

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

¹Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ 08542

8

9

Submitted to *J. of Climate*

10

11

Version: Revised January 2, 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,
25 weighed against an alternative scenario involving only natural forcings.

26 Variability in the models is assessed via inspection of control run time series, standard deviation
27 maps, spectral analyses, and low-frequency variance consistency tests applied to individual
28 models. The models are found to provide plausible representations of internal climate
29 variability, though there is room for improvement. The influence of observational uncertainty on
30 the trends is assessed, and found to be generally small compared to intrinsic climate variability.

31 Observed temperature trends over 1901-2010 are found to contain detectable anthropogenic
32 warming components over a large fraction (about 75%) of the analyzed global area. In several
33 regions, the observed warming is significantly underestimated by the models, including parts of
34 the southern Ocean, south Atlantic, and far west Pacific. Regions without detectable warming
35 signals include the high latitude North Atlantic, the eastern U.S., and parts of the eastern Pacific.
36 For 1981-2010, the observed warming trends over about 45% of the globe are found to contain a
37 detectable anthropogenic warming; this includes much of the globe within about 40-45 degrees
38 of the equator, except for the eastern Pacific.

39

40 **1. Introduction**

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

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

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

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

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

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

94

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

102

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

104

105 *a. Observed data*

106

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

114

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

122

123 *b. CMIP3 and CMIP5 models*

124

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

141

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

148 “Volcanic” and 11 “Non-Volcanic” CMIP3 models, respectively, and distinguish between results
149 for the two types of historical runs in our analysis. For instance, in cases where we include both
150 sets (19 models), we used the term “Volcanic and Non-Volcanic” models. All 23 of the CMIP5
151 models included in this study included volcanic forcing in their 20C3M runs. However, only
152 seven of the 23 CMIP5 models had Natural Forcing Only runs that extended to 2010. These
153 Natural Forcing runs extending to 2010 were necessary for some of our detection and attribution
154 analyses concerning anthropogenic forcing.

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 are estimates of the internal variability of the climate system derived from the
164 observations in combination with the climate models’ response to estimated historical forcing.

165

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

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

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

186

187 While some of the trends in the CMIP3 and CMIP5 control runs (Fig. 1) approach the observed
188 trend in terms of general magnitude, those are associated with either the long-term drifts
189 discussed above or with a few spurious discontinuity issues (e.g., IAP_fgoals1.0.g). Controlling
190 for these apparent problems, none of the control runs in the CMIP3 or CMIP5 samples exhibit a

191 centennial scale trend as large as the trend in the observations. On the other hand, the variability
192 of observed residual series appears roughly similar in scale to that from several of the control
193 runs. Three of the CMIP3 control runs illustrated in Fig. 1 (GISS_aom, GISS_model_e_h, and
194 GISS_model_e_f) have much lower levels of global surface temperature variability than in the
195 observed residual series. For some sensitivity tests on the multi-model assessments, we have
196 excluded these three models to test for robustness.

197 *b. Geographical distribution of variability*

198
199 The geographical distribution of the standard deviation of annual mean surface air temperature is
200 shown in Fig. 2. for CMIP3 models and Fig. 3 for CMIP5 models. These use the full available
201 time series from each control run. The time series have had the long-term drift removed as
202 discussed in section (3a). Two features that stand out in Figs. 2 and 3 are the enhanced
203 variability over land regions and in the eastern Equatorial Pacific. These general features (and
204 magnitudes of standard deviation) are also seen in the observed residual variability map shown in
205 each figure, giving us some confidence in the models' ability to simulate broad-scale features of
206 surface temperature variability. Note that while the observed residual standard deviation map is
207 also shown here for reference, it should be compared to the model variability with caution. For
208 example, the available observational record is relatively short, compared with many of the
209 model control runs, and there are uncertainties in removing the forced variability component
210 from observations to create the observed residual and thus uncertainties in the observed internal
211 variability estimate used for comparison to the model control runs. With these caveats in mind,
212 one can compare the modeled and observed standard deviation fields using the spatial correlation
213 coefficient (shown above each figure panel, upper right). These vary from about 0.5 to 0.7 for

214 the models shown, indicating relatively good agreement of overall spatial structure of the
215 variability maps. Further, by comparing the “All-model mean” standard deviation fields
216 (average across all of the individual model panels) to the observed field, there is some suggestion
217 that models tend to underestimate the interannual sea surface temperature variability over many
218 parts of the globe. The simulated variability in the ENSO region is very different among the
219 models, with some models clearly displaying much less variability than in the observed map, and
220 other models with apparently excessive variability. However, as noted by Wittenberg (2009) and
221 Vecchi and Wittenberg (2010), long-running control runs suggest that the SST variability in the
222 ENSO region can vary substantially between different 60-yr periods (the length of record used
223 here for observations), which again emphasizes the caution that must be placed on comparisons
224 of modeled vs. observed variability based on records of relatively limited duration.
225 Versions of the control run standard deviation map which use low pass (> 10 year) filtered data
226 have also been examined (not shown). These have similar limitations to Fig. 2 and 3 when one
227 attempts to compare models with observations. The model fields indicate that most CMIP3 and
228 CMIP5 models have their strongest low-frequency (> 10 year) variability in the polar regions and
229 marginal sea ice areas near Antarctica, Greenland, and the periphery of the Arctic Ocean.
230

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

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

233 The global mean time series of surface temperature from the 20C3M historical runs are
234 compared with observations (black curves) in Fig. 4 in a form similar to that presented by Hegerl
235 et al. (2007). The historical dates of large volcanic eruptions are shown by vertical brown lines.
236 An analysis of the model time series for the CMIP3 and CMIP5 All-Forcing experiments is

237 presented in Figs. 4a-c, and for the CMIP5 Natural Forcing Only experiments in Fig. 4d. The
238 large shaded region on each plot shows the 5th to 95th percentile range of a single model
239 realization from the multi-model sample. The multi-model sample is formed by combining the
240 distributions of each of the models, with each model having an equal probability weight in the
241 multi-model distribution. The sub-distribution from each model is centered on that model's
242 ensemble mean with the distribution about that mean based on the control run for that model.
243 Thus the multi-model distribution incorporates the uncertainty due to differences between the
244 model ensemble means (i.e., forcing and response-to-forcing uncertainties) and uncertainties due
245 to internal variability for each model.

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

253 The narrower shaded region between the two thick red lines (a-c) depicts the 5th to 95th percentile
254 range of the multi-model ensemble mean. This is fairly narrow, indicating that the multi-model
255 ensemble means of these particular sets of models are fairly well-constrained, with relatively
256 small uncertainty. The ensemble means of the CMIP3 and CMIP5 volcanic models (Fig. 4 a,c)
257 track the observations remarkably well although the apparent volcanically induced temporary
258 dips are not in full agreement with the observed behavior for those periods. For example, in Fig.
259 4a, and 4c, the multi-model responses to the Pinatubo and Krakatau eruptions appear to be larger

260 than in observations. These apparent discrepancies in the volcanic responses will require further
261 analysis (see e.g. Stenchikov et al. 2009) and are not a focus of the present study. For example,
262 one must carefully assess the role of internal climate variability in judging whether these
263 differences are significant or not.

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

270 *b. Spectra of global mean surface temperature*

271 Figure 5 (a,b) shows the variance spectra of observed global mean temperature (black curves,
272 with shaded range for the 90% confidence intervals) and of the individual CMIP3 and CMIP5
273 “Volcanic forcing” historical runs (red curves) from Fig. 4 (a, c), using data from the years 1880-
274 2010. The data were not detrended prior to computing the spectra. Before plotting, the raw
275 spectra were smoothed using a non-overlapping sliding boxcar window that groups the raw
276 spectra into groups of three calculable frequencies. The 90% confidence intervals on the
277 observed spectrum assume six degrees of freedom for each spectral estimate (group of three)
278 shown. The sum of the variance is plotted at the central frequency of the sliding boxcar window.
279 The enhanced power at low frequencies in (a,b) relative to (c,d) is associated with the strong
280 warming trend in both observations and the All Forcing model runs. There is a strong tendency

281 for the model spectra to lie within the 90% confidence intervals of the observed spectra,
282 particularly at periods longer than 10 yr (frequency $< 0.1 \text{ yr}^{-1}$).

283 The spectra in Fig 5 (c) and (d) are based on residual time series from observations or model
284 historical runs, where the multi-model ensemble surface temperature time series from the
285 20C3M volcanically forced historical runs is first subtracted from the observed global mean
286 temperature series or from the individual model historical runs to form residual time series. As a
287 result of this filtering procedure, most of the long-term warming trend (e.g., Fig. 4 a, c) is
288 removed from the time series. The agreement between variance spectra of model and observed
289 residual time series in Fig. 5 (c,d) is not as good as for the original unfiltered spectra (Fig. 5 a,b),
290 particularly for the CMIP3.

291

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

302

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

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

305 In this section we compare the observed and simulated historical (20C3M) temperature trends
306 obtained from global or regional averages, to assess whether a linear trend signal has emerged
307 from the “background noise” of internal or natural climate variability, as estimated by the
308 models. The primary focus is on the seven CMIP5 models that have Natural Forcing Only runs
309 extending to 2010 and the larger set of 23 CMIP5 models that have All Forcing runs that we can
310 extend to 2010, when necessary, using RCP4.5 projections. We can use these sets of CMIP5
311 model runs together to assess whether the observed trends have emerged from the background of
312 natural variability and whether they contain an attributable anthropogenic component. We use a
313 larger sample (23 models) for our All Forcing analysis because we want to compute the best
314 estimate possible, based on the available models, of the multi-model ensemble mean and its
315 uncertainty, even though we do not have as many models available to use for the Natural Forcing
316 Only estimates. For the eight CMIP3 models that include volcanic forcing (but for which we
317 generally do not have Natural Forcing Only runs), we can ask a more limited set of questions,
318 namely whether the linear trend signal in the observations has emerged from the background of
319 internal climate variability and whether the All Forcing run trends are consistent with the
320 observed trends. We can also address this more limited set of questions using the full set of 23
321 CMIP5 models.

322 We assess the trends across a wide “sliding range” of start years beginning as early as 1861. All
323 trends in the analysis use 2010 as the end year. The general procedure we use is illustrated in
324 Fig. 7 (a) for global mean surface temperature. The black shaded curve in the figure shows the

325 value of the linear trend in observed global mean temperature for each beginning year from 1880
326 to 2000, in each case with the trend ending in the year 2010. The HadCRUT4 observed data set
327 contains an ensemble of 100 estimates, and these are used to create an ensemble of observed
328 trend estimates. The black shading depicts the 5th to 95th percentile range of this ensemble. The
329 first year plotted for global mean temperature was 1880 because the areal coverage and temporal
330 coverage requirements for a trend to 2010 were reached in that year. The observed temperature
331 trend to 2010 is about 0.5°C/100 yr beginning early in the record (late 1800s) and increases to
332 about 2°C / 100yr by around 1980. The observed trend has decreased for more recent start
333 dates, falling below 1°C/100 yr for trends beginning in the late 1990s.

334 The blue curve in Fig. 7a shows the “mean of ensemble mean trends” for the Natural Forcing
335 Only runs of the following seven CMIP5 models, each of which has at least one Natural Forcing
336 Only run in the CMIP5 archive extending to 2010: CanESM2, CNRM-CM5, CSIRO-Mk3-6-0,
337 FGOALS-g2, HadGEM2-ES, IPSL-CM5A-LR, and NorESM1-M. Each of the seven models is
338 weighted equally in the mean of ensemble means, even if a modeling center provided a greater or
339 smaller than average number of within-model ensemble members. The light blue shading in Fig.
340 7 (a) shows the 5th to 95th percentile range of trend values for the Natural Forcing Only runs,
341 which is constructed using the long-term drift-adjusted control run variability (Fig. 1 c,d) from
342 each model. Under an assumption that internal variability in the control run is not substantially
343 different from that in the forced runs, we can use the long control run for each model to estimate
344 the component of inter-realization uncertainty that would be present in the forced trends; this is
345 helpful, since most centers did not provide enough ensemble members to precisely assess this
346 component of the uncertainty.

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

355 Each randomly selected control run trend (from the seven models used) is combined with that
356 model’s ensemble-mean Natural Forcing Only trend for that trend length, to create a distribution
357 of historical Natural Forcing Only trends that include the uncertainty due to both internal
358 variability and the forced response. The blue region is the 5th to 95th percentile range of this
359 distribution of trends, and thus relates to the uncertainty of single ensemble members (which
360 mimics the real world, itself a “single ensemble member”). Therefore, the distribution of trends
361 used to construct the percentile range includes uncertainty due to both the different natural
362 forcings and responses of the individual models, and the uncertainty due to the internal
363 variability as simulated in the control runs. The random resampling approach is necessary
364 because the available control runs for the various models are of different lengths and yet we
365 purposely chose to give each available model an equal “vote” in estimating internal variability.
366 The samples are drawn from the control runs in the form of 150-yr samples with randomly
367 chosen start dates, and each sample is masked with the observed mask of missing data over the
368 period 1861-2010 to create data sets with missing data characteristics that are similar to those of
369 the observations. The analysis in Fig. 7 (a) shows that observed global temperature trends-to-

370 2010 of almost any length are detectable compared to the CMIP5 Natural Forcing Only runs and
371 simulated internal variability—even for trends as short as those beginning around 1990.

372 The dark red curve and light pink shading in Fig. 7 (a) depict the inter-model mean of ensemble
373 means and the 5th to 95th percentile uncertainty range for the All Forcing runs (i.e., natural and
374 anthropogenic forcings combined) and control runs of the full set of 23 available CMIP5 models.
375 These are constructed in an analogous way to the Natural Forcing Only curves and blue shading,
376 and thus depict the uncertainty due to both internal variability and to the different models’
377 responses to historical climate forcing agents (All Forcings, in this case). The violet shading in
378 the plot is the region where the pink and blue shading overlap, indicating that the 5th to 95th
379 percentile ranges of the All Forcing and the Natural Forcing simulated trends at least partially
380 overlap.

381 In Fig. 7 (a), the black (observed) curve is always within the pink (or violet) shaded region,
382 meaning that global mean temperature trends are not significantly different from the CMIP5
383 historical All Forcing run ensemble on any time scale, including for the most recent ‘weak
384 trends’.

385 When the black shaded curve in Fig. 7a lies entirely within (or above) the pink shaded region and
386 entirely outside of the blue shaded region, we conclude that the trend from that point to 2010 has
387 a detectable anthropogenic component. Given that the observed global mean surface temperature
388 trends with start dates through about the mid-1990s lie within this region of then graph, we
389 conclude that the observed global surface temperature warming to 2010 is at least partially
390 attributable to anthropogenic forcing according to these model data and observations. Inspection
391 of Fig. 7a further indicates that the detection and attribution result is sufficiently strong that the

392 uncertainty associated with the combined effects of internal climate variability, uncertainty in the
393 model responses to natural forcing, and the uncertainty in the observed ensemble could be a
394 factor of two larger than shown here and the same conclusion would still hold. for start dates
395 from the late 1800s to about the mid-20th century. Our attribution conclusion for anthropogenic
396 forcing and global mean temperature is not as strong as in IPCC AR4 (Hegerl et al. 2007), partly
397 because we are not focusing in this study on quantifying the magnitude or fractional contribution
398 of the anthropogenic forcing. Rather, our focus is on evaluating the evidence for detectable and
399 attributable anthropogenic influence on surface temperature in various regions around the globe
400 and eventually focusing down to the spatial scale of individual gridpoints in the next section of
401 the study.

402

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

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

407 1) MAJOR LARGE-SCALE REGIONAL INDICES

408 For **global sea surface temperature (SST)** (Fig. 7b), trends to 2010 are clearly detectable for
409 starting years up to about 1990. The observed trends are only marginally attributable to
410 anthropogenic forcing for trends beginning around the mid-20th century, otherwise an
411 attributable anthropogenic signal is clearly apparent for the detectable trends. For **global land**
412 **surface temperature** (Fig. 7c) an attributable anthropogenic signal is clearly seen in the
413 observed trends for all start dates from about 1885 up to about 1990, so the case for attribution is

414 slightly more robust than for global sea surface temperature. The anthropogenic warming signal
415 is so much stronger over land than over ocean, that it readily detectable and attributable despite
416 the greater intrinsic variability over land than over ocean. **Northern hemisphere temperature**
417 (Fig. 7d) roughly mirrors the results for global temperature and global land temperature, with
418 robust detection and attribution for start years up to about 1990. **Southern hemisphere**
419 **temperature** (Fig. 7e) results are similar though not quite as robust as for the Northern
420 hemisphere, as the start dates with attributable anthropogenic influence extending up to about
421 1980, rather than 1990.

422 The **northern hemisphere extratropics** (30°-90°N) series (Fig. 7f) has robust detection and
423 attribution up to around a 1990 start date, but the **southern hemisphere extratropics** (30°-90°S;
424 Fig. 7g) is slightly less robust than the northern hemisphere, as detection/attribution extends to
425 starts dates up to about 1980. The trends for the southern extratropics are relatively constant over
426 a range of start dates from 1900 to 1970, in contrast to northern hemisphere series which shows a
427 period of higher warming trend rates for trends to 2010 beginning in the second half of the 20th
428 century. The southern extratropics trends from 1900 are marginally consistent with the All
429 Forcing model trends, as they are near the upper edge (95th percentile) of the modeled
430 distribution. **Tropical surface temperatures**, which combine land and ocean (Fig. h) regions,
431 show robust detection and attribution for trends to 2010 with start dates as late as about the late
432 1970s.

433 2) REGIONAL SEA SURFACE TEMPERATURE INDICES

434 **Tropical SST's** (20°N-20°S; Fig. 7i) show similar robust detection and attribution results (for
435 start dates as late as about the 1970s) to those for the tropical surface temperature as a whole.

436 **Indian Ocean SSTs** (Fig. 7j; see Fig. 6 to identify region IO) exhibit robust detection and
437 attribution for start dates up to about 1990, despite a larger observational uncertainty, particularly
438 for trends beginning from the 1940s through the 1980s. The **tropical west Pacific** (Fig. 7k) and
439 the **tropical east Pacific** (Fig. 7l) both show a detectable anthropogenic component for trends to
440 2010 beginning from the 1880s to about 1920. However, trends beginning from 1920 to 1970
441 are only marginally detectable as the black region (observations, including uncertainties) is not
442 clearly outside of the blue (natural forcing) region. **North Pacific SSTs** (25° - 45° N, Fig. 7m, see
443 Fig. 6 to identify region), have a detectable anthropogenic component but only for start dates up
444 to about 1910. A marginally detectable signal is found for start dates up to about 1930 and for a
445 narrow range of start years in the 1970s. Otherwise, the trends are not detectable according to
446 our analysis. The **tropical Indian Ocean / western Pacific “warm pool” region** (Fig. 8o) is an
447 important region as it is a dominant large-scale region for tropical convection. This region has a
448 detectable anthropogenic warming trend to 2010 for start dates as late as about 1990.

449 We analyzed four separate regions of the Atlantic Ocean, as this basin is noted for pronounced
450 multi-decadal variability. **North Atlantic SSTs** (45° - 60° N; Fig. 7n) exhibit no detectable trends
451 outside of the range of natural variability for any start dates, according to our analysis. This
452 region is notable for having probably the least detectable signal of any of our study regions
453 around the globe. Despite the lack of detectable trends, the observed trends are at least
454 consistent with the All Forcing runs, which have a very wide 5th to 95th percentile range of trends
455 due to the large simulated internal variability, as will be shown later in this section. In the
456 **subtropical north Atlantic** (20° - 45° N; Fig. 7o) an anthropogenic signal is detected for start
457 dates from about 1890 to 1920 and around 1970, but otherwise is only borderline detectable up
458 to about 1980. In the **tropical North Atlantic “main development region”** for Atlantic tropical

459 cyclones (Fig. 8a), there is a detectable anthropogenic warming to 2010 for start dates up to
460 about 1960, and then intermittently for start dates up to about 1990. In the **South Atlantic** (Fig.
461 8b), there is a detectable anthropogenic warming for start dates up to the late 1970s. An
462 interesting feature in this region is that warming trends from the 1890s are slightly higher than
463 even the 95th percentile of the model simulations.

464 3) MAJOR LAND REGION TEMPERATURE INDICES

465 We now summarize the characteristics of surface temperature trends in major continental
466 regions, beginning with Eurasia, Africa, and Australia. The **Europe** temperature index (Fig.
467 8c) has detectable anthropogenic warming trends for start dates up to about 1990. An interesting
468 feature of the Europe trends is that there is no start date for which the 5th to 95th percentile range
469 of the All Forcing and Natural Forcing Only simulated trends are not at least partially
470 overlapping. That is, in some sense the All Forcing and Natural Forcing trends from the models
471 are not completely distinguishable from each other. Nonetheless, the observed trends (even
472 accounting for observational uncertainty in the HadCRUT4 data set) are clearly outside of the
473 range of the Natural Forcing trends but lie well within the range for the All Forcing trends. The
474 **Africa** index (Fig. 8d) has detectable anthropogenic warming trends for start dates up to about
475 the year 2000. Our analysis of **African** temperature trends only extends back to start dates
476 beginning in the mid-1920s, due to more limited data coverage. For **northern Asia** (Fig. 8e),
477 our start dates extend back to the early 1900s and show a clear detectable anthropogenic
478 warming signal for start dates extending from there up to about 1980. For **southern Asia** (Fig.
479 8f) there is a similarly strong detectable anthropogenic warming signal for start dates extending
480 from the late 1800s through about 1990. An interesting feature of the **African** and **southern**
481 **Asia** results is that the 5th to 95th percentile range of the All Forcing trends from much of the 20th

482 century is much wider than the range for the Natural Forcing runs. Since the contribution from
483 internal variability (estimated from the control runs) is the same for the two sets of trend results,
484 the uncertainty range of the All Forcing ensemble mean trends across the models must be
485 comparable to or substantially larger than the uncertainty due to internal climate variability
486 alone. The **Australia** temperature index (Fig. 8g) shows detectable anthropogenic warming
487 trends for start dates from the late 1800s to about 1970.

488 Considering now the land regions of North and South America, the index for **Canada** (Fig. 8h)
489 shows detectable anthropogenic warming trends for start dates up to about 1970. In contrast, for
490 the **Alaska** index (Fig. 8i), a detectable anthropogenic warming trend to 2010 is most clear for
491 start dates over the more limited range of 1940 to 1970. Trends for post-1970 start dates are
492 generally not detectable, and trends for start dates from about 1910 to 1940 are only marginally
493 detectable. For the **continental United States** (Fig. 8j) an anthropogenic warming trend to 2010
494 is detectable for start dates of about 1900 to 1975. For start dates of about 1860 to 1900, the
495 warming signal is only marginally detectable. The temperature index for **Mexico** (Fig. 8k)
496 indicates that observational uncertainties play an important role for detection and attribution
497 results in this region. A detectable anthropogenic warming trend is seen for start dates of about
498 1910-1920 and about 1965-1980, otherwise the trends are not detectable. In contrast, for the
499 **South America** index (Fig. 8 l), the temperature trends to 2010 are mostly detectable for start
500 dates from about 1910 go 1950, but are not necessarily attributable to anthropogenic forcing for
501 these periods because the observed trends are not within the pink region (range of All Forcing
502 simulated trends). Rather, they appear systematically smaller than the simulated trends, after
503 accounting for observational uncertainties. Anthropogenic warming trends to 2010 are detectable
504 for the **South America** index but only for a limited set of start years in the early 1970s.

505 Temperature trends for the **southeastern United States** index (Fig. 8n) are of particular interest
506 because the trend behavior in this region is different from most other land regions around the
507 globe, as has been pointed out in a number of previous studies (e.g., Hegerl et al. 2007; Knutson
508 et al. 1999, 2006). According to our present analysis, trends to 2010 in this index are detectable
509 only for a limited range of start years (mid-1950s to the mid-1970s). For that limited set of start
510 years, an anthropogenic warming trend to 2010 is detectable in our analysis. The trends in the
511 index to 2010 at least are consistent with All Forcing runs for all start years after about 1940, but
512 the warming trends even after 1940 are for the most part not strong enough to be detectable
513 against the background of natural forcing and internal climate variability. This behavior
514 contrasts with the index for the **rest of the continental United States** (that lies outside of the
515 southeastern U.S.) (Fig. 8 m), where an anthropogenic warming trend to 2010 is broadly
516 detectable for start years ranging from about 1870 to the mid-1970s.

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

518 Our regional temperature indices analysis in subsections 5(a) and 5(b) (i.e., Figs. 7 and 8)
519 focused on the subset of seven CMIP5 models that had Natural Forcing Only runs that extended
520 to 2010 and on the full set of 23 CMIP5 models that had All-Forcing runs available. Here we
521 conduct a complimentary assessment (for a more limited set of regions) that compares the eight
522 CMIP3 models (All Forcing and control runs) with the full set of 23 CMIP5 models (All Forcing
523 and control runs). Where necessary, the All-Forcing 20C3M runs were extended to 2010 using
524 A1B (CMIP3) or RCP4.5 (CMIP5) projection runs; this procedure was not tenable for the
525 Natural Forcing Only runs due to the strong differences in forcing between Natural Only and the
526 A1B or RCP4.5 scenarios in the early 2000s.. Our analyses for the CMIP3 models (and the
527 CMIP5 models as shown in the middle column of Fig. 9) only compare internal climate

528 variability (control runs) with All Forcing historical runs. Thus, we cannot use these results to
529 draw firm conclusions about detection of anthropogenic trends, because an important alternative
530 hypothesis (Natural Forcing) is not being thoroughly tested in this case. Nonetheless, we can
531 draw some conclusions about detection of significant trends (against a background of internal
532 climate variability) and about consistency of observed trends versus the trends in the All Forcing
533 20C3M experiments.

534 Our procedure is illustrated for the **global temperature** analysis in the top row of Fig. 9 (a-c).
535 Figure 9c is identical to Fig. 7a and is repeated here for reference only. Figure 9a shows the 5th
536 to 95th percentile range for the observed trends to 2010 (black shading); the 5th to 95th percentile
537 range for the All Forcing runs from the eight CMIP3 models (pink shading, with the red curve
538 depicting the ensemble mean); and the 5th to 95th percentile range of control run trends from the
539 same eight CMIP3 models (green shading). Violet shading illustrates regions of overlap of the
540 pink and green shaded regions. Where the black curve lies outside of the green shaded region,
541 the observed trend is detectable compared to internal climate variability in the CMIP3 runs.
542 Where the observed curve lies within the pink shading, the observed trend is assessed as
543 consistent with the CMIP3 All Forcing ensemble of runs.

544 Figure 9a (CMIP3) indicates that the observed **global mean temperature** trends to 2010 are
545 detectable (inconsistent with internal climate variability in the eight CMIP3 models) for start
546 dates from about 1880 to the mid-1990s, and are consistent with the CMIP3 All Forcing run
547 trends to 2010 for essentially all start dates from 1880 to 2000. Similar conclusions are evident
548 for the 23 CMIP5 models as shown in Fig. 9b. As noted earlier, similar results are seen for the
549 CMIP5 models when we incorporate the Natural Forcing Only runs in the tests (Fig. 9c),

550 although there the detectability of the observed trend extends to start dates as late as about 1990,
551 rather than the mid-1990s.

552 For **tropical SST** (Fig. 9d-f) the CMIP5 models, including the seven model subset with Natural
553 Forcing Only runs to 2010 (Fig. 9 f), indicate robust detection and attribution for trends to 2010
554 with start dates as late as about the late 1970s, as discussed earlier. The consistency with the All
555 Forcing runs (all 23 CMIP5 models) is only marginal for a period of start dates around 1960. A
556 similar consistency result is seen for the 23 CMIP5 models (Fig. 9e) where we compare their All
557 Forcing runs with their control variability. The observed trends to 2010 appear to be detectable
558 against the internal variability (control run) background of the 23 CMIP5 models for start dates
559 as late as about 1990. For the eight CMIP3 models (Fig. 9d), the observed trends to 2010 are
560 detectable for start dates up to 1990, similar to the CMIP5 models (Fig. 9e). However, the eight
561 CMIP3 All Forcing runs are not assessed as being as consistent with the observed trends to 2010
562 as are the 23 CMIP5 All Forcing runs. In fact the CMIP3 All Forcing runs appear only
563 marginally consistent with the observed trends to 2010 for most of the start dates from 1880
564 through about 1980.

565

566 The **North Atlantic** (45° - 65° N) was highlighted earlier as a region with no detectable trends
567 compared with the CMIP5 Natural Forcing Only runs and internal climate variability combined
568 (Fig. 9i). This is perhaps not surprising, given the substantial intrinsically-generated fluctuations
569 on multi-decadal time scales in this region (see e.g. Yang et al. 2013). We see from the green
570 and violet shaded regions in Figs. 9 g,h that the range of trends to 2010 due to internal climate
571 variability alone in the CMIP3 and CMIP5 models is quite large and appears to largely account

572 for a similar wide range of simulated trends in the All Forcing runs. This also helps allow the
573 observed trends to 2010 to be consistent with the CMIP3 and CMIP5 All Forcing trends for all of
574 the start dates examined, despite the fact that the observed trends are not detectable (i.e., not
575 distinguishable from control run variability alone).

576 For the **southeastern United States** index (Fig. 9 j-l) there is slightly more evidence for
577 detectable trends to 2010 versus the internal variability samples in Fig. 9 j,k (start years 1950 to
578 1980) than versus the combined Natural Forcing/internal variability sample of trends (blue
579 shading in Fig. 9 (l)) with marginally detectable trends for start dates from the mid-1950s to the
580 mid-1970s). For start years prior to about 1940, the observations lie near the edge and even
581 outside of this 5th to 95th percentile range for the All Forcing runs (pink/violet shaded
582 envelopes), especially for CMIP3 (Fig. 9j). We thus conclude that even accounting for internal
583 variability, the CMIP3 and CMIP5 historical runs trends-to-2010 tend to be inconsistent or only
584 marginally consistent with the observed southeastern U.S. surface temperature trends for starting
585 dates before about 1940. This means that the CMIP3 and CMIP5 All Forcing runs can be
586 falsified, at least for this relatively small region, and further implies that there remain as yet
587 unexplained discrepancies between the historical simulations and observations for trends in this
588 region.

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

594

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

596 *a. Multi-model ensemble assessment*

597 1) 1901-2010 TRENDS

598 The procedures in Section 5 that were used to categorize observed trends at individual grid
599 points as detectable, attributable in part to anthropogenic forcing, consistent with All Forcing
600 runs, etc. can be applied at the grid-point scale, and the categories displayed in map form, for a
601 selected trend period. For example, Fig. 10 shows the results of such a category analysis for the
602 observed vs modeled trends for 1901-2010, with the bottom row showing category maps for the
603 CMIP3 All-Forcing runs (e) and CMIP5 All-Forcing and Natural Only Forcing runs (f). The
604 linear trend maps for observed temperature (1901-2010) and the CMIP3 and CMIP5 All Forcing
605 ensemble means are shown in Fig. 10 (a-d) for reference. The observed trend map shows broad-
606 scale warming trends since 1901 at almost all locations around the globe, with areas of cooling in
607 only a few regions, mainly in the high latitude North Atlantic and the southeastern United States.
608 The CMIP3 and CMIP5 multi-model ensemble trends show broadly similar magnitude and
609 pattern of cooling to that observed, where the agreement can be quantitatively tested by our
610 consistency tests as described in the previous section. For the tests described in this section, we
611 use only the ensemble mean observed trend and thus do not consider observational uncertainty,
612 which was examined in the previous section.

613 Figure 10 (f), for the 23 CMIP5 models with All Forcing and the seven CMIP5 models with
614 Natural Forcing Only runs to 2010, builds upon the regional time series analysis shown in Figs.
615 7-8. The white regions in Fig. 10 (f) indicate where the observed trend is not detectable

616 compared to the Natural Forcing only runs (where the uncertainty estimates incorporate both
617 simulated internal climate variability from the seven control runs and uncertainties in the Natural
618 Forcing Only ensemble mean). The dark grey regions in Fig. 10 (f) do not have sufficient data
619 coverage for our tests. (To determine if a grid point has “sufficient coverage” to include in our
620 maps and analyzed area, we divide a given trend period (e.g., 1901-2010) into five roughly equal
621 periods, and require that each of the five periods has at least 20% temporal coverage in the
622 monthly anomaly data.) The various colored (non-white, non-grey) regions in Fig. 10 (f)
623 indicate where the trends are detectable, with the category identified on the legend. The orange
624 regions show where the warming trend is detectable but still less than the lower end (5th
625 percentile) of the All Forcing trend distribution. The light and dark red regions indicate where
626 the observed trend has a detectable anthropogenic component; for the darkest red regions the
627 observed warming trend is so large that it exceeds the 95th percentile of the modeled distribution,
628 but here we still interpret this as implying a detectable anthropogenic component. For cooling
629 trends (blue regions), we have analogous terms to those used for the various warming cases,
630 although these cases are almost absent for the 1901-2010 trends in our analysis.

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

639 A similar analysis for the CMIP3 All-Forcing runs (eight models with volcanic forcing) is shown
640 in the left column of Fig. 10 (a,c,e). The category names for the assessment (Fig. 10 e) are
641 different than for the CMIP5 models (Fig. 10 f) because a Natural Forcing Only ensemble is not
642 available in the archive for the CMIP3 models. Therefore, our categories for CMIP3 (see
643 legend) are limited to assessing consistency, either with the internal variability of the control
644 runs or with the All-Forcing runs, and we do not assess the question of attribution to
645 anthropogenic forcing. The observed widespread warming trends shown in Fig. 10 (a) are
646 assessed as detectable (compared with control run or internal climate variability) over most of
647 the global region with sufficient coverage. Only in some regions of the North Atlantic and North
648 Pacific (white regions in Fig. 10 (e) is the observed trend not detectable. In only a very minor
649 fraction of the analyzed area is there a detectable cooling trend since 1901 (blue shading in Fig.
650 10 e), according to our analysis. Orange regions (where the warming trend is detectable but less
651 than simulated) occur preferentially in the lower latitudes, while regions with significantly
652 greater than observed warming trends tend to occur more in the extratropics. This feature is
653 clearer for the CMIP3 ensemble (Fig. 10 e) than the CMIP5 (Fig. 10 f).

654 2) 1951-2010 TRENDS

655 Figures 11 explore how the results seen for 1901-2010 trends in Fig. 10 are altered when we
656 analyze the trends for 1951-2010 (Fig. 11). The observed trend map (Fig. 11 a or b) shows much
657 more spatial structure than the trend map for 1901-2010 (Fig. 10 a or b). The Asian and North
658 American extratropical land regions have warmed more since 1951 than the oceanic regions.
659 This amplification of warming over land since 1951 is also evident in the All Forcing 20C3M
660 ensemble means for both the CMIP3 (8 models) and CMIP5 (23 and 7 models with
661 accompanying All Forcing runs and Natural Forcing Only runs to 2010, respectively)—although

662 the contrast between the continental and oceanic regions is more pronounced in the observed
663 trend map than in the multi-model ensembles, especially for CMIP3. This is also seen in the
664 category maps (Fig. 11 e, f) where dark red shading (observed warming significantly greater than
665 simulated) is more prevalent over Asia in the CMIP3 assessment (e) than in the CMIP5
666 assessment (f).

667 The observed trend map (Fig. 11 a, b) shows a region of notable cooling over the mid-latitude
668 North Pacific and a smaller region of cooling trends in the high latitude North Atlantic south of
669 Greenland. These cooling regions are assessed as having no detectable change, meaning that the
670 cooling trends lie within the 5th to 95th percentile range of the simulated trends from the model
671 control runs (CMIP3) or combined control run/Natural Forcing runs (CMIP5). Non-detectable
672 trends for 1951-2010 (white category, Fig. 11 e,f) are found over large regions of the North
673 Pacific, the central equatorial Pacific, the mid- to high-latitude North Atlantic, the far Southern
674 Ocean near Antarctica, and in a few scattered continental regions such as the south-central
675 United States.

676 Figure 11 (f) indicates where observed trends (1951-2010) are attributable, at least in part, to
677 anthropogenic forcing (light and dark red regions). These regions cover most of the global area
678 that has detectable trends, and for the 1951-2010 trends are comprised predominantly of regions
679 where the trends are consistent with the All Forcing ensemble for CMIP5 (light red). Smaller
680 regions of Asia, the tropical Indian Ocean and South Pacific have strong warming trends that are
681 attributable in part to anthropogenic forcing but are also significantly larger than simulated in the
682 CMIP5 All Forcing runs (dark red shading). The category results for the eight CMIP3 models
683 (Fig. 11 e) are similar to those for the CMIP5, although the categories in Fig. 11 (e) do not

684 include attribution to anthropogenic forcing (see legend), since the CMIP3 set of models does
685 not include Natural Forcing Only runs that are necessary for such an attribution.

686 Regions in Fig.11 (e, f) with warming trends that are detectable but significantly less than
687 simulated in the All Forcing runs (orange regions) are mainly found in the tropical and
688 subtropical latitudes. This, combined with the greater prevalence of dark red (stronger than
689 simulated warming) in the higher latitudes, implies that for the 1951-2010 trends overall, the All
690 Forcing runs (CMIP3 and CMIP5) tend to exhibit too strong a warming trend at lower latitudes
691 but too little warming in high-latitudes.

692 3) 1981-2010 TRENDS

693 The trend assessment results for 1981-2010 are presented in Fig. 12. The observed trend map
694 (Fig. 12 a) has much more spatial structure than for the longer trend periods in Figs. 10a and 11a.
695 Since 1981 there have been large regions of cooling trends over the tropical and subtropical
696 eastern Pacific, Gulf of Alaska, and the high latitude Southern Ocean. The analysis shows that
697 for the most part, the cooling trends in these regions are not detectable. In fact, since less than
698 5% of the globe has “detectable” cooling trends, the percent of occurrence of the blue regions is
699 not significantly different from what could occur from sampling variability alone.

700 The large expanse of the globe without detectable trends (1981-2010) in Fig. 12 contrasts with
701 the earlier finding of detectable warming in most analyzed regions for the longer trend analyses
702 (Figs. 10, 11). The loss of a detectable signal, as one proceeds to later start dates in the 20th
703 century--and shorter trend periods--is not unexpected. For example, the results in Figs. 7-9
704 showed how the trend *rates* for internally generated trends in the model become higher for
705 shorter trend periods, as the models can produce strong internally generated trend rates over

706 relatively short periods. Comparing the category maps for different start dates (Fig. 10-12), the
707 loss of detectability, as one proceeds to later start dates, occurs first in the extratropical North
708 Atlantic (north of 40°N) and over large parts of the North Pacific, extending into the tropics, as
709 seen for the 1951-2010 trends (Figs. 11). For the late 20th century start dates (e.g., 1981-2010;
710 Fig. 12) the region of no detectable warming expands to cover most of the southern oceans, south
711 of 40°S, and extending south from 20°S in the South Atlantic. This non-detection region also
712 expands to include most of the eastern tropical and subtropical Pacific and much of the northern
713 extratropics over Eurasia, North America, and the North Pacific.

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

728 There is relatively little orange area (i.e., detectable warming, but significantly less than
729 simulated) on the assessment maps for 1981-2010 (Fig. 12 e, f). The infrequent occurrence of
730 this category for the later trend start dates can be explained by referring to the sliding trend
731 analyses in Figs. 7-9. The unshaded area on those graphs between the pink- and blue-shaded
732 envelopes, corresponding to detectable warming that is less than simulated, systematically
733 shrinks as one progresses to later start dates. That is, for shorter trend periods, it becomes much
734 more difficult to distinguish the simulated All Forcing trend distribution from the trend
735 distribution of the Natural Forcing Only runs (CMIP5) or from the control runs (CMIP3).

736

737 4) ENSEMBLE MEAN ASSESSMENT STATISTICS ACROSS TIME

738 In Fig. 13, we explore how the percent of analyzed areas with various category classifications
739 changes for different start years (all for trends ending in 2010). Figure 13(b) shows the
740 aggregate percent area results for the CMIP5 models, using the 23 models that have All Forcing
741 runs, and the seven model subset with Natural Forcing Only runs extending to 2010. The total
742 percent of analyzed area (i.e., regions with sufficient data coverage) that was assessed as having
743 attributable anthropogenic warming trends (black curve) was about 75% for trends over the
744 period 1901-2010. This drops to about 60% for start dates from 1931 to 1961, then temporarily
745 increases again to over 65% for trends 1971-2010, before dropping sharply to about 25% for the
746 shortest period (1991-2010). The temporary increase in percent of area with attributable
747 anthropogenic warming for the 1971 start date, is apparently due to the temporary pause in
748 global warming from about 1940 to 1970. The end of this pause, around 1970, is a time period
749 during which the prospects for detection of a warming signal are at least temporarily enhanced

750 against a backdrop of a gradually declining percentage as the start date is moved forward through
751 the 20th century. The green curve in Fig. 13b (percent of analyzed area with no detectable
752 change) shows generally opposite behavior to the black curve, increasing from a low of about
753 10%, for 1901-2010 trends, to a high point of over 60% for the latest start period analyzed
754 (1991-2010). The analysis thus illustrates the advantages of a long record for detectability of
755 the warming trend. The light green curve (warming that is detected but less than simulated) is
756 roughly 15% of the analyzed area for start dates through about 1941, then declines for later start
757 dates as the increasing dominance of internal variability for short trend periods makes it much
758 more difficult to distinguish the All Forcing and Natural Forcing trend distributions and thus
759 more difficult for a trend to lie between the two distributions as discussed earlier. The percent
760 of area with trends that are attributable to anthropogenic forcing but significantly greater than
761 simulated (red curve) also diminishes as the start dates move later in the century, possibly
762 because of the growing width of the simulated trend distributions associated with internal climate
763 variability, implying that it becomes difficult for an observed trend to be large enough to be
764 distinguishable from the All Forcing distributions on the high side.

765

766 Figure 13 (a) summarizes the comparison between the CMIP3 and CMIP5 results (solid lines vs.
767 dashed lines) for various common categories. To construct this figure we use the percent of
768 analyzed areas from the CMIP3 eight model ensemble (with volcanic forcing) as shown in Figs.
769 10-12 (panels a, c, e). For the CMIP5, we use results for all 23 models that have volcanic
770 forcing, since a Natural Forcing Only experiment (extending to 2010) is not required for the
771 comparison in Fig. 13 (a), and thus we are not limited to the seven CMIP5 model subset. The
772 percent area where the warming is detected and either consistent or greater than simulated (black

773 curves) is about 70% (CMIP3) and over 75% (CMIP5) for the 1901-2010 period, then decreases
774 for start dates of 1931 or 1941, before rising to a temporary peak of about 70% for the 1971 start
775 date and then falling again for later start dates. As discussed earlier, temporary rise for mid-
776 century start dates is likely due to the enhanced detectability of trends that start within the
777 “relative trough” or temporary interruption of global warming that occurred around this time
778 following the relative peak in global temperatures around 1940. For start dates up to about
779 1931, the black curve for the 23 CMIP5 models (dashed) is about 5% higher on average than the
780 (solid) one for the eight CMIP3 models. Thus, the 23 CMIP5 model All Forcing runs appear at
781 least slightly more consistent with observed trends than the eight CMIP3 All Forcing runs, at
782 least for the case of trends to 2010 starting earlier than 1940. However, for trends with start
783 dates from 1941 through about 1971, the opposite is true, and the CMIP3 All Forcing runs
784 appear modestly more consistent with observations. Other features in Fig. 13 (a) are generally
785 similar to those described for the seven CMIP5 models (Fig. 13 b), although the category
786 descriptions (conclusions about attribution) are necessarily different. The general temporal
787 behavior of the various curves through time is remarkably similar between the solid (CMIP3)
788 and dashed (CMIP5) models in Fig. 13 (a).

789 *b. Model by model trend assessment*

790 In contrast to the analyses in the previous subsection (Figs. 10-13) which focused on the multi-
791 model ensemble means vs. observations, in this subsection we consider the individual models
792 within the CMIP3 and CMIP5 ensembles and assess what percentage of individual models meet
793 certain criteria. That is, the determination of whether a given CMIP3 or CMIP5 individual
794 model is included in a category (e.g., “warming- detectable and consistent”) for a given grid
795 point is based on the evaluation of the historical runs and control runs for that model alone. In

796 this section, we also introduce and apply a variance consistency test as an addition consistency
797 test for the models vs. observations.

798 We will introduce and describe the various tests as we discuss the different panels in Fig. 14,
799 which contains the analysis of the eight CMIP3 models (with volcanic forcing) vs. observations
800 for linear trends over the period 1901-2010. Figure 14 (a) and (b) present the observed and
801 multi-model ensemble mean trend maps for reference; these were discussed earlier for Fig. 10.
802 Figure 14 (c) shows the fraction (or percent) of models, at each grid point, that have no
803 detectable trend. The area-weighted global average of this fraction is 0.09, and the most
804 prominent regions with no detectable trend are in the North Atlantic (south of Greenland), the
805 mid-latitude North Pacific, and the southeastern United States. Figure 14 (d) shows the fraction
806 of models at each grid point with warming that is detectable but less than simulated in the All
807 Forcing runs. The global average fraction is 0.22, and the most prominent region of occurrence
808 is in the tropics, meaning that the models tend to simulate too rapid a century-scale warming in
809 the tropics. The warming is detectable and consistent with the All Forcing runs for a global
810 average fraction of 0.34 of the models (Fig. 14 e), with a spatial pattern that is fairly evenly
811 distributed around the analyzed areas of the globe. The warming is detectable and significantly
812 greater than simulated for a global average fraction of 0.32 of the models (Fig. 14 f), with the
813 most prominent occurrence of this category being in the mid- to high latitudes of both
814 hemispheres. Warming is detectable for about 89% of the models, on average around the globe
815 (Fig. 14 g)—essentially the inverse of the results in Fig. 14 (c). Warming is detectable and
816 consistent or greater than simulated for two thirds of the models, on average, (Fig. 14 h) which
817 shows essentially the inverse of the pattern in Fig. 14 (d), and indicates that the simulated
818 warming tends to be too weak in mid to higher latitudes in the CMIP3 All Forcing runs. The

819 observed and CMIP3 simulated (All Forcing) trends are assessed as consistent for 39% of the
820 models on average (Fig. 14 i); this category includes cases where the trend is not detectable, but
821 still consistent with the All Forcing runs. The fraction field for the models has a fairly even
822 spatial distribution over the global analyzed area.

823 One limitation of our approach is that models with unrealistically large internal variability have
824 some advantage over models with more realistic variability, in that it is easier for high-variability
825 models to have trends that are consistent with observations, since the margin of error is greater.
826 To address this concern, we apply a second test (a variance consistency test) to the models. Then
827 a model that has both a consistent trend and consistent variability, compared with observed
828 estimates, will be ranked more highly in a metric test compared with a model with consistent
829 trends but inconsistent variability. In other words, this expands our consistency tests into a two-
830 dimensional space (trend and internal variability).

831 The variance consistency test for the eight CMIP3 models with volcanic forcing is constructed as
832 follows. For each grid point, we form the observed residual time series for the period 1901-
833 2010, which is defined as the observed minus the CMIP3 eight-model ensemble mean All
834 Forcing series. We filter this residual time series with a low pass smoother that transmits
835 variance on time scales of about 10 years or greater, which are the time scales most relevant to
836 the issue of long-term trends. We then compute the standard deviation of the low-pass filtered
837 observed residual series. For each of the CMIP3 models, we draw many 110-yr samples from
838 the drift-adjusted control runs (see Section 3 a) and with the samples having random start dates.
839 For each of these 110-yr segments, the control model data is masked with the observed mask for
840 the given grid point, then low-pass filtered, and the standard deviation computed. A distribution
841 of the standard deviations for the control run is computed for each grid point and model. If the

842 standard deviation of the observed residual series lies within the 5th to 95th percentile range of the
843 control run distributions, the model is assessed as having low-frequency internal climate
844 variability that is consistent with the observations according to this test. There are some
845 important limitations of this test, which we recognize at the outset. When applied to a single
846 model, as done here, a single model's control run may not be long enough to provide an adequate
847 sample of the 5th to 95th percentile range of low-frequency (>10 yr) variance estimates; indeed,
848 this is an important reason to advocate for longer control runs (or larger ensemble sizes) in future
849 CMIP designs. In addition, the observed residual, which is needed for comparisons with control
850 run variability, has some uncertainties, as the multi-model ensemble mean forced response only
851 approximately removes the forced climate signal from the observations.

852 Figure 14 (j) illustrates the results of applying this test. On average, 25% (two of the eight)
853 CMIP3 models have variability consistent with observations, according to the test. The
854 occurrence of the consistent model and variance has a fairly even distribution around the globe,
855 although the fraction is notably low in the southeastern Pacific and south Atlantic basins.

856 Figure 14 (k) shows the map of the fraction of the CMIP3 models where both the variability and
857 trend are consistent with observations according to our tests. The global average fraction is 11%,
858 indicating that achieving consistency with both tests simultaneously at the grid point scale is a
859 challenge for most models. The variance consistency test can also be applied to the global mean
860 temperature series (e.g., Figs. 4b, 5c, and 9a). We find that six of the eight CMIP3 models
861 (75%) have low-frequency variance for their global mean temperature that is consistent with the
862 observed residual, according to our test.

863 Figures 15 and 16 present the same analysis as Fig. 14, but for the 23 CMIP5 models with All
864 Forcing runs (Fig. 15), and for the subset of seven CMIP5 models that have at least one Natural
865 Forcing Only run extending to 2010 (Fig. 16). The mapped results for the 23 CMIP5 models
866 (Fig. 15) are rather similar overall and have similar spatial features to those for the CMIP3
867 models (Fig. 14) discussed above. One notable difference is that the CMIP5 models in both Fig.
868 15 and Fig. 16 have a substantially greater global mean fraction of models with consistent low-
869 frequency variance (0.36-0.37) than the CMIP3 models in Fig. 14 (0.25). Consequently, the
870 globally averaged fraction of models that have both consistent trend and variance (panel k) is
871 modestly higher in CMIP5 (0.14-0.15) than in the CMIP3 sample (0.11). Figure 16, for the
872 seven model subset of CMIP5 models, can be used to assess whether trends contain attributable
873 anthropogenic trend contributions. The analysis indicates that the globally averaged percent of
874 the seven CMIP5 models with attributable anthropogenic warming at the grid point scale over
875 the 1901-2010 period is 70% (Fig. 16 h). The globally averaged percentage of models with both
876 attributable anthropogenic warming and consistent low-frequency variance is 27%, according to
877 the tests described above (Fig. 16 l).

878 The variance consistency test can also be applied to the global mean temperature series for both
879 the full set of 23 CMIP5 models and the seven model subset of CMIP5 models. This test
880 indicates that 30% of the 23 CMIP5 models, and 14% of the CMIP5 model subset (one of
881 seven), have low-frequency variance that is consistent with observations.

882 As mentioned earlier, there are important limitations of our variance consistency test. We hope
883 to improve on the variance consistency tests in a future study; for example, and there are other
884 model-observation comparison paradigms that can be explored (e.g., Annan and Hargreaves
885 2010). Meanwhile, we stress the need for longer control runs and/or greater numbers of

886 independent ensemble members from the models in order to more robustly assess the various
887 models' low-frequency variability.

888 Figure 17 displays several globally averaged trend consistency metrics as a function of trend start
889 year for the individual models in the CMIP3 and CMIP5 samples. Fig. 17 (c, d, and f) also
890 assess the consistency of the models' low frequency variability, as these include both the trend
891 consistency test and the variability consistency test. In the various panels of Figure 17, we
892 compare, across the models, the fraction of analyzed area where there is both a detectable change
893 in observations and where this detectable change is consistent with the individual climate
894 models. Note that the metrics examined do not include the fraction of area where a climate
895 model is consistent with observations but there is not a detectable trend.

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

903 The results in Fig. 17 (a, b) show that the individual CMIP3 and CMIP5 models have rather
904 similar behavior in terms of fraction of globally analyzed area with consistent detectable trends
905 (typically ranging from 20 to 50%). There is somewhat more spread among the CMIP5 models,
906 although there are more models in the CMIP5 sample as well. This trend consistency metric
907 tends to reach a peak value around 1960-1970 start dates before declining for later start dates, for

908 reasons discussed for Fig. 13. When a variance consistency test is added (Fig. 17 c,d), the
909 percent of analyzed global area with both consistent trends and consistent low frequency
910 variance drops substantially, to typically about 10 to 20%.

911 For the seven-model CMIP5 sample (Fig. 17 e), the percent of analyzed global area with
912 attributable anthropogenic trends is close to 80% for 1901-2010 trends, for five of the seven
913 models, with the remaining two models having lower percent area (40-60%). All seven models
914 end up in the range of 35-65% for this metric for the latest starting date analyzed (1991). The
915 metric that tests for both attributable anthropogenic trend and consistent low-frequency variance
916 (Fig. 17 f), indicates that the seven models have a range of percent area of 20-35% for the 1901-
917 2010 trends, but this range decreases to 10-18% for the 1991-2010 trends. Clearly the variance
918 consistency test proposed here can pose a challenging test for the current models. We have plans
919 to explore other types of variance consistency tests in our future work.

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

921 The analysis presented in this study introduces a framework for trend analysis that has many
922 possible applications and extensions. For surface temperature, there are many figures that are
923 variations on the ones presented here, but were too numerous to include in this article.

924 Therefore, we have created a web site based largely on this analysis, but which contains
925 additional supplemental figures (<http://www.gfdl.noaa.gov/surface-temperature-trends>). For
926 example, the web site contains plots for individual seasons that complement the annual-averaged
927 analysis in this study. We show plots using alternative percentiles (97.5th and 2.5th) instead of
928 95th and 5th, and plots excluding certain low variability models from the analysis, etc. Additional
929 regional plots like Figs. 7-9, including ones for individual seasons, are available, as well as maps

930 for different trend start dates. In addition, a number of plots based on analysis of individual
931 CMIP3 or CMIP5 models, as opposed to multi-model ensemble means, are available.

932

933

934 **8. Summary and Conclusions**

935 The purpose of this analysis has been to introduce and apply a framework for assessing regional
936 surface temperature trends using both the CMIP3 and CMIP5 models and using a multi-model
937 sampling approach. We examined the behavior of the various control runs for the CMIP3 and
938 CMIP5 models, and used the control run variability to help assess whether observed trends were
939 unusual or not compared with the models' internally generated variability. We also used the
940 control run variability to help assess whether observed trends were consistent with trends from
941 the historical (20C3M) simulations—either runs with All Forcings or runs with Natural Forcing
942 Only. In cases for the CMIP5 models where trends were demonstrated to be inconsistent with
943 Natural Forcing Only, but consistent with the All Forcing runs, we conclude that an attributable
944 anthropogenic component is present in the observed trend. For cases, such as the CMIP3 model
945 assessments, where Natural Forcing Only runs are generally not available, we test for detectable
946 trends (compared to internal climate variability) and for consistency between observed and All
947 Forcing historical (20C3M) runs.

948 In the separate CMIP3 and CMIP5 analyses, we generally attempt to give different models equal
949 weight, even when a modeling center provides fewer ensemble members or shorter control runs.
950 Tests are applied at global and regional scales, as well as at individual grid points on the
951 observed data grid where there is sufficient data coverage over the period of the trend. Results

952 are summarized using classification maps and global percent area statistics. Our analysis
953 contains a substantial assessment of the variability in the models, including control run time
954 series for visual inspection, standard deviation maps, spectral analysis, and a low-frequency
955 variance consistency test that is applied to individual models.

956 One of the most important results from the assessment is the identification of regions—and even
957 grid points--where an anthropogenic warming signal is detectable in the observed temperature
958 records. For trends over the period 1901-2010, a large fraction (about 75%) of the global area
959 (with sufficient data coverage over time) has a detectable anthropogenic warming signal.

960 Regions where the observed warming seems to be most commonly underestimated by the models
961 include the southern Ocean, south Atlantic, and off the east coast of Asia. The main regions
962 without detectable warming signals include the high latitude North Atlantic, the eastern U.S., and
963 parts of the eastern Pacific. Moving forward in time, for the much shorter period (1981-2010)
964 the observed warming trends over about 45% of the globe are assessed as having a detectable
965 anthropogenic contribution. These regions include parts of the tropics, subtropics, and mid-
966 latitudes (within about 40-45 degrees of the equator), and a narrow zonally oriented band near
967 the Arctic Ocean. Areas without detectable trends (1981-2010) include much of the eastern
968 Pacific--which is a region influenced by strong interannual variability associated with ENSO--
969 and many extratropical regions poleward of about 40°N and 40°S. The CMIP3 models and the
970 larger sample (23 models) of CMIP5 models yield results similar to those described above,
971 although for these samples we assess only the consistency of trends, and not whether they
972 contain an attributable anthropogenic component (due to the lack of Natural Forcing runs with
973 which to do such an assessment).

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

979 There are numerous examples of modeled trends or variability that are inconsistent with
980 observations in our study. As has been noted in a previous paper using a similar methodology
981 with two climate models (Knutson et al. 2006), disagreement between modeled and observed
982 trends in this type of analysis can occur due to shortcomings of models (internal variability
983 simulation; response to forcing), shortcomings of the specified historical forcings, or problems
984 with the observed data. As one example, Wu and Karoly (2007) and Wu (2010) have noted that
985 disagreement between simulated and observed regional surface temperature trends can result
986 from shortcomings of models in simulating the observed warming associated with the changes of
987 the leading climate variability modes (such as the Arctic Oscillation). Concerning observational
988 uncertainty, the HadCRUT4 data set (Morice et al. 2012) contains 100 ensemble members that
989 attempt to characterize the uncertainties in the observations. We have performed some
990 preliminary tests using these ensembles to assess the spread of observed trend estimates. These
991 tests thus far indicate that even at the regional scale, the spread in trend estimates due to
992 observational uncertainties, as contained in the ensembles, is generally much smaller than the
993 spread in model simulated trends due to the internal variability and differences in forced
994 responses in the historical runs (e.g., Figs. 7-9). However, in some regions (e.g., Mexico), the
995 uncertainty in the observations plays an important role in the assessment of detectable
996 anthropogenic contributions to trends

997 We have attempted to at least partially address the issue of model uncertainties in the simulation
998 of internal climate variability and in the response to historical forcing by using multi-model
999 ensembles and by assessing consistency of both trends and low-frequency variability. When we
1000 apply a two-dimensional screening test (assessing consistency of both the trend and low-
1001 frequency variability) we find that most models tend to be challenged to be consistent on both
1002 tests. Overall, our variance consistency tests suggest that while the CMIP3 and CMIP5 models
1003 provide a plausible representation of internal climate variability, there is considerable scope for
1004 improvement in the model simulations of internal climate variability, apart from their simulation
1005 of trends and variability in response to various forcing agents. From a different perspective, Shin
1006 and Sardeshmukh (2011) have concluded that the CMIP3 models do not simulate historical
1007 trends of temperature and precipitation as realistically as do atmospheric models forced by
1008 observed trends in tropical SSTs—a problem they attribute to model errors as opposed to climate
1009 noise (internal variability).

1010 The CMIP3 and CMIP5 simulations used here represent “ensembles of opportunity” which
1011 cannot necessarily be expected to represent the true structural uncertainty in results, due to
1012 shortcomings/uncertainties in the models and climate forcings. The procedures in our paper
1013 assume that the intrinsic internal variability of climate has not changed significantly since pre-
1014 industrial times, as we are using control run variability from pre-industrial control runs for our
1015 forced-run consistency tests. If anthropogenic forcing had actually *weakened* the intrinsic
1016 variability in the real world, then our estimated uncertainty range around the All-Forcing model
1017 responses would be too wide -- making it overly difficult to conclude that observations were
1018 inconsistent with the All-Forcing runs. Similarly, if anthropogenic forcing had actually
1019 *strengthened* the intrinsic variability in the real world, then our estimated uncertainty range

1020 around the All-Forcing model responses would be too narrow -- making it too easy to conclude
1021 that the observations were inconsistent with the All-Forcing runs.

1022 While the above uncertainty issues lack a final resolution, the methodology shown here can at
1023 least help to quantify the uncertainties associated with the climate change detection and
1024 attribution problem. The results show that when CMIP3 and CMIP5 historical runs are
1025 confronted with observed surface temperature trends, across a wide range of trend start dates, at
1026 various geographical locations around the globe, and even down to the grid point scale, a
1027 pervasive warming signal is found that is generally much more consistent with simulations that
1028 include anthropogenic forcing than with simulations that include either no forcing changes
1029 (control runs) or that include only natural forcing agents (solar, volcanic). Our conclusions about
1030 detectable anthropogenic contributions to the trends provides further support for the claim of a
1031 substantial human influence on climate, via anthropogenic forcing agents such as increased
1032 greenhouse gases. A future enhancement of our analysis would include an attempt to quantify
1033 the contributions of different natural and anthropogenic forcing agents in the CMIP5 All-Forcing
1034 and Natural Forcing Only historical runs. This would provide a more direct assessment of the
1035 relative influence of different forcing agents on the observed temperature trends at the regional
1036 scale.

1037

1038 Acknowledgements. We thank the Met Office Hadley Centre and the Climatic Research Unit,
1039 Univ. of East Anglia, for making the HadCRUT4 data set available to the research community.
1040 We thank the modeling groups participating in CMIP3 and CMIP5, and PCMDI for generously

1041 making the model output used in our report available to the community, and we thank three
1042 anonymous reviewers for their helpful comments on the manuscript.

1043

1044

1045

1046

1047 **References**

1048 Allen, M. R., and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting. Part
1049 I: Theory. *Clim. Dyn.*, **21**, 477-491.

1050

1051 Allen, M. R., and S.F.B. Tett, 1999: Checking for model consistency in optimal fingerprinting.
1052 *Clim. Dyn.*, **15**, 419-434.

1053

1054 Annan, J. D. and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res.*
1055 *Lett.*, **37**, L02703, doi:10.1029/2009GL041994.

1056

1057 Hasselmann, K., 1997: Multi-pattern fingerprint method for detection and attribution of climate
1058 change. *Clim. Dyn.*, **13**, 601-612.

- 1059 Hegerl, G. C., F. W. Zwiers, P. Braconnot, N. P. Gillett, Y. Luo, J. A. Marengo Orsini, N.
1060 Nicholls, J. E. Penner, and P. A. Stott, 2007: Understanding and attributing climate change. In
1061 *Climate Change 2007: The Physical Science Basis*. [Solomon, S., D. Qin, M. Manning, Z. Chen,
1062 M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press,
1063 Cambridge, United Kingdom and New York, NY, USA, 996 pp.
- 1064
- 1065 Hegerl, G. C., et al. 2009: Good practice guidance paper on detection and attribution related to
1066 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)
1067 [material/ipcc_good_practice_guidance_paper_anthropogenic.pdf](http://www.ipcc.ch/pdf/supporting-material/ipcc_good_practice_guidance_paper_anthropogenic.pdf)
- 1068 Hegerl, G.C., H. v. Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996:
1069 Detecting greenhouse gas induced climate change with an optimal fingerprint method. *J.*
1070 *Climate*, **9**, 2281-2306.
- 1071 Karoly, D.J., and Q. Wu, 2005: Detection of regional surface temperature trends. *J. Clim.*, **18**,
1072 4337–4343.
- 1073 Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: Reassessing
1074 biases and other uncertainties in sea-surface temperature observations measured in situ
1075 since 1850, part 2: biases and homogenization. *J. Geophys. Res.*, **116**, D14104,
1076 doi:10.1029/2010JD015220.
- 1077 Knutson, T.R., T.L. Delworth, K.W. Dixon, and R.J. Stouffer, 1999: Model assessment of
1078 regional surface temperature trends (1949-1997). *J. Geophys. Res.*, **104**, 30981–30996.
- 1079

1080 Knutson, T.R., et al., 2006: Assessment of twentieth-century regional surface temperature trends
1081 using the GFDL CM2 coupled models. *J. Clim.*, **19**, 1624–1651.

1082

1083 Mahlstein, I., R. Knutti, S. Solomon, and R. W. Portmann, 2011: Early onset of significant local
1084 warming in low latitude countries. *Environ. Res. Lett.*, **6**, 034009, doi:10.1088/1748-
1085 9326/6/034009.

1086

1087 Meehl, G. A. et al., 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change
1088 research. *Bull. Amer. Meteor. Soc.* **88**, 1383–1394.

1089

1090 Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in
1091 global and regional temperature change using an ensemble of observation estimates: The
1092 HadCRUT5 data set. *J. Geophys. Res.*, 117, D08101, doi:10.1029/2011JD017187.

1093

1094 Rind, D., M. Chin, G. Feingold, D. Streets, R. A. Kahn, S. E. Schwartz, and H. Yu, 2009:
1095 Modeling the effects of aerosols on climate. In *Atmospheric Aerosol Properties and*
1096 *Climate Impacts, A Report by the U.S. Climate Change Science Program and the*
1097 *Subcommittee on Global Change Research*. [Mian Chin, Ralph A. Kahn, and Stephen E.
1098 Schwartz (eds.)]. National Aeronautics and Space Administration, Washington, D.C.,
1099 USA.

1100

- 1101 Sakaguchi, K. X. Zeng, and M. A. Brunke, 2012: Temporal- and Spatial-scale dependence of
1102 three CMIP3 climate models in simulating the surface temperature trend in the twentieth
1103 century. *J. Climate*, **25**, 2456-2470.
- 1104
- 1105 Santer, B. D., T. M. L. Wigley, and P. D. Jones, 1993: Correlation method in fingerprint
1106 detection studies. *Clim. Dyn.*, **8**, 265-276.
- 1107
- 1108 Schneider, T., and I.M. Held, 2001: Discriminants of twentieth-century changes in Earth surface
1109 temperatures. *J. Clim.*, **14**, 249–254.
- 1110
- 1111 Shin, S.-I., and P. D. Sardeshmuhk, 2011: Critical influence of the pattern of tropical ocean
1112 warming on remote climate trends. *Clim. Dyn.*, **36**, 1577-1591.
- 1113
- 1114 Stenchikov, G., T. L. Delworth, V. Ramaswamy, R. J. Stouffer, A. Wittenberg, and F. Zeng,
1115 2009: Volcanic signals in oceans. *J. Geophys. Res.*, **114**, D16104. doi:
1116 10.1029/2008JD011673.
- 1117
- 1118 Stouffer, R. J., S. Manabe, and K. Y. Vinnikov, 1994: Model assessment of the role of natural
1119 variability in recent global warming. *Nature*, **367**, 634-636.
- 1120
- 1121 Stouffer R. J., Hegerl G. C. and Tett S. F. B. (2000): A comparison of Surface Air Temperature
1122 Variability in Three 1000-Year coupled Ocean-Atmosphere Model Integrations. *J.*
1123 *Climate*, **13**, 513-537.

- 1124
- 1125 Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An overview of CMIP5 and the experiment
1126 design. *Bull. Amer. Meteor. Soc.*, **93**, 485-498
- 1127 .
- 1128 Vecchi, G. A., and A. T. Wittenberg, 2010: El Nino and our future climate: Where do we stand?
1129 *Wiley Interdisciplinary Reviews: Climate Change*, **1**, 260-270. doi: 10.1002/wcc.33.
- 1130
- 1131 Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations?
1132 *Geophys. Res. Lett.*, **36**, L12702. doi: 10.1029/2009GL038710.
- 1133
- 1134 Wu, Q., and D. J. Karoly (2007): Implications of changes in the atmospheric circulation on the
1135 detection of regional surface air temperature trends, *Geophys. Res. Lett.*, **34**, L08703,
1136 doi:10.1029/2006GL028502.
- 1137
- 1138 Wu, Q. (2010): Associations of diurnal temperature range change with the leading climate
1139 variability modes during the Northern Hemisphere wintertime and their implication on
1140 the detection of regional climate trends, *J. Geophys. Res.*, **115**, D19101,
1141 doi:10.1029/2010JD014026.
- 1142

1143 Yang, X., A. Rosati, S. Zhang, T. L. Delworth, R. G. Gudgel, R. Zhang, G. Vecchi, W.
1144 Anderson, Y.-S. Chang, T. DelSole, K. Dixon, R. Msadek, W. F. Stern, A. Wittenberg,
1145 and F. Zeng, 2013: A predictable AMO-like pattern in GFDL's fully-coupled ensemble
1146 initialization and decadal forecasting system. *J. Climate*, in press. doi: 10.1175/JCLI-D-
1147 12-00231.1.

1148

1149

1150

1151

1152 Figure Captions

1153

1154 Fig. 1. Time series of global-mean annual-mean surface air temperature (2 m) anomalies from
1155 the CMIP3 (a, b) and CMIP5 (c, d) preindustrial control runs (black curves). Observed global
1156 mean surface temperature (HadCRUT4, combining SST and land surface air temperature
1157 anomalies) is also shown in blue on the diagrams for comparison. The blue curves labeled
1158 “Residual (HadCRUT4...” were created by subtracting the multi-model ensemble mean surface
1159 temperature (from masked SSTs and land surface air temperatures from the 20C3M All Forcing
1160 historical runs for either CMIP3 or CMIP5) from the observed temperature. Orange straight
1161 lines (one or two segments) through the control run time series depict the long term linear drift.
1162 The long term drift over the year range shown is calculated at each grid point and then subtracted
1163 from the model control run series before performing further analysis in our study. Short vertical
1164 orange segments denote two places where control runs were divided into two separate segments
1165 and the linear drift computed separately for each segment. In that case, the residuals from the
1166 drift were formed and then combined back into a single series. The various curves in the figure
1167 have been displaced vertically by arbitrary constants for visual clarity. Curves labeled with a ‘*’
1168 denote CMIP3 models that did not include volcanic forcing in their historical runs. Curves
1169 labeled with a ‘(0)’ were excluded from the remainder of our analysis due to various issues such
1170 as discontinuities in time series, short record length, or unavailable sea surface temperature data
1171 in the CMIP3 archive. Vertical axis tick mark spacing is 0.2°C.

1172

1173 Fig. 2. Standard deviation ($^{\circ}\text{C}$) of annual mean surface air temperature from the CMIP3 pre-
1174 industrial control runs. The long term linear drifts (time periods identified by the orange line
1175 segments in Fig. 1 a,b) were removed prior to computing the standard deviations. The individual
1176 plots are labeled with the name of the model/center and classified as “Non-V” (non-volcanic) or
1177 “V” (volcanic) depending on whether that model’s historical run used in this study included
1178 volcanic forcing. Note that the control runs on which the figure are based do not have episodic
1179 volcanic forcing and have been masked for observed missing data periods. The panel labeled
1180 “All-model mean” is the average of the individual model panels. The final panel (“HadCRU4
1181 Observed”) is an observational estimate of internal variability of SST (oceanic regions) and
1182 surface air temperature (land regions) constructed by removing the CMIP3 eight-model
1183 ensemble (All Forcing; Volcanic models only) estimate of the forced climate response from the
1184 observed temperature record over 1949-2010. The number at upper right on each panel is the
1185 spatial correlation of that model’s field with the observed standard deviation field.

1186

1187 Fig. 3. As in Fig. 2 but for the 23 CMIP5 models analyzed in this study. The final panel
1188 (“HadCRUT4 Observed”) is an observational estimate of internal variability of SST (oceanic
1189 regions) and surface air temperature (land regions) constructed by removing the CMIP5 23-
1190 model ensemble (All Forcing) estimate of the forced climate response from the observed
1191 temperature record over 1949-2010.

1192

1193 Fig. 4. Time series of global mean surface temperature anomalies (combined SST and land
1194 surface air temperature) from observations (HadCRUT4; black curves) in degrees Celsius. The

1195 red curves in a-c depict the 5th and 95th percentiles of annual mean anomalies for the multi-model
1196 mean (thick) or of single model realizations (thin lines, gray stippling) for the CMIP3 (a, b) or
1197 CMIP5 (c) 20C3M historical All Forcing runs in degrees Celsius. The mean curve is not shown
1198 but lies approximately midway between the 5th and 95th percentiles. The series in (a) are from
1199 eight CMIP3 models run with volcanic forcing. The historical runs in (b) include 19 CMIP3
1200 models with and without volcanic forcing (as identified in Fig. 1 (a,b). All of the 23 CMIP5
1201 model runs included in the computations (c) incorporated volcanic forcing. In (d) the blue
1202 curves are based on seven CMIP5 models that had Natural Forcing Only runs extending through
1203 2010. See text for description of how the confidence limits were computed. The time series
1204 have been re-centered so that the ensemble mean value, averaged for the years 1881-1920, is
1205 zero. Model data were masked with the observed spatially and temporally evolving missing data
1206 mask. The total number of individual experiments included in each panel was: a) 26; b) 51; c)
1207 79; and d) 25.

1208

1209 Fig. 5. Variance spectra as a function of frequency for observed global mean surface
1210 temperature (combined SST and land surface air temperature), in black with 90% confidence
1211 intervals shown by the shading, plotted against spectra for the individual (a) CMIP3 and (b)
1212 CMIP5 All Forcing historical runs with Volcanic forcing (red) based on the time series in Fig. 4
1213 (a,c). The spectra in (c) and (d) are based on residual observed or model historical run time
1214 series, where the multi-model ensemble surface temperature from the 20C3M All Forcing (with
1215 volcanic) historical runs (CMIP3 or CMIP5) is subtracted from the observed and from each
1216 model's global mean temperature series to form residual time series prior to computing the
1217 spectra (see text for details).

1218

1219 Fig. 6. Map illustrating averaging regions examined in Figs. 7-9. Regions abbreviations
1220 including: Euro = Europe; NAs = Northern Asia; SAs = Southern Asia; Afr = Africa; IO =
1221 Indian Ocean; Aus = Australia; TWP = Tropical western Pacific; TEP = Tropical eastern Pacific;
1222 IOWP = Tropical Indian Ocean/western Pacific warm pool; NP = North Pacific; AL = Alaska;
1223 SEUS = Southeastern United States; ConUS = Continental United States; RofUS = rest of
1224 continental United States, other than SEUS; SAmer = South America; Can = Canada; NAtl =
1225 North Atlantic; SNA = Subtropical North Atlantic; TNA = Tropical North Atlantic (Main
1226 Development Region); SAtl = South Atlantic.

1227

1228 Fig. 7. Trends ($^{\circ}\text{C}/100\text{ yr}$) in area-averaged annual-mean surface temperature as a function of
1229 starting year, with all trends ending in 2010. The black curves are trends from observations
1230 (HadCRUT4), where observational uncertainty is depicted as a range showing the 5th to 95th
1231 percentile ranges of trends obtained using the 100-member HadCRUT4 ensemble. Red curves
1232 are ensemble means of the All-Forcing runs from 23 CMIP5 models. Blue curves are ensemble
1233 means for Natural Forcing Only runs using a subset of seven CMIP5 models that had Natural
1234 Forcing runs extending to 2010. See Fig. 6 for definitions of averaging regions. The different
1235 models are weighted equally for the multi-model ensemble means, regardless of the number of
1236 ensemble members they had. The pink shading shows the 5th to 95th percentile range of the
1237 distribution of trends obtained by combining random samples from each of the 23 CMIP5 model
1238 control runs together with the corresponding model's ensemble-mean forced trend (All Forcing
1239 runs) to create a total multi-model distribution of trends that reflects uncertainty in both the

1240 forced response and the influence of internal climate variability. The blue-shaded region shows
1241 the same, but for the seven models with Natural Forcing Only runs and their seven control runs.
1242 Violet shading indicates where the pink- and blue-shaded regions overlap. Gaps in the curves
1243 indicate inadequate data coverage for a trend-to-2010 for those start years. Requirements
1244 include: 33% areal coverage to define an index time series point for a month, 40% of months
1245 available for a year to be non-missing, and 20% of all years available in each of five equal
1246 segments for a time series have adequate coverage for a trend. The seven CMIP5-model subset
1247 used here and in subsequent assessment figures that incorporate Natural Forcing runs include:
1248 CanESM2, CNRM-CM5, CSIRO-Mk3-6-0, FGOALS-g2, HadGEM2-ES, IPSL-CM5A-LR, and
1249 NorESM1-M.

1250

1251 Fig. 8. As in Fig. 7, but for additional regions as labeled (see Fig. 6).

1252

1253 Fig. 9. As in Fig. 7, but for additional regions as labeled (see Fig. 6). The left column is based
1254 on All Forcing runs from eight CMIP3 models that include volcanic forcing in their historical
1255 simulations, and the eight corresponding control runs (without volcanic forcing). The middle
1256 column is based on All Forcing and control runs from all 23 CMIP5 models. The right column is
1257 based on All Forcing, Natural Forcing Only, and control runs from the same sets of CMIP5
1258 models as used in Figs. 7 and 8 (see Fig. 7 caption).

1259

1260 Fig. 10. Geographical distribution of surface temperature trends (1901-2010) in: (a,b)
1261 HadCRUT4 observations; (c) CMIP3 eight-model ensemble mean (All Forcing, volcanic
1262 models); d) CMIP5 23-model ensemble mean (All Forcing, volcanic models). Unit: degrees C
1263 per 100 yr. In (e, f) the observed trend is assessed in terms of the multi-model ensemble mean
1264 trends and variability in the historical forcing and control runs (CMIP3 and CMIP5). The
1265 different colors in (e, f) depict different categories of assessment result; the categories are listed
1266 in the legends below panels e and f. Panel (e) compares observed trends with trends from 8
1267 CMIP3 All Forcing models and their 8 control runs. Panel (f) compares observed trends with
1268 trends from 23 All-Forcing CMIP5 models and their 23 control runs and with the 7 All Forcing
1269 CMIP5 model subset and their 7 control runs.

1270

1271 Fig. 11. Same as Fig. 10 but for trends from 1951 to 2010.

1272

1273 Fig. 12. Same as Fig. 10 but for trends from 1981 to 2010.

1274

1275 Fig. 13. Summary assessment of observed vs. model ensemble-mean trends-to-2010. The
1276 percent of global analyzed areas meeting certain criteria (see graph labels) are shown as a
1277 function of start year (all trends ending in 2010). a) Assessments of the 8 CMIP3 (solid lines) vs.
1278 the 23 CMIP5 (dashed lines) multi-model ensemble mean (historical 20C3M All-Forcing runs
1279 with volcanic forcing and associated control runs). b) Assessment of the CMIP5 multi-model
1280 ensemble means and control runs using all 23 CMIP5 models and their 23 control runs for the
1281 All-Forcing assessment and the seven-model subset of CMIP5 models (with Natural Forcing

1282 Only runs extending to 2010) and their seven control runs for the Natural-Forcing assessment.
1283 The black curves are the sum of the red and orange curves; the sum of black + light green + dark
1284 green + blue = 100%.

1285

1286 Fig. 14. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP3 multi-model
1287 (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100
1288 yr. The observed trend is assessed in terms of the eight individual CMIP3 models (trends and
1289 variability) in (c-k). Panels (c-k) show the fraction of the eight individual CMIP3 models whose
1290 historical All Forcing runs meet the criteria listed above each panel. The criteria are: c) no
1291 detectable change; d) warming that is detectable but significantly greater than simulated in the
1292 All Forcing runs; e) warming that is detectable and consistent with the All Forcing runs; f)
1293 warming that is detectable but significantly less than simulated in the All Forcing runs; g)
1294 warming that is detectable; h) warming that is detectable and either consistent with or greater
1295 than the simulated (All Forcing) runs; i) observed and simulated trends are consistent (though the
1296 observed trend may not be detectable); j) observed and simulated internal low-frequency
1297 variability are consistent; and k) conditions for (i) and (j) are both satisfied (i.e., the simulated
1298 variability and trend are both consistent with observations. The white numbers at the bottom of
1299 maps c-k indicate the area-weighted global average of the mapped fields.

1300

1301 Figure 15. Same as Fig. 14, but for 23 CMIP5 models with volcanic forcing.

1302

1303 Fig. 16. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP5 multi-model
1304 (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100
1305 yr. The observed trend is assessed in terms of trend and variability using the seven CMIP5
1306 models that had available an All Forcing ensemble and an ensemble of Natural Forcing Only
1307 runs extending to 2010. Panels (c-l) show the fraction of the seven individual CMIP5 models at
1308 each grid point whose All Forcing, Natural Forcing Only, and control runs together meet the
1309 criteria listed above the panel. The criteria are: c) no detectable change; d) warming that is
1310 detectable (inconsistent with Natural Forcing runs) but significantly less than simulated in the
1311 All Forcing runs; e) attributable anthropogenic warming that is detectable (inconsistent with
1312 Natural Forcing Only runs) and consistent with the All Forcing runs; f) attributable
1313 anthropogenic warming that is significantly greater than simulated in the All Forcing runs; g)
1314 warming that is detectable; h) total attributable to anthropogenic warming (i.e., sum of (e) and
1315 (f)); i) observed and simulated trends are consistent (though the observed trend may not be
1316 detectable); j) observed and simulated internal low-frequency variability are consistent; k)
1317 conditions for (i) and (j) are both satisfied (i.e., the simulated variability and trend are both
1318 consistent with observations; and l) conditions for (h) and (j) are both satisfied (i.e., there is
1319 attributable anthropogenic warming and low-frequency variance is consistent with observations).

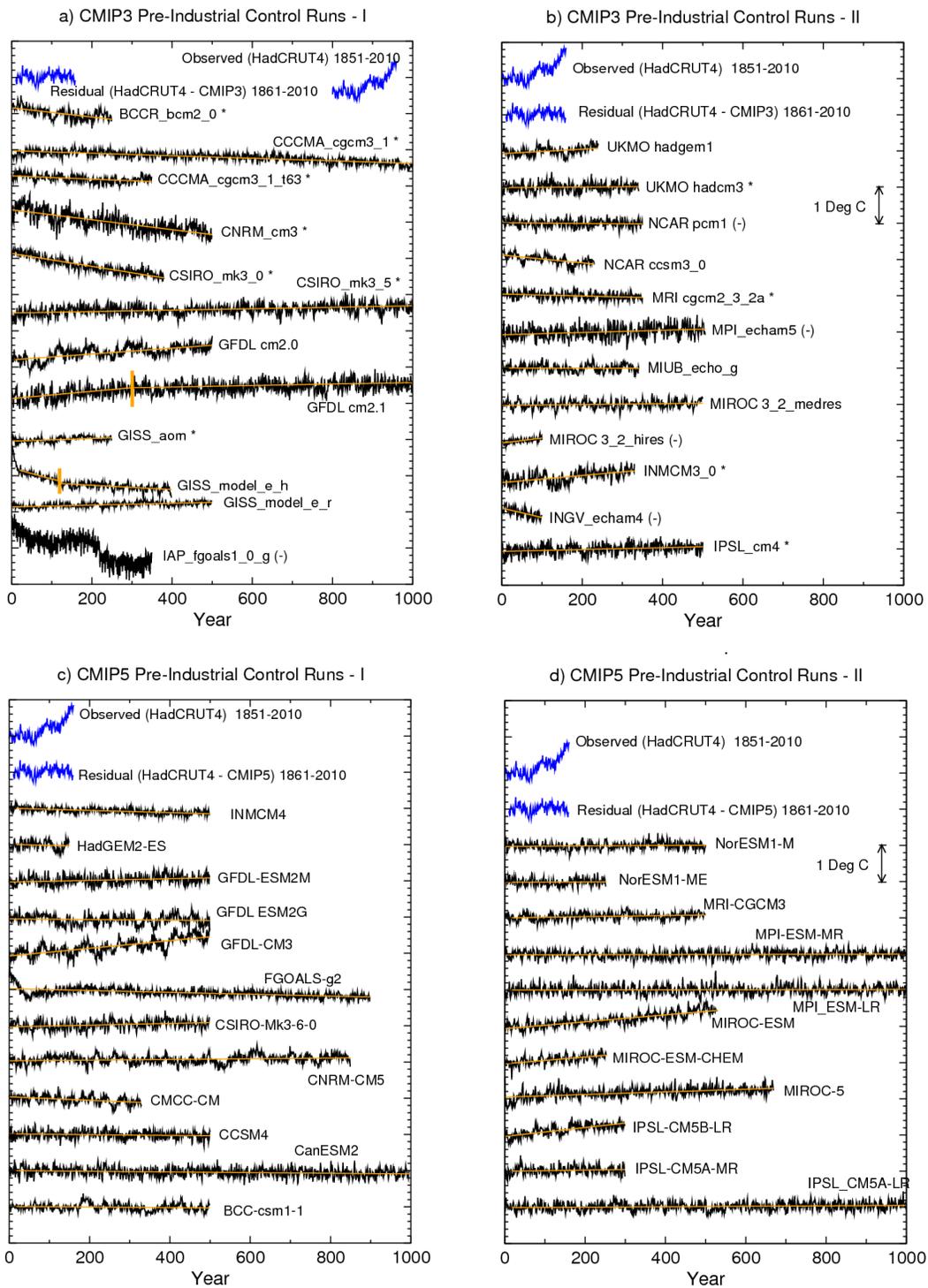
1320

1321 Fig. 17. Individual CMIP3 (a, c) and CMIP5 (b, d, e, f) models are assessed for consistency with
1322 detectable observed surface temperature trends-to-2010 (a-d), for attributable anthropogenic
1323 trends (e, f), and for consistency of simulated internal variability (c, d, f). Trend results are
1324 shown for start years from 1901 to 1991 (all trends ending in 2010). Plotted is the percent of
1325 analyzed global area where each individual model's (legend) multi-realization ensemble mean

1326 forced trend and internal variability meet the criteria listed above the panel. The trends are
1327 analyzed at each grid point where there is sufficient temporal data coverage for the trend in
1328 question (see text). Note that panel (f) includes areas where the observed trend is detectable but
1329 greater than simulated, whereas panel (d) includes only areas with trends that are detectable and
1330 consistent with simulations.

1331

Model control runs: simulated internal variability of global temperature



1333 Fig. 1. Time series of global-mean annual-mean surface air temperature (2 m) anomalies from the CMIP3
1334 (a, b) and CMIP5 (c, d) preindustrial control runs (black curves). Observed global mean surface
1335 temperature (HadCRUT4, combining SST and land surface air temperature anomalies) is also shown in
1336 blue on the diagrams for comparison. The blue curves labeled “Residual (HadCRUT4...)” were created
1337 by subtracting the multi-model ensemble mean surface temperature (from masked SSTs and land surface
1338 air temperatures from the 20C3M All Forcing historical runs for either CMIP3 or CMIP5) from the
1339 observed temperature. Orange straight lines (one or two segments) through the control run time series
1340 depict the long term linear drift. The long term drift over the year range shown is calculated at each grid
1341 point and then subtracted from the model control run series before performing further analysis in our
1342 study. Short vertical orange segments denote two places where control runs were divided into two
1343 separate segments and the linear drift computed separately for each segment. In that case, the residuals
1344 from the drift were formed and then combined back into a single series. The various curves in the figure
1345 have been displaced vertically by arbitrary constants for visual clarity. Curves labeled with a ‘*’ denote
1346 CMIP3 models that did not include volcanic forcing in their historical runs. Curves labeled with a ‘(0)’
1347 were excluded from the remainder of our analysis due to various issues such as discontinuities in time
1348 series, short record length, or unavailable sea surface temperature data in the CMIP3 archive. Vertical
1349 axis tic mark spacing is 0.2°C.

1350

1351

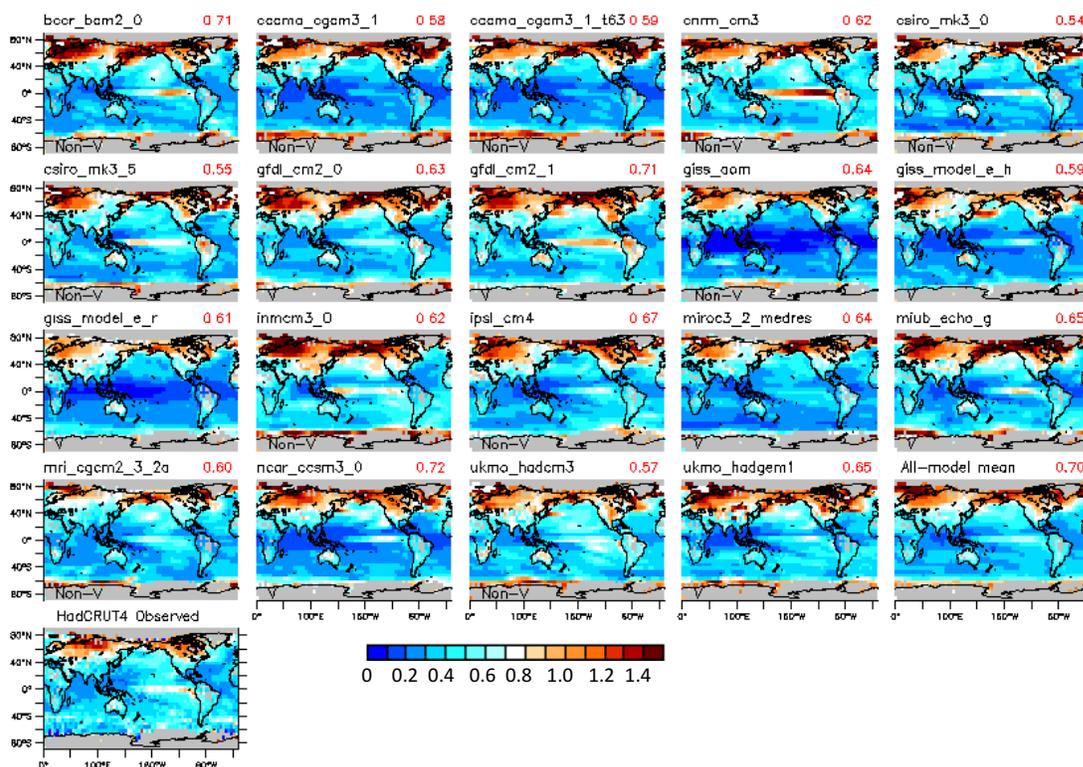


Fig. 2. Standard deviation ($^{\circ}\text{C}$) of annual mean surface air temperature from the CMIP3 pre-industrial control runs. The long term linear drifts (time periods identified by the orange line segments in Fig. 1 a,b) were removed prior to computing the standard deviations. The individual plots are labeled with the name of the model/center and classified as “Non-V” (non-volcanic) or “V” (volcanic) depending on whether that model’s historical run used in this study included volcanic forcing. 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. The panel labeled “All-model mean” is the average of the individual model panels. The final panel (“HadCRUT4 Observed”) is an observational estimate of internal variability of SST (oceanic regions) and surface air temperature (land regions) constructed by removing the CMIP3 eight-model ensemble (All Forcing; Volcanic models only) estimate of the forced climate response from the observed temperature record over 1949-2010. The number at upper right on each panel is the spatial correlation of that model’s field with the observed standard deviation field.

1352

1353

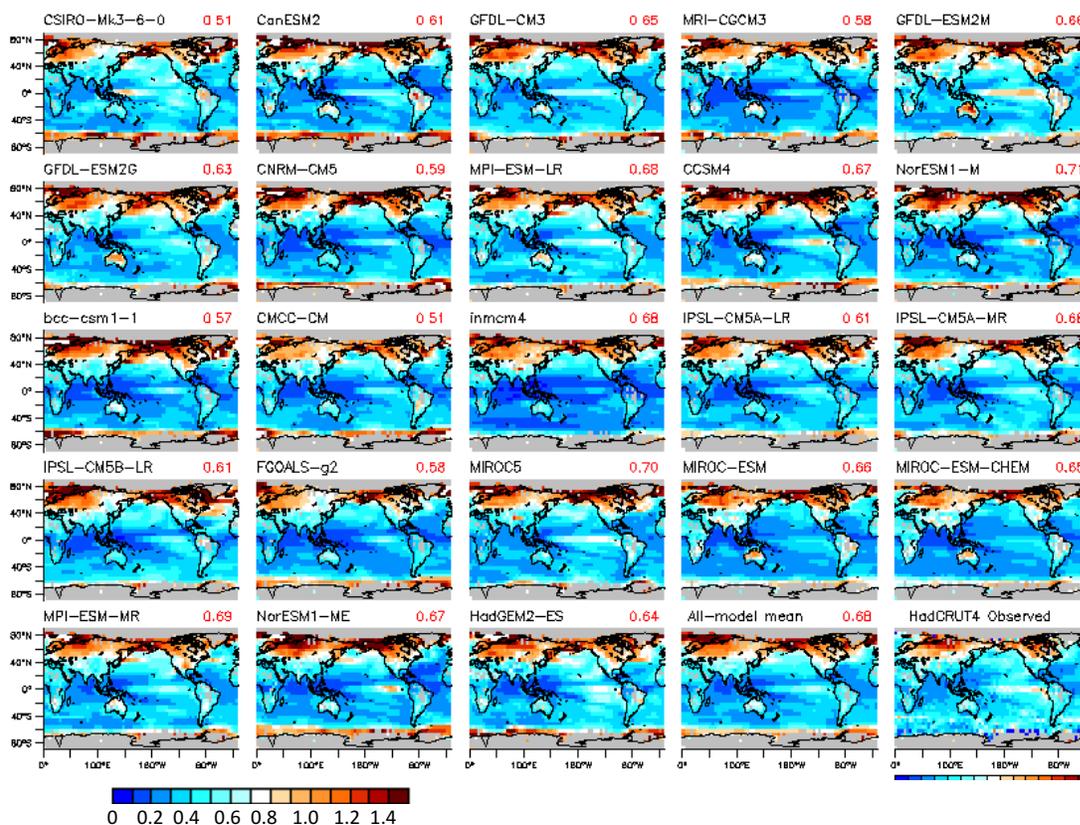


Fig. 3. As in Fig. 2 but for the 23 CMIP5 models analyzed in this study. The final panel (“HadCRUT4 Observed”) is an observational estimate of internal variability of SST (oceanic regions) and surface air temperature (land regions) constructed by removing the CMIP5 23-model ensemble (All Forcing) estimate of the forced climate response from the observed temperature record over 1949-2010.

1354

Global Mean Surface Temperature Anomalies

1355

1356

1357

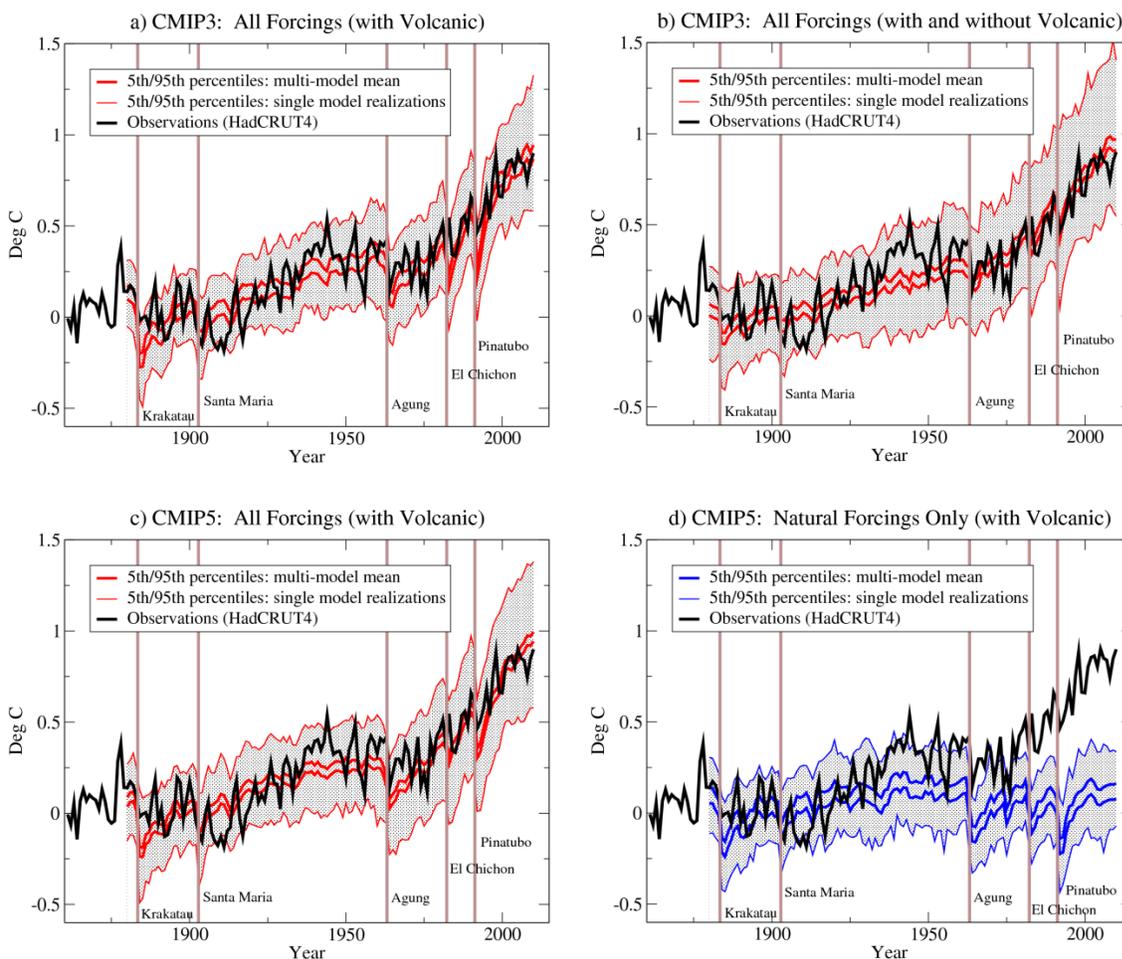


Fig. 4. Time series of global mean surface temperature anomalies (combined SST and land surface air temperature) from observations (HadCRUT4; black curves) in degrees Celsius. The red curves in a-c depict the 5th and 95th percentiles of annual mean anomalies for the multi-model mean (thick) or of single 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.

1358

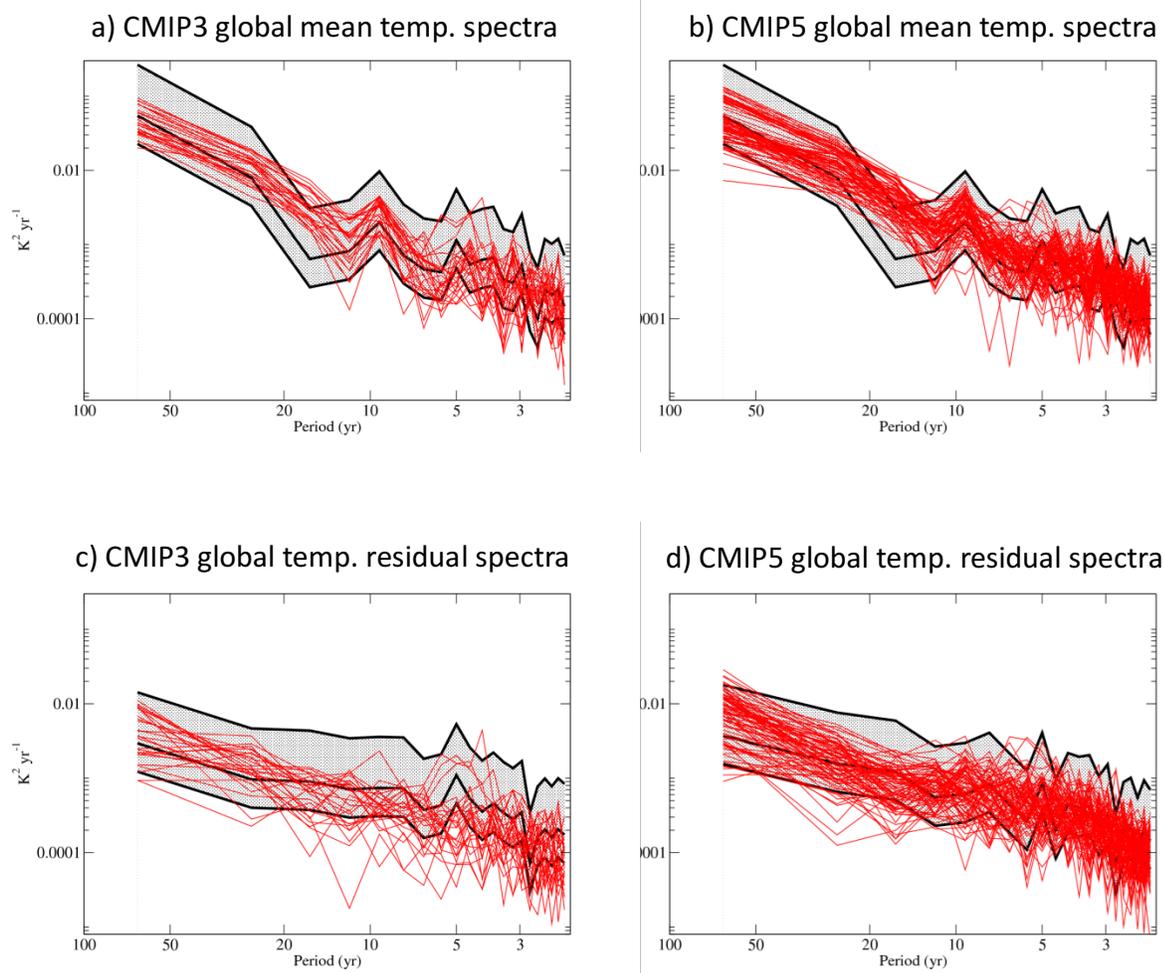
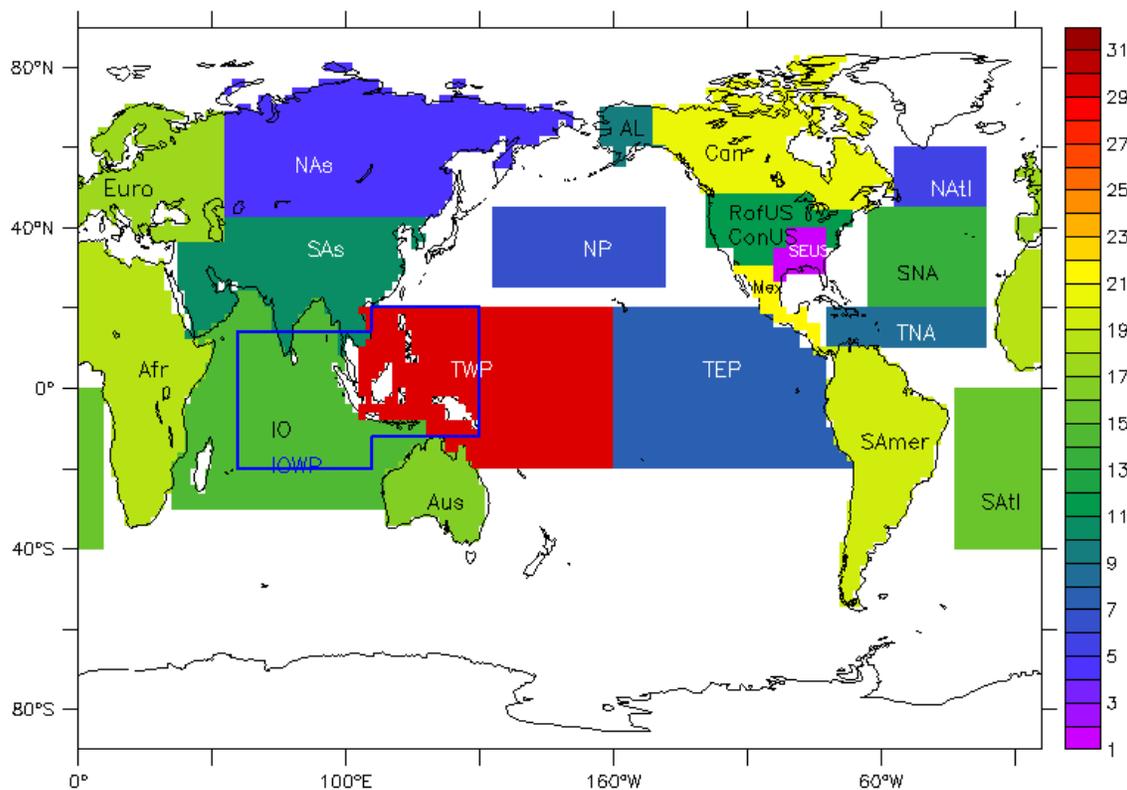


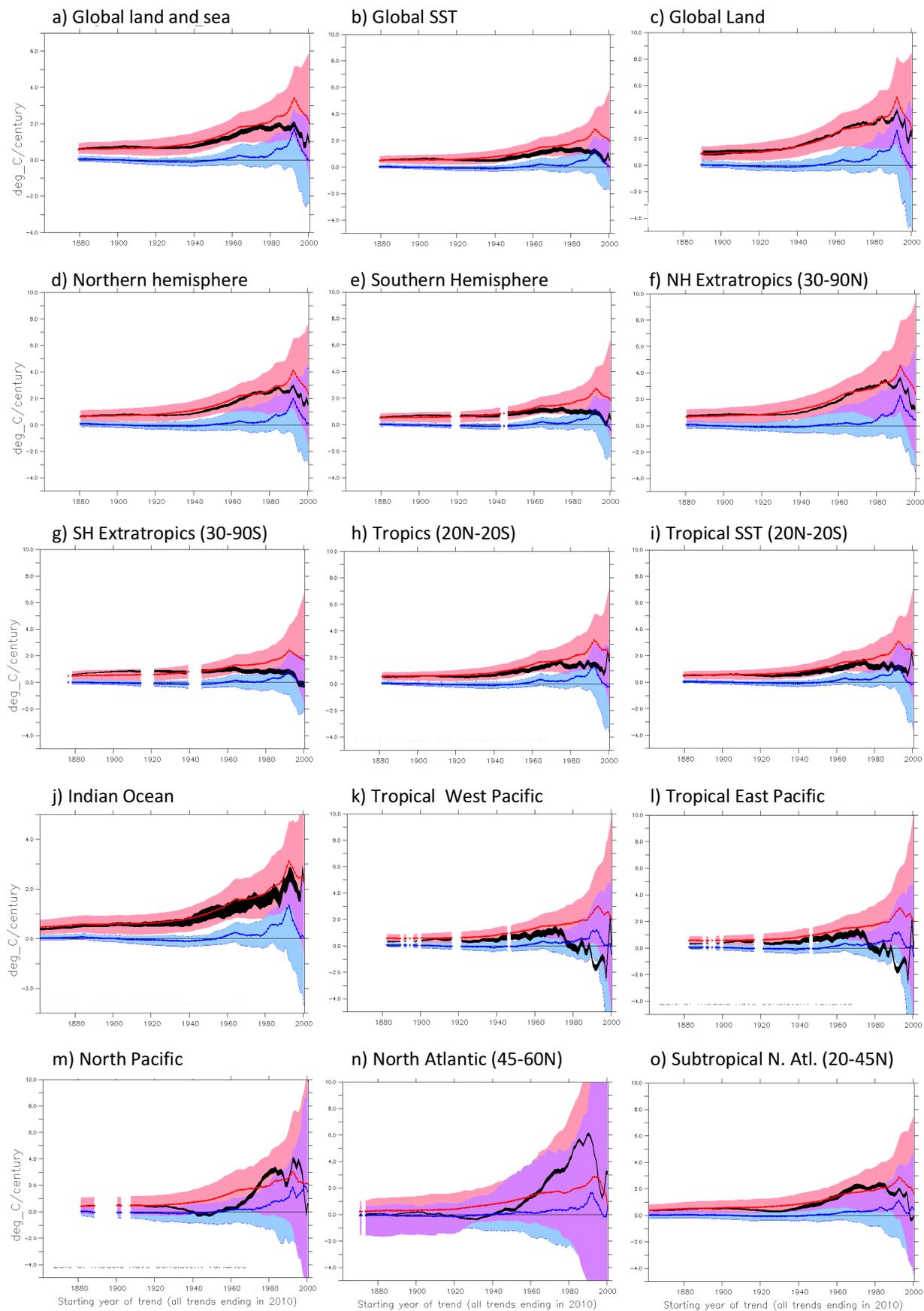
Fig. 5. Variance spectra as a function of frequency for observed global mean surface temperature (combined SST and land surface air temperature), in black with 90% confidence intervals shown by the shading, plotted against spectra for the individual (a) CMIP3 and (b) CMIP5 All Forcing historical runs 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).



1359

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.

1360



1361 Fig. 7. Trends ($^{\circ}\text{C}/100$ yr) in area-averaged annual-mean surface temperature as a function of
1362 starting year, with all trends ending in 2010. The black curves are trends from observations
1363 (HadCRUT4), where observational uncertainty is depicted as a range showing the 5th to 95th
1364 percentile ranges of trends obtained using the 100-member HadCRUT4 ensemble. Red curves
1365 are ensemble means of the All-Forcing runs from 23 CMIP5 models. Blue curves are ensemble
1366 means for Natural Forcing Only runs using a subset of seven CMIP5 models that had Natural
1367 Forcing runs extending to 2010. See Fig. 6 for definitions of averaging regions. The different
1368 models are weighted equally for the multi-model ensemble means, regardless of the number of
1369 ensemble members they had. The pink shading shows the 5th to 95th percentile range of the
1370 distribution of trends obtained by combining random samples from each of the 23 CMIP5 model
1371 control runs together with the corresponding model's ensemble-mean forced trend (All Forcing
1372 runs) to create a total multi-model distribution of trends that reflects uncertainty in both the
1373 forced response and the influence of internal climate variability. The blue-shaded region shows
1374 the same, but for the seven models with Natural Forcing Only runs and their seven control runs.
1375 Violet shading indicates where the pink- and blue-shaded regions overlap. Gaps in the curves
1376 indicate inadequate data coverage for a trend-to-2010 for those start years. Requirements
1377 include: 33% areal coverage to define an index time series point for a month, 40% of months
1378 available for a year to be non-missing, and 20% of all years available in each of five equal
1379 segments for a time series have adequate coverage for a trend. The seven CMIP5-model subset
1380 used here and in subsequent assessment figures that incorporate Natural Forcing runs include:
1381 CanESM2, CNRM-CM5, CSIRO-Mk3-6-0, FGOALS-g2, HadGEM2-ES, IPSL-CM5A-LR, and
1382 NorESM1-M.

1383

1384

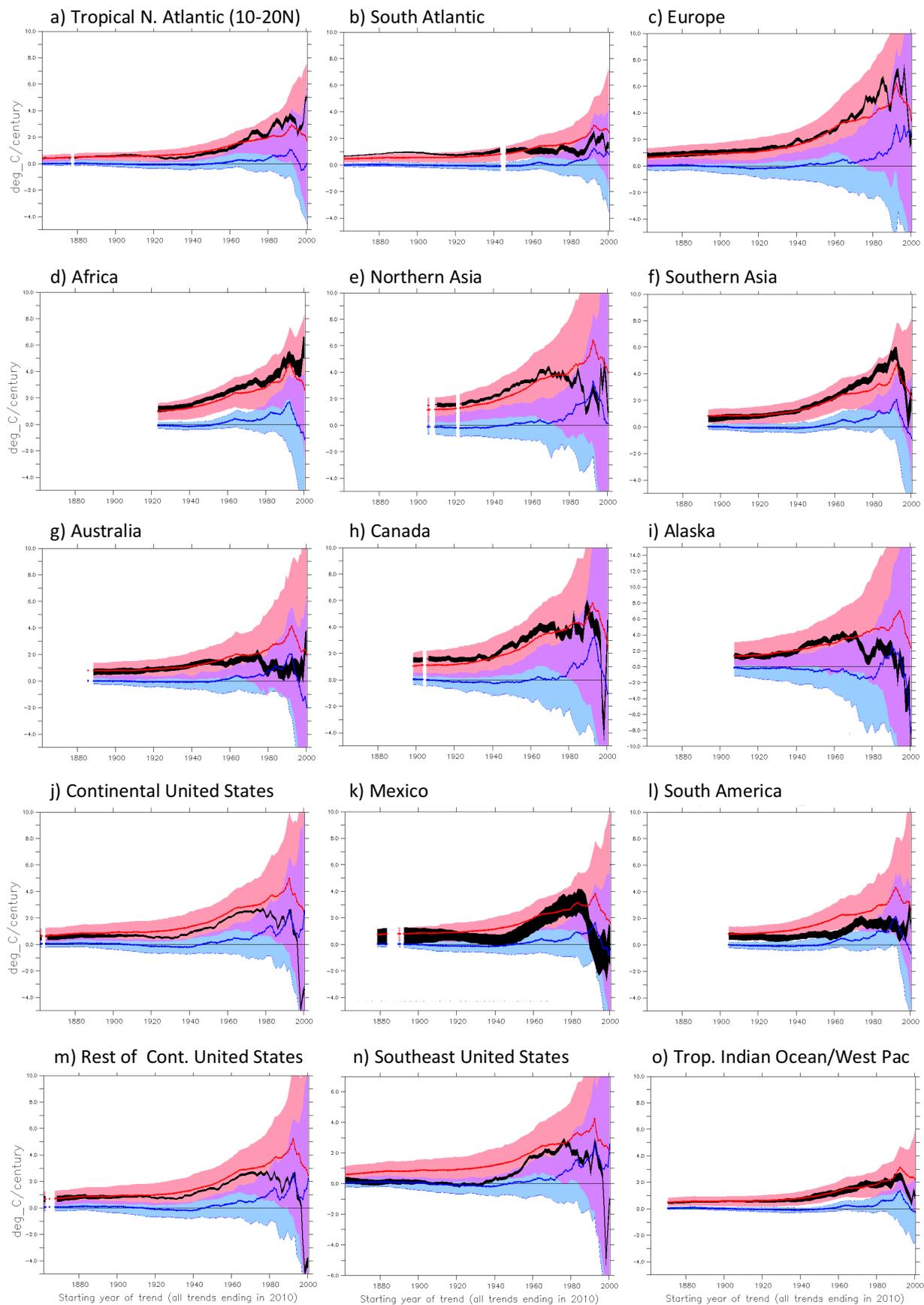
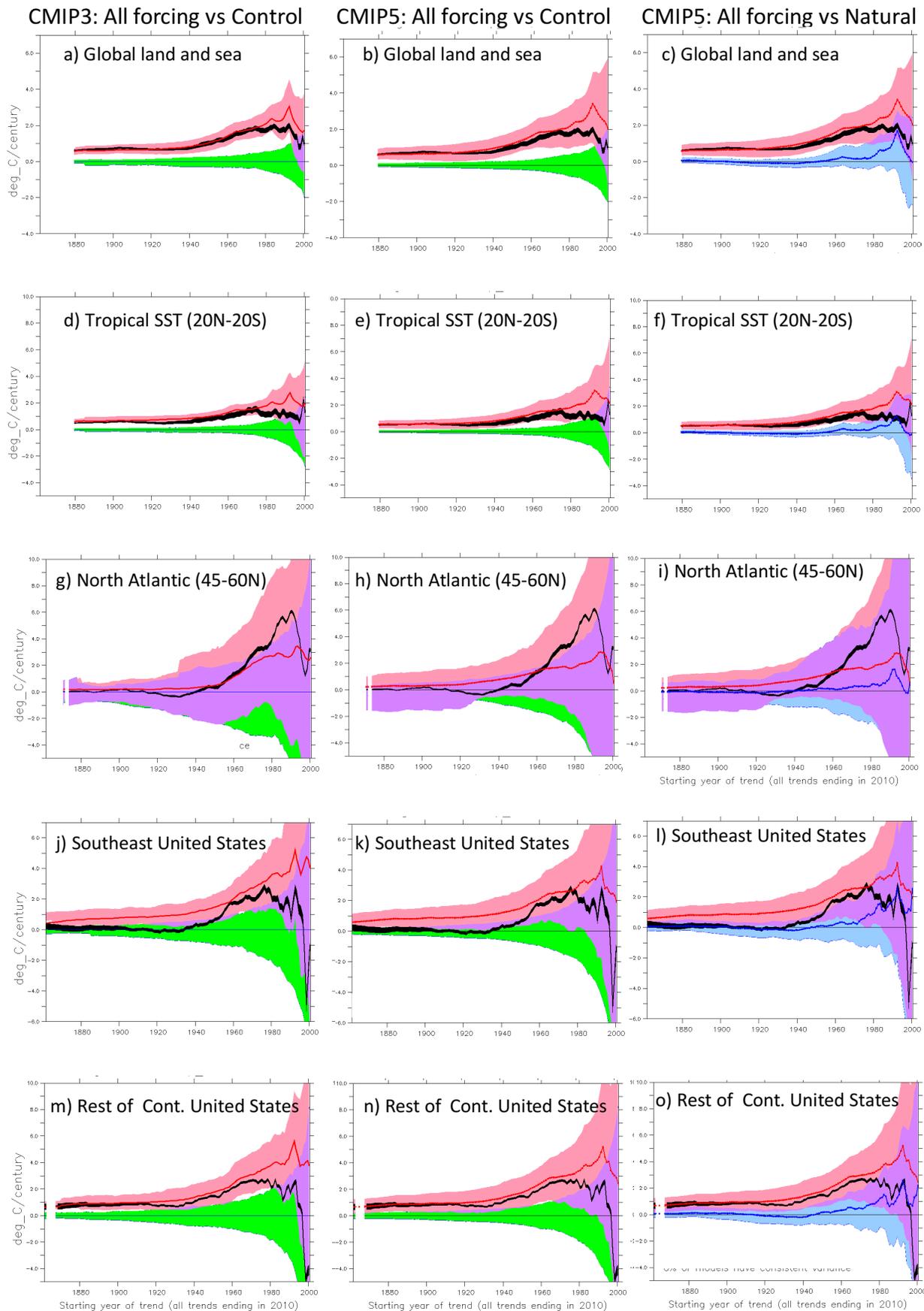


Fig. 8. As in Fig. 7, but for additional regions as labeled (see Fig. 6).

1385



1386 Fig. 9. As in Fig. 7, but for additional regions as labeled (see Fig. 6). 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. 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).

1901-2010 Surface Temperature Trends

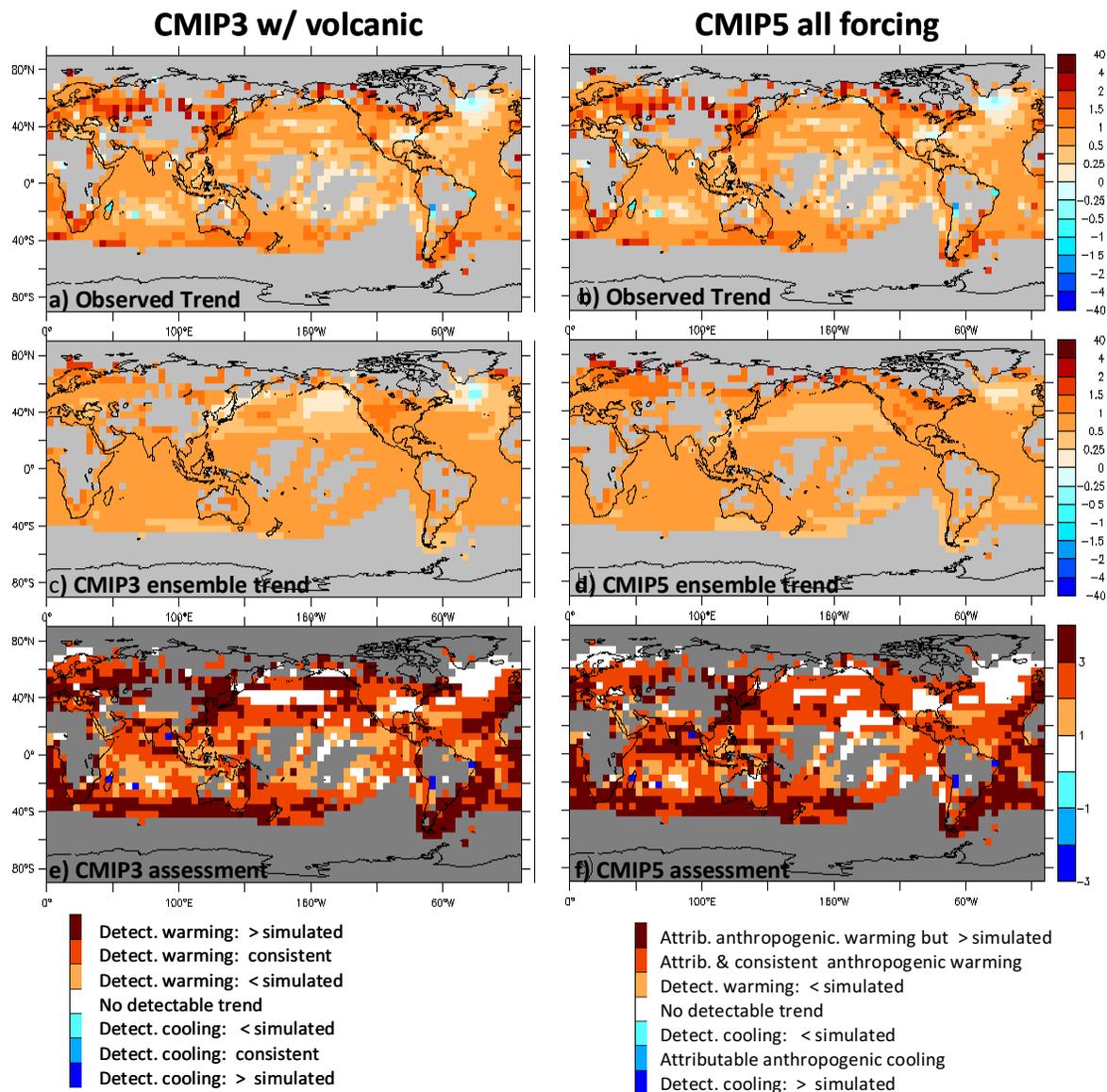


Fig. 10. Geographical distribution of surface temperature trends (1901-2010) in: (a,b) HadCRUT4 observations; (c) CMIP3 eight-model ensemble mean (All Forcing, volcanic models); (d) CMIP5 23-model ensemble mean (All Forcing, volcanic models). Unit: degrees C per 100 yr. In (e, f) the observed trend is 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 8 CMIP3 All Forcing models and their 8 control runs. Panel (f) compares observed trends with trends from 23 All-Forcing CMIP5 models and their 23 control runs and with the 7 All Forcing CMIP5 model subset and their 7 control runs.

1951-2010 Surface Temperature Trends

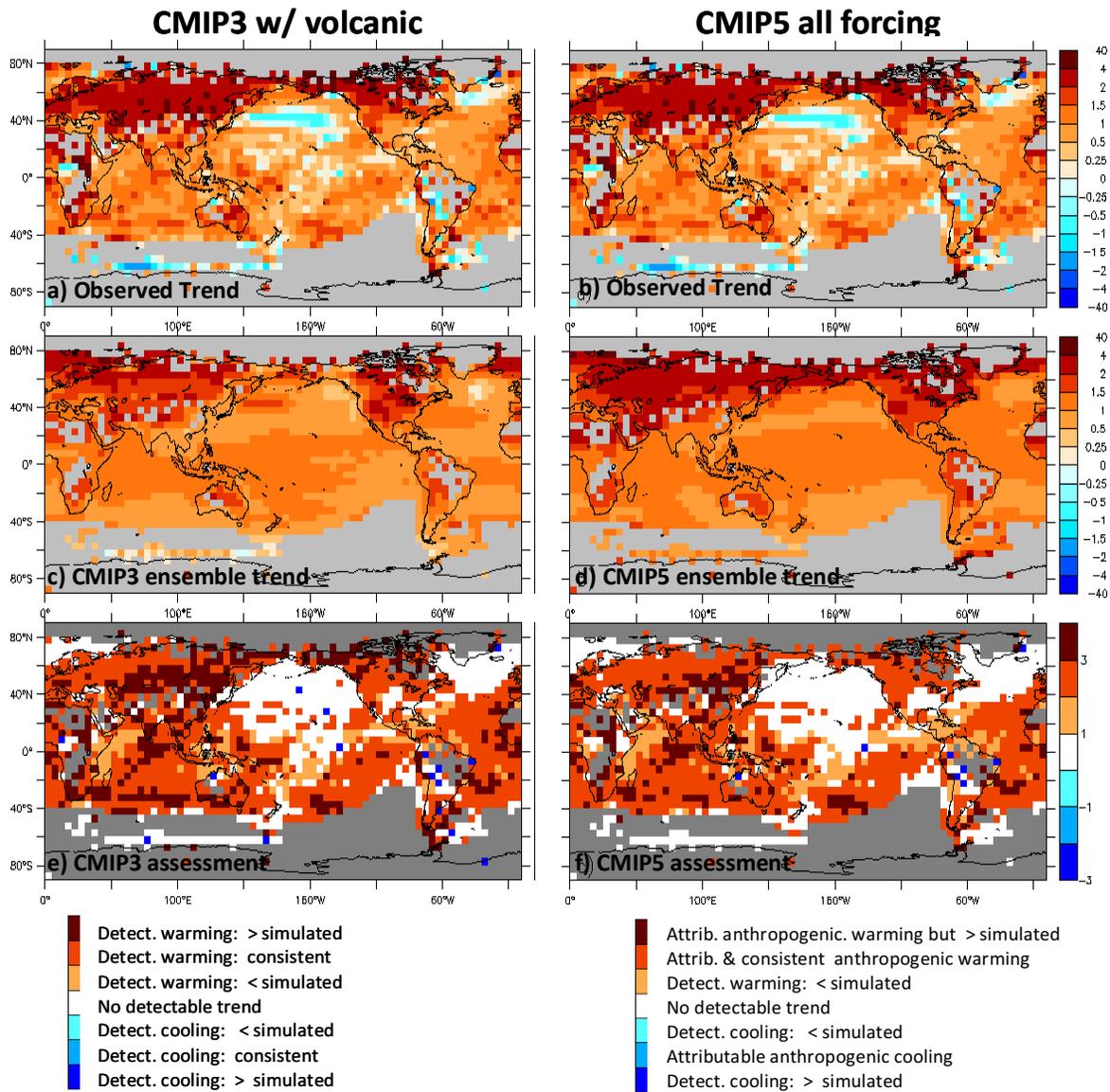


Fig. 11. Same as Fig. 10 but for trends from 1951 to 2010.

1981-2010 Surface Temperature Trends

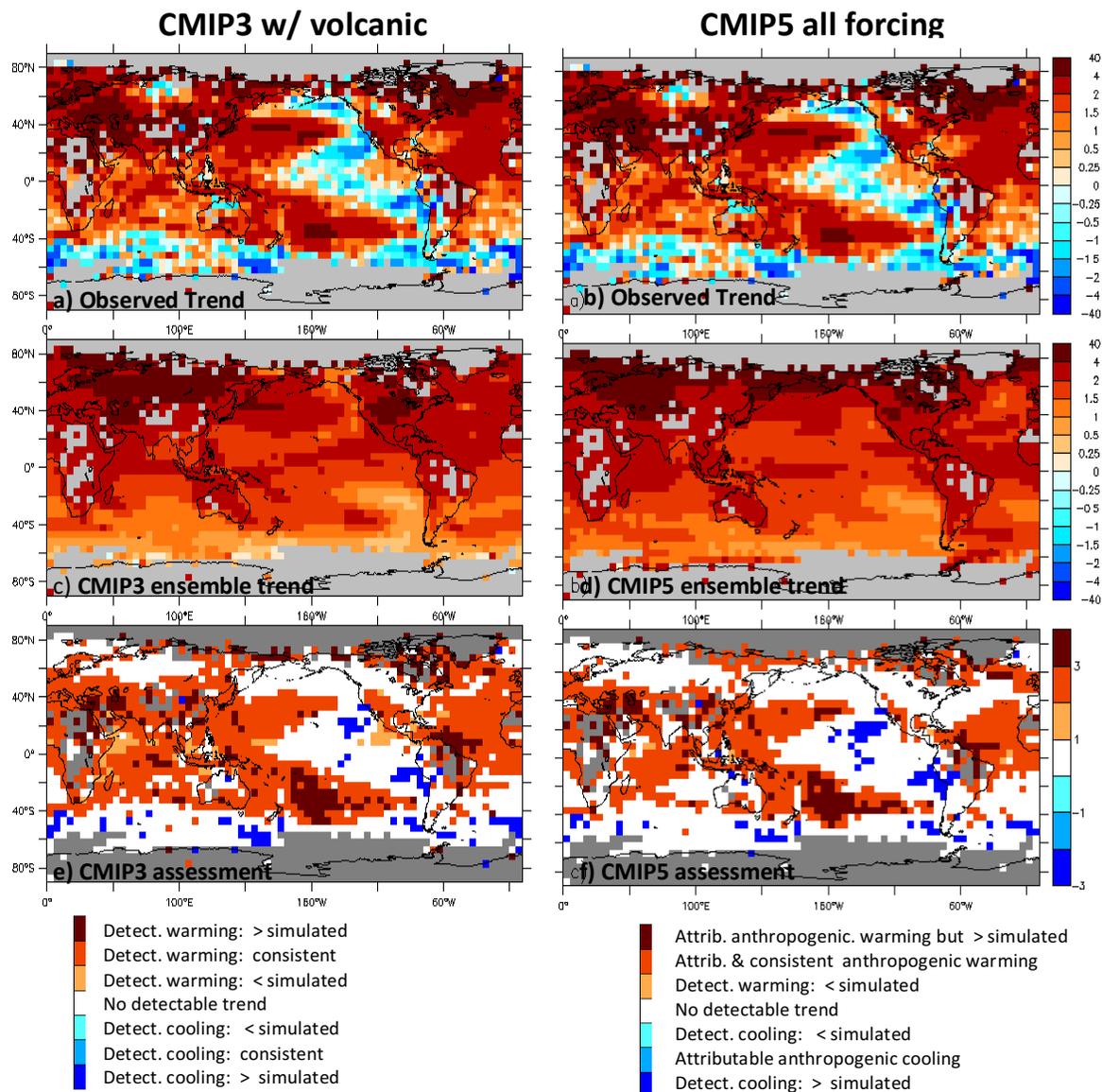
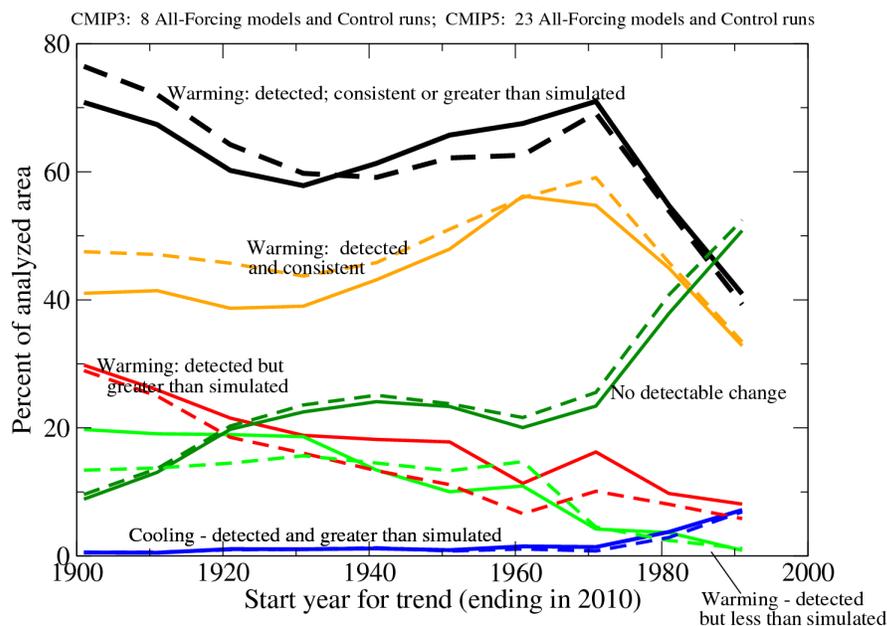


Fig. 12. Same as Fig. 10 but for trends from 1981 to 2010.

1390

a) Assessment of CMIP3 (solid) vs. CMIP5 (dashed) Multi-model Ensemble Means



b) Assessment of CMIP5 Natural vs. All-Forcing Multi-model Ensemble Means

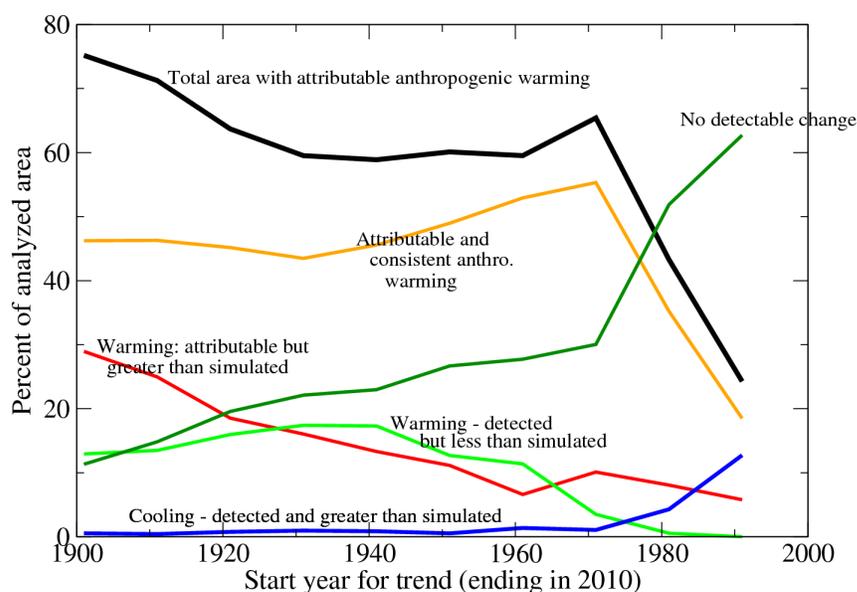


Fig. 13. Summary assessment of observed vs. model ensemble-mean trends-to-2010. The percent of global analyzed areas meeting certain criteria (see graph labels) are shown as a function of start year (all trends ending in 2010). a) Assessments of the 8 CMIP3 (solid lines) vs. the 23 CMIP5 (dashed lines) multi-model ensemble mean (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 all 23 CMIP5 models and their 23 control runs for the All-Forcing assessment and the seven-model subset of CMIP5 models (with Natural Forcing Only runs extending to 2010) and their seven control runs for the Natural-Forcing assessment. The black curves are the sum of the red and orange curves; the sum of black + light green + dark green + blue = 100%.

1391

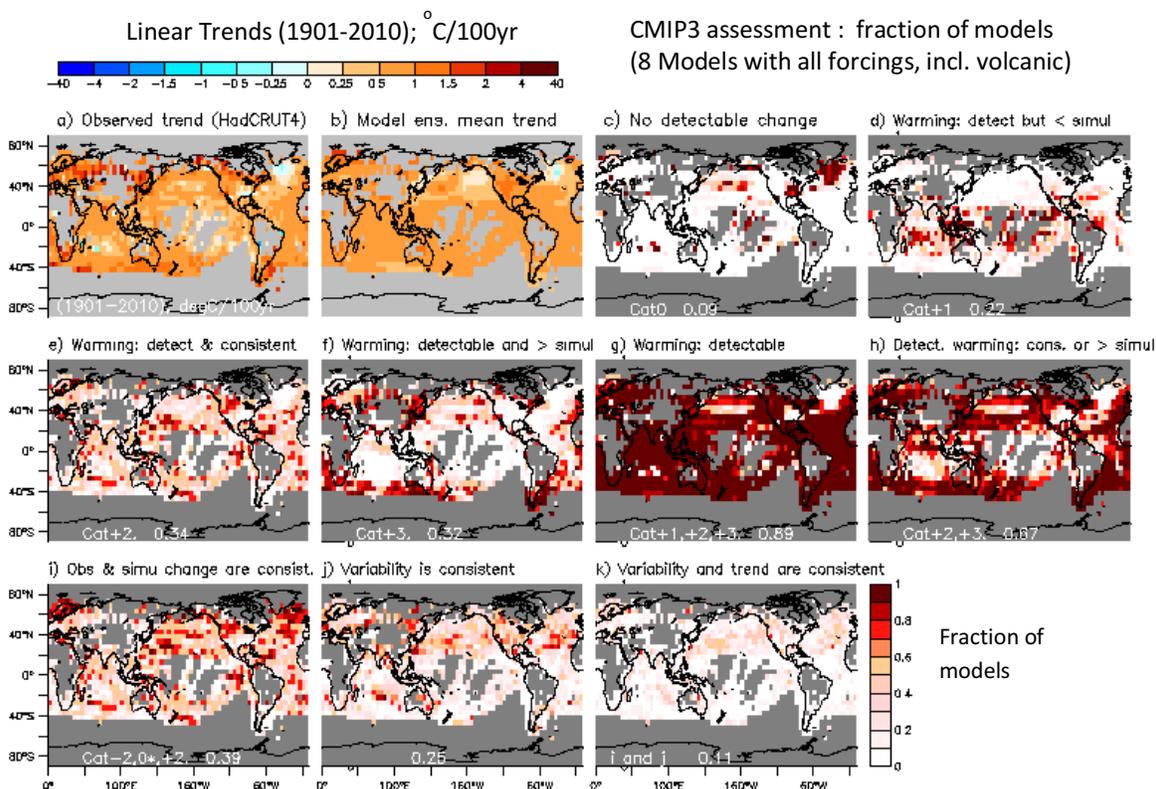


Fig. 14. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP3 multi-model (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed trend is assessed in terms of the eight individual CMIP3 models (trends and variability) in (c-k). Panels (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 greater 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 less 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.

1392

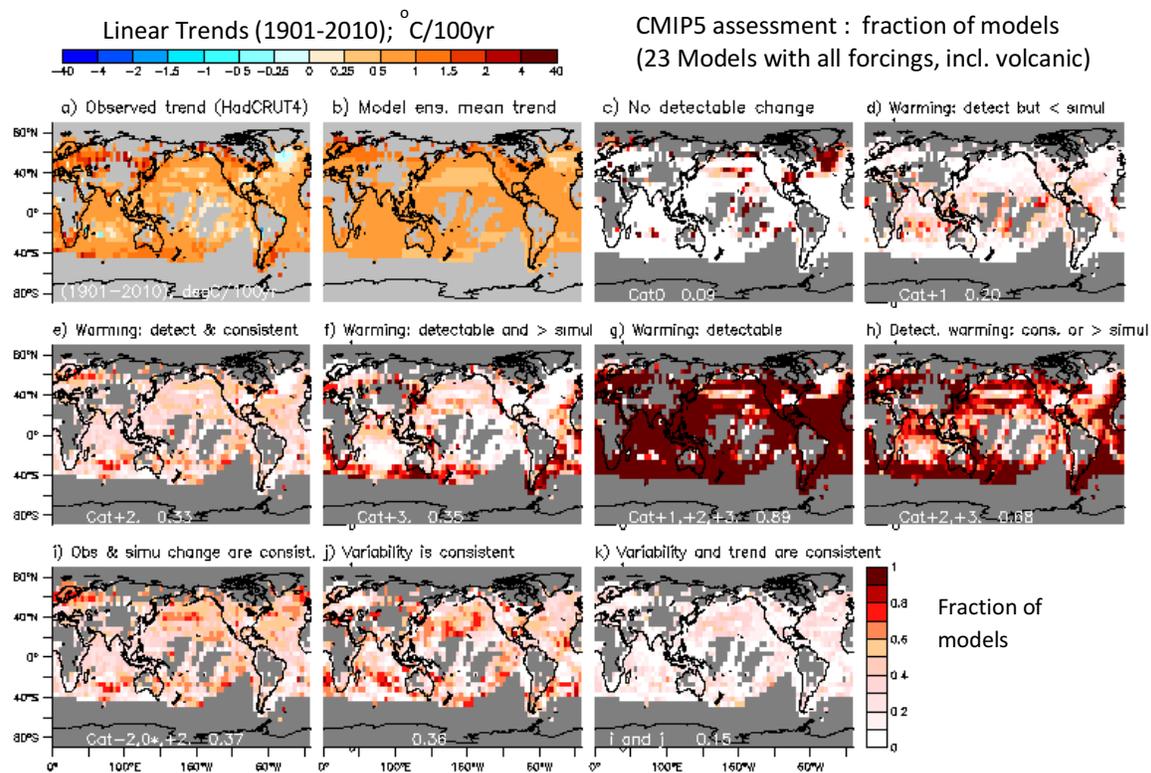


Figure 15. Same as Fig. 14, but for 23 CMIP5 models with volcanic forcing.

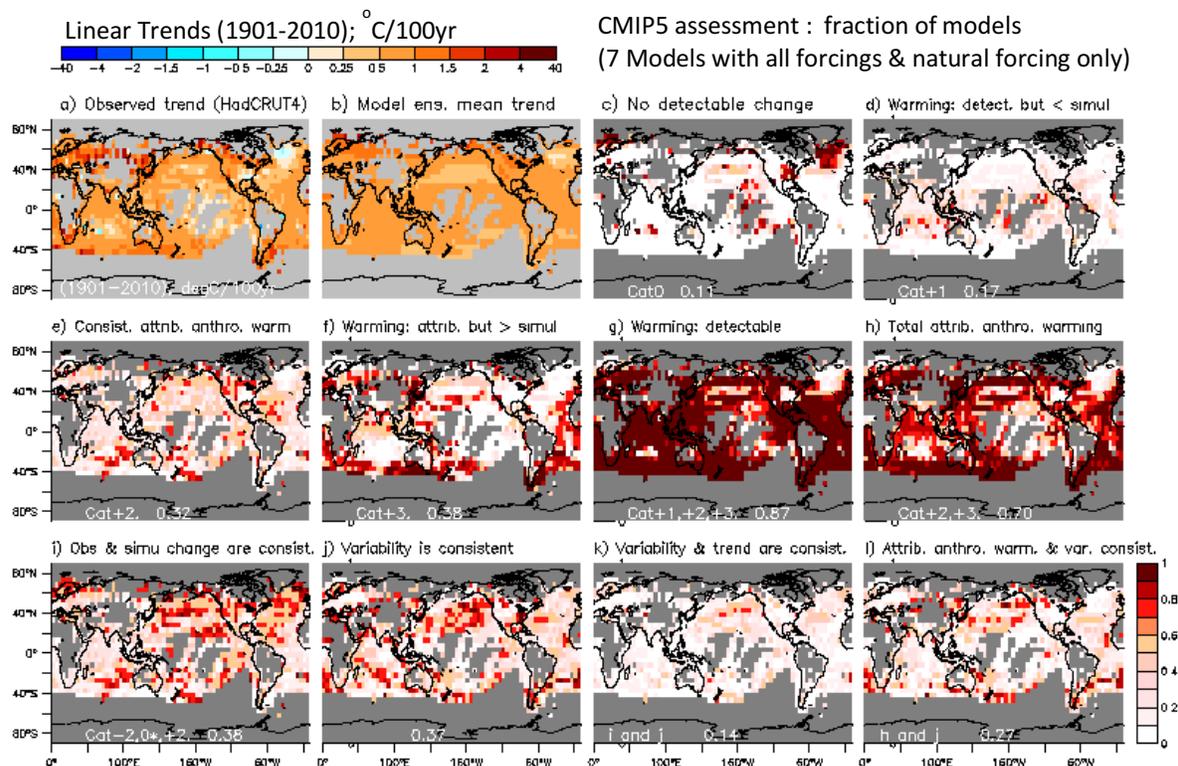


Fig. 16. Geographical distribution of: (a) HadCRUT4 observed or (b) CMIP5 multi-model (volcanic models) ensemble mean surface temperature trends (1901-2010) in degrees C per 100 yr. The observed trend is assessed in terms of trend and variability using the seven CMIP5 models that had available an All Forcing ensemble and an ensemble of 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).

1394

1395

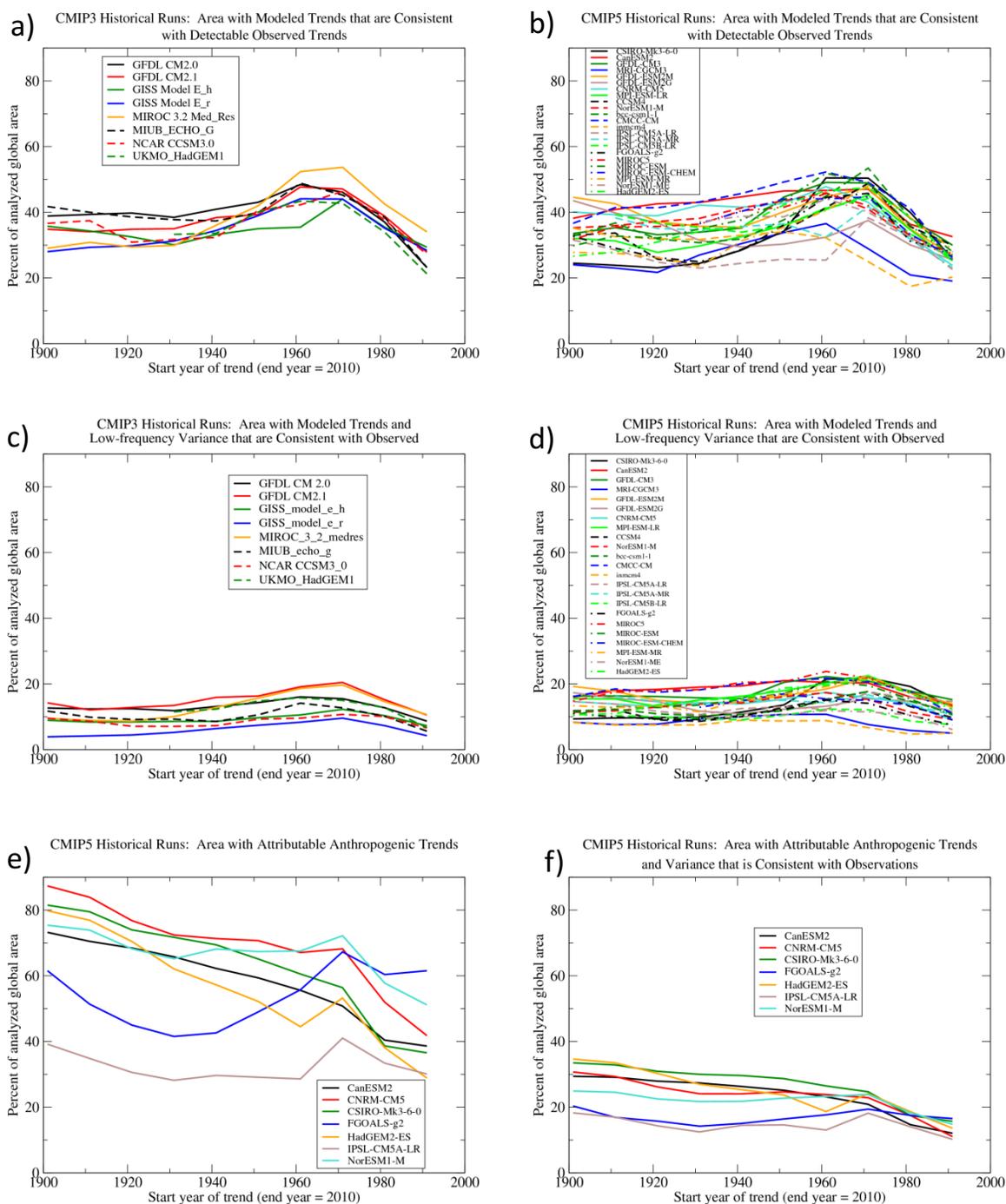


Fig. 17. Individual CMIP3 (a, c) and CMIP5 (b, d, e, f) models are assessed for consistency with detectable observed surface temperature trends-to-2010 (a-d), for attributable anthropogenic trends (e, f), and for consistency of simulated internal variability (c, d, f). Trend results are shown for start years from 1901 to 1991 (all trends ending in 2010). Plotted is the percent of analyzed global area where each individual model's (legend) 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 panel (f) includes areas where the observed trend is detectable but greater than simulated, whereas panel (d) includes only areas with trends that are detectable and consistent with simulations.