Revisiting Whether Recent Surface Temperature Trends Agree with the CMIP5 Ensemble

Revisiting Whether Recent Surface Temperature Trends Agree with the CMIP5 Ensemble

MARENA LIN AND PETER HUYBERS

Harvard University, Cambridge, Massachusetts

(Manuscript received 8 February 2016, in final form 11 May 2016)

ABSTRACT

In an earlier study, a weaker trend in global mean temperature over the past 15 years relative to the preceding decades was characterized as significantly lower than those contained within the phase 5 of the Coupled Model Intercomparison Project (CMIP5) ensemble. In this study, divergence between model simulations and observations is estimated using a fixed-intercept linear trend with a slope estimator that has one-third the noise variance compared to simple linear regression. Following the approach of the earlier study, where intermodel spread is used to assess the distribution of trends, but using the fixed-intercept trend metric demonstrates that recently observed trends in global mean temperature are consistent ($p > 0.1$) with the CMIP5 ensemble for all 15-yr intervals of observation–model divergence since 1970. Significant clustering of global trends according to modeling center indicates that the spread in CMIP5 trends is better characterized using ensemble members drawn across models as opposed to using ensemble members from a single model. Despite model–observation consistency at the global level, substantial regional discrepancies in surface temperature trends remain.

1. Introduction

Much attention has been focused on the fact that recent trends in global warming are slower than those predicted in many climate simulations. One class of explanation for this model–data disagreement is models not capturing internal variations in the surface energy balance (Kosaka and Xie 2013; England et al. 2014), possibly associated with increased deep ocean heat uptake (Meehl et al. 2011; Trenberth and Fasullo 2013). Changes to external radiative forcing specifications could also reduce model warming trends (Solomon et al. 2010; Santer et al. 2014; Schmidt et al. 2014; Huber and Knutti 2014), as would downward revision of a model's transient climate sensitivity (Otto et al. 2013). Another class of explanation involves changes to global temperature estimates. Inclusion of arctic surface temperature estimates (Cowtan and Way 2014), revision of temperature buoy offsets (Karl et al. 2015), and adjusting for air–sea temperature differences (Cowtan et al. 2015) can all incline recent warming observations nearer to the models. Other studies reconcile observed trends with either statistical properties of individual models (Thorne et al. 2015) or specific phases of modeled internal variability (Risbey et al. 2014) within phase 5 of the Coupled Model Intercomparison Project (CMIP5) ensemble (Taylor et al. 2012).

Both improved mechanistic understanding of decadal temperature variability and more accurate global temperature estimates are of obvious value. There is also utility in addressing whether differences between recent temperature trends and model projections are statistically significant. Findings of significant differences would give grounds for concluding that models are missing major components of internal variability, that data-based estimates are biased, or that other sources of uncertainties are too narrowly construed.

A variety of approaches have been employed in assessing statistical significance of recent warming trends. Rajaratnam et al. (2015) tested whether recent warming rates were slower than those between 1950 and 1997 and found no evidence for significant slowing. Brown et al. (2015) assessed recent trends in global temperature against a combination of model- and empirically derived variability, finding consistency when using CMIP5

Denotes Open Access content.

Corresponding author address: Marena Lin, Earth and Planetary Sciences, Harvard University, 20 Oxford St., Cambridge, MA 02138.
E-mail: lin8@fas.harvard.edu

DOI: 10.1175/JCLI-D-16-0123.1

regional concentration pathways (RCPs) 4.5 or 6 (Taylor et al. 2012), but they observed that decadal trends fall below the 5th percentile of distributions when using RCP8.5. Fyfe et al. (2013, hereafter Fyfe2013) also assessed observed trends relative to CMIP5 projections and found them to generally reside below the 5th percentile of simulations when using RCP4.5. A similar analysis is presented in box 9.2 of IPCC AR5 (Flato et al. 2013).

To our knowledge, Fyfe2013's analysis represents the strongest published claim for the statistical significance of the hiatus, and here we take up two major elements of that analysis in further detail. First, as Fyfe2013 document, their results are sensitive to selection of specific intervals. For example, trends in observed global temperature computed using the Hadley Centre/Climatic Research Unit version 4 (HadCRUT4) gridded compilation (Morice et al. 2012) range from 0° to 0.07°C decade^{-1} when started between 1998 and 2002, all ending in 2014. If individual intervals are then examined in isolation, statistical significance varies, with trends indicated as highly anomalous ($p < 0.05$) or consistent with CMIP5 trends ($p > 0.10$). This sensitivity to interval selection is not surprising given the shortness of the examined trends (Wunsch 1999), but it introduces an element of arbitrariness insomuch as a basis for choosing between results is lacking. Further, finding an interval falling outside of a 95% confidence interval becomes increasingly likely with the number of distinct intervals examined (e.g., Marotzke and Forster 2015).

A second issue is how the model ensemble ought to be statistically interpreted for purposes of constructing a null distribution. A truth-plus-error approach posits temperature trends as involving a deterministic component plus biases, whereas an exchangeable approach posits that actual climate and individual ensemble members share equivalent statistical properties (e.g., Annan and Hargreaves 2010). Although Rougier et al. (2011) showed that these approaches can be statistically equivalent, the implementation of the truth-plus-error approach by Fyfe2013 generally indicates differences between simulations and observations that are significant, whereas the exchangeable implementation only indicates significant differences for some of the most recent intervals considered. Determining which representation of the null is better suited to the present test would also reduce arbitrariness in the interpretation of the results.

In the following, we introduce a more stable metric of divergence in trend between observations and simulations. This metric differs from the typical least squares linear metric because it is fit with a fixed intercept, reducing the added variance from interval selection. With the exception of this modified trend metric, we replicate the hypothesis testing of Fyfe2013 and identify why the null distributions inferred from the truth-plus-error and exchangeable approaches differ in implementation. On these bases, a consistent interpretation emerges whereby no significant difference is found between observed global trends and the CMIP5 ensemble using RCP4.5.

2. Data

Observational temperature estimates are from the 5° × 5° HadCRUT4 gridded compilation of instrumental temperatures (Morice et al. 2012). Missing monthly data are infilled using the annual average if at least 10 months of observations are present in the year. Only grid boxes having at least 90% of monthly data coverage between 1950 and 2015 are included, covering 71% of the global surface area. For included grid boxes, data that are still missing are filled with the total time series average. All values are monthly anomalies with respect to 1950–2015 average seasonal cycle.

Simulations of surface temperature are from the CMIP5 historical ensemble conjoined with matching members from the RCP4.5 ensemble (Taylor et al. 2012; van Oldenborgh 2015). The ensemble comprises 22 modeling centers, 38 models, and 108 simulations (Table 1). Our ensemble differs from that of Fyfe 2013 by inclusion of the EC-EARTH and INM-CM4.0 models, addition of 21 NASA Goddard Institute of Space Studies (GISS) ensemble members, and omission of HadCM3 owing to the lack of complete RCP4.5 runs. Models have varying numbers of representative ensemble members, with, for example, NASA GISS contributing 34 ensemble members, CSIRO–Queensland Climate Change Centre of Excellence (QCCCE) contributing 10, and MRI contributing 1. With the exception of two GISS simulations, different simulations from the same model will differ at least in their initialization time within a control run for a given physics parameterization.

Analyses are performed on annual averages, where the July–June year is used in order to better contain ENSO anomalies within a given year, with the year associated with the January–June portion of the average reported. Monthly averages are weighted according to the number of days in a month, which differs across models. Models variously employ the standard Gregorian calendar with leap years, a fixed 365-day no-leap-year calendar, and a fixed 360-day calendar. For all model–data comparisons, simulations were regridded to the observational grid by taking the area-weighted average of simulation grid boxes contained within each uncensored observational grid box. Global mean temperatures are determined as the area-weighted average across all uncensored grid boxes on the native

TABLE 1. List of CMIP5 ensemble members assessed, sorted by center, model, and ensemble number. Columns are given in descending order of hierarchical model lineage: the center number and name, the model number and name, and ensemble members comprising each model. Note that numbering is internal to this study. Expansions of institutions and model names are available online at http://www.ametsoc.org/PubsAcronymList.

Center No.	Center name	Model No.	Model name	Ensemble No.
1	National Institute of Meteorological Research/Korea Meteorological Administration (NIMR/KMA)	1	HadGEM2-AO	1
2	BCC	2	BCC_CSM1.1	2
		3	BCC_CSM1.1(m)	3
3	College of Global Change and Earth System Science (GCESS)	4	BNU-ESM	4
4	CCCma	5	CanESM2	5–9
5	NCAR	6	CCSM4	10–15
6	NSF–DOE–NCAR	7	CESM1(BGC)	16
		8	CESM1(CAM5)	17–19
7	CMCC	9	CMCC-CM	20
		10	CMCC-CMS	21
8	CNRM–CERFACS	11	CNRM-CM5	22
9	CSIRO–QCCCE	12	CSIRO Mk3.6.0	23–32
10	EC-EARTH	13	EC-EARTH	33–39
11	LASG–IAP	14	FGOALS-g2	40
12	FIO	15	FIO-ESM	41–43
13	NOAA/GFDL	16	GFDL CM3	44
		17	GFDL-ESM2G	45
		18	GFDL-ESM2M	46
14	NASA GISS	19	GISS-E2-H	47–61
		20	GISS-E2-H-CC	62
		21	GISS-E2-R	63–79
		22	GISS-E2-R-CC	80
15	MOHC	23	HadGEM2-CC	81
		24	HadGEM2-ES	82–85
16	INM	25	INM-CM4.0	86
17	IPSL	26	IPSL-CM5A-LR	87–90
		27	IPSL-CM5A-MR	91
		28	IPSL-CM5B-LR	92
18	MIROC	29	MIROC5	93–95
		30	MIROC-ESM	96
		31	MIROC-ESM-CHEM	97
19	MPI-M	32	MPI-ESM-LR	98–100
		33	MPI-ESM-MR	101–103
20	MRI	34	MRI-CGCM3	104
21	Norwegian Climate Centre (NCC)	35	NorESM1-M	105
		36	NorESM1-ME	106
22	CSIRO–BoM	37	ACCESS1.0	107
		38	ACCESS1.3	108

grid of the simulation. Results are unchanged to two significant figures if simulations are instead regridded using linear interpolation.

3. Methods

a. Measuring the divergence

We test the null hypothesis H_0 that recent global temperature trends are consistent with the CMIP5 multimodel ensemble. It is useful to focus on the trend metric used in evaluating H_0 because of its implications for the stability of the test. Global temperature trends are often quantified using an estimate of slope s in the following simple linear regression equation:

$$T_k = s(t_k - t_0) + b + \varepsilon_k, \quad t \geq t_0, \tag{1}$$

in which both s and b are estimated in the least squares sense, by minimizing $\sum_{k=0}^{N} \varepsilon_k^2$. If ε_k is assumed uncorrelated and normally distributed with standard deviation σ, the expected variance of the slope estimator is $\mathrm{VAR}(s) = 12\sigma^2/(L^3 - L)$, where L is the number of data points that comprise the trend. This formulation is common (e.g., Thompson et al. 2015) and derived in appendix B.

Many statistical models can be fit to quantify trends (e.g., Visser et al. 2015), and we consider further formulations according to several characteristics: similarity to foregoing approaches [i.e., Eq. (1)], suitability for describing previous agreement but recent divergence between models and data, and low sensitivity to choice of interval. Although not considered by Visser et al. (2015), a piecewise fit to a time series of the difference between two temperature estimates appears apt under these criteria. Specifically, the difference between two time series T' is fit using a constant offset c followed by a linear trend δ that is piecewise continuous:

$$T'_k = \begin{cases} c + \eta_k, & t_k \le t_0, \\ \delta(t_k - t_0) + c + \eta_k, & t_k > t_0. \end{cases} \quad (2)$$

The constant c is estimated as the average of T' for $t \le t_0$, and δ is estimated in a least squares sense, as for s. Estimates of δ differ from s in having an intercept fixed at (t_0, c), which acts as a hinge point preceded by a constant and followed by a trend diverging from that constant. Different sets of time series used to calculate T' are defined in the context of various tests that follow.

The expected variance of δ is smaller than that of s when each is fit to the same T' time series. Assuming that η_k and ε_k have equivalent distributions and that the variance of c is small on the basis of being constrained by a relatively long sequence of T' permits for writing $\mathrm{VAR}(\delta) = 6\sigma^2/[L(L-1)(2L-1)]$. The ratio of variances between δ and s is then $(L+1)/(4L-2)$. Appendixes A and B give derivations of these variances and further calculations involving variance contributions from c. We find that δ has 0.35 times the variance of s when applied to global temperature trends over 15-yr intervals.

Importantly, δ is also more stable than s across application to different intervals. The variance of the difference in trends fit to L-length intervals with consecutive start years using δ, relative to that for s is given by $\mathrm{VAR}(\delta_2 - \delta_1)/\mathrm{VAR}(s_2 - s_1) = L(L+1)/[2(2L-1)^2]$. This ratio is 0.143 for an interval length of $L = 15$. Contributions from variance in c are neglected in the foregoing expression but are minor as long as the interval over which c is calculated ($\le t_0$) exceeds that over which δ is defined ($> t_0$; see appendix C). The lower variance of δ associated with interval selection reduces the potential for false positives that otherwise occur when conducting multiple tests for trend significance over various intervals.

Stability of the δ estimator is also empirically indicated by its more smoothly varying as a function of start year (Fig. 1). When T' is defined as the difference between HadCRUT4 global average temperature and the CMIP5 ensemble average, values of s equal $-0.172°$, $-0.128°$, and $-0.189°\mathrm{C\,decade^{-1}}$ for 1998–2014, 2000–14, and 2002–14 (Fig. 1b). In contrast, estimates of δ fit to the same time series change monotonically when computed over the same intervals, having values of δ equal to $-0.140°$, $-0.168°$, and $-0.205°\mathrm{C\,decade^{-1}}$ (Fig. 1c). Unless explicitly indicated otherwise, regional and global estimates of CMIP5 ensemble average temperature are always computed as the average across equally weighted modeling centers and include only grid boxes corresponding to observations.

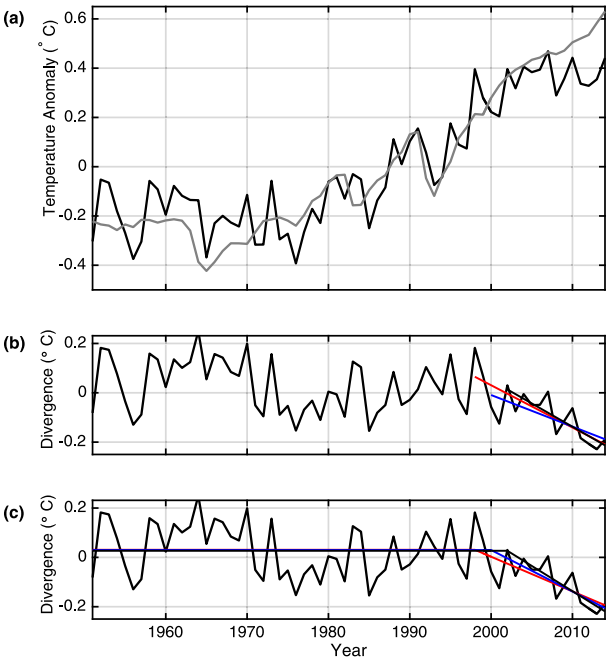

FIG. 1. Global mean temperatures from observations and the CMIP5 ensemble average, and their difference. (a) HadCRUT4 global mean temperature time series (black) and CMIP5 ensemble average time series (gray). (b) Difference between observations and the CMIP5 ensemble average (black) with trends estimated over intervals of 1998–2014, 2000–14, and 2002–14 having values of $s = -0.172°$, $-0.128°$, and $-0.189°\mathrm{C\,decade^{-1}}$, respectively. (c) Similar to (b), but using the δ trend estimator and obtaining values of $-0.140°$, $-0.168°$, and $-0.20°\mathrm{C\,decade^{-1}}$ for the same intervals.

Computing δ on a gridbox basis—again using CRU observations relative to the CMIP5 ensemble average—shows a coherent pattern of cooling in the eastern Pacific consistent with the negative phase of the Pacific decadal oscillation (PDO; Zhang et al. 1997; Fig. 2). Further, a cooling trend that is prominent in the midlatitudes of Eurasia when using s is suppressed when using δ, consistent with findings that these trends result from short-term internal variability (Li et al. 2015; Cohen et al. 2012) and that δ is less volatile. That δ yields more physically interpretable patterns, to which we return later, also supports its being a more suitable metric for

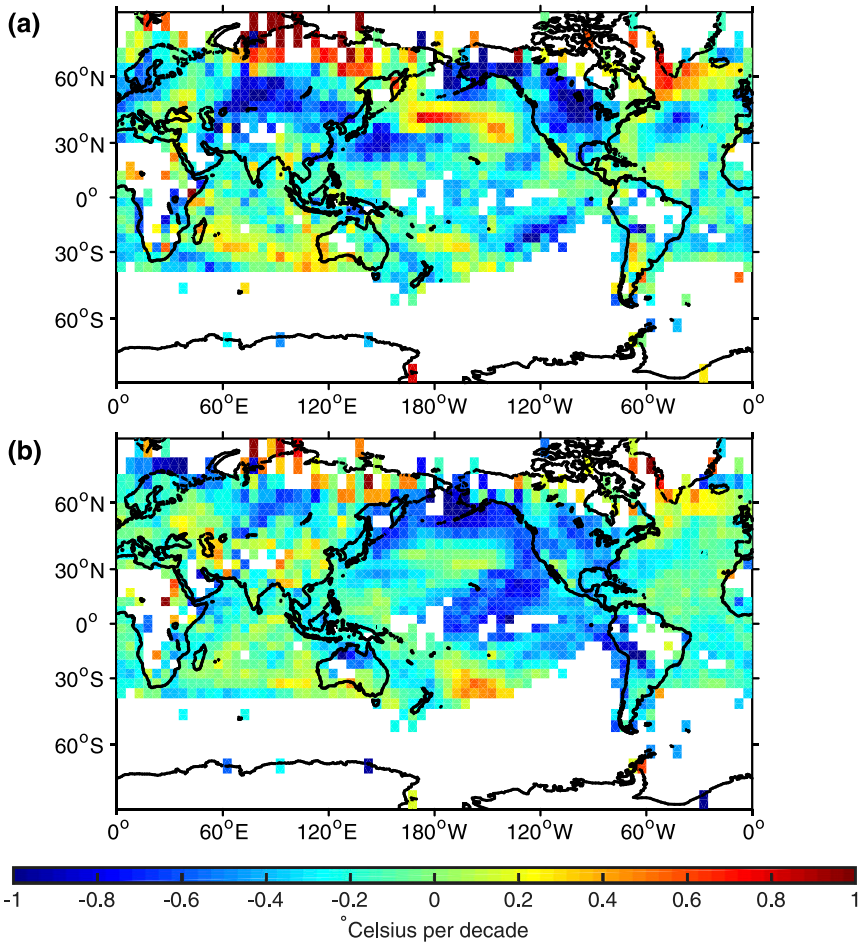

FIG. 2. Maps of trend divergence between HadCRUT4 observations and the CMIP5 ensemble average using (a) the standard trend estimator s and (b) fixed-intercept trend estimator δ for 2000–14. Trends are computed on a gridbox basis, with model results first being regridded to the 5° × 5° grid of the HadCRUT4 observations.

interpreting divergence in recent trends. From both theoretical variance properties and empirical application, we find δ to be less volatile than s and, therefore, expect it to yield more consistent results when applied in testing for model–observational discrepancy.

b. *Formulating a null hypothesis*

We evaluate a null hypothesis H_0 that recent global temperature trends are consistent with the CMIP5 multimodel ensemble using the δ metric. Contributions to uncertainty in comparing observations and models are captured through combining different realizations of three time series: A, a version of global mean temperature anomalies from observations; B, a version of global mean temperature anomalies averaged across the CMIP5 ensemble; and C, a time series representing the variability associated with an individual CMIP5 simulation. The null hypothesis H_0 is assessed by calculating the degree to which the distribution of trends derived from realizations of $A - B + C$ contain zero and, therefore, reflect consistency between model and observations. This hypothesis test is modeled exactly upon that of Fyfe2013 in order to allow for direct comparison of results. We evaluate 15-yr trends, as this was the chosen length of Fyfe2013 and Marotzke and Forster (2015), and Fyfe2013 rejected H_0 using s for the 15-yr interval 1998–2012.

Realizations of global temperature from observations A have uncertainties that include observational noise, issues associated with computing global averages from a limited network, and systematic errors from switching between observing methods. These uncertainties are expressed for the HadCRUT4 observations through an ensemble with 100 members, each perturbed with noise realizations (Morice et al. 2012). The mean and standard deviation of s computed between 2000 and 2014 from

the HadCRUT4 observational ensemble is 0.076° ± 0.006°C decade^{-1}. Fyfe2013 compute realizations of observational trends by averaging over 100 draws of the HadCRUT4 ensemble taken with replacement. This approach suppresses the standard deviation of s by a factor of 10 and, in our view, seems unwarranted since each ensemble member is meant to indicate a plausible realization. Furthermore, note that the bias correction for ocean buoy data suggested by Karl et al. (2015) results in a trend estimate of 0.116°C decade^{-1} between 2000 and 2014 that exceeds all HadCRUT4 ensemble trends, suggesting that the uncertainty estimates in the ensemble are too narrow or that the correction of Karl et al. (2015) is too large. A question noted, albeit not otherwise addressed here, is whether the observational record of surface temperature is known with sufficient accuracy to provide a stringent test of the climate models. We proceed using Fyfe2013's estimate of temperature trends and their spread from the uncorrected HadCRUT4 ensemble in order to illustrate that it is difficult to reject H_0 even under Fyfe2013's representation of low uncertainty.

Realizations of global temperature across the CMIP5 ensemble B are obtained by sampling the 38 CMIP5 models with replacement, randomly selecting an ensemble member for each, and taking the average across samples. This approach gives equal weight to each model, as opposed to weighting according to the number of submitted ensemble members. Averaging across models in order to obtain a more stable estimate is more defensible than averaging across HadCRUT4 temperature realizations because models are developed semi-independently from one another and contain independent realizations of internal variability, whereas there is only a single set of temperature observations. There is evidence, however, that the spread across various model ensembles is suppressed (Huybers 2010; Masson and Knutti 2011), possibly because of anchoring effects or suppression of outliers (Cess et al. 1996).

Finally, a realization of the variability associated with a simulation C is obtained by sampling ensemble members in a similar manner as for B and then computing the difference between a single one of the sampled ensemble members and the average across the sample. In accord with the statistical approach that models and observations are exchangeable realizations of climate (Annan and Hargreaves 2010), the observations, represented by a single sample from the HadCRUT4 ensemble, are afforded the weight of one model and pooled with the 38 CMIP5 models for a total of 39 sampled units. In practice, however, we find that the inclusion of observations in this sampling process does not affect estimates of statistical significance.

Each realization of time series $A - B + C$ then describes a departure of observed temperatures A from a CMIP5 average B with variability associated with a single ensemble member C. The distribution of H_0 is estimated from trends fit to 10^5 realizations of $A - B + C$, and H_0 is rejected if fewer than 5% of realizations are greater than or less than zero; that is, tests are performed as two sided at the $p = 0.1$ level.

c. Clustering and rejection of an alternate test approach

Fyfe2013 also employ a methodology loosely motivated by the truth-plus-error framework where B is meant to represent true climate and a quantity C' represents internal variability. Following Fyfe2013, we realize C' by selecting one of the 13 CMIP5 models associated with more than one ensemble member and computing the difference between one of these ensemble members and the average time series of that model. This approach yields a more narrow distribution of trends and more frequent rejection of the null hypothesis than that based on the exchangeable hypothesis. The origin of this discrepancy becomes evident when assessing whether simulations cluster according to model.

Clustering between members of a sample tends to narrow the inferred distribution of trends, if not otherwise accounted for (Pennell and Reichler 2011). Several studies document correlations of temperature or precipitation patterns according to model (Masson and Knutti 2011; Knutti and Sedlácek 2013; Sanderson et al. 2015). We use an Ansari–Bradley rank-sum approach (Ansari and Bradley 1960) to specifically test whether the values of δ cluster according to model. Values of δ are assessed for each ensemble member relative to the CMIP5 ensemble average temperatures for 2000–14. The test is one sided and performed at the $p = 0.05$ level.

Of the 16 modeling centers that contribute multiple ensembles, 7 centers comprising a total of 58 ensemble members each have significantly smaller dispersion than the remainder of the ensemble at the $p = 0.05$ level (Table 2). The 34 ensemble members contributed by NASA GISS show particularly significant clustering at $p < 0.01$. A bootstrapped version of the Ansari–Bradley rank-sum test that accounts for the correlation between sample medians gives equivalent results. Figure 3 also illustrates this clumping of trends according to modeling center and that the empirical histogram of trends is broader when each modeling center is equally weighted.

Smaller dispersion of members from a single model relative to those from across models is not surprising considering that different models may include different physics (e.g., Watanabe et al. 2012), entail different

TABLE 2. Test results for whether ensemble members from a single modeling center significantly cluster. Specifically, a null hypothesis that the dispersion of δ slopes associated with a modeling center are equal to or greater than that across all other modeling centers is tested using an Ansari–Bradley rank-sum approach. The number of ensemble members associated with each modeling center is given in the last column.

Modeling center	P value	No. of ensemble members
BCC	0.02	2
CCCma	0.01	5
NCAR	0.07	6
NSF–DOE–NCAR	0.17	4
CMCC	<0.01	2
CSIRO–QCCCE	0.18	10
EC-EARTH	0.01	7
FIO	<0.05	3
NOAA/GFDL	0.32	3
NASA GISS	<0.01	34
MOHC (additional realizations by INPE)	0.55	5
IPSL	0.17	6
MIROC	<0.01	5
MPI-M	0.09	6
NCC	0.21	2
CSIRO–BoM	0.51	2
No. of ensemble members with $p < 0.05$		58

parameterizations (e.g., Collins et al. 2011), and contain different forcing (e.g., Tebaldi and Knutti 2007). Smaller dispersion explains why the null distribution estimated using C', which depends on intramodel differences, is narrower than C, which depends on intermodel differences. Simulations may also cluster along axes not entirely described according to model center—inclusive of aspects of physics, parameterization, and forcing—though such dependencies are not directly relevant to the distinction between the C and C' null models considered here. Our view is that evaluation of whether observations are consistent with simulations should include all relevant sources of uncertainty in model simulations and that the intermodel comparisons associated with C are more representative of this uncertainty. In the following sections, we thus rely exclusively on the exchangeable approach in order to gauge consistency between CMIP5 simulations and observations.

4. Results

We are systematically unable to reject H_0 for all 15-yr intervals that we examine, with start years ranging between 1970 and 2000 ($p > 0.1$; Fig. 4). Note that

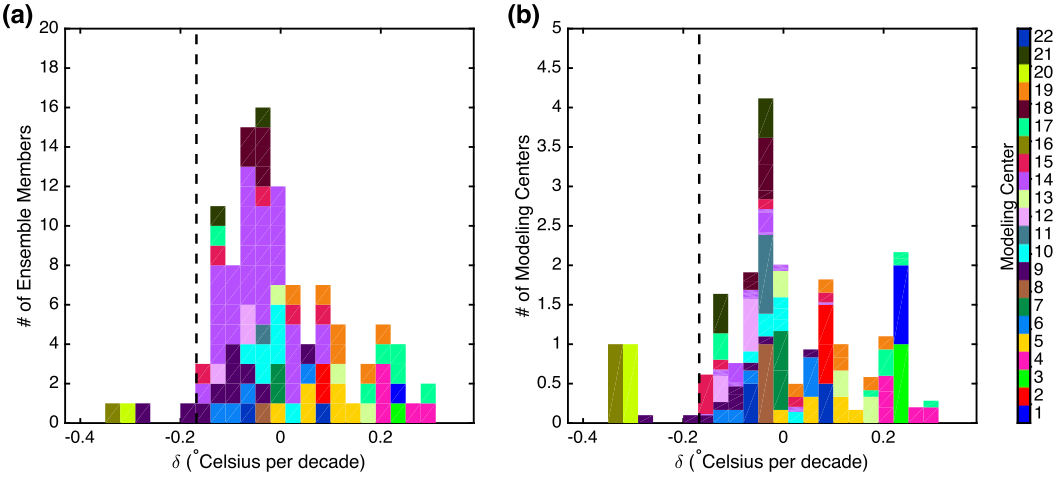

FIG. 3. Histograms of δ illustrating different model weights. (a) Equal weighting of ensemble members gives a more narrow distribution than when (b) each modeling center is equally weighted. All values of δ are calculated for 2000–14. Color bar indicates modeling centers according to the numbering in Table 1.

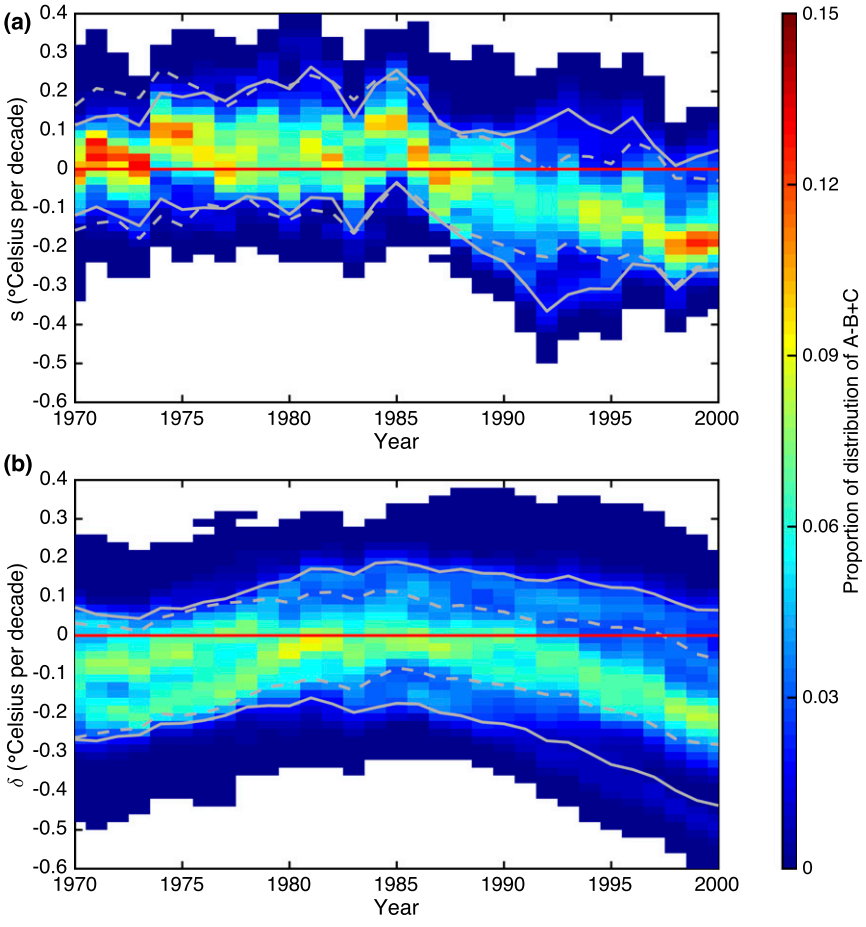

FIG. 4. Significance of observed slopes judged relative to null distributions. Slopes are measured using either (a) s or (b) δ, begin at the values labeled on the abscissa, and span 15 years. Distributions are estimated from 10^5 realizations of $A - B + C$, where A is from observations and B and C are from CMIP5. The 90% confidence intervals are shown for $A - B + C$ (solid gray line) as well as the truth-plus-error approach of $A - B + C'$ (dashed). Trend distributions with fewer than 5% of realizations greater or less than zero would indicate that observations are inconsistent with the CMIP5 ensemble. Using δ to assess the trend of $A - B + C$, H_0 cannot be rejected for any 15-yr interval between 1970 and 2014. Note that Marotzke and Forster (2015) show a similar figure but use different metrics and null distributions.

this failure to reject is obtained using a p value of 0.1, as opposed to 0.05, and using the average across HadCRUT4 ensemble members in estimating A, both of which make it easier to reject H_0. The differences between our results and Fyfe2013 relates to our use of δ instead of the more volatile metric s and our employing only the exchangeable approach C as opposed to the more narrowly distributed C' obtained from a truth-plus-error approach.

Although we do not favor other approaches, for completeness we note that if instead a combination of C and s were employed, results would be more variable and rejection of H_0 becomes almost, but still not quite, possible for the 1998–2012 interval. Using C' and δ would lead to rejections for 15-yr intervals starting between 1998 and 2000. Finally, using C' and s would lead to intermittent rejection of four different 15-yr intervals, starting at 1992, 1998, 1999, and 2000 (Fig. 4a), consistent with the combined effects of a null that is more narrowly distributed and a metric that is more variable between different fitted intervals. We note that the variability of p values across different 15-yr intervals is smooth in Fig. 2d of Fyfe2013, which we have not been able to reproduce.

Our basic result, that H_0 cannot be rejected using C and δ, is insensitive to three other relevant variants. First, Fyfe2013 do not include INM-CM4.0 in their ensemble,

and the single ensemble member associated with this model shows one of the most negative δ trends. Whereas repeating our test excluding this model lowers estimated p values, they nowhere become lower than 0.1.

Second, if C is interpreted as uncertainty that is equally applicable to the observations, represented by A, and to the CMIP5 ensemble average, represented by B, it appears arbitrary whether it is added to or subtracted from the quantity $A - B$. However, C is asymmetric with a positive skew toward larger values (Fig. 4) such that p values are smaller when C is added, though again never lower than 0.1. It is unclear whether this asymmetry is indicative of physical processes that make positive anomalies from the mean more likely than negative ones, as implied by the asymmetric nature of feedback sensitivity (Roe and Baker 2007), or merely results from the small sample population.

Finally, when using the spatially interpolated HadCRUT4 produced by Cowtan and Way (2014) and the spatially complete estimates of global mean temperatures for the CMIP5 ensemble, it becomes even more difficult to reject H_0 using s and δ. These latter results are expected given the rapid warming in polar regions (Cowtan and Way 2014).

5. Discussion and conclusions

Meehl et al. (2013) demonstrated that intervals of slow temperature rise in CCSM4 RCP4.5 projections also generally feature negative PDO patterns along with anomalously positive rates of ocean heat uptake. In our results, regional δ between CRU observations and the CMIP5 ensemble average show a clear negative phase of the PDO (Fig. 2b). Furthermore, the simulation having the second-closest global δ to the observations (ensemble member 24; see Table 1) shows a regional δ pattern resembling the negative phase of the PDO (Fig. 5a), suggesting that at least some ensemble members produce cooling for physically similar reasons. A systematic exploration of the manifestation of the PDO in each ensemble member, however, reveals no clear relationship between global values of δ and the pattern or phase of the PDO.

Rather than providing an explanation in terms of the PDO, ensemble members with global δ values similar to the observations generally have anomalously low temperature trends at high northern latitudes over continents (e.g., CSIRO–QCCCE model numbers 24 and 32; Figs. 5a,b). This congruence between regional and global δ is consistent with findings that cooling trends across northern regions are tied to the slowdown in global temperature trends (Cohen et al. 2012). The presence of such negative regional δ values within the ensemble follows from northern continental regions having high variance in δ across ensemble members (Fig. 5c).

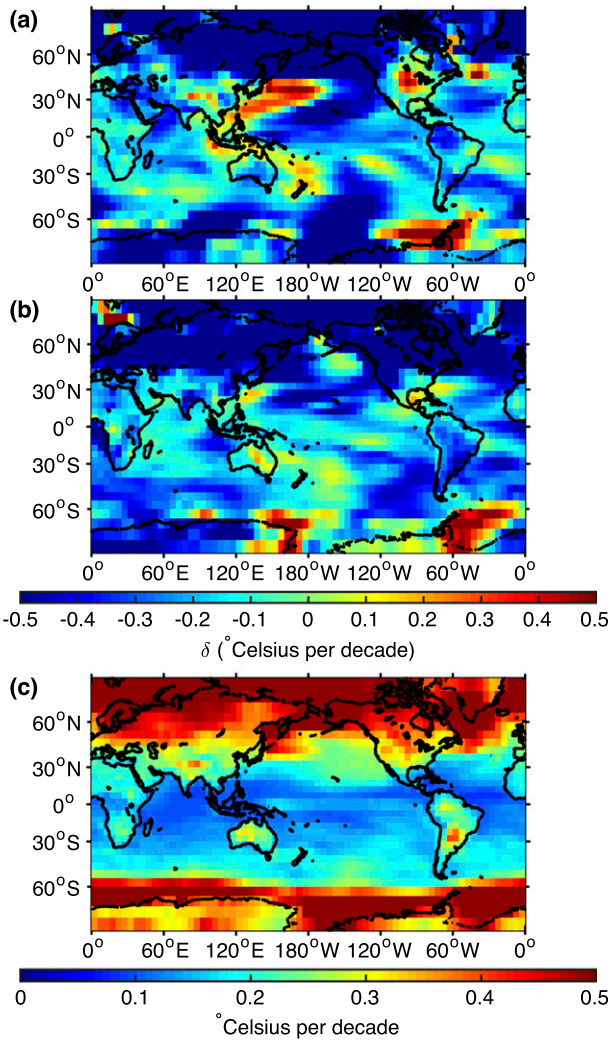

FIG. 5. Maps of δ for two CMIP5 ensemble members and of the standard deviation of δ across the CMIP5 ensemble. The two ensemble members are selected for having a global δ at least as negative as that observed: (a) CSIRO–QCCCE (model 24 in Table 1) with a global average δ of 0.219°C decade^{-1} and (b) CSIRO–QCCCE (model 32) with a global average δ of 0.278°C decade^{-1}. Similar to Fig. 2, δ is computed on a gridbox basis, with differences taken between the CMIP5 ensemble average and individual ensemble members. (c) Standard deviation of δ across 108 ensemble members at each grid box. All δ values are computed over the interval 2000–14.

High variance in northern continental regions has been found in other simulations (e.g., Deser et al. 2012) and is presumably associated with positive feedbacks that amplify arctic warming (e.g., Feldl and Roe 2013) and low thermal buffering at high northern regions (e.g., Kim and North 1991). Negative trends in northern regions are, however, inconsistent with the observed neutral or positive δ values found at most high latitudes in observations (Fig. 2).

Further discrepancies exist in the eastern equatorial Pacific, where CMIP5 trends are uniformly higher than

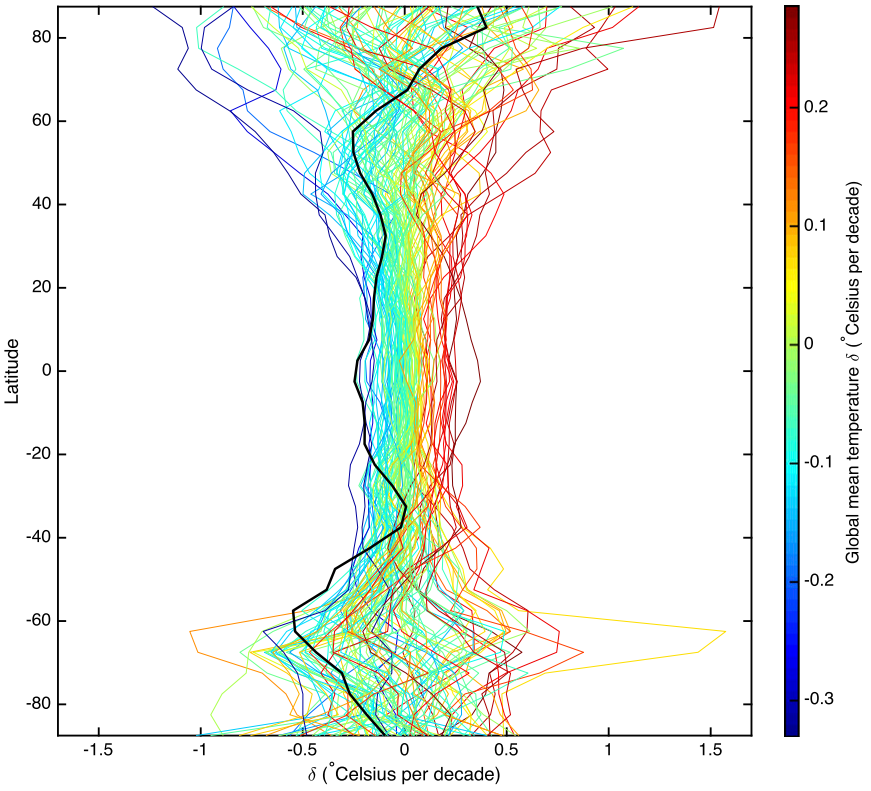

FIG. 6. Zonal average δ for the 108 CMIP5 ensemble members (color lines) and observations (black line). The δ values are calculated between each ensemble member and the CMIP5 ensemble average on a zonal basis over the 2000–14 interval. Results from each ensemble member are also colored according to their global δ that are calculated using global averages. Also shown in black are zonal average δ values between the HadCRUT4 gridded observations and the CMIP5 ensemble average. In this case, only a spatially complete observational estimate from an interpolation of the HadCRUT4 gridded product (Cowtan and Way 2014) is used to allow for comparisons across all latitudes.

the observations across all ensemble members (Fyfe and Gillett 2014), a conclusion that holds no matter how the slope or its significance is determined. For example, although ensemble member 24 produces a North Pacific pattern of divergence that is unusually consistent with the observations, there is disagreement in δ values in the eastern equatorial Pacific, where the simulation has a more positive trend than in the observations. Kosaka and Xie (2013) show that restoring a model toward observed temperatures in the eastern equatorial Pacific results in global temperature variations very similar to observations, along with regional warming in arctic Eurasia and northeastern Canada similar to observations. Ding et al. (2014) demonstrate in a slab ocean model that a pattern of tropical temperature trends involving cooling in the eastern equatorial Pacific leads to increased warming focused in northeastern Canada as well as Greenland, also consistent with observed trends.

The apparent importance of the eastern equatorial Pacific for governing global temperature and the clear model–observation discrepancies in this region creates some tension with our conclusion that CMIP5 simulations are consistent with observations at the global scale. Indeed, the observed pattern of strong negative divergence in the tropics and positive divergence at high northern latitudes is not found among CMIP5 ensemble members (Fig. 6). Ensemble members with the most negative global δ values show negative divergences both in the tropics and at high northern latitudes. Thus, although the CMIP5 ensemble contains recent global temperature trends similar in magnitude to the observations, they are composed of differing regional patterns.

Our finding of global consistency but regional discrepancy between simulations and observations also reflects findings using earlier periods of the observational record. Examination of SST variability showed that, whereas global-scale decadal variability is consistent, decadal SST variability in 5° × 5° gridded observations is significantly larger than that found at comparable spatial scales in CMIP5 (Laepple and Huybers 2014a).

Examinations of decadal variability at the scale of the eastern equatorial Pacific, however, show consistency between CMIP5 and observations (Ault et al. 2013; Fyfe and Gillett 2014). Further study of these issues is warranted to include how simulations and observations compare as a function of spatial scale (e.g., Stott and Tett 1998), how irregular sampling and noise influence estimates of SST variability (e.g., Laepple and Huybers 2014b), and how model specification and resolution influence simulated variability (e.g., Stammer 2005).

Our results differ from Fyfe2013 in that we find no significant differences between global-scale temperature trends and those in the CMIP5 ensemble across all 15-yr intervals since 1970. Given that we otherwise replicate the hypothesis testing of Fyfe2013, the stability of our results comes from using a metric of divergence in trend that pivots from a long-term mean as well as exclusive use of a null distribution that accounts for intermodel spread. This reevaluation brings comparison of observed and CMIP5 trends into agreement with other analyses (e.g., Brown et al. 2015). At some level, uncovering flaws in models on the basis of observations would be an important scientific accomplishment, demonstrating the capacity to test climate model predictions and helping to focus future work, but our findings for global mean temperature demonstrate mere consistency. Regional discrepancies, however, highlight the continued utility of improving observational estimates, developing techniques for better comparing observations and models, and continued inquiry into the causes of regional trends.

Acknowledgments. We thank Geert Jan van Oldenborgh for facilitating access to CMIP5 simulation output through the KNMI climate explorer website and are grateful for comments provided by Lauren Kuntz, Thomas Laepple, Karen McKinnon, Cristian Proistosescu, Andrew Rhines, Daniel Schrag, Eric Stansifer, and Giuseppe Torri. Funding was provided by NSF Award 1304309 and an NSF Graduate Research Fellowship.

APPENDIX A

Derivation of Trend Estimators

Estimators for the simple linear regression slope s and intercept b are derived from the linear equation, where T'_k and ε_k are random variables:

$$T'_k = s(t_k - t_0) + b + \varepsilon_k, \quad k = 0, 1, 2, \ldots N. \tag{A1}$$

To simplify the subsequent algebraic expressions, N is defined as the number of data points after t_0, and $L = N + 1$. For a trend of length L years, s is fit to $k = 0 \ldots N$. The sum of the residual variance is defined as follows:

$$\sum_{k=0}^{N} \varepsilon_k^2 = \sum_{k=0}^{N} [T'_k - (sk + b)]^2. \tag{A2}$$

Setting the partial derivative with respect to s equal to zero,

$$\frac{\partial\left(\sum_{k=0}^{N} \varepsilon_k^2\right)}{\partial s} = \sum_{k=0}^{N} \{-2k[T'_k - (sk + b)]\} = 0, \tag{A3}$$

which yields the following expression for s:

$$s = \frac{-b\sum_{k=0}^{N} k + \sum_{k=0}^{N} T'_k k}{\sum_{k=0}^{N} k^2}. \tag{A4}$$

An expression for b follows similarly:

$$\frac{\partial\left(\sum_{k=0}^{N} \varepsilon_k^2\right)}{\partial b} = \sum_{k=0}^{N} [2(sk + b) - 2T'_k] = 0, \tag{A5}$$

$$b = \frac{1}{N+1}\left(\sum_{k=0}^{N} T'_k - s\sum_{k=0}^{N} k\right). \tag{A6}$$

Substituting Eq. (A6) into Eq. (A4) yields the least squares solution for slope s:

$$s = \frac{-\frac{1}{N+1}\sum_{k=0}^{N} T'_k \sum_{k=0}^{N} k + \sum_{k=0}^{N} T'_k k}{\sum_{k=0}^{N} k^2 - \frac{1}{N+1}\left(\sum_{k=0}^{N} k\right)^2}. \tag{A7}$$

A similar approach is used to find δ, where T'_k and η_k are random variables:

$$T'_k = \delta(t_k - t_0) + c + \eta_k, \quad k = 1, 2, \ldots N. \tag{A8}$$

Because the intercept at $k = 0$ is fixed at $T'_0 = c$, δ is computed using $k = 1 \ldots N$. The sum of the residual variance is defined as follows:

$$\sum_{k=1}^{N} \eta_k^2 = \sum_{k=1}^{N} [T'_k - (\delta k + c)]^2. \tag{A9}$$

Analogous to the derivation of s [Eq. (A4)], δ can be expressed in terms of c:

$$\delta = \frac{-c\sum_{k=1}^{N} k + \sum_{k=1}^{N} T'_k k}{\sum_{k=1}^{N} k^2}. \tag{A10}$$

Equation (A10) can be rewritten to include the computation of c. In the following, indexing of T' is shifted by $M - N$ to include the interval over which c is computed, where M is the total length of the time series T'. We define $c = \sum_{k=1}^{M-N} T'_k/(M-N)$, and substituting this into δ gives the following:

$$\delta = \frac{-\dfrac{\sum_{k=1}^{M-N} T'_k}{M-N}\sum_{k=1}^{N} k + \sum_{k=1}^{N} T'_{M-N+k}k}{\sum_{k=1}^{N} k^2}. \tag{A11}$$

APPENDIX B

Derivation of Trend Estimator Variances

The variance of the trend estimator given in Eq. (A4) can be expressed as follows:

$$\mathrm{VAR}(s) = \frac{\mathrm{VAR}\left\{\sum_{k=0}^{N}\left[T'_k\left(k - \frac{1}{N+1}\sum_{k=0}^{N} k\right)\right]\right\}}{\left[\sum_{k=0}^{N} k^2 - \frac{1}{N+1}\left(\sum_{k=0}^{N} k\right)^2\right]^2}. \tag{B1}$$

This equation can be substantially simplified. Defining the variance of T'_k as σ^2 and noting that $[1/(N+1)]\sum_{k=0}^{N} k$ is the average $\overline{k}$,

$$\mathrm{VAR}(s) = \frac{\sigma^2\sum_{k=0}^{N}(k-\overline{k})^2}{\left[\sum_{k=0}^{N} k^2 - \frac{1}{N+1}\left(\sum_{k=0}^{N} k\right)^2\right]^2}. \tag{B2}$$

Further simplification comes from the equality $\sum_{k=0}^{N}(k-\overline{k})^2 = \sum_{k=0}^{N} k^2 - [1/(N+1)](\sum_{k=0}^{N} k)^2$, yielding

$$\mathrm{VAR}(s) = \frac{\sigma^2}{\sum_{k=0}^{N} k^2 - \frac{1}{N+1}\left(\sum_{k=0}^{N} k\right)^2}. \tag{B3}$$

Finally, the integer addition identities $\sum_{k=1}^{N} k^2 = [N(N+1)(2N+1)]/6$ and $\sum_{k=1}^{N} k = [N(N+1)]/2$ allow for writing

$$\mathrm{VAR}(s) = \frac{12\sigma^2}{N(N+1)(N+2)} = \frac{12\sigma^2}{L^3 - L}. \tag{B4}$$

An expression can similarly be derived for the δ estimator given in Eq. (A10) if the variance of the intercept c is assumed negligible:

$$\mathrm{VAR}(\delta) = \frac{6\sigma^2}{L(L-1)(2L-1)}. \tag{B5}$$

The ratio of variances between the fixed-intercept and standard trend estimators is given by

$$\frac{\mathrm{VAR}(\delta)}{\mathrm{VAR}(s)} = \frac{L+1}{4L-2}. \tag{B6}$$

If, instead, the variance of the intercept estimator c is accounted for, the overall variance associated with δ is given by

$$\mathrm{VAR}(\delta) = \sigma^2\left[\frac{\frac{1}{M-N}\left(\sum_{k=1}^{M-N} k\right)^2}{\left(\sum_{k=1}^{N} k^2\right)^2} + \frac{1}{\sum_{k=1}^{N} k^2}\right], \tag{B7}$$

and the variance ratio becomes

$$\frac{\mathrm{VAR}(\delta)}{\mathrm{VAR}(s)} = \frac{3}{4(M-L+1)}\frac{L(L-1)(L+1)}{(2L-1)^2} + \frac{L+1}{4L-2}. \tag{B8}$$

For the 2000–14 trend examined in the main text, $L = 15$ and $M = 64$, which gives a variance ratio of 0.336. This ratio is only slightly higher in the presence of autocorrelation. For example, numerical simulations wherein ε_k and η_k are represented as a first-order autoregressive process fit to each of the T' calculated from the CMIP5 ensemble gives an average variance ratio of 0.349.

APPENDIX C

Derivation of Variances of Trend Differences between Intervals with Consecutive Start Years

Our choice of the δ estimator is guided by its greater stability than s between trend estimates of same-length intervals with consecutive start years. Stability can be demonstrated analytically by comparing the variances of the differences in trends between overlapping intervals offset by one year. The difference in consecutive trends s is first treated, where s^+ corresponds to an interval that increments all years of s by one:

$$s^+ - s = \frac{-\left(\frac{1}{N+1}\sum_{k=0}^{N} k\right)\left(\sum_{k=0}^{N} T'_{k+1} - \sum_{k=0}^{N} T'_k\right) + \sum_{k=0}^{N} T'_{k+1}k - \sum_{k=0}^{N} T'_k k}{\sum_{k=0}^{N} k^2 - \frac{1}{N+1}\left(\sum_{k=0}^{N} k\right)^2}, \tag{C1}$$

$$= \frac{-\left(\frac{1}{N+1}\sum_{k=0}^{N} k\right)(T'_{N+1} - T'_0) + NT'_{N+1} - \sum_{k=0}^{N-1} T'_{k+1}}{\sum_{k=0}^{N} k^2 - \frac{1}{N+1}\left(\sum_{k=0}^{N} k\right)^2}. \tag{C2}$$

Applying the variance operator and foregoing integer addition identities gives the following:

$$\mathrm{VAR}(s^+ - s) = \frac{144\sigma^2\left[\left(N - \frac{1}{N+1}\sum_{k=0}^{N} k\right)^2 + \left(\frac{\sum_{k=0}^{N} k}{N+1}\right)^2 + N\right]}{N^2(N+1)^2(N+2)^2} \tag{C3}$$

$$= \frac{72\sigma^2}{L^2(L-1)(L+1)}. \tag{C4}$$

The variance of consecutive δ slope estimates can be determined analogously if c is assumed the same for both intervals:

$$\mathrm{VAR}(\delta^+ - \delta) = \frac{36\sigma^2}{L(L-1)(2L-1)^2}. \tag{C5}$$

The ratio of the two variances is then given by

$$\frac{\mathrm{VAR}(\delta^+ - \delta)}{\mathrm{VAR}(s^+ - s)} = \frac{L(L+1)}{2(2L-1)^2}, \tag{C6}$$

having a value of 0.143 for $L = 15$.

Including the variance contributions from estimating c over an interval of length $M - N$ leads to a longer expression, where δ and δ^+ are estimated over $T'_{M-N+1} \ldots T'_M$ and $T'_{M-N+2} \ldots T'_{M+1}$, respectively:

$$\delta^+ - \delta = \frac{-\frac{1}{M-N}(T'_{M-N+1} - T'_1)\sum_{k=1}^{N} k + T'_{M+1}N - \sum_{k=1}^{N-1} T'_{M-N+k+1}T'_{M-N+1}}{\sum_{k=1}^{N} k^2}. \tag{C7}$$

The variance of this difference is given by

$$\mathrm{VAR}(\delta^+ - \delta) = \sigma^2\left[\frac{18}{(M-L+1)^2(2L-1)^2} + \frac{36}{L(L-1)(2L-1)^2} + \frac{36}{(M-L+1)L(L-1)(2L-1)^2}\right]. \tag{C8}$$

Note that the interval over which c is computed for δ^+ is equivalent in length but incremented by one year relative to that of δ. The variance ratio is then given by

$$\frac{\mathrm{VAR}(\delta^+ - \delta)}{\mathrm{VAR}(s^+ - s)} = \frac{L^2(L-1)(L+1)}{4(M-L+1)^2(2L-1)^2} + \frac{L(L+1)}{2(2L-1)^2} + \frac{L(L+1)}{2(M-L+1)(2L-1)^2} \tag{C9}$$

and is only slightly greater at 0.151 than when variance in c is neglected, given $M = 64$ and $L = 15$. The greater stability between estimates associated with δ reduces the potential for multiple tests involving different intervals to produce false positives.

REFERENCES

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

Ansari, A. R., and R. A. Bradley, 1960: Rank-sum tests for dispersions. *Ann. Math. Stat.*, **31**, 1174–1189, doi:10.1214/aoms/1177705688.

Ault, T., C. Deser, M. Newman, and J. Emile-Geay, 2013: Characterizing decadal to centennial variability in the equatorial Pacific during the last millennium. *Geophys. Res. Lett.*, **40**, 3450–3456, doi:10.1002/grl.50647.

Brown, P., W. Li, E. Cordero, and S. Mauget, 2015: Comparing the model-simulated global warming signal to observations using empirical estimates of forced noise. *Sci. Rep.*, **5**, 9957, doi:10.1038/srep09957.

Cess, R., and Coauthors, 1996: Cloud feedback in atmospheric general circulation models: An update. *J. Geophys. Res.*, **101**, 12 791–12 794, doi:10.1029/96JD00822.

Cohen, J., J. Furtado, M. Barlow, V. A. Alexeev, and J. Cherry, 2012: Asymmetric seasonal temperature trends. *Geophys. Res. Lett.*, **39**, L04705, doi:10.1029/2011GL050582.

Collins, M., B. B. B. Booth, B. Bhaskaran, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb, 2011: Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Climate Dyn.*, **36**, 1737–1766, doi:10.1007/s00382-010-0808-0.

Cowtan, K., and R. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quart. J. Roy. Meteor. Soc.*, **140**, 1935–1944, doi:10.1002/qj.2297.

——, and Coauthors, 2015: Robust comparison of climate models with observations using blended air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534, doi:10.1002/2015GL064888.

Deser, C., A. Philips, V. Bourdette, and H. Teng, 2012: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, doi:10.1007/s00382-010-0977-x.

Ding, Q., J. M. Wallace, D. S. Battisti, E. J. Steig, A. J. Gallant, H.-J. Kim, and L. Geng, 2014: Tropical forcing of the recent rapid arctic warming in northeastern Canada and Greenland. *Nature*, **509**, 209–212, doi:10.1038/nature13260.

England, M., and Coauthors, 2014: Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nat. Climate Change*, **4**, 222–227, doi:10.1038/nclimate2106.

Feldl, N., and G. H. Roe, 2013: The nonlinear and nonlocal nature of climate feedbacks. *J. Climate*, **26**, 8289–8304, doi:10.1175/JCLI-D-12-00631.1.

Flato, G., and Coauthors, 2013: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 741–866.

Fyfe, J., and N. Gillett, 2014: Recent observed and simulated warming. *Nat. Climate Change*, **4**, 150–151, doi:10.1038/nclimate2111.

——, ——, and F. Zwiers, 2013: Overestimated global warming over the past 20 years. *Nat. Climate Change*, **3**, 767–769, doi:10.1038/nclimate1972.

Huber, M., and R. Knutti, 2014: Natural variability, radiative forcing and climate response in the recent hiatus reconciled. *Nat. Geosci.*, **7**, 651–656, doi:10.1038/ngeo2228.

Huybers, P., 2010: Compensation between model feedbacks and curtailment of climate sensitivity. *J. Climate*, **23**, 3009–3018, doi:10.1175/2010JCLI3380.1.

Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348**, 1469–1472, doi:10.1126/science.aaa5632.

Kim, K.-Y., and G. R. North, 1991: Surface temperature fluctuations in a stochastic climate model. *J. Geophys. Res.*, **96**, 18 573–18 580, doi:10.1029/91JD01959.

Knutti, R., and J. Sedlácek, 2013: Robustness and uncertainties in the new CMIP5 climate model projections. *Nat. Climate Change*, **3**, 369–373, doi:10.1038/nclimate1716.

Kosaka, Y., and S.-P. Xie, 2013: Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, **501**, 403–407, doi:10.1038/nature12534.

Laepple, T., and P. Huybers, 2014a: Global and regional variability in marine surface temperatures. *Geophys. Res. Lett.*, **41**, 2528–2534, doi:10.1002/2014GL059345.

——, and ——, 2014b: Ocean surface temperature variability: Large model–data differences at decadal and longer periods. *Proc. Natl. Acad. Sci. USA*, **111**, 16 682–16 687, doi:10.1073/pnas.1412077111.

Li, C., B. Stevens, and J. Marotzke, 2015: Eurasian winter cooling in the warming hiatus of 1998–2012. *Geophys. Res. Lett.*, **42**, 8131–8139, doi:10.1002/2015GL065327.

Marotzke, J., and P. M. Forster, 2015: Forcing, feedback, and internal variability in global temperature trends. *Nature*, **517**, 565–570, doi:10.1038/nature14117.

Masson, D., and R. Knutti, 2011: Climate model genealogy. *Geophys. Res. Lett.*, **38**, L08703, doi:10.1029/2011GL046864.

Meehl, G. A., J. Arblaster, J. Fasullo, A. Hu, and K. Trenberth, 2011: Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods. *Nat. Climate Change*, **1**, 360–364, doi:10.1038/nclimate1229.

——, A. Hu, J. Arblaster, J. Fasullo, and K. E. Trenberth, 2013: Externally forced and internally generated decadal climate internal variability associated with the interdecadal Pacific oscillation. *J. Climate*, **26**, 7298–7310, doi:10.1175/JCLI-D-12-00548.1.

Morice, C., J. Kennedy, N. Rayner, and P. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset. *J. Geophys. Res.*, **117**, D08101, doi:10.1029/2011JD017187.

Otto, A., and Coauthors, 2013: Energy budget constraints on climate response. *Nat. Geosci.*, **6**, 415–416, doi:10.1038/ngeo1836.

Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*, **24**, 2358–2367, doi:10.1175/2010JCLI3814.1.

Rajaratnam, B., J. Romano, M. Tsiang, and N. S. Diffenbaugh, 2015: Debunking the climate hiatus. *Climatic Change*, **133**, 129–140, doi:10.1007/s10584-015-1495-y.

Risbey, J., S. Lewandowsky, C. Langlais, D. Monselesan, T. O'Kane, and N. Orekes, 2014: Well-estimated global surface warming in climate projections selected for ENSO phase. *Nat. Climate Change*, **4**, 835–840, doi:10.1038/nclimate2310.

Roe, G., and M. Baker, 2007: Why is climate sensitivity so unpredictable? *Science*, **318**, 629–632, doi:10.1126/science.1144735.

Rougier, J., M. Goldstein, and L. House, 2011: Second-order exchangeability analysis for multimodel ensembles. *J. Amer. Stat. Assoc.*, **108**, 852–863, doi:10.1080/01621459.2013.802963.

Sanderson, B., R. Knutti, and P. Caldwell, 2015: A representative democracy to reduce interdependency in multimodel ensemble. *J. Climate*, **28**, 5171–5194, doi:10.1175/JCLI-D-14-00362.1.

Santer, B., and Coauthors, 2014: Volcanic contributions to decadal changes in tropospheric temperature. *Nat. Geosci.*, **7**, 185–189, doi:10.1038/ngeo2098.

Schmidt, G., D. Shindell, and K. Tsigaridis, 2014: Reconciling warming trends. *Nat. Geosci.*, **7**, 158–160, doi:10.1038/ngeo2105.

Solomon, S., K. Rosenlof, R. Portman, J. Daniel, S. Davis, T. Sanford, and G. Plattner, 2010: Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science*, **327**, 1219–1223, doi:10.1126/science.1182488.

Stammer, D., 2005: Adjusting internal model errors through ocean state estimation. *J. Phys. Oceanogr.*, **35**, 1143–1153, doi:10.1175/JPO2733.1.

Stott, P. A., and S. F. Tett, 1998: Scale-dependent detection of climate change. *J. Climate*, **11**, 3282–3294, doi:10.1175/1520-0442(1998)011<3282:SDDOCC>2.0.CO;2.

Taylor, K., R. Stouffer, and G. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, doi:10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London*, **365**, 2053–2075, doi:10.1098/rsta.2007.2076.

Thompson, D., E. Barnes, C. Deser, W. Foust, and A. Phillips, 2015: Quantifying the role of climate internal variability in future climate trends. *J. Climate*, **28**, 6443–6456, doi:10.1175/JCLI-D-14-00830.1.

Thorne, P., S. Outten, I. Bethke, and O. Seland, 2015: Investigating the recent apparent hiatus in surface temperature increases: 2. Comparison of model ensembles to observations. *J. Geophys. Res. Atmos.*, **120**, 8597–8620, doi:10.1002/2014JD022805.

Trenberth, K., and J. Fasullo, 2013: An apparent hiatus in global warming? *Earth's Future*, **1**, 19–32, doi:10.1002/2013EF000165.

van Oldenborgh, G. J., 2015: KNMI climate explorer. [Available online at http://climexp.knmi.nl/start.cgi?id=someone@somewhere.]

Visser, H., S. Dangendorf, and A. Petersen, 2015: A review of trend models applied to sea level data with reference to the ''acceleration-deceleration debate.'' *J. Geophys. Res. Oceans*, **120**, 3873–3895, doi:10.1002/2015JC010716.

Watanabe, M., and Coauthors, 2012: Using a multiphysics ensemble for exploring diversity in cloud–shortwave feedback in GCMs. *J. Climate*, **25**, 5416–5431, doi:10.1175/JCLI-D-11-00564.1.

Wunsch, C., 1999: The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bull. Amer. Meteor. Soc.*, **80**, 245–255, doi:10.1175/1520-0477(1999)080<0245:TIOSCR>2.0.CO;2.

Zhang, Y., J. Wallace, and D. Battisti, 1997: ENSO-like interdecadal variability: 1900–93. *J. Climate*, **10**, 1004–1020, doi:10.1175/1520-0442(1997)010<1004:ELIV>2.0.CO;2.