010 of almost any length are detectable compared to the CMIP5
431 Natural-Forcing-Only runs and simulated internal variability—even for trends as short as those
432 beginning around 1990. Note that the spread of uncertainty expands for shorter trends, reflecting
433 the fact that the model can internally produce relatively larger-magnitude trend *rates* over
434 relatively short periods.

435 The dark red curve and light pink shading in Fig. 7 (a) depict the inter-model mean of ensemble
436 means and the 5th to 95th percentile uncertainty range for the All-Forcing runs (i.e., natural and
437 anthropogenic forcings combined) and control runs for the seven-model CMIP5 subset. These

438 are constructed in an analogous way to the Natural-Forcing-Only curves and blue shading, and
439 thus depict the uncertainty due to both internal variability and to the different models' responses
440 to historical climate forcing agents (All Forcings, in this case). The violet shading in the plot is
441 the region where the pink and blue shading overlap, indicating that the 5th to 95th percentile
442 ranges of the All-Forcing and the Natural-Forcing simulated trends at least partially overlap.

443 In Fig. 7 (a), the black (observed) curve is always within the pink- (or violet-) shaded region,
444 meaning that global mean temperature trends are not significantly different from the CMIP5
445 historical All-Forcing run ensemble on any time scale, including the most recent 'weak trends'
446 beginning in the late 1990s.

447 When the black-shaded curve in Fig. 7a lies entirely within (or above) the pink-shaded region
448 and entirely outside of the blue-shaded region, we conclude that the trend from that point to 2010
449 has a detectable anthropogenic component. Given that the observed global mean surface
450 temperature trends with start dates through about the mid-1990s lie within this region of the
451 graph, we conclude that the observed global surface temperature warming to 2010 is at least
452 partially attributable to anthropogenic forcing according to these model data and observations.

453 Inspection of Fig. 7a further indicates that this detection and attribution result is sufficiently
454 strong that the uncertainty associated with the combined effects of internal climate variability,
455 uncertainty in the model responses to natural forcing, and the uncertainty in the observed
456 ensemble could be a factor of two larger than shown here and the same conclusion would still
457 hold for start dates from the late 1800s to about the mid-20th century. Our attribution conclusion
458 for anthropogenic forcing and global mean temperature is not as strong as in IPCC AR4 (Hegerl
459 et al. 2007), partly because we are not focusing in this study on quantifying the magnitude or
460 fractional contribution of the anthropogenic forcing. Also, our technique does not use

461 information in spatial patterns to distinguish between different forcings or to quantify the effect
462 of individual forcings (e.g., greenhouse gases). Rather, our focus is on evaluating the evidence
463 for detectable and attributable net anthropogenic influence on surface temperature in various
464 regions around the globe, using the ‘best estimates’ as provided by current models (without any
465 rescaling). We essentially compare two alternative hypotheses (natural and anthropogenic
466 forcings vs. natural forcings only) and focus down even to the scale of individual $5^{\circ} \times 5^{\circ}$ grid
467 boxes, which is important for regional climate change assessment.

468 There are some important caveats to the approach that we use, aside from the obvious one that
469 we rely on models to estimate the internal climate variability levels (which are compared to an
470 derived observed estimate Obs. St. Dev.* in Section 3b). The limited number of ensemble
471 members for the individual models means that there is additional variance in the grand
472 distributions of trends (i.e., pink- and blue-shaded regions) due to our imperfect knowledge of
473 each model’s forced response. However, the net impact of this limitation on the spread of the
474 total distribution is a complicated function of several factors. These include the following four
475 factors: 1) the number of ensemble members a particular model has (which we now show in Fig.
476 1; the larger the number of ensemble members, the smaller the overestimate of variance); 2)
477 where the models with few ensemble members sit in the distribution (if they are close to the
478 outer edge, the overestimate can be greater than if they are near the middle of the distribution); 3)
479 what is the variance of the model with few ensemble members or that sits at the outer edge of the
480 distribution; and 4) what is the relative size of the spread of the individual model ensemble
481 responses vs. the internal variability of the models near the outer edge of the distribution.

482 We can also estimate an upper limit on the overestimate of the standard deviation, based on the
483 number of ensemble members we use, as about 15-40% at most, with the worst case being for a

484 single ensemble member, where the variance is as much as doubled, so the standard deviation is
485 40% overestimated. However, given the four factors mentioned above, the effect will typically
486 be considerably smaller than this.

487 It is also worth noting that the effect of an overestimation of variance in our framework is to
488 make trends too difficult to detect (compared to internal variability or to the internal variability
489 plus natural forcing), but to also make it too easy for All-Forcing trends to be consistent with
490 observations.

491 We could in principle attempt a simulation to essentially estimate confidence intervals on our
492 confidence intervals, but these would be situation dependent and would vary for different
493 locations around the globe, time period, etc. We have chosen to leave this extension for further
494 studies, but note that the above issues should be considered in evaluating our results.

495

496 *b. Detection/attribution findings for various regional indices*

497 The sliding trend/ detection and attribution analysis discussed above for global mean temperature
498 can be applied to various regions around the globe. Here we briefly summarize the findings of
499 such an application (panels shown in Figs. 7 and 8).

500 1) MAJOR LARGE-SCALE REGIONAL INDICES

501 For **global sea surface temperature (SST)** (Fig. 7b), trends to 2010 are clearly detectable for
502 starting years up to about 1990. The observed trends are only marginally attributable to
503 anthropogenic forcing for trends beginning around the mid-20th century, otherwise an
504 attributable anthropogenic signal is clearly apparent for the detectable trends. For **global land**

505 **surface temperature** (Fig. 7c) an attributable anthropogenic signal is clearly seen in the
506 observed trends for all start dates from about 1885 up to about 1990, so the case for attribution is
507 slightly more robust than for global sea surface temperature. The anthropogenic warming signal
508 is so much stronger over land than over ocean, that it readily detectable and attributable despite
509 the greater intrinsic variability over land than over ocean. **Northern hemisphere temperature**
510 (Fig. 7d) roughly mirrors the results for global temperature and global land temperature, with
511 robust detection and attribution for start years up to about 1990. **Southern hemisphere**
512 **temperature** (Fig. 7e) results are similar though not quite as robust as for the Northern
513 hemisphere, as the start dates with attributable anthropogenic influence extending up to about
514 1980, rather than 1990.

515 The **northern hemisphere extratropics** (30°-90°N) series (Fig. 7f) has robust detection and
516 attribution up to around a 1990 start date, but the **southern hemisphere extratropics** (30°-90°S;
517 Fig. 7g) is slightly less robust than the northern hemisphere, as detection/attribution extends to
518 starts dates up to about 1980. The trends for the **southern extratropics** are relatively constant
519 over a range of start dates from 1900 to 1970, in contrast to **northern hemisphere** series which
520 shows a period of higher warming trend rates for trends to 2010 beginning in the second half of
521 the 20th century. The **southern extratropics** trends from 1900 are marginally consistent with the
522 All-Forcing model trends, as they are near the upper edge (95th percentile) of the modeled
523 distribution. An interesting feature of the **northern extratropics** and **southern extratropics**
524 trends is that there is essentially no start date for which the 5th to 95th percentile range of the All-
525 Forcing and Natural-Forcing-Only simulated trends are not at least partially overlapping. That
526 is, in some sense the All-Forcing and Natural-Forcing trends from the models are not completely
527 distinguishable from each other. The same will be true for many of the subsequent regional series

528 analyzed, especially for land regions and ocean regions with pronounced multi-decadal
529 variability. **Tropical surface temperatures**, which combine land and ocean (Fig. h) regions,
530 show robust detection and attribution for trends to 2010 with start dates as late as about the late
531 1970s.

532 2) REGIONAL SEA SURFACE TEMPERATURE INDICES

533 **Tropical SST's** (20°N-20°S; Fig. 7i) show similar robust detection and attribution results (for
534 start dates as late as about the 1970s) to those for the tropical surface temperature as a whole.
535 **Indian Ocean SSTs** (Fig. 7j; see Fig. 6 to identify region IO) exhibit robust detection and
536 attribution for start dates up to about 1990, despite a larger observational uncertainty, particularly
537 for trends beginning from the 1940s through the 1980s. A similar result is seen for the **tropical**
538 **Indian Ocean/western Pacific** warm pool index (Fig. 7k) and for the **tropical west Pacific** (Fig.
539 7l), which are important regions as they are dominant large-scale regions for tropical convection;
540 these have a detectable anthropogenic component for trends beginning up until about 1980. The
541 **tropical east Pacific** (Fig. 7m) shows a detectable anthropogenic component for trends to 2010
542 beginning from the 1880s to about 1920. However, trends beginning from 1920 to 1970 are only
543 marginally detectable as the black region (observations, including uncertainties) is not clearly
544 outside of the blue (natural forcing) region. **North Pacific SSTs** (25°-45°N, Fig. 7n, see Fig. 6 to
545 identify region), have a detectable anthropogenic component but only for start dates up to about
546 1910. A marginally detectable signal is found for start dates up to about 1930 and for a narrow
547 range of start years in the 1970s. Otherwise, the trends are not detectable according to our
548 analysis.

549 We analyzed four separate regions of the Atlantic Ocean, as this basin is noted for pronounced
550 multi-decadal variability. In the **South Atlantic** (Fig. 7o), there is a detectable anthropogenic
551 warming for start dates up to the late 1970s. An interesting feature in this region is that warming
552 trends from the 1890s are slightly higher than even the 95th percentile of the model simulations.
553 **North Atlantic SSTs** (45°-60°N; Fig. 8a) exhibit no detectable trends outside of the range of
554 natural variability for any start dates, according to our analysis. This region is notable for having
555 probably the least detectable signal of any of our study regions around the globe. Despite the
556 lack of detectable trends, the observed trends are at least consistent with the All-Forcing runs,
557 which have a very wide 5th to 95th percentile range of trends due to the large simulated internal
558 variability, as will be shown later in this section. In the **subtropical north Atlantic** (20°-45°N;
559 Fig. 8b) an anthropogenic signal is detected for start dates from about 1890 to 1920 and around
560 1970, but otherwise is only borderline detectable up to about 1980. In the **tropical North**
561 **Atlantic “main development region”** for Atlantic tropical cyclones (Fig. 8c), there is a
562 detectable anthropogenic warming to 2010 for start dates up to about 1960, and then only
563 intermittently for start dates up to about 1990.

564 MAJOR LAND REGION TEMPERATURE INDICES

565 We now summarize the characteristics of surface temperature trends in major continental
566 regions, beginning with Eurasia, Africa, and Australia. The **Europe** temperature index (Fig.
567 8d) has detectable anthropogenic warming trends for start dates up to about 1990, as the
568 observed trends (even accounting for observational uncertainty in the HadCRUT4 data set) are
569 outside of the range of the Natural-Forcing trends but lie well within the range for the All-
570 Forcing trends. The **Africa** index (Fig. 8e) has detectable anthropogenic warming trends for start
571 dates up to about the year 2000. Our analysis of **African** temperature trends only extends back

572 to start dates beginning in the mid-1920s, due to more limited data coverage. For **northern Asia**
573 (Fig. 8f), our start dates extend back to the early 1900s and show a clear detectable
574 anthropogenic warming signal for start dates extending from there up to about 1980. For
575 **southern Asia** (Fig. 8g) our analysis shows a similarly strong detectable anthropogenic warming
576 signal for start dates extending from the late 1800s through about 1990. An interesting feature of
577 the **African** and **southern Asia** results is that the 5th to 95th percentile range of the All-Forcing
578 trends from much of the 20th century is much wider than the range for the Natural-Forcing runs.
579 Since the contribution from internal variability (estimated from the control runs) is the same for
580 the two sets of trend results, the uncertainty range of the All-Forcing ensemble mean trends
581 across the models must be comparable to or substantially larger than the uncertainty due to
582 internal climate variability alone. The **Australia** temperature index (Fig. 8h) shows detectable
583 anthropogenic warming trends for start dates from the late 1800s to about 1970.

584 Considering now the land regions of North and South America, the index for **Canada** (Fig. 8i)
585 shows detectable anthropogenic warming trends for start dates up to about 1970. In contrast, for
586 the **Alaska** index (Fig. 8j), a detectable anthropogenic warming trend to 2010 is most clear for
587 start dates over the more limited range of 1940 to 1970. Trends for post-1970 start dates are
588 generally not detectable, and trends for start dates from about 1910 to 1940 are only marginally
589 detectable. For the **continental United States** (Fig. 8k) an anthropogenic warming trend to 2010
590 is detectable for start dates of about 1900 to 1975. For start dates of about 1860 to 1900, the
591 warming signal is only marginally detectable. The temperature index for **Mexico** (Fig. 8 l)
592 indicates that observational uncertainties play an important role for detection and attribution
593 results in this region. A detectable anthropogenic warming trend is seen for start dates of about
594 1910-1920 and about 1965-1980, otherwise the trends are not detectable. In contrast, for the

595 **South America** index (Fig. 8 m), the temperature trends to 2010 are mostly detectable for start
596 dates from about 1910 to 1950, but are not necessarily attributable to anthropogenic forcing for
597 these periods because the observed trend range is not entirely within the pink region (range of
598 All-Forcing simulated trends). Rather, they appear systematically smaller than the simulated
599 trends, after accounting for observational uncertainties. Anthropogenic warming trends to 2010
600 are detectable for the **South America** index but only for a limited set of start years in the early
601 1970s.

602 Temperature trends for the **southeastern United States** index (Fig. 8o) are of particular interest
603 because the trend behavior in this region is different from most other land regions around the
604 globe, as has been pointed out in a number of previous studies (e.g., Knutson et al. 1999, 2006;
605 Portmann et al. 2009). According to our present analysis, trends to 2010 in this index are
606 detectable only for a limited range of start years (mid-1950s to the mid-1970s). For that limited
607 set of start years, an anthropogenic warming trend to 2010 is detectable in our analysis. The
608 trends in the index to 2010 at least are consistent with All-Forcing runs for all start years after
609 about 1940, but the warming trends even after 1940 are for the most part not strong enough to be
610 detectable against the background of natural forcing and internal climate variability. This
611 behavior contrasts with the index for the **rest of the continental United States** (that lies outside
612 of the southeastern U.S.) (Fig. 8 o), where an anthropogenic warming trend to 2010 is broadly
613 detectable for start years ranging from about 1870 to the mid-1970s.

614 *c. Consistency test findings using CMIP3 and CMIP5 models*

615 Our regional temperature indices analysis in subsections 5(a) and 5(b) (i.e., Figs. 7 and 8)
616 focused on the subset of seven CMIP5 models that had Natural-Forcing-Only runs that extended

617 to 2010. Here we conduct a complimentary assessment (for a more limited set of regions) that
618 compares these results with similar analyses for the eight CMIP3 models (All-Forcing and
619 control runs) and with the full set of 23 CMIP5 models (All-Forcing and control runs). Where
620 necessary, the All-Forcing 20C3M runs were extended to 2010 using A1B (CMIP3) or RCP4.5
621 (CMIP5) projection runs; this procedure was not tenable for the Natural-Forcing-Only runs due
622 to the strong differences in forcing between Natural-Only and the A1B or RCP4.5 scenarios for
623 the extension years to 2010. Our analyses for the CMIP3 models (and the 23 CMIP5 models as
624 shown in the middle column of Fig. 9) therefore only compare internal climate variability
625 (control runs) with All-Forcing historical runs. Thus, we cannot use these results to draw firm
626 conclusions about detection of anthropogenic trends, because the alternative hypothesis (Natural-
627 Forcing) is not available through 2010 for all of the models. Nonetheless, we can draw some
628 conclusions about detection of significant trends (against a background of internal climate
629 variability) and about consistency of observed trends versus the trends in the All-Forcing 20C3M
630 experiments.

631 Our procedure is illustrated for the **global temperature** analysis in the top row of Fig. 9 (a-c).
632 Figure 9c is identical to Fig. 7a and is repeated here for reference only. Figure 9a shows the 5th
633 to 95th percentile range for the observed trends to 2010 (black shading); the 5th to 95th percentile
634 range for the All-Forcing runs from the eight CMIP3 models (pink shading, with the red curve
635 depicting the ensemble mean); and the 5th to 95th percentile range of control run trends from the
636 same eight CMIP3 models (green shading). Violet shading illustrates regions of overlap of the
637 pink- and green-shaded regions. Where the black curve lies outside of the green-shaded region,
638 the observed trend is detectable compared to internal climate variability in the CMIP3 runs.

639 Where the observed curve lies within the pink shading, the observed trend is assessed as
640 consistent with the CMIP3 All-Forcing ensemble of runs.

641 Figure 9a (CMIP3) indicates that the observed **global mean temperature** trends to 2010 are
642 detectable (inconsistent with internal climate variability in the eight CMIP3 models) for start
643 dates from about 1880 to the mid-1990s, and are consistent with the CMIP3 All-Forcing run
644 trends to 2010 for essentially all start dates from 1880 to 2000. Similar conclusions are evident
645 for the 23 CMIP5 models as shown in Fig. 9b. As noted earlier, similar results are seen for the
646 seven CMIP5 models when we incorporate the Natural-Forcing-Only runs in the tests (Fig. 9c),
647 although there the detectability of the observed trend extends to start dates as late as about 1990,
648 rather than into the mid-1990s.

649 For **tropical SST** (Fig. 9d-f) the CMIP5 models, including the seven model subset with Natural-
650 Forcing-Only runs to 2010 (Fig. 9 f), indicate robust detection and attribution for trends to 2010
651 for almost all start dates as late as about the late 1970s, as discussed earlier. The consistency
652 with the All-Forcing runs (all 23 CMIP5 models) is only marginal for a period of start dates
653 around 1960. A similar consistency result is seen for the 23 CMIP5 models (Fig. 9e) where we
654 compare their All-Forcing runs with their control variability. The observed trends to 2010
655 appear to be detectable against the internal variability (control run) background of the 23 CMIP5
656 models for start dates as late as about 1990. For the eight CMIP3 models (Fig. 9d), the observed
657 trends to 2010 are detectable for start dates up to 1990, similar to the CMIP5 models (Fig. 9e).
658 However, the eight CMIP3 All-Forcing runs are not as consistent with the observed trends to
659 2010 as the 23 CMIP5 All-Forcing runs. In fact the CMIP3 All-Forcing runs appear only
660 marginally consistent with the observed trends to 2010 for most of the start dates from 1880
661 through about 1980. This illustrates that the relatively modest levels of estimated internal

662 variability in this basin lead to a strongly detectable warming signal, but also make it difficult for
663 a model to be assessed as consistent with the observations, as the margin for error is relatively
664 small.

665
666 The **North Atlantic** (45° - 65° N) was highlighted earlier as a region with no detectable trends
667 compared with the CMIP5 Natural-Forcing-Only runs and internal climate variability combined
668 (Fig. 9i). This is perhaps not surprising, given the substantial intrinsically-generated fluctuations
669 on multi-decadal time scales in this region (see e.g. Yang et al. 2013). We see from the green
670 and violet shaded regions in Figs. 9 g,h that the range of trends to 2010 due to internal climate
671 variability alone in the CMIP3 and CMIP5 models is quite large and appears to largely account
672 for a similar wide range of simulated trends in the All-Forcing runs. This also helps allow the
673 observed trends to 2010 to be consistent with the CMIP3 and CMIP5 All-Forcing trends for all
674 of the start dates examined, despite the fact that the observed trends are not detectable (i.e., not
675 distinguishable from control run variability alone).

676 For the **southeastern United States** index (Fig. 9 j-l) there is slightly more evidence for
677 detectable trends to 2010 versus the internal variability samples in Fig. 9 j,k (start years 1950 to
678 1980) than versus the combined Natural-Forcing/internal variability sample of trends from the
679 seven CMIP5 models (blue shading in Fig. 9 (l)) with the latter having only marginally
680 detectable trends and only for start dates from the mid-1950s to the mid-1970s). For start years
681 prior to about 1940, the distribution of observed trends lies near the edge and even outside of this
682 5th to 95th percentile range for the All-Forcing runs (pink/violet shaded envelopes), especially
683 for the CMIP3 model sample (Fig. 9j). We thus conclude that even accounting for internal

684 variability, the CMIP3 and CMIP5 historical runs trends-to-2010 tend to be inconsistent or only
685 marginally consistent with the observed southeastern U.S. surface temperature trends,
686 particularly for starting dates in the early 20th century. This means that the CMIP3 and CMIP5
687 All-Forcing runs can be falsified, at least for this relatively small region, and further implies that
688 there remain as yet unexplained discrepancies between the historical simulations and
689 observations for trends in this region. We note that our tests are conducted on a large sample of
690 at least partly independent regions, and thus we would expect some fraction of the area to have
691 values that are too high or low due to chance. Further discussion of this issue in the context of
692 “global significance testing”, can be found, for example, in Knutson et al. (1999).

693 The results for the **rest of the continental United States** index (outside of the southeastern
694 United States; Fig. 9 m-o) are fairly consistent between the CMIP3 (m) and the CMIP5 models
695 (n, o), although as discussed above, the nature of our conclusions are different for Fig. 9 (m and
696 n) than for Fig. 9 (o), with the latter one including also the ensemble mean and additional
697 uncertainty range associated with the different model responses to Natural Forcings.

698

699 **6. Grid point-scale detection and attribution tests**

700 *a. Multi-model ensemble assessment*

701 1) 1901-2010 TRENDS

702 The procedures in Section 5 that were used to categorize observed trends at individual grid
703 points as detectable, attributable in part to anthropogenic forcing, consistent with All-Forcing
704 runs, etc. can be applied at the grid-point scale, and the categories displayed in map form, for a

705 selected trend period. For example, Fig. 10 shows the results of such a category analysis for the
706 observed vs modeled trends for 1901-2010, with the bottom row showing category maps for the
707 CMIP3 All-Forcing runs (e) and CMIP5 All-Forcing and Natural-Forcing-Only runs (f). The
708 linear trend maps for observed temperature (1901-2010) and the CMIP3 and CMIP5 All-Forcing
709 ensemble means are shown in Fig. 10 (a-d) for reference. The observed trend map shows broad-
710 scale warming trends since 1901 at almost all locations around the globe, with areas of cooling in
711 only a few regions, mainly in the high latitude North Atlantic and the southeastern United States.
712 The CMIP3 and CMIP5 multi-model ensemble trends show broadly similar magnitude and
713 pattern of cooling to observations, where the agreement can be quantitatively tested by our
714 consistency tests as described in the previous section. For the tests described in this section, we
715 use only the ensemble mean observed trend and thus do not consider observational uncertainty,
716 which was examined in the previous section.

717 Figure 10 (f), for the seven CMIP5 models with both All-Forcing runs and Natural-Forcing-Only
718 runs to 2010, builds upon the regional time series analysis shown in Figs. 7-8. The white regions
719 in Fig. 10 (f) indicate where the observed trend is not detectable compared to the Natural-
720 Forcing-Only runs (where the uncertainty estimates incorporate both simulated internal climate
721 variability from the seven control runs and uncertainties in the Natural-Forcing-Only ensemble
722 mean). The dark grey regions in Fig. 10 (f) do not have sufficient data coverage for our tests.
723 To determine if a grid point has “sufficient coverage” to include in our maps and analyzed area,
724 we divide a given trend period (e.g., 1901-2010) into five roughly equal periods, and require that
725 each of the five periods has at least 20% temporal coverage for annual means, where an annual
726 mean is considered available if at least 40% of the months are available for the year. The various
727 colored (non-white, non-grey) regions in Fig. 10 (f) indicate where the trends are detectable, with

728 the category identified on the legend. The yellow-orange regions show where the warming trend
729 is detectable but still less than the lower end (5th percentile) of the All-Forcing trend distribution.
730 The light-red and dark-red regions indicate where the observed trend has a detectable
731 anthropogenic component; for the darkest red regions the observed warming trend is so large that
732 it exceeds the 95th percentile of the modeled distribution, but here we still interpret this as
733 implying a detectable anthropogenic component. For cooling trends (blue regions), we have
734 analogous terms to those used for the various warming cases, although these cases are almost
735 absent for the 1901-2010 trends in our analysis.

736 The results for Fig. 10 (f) show that most of the global area with sufficient temporal coverage is
737 categorized as having attributable anthropogenic warming (either consistent in magnitude or
738 significantly larger than in the All-Forcing runs). The larger-than-simulated warming trends
739 occur preferentially in the extratropical South Pacific, the South Atlantic, the far eastern Atlantic
740 and the far western Pacific. In only a relatively small percentage of the globe is the observed
741 trend classified as not a detectable change (white regions in Fig. 10 f). These include mainly the
742 mid- to high-latitude North Atlantic, eastern United States, and parts of the eastern tropical and
743 subtropical Pacific.

744 A similar analysis for the CMIP3 All-Forcing runs (eight models with volcanic forcing) is shown
745 in the left column of Fig. 10 (a,c,e). The category names for the assessment (Fig. 10 e) are
746 different than for the CMIP5 models (Fig. 10 f) because a Natural-Forcing-Only ensemble is not
747 available in the archive for the CMIP3 models. Therefore, our categories for CMIP3 (see
748 legend) are limited to assessing consistency, either with the internal variability of the control
749 runs or with the All-Forcing runs, and we do not address the question of attribution to
750 anthropogenic forcing. The observed widespread warming trends shown in Fig. 10 (a) are

751 assessed as detectable (compared with control run or internal climate variability) over most of
752 the global region with sufficient coverage. Only in some regions of the North Atlantic, eastern
753 United States, and North Pacific (white regions in Fig. 10 (e) is the observed trend not
754 detectable. In only a very minor fraction of the analyzed area is there a detectable cooling trend
755 since 1901 (blue shading in Fig. 10 e), according to our analysis. Yellow-orange regions (where
756 the warming trend is detectable but less than simulated) occur preferentially in the lower
757 latitudes, and are more common in the CMIP3 assessment than the CMIP5 assessment. Regions
758 with significantly greater than observed warming trends (dark red) tend to occur more outside of
759 the tropics for the CMIP3 assessment (Fig. 10e), but are fairly common even in the tropics for
760 the CMIP5 assessment (Fig. 10f).

761 2) 1951-2010 TRENDS

762 Figure 11 explores how the results seen for 1901-2010 trends in Fig. 10 are altered when we
763 analyze the trends for 1951-2010. The observed trend map (Fig. 11 a) shows a more spatially
764 varying structure than the trend map for 1901-2010 (Fig. 10 a). The Asian and North American
765 extratropical land regions have warmed substantially more than oceanic regions since 1951. This
766 amplification of warming over land since 1951 is also evident in the All-Forcing 20C3M
767 ensemble means for both the CMIP3 eight-model set (Fig. 11c) and the CMIP5 seven-model
768 (Fig. 11d)—although the contrast between the continental and oceanic regions is more
769 pronounced in the observed trend map than in the multi-model ensembles, especially for CMIP3.
770 This is also seen in the category maps (Fig. 11 e, f) where dark-red shading (observed warming
771 significantly greater than simulated) is more prevalent over Asia and Alaska in the CMIP3
772 assessment (e) than in the CMIP5 assessment (f).

773 The observed trend map (Fig. 11 a, b) shows a region of notable cooling over the mid-latitude
774 North Pacific and a smaller region of cooling trends in the high-latitude North Atlantic just south
775 of Greenland. These cooling regions are assessed as having no detectable change (Fig. 11e, f),
776 meaning that the cooling trends lie within the 5th to 95th percentile range of the simulated trends
777 from the model control runs (CMIP3) or combined control run/Natural-Forcing runs (CMIP5).
778 Non-detection of trends for 1951-2010 (white category, Fig. 11 e,f) are also found over large
779 regions of the North Pacific, the central equatorial Pacific, the mid- to high-latitude North
780 Atlantic, the far Southern Ocean near Antarctica, and in a few scattered continental regions such
781 as the south-central or southeastern United States.

782 Figure 11 (f) indicates where observed trends (1951-2010) are attributable, at least in part, to
783 anthropogenic forcing (light-red and dark-red regions). These regions cover most of the global
784 area that has detectable trends, and for the 1951-2010 trends are comprised predominantly of
785 regions where the trends are consistent with the All-Forcing ensembles (i.e., light red). In
786 addition to the land regions (parts of Asia, Alaska) mentioned earlier, parts of the tropical Indian
787 Ocean and South Pacific also have warming trends that are not only attributable in part to
788 anthropogenic forcing but are even significantly larger than simulated in the CMIP5 All-Forcing
789 runs (dark-red shading). The category results for the eight CMIP3 models (Fig. 11 e) are
790 generally similar overall to those for the CMIP5, although the categories in Fig. 11 (e) do not
791 include attribution to anthropogenic forcing (see legend), since the CMIP3 set of models does
792 not include Natural-Forcing-Only runs that are necessary for such an attribution.

793 Regions in Fig.11 (e, f) with warming trends that are detectable but significantly less than
794 simulated in the All-Forcing runs (yellow-orange regions) are not that common, but are mainly
795 found in the tropical and subtropical latitudes. This, combined with the greater prevalence of

796 dark red (stronger than simulated warming) in the higher latitudes, implies that for the 1951-
797 2010 trends overall, the All-Forcing runs (CMIP3 and CMIP5) tend to exhibit too strong a
798 warming trend at lower latitudes but too little warming in high-latitudes.

799 3) 1981-2010 TRENDS

800 The trend assessment results for the much shorter period 1981-2010 are presented in Fig. 12.
801 The observed trend map (Fig. 12 a) has much more spatial structure than for either of the longer
802 trend periods in Figs. 10a and 11a. Since 1981 there have been extensive regions of cooling
803 trends over the tropical and subtropical eastern Pacific, Gulf of Alaska, and much of the high
804 latitude Southern Ocean. The trend assessment (Fig. 12 e, f) shows that for the most part, the
805 cooling trends in these regions are not detectable. In fact, since less than 5% of the globe has
806 “detectable” cooling trends, the percent of occurrence of the blue regions is not significantly
807 different from what could occur from sampling variability alone.

808 The large expanses of the globe without detectable trends (1981-2010) in Fig. 12 contrasts with
809 the earlier finding of detectable warming in most analyzed regions for the longer trend analyses
810 (Figs. 10, 11). The loss of a detectable signal, as one proceeds to later start dates in the 20th
811 century--and shorter trend periods--is not unexpected. For example, the results in Figs. 7-9
812 showed how the trend *rates* for internally generated trends in the model become higher for
813 shorter trend periods, as the models can produce strong internally generated trend rates over
814 relatively short periods. Comparing the category maps for different start dates (Fig. 10-12), the
815 loss of detectability, as one proceeds to later start dates, occurs first in the extratropical North
816 Atlantic (north of 40°N) and over large parts of the North Pacific, extending into the tropics, as
817 seen for the 1951-2010 trends (Figs. 11). For the late 20th century start dates (e.g., 1981-2010;

818 Fig. 12) the region of no detectable warming expands to cover most of the southern oceans, south
819 of 40°S, and extending south from 20°S in the South Atlantic. This non-detection region also
820 expands to include most of the eastern tropical and subtropical Pacific and much of the northern
821 extratropics over Eurasia, North America, and the North Pacific.

822 Of the regions with detectable trends for 1981-2010 (Fig. 12 e, f), the vast majority of grid points
823 have trends that are consistent with the models (light red) and thus are at least partly attributable
824 to anthropogenic forcing (CMIP5; Fig. 12f) or, in the case of the CMIP3 models (Fig. 12 e), at
825 least consistent with All-Forcing runs. These areas include large regions of the tropics,
826 subtropics, and mid-latitudes within about 40-50 degrees of the equator (except for the eastern
827 Pacific). The relatively robust emergence of a significant warming signal over a relatively short
828 time period (30 years) in the lower latitudes, as in Fig. 12 (f), is reminiscent of the recent study
829 of Mahlstein et al. (2011), who conclude that the earliest emergence of significant greenhouse
830 warming will occur in the summer season in low-latitude countries. They examined land regions
831 and looked at signal emergence for particular seasons (whereas we examine land and ocean
832 regions and focus on annual means). However, both studies point toward early emergence of
833 anthropogenic warming signals in lower latitudes, as opposed to most high latitude continental
834 regions. Some exceptions we note in Fig. 12 (f) include the significant anthropogenic warming
835 trends (1981-2010) in the vicinity of Greenland and in some land regions near the edge of the
836 Arctic Ocean.

837 There is relatively little yellow-orange area (which in our convention designates warming that is
838 detectable but significantly less than simulated) on the assessment maps for 1981-2010 (Fig. 12
839 e, f). The rare occurrence of this category for the later trend start dates can be explained by
840 referring to the sliding trend analyses in Figs. 7-9. The unshaded area on those graphs between

841 the pink- and blue-shaded “envelopes” corresponds to detectable warming that is less than
842 simulated. However, this region typically systematically shrinks as one progresses to later start
843 dates. That is, for shorter trend periods, it becomes much more difficult to distinguish the
844 simulated All-Forcing trend distribution from the trend distribution of the Natural-Forcing-Only
845 runs (CMIP5) or from the control runs (CMIP3).

846

847 4) ENSEMBLE MEAN ASSESSMENT STATISTICS ACROSS TIME

848 In Fig. 13, we explore how the percent of analyzed area with various category classifications
849 changes for different start years (all for trends ending in 2010). Figure 13(b) shows the
850 aggregate percent area results for the CMIP5 models, using the seven models that have Natural-
851 Forcing-Only runs extending to 2010. The total percent of analyzed area (i.e., regions with
852 sufficient data coverage) that was assessed as having attributable anthropogenic warming trends
853 (black curve) was about 80% for trends over the period 1901-2010. This drops to about 65% for
854 start dates from 1931 to 1971, before dropping sharply to about 25% for the shortest period
855 (1991-2010). There is a temporary increase in percent of area with attributable anthropogenic
856 warming for the 1971 start date, which is apparently due to the temporary pause in global
857 warming from about 1940 to 1970, which was preceded by a relatively strong rate of global
858 warming during early 20th century (Delworth and Knutson 2000). The end of this pause, around
859 1970, is a time period during which the prospects for detection of a warming signal are at least
860 temporarily enhanced against a backdrop of a gradually declining percentage as the start date is
861 moved forward through the 20th century. The blue curve in Fig. 13b (percent of analyzed area
862 with no detectable change) shows generally opposite behavior to the black curve, increasing

863 from a low of about 10%, for 1901-2010 trends, to a high point of over 60% for the latest start
864 period analyzed (1991-2010). The analysis thus illustrates the advantages of a long record for
865 detectability of the warming trend. The green curve shows that roughly 15 % of the analyzed
866 area has warming that is detected but less than simulated, for start dates through about 1941.
867 This percentage then declines for later start dates as the increasing dominance of internal
868 variability for short trend periods makes it much more difficult to distinguish the All-Forcing and
869 Natural-Forcing trend distributions and thus more difficult for a trend to lie between the two
870 distributions as discussed earlier. The percent of area with trends that are attributable to
871 anthropogenic forcing but significantly greater than simulated (red curve) also diminishes as the
872 start dates move later in the century, possibly because of the growing width of the simulated
873 trend distributions associated with internal climate variability, implying that it becomes difficult
874 for an observed trend to be large enough to be inconsistent with the All-Forcing distributions on
875 the high side.

876

877 Figure 13 (a) summarizes the comparison between the CMIP3 (eight-models with volcanic
878 forcing) and CMIP5 (23-model) results (solid lines vs. dashed lines) for various common
879 categories. This figure shows the percent areas corresponding to the maps in Figs. 10-12 (a, c, e)
880 for the CMIP3 models, but for a range of start dates. For the CMIP5, we use results for all 23
881 models that have volcanic forcing, since a Natural-Forcing-Only experiment (extending to 2010)
882 is not required for the comparisons shown in Fig. 13 (a), and thus we are not limited to the
883 seven-model subset of CMIP5. The percent area where the warming for the period 1901-2010 is
884 detected and either consistent or greater than simulated (black curves) is about 70% for CMIP3
885 and over 75% for CMIP5. This percentage decreases for start dates of 1931 or 1941, before

886 rising to a temporary peak of about 70% for the 1971 start date and then falling again for later
887 start dates. As discussed earlier, temporary rise for mid-century start dates is likely due to the
888 enhanced detectability of trends that start within the “relative trough” or temporary interruption
889 of global warming that occurred around this time following the relative peak in global
890 temperatures around 1940. For start dates up to about 1931, the black curve for the CMIP5
891 models (dashed) is about 5% higher on average than the (solid) one for the CMIP3 models.
892 Thus, the 23 CMIP5 model All-Forcing runs appear at least slightly more consistent with
893 observed trends than the eight CMIP3 All-Forcing runs, at least for the case of trends to 2010
894 starting earlier than 1940. However, for trends with start dates from 1941 through about 1971,
895 the opposite is true, and the CMIP3 All-Forcing runs appear modestly more consistent with
896 observations. Other features in Fig. 13 (a) are generally similar to those described for the seven
897 CMIP5 models (Fig. 13 b), although the category descriptions (conclusions about attribution) are
898 necessarily different. The general temporal behavior of the various curves through time is
899 remarkably similar between the solid (CMIP3) and dashed (CMIP5) models in Fig. 13 (a).

900 *b. Model by model trend assessment*

901 In contrast to the analyses in the previous subsection (Figs. 10-13) which focused on the multi-
902 model ensemble means vs. observations, in this subsection we consider the individual models
903 within the CMIP3 and CMIP5 ensembles and assess what percentage of individual models meet
904 certain criteria. That is, the determination of whether a given CMIP3 or CMIP5 individual
905 model is included in a category (e.g., “warming- detectable and consistent”) for a given grid
906 point is based on the evaluation of the historical runs and control runs for that model alone. In
907 this section, we also introduce and apply a variance consistency test as an addition consistency
908 test for the models vs. observations.

909 We will introduce and describe the various tests as we discuss the different panels in Fig. 14,
910 which contains the analysis of the eight CMIP3 models (with volcanic forcing) vs. observations
911 for linear trends over the period 1901-2010. Figure 14 (a) and (b) present the observed and
912 multi-model ensemble mean trend maps for reference; these were discussed earlier for Fig. 10.
913 Figure 14 (c) shows the fraction (or percent) of models, at each grid point, that have no
914 detectable trend. The area-weighted global average of this fraction is 0.09, and the most
915 prominent regions with no detectable trend are in the North Atlantic (south of Greenland), the
916 mid-latitude North Pacific, and the southeastern United States. Figure 14 (d) shows the fraction
917 of models at each grid point with warming that is detectable but less than simulated in the All-
918 Forcing runs. The global average fraction is 0.22, and the most prominent regions of occurrence
919 are in the tropics, meaning that the eight CMIP3 models, viewed independently, have a tendency
920 to simulate too rapid a century-scale warming in the tropics. The warming is detectable and
921 consistent with the All-Forcing runs for a global average fraction of 0.34 of the models (Fig. 14
922 e), with a spatial pattern that is fairly evenly distributed around the analyzed areas of the globe.
923 The warming is detectable and significantly greater than simulated for a global average fraction
924 of 0.32 of the models (Fig. 14 f), with the most prominent occurrence of this category being in
925 the mid- to high latitudes of both hemispheres. Warming is detectable for about 89% of the
926 models, on average around the globe (Fig. 14 g)—essentially the inverse of the results in Fig. 14
927 (c). Warming is detectable and consistent or greater than simulated for two thirds of the models,
928 on average, (Fig. 14 h) which shows essentially the inverse of the pattern in Fig. 14 (d), and
929 indicates that the simulated warming tends to be too weak in mid- to higher latitudes in the
930 CMIP3 All-Forcing runs. The observed and CMIP3 simulated (All-Forcing) trends are assessed
931 as consistent for 39% of the models on average (Fig. 14 i); this category includes cases where the

932 trend is not detectable, but still consistent with the All-Forcing runs (see Figs. 7-9 for example).
933 This fraction field (Fig. 14i) has a fairly even spatial distribution over the global analyzed area.
934 One limitation of our approach is that models with unrealistically large internal variability have
935 some advantage over models with more realistic variability, in that it is easier for high-variability
936 models to have trends that are consistent with observations, since the margin of error is greater.
937 To address this concern, here we apply a second test (a variance consistency test) to the models.
938 Then a model that has both a consistent trend and consistent variability, compared with observed
939 estimates, will be ranked more highly in a metric test compared with a model with consistent
940 trends but inconsistent variability. In other words, this expands our consistency tests into a two-
941 dimensional space (trend and internal variability).

942 The variance consistency test for the eight CMIP3 models with volcanic forcing (Fig. 14 j) is
943 constructed as follows. For each grid point, we estimate the adjusted observed standard
944 deviation of low-frequency (>10 yr) internal variability (Obs. St. Dev.*) as discussed in Section
945 3b. This variability is compared with a distribution of low-frequency standard deviations from
946 the model control run, which is obtained by drawing 50 random 110-yr samples of combined
947 SST and surface air temperature from the drift-adjusted control run (see Section 3 a), masking
948 missing data periods with the observed mask for the given grid point, low-pass filtering, and
949 computing the 50 standard deviation estimates. If the Obs. St. Dev.* value for the grid point lies
950 within the 5th to 95th percentile range of the combined control run distribution, the model is
951 assessed as having low-frequency internal climate variability that is consistent with the
952 observations according to this test. There are important limitations of this test, which we
953 recognize at the outset. When applied to a single model, as done here, a single model's control
954 run may not be long enough to provide an adequate sample of the 5th to 95th percentile range of

955 low-frequency (>10 yr) variance estimates; indeed, this is an important reason to advocate for
956 longer control runs (or larger ensemble sizes) in future CMIP designs. In addition, the observed
957 residual, which is needed for comparisons with control run variability, has some uncertainties, as
958 the multi-model ensemble mean forced response only approximately removes the forced climate
959 signal from the observations. Our adjustment procedure used to create Obs. St. Dev.*, described
960 in Section 3b, attempts to account for this uncertainty.

961 Figure 14 (j) illustrates the results of applying the test. On average, 26% of CMIP3 models have
962 variability consistent with observations, according to the test. Locations where the modeled low-
963 frequency variability is consistent with observations are fairly evenly distributed around the
964 globe, although the fraction is notably low in the southeastern Pacific and south Atlantic basins.

965 Figure 14 (k) shows the map of the fraction of the CMIP3 models where both the variability and
966 trend are consistent with observations on a grid point basis according to our tests. The global
967 average fraction is 0.11, indicating that achieving consistency with both tests simultaneously at
968 the grid point scale is a challenge for the CMIP3 models.

969 The variance consistency test can also be applied to the global mean temperature series (e.g.,
970 Figs. 4b, 5c, and 9a). We find that seven of the eight CMIP3 models (88%) have low-frequency
971 variance for their global mean temperature that is consistent with the observed residual,
972 according to our test, with one model having variance that is significantly too low.

973 Figures 15 and 16 present the same analysis as Fig. 14, but for the 23 CMIP5 models with All-
974 Forcing runs (Fig. 15), and for the subset of seven CMIP5 models that have at least one Natural-
975 Forcing-Only run extending to 2010 (Fig. 16). The mapped results for the 23 CMIP5 models
976 (Fig. 15) are rather similar overall and have similar spatial features to those for the eight CMIP3

977 models (Fig. 14) discussed above. One notable difference is that the CMIP5 models in both Fig.
978 15 and Fig. 16 have a greater global mean fraction of models with consistent low-frequency
979 variance (0.30-0.31) than the CMIP3 models in Fig. 14 (0.26). The globally averaged fraction of
980 models that have both consistent trend and variance (panel k) is about the same in CMIP5 (0.12)
981 as in the CMIP3 sample (0.11). Figure 16, for the seven model subset of CMIP5 models, shows
982 where trends are assessed as containing attributable anthropogenic trend contributions. The
983 analysis indicates that the globally averaged percent of the seven CMIP5 models with
984 attributable anthropogenic warming at the grid point scale over the 1901-2010 period is 70%
985 (Fig. 16 h). The globally averaged percentage of models with both attributable anthropogenic
986 warming and consistent low-frequency variance is 22%, according to the tests described above
987 (Fig. 16 l).

988 The variance consistency test can also be applied to the global mean temperature series for both
989 the full set of 23 CMIP5 models and the seven model subset of CMIP5 models. This test
990 indicates that 15 of the 23 CMIP5 models (65%), and four of the seven-model CMIP5 model
991 subset (57%), have global mean low-frequency variance that is consistent with observations.
992 This is a smaller fraction than for the CMIP3 models (88%). In cases of inconsistency, the
993 model variance is too low more often than too high (six low vs. two high for the CMIP5 23
994 models; and two low vs. one high for the CMIP5 seven-model subset). For cases other than the
995 global mean, Figures 2 and 3 depict where the low-frequency variability of individual models, or
996 the ensemble-average low-frequency variability across the models, tends to be either too low or
997 too high, compared to the adjusted observed internal variability estimate (Obs. St. Dev.*).

998 As has been discussed mentioned earlier, there are a number of limitations in our trend variance
999 estimates and consistency tests. We hope to improve on the variance consistency tests in a future

1000 study; for example, there are other model-observation comparison paradigms that can be
1001 explored (e.g., Annan and Hargreaves 2010). Meanwhile, we stress the need for longer control
1002 runs and/or greater numbers of independent ensemble members from the models in order to more
1003 robustly assess the various models' low-frequency variability.

1004 Figure 17 summarizes several globally averaged trend consistency metrics as a function of trend
1005 start year for the individual models in the CMIP3 and CMIP5 samples. Fig. 17 (b, d, and f) also
1006 assess the consistency of the models' low frequency variability, as these include both a trend
1007 consistency test and a variability consistency test. In the various panels of Figure 17, we
1008 compare, across the models, the fraction of analyzed area where there is both a detectable change
1009 in observations and where this detectable change is consistent with the individual climate
1010 models. Note that these metrics do not include the fraction of area where a climate model is
1011 consistent with observations but there is not a detectable trend.

1012 While all metrics have shortcomings, the particular metrics in Fig. 17 have at least some useful
1013 compensation properties. For example, for a model with unrealistically large internal variability,
1014 the enhanced potential for consistency of modeled and observed trends due simply to the larger
1015 internal variability is partly compensated by a reduction in the area assessed as having detectable
1016 trends according to that model. The two-dimensional (trend and low-frequency variance)
1017 consistency tests provide for an even greater compensating balance against the potential metric
1018 problem mentioned above.

1019 The results in Fig. 17 (a, c) show that the individual CMIP3 and CMIP5 models have rather
1020 similar behavior in terms of fraction of globally analyzed area with consistent detectable trends
1021 (typically ranging from 20 to 50%). There is somewhat more spread among the CMIP5 models,

1022 although there are more models in the CMIP5 sample as well. This trend consistency metric
1023 tends to reach a peak value around 1960-1970 start dates before declining for later start dates, for
1024 reasons discussed for Fig. 13. When a variance consistency test is added (Fig. 17 b,d), the
1025 percent of analyzed global area with both consistent trends and consistent low frequency
1026 variance drops substantially, to typically about 10 to 20%. Clearly the variance consistency test
1027 proposed here can pose a challenging test for the current models. We have plans to explore other
1028 types of variance consistency tests in our future work.

1029

1030 For the seven-model CMIP5 sample (Fig. 17 e), the percent of analyzed global area with
1031 attributable anthropogenic trends (including trends that are detectable but greater than simulated)
1032 is close to 80% for 1901-2010 trends, for five of the seven models, with the remaining two
1033 models having lower percent area (40-60%). All seven models end up in the range of 40-70%
1034 for this metric for the latest starting date analyzed (1991). The metric that tests for both
1035 attributable anthropogenic trend and consistent low-frequency variance (Fig. 17 f), indicates that
1036 the seven models have a range of percent area of 17-35% for the 1901-2010 trends, but this range
1037 decreases to about 10% or less for the 1991-2010 trends.

1038 **7. Supplemental material and further sensitivity studies**

1039 The analysis presented in this study introduces a framework for trend analysis that has many
1040 possible applications and extensions. For surface temperature, there are many figures that are
1041 variations on the ones presented here, but were too numerous to include in this article.
1042 Therefore, we have created a web site based largely on this analysis, but which contains
1043 additional supplemental figures (<http://www.gfdl.noaa.gov/surface-temperature-trends>). For

1044 example, the web site contains plots for individual seasons that complement the annual-averaged
1045 analysis in this study. We show plots using alternative percentiles (97.5th and 2.5th) instead of
1046 95th and 5th, and plots excluding certain low variability models from the analysis, etc. Additional
1047 regional plots like Figs. 7-9, including ones for individual seasons, are available, as well as maps
1048 for different trend start dates. In addition, a number of plots based on analysis of individual
1049 CMIP3 or CMIP5 models, as opposed to multi-model ensemble means, are available.

1050

1051

1052 **8. Summary and Conclusions**

1053 The purpose of this analysis has been to introduce and apply a framework for assessing regional
1054 surface temperature trends using both the CMIP3 and CMIP5 models and using a multi-model
1055 sampling approach. We examined the behavior of the various control runs for the CMIP3 and
1056 CMIP5 models, and used the control run variability to help assess whether observed trends were
1057 unusual or not compared with the models' internally generated variability. We also used the
1058 control run variability to help assess whether observed trends were consistent with trends from
1059 the historical (20C3M) simulations—either All-Forcing runs or Natural-Forcing-Only runs. In
1060 cases for the CMIP5 models where trends were demonstrated to be inconsistent with Natural-
1061 Forcing-Only, but consistent with the All-Forcing runs, we conclude that an attributable
1062 anthropogenic component is present in the observed trend. For cases, such as the CMIP3 model
1063 assessments, where Natural-Forcing-Only runs are generally not available, we tested for
1064 detectable trends (compared to internal climate variability) and for consistency between observed
1065 and All-Forcing historical (20C3M) runs.

1066 In the separate CMIP3 and CMIP5 analyses, we generally attempt to give different models equal
1067 weight, even when a modeling center provides fewer ensemble members or shorter control runs.
1068 Tests are applied at global and regional scales, as well as at individual grid points on the
1069 observed data grid where there is sufficient data coverage over the period of the trend. Results
1070 are summarized using classification maps and global percent area statistics. Our analysis
1071 contains a substantial assessment of the variability in the models, including control run time
1072 series for visual inspection, standard deviation maps of low-pass filtered data, spectral analysis,
1073 and a low-frequency variance consistency test that is applied to individual models.

1074 One of the most important results from the assessment is the identification of regions—and even
1075 grid points--where an anthropogenic warming signal is detectable in the observed temperature
1076 records. For trends over the period 1901-2010, a large fraction (about 80%) of the global area
1077 (with sufficient data coverage over time) has a detectable anthropogenic warming signal.
1078 Regions where the observed warming seems to be most commonly underestimated by the models
1079 include the southern Ocean, south Atlantic, the far eastern North Atlantic, and off the east coast
1080 of Asia. The main regions without detectable warming signals include the high latitude North
1081 Atlantic, the eastern U.S., and parts of the eastern and North Pacific. Moving forward in time,
1082 for the much shorter period (1981-2010) the observed warming trends over about 45% of the
1083 globe are assessed as having a detectable anthropogenic contribution. These regions include
1084 parts of the tropics, subtropics, and mid-latitudes (within about 40-45 degrees of the equator),
1085 and a narrow zonally oriented band near the Arctic Ocean. Areas without detectable trends
1086 (1981-2010) include much of the eastern Pacific--which is a region influenced by strong
1087 interannual variability associated with ENSO--and many extratropical regions poleward of about
1088 40°N and 40°S. The CMIP3 models and the larger sample (23) of CMIP5 models yield results

1089 similar to those described above, although for these samples we assess only the consistency of
1090 trends, and not whether they contain an attributable anthropogenic component (due to the lack of
1091 Natural-Forcing runs with which to do such an assessment).

1092 The reduced global area with detectable anthropogenic trends as one examines later start dates
1093 for trends in the record (all trends ending in 2010) illustrates the advantages of long records for
1094 trend detection in the context of this model-based assessment. In general, the shorter the epoch,
1095 the larger is the potential contribution of internal variability to the trend, leading to a greater
1096 spread (uncertainty) for sampled trends.

1097 There are numerous examples of modeled trends or variability that are inconsistent with
1098 observations in our study. As has been noted in a previous paper using a similar methodology
1099 with two climate models (Knutson et al. 2006), disagreement between modeled and observed
1100 trends in this type of analysis can occur due to shortcomings of models (internal variability
1101 simulation; response to forcing), shortcomings of the specified historical forcings, or problems
1102 with the observed data. A certain fraction of area should be expected to have inconsistent results
1103 due to chance alone (see Knutson et al. 1999 for further discussion of global significance testing
1104 in this context). As a further example, Wu and Karoly (2007) and Wu (2010) have noted that
1105 disagreement between simulated and observed regional surface temperature trends can result
1106 from shortcomings of models in simulating the observed warming associated with the changes of
1107 the leading climate variability modes (such as the Arctic Oscillation).

1108 Concerning observational uncertainty, the HadCRUT4 data set (Morice et al. 2012) contains 100
1109 ensemble members that attempt to characterize the uncertainties in the observations. We have
1110 performed some preliminary tests using these ensembles to assess the spread of observed trend

1111 estimates. These tests thus far indicate that even at the regional scale, the spread in trend
1112 estimates due to observational uncertainties, as contained in the ensembles, is generally much
1113 smaller than the spread in model simulated trends due to the internal variability and differences
1114 in forced responses in the historical runs (e.g., Figs. 7-9). However, in some regions (e.g.,
1115 Mexico), the uncertainty in the observations plays an important role in the assessment of
1116 detectable anthropogenic contributions to trends.

1117 We have attempted to at least partially address the issue of model uncertainties in the simulation
1118 of internal climate variability and in the response to historical forcing by using multi-model
1119 ensembles and by assessing consistency of both trends and low-frequency variability. When we
1120 apply a two-dimensional screening test (assessing simultaneously the consistency of the trend
1121 and low-frequency variability) we find that most models tend to be challenged to be consistent
1122 on both tests. Overall, our variance consistency tests suggest that while the CMIP3 and CMIP5
1123 models provide a plausible representation of internal climate variability, there is considerable
1124 scope for improvement in the model simulations of internal climate variability, apart from their
1125 simulation of trends and variability in response to various forcing agents. From a different
1126 perspective, Shin and Sardeshmukh (2011) have concluded that the CMIP3 models do not
1127 simulate historical trends of temperature and precipitation as realistically as do atmospheric
1128 models forced by observed trends in tropical SSTs—a problem they attribute to model errors as
1129 opposed to climate noise (internal variability).

1130 The CMIP3 and CMIP5 simulations used here represent “ensembles of opportunity” which
1131 cannot necessarily be expected to represent the true structural uncertainty in the results, due to
1132 shortcomings/uncertainties in the models and climate forcings. The procedures in our paper
1133 assume that the intrinsic internal variability of climate has not changed significantly since pre-

1134 industrial times, as we are using control run variability from pre-industrial control runs for our
1135 forced-run consistency tests. If anthropogenic forcing had actually *weakened* the intrinsic
1136 variability in the real world, then our estimated uncertainty range around the All-Forcing model
1137 responses would be too wide -- making it overly difficult to conclude that observations were
1138 inconsistent with the All-Forcing runs. Similarly, if anthropogenic forcing had actually
1139 *strengthened* the intrinsic variability in the real world, then our estimated uncertainty range
1140 around the All-Forcing model responses would be too narrow -- making it too easy to conclude
1141 that the observations were inconsistent with the All-Forcing runs.

1142 While the above uncertainty issues lack a final resolution, the methodology shown here can at
1143 least help to quantify the uncertainties associated with the climate change detection and
1144 attribution problem. The results show that when CMIP3 and CMIP5 historical runs are
1145 confronted with observed surface temperature trends, across a wide range of trend start dates, at
1146 various geographical locations around the globe, and even down to the grid point scale, a
1147 pervasive warming signal is found that is generally much more consistent with simulations that
1148 include anthropogenic forcing than with simulations that include either no forcing changes
1149 (control runs) or that include only natural forcing agents (solar, volcanic). Our conclusions about
1150 detectable anthropogenic contributions to the trends provide further support for the claim of a
1151 substantial human influence on climate, via anthropogenic forcing agents such as increased
1152 greenhouse gases. A future enhancement of our analysis would include an attempt to quantify
1153 the contributions of specific natural and anthropogenic forcing agents, or subsets of agents, in the
1154 CMIP5 All-Forcing and Natural-Forcing-Only historical runs. This would provide a more direct
1155 assessment of the relative influence of different forcing agents on the observed temperature
1156 trends at the regional scale.

1157

1158 Acknowledgments. We thank the Met Office Hadley Centre and the Climatic Research Unit,
1159 Univ. of East Anglia, for making the HadCRUT4 data set available to the research community.
1160 We thank the modeling groups participating in CMIP3 and CMIP5, and PCMDI for generously
1161 making the model output used in our report available to the community, and we thank three
1162 anonymous reviewers for their helpful comments on the manuscript.

1163

1164

1165

1166

1167 **References**

1168 Allen, M. R., and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting. Part
1169 I: Theory. *Clim. Dyn.*, **21**, 477-491.

1170

1171 Allen, M. R., and S.F.B. Tett, 1999: Checking for model consistency in optimal fingerprinting.
1172 *Clim. Dyn.*, **15**, 419-434.

1173 Delworth, T. L., and T. R. Knutson, 2000: Simulation of early 20th Century global warming.
1174 *Science*, **287**(5461), 2246-2250.

1175 Annan, J. D. and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble, *Geophys. Res.*
1176 *Lett.*, **37**, L02703, doi:10.1029/2009GL041994.

- 1177
- 1178 Hasselmann, K., 1997: Multi-pattern fingerprint method for detection and attribution of climate
1179 change. *Clim. Dyn.*, **13**, 601-612.
- 1180 Hegerl, G. C., F. W. Zwiers, P. Braconnot, N. P. Gillett, Y. Luo, J. A. Marengo Orsini, N.
1181 Nicholls, J. E. Penner, and P. A. Stott, 2007: Understanding and attributing climate change. In
1182 *Climate Change 2007: The Physical Science Basis*. [Solomon, S., D. Qin, M. Manning, Z. Chen,
1183 M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press,
1184 Cambridge, United Kingdom and New York, NY, USA, 996 pp.
- 1185
- 1186 Hegerl, G. C., et al. 2009: Good practice guidance paper on detection and attribution related to
1187 anthropogenic climate change. Available from IPCC: [www.ipcc.ch/pdf/supporting-](http://www.ipcc.ch/pdf/supporting-material/ipcc_good_practice_guidance_paper_anthropogenic.pdf)
1188 [material/ipcc_good_practice_guidance_paper_anthropogenic.pdf](http://www.ipcc.ch/pdf/supporting-material/ipcc_good_practice_guidance_paper_anthropogenic.pdf)
- 1189 Hegerl, G.C., H. v. Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996:
1190 Detecting greenhouse gas induced climate change with an optimal fingerprint method. *J.*
1191 *Climate*, **9**, 2281-2306.
- 1192 Karoly, D.J., and Q. Wu, 2005: Detection of regional surface temperature trends. *J. Clim.*, **18**,
1193 4337–4343.
- 1194 Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: Reassessing
1195 biases and other uncertainties in sea-surface temperature observations measured in situ
1196 since 1850, part 2: biases and homogenization. *J. Geophys. Res.*, **116**, D14104,
1197 doi:10.1029/2010JD015220.

- 1198 Knutson, T.R., T.L. Delworth, K.W. Dixon, and R.J. Stouffer, 1999: Model assessment of
1199 regional surface temperature trends (1949-1997). *J. Geophys. Res.*, **104**, 30981–30996.
1200
- 1201 Knutson, T.R., et al., 2006: Assessment of twentieth-century regional surface temperature trends
1202 using the GFDL CM2 coupled models. *J. Clim.*, **19**, 1624–1651.
1203
- 1204 Mahlstein, I., R. Knutti, S. Solomon, and R. W. Portmann, 2011: Early onset of significant local
1205 warming in low latitude countries. *Environ. Res. Lett.*, **6**, 034009, doi:10.1088/1748-
1206 9326/6/034009.
1207
- 1208 Meehl, G. A. et al., 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change
1209 research. *Bull. Amer. Meteor. Soc.* **88**, 1383–1394.
1210
- 1211 Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in
1212 global and regional temperature change using an ensemble of observation estimates: The
1213 HadCRUT5 data set. *J. Geophys. Res.*, 117, D08101, doi:10.1029/2011JD017187.
1214
- 1215 Portmann, R. W., S. Solomon and G.C. Hegerl, 2009: Spatial and seasonal patterns in climate
1216 change, temperature, and precipitation across the United States. *Proc. Nat. Acad. Sci.*,
1217 www.pnas.org/cgi/doi/10.1073/pnas.0808533106
1218
- 1219 Rind, D., M. Chin, G. Feingold, D. Streets, R. A. Kahn, S. E. Schwartz, and H. Yu, 2009:
1220 Modeling the effects of aerosols on climate. In *Atmospheric Aerosol Properties and*

- 1221 *Climate Impacts, A Report by the U.S. Climate Change Science Program and the*
1222 *Subcommittee on Global Change Research.* [Mian Chin, Ralph A. Kahn, and Stephen E.
1223 Schwartz (eds.)]. National Aeronautics and Space Administration, Washington, D.C.,
1224 USA.
- 1225
- 1226 Sakaguchi, K. X. Zeng, and M. A. Brunke, 2012: Temporal- and Spatial-scale dependence of
1227 three CMIP3 climate models in simulating the surface temperature trend in the twentieth
1228 century. *J. Climate*, **25**, 2456-2470.
- 1229
- 1230 Santer, B. D., T. M. L. Wigley, and P. D. Jones, 1993: Correlation method in fingerprint
1231 detection studies. *Clim. Dyn.*, **8**, 265-276.
- 1232
- 1233 Schneider, T., and I.M. Held, 2001: Discriminants of twentieth-century changes in Earth surface
1234 temperatures. *J. Clim.*, **14**, 249–254.
- 1235
- 1236 Shin, S.-I., and P. D. Sardeshmuhk, 2011: Critical influence of the pattern of tropical ocean
1237 warming on remote climate trends. *Clim. Dyn.*, **36**, 1577-1591.
- 1238
- 1239 Stenchikov, G., T. L. Delworth, V. Ramaswamy, R. J. Stouffer, A. Wittenberg, and F. Zeng,
1240 2009: Volcanic signals in oceans. *J. Geophys. Res.*, **114**, D16104. doi:
1241 10.1029/2008JD011673.
- 1242

- 1243 Stouffer, R. J., S. Manabe, and K. Y. Vinnikov, 1994: Model assessment of the role of natural
1244 variability in recent global warming. *Nature*, **367**, 634-636.
- 1245
- 1246 Stouffer R. J., Hegerl G. C. and Tett S. F. B. (2000): A comparison of Surface Air Temperature
1247 Variability in Three 1000-Year coupled Ocean-Atmosphere Model Integrations. *J.*
1248 *Climate*, **13**, 513-537.
- 1249
- 1250 Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An overview of CMIP5 and the experiment
1251 design. *Bull. Amer. Meteor. Soc.*, **93**, 485-498
- 1252 .
- 1253 Vecchi, G. A., and A. T. Wittenberg, 2010: El Nino and our future climate: Where do we stand?
1254 *Wiley Interdisciplinary Reviews: Climate Change*, **1**, 260-270. doi: 10.1002/wcc.33.
- 1255
- 1256 Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations?
1257 *Geophys. Res. Lett.*, **36**, L12702. doi: 10.1029/2009GL038710.
- 1258
- 1259 Wu, Q., and D. J. Karoly (2007): Implications of changes in the atmospheric circulation on the
1260 detection of regional surface air temperature trends, *Geophys. Res. Lett.*, **34**, L08703,
1261 doi:10.1029/2006GL028502.
- 1262

1263 Wu, Q. (2010): Associations of diurnal temperature range change with the leading climate
1264 variability modes during the Northern Hemisphere wintertime and their implication on
1265 the detection of regional climate trends, *J. Geophys. Res.*, **115**, D19101,
1266 doi:10.1029/2010JD014026.

1267

1268 Yang, X., A. Rosati, S. Zhang, T. L. Delworth, R. G. Gudgel, R. Zhang, G. Vecchi, W.
1269 Anderson, Y.-S. Chang, T. DelSole, K. Dixon, R. Msadek, W. F. Stern, A. Wittenberg,
1270 and F. Zeng, 2013: A predictable AMO-like pattern in GFDL's fully-coupled ensemble
1271 initialization and decadal forecasting system. *J. Climate*, in press. doi: 10.1175/JCLI-D-
1272 12-00231.1.

1273

1274

1275

1276

1277 **Figure Captions**

1278

1279 Fig. 1. Time series of global-mean annual-mean surface air temperature (2 m) anomalies from
1280 the CMIP3 (a, b) and CMIP5 (c, d) preindustrial control runs (black curves). Observed global
1281 mean surface temperature (HadCRUT4, combining SST and land surface air temperature
1282 anomalies) is also shown in blue on the diagrams for comparison. The blue curves labeled
1283 “Residual (HadCRUT4...” were created by subtracting the multi-model ensemble mean surface
1284 temperature (using masked SSTs and land surface air temperatures from the 20C3M All-Forcing
1285 historical runs for either CMIP3 or CMIP5) from the observed temperature. Orange straight
1286 lines (one or two segments) through the control run time series depict the long term linear drift.
1287 The long term drift over the year range shown is calculated at each grid point and then subtracted
1288 from the model control run series before performing further analysis in our study. Short vertical
1289 orange segments denote two places where specific control runs were divided into two separate
1290 segments and the linear drift computed separately for each segment. In those cases, the residuals
1291 from the drift were formed and then combined back into a single series. The various curves in the
1292 figure have been displaced vertically by arbitrary constants for visual clarity. Curves labeled
1293 with a ‘*’ denote CMIP3 models that did not include volcanic forcing in their historical runs.
1294 The number in brackets by each model name denotes how many All-Forcing ensemble members
1295 were available; when there are two numbers in brackets, the second refers to the number of
1296 Natural-Forcing ensemble members. Curves labeled with a ‘(0)’ were excluded from the
1297 remainder of our analysis due to various issues such as discontinuities in time series, short record
1298 length, or unavailable sea surface temperature data in the CMIP3 archive. Vertical axis tic mark
1299 spacing is 0.2°C.

1300

1301 Fig. 2. Standard deviation ($^{\circ}\text{C}$) of low-pass (>10 yr) filtered internal variability of surface
1302 temperature derived from CMIP3 or CMIP5 pre-industrial control runs (a, d, g), an observed
1303 estimate (Obs. St. Dev.*; b, e, h), and the difference between the control runs and the observed
1304 estimate (c, f, i). The long-term linear drifts (time periods identified by the orange line segments
1305 in Fig. 1 a,b) were removed prior to computing the control run standard deviations. The model
1306 control run results are based on the mean standard deviation of a) eight CMIP3 models that have
1307 All-Forcing runs with volcanic forcing; b) all 23 CMIP5 models; and c) seven CMIP5 models
1308 that included at least one experiment with Natural Forcing only and extending to 2010. Note
1309 that the control runs on which the figure are based do not have episodic volcanic forcing and
1310 have been masked for observed missing data periods. Therefore, the observational estimate of
1311 internal variability (Obs. St. Dev.*) is derived from observations with adjustments for variance
1312 associated with various natural or anthropogenic forcing agents. See text for details of the
1313 adjustment.

1314

1315 Fig. 3. As in Figure 2 (c, f, i) except for individual models in the (a) CMIP3 or (b) CMIP5 sets
1316 of models used in Fig. 2. The red number at upper right above each figure lists the spatial
1317 correlation of the model's low-pass filtered standard deviation field vs. the observational
1318 estimate (Obs. St. Dev.*) in Fig. 2. For the seven-model subset of CMIP5 models, the
1319 comparison is with the observations adjusted according to just those seven models and their
1320 respective All-Forcing and control runs. See text for further details.

1321

1322 Fig. 4. Time series of global mean surface temperature anomalies (combined SST and land
1323 surface air temperature) from observations (HadCRUT4; black curves) in degrees Celsius. The
1324 red curves in a-c depict the 5th and 95th percentiles of annual mean anomalies for the multi-model
1325 mean (thick) or of single model realizations (thin lines, gray stippling) for the CMIP3 (a, b) or
1326 CMIP5 (c) 20C3M historical All-Forcing runs in degrees Celsius. The mean curve is not shown
1327 but lies approximately midway between the 5th and 95th percentiles. The series in (a) are from
1328 eight CMIP3 models run with volcanic forcing. The historical runs in (b) include 19 CMIP3
1329 models with and without volcanic forcing (as identified in Fig. 1 (a,b). All of the 23 CMIP5
1330 model runs included in the computations (c) incorporated volcanic forcing. In (d) the blue
1331 curves are based on seven CMIP5 models that had Natural-Forcing-Only runs extending through
1332 2010. See text for description of how the confidence limits were computed. The time series
1333 have been re-centered so that the ensemble mean value, averaged for the years 1881-1920, is
1334 zero. Model data were masked with the observed spatially and temporally evolving missing data
1335 mask. The total number of individual experiments included in each panel was: a) 26; b) 51; c)
1336 79; and d) 25.

1337

1338 Fig. 5. Variance spectra as a function of frequency for observed global mean surface
1339 temperature (combined SST and land surface air temperature), in black with 90% confidence
1340 intervals shown by the shading, plotted against spectra for the individual (a) CMIP3 and (b)
1341 CMIP5 All-Forcing historical runs with Volcanic forcing (red) based on the time series in Fig. 4
1342 (a,c). The spectra in (c) and (d) are based on residual observed or model historical run time
1343 series, where the multi-model ensemble surface temperature from the 20C3M All-Forcing (with
1344 volcanic) historical runs (CMIP3 or CMIP5) is subtracted from the observed and from each

1345 model's global mean temperature series to form residual time series prior to computing the
1346 spectra (see text for details).

1347

1348 Fig. 6. Map illustrating averaging regions examined in Figs. 7-9. Regions abbreviations
1349 including: Euro = Europe; NAs = Northern Asia; SAs = Southern Asia; Afr = Africa; IO =
1350 Indian Ocean; Aus = Australia; TWP = Tropical western Pacific; TEP = Tropical eastern Pacific;
1351 IOWP = Tropical Indian Ocean/western Pacific warm pool; NP = North Pacific; AL = Alaska;
1352 SEUS = Southeastern United States; ConUS = Continental United States; RofUS = rest of
1353 continental United States, other than SEUS; SAmer = South America; Can = Canada; Natl =
1354 North Atlantic; SNA = Subtropical North Atlantic; TNA = Tropical North Atlantic (Main
1355 Development Region); SAtl = South Atlantic.

1356

1357 Fig. 7. Trends ($^{\circ}\text{C}/100$ yr) in area-averaged annual-mean surface temperature as a function of
1358 starting year, with all trends ending in 2010. The black curves are trends from observations
1359 (HadCRUT4), where observational uncertainty is depicted as a range showing the 5th to 95th
1360 percentile ranges of trends obtained using the 100-member HadCRUT4 ensemble. Blue curves
1361 are ensemble means for Natural-Forcing-Only runs using a subset of seven CMIP5 models that
1362 had Natural-Forcing runs to 2010. Red curves are ensemble means of the All-Forcing runs from
1363 the same seven CMIP5 models. See Fig. 6 for definitions of averaging regions. The different
1364 models are weighted equally for the multi-model ensemble means, regardless of the number of
1365 ensemble members they had. The pink shading shows the 5th to 95th percentile range of the
1366 distribution of trends obtained by combining random samples from each of the seven CMIP5

1367 model control runs together with the corresponding model's ensemble-mean forced trend (All-
1368 Forcing runs) to create a total multi-model distribution of trends that reflects uncertainty in both
1369 the forced response and the influence of internal climate variability. The blue-shaded region
1370 shows the same, but for the Natural-Forcing-Only runs. Violet shading indicates where the pink-
1371 and blue-shaded regions overlap. Gaps in the curves indicate inadequate data coverage for a
1372 trend-to-2010 for those start years. Requirements include: 33% areal coverage to define an index
1373 time series point for a month, 40% of months available for a year to be non-missing, and 20% of
1374 all years available in each of five equal segments for a time series have adequate coverage for a
1375 trend. The seven-model CMIP5 subset used here and in subsequent assessment figures that
1376 incorporate Natural-Forcing runs include: CanESM2, CNRM-CM5, CSIRO-Mk3-6-0, FGOALS-
1377 g2, HadGEM2-ES, IPSL-CM5A-LR, and NorESM1-M.

1378

1379 Fig. 8. As in Fig. 7, but for additional regions as labeled (see Fig. 6).

1380

1381 Fig. 9. As in Fig. 7, except the left column is based on All-Forcing runs from eight CMIP3
1382 models that include volcanic forcing in their historical simulations, and the eight corresponding
1383 control runs (without volcanic forcing); the middle column is based on All-Forcing and control
1384 runs from all 23 CMIP5 models; and the right column is based on All-Forcing, Natural-Forcing-
1385 Only, and control runs from the same sets of CMIP5 models as used in Figs. 7 and 8 (see Fig. 7
1386 caption).

1387

1388 Fig. 10. Geographical distribution of surface temperature trends (1901-2010) in: (a,b)
1389 HadCRUT4 observations; (c) CMIP3 eight-model ensemble mean (All-Forcing, volcanic
1390 models); d) CMIP5 seven-model ensemble mean (All-Forcing, volcanic models). Unit: degrees
1391 C per 100 yr. In (e, f) the observed trend is assessed in terms of the multi-model ensemble mean
1392 trends and variability in the historical forcing and control runs (CMIP3 and CMIP5). The
1393 different colors in (e, f) depict different categories of assessment result; the categories are listed
1394 in the legends below panels e and f. Panel (e) compares observed trends with trends from eight
1395 CMIP3 All-Forcing models and their eight control runs. Panel (f) compares observed trends
1396 with trends from the CMIP5 seven-model subset, including All-Forcing, Natural-Forcing, and
1397 control runs.

1398

1399 Fig. 11. Same as Fig. 10 but for trends from 1951 to 2010.

1400

1401 Fig. 12. Same as Fig. 10 but for trends from 1981 to 2010.

1402

1403 Fig. 13. Summary assessment of observed vs. model ensemble-mean trends-to-2010. The
1404 percent of global analyzed areas meeting certain criteria (see graph labels) are shown as a
1405 function of start year (all trends ending in 2010). a) Assessments of the eight CMIP3 (solid
1406 lines) vs. the 23 CMIP5 (dashed lines) multi-model ensemble means (historical 20C3M All-
1407 Forcing runs with volcanic forcing and associated control runs). b) Assessment of the CMIP5
1408 multi-model ensemble means and control runs using the seven-model subset of CMIP5 models
1409 (with Natural-Forcing-Only runs extending to 2010), the All-Forcing runs from the same seven

1410 models, and their seven control runs. The black curves are the sum of the red and orange curves;
1411 the sum of black + cyan + green + blue = 100%.

1412

1413 Fig. 14. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP3 multi-model
1414 (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100
1415 yr. The observed trend is assessed in terms of the eight individual CMIP3 models (trends and
1416 variability) in (c-k). Panels (c-k) show the fraction of the eight individual CMIP3 models whose
1417 historical All-Forcing runs meet the criteria listed above each panel. The criteria are: c) no
1418 detectable change; d) warming that is detectable but significantly less than simulated in the All-
1419 Forcing runs; e) warming that is detectable and consistent with the All-Forcing runs; f) warming
1420 that is detectable but significantly greater than simulated in the All-Forcing runs; g) warming
1421 that is detectable; h) warming that is detectable and either consistent with or greater than the
1422 simulated (All-Forcing) runs; i) observed and simulated trends are consistent (though the
1423 observed trend may not be detectable); j) observed and simulated internal low-frequency
1424 variability are consistent; and k) conditions for (i) and (j) are both satisfied (i.e., the simulated
1425 variability and trend are both consistent with observations). The white numbers at the bottom of
1426 maps c-k indicate the area-weighted global average of the mapped fields.

1427

1428 Figure 15. Same as Fig. 14, but for 23 CMIP5 models with volcanic forcing.

1429

1430 Fig. 16. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP5 multi-model
1431 ensemble-mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed
1432 trend is assessed in terms of trend and variability using the seven CMIP5 models that had
1433 available an All-Forcing ensemble and Natural-Forcing-Only runs extending to 2010. Panels (c-
1434 l) show the fraction of the seven individual CMIP5 models at each grid point whose All-Forcing,
1435 Natural-Forcing-Only, and control runs together meet the criteria listed above the panel. The
1436 criteria are: c) no detectable change; d) warming that is detectable (inconsistent with Natural-
1437 Forcing runs) but significantly less than simulated in the All-Forcing runs; e) attributable
1438 anthropogenic warming that is detectable (inconsistent with Natural-Forcing Only runs) and
1439 consistent with the All-Forcing runs; f) attributable anthropogenic warming that is significantly
1440 greater than simulated in the All-Forcing runs; g) warming that is detectable; h) total attributable
1441 to anthropogenic warming (i.e., sum of (e) and (f)); i) observed and simulated trends are
1442 consistent (though the observed trend may not be detectable); j) observed and simulated internal
1443 low-frequency variability are consistent; k) conditions for (i) and (j) are both satisfied (i.e., the
1444 simulated variability and trend are both consistent with observations; and l) conditions for (h)
1445 and (j) are both satisfied (i.e., there is attributable anthropogenic warming and low-frequency
1446 variance is consistent with observations).

1447

1448 Fig. 17. Individual CMIP3 (a, b) and CMIP5 (c-f) models are assessed for consistency with
1449 detectable observed surface temperature trends-to-2010 (a-d), for attributable anthropogenic
1450 trends (e, f), and for consistency of both simulated trend and internal variability with observed
1451 estimates (b, d, f). Trend results are shown for start years from 1901 to 1991 (all trends ending
1452 in 2010). Plotted is the percent of analyzed global area where each individual model's (see

1453 legends) multi-realization ensemble mean forced trend and internal variability meet the criteria
1454 listed above the panel. The trends are analyzed at each grid point where there is sufficient
1455 temporal data coverage for the trend in question (see text). Note that panels (e, f) include areas
1456 where the observed trend is detectable and either consistent with or greater than simulated,
1457 whereas panels (c, d) include only areas with observed trends that are detectable and consistent
1458 with simulations.

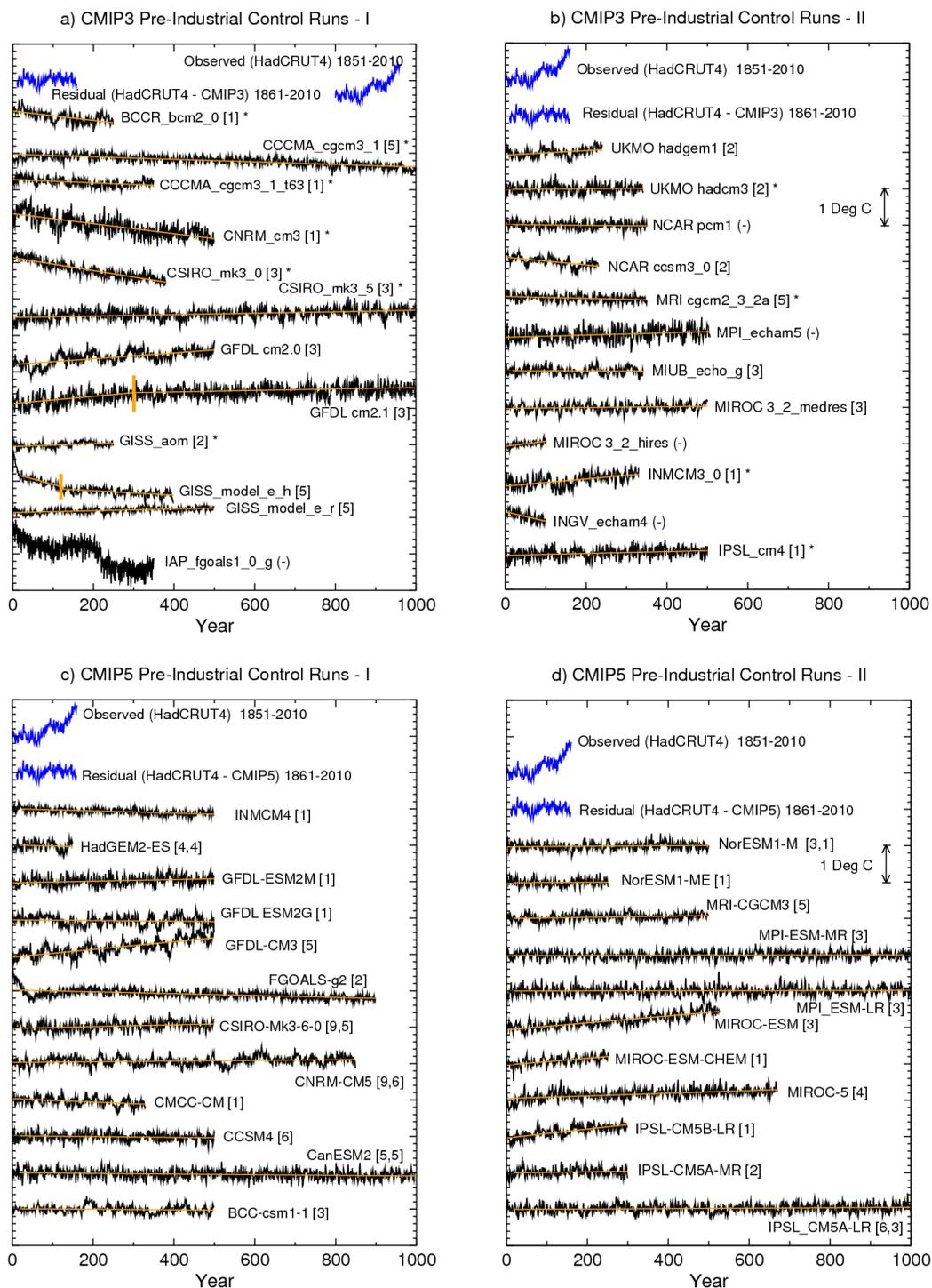
1459

1460

Fig. 1

1461

Model control runs: simulated internal variability of global temperature



1462 Fig. 1. Time series of global-mean annual-mean surface air temperature (2 m) anomalies from
1463 the CMIP3 (a, b) and CMIP5 (c, d) preindustrial control runs (black curves). Observed global
1464 mean surface temperature (HadCRUT4, combining SST and land surface air temperature
1465 anomalies) is also shown in blue on the diagrams for comparison. The blue curves labeled
1466 “Residual (HadCRUT4...)” were created by subtracting the multi-model ensemble mean surface
1467 temperature (using masked SSTs and land surface air temperatures from the 20C3M All-Forcing
1468 historical runs for either CMIP3 or CMIP5) from the observed temperature. Orange straight
1469 lines (one or two segments) through the control run time series depict the long term linear drift.
1470 The long term drift over the year range shown is calculated at each grid point and then subtracted
1471 from the model control run series before performing further analysis in our study. Short vertical
1472 orange segments denote two places where specific control runs were divided into two separate
1473 segments and the linear drift computed separately for each segment. In those cases, the residuals
1474 from the drift were formed and then combined back into a single series. The various curves in the
1475 figure have been displaced vertically by arbitrary constants for visual clarity. Curves labeled
1476 with a ‘*’ denote CMIP3 models that did not include volcanic forcing in their historical runs.
1477 The number in brackets by each model name denotes how many All-Forcing ensemble members
1478 were available; when there are two numbers in brackets, the second refers to the number of
1479 Natural-Forcing ensemble members. Curves labeled with a ‘(0)’ were excluded from the
1480 remainder of our analysis due to various issues such as discontinuities in time series, short record
1481 length, or unavailable sea surface temperature data in the CMIP3 archive. Vertical axis tic mark
1482 spacing is 0.2°C.

1483

1484

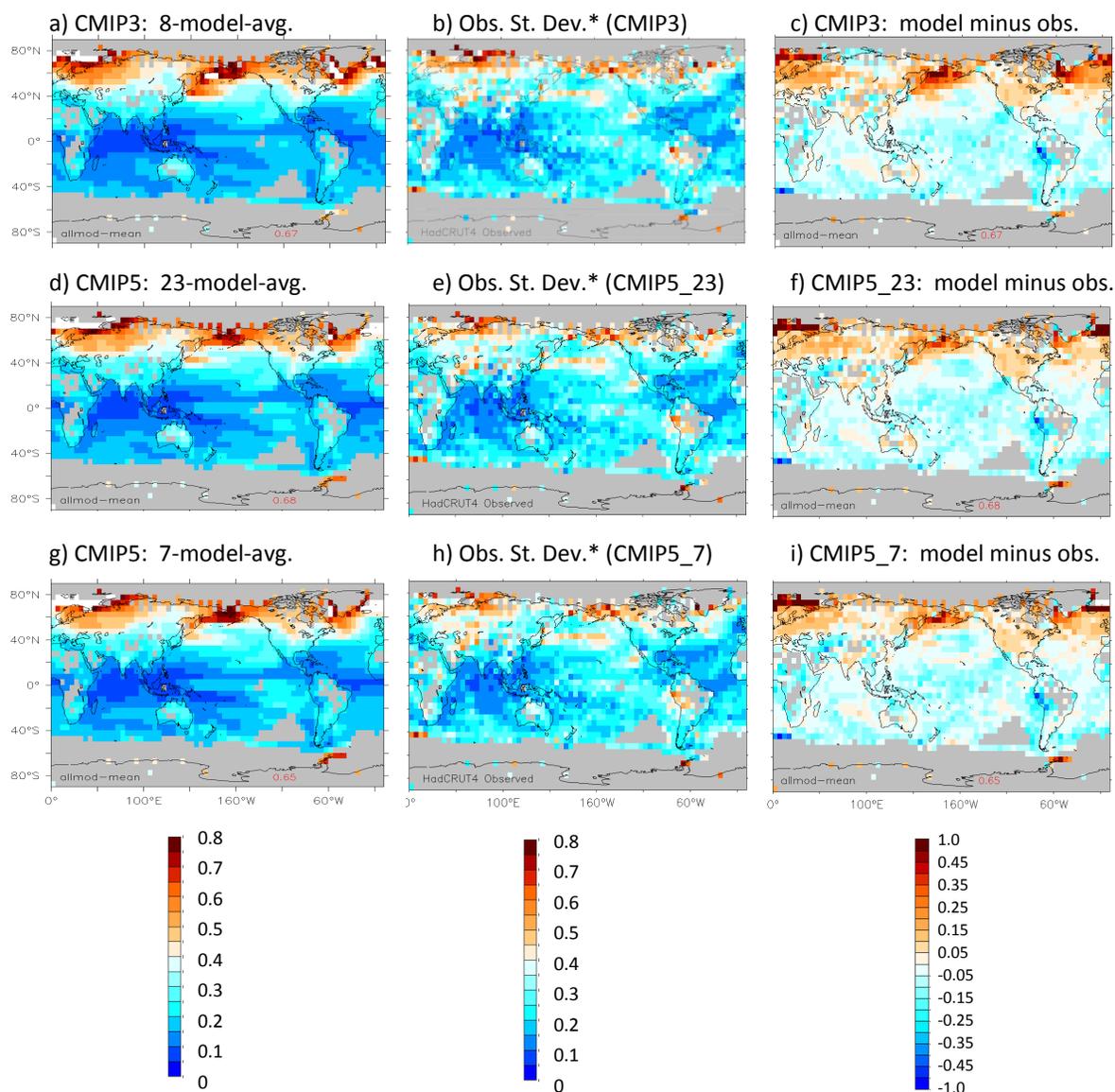


Fig. 2. Standard deviation ($^{\circ}\text{C}$) of low-pass (>10 yr) filtered internal variability of surface temperature derived from CMIP3 or CMIP5 pre-industrial control runs (a, d, g), an observed estimate (Obs. St. Dev.*; b, e, h), and the difference between the control runs and the observed estimate (c, f, i). The long-term linear drifts (time periods identified by the orange line segments in Fig. 1 a,b) were removed prior to computing the control run standard deviations. The model control run results are based on the mean standard deviation of a) eight CMIP3 models that have All-Forcing runs with volcanic forcing; b) all 23 CMIP5 models; and c) seven CMIP5 models that included at least one experiment with Natural Forcing only and extending to 2010. Note that the control runs on which the figure are based do not have episodic volcanic forcing and have been masked for observed missing data periods. Therefore, the observational estimate of internal variability (Obs. St. Dev.*) is derived from observations with adjustments for variance associated with various natural or anthropogenic forcing agents. See text for details of the adjustment.

1485

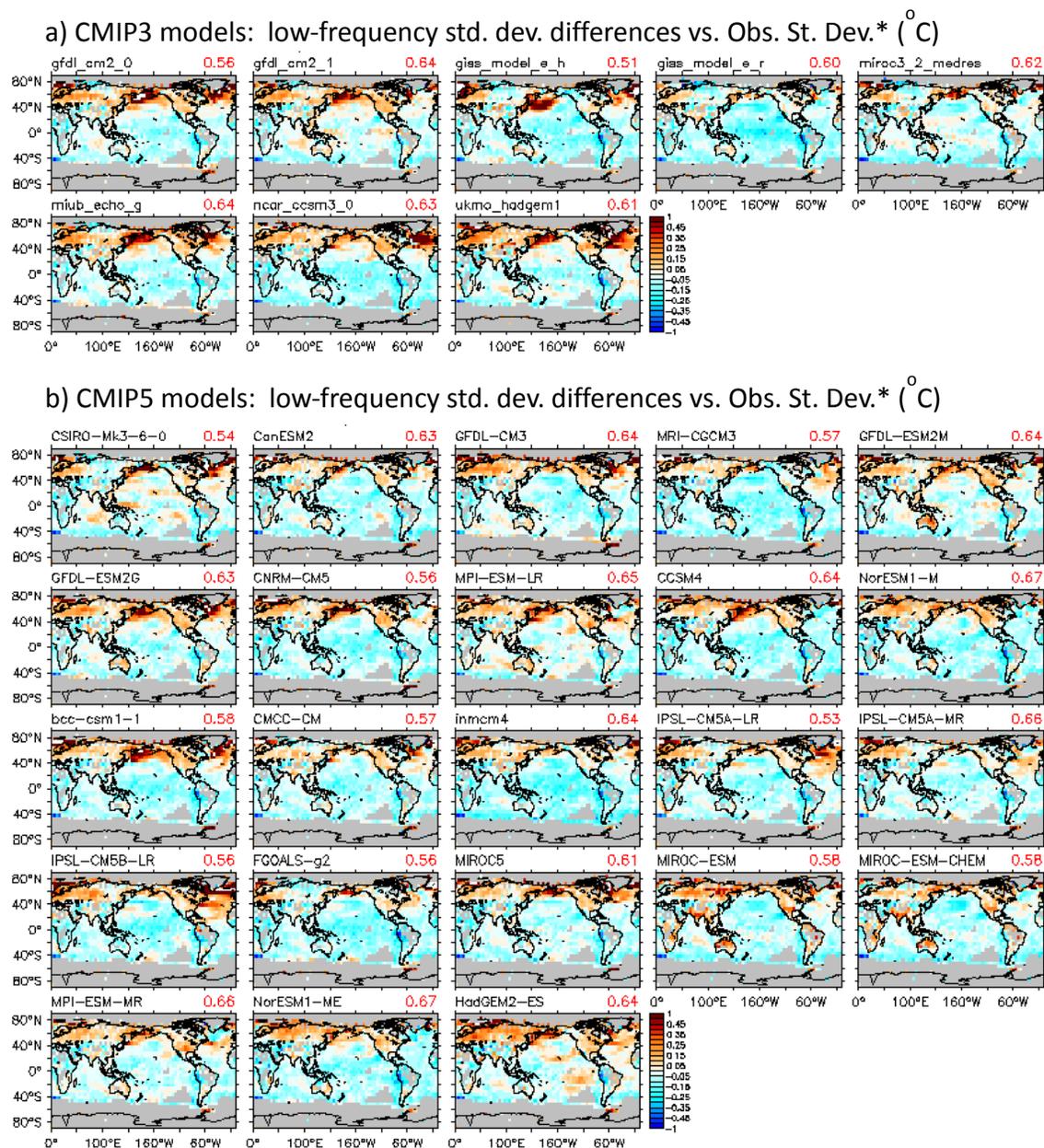


Fig. 3. As in Figure 2 (c, f, i) except for individual models in the (a) CMIP3 or (b) CMIP5 sets of models used in Fig. 2. The red number at upper right above each figure lists the spatial correlation of the model's low-pass filtered standard deviation field vs. the observational estimate (Obs. St. Dev.*) in Fig. 2. For the seven-model subset of CMIP5 models, the comparison is with the observations adjusted according to just those seven models and their respective All-Forcing and control runs. See text for further details.

1486

1487

Global Mean Surface Temperature Anomalies

1488

1489

1490

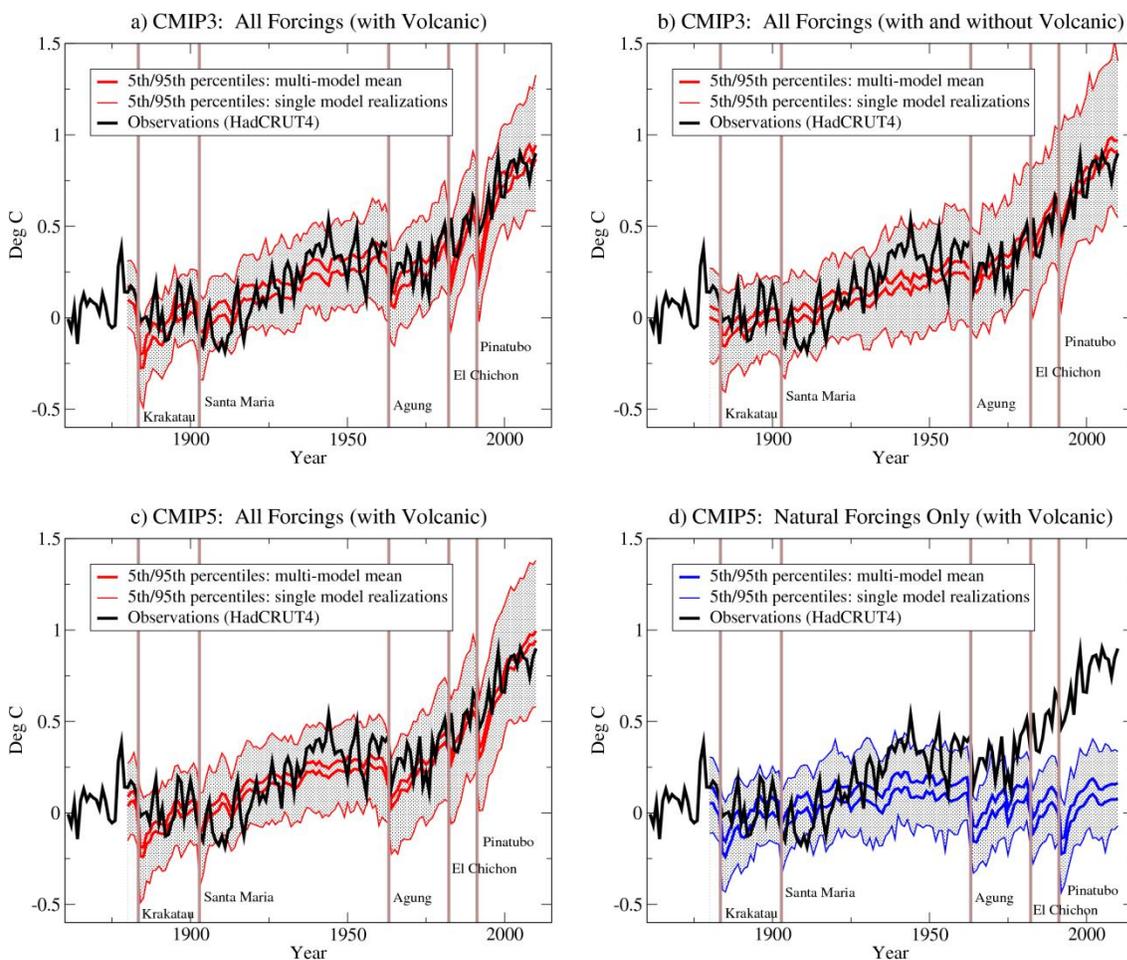


Fig. 4. Time series of global mean surface temperature anomalies (combined SST and land surface air temperature) from observations (HadCRUT4; black curves) in degrees Celsius. The red curves in a-c depict the 5th and 95th percentiles of annual mean anomalies for the multi-model mean (thick) or of single model realizations (thin lines, gray stippling) for the CMIP3 (a, b) or CMIP5 (c) 20C3M historical All-Forcing runs in degrees Celsius. The mean curve is not shown but lies approximately midway between the 5th and 95th percentiles. The series in (a) are from eight CMIP3 models run with volcanic forcing. The historical runs in (b) include 19 CMIP3 models with and without volcanic forcing (as identified in Fig. 1 (a,b)). All of the 23 CMIP5 model runs included in the computations (c) incorporated volcanic forcing. In (d) the blue curves are based on seven CMIP5 models that had Natural-Forcing-Only runs extending through 2010. See text for description of how the confidence limits were computed. The time series have been re-centered so that the ensemble mean value, averaged for the years 1881-1920, is zero. Model data were masked with the observed spatially and temporally evolving missing data mask. The total number of individual experiments included in each panel was: a) 26; b) 51; c) 79; and d) 25.

1491

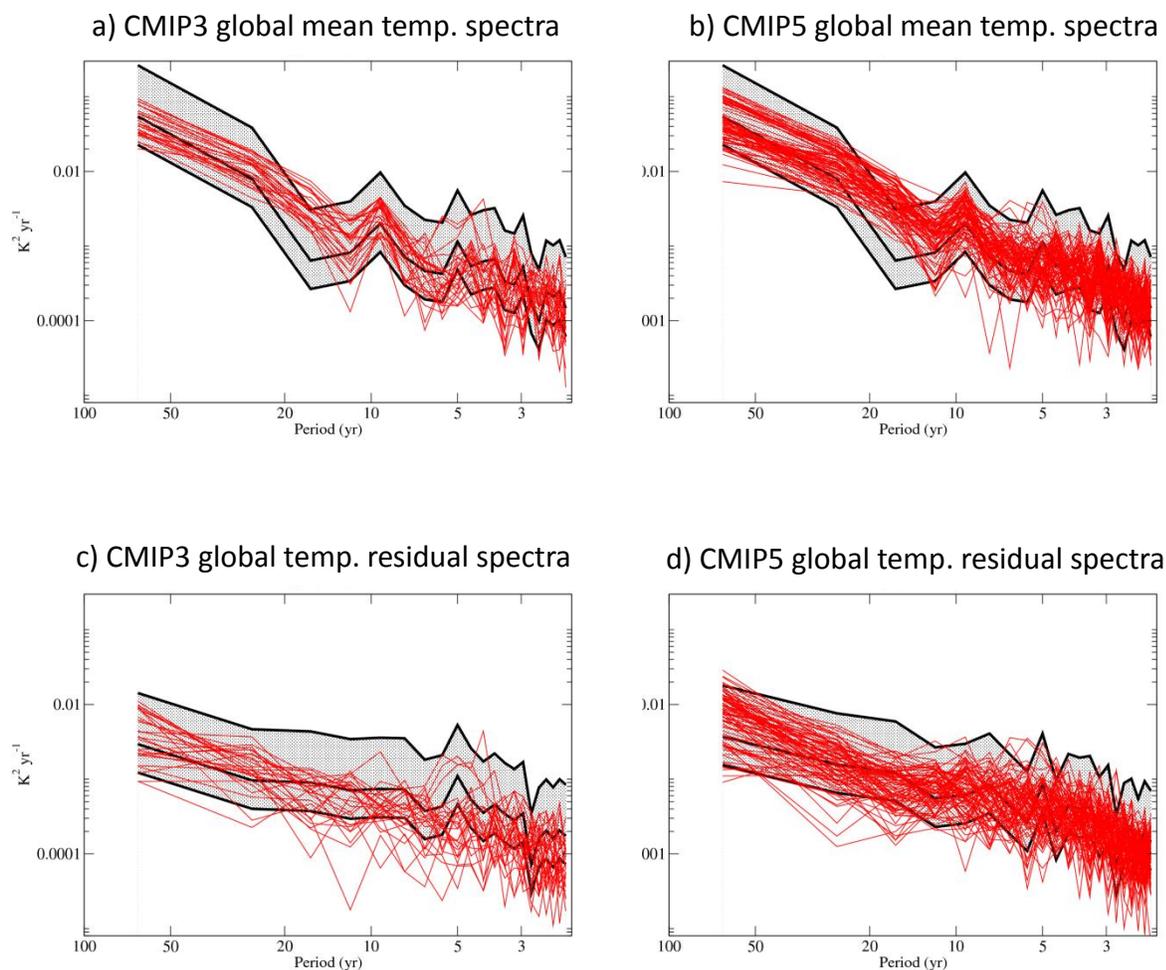
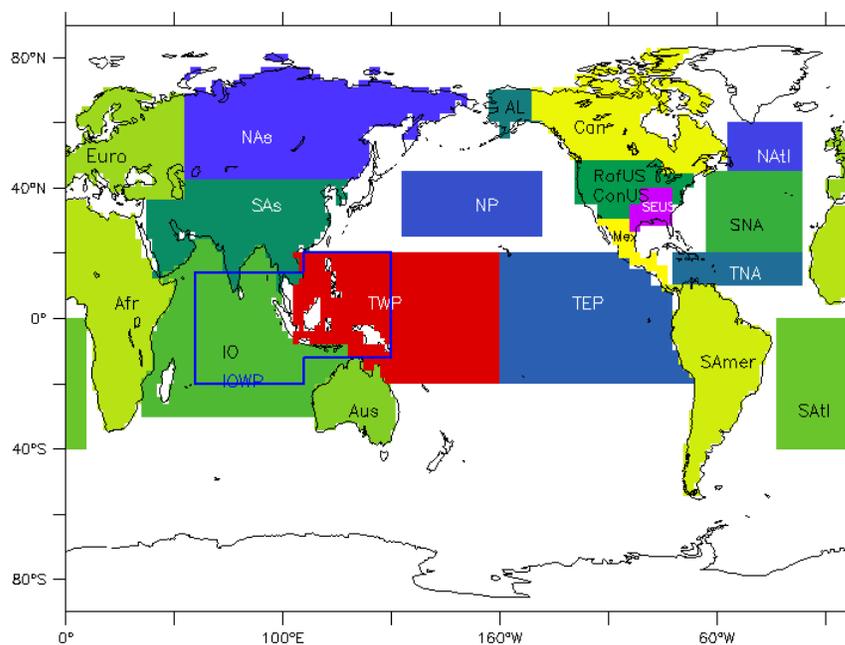


Fig. 5. Variance spectra as a function of frequency for observed global mean surface temperature (combined SST and land surface air temperature), in black with 90% confidence intervals shown by the shading, plotted against spectra for the individual (a) CMIP3 and (b) CMIP5 All-Forcing historical runs with Volcanic forcing (red) based on the time series in Fig. 4 (a,c). The spectra in (c) and (d) are based on residual observed or model historical run time series, where the multi-model ensemble surface temperature from the 20C3M All-Forcing (with volcanic) historical runs (CMIP3 or CMIP5) is subtracted from the observed and from each model's global mean temperature series to form residual time series prior to computing the spectra (see text for details).

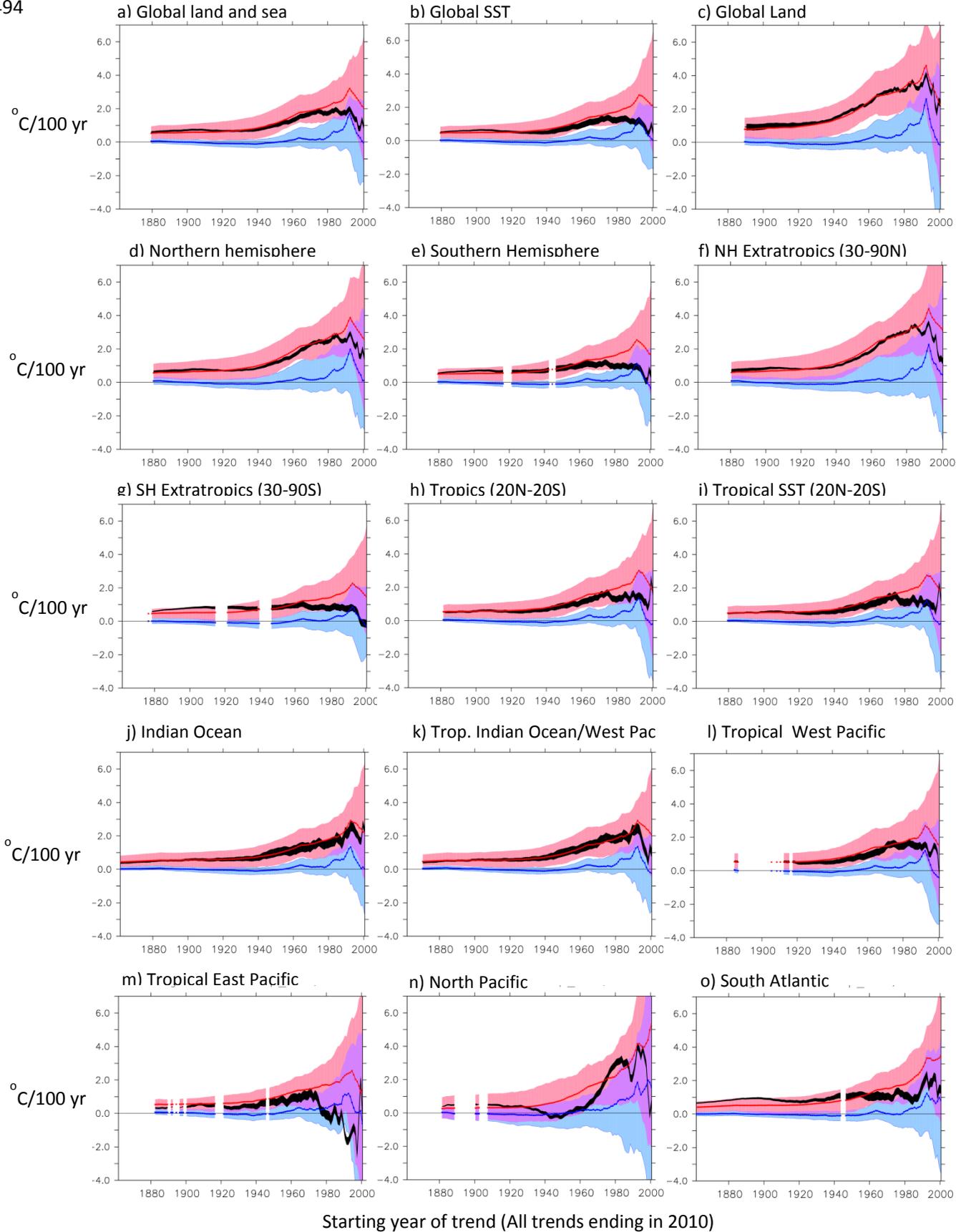
1492



1493

Fig. 6. Map illustrating averaging regions examined in Figs. 7-9. Regions abbreviations including: Euro = Europe; NAs = Northern Asia; SAs = Southern Asia; Afr = Africa; IO = Indian Ocean; Aus = Australia; TWP = Tropical western Pacific; TEP = Tropical eastern Pacific; IOWP = Tropical Indian Ocean/western Pacific warm pool; NP = North Pacific; AL = Alaska; SEUS = Southeastern United States; ConUS = Continental United States; RofUS = rest of continental United States, other than SEUS; SAmer = South America; Can = Canada; NATl = North Atlantic; SNA = Subtropical North Atlantic; TNA = Tropical North Atlantic (Main Development Region); SATl = South Atlantic.

1494



1495 Fig. 7. Trends ($^{\circ}\text{C}/100\text{ yr}$) in area-averaged annual-mean surface temperature as a function of
1496 starting year, with all trends ending in 2010. The black curves are trends from observations
1497 (HadCRUT4), where observational uncertainty is depicted as a range showing the 5th to 95th
1498 percentile ranges of trends obtained using the 100-member HadCRUT4 ensemble. Blue curves
1499 are ensemble means for Natural-Forcing-Only runs using a subset of seven CMIP5 models that
1500 had Natural-Forcing runs to 2010. Red curves are ensemble means of the All-Forcing runs from
1501 the same seven CMIP5 models. See Fig. 6 for definitions of averaging regions. The different
1502 models are weighted equally for the multi-model ensemble means, regardless of the number of
1503 ensemble members they had. The pink shading shows the 5th to 95th percentile range of the
1504 distribution of trends obtained by combining random samples from each of the seven CMIP5
1505 model control runs together with the corresponding model's ensemble-mean forced trend (All-
1506 Forcing runs) to create a total multi-model distribution of trends that reflects uncertainty in both
1507 the forced response and the influence of internal climate variability. The blue-shaded region
1508 shows the same, but for the Natural-Forcing-Only runs. Violet shading indicates where the pink-
1509 and blue-shaded regions overlap. Gaps in the curves indicate inadequate data coverage for a
1510 trend-to-2010 for those start years. Requirements include: 33% areal coverage to define an index
1511 time series point for a month, 40% of months available for a year to be non-missing, and 20% of
1512 all years available in each of five equal segments for a time series have adequate coverage for a
1513 trend. The seven-model CMIP5 subset used here and in subsequent assessment figures that
1514 incorporate Natural-Forcing runs include: CanESM2, CNRM-CM5, CSIRO-Mk3-6-0, FGOALS-
1515 g2, HadGEM2-ES, IPSL-CM5A-LR, and NorESM1-M.

1516

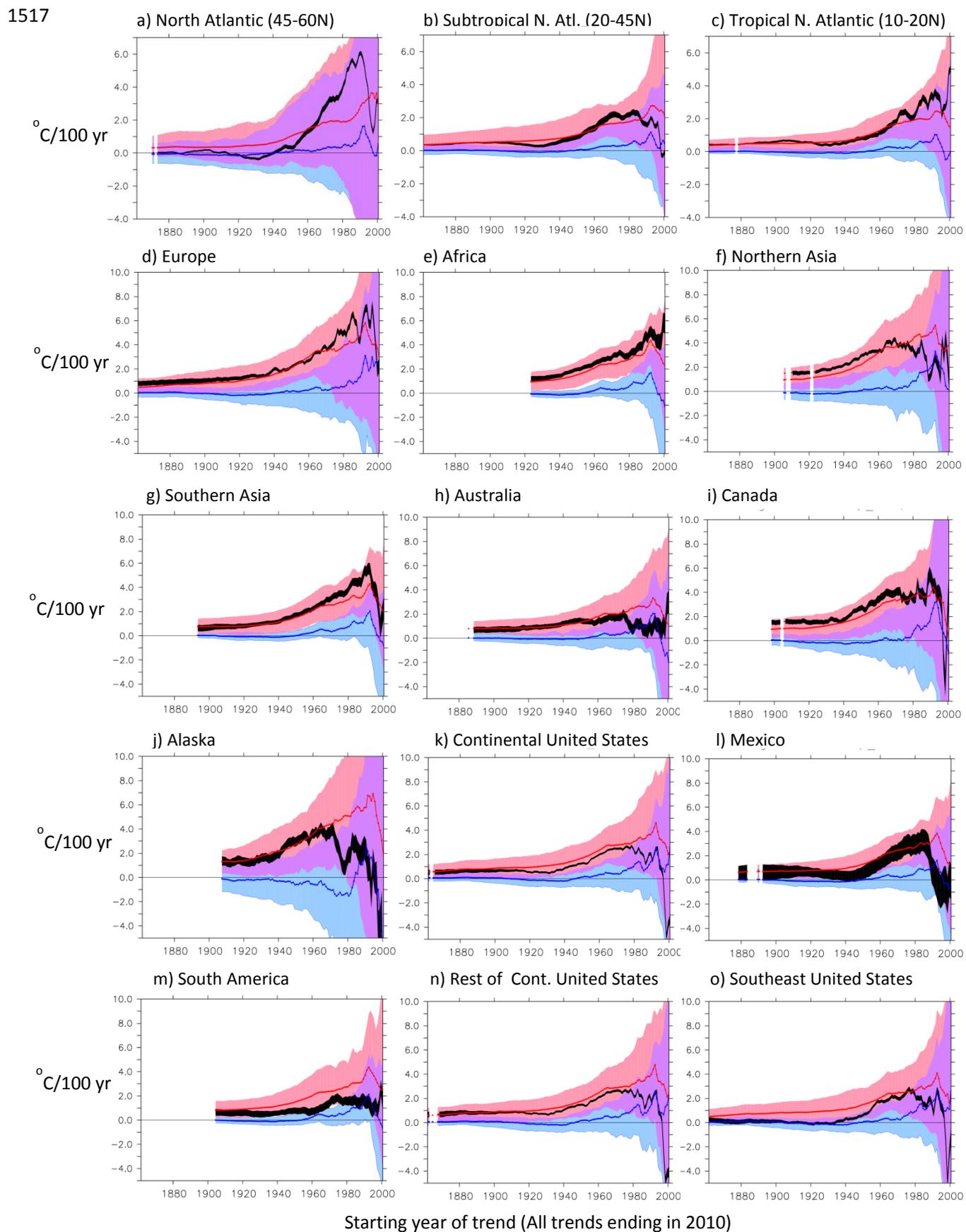
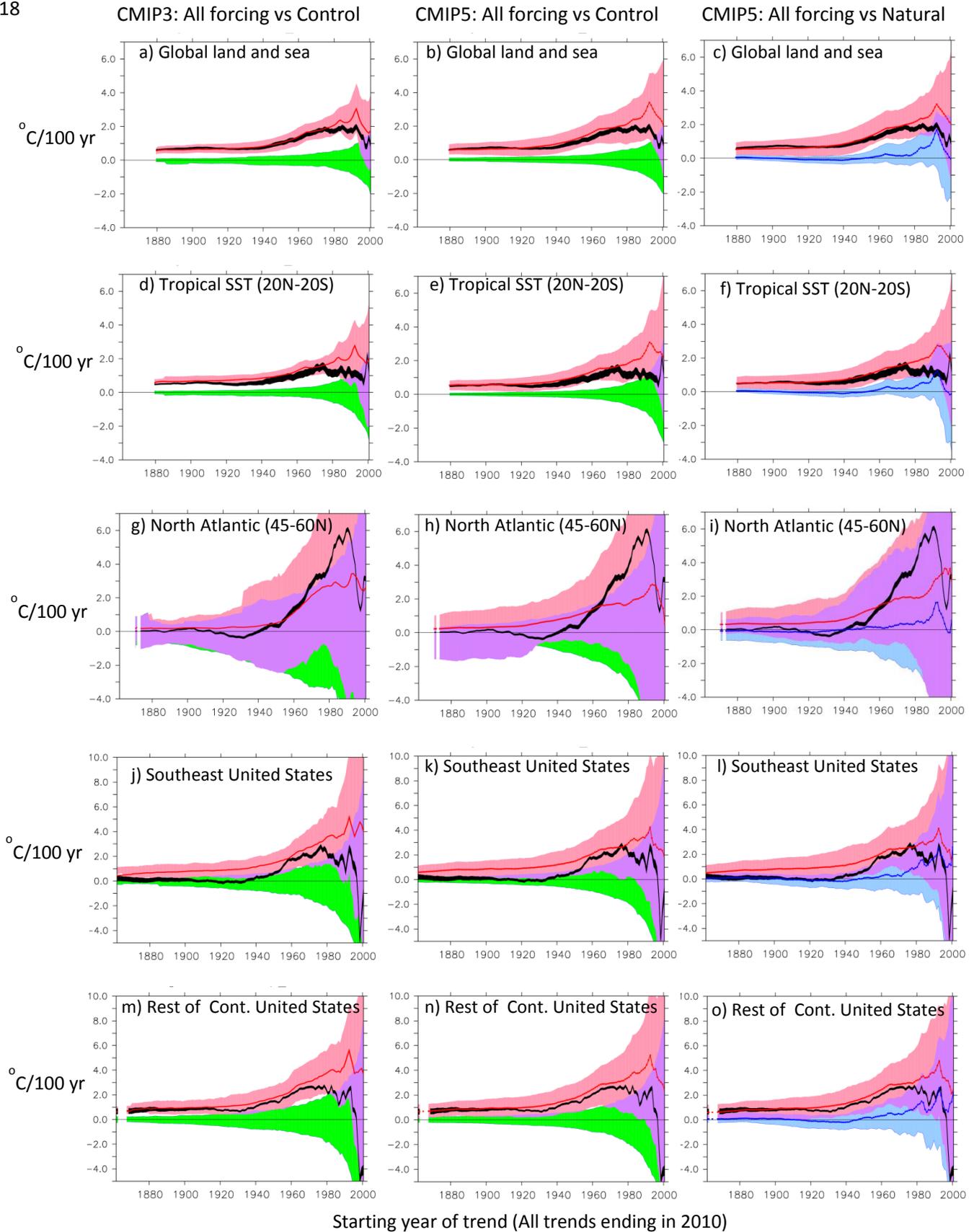


Fig. 8. As in Fig. 7, but for additional regions as labeled (see Fig. 6).

1518



1519

Fig. 9. As in Fig. 7, except the left column is based on All-Forcing runs from eight CMIP3 models that include volcanic forcing in their historical simulations, and the eight corresponding control runs (without volcanic forcing); the middle column is based on All-Forcing and control runs from all 23 CMIP5 models; and the right column is based on All-Forcing, Natural-Forcing-Only, and control runs from the same sets of CMIP5 models as used in Figs. 7 and 8 (see Fig. 7 caption).

1520

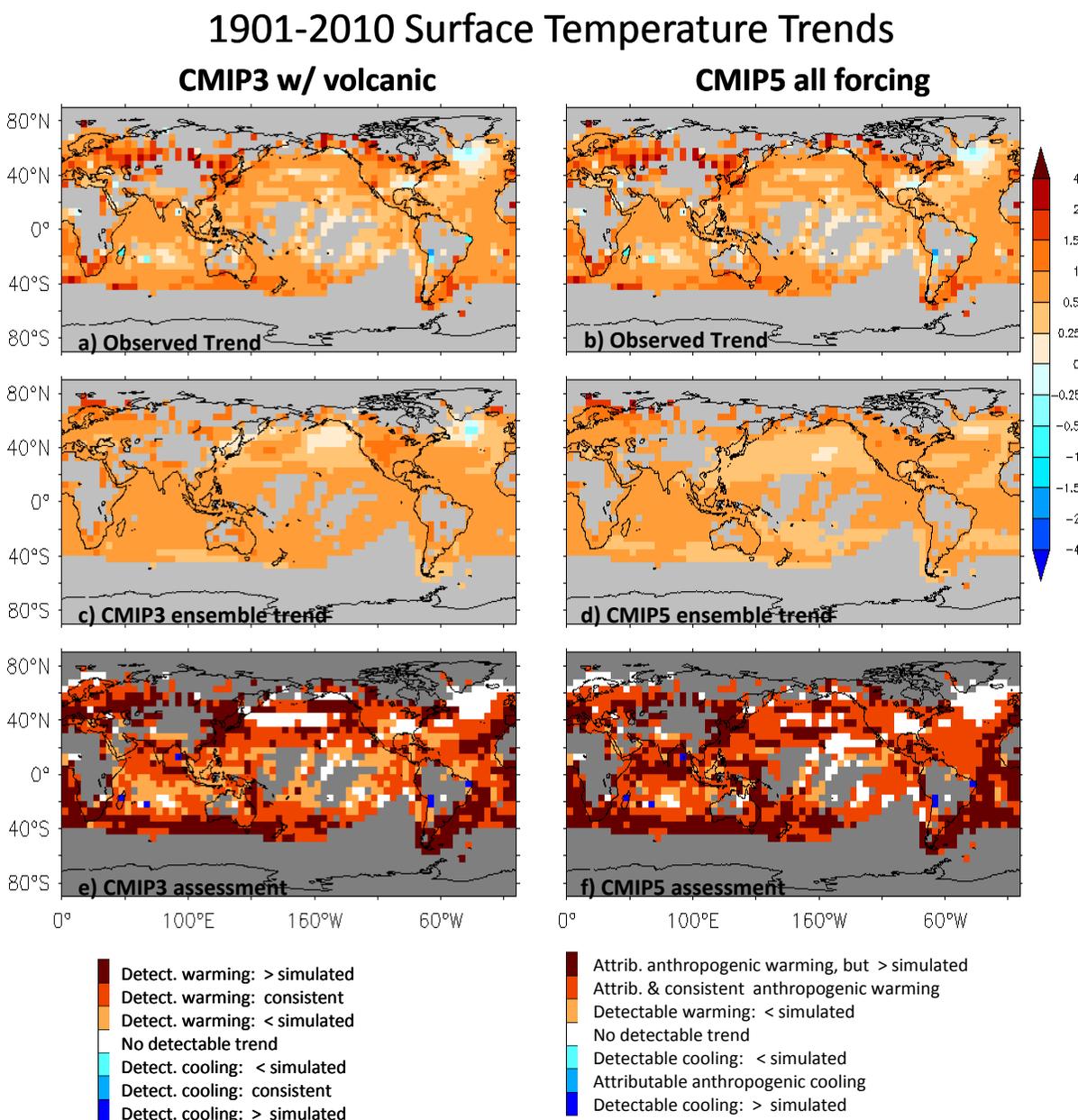


Fig. 10. Geographical distribution of surface temperature trends (1901-2010) in: (a,b) HadCRUT4 observations; (c) CMIP3 eight-model ensemble mean (All-Forcing, volcanic models); (d) CMIP5 seven-model ensemble mean (All-Forcing, volcanic models). Unit: degrees C per 100 yr. In (e, f) the observed trend is assessed in terms of the multi-model ensemble mean trends and variability in the historical forcing and control runs (CMIP3 and CMIP5). The different colors in (e, f) depict different categories of assessment result; the categories are listed in the legends below panels e and f. Panel (e) compares observed trends with trends from eight CMIP3 All-Forcing models and their eight control runs. Panel (f) compares observed trends with trends from the CMIP5 seven-model subset, including All-Forcing, Natural-Forcing, and control runs.

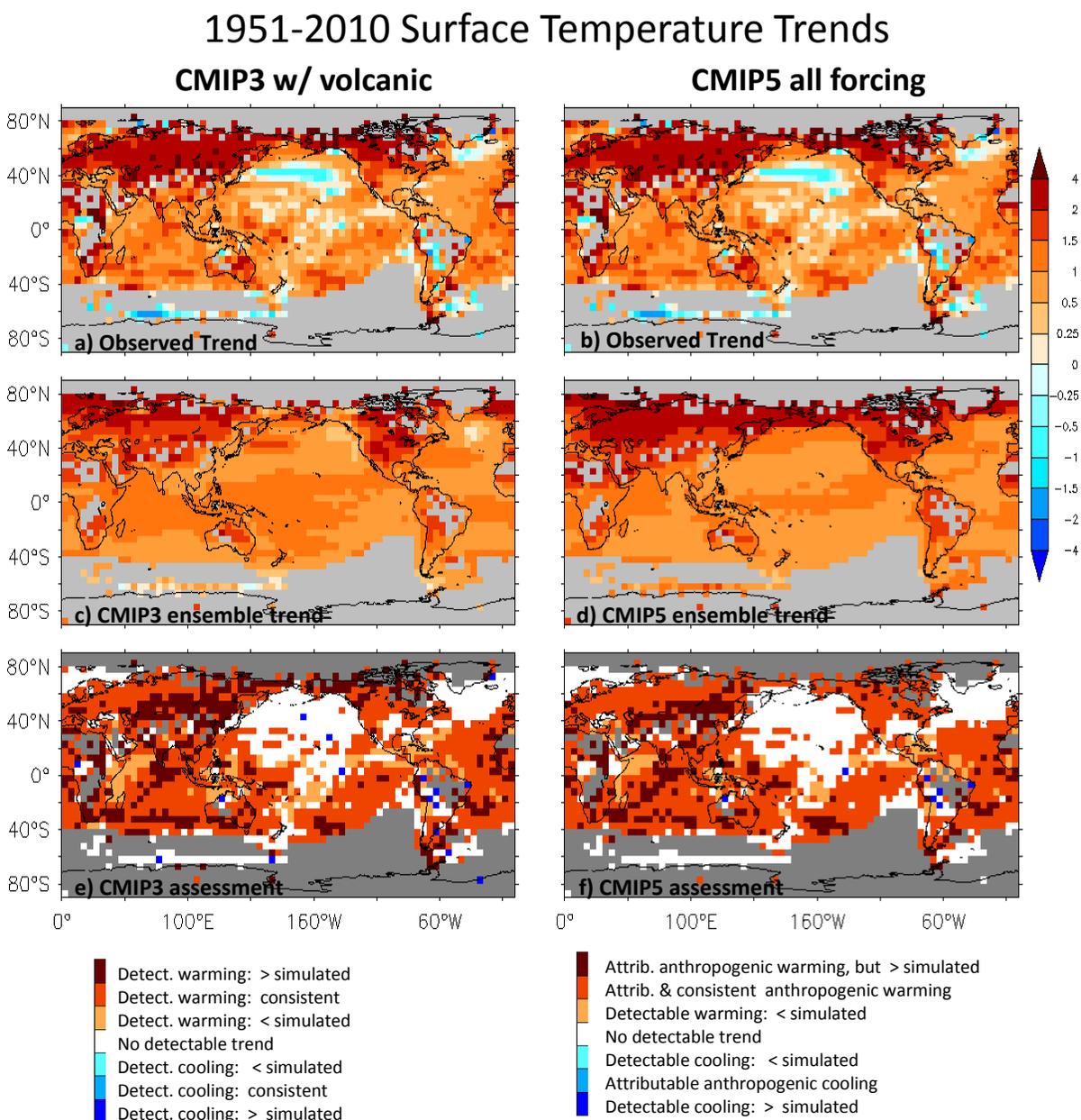


Fig. 11. Same as Fig. 10 but for trends from 1951 to 2010.

1522

1981-2010 Surface Temperature Trends

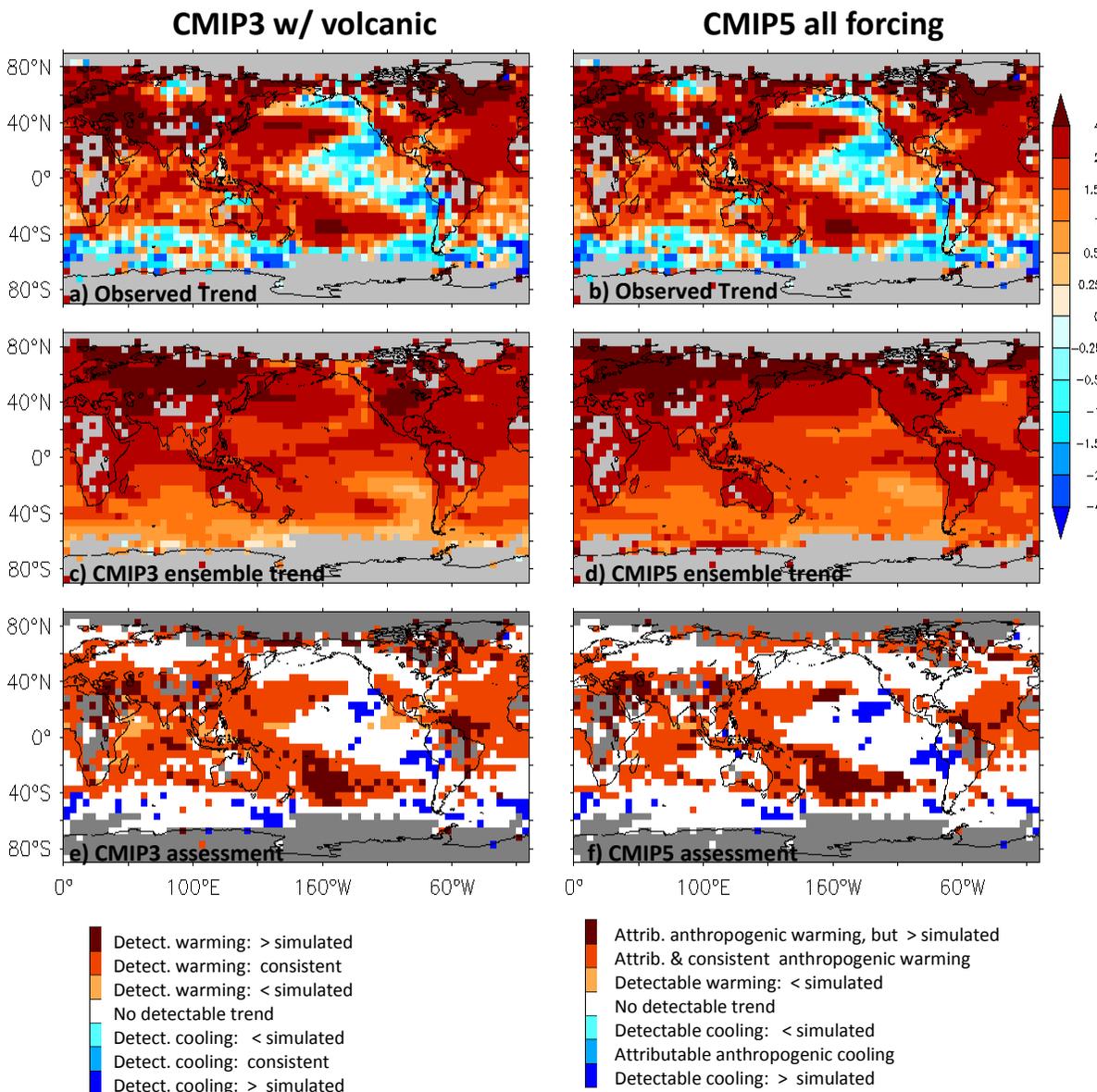
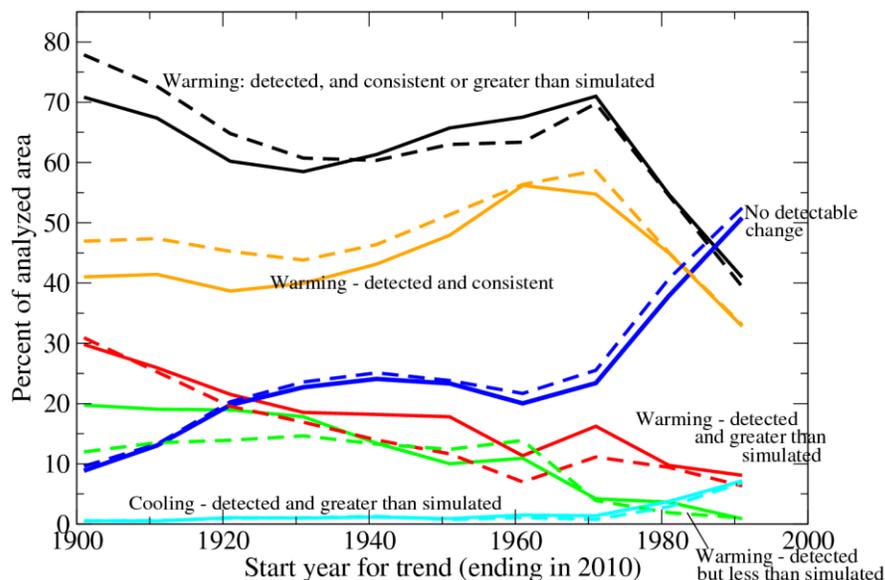


Fig. 12. Same as Fig. 10 but for trends from 1981 to 2010.

1523

a) Assessment of CMIP3 (solid) vs. CMIP5 (dashed) Multi-model Ensemble Means
 CMIP3: 8 All-Forcing models and Control runs; CMIP5: 23 All-Forcing models and Control runs



b) Assessment of CMIP5 Natural vs. All-Forcing Multi-model Ensemble Means
 Seven All-Forcing models; seven Natural Forcing models; seven Control runs

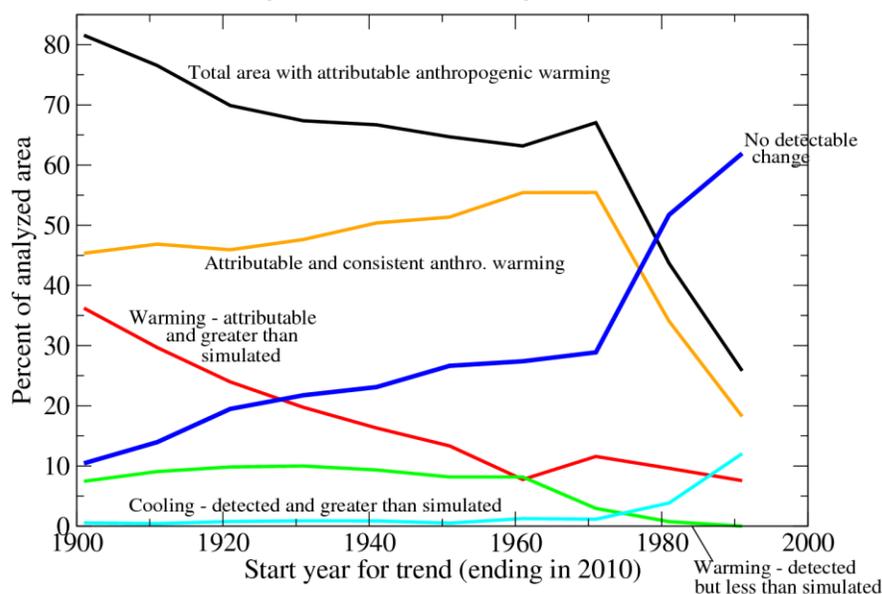


Fig. 13. Summary assessment of observed vs. model ensemble-mean trends-to-2010. The percent of global analyzed areas meeting certain criteria (see graph labels) are shown as a function of start year (all trends ending in 2010). a) Assessments of the eight CMIP3 (solid lines) vs. the 23 CMIP5 (dashed lines) multi-model ensemble means (historical 20C3M All-Forcing runs with volcanic forcing and associated control runs). b) Assessment of the CMIP5 multi-model ensemble means and control runs using the seven-model subset of CMIP5 models (with Natural-Forcing-Only runs extending to 2010), the All-Forcing runs from the same seven models, and their seven control runs. The black curves are the sum of the red and orange curves; the sum of black + cyan + green + blue = 100%.

1524

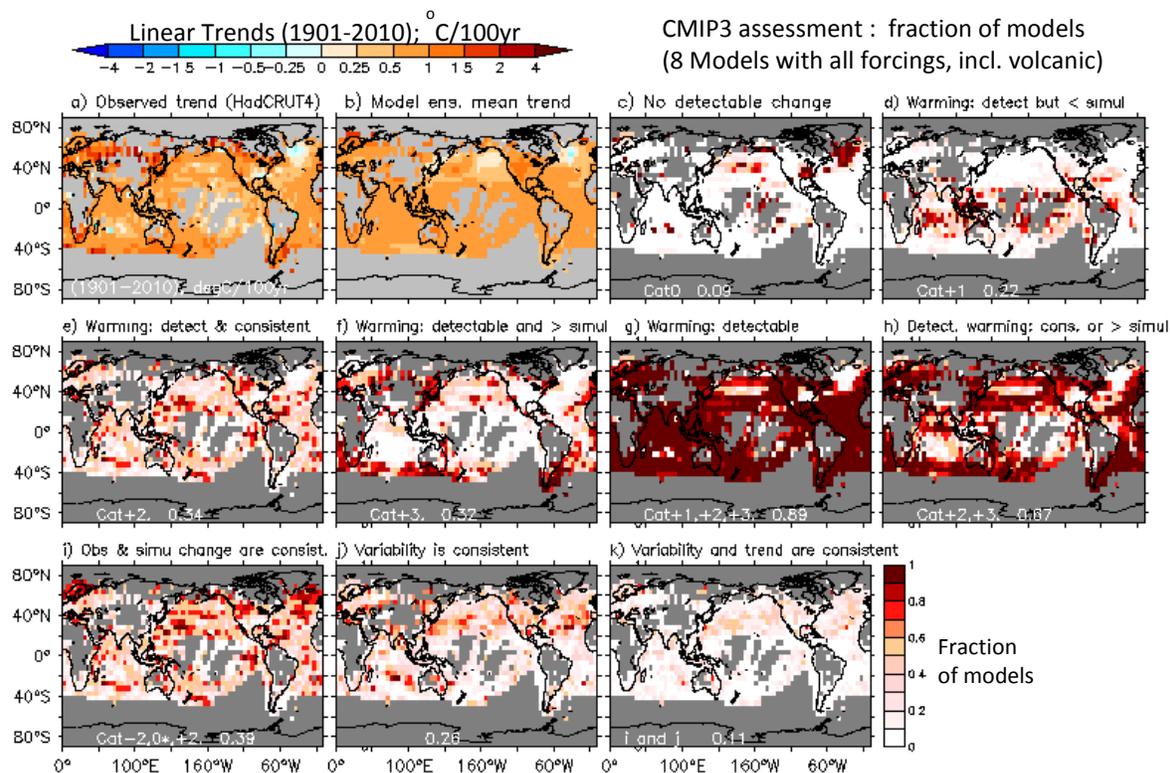


Fig. 14. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP3 multi-model (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed trend is assessed in terms of the eight individual CMIP3 models (trends and variability) in (c-k). Panels (c-k) show the fraction of the eight individual CMIP3 models whose historical All-Forcing runs meet the criteria listed above each panel. The criteria are: c) no detectable change; d) warming that is detectable but significantly less than simulated in the All-Forcing runs; e) warming that is detectable and consistent with the All-Forcing runs; f) warming that is detectable but significantly greater than simulated in the All-Forcing runs; g) warming that is detectable; h) warming that is detectable and either consistent with or greater than the simulated (All-Forcing) runs; i) observed and simulated trends are consistent (though the observed trend may not be detectable); j) observed and simulated internal low-frequency variability are consistent; and k) conditions for (i) and (j) are both satisfied (i.e., the simulated variability and trend are both consistent with observations). The white numbers at the bottom of maps c-k indicate the area-weighted global average of the mapped fields.

1525

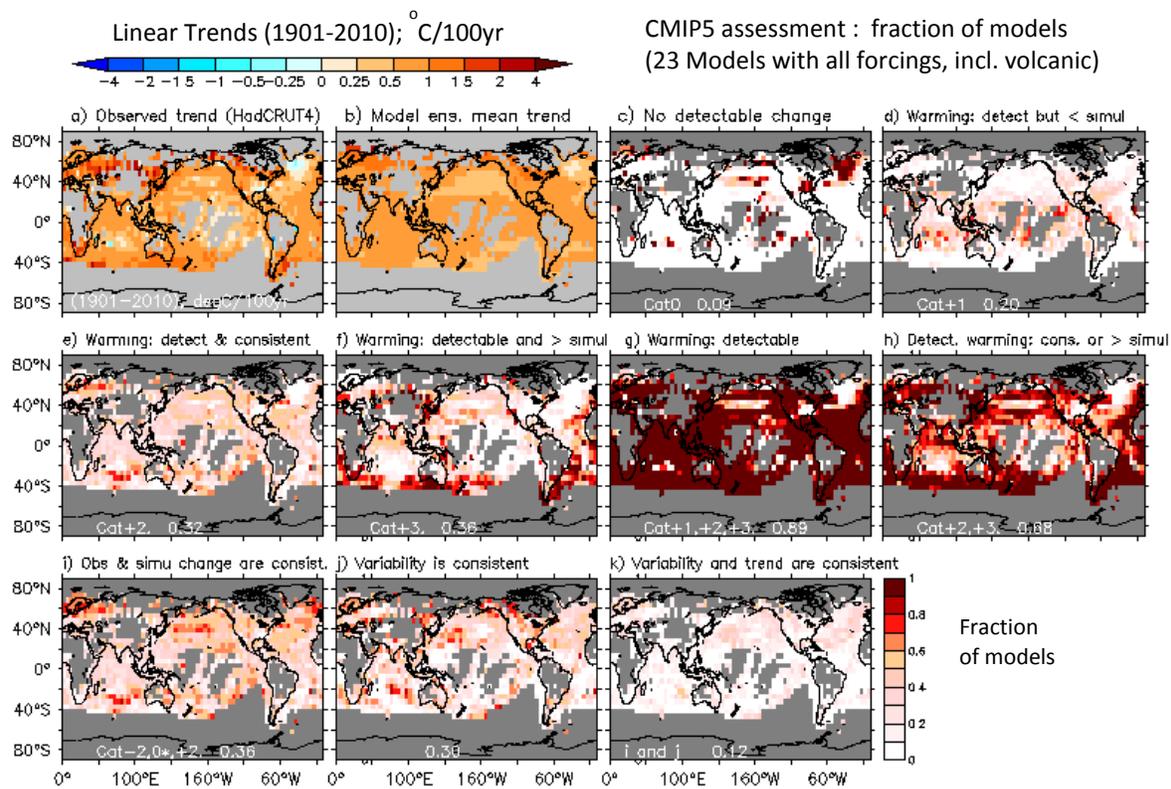


Figure 15. Same as Fig. 14, but for 23 CMIP5 models with volcanic forcing.

1526

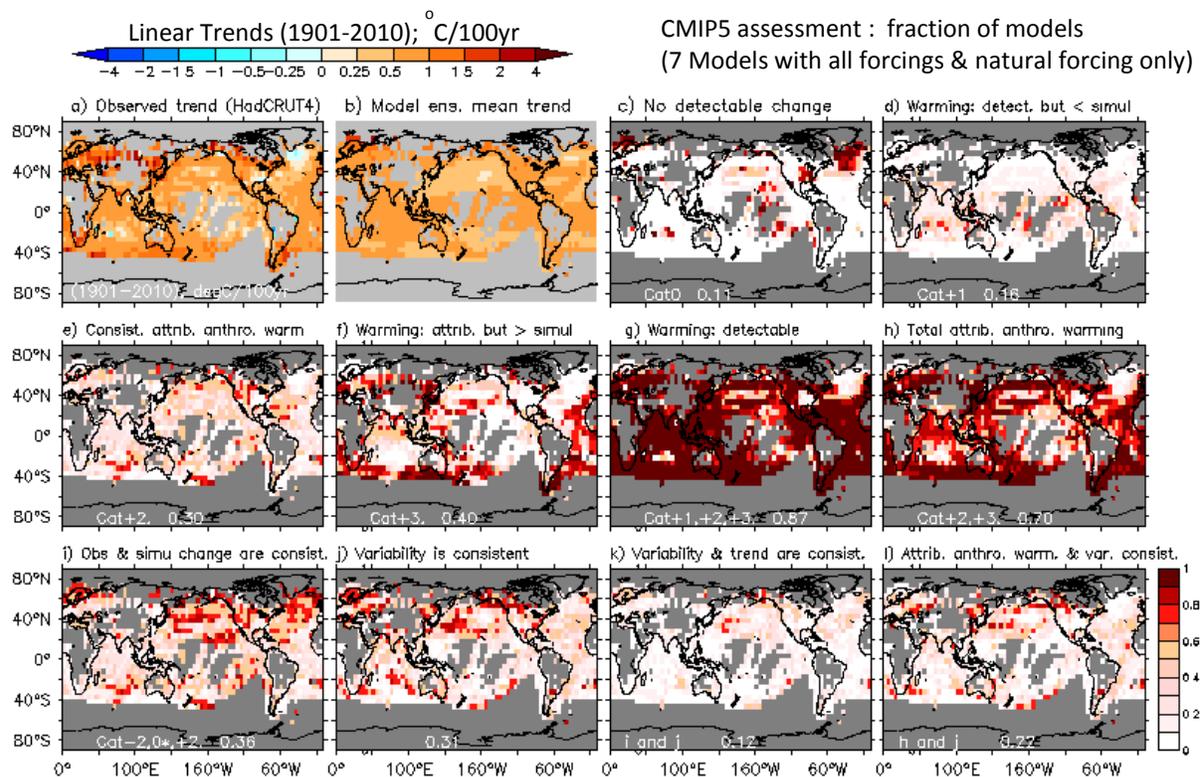


Fig. 16. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP5 multi-model ensemble-mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed trend is assessed in terms of trend and variability using the seven CMIP5 models that had available an All-Forcing ensemble and Natural-Forcing-Only runs extending to 2010. Panels (c-l) show the fraction of the seven individual CMIP5 models at each grid point whose All-Forcing, Natural-Forcing-Only, and control runs together meet the criteria listed above the panel. The criteria are: c) no detectable change; d) warming that is detectable (inconsistent with Natural-Forcing runs) but significantly less than simulated in the All-Forcing runs; e) attributable anthropogenic warming that is detectable (inconsistent with Natural-Forcing Only runs) and consistent with the All-Forcing runs; f) attributable anthropogenic warming that is significantly greater than simulated in the All-Forcing runs; g) warming that is detectable; h) total attributable to anthropogenic warming (i.e., sum of (e) and (f)); i) observed and simulated trends are consistent (though the observed trend may not be detectable); j) observed and simulated internal low-frequency variability are consistent; k) conditions for (i) and (j) are both satisfied (i.e., the simulated variability and trend are both consistent with observations; and l) conditions for (h) and (j) are both satisfied (i.e., there is attributable anthropogenic warming and low-frequency variance is consistent with observations).

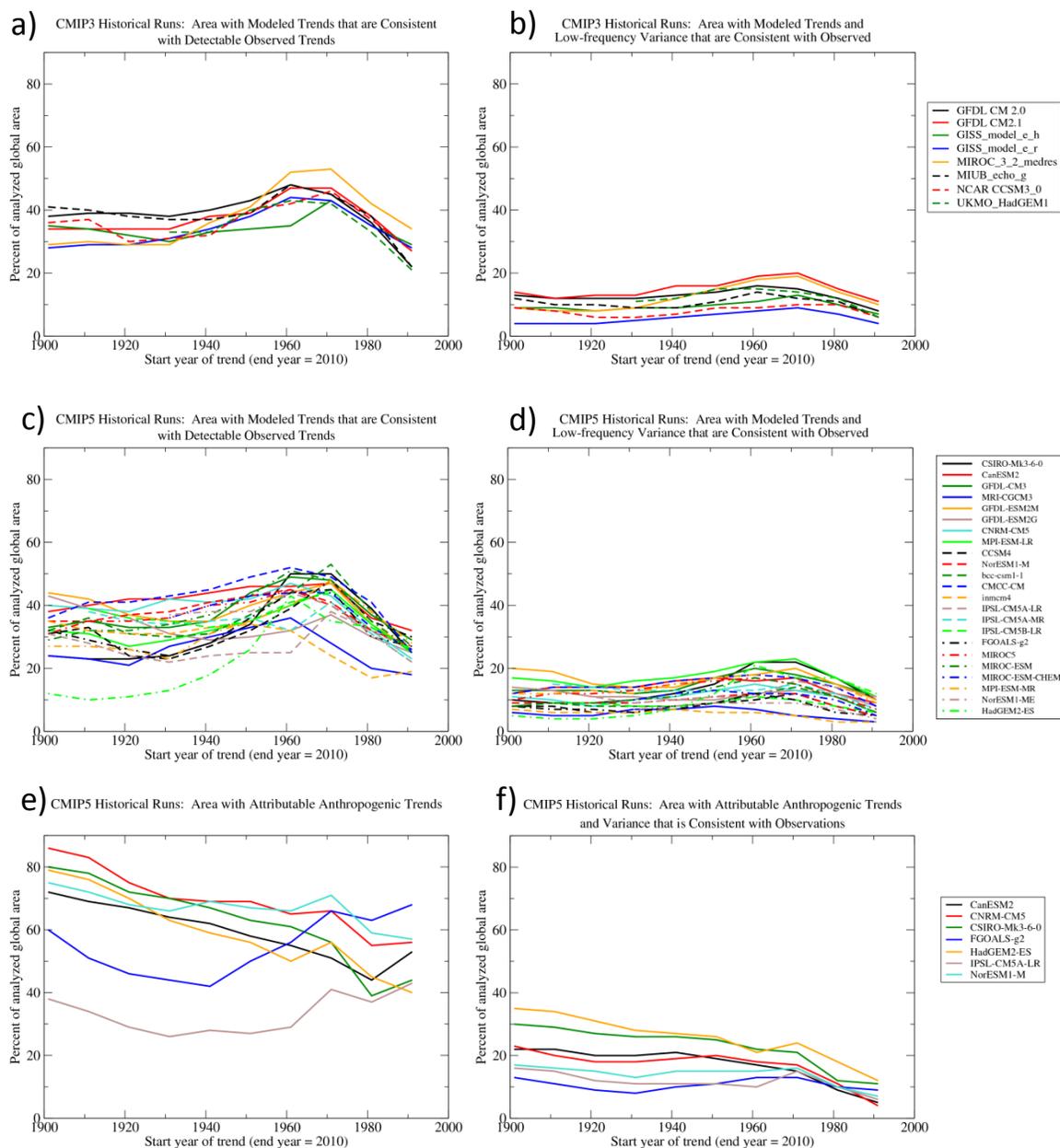


Fig. 17. Individual CMIP3 (a, b) and CMIP5 (c-f) models are assessed for consistency with detectable observed surface temperature trends-to-2010 (a-d), for attributable anthropogenic trends (e, f), and for consistency of both simulated trend and internal variability with observed estimates (b, d, f). Trend results are shown for start years from 1901 to 1991 (all trends ending in 2010). Plotted is the percent of analyzed global area where each individual model's (see legends) multi-realization ensemble mean forced trend and internal variability meet the criteria listed above the panel. The trends are analyzed at each grid point where there is sufficient temporal data coverage for the trend in question (see text). Note that panels (e, f) include areas where the observed trend is detectable and either consistent with or greater than simulated, whereas panels (c, d) include only areas with observed trends that are detectable and consistent with simulations.

Non-Rendered Figure 1

[Click here to download Non-Rendered Figure: Knutson_Fig1.pdf](#)

Non-Rendered Figure 2

[Click here to download Non-Rendered Figure: Knutson_Fig2.pdf](#)

Non-Rendered Figure 3

[Click here to download Non-Rendered Figure: Knutson_Fig3.pdf](#)

Non-Rendered Figure 4

[Click here to download Non-Rendered Figure: Knutson_Fig4.pdf](#)

Non-Rendered Figure 5

[Click here to download Non-Rendered Figure: Knutson_Fig5.pdf](#)

Non-Rendered Figure 6

[Click here to download Non-Rendered Figure: Knutson_Fig6.pdf](#)

Non-Rendered Figure 7

[Click here to download Non-Rendered Figure: Knutson_Fig7.pdf](#)

Non-Rendered Figure 8

[Click here to download Non-Rendered Figure: Knutson_Fig8.pdf](#)

Non-Rendered Figure 9

[Click here to download Non-Rendered Figure: Knutson_Fig9.pdf](#)

Non-Rendered Figure 10

[Click here to download Non-Rendered Figure: Knutson_Fig10.pdf](#)

Non-Rendered Figure 11

[Click here to download Non-Rendered Figure: Knutson_Fig11.pdf](#)

Non-Rendered Figure 12

[Click here to download Non-Rendered Figure: Knutson_Fig12.pdf](#)

Non-Rendered Figure 13

[Click here to download Non-Rendered Figure: Knutson_Fig13.pdf](#)

Non-Rendered Figure 14

[Click here to download Non-Rendered Figure: Knutson_Fig14.pdf](#)

Non-Rendered Figure 15

[Click here to download Non-Rendered Figure: Knutson_Fig15.pdf](#)

Non-Rendered Figure 16

[Click here to download Non-Rendered Figure: Knutson_Fig16.pdf](#)

Non-Rendered Figure 17

[Click here to download Non-Rendered Figure: Knutson_Fig17.pdf](#)