AGU100 ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

RESEARCH ARTICLE
10.1029/2018EA000401

Key Points:
- Climate models embed countless minor hypotheses and at least one major, testable hypothesis related to equilibrium sensitivity
- We outline four criteria that identify a testable prediction of the major hypothesis, namely, tropical 200- to 300-hPa warming
- CMIP5 models show a large, significant, and uniform warm bias in that layer of sufficient magnitude to reject the major hypothesis

Supporting Information:
- Supporting Information S1

Correspondence to:
R. McKitrick,
ross.mckitrick@uoguelph.ca

Citation:
McKitrick, R., & Christy, J. (2018). A test of the tropical 200- to 300-hPa warming rate in climate models. *Earth and Space Science*, *5*, 529–536. https://doi.org/10.1029/2018EA000401

Received 3 APR 2018
Accepted 22 JUN 2018
Accepted article online 6 JUL 2018
Published online 21 SEP 2018

A Test of the Tropical 200- to 300-hPa Warming Rate in Climate Models

Ross McKitrick[1] and **John Christy**[2]

[1]Department of Economics and Finance, University of Guelph, Guelph, Canada, [2]Earth System Science Center, University of Alabama in Huntsville, Huntsville, AL, USA

Abstract Overall climate sensitivity to CO_2 doubling in a general circulation model results from a complex system of parameterizations in combination with the underlying model structure. We refer to this as the model's *major hypothesis*, and we assume it to be testable. We explain four criteria that a valid test should meet: measurability, specificity, independence, and uniqueness. We argue that temperature change in the tropical 200- to 300-hPa layer meets these criteria. Comparing modeled to observed trends over the past 60 years using a persistence-robust variance estimator shows that all models warm more rapidly than observations and in the majority of individual cases the discrepancy is statistically significant. We argue that this provides informative evidence against the major hypothesis in most current climate models.

1. Introduction

The purpose of this paper is to isolate a measurable feature of the climate that can serve as a testable index of the major hypothesis of the atmospheric component of general circulation models (GCMs). By *major hypothesis* we refer to the process of central interest both to most modelers and to the many users of GCM output, namely, the parameterized representation of moist thermodynamics and convection that, in combination with the underlying model structure, yields amplified warming of the atmosphere from greenhouse gases consistent with mainstream magnitudes of equilibrium climate sensitivity (ECS), namely 3.0 ± 1.5°C per doubling of carbon dioxide-equivalent. GCMs embed countless minor hypotheses subject to continual testing and revision. Due to the sheer complexity of the climate itself, and of climate models, any number of observed discrepancies between projections from an individual model and some local feature of the real world can be accommodated, rationalized, or ignored, without calling into question the model itself, since the rejected component could be removed without the rest of the model ceasing to be a member of the class of GCMs. We start from the assumption that there must, in principle, be at least one testable major hypothesis, the rejection of which would constitute failure of the model itself, in the sense that were the failed component to be removed, what remains would no longer be a GCM. We also start from the assumption that the measure we seek represents an emergent behavior from models based on both physical theory and modeler judgment, so that model integrations are genuinely expressions of a hypothesis (as opposed to computations of a known constant). The ensemble mean would thus represent the central tendency of modeler assumptions and is itself a testable quantity.

There are many identifiable predictions generated by climate models that could serve as test targets, but we propose four conditions that help us narrow the field down to a truly informative one: measurability, specificity, independence, and uniqueness. The first refers to the point that a prediction must refer to a target that is well measured over a long time span. This rules out testing ECS directly since neither it nor its transient counterpart are observable. It also limits the choice of temperature fields. Remote places like the polar surfaces are poorly sampled, creating known problems for assembling complete, homogeneous long-term temperature estimates. Many regions of the ocean are also poorly sampled or are subject only to recent measurement. Behavior of a target over a relatively short time span may be strongly affected by climate system internal variability or by exceptional events. An example of the latter is the influence of 1983 and 1992 volcanoes on the stratospheric temperature record, each of which led to temporary spikes twice as large as the multidecadal trend, thus making trend identification difficult over the post-1979 satellite record. Similar problems limit the usefulness of many other potential targets for furnishing testable predictions.

The second condition is that it must be a specific prediction, namely, one that reliably emerges across runs and across all models, on a specific temporal scale. To the extent the governing mechanisms in models

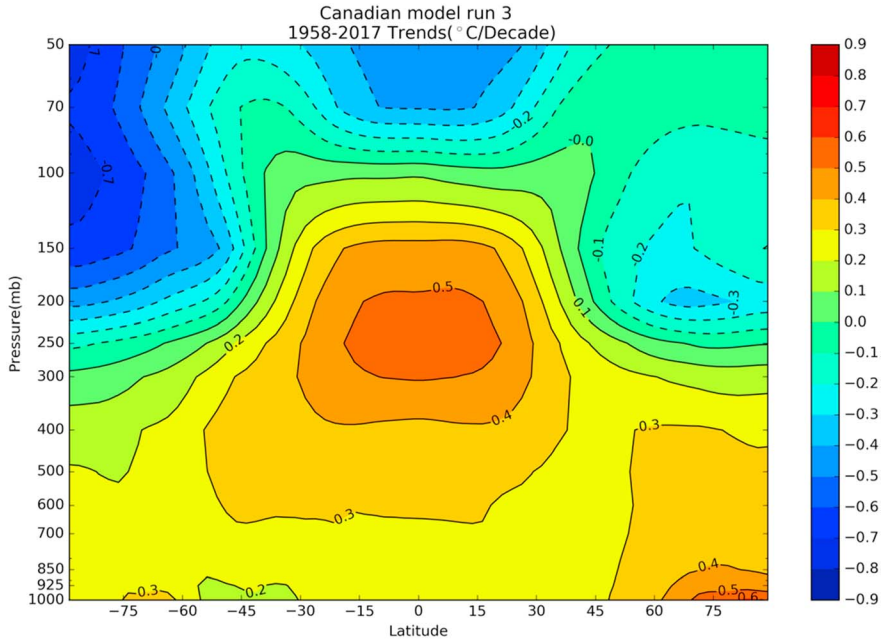

Figure 1. Warming pattern in Canadian model 1958–2017. Horizontal axis shows latitude, vertical axis shows altitude, and color shows warming trend magnitude.

reflect shared hypotheses among independent modeling teams we should expect to see coherent behavior in the target variable across independent runs, varying by magnitude but not by sign. The issue of timescale is equally important. One could endlessly shield a GCM from testing by arguing that while the magnitude of a projected change is precisely forecast, the timing is unknown to plus or minus several decades or centuries, so the failure to observe an expected change even in a lengthy data set only means that it is delayed. To avoid this dead end we confine attention to large, well-measured atmospheric regions where GCMs predict, more or less in unison, not only specific magnitudes of change but also on a specific (and reasonably rapid) timescale.

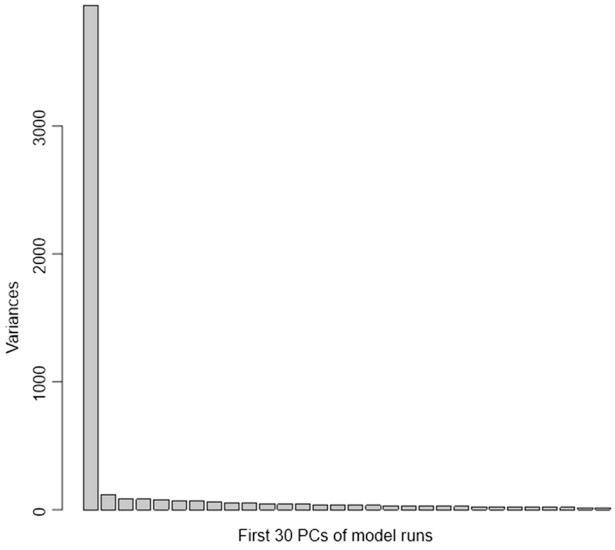

Figure 2. Scree plot of first 30 principal components of 102 modeled temperature simulations over 1958–2017.

Third, the independence criterion means that the target of the prediction must not be an input to the empirical tuning of the model. Once a model has been tuned to match a target, its reproduction of the target is no longer a test of its validity. In the case of GCMs, this rules out using the global average surface temperature record for testing, since during development models are often adjusted to broadly match its evolution over time. If the model structure is otherwise valid, such tuning practices should improve empirical fidelity, and the result should be that the model now makes more accurate predictions about other features of the atmosphere, measurements of which were not inputs to the tuning process. A good test ought therefore to focus on those other measures.

Finally, uniqueness refers to the causality behind the observed change. If the model predicts that greenhouse gases (GHGs) will cause the target to warm, but also predicts that many other factors could cause the target to warm, an observed warming would be less informative, since it is consistent both with a successful prediction and with a failed prediction coupled with the coincidental action of other causes. Ideally, then, we look for a prediction uniquely tied to the underlying causal mechanism of interest.

Air temperature in the 200- to 300-hPa layer of the tropical troposphere meets all four test conditions, pretty much uniquely in the climate

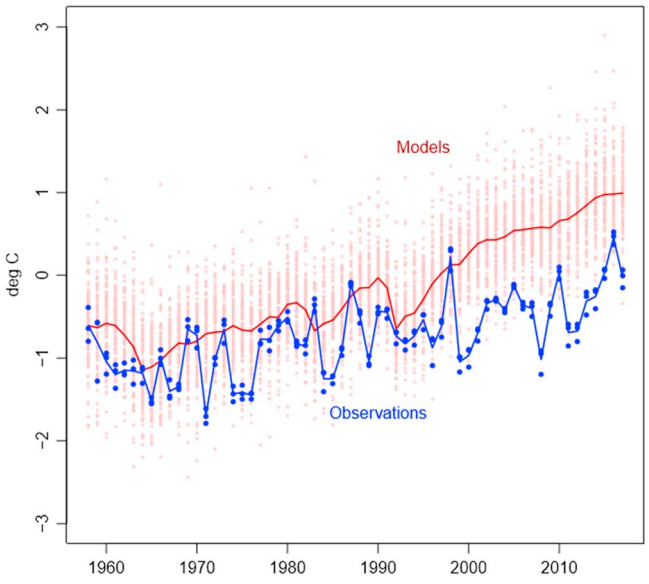

Figure 3. Model and observational data.

system as far as we are aware. First, homogenized measurements from more than one independent source are available over a 60-year span from 1958 to 2017. This is twice the length of the customary 30-year interval usually thought to be necessary for identifying a climatological phenomenon and more than enough compared to the response timescale in GCMs. The time span encompasses several major volcanoes and strong El Niño events, and the Pacific climate shift (PCS) of the late 1970s but is long enough to allow distinct identification of an underlying smooth trend, if one exists. Also, since it is part of the well-mixed free troposphere layer, there are fewer problems in obtaining a credible tropical-scale sample than is the case with surface measurements. For instance, Figure 17 in Christy et al. (2018) compares a variety of trends in tropical midtroposphere data products over the 1979-2016 interval. The radiosonde products, covering large parts of the tropical grid, yield results nearly identical to reanalysis products, which cover the entire grid.

Second, as was noted in the 2007 Fourth IPCC Assessment report (Meehl et al., 2007, Ch. 10), GCMs unanimously project that warming will reach a global maximum in the tropics near the 200- to 300-hPa layer, due to the so-called negative lapse rate feedback (National Academy of Sciences, 2003) and that the warming will occur rapidly in response to increased greenhouse forcing. Figure 1 shows the simulated 1958–2017 warming rates from the IPCC AR5 Canadian model, with the target zone visible as the red bullseye in the middle. Similar figures from models developed in the United States, the UK, and Germany are shown in the supporting information Figures S1–S3. Model representations of this layer's annual temperature series over our sample span are very coherent. Ninety-four percent of the possible cross-correlations among model runs exceed 0.5, and 77% exceed 0.6. The first principal component (PC) explains 73% of the variance across all 102 runs. What remains in the data is largely model-specific noise. Figure 2 shows the Scree plot of the first 30 PCs. After PC1 the next four each explain 2% or less and the remainder taper off quickly to very small levels, indicating that there is only one dominant signal common across models and across all runs by each model. The timescale is also well constrained. The average projected warming rate over 1958–2017 in the target layer is 0.33°C per decade, with a range spanning 0.18–0.51 °C per decade. Hence, models project on average that the total amount of warming in the target zone since 1958 should have been about 2 °C by now, a magnitude well within observational capability, and that the trends should be well established, thus specifying both a magnitude and a timescale.

Third, by focusing on the 200- to 300-hPa layer we avoid contaminating the test by searching for a signal to which the models were already tuned. The surface temperature record is ruled out for this reason, but satellite-based lower- and middle-troposphere composites are also somewhat contaminated since they

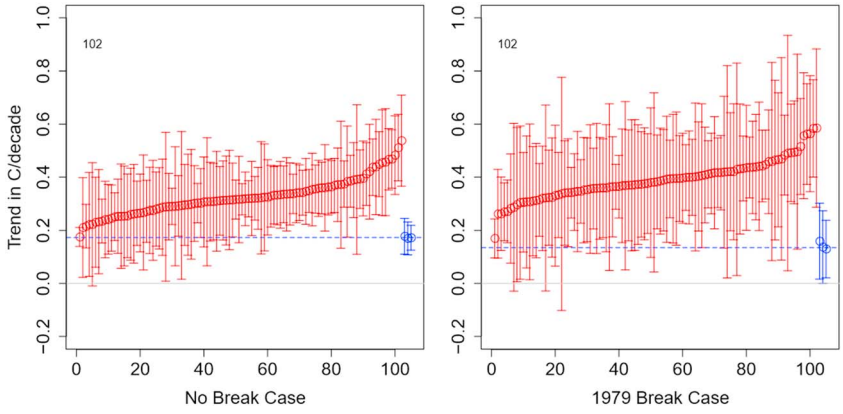

Figure 4. Trend magnitudes and 95% confidence intervals. Number in upper left corner indicates number of model trends (out of 102) that exceed observed average trend.

Table 1
Statistical Test Results

	Null hypothesis	Test score	*p* value
Restricted	RICH trend = 0	293.358	<0.0001
(No break)	RAOBCORE trend = 0	312.901	<0.0001
	RATPAC trend = 0	573.513	<0.0001
	Avg GCM trend = Avg Obs trend	81.114	0.013
General	RICH trend = 0	62.849	0.032
(With break)	RAOBCORE trend = 0	50.551	0.050
	RATPAC trend = 0	72.437	0.023
	Avg GCM trend = Avg Obs trend	260.698	0.0003

include the near-surface layer in their weighting functions. Radiosonde samples measure each layer of the atmosphere independently, not simply as a gradient against the surface.

Fourth, simulations in the IPCC AR4 Chapter 9 (Hegerl et al., 2007) indicate that, within the framework of mainstream GCMs, greenhouse forcing provides the only explanation for a strong warming trend in the target region. AR4 Figure 9.1 illustrates 20th-century climatic reconstructions applying one-at-a-time individual forcings from observed solar, volcanic, GHG, stratospheric ozone, and sulfate aerosol changes. Solar forcing yields an amplified warming aloft in the tropics, but the magnitude of change is very small, and the IPCC elsewhere emphasizes that actual historical trends in solar output have been too small to cause much atmospheric warming (AR4 Sct. 2.7, Forster et al., 2007). Only GHG forcing yields a large modeled warming pattern in the tropical 200- to 300-mb layer, which accords with the finding above that PC decomposition identifies only one common signal. Such a warming trend in the atmosphere, were it to be observed, would thus have only one explanation; likewise, its absence would conflict with only one major hypothesis of the model, namely, the set of parameterizations that yield amplified GHG-induced warming.

We make use herein of the latest releases of three radiosonde data sets, the U.S. National Oceanic and Atmospheric Administration's Radiosonde Atmospheric Temperature Products for Assessing Climate (RATPAC-A v2, Durre & Yin, 2011), as well as the University of Wien's RAdiosonde Observation Correction using Reanalyses (RAOBCORE v1.5) and Radiosonde Innovation Composite Homogenization (RICH v1.5, Haimberger et al., 2012). All data begin in 1958 when radiosonde coverage expanded around the globe for the International Geophysical Year and continue to the end of 2017. These series are compared against the complete ensemble of 102 model runs prepared for the Climate Model Intercomparison Project Number 5 (CMIP5) used in the most recent IPCC report (see Flato et al., 2013). The model output was obtained and used as-is from the Koninklijk Nederlands Meteorologisch instituut Climate Data Explorer site (van Oldenborg, 2016). Since autocorrelation structures in climate data can be complex and may differ among data types (see, e.g., Varotsos et al., 2013) we use a variance estimation methodology robust to general forms of heteroskedasticity and autocorrelation (McKitrick & Vogelsang, 2014; Vogelsang & Franses, 2005). We also allow for a possible break term at 1979 associated with the Pacific climate shift (see Seidel & Lanzante, 2004; Tsonis et al., 2007; Powell & Xu, 2011, and extensive references therein). What we refer to as the general trend model allows for a step change at 1979, while the restricted model does not.

Our test is directed at the response size, rather than the sign. Our real-world analogue exhibits a change in the predicted direction, so we focus our testing on whether the magnitude of change is consistent with the model prediction. As we will show, all 102 model runs warm more rapidly than observations, whether or not we allow for a break term. Most of the divergences are individually significant. We reject the hypothesis that the average model trend matches the average observed trend, regardless of the inclusion of a break term. Thus, the observed data are inconsistent with the major hypothesis of GCMs as represented by the selected target variable.

2. Data and Methods

2.1. Data Sources

We use the temperature data collected at the tropical (20S to 20N) 200–300 hPa as found in three data products (RAOBCORE, RICH, and RATPAC), taking annual averages as this is the finest time resolution available from RATPAC for the pressure-level quantity we require. Our test metric is the simple average of the 200-, 250-, and 300-hPa temperature value provided at those levels in both the radiosonde data sets and models. Radiosonde discontinuities, for which adjustments were applied in the homogenization process, generally occur in the stratosphere where solar heating of the instrument package led to spurious readings. Tropical radiosonde observations of the 200- to 300-hPa layer were also affected by this problem at several of the tropical stations. One method to address it is to compare the daytime readings (affected) with the nighttime readings (unaffected, e.g., Sherwood et al., 2005). Corrections for this and other issues were calculated and applied to the data sets by their authors.

Table 2
Model-Specific Test Scores

	No brk	w/ brk
ACCESS1-0	26.6	198.1
ACCESS1-3	5.8	97.2
bcc-csm1-1	220.1	593.3
bcc-csm1-1-m	142.6	324.8
BNU-ESM	174.3	162.2
CanESM2	410.5	511.1
CanESM2	110.1	59.2
CanESM2	410.3	534.4
CanESM2	114.9	256.7
CanESM2	393.0	270.9
CCSM4	103.1	212.2
CCSM4	270.9	359.2
CCSM4	223.3	266.7
CCSM4	258.1	430.6
CCSM4	193.3	242.2
CCSM4	244.9	243.3
CESM1-BGC	180.3	181.4
CESM1-CAM5	27.0	79.3
CESM1-CAM5	13.0	65.7
CESM1-CAM5	14.6	29.5
CMCC-CM	114.3	186.8
CMCC-CMS	100.7	65.3
CNRM-CM5	32.2	149.7
CSIRO-Mk3-6-0	3.7	22.1
CSIRO-Mk3-6-0	17.5	16.6
CSIRO-Mk3-6-0	16.0	120.3
CSIRO-Mk3-6-0	9.6	40.9
CSIRO-Mk3-6-0	2.5	25.6
CSIRO-Mk3-6-0	10.1	95.3
CSIRO-Mk3-6-0	4.4	67.2
CSIRO-Mk3-6-0	32.3	50.1
CSIRO-Mk3-6-0	2.2	26.1
CSIRO-Mk3-6-0	9.4	81.2
EC-EARTH	296.0	222.5
FGOALS-g2	180.5	229.6
FIO-ESM	224.2	143.0
FIO-ESM	119.0	39.8
FIO-ESM	129.2	310.9
GFDL-CM3	31.0	49.8
GFDL-ESM 2G	131.8	109.3
GFDL-ESM 2M	156.4	183.5
GISS-E2-H	49.5	153.3
GISS-E2-H	157.3	444.8
GISS-E2-H	49.2	325.4
GISS-E2-H	119.0	454.2
GISS-E2-H	164.9	191.1
GISS-E2-H	45.9	147.6
GISS-E2-H	23.4	75.1
GISS-E2-H	42.5	220.9
GISS-E2-H	18.6	104.9
GISS-E2-H	34.1	116.9
GISS-E2-H	40.8	406.0
GISS-E2-H	51.4	126.3
GISS-E2-H	63.7	190.8
GISS-E2-H	56.2	148.5
GISS-E2-H	24.0	79.8
GISS-E2-H-CC	139.0	468.5
GISS-E2-R	62.8	146.1
GISS-E2-R	236.3	270.8
GISS-E2-R	106.5	154.2

The climate model simulations utilized the "representative concentration pathway 4.5" (rcp4.5), which employs the best estimate of historical forcings through 2006, then anticipated forcings through 2100. For our purposes here, there is no difference through 2017 between rcp4.5 and the other rcps.

The time series data are shown in Figure 3. The light red dots show the complete year-by-year array of individual anomaly values from CMIP5. The red line is the annual mean of CMIP5 anomalies. The blue line is the mean of the three observational series, which are shown individually as blue dots. These are positioned so that the year-by-year observational mean starts at the same value as the corresponding model mean. Our statistical analysis focuses on a comparison of trends, which is not affected by the choice of crossing point for the data.

2.2. Statistical Methods

There are $i = 1, \ldots, 105$ temperature series of interest, of which 102 are model runs and three are observations. The general linear trend model is

$$y_t^i = a^i + b^i t + d^i D(\lambda) + e_t^i \quad (1)$$

where y_t^i is temperature series i over the time interval $t = 1, \ldots, T$, $D(\lambda)$ is a T-length indicator variable equal to zero for the first fraction λ of the sample time period and one thereafter, which thereby captures a possible step change in the sample mean; e_t^i is an error term assumed to be autocorrelated to an unknown extent up to but not including a unit root; and the parameters to be estimated are the constant a^i, the trend term b^i, and the break term d^i. In the restricted model the latter parameter is set equal to zero and equation (1) reduces to a simple linear trend. Since a break term was identified by McKitrick and Vogelsang (2014) at 1979, which coincides with prior information from other sources regarding the PCS, we impose $\lambda = 0.35$ based on the length of our sample.

We test null hypotheses of the form $H_0 : Rb = r$, where R is a 105-length restriction vector, b is the vector of slope coefficients, and r is a constant. A test that the average model trend equals the average observed trend would be formed by setting each of the first 102 elements of R to 1/102 and the final three each to $-1/3$, and $r = 0$. The test statistic VF, based on Vogelsang and Franses (2005), is

$$VF = \left(R\widehat{b} - r\right)' \left[\left(\sum_{t=1}^{T} \widetilde{t}^2\right)^{-1} R\widehat{\Omega}_T R'\right]^{-1} \left(R\widehat{b} - r\right) \quad (2)$$

where $\widehat{\ }$ denotes a least-squares estimator, $\widetilde{t}$ is the residual vector from a regression of t on $a^i + d^i D(\lambda)$, and $\widehat{\Omega}_T$ is the heteroskedasticity- and autocorrelation-consistent variance-covariance matrix for $\widehat{b}$ as derived in McKitrick and Vogelsang (2014). The VF test statistic has attractive size and power characteristics for tests of the kind we are doing here. If the true error process is a short-memory, one-lag autocorrelation it performs very similarly to a standard AR1 model. As the true lag structure grows, VF preserves power while minimizing size distortions (tendency to overreject) compared to other test options. VF is like a standard F statistic but follows a nonstandard distribution. If $\lambda = 0$ the critical values are as given in Vogelsang and Franses (2005). In our application $\lambda = 0.35$ and to generate appropriate critical values we apply the bootstrap methodology outlined in McKitrick and Vogelsang (2014).

Table 2 *(continued)*

	No brk	w/ brk
GISS-E2-R	254.7	258.7
GISS-E2-R	382.4	237.7
GISS-E2-R	63.9	149.2
GISS-E2-R	15.5	87.1
GISS-E2-R	14.2	119.9
GISS-E2-R	21.6	58.9
GISS-E2-R	12.6	98.2
GISS-E2-R	8.3	80.5
GISS-E2-R	29.9	93.8
GISS-E2-R	38.4	126.1
GISS-E2-R	83.4	179.9
GISS-E2-R	31.0	92.6
GISS-E2-R	18.7	151.1
GISS-E2-R	117.0	259.1
GISS-E2-R-CC	97.4	187.3
HadGEM2-AO	36.9	64.3
HadGEM2-CC	38.9	48.7
HadGEM2-ES	50.0	575.4
HadGEM2-ES	14.9	49.1
HadGEM2-ES	41.8	55.7
HadGEM2-ES	44.3	77.6
INMCM4	0.0	2.9
IPSL-CM5A-LR	101.7	145.8
IPSL-CM5A-LR	214.7	61.8
IPSL-CM5A-LR	330.7	109.9
IPSL-CM5A-LR	281.8	473.8
IPSL-CM5A-MR	114.4	121.2
IPSL-CM5B-LR	182.2	134.6
MIROC5	8.8	44.4
MIROC5	62.6	67.0
MIROC5	20.1	228.6
MIROC-ESM	38.4	121.4
MIROC-ESM-CHEM	61.2	57.9
MPI-ESM-LR	332.1	188.4
MPI-ESM-LR	76.4	41.1
MPI-ESM-LR	105.0	176.6
MPI-ESM-MR	177.2	212.9
MPI-ESM-MR	112.5	222.3
MPI-ESM-MR	308.9	215.0
MRI-CGCM3	16.4	120.2
NorESM1-M	56.2	55.2
NorESM1-ME	25.6	35.3

Note. First column: test score for restricted case (no break). Score is significant at 5% if it exceeds 41.53. Second column: test score for unrestricted case (with break at 1979). Score is significant at 5% if it exceeds 50.48.

As with *F*-type statistics generally, when only one restriction is being tested (as is the case herein), a *t*-type statistic can be formed by taking the square root of equation (2). Its denominator times the corresponding 0.025 critical value can be used to determine the width of the 95% confidence interval. This is the method used whenever we report such confidence intervals herein.

Regarding the break parameter, if one is needed but is omitted from the trend model, the slope parameter will be biased. If it is not needed but one is included anyway, the trend coefficient will not be biased but the *VF* test statistic will exhibit a slight loss of power (McKitrick & Vogelsang, 2014); in other words, it will tend to underreject the null hypothesis in equation (2). Consequently, for our application herein, inclusion of the break term is a conservative option since it will yield a tendency to underestimate the significance of model-observational trend discrepancies.

In addition to the hypothesis tests using *VF* we construct 102 divergence terms using equation (1) as above after replacing y_t^i with $\left(m_t^i - \bar{r}_t\right)$, where m_t^i is the temperature time series from model *i* and $\bar{r}_t$ is the average of the three corresponding radiosonde series for that layer. The rationale is that, if observations and model runs share common cyclical or aperiodic residual patterns, the difference series will remove them and reveal any structural divergence more precisely. For example, both the observations and the model simulations include the cooling impact of volcanoes in 1963, 1982, and 1991 whose ephemeral temporal structure is reasonably represented in both. In addition, the 60-year temperature impact on the real atmospheric layer by extra greenhouse gases will be removed by this procedure. A positive trend in the remaining divergence series will thus indicate a structural bias in the model related to long-term changes.

All computations were done using the R programming language. Data and code are available from the authors on request.

3. Results

All series-specific trends and confidence intervals are reported in the supporting information Table S1. The mean restricted trend (without a break term) is 0.325 ± 0.132°C per decade in the models and 0.173 ± 0.056°C per decade in the observations. With a break term included they are 0.389 ± 0.173°C per decade (models) and 0.142 ± 0.115°C per decade (observed). Figure 4 shows the individual trend magnitudes. The red circles and confidence interval whiskers are from models, and the blue are observed.

Table 1 reports the main statistical test scores. The restricted trends in all three data sets are significant, as are the general trends, although the RAOBCORE 95% trend confidence interval has a lower bound essentially at zero when the break term is included. The *p* value of the test of equivalence between the GCM ensemble average trend and that of the observational average is 0.013 (restricted case) and 0.0003 (general case), clearly rejecting the null hypothesis of trend equivalence.

Table 2 lists the model-specific comparisons against the average observational series. In the restricted case, 62 of 102 models reject, while in the general case, 87 of 102 models reject. It is striking that all model runs exhibit too much warming and in a clear majority of cases the discrepancies are statistically significant.

Figure 5 shows the histograms of trends in the divergence terms. If models accurately represented the magnitude of 200- to 300-hPa warming with only nonsystematic errors contributing noise, these distributions would be centered on zero. Clearly, they are centered above zero, in fact in both the restricted and general cases, the entire distribution is above zero. Table S2 presents individual run test results. In the restricted case,

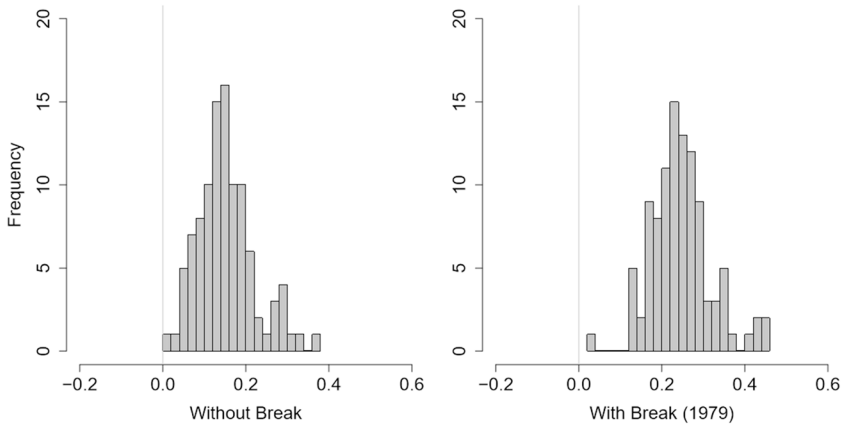

Figure 5. Histograms of divergence trend terms.

62 of the 102 divergence terms are significant, while in the general case, 87 of 102 are. The model-observational discrepancy is not simple uncertainty or random noise but represents a structural bias shared across models.

4. Discussion and Conclusions

We propose four conditions that a prediction test must meet to be informative regarding the major hypothesis embedded within GCMs concerning climate sensitivity to GHGs: measurability, specificity, independence, and uniqueness. Temperatures in the tropical 200- to 300-hPa layer meet all four conditions. We present a trend model robust to general forms of autocorrelation and the possible existence of a step change associated with the 1979 PCS. Comparing observed trends to those predicted by models over the past 60 years reveals a clear and significant tendency on the part of models to overstate warming. All 102 CMIP5 model runs warm faster than observations, in most individual cases the discrepancy is significant, and on average the discrepancy is significant. The test of trend equivalence rejects whether or not we include a break at 1979 for the PCS, though the rejections are stronger when we control for its influence. Measures of series divergence are centered at a positive mean and the entire distribution is above zero. While the observed analogue exhibits a warming trend over the test interval it is significantly smaller than that shown in models, and the difference is large enough to reject the null hypothesis that models represent it correctly, within the bounds of random uncertainty.

Swanson (2013) noted that the changes in model output between CMIP3 and CMIP5 improved the fit to Arctic warming but worsened it everywhere else, raising the possibility that the models were getting the Arctic right for the wrong reasons. In the same vein we argue that to the extent GCMs are getting some features of the surface climate correct as a result of their current tuning, they are doing so with a flawed structure. If tuning to the surface added empirical precision to a valid physical representation, we would expect to see a good fit between models and observations at the point where the model predicts the clearest and strongest thermodynamic response to greenhouse gases. Instead, we observe a discrepancy across all runs of all models, taking the form of a warming bias at a sufficiently strong rate as to reject the hypothesis that the models are realistic. Our interpretation of the results is that the major hypothesis in contemporary climate models, namely, the theoretically based negative lapse rate feedback response to increasing greenhouse gases in the tropical troposphere, is incorrect. Further diagnosis of the nature of the inaccuracy is beyond this analysis: For discussion see, for example, Spencer and Braswell (2014), Lewis and Curry (2014), and Christy and McNider (2017).

Acknowledgments
No funding was received for this research. All data and R code will be available upon publication at rossmckitrick.com and are available on request from the corresponding author. First author additional affiliations are as follows: Senior Fellow, Fraser Institute, Adjunct Scholar, Cato Institute.

References

Christy, J. R., & McNider, R. T. (2017). Satellite bulk tropospheric temperatures as a metric for climate sensitivity. *Asia-Pacific Journal of Atmospheric Sciences*, *53*(4), 511–518. https://doi.org/10.1007/s13143-017-0070-z

Christy, J. R., Spencer, R. W., Braswell, W. D., & Junod, R. (2018). Examination of space-based bulk atmospheric temperatures for climate research. *International Journal of Remote Sensing*, *39*(11), 3580–3607. https://doi.org/10.10180/01431161.2018.1444293

Durre, I., & Yin, X. (2011). Enhancements of the dataset of sounding parameters derived from the Integrated Global Radiosonde Archive. 23rd Conference on Climate Variability and Change, Seattle, WA, 25 January 2011. Retrieved from https://ams.confex.com/ams/91Annual/webprogram/Paper179437.html

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of Climate Models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, et al. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741–866). Cambridge, UK and New York: Cambridge University Press.

Forster, P., Ramaswamy, V., Artaxo, P., Berntsen, T., Betts, R., Fahey, D. W., et al. (2007). Changes in Atmospheric Constituents and in Radiative Forcing. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, et al. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 129–234). Cambridge, UK and New York: Cambridge University Press.

Haimberger, L., Tavolato, C., & Sperka, S. (2012). Homogenization of the global radiosonde temperature dataset through combined comparison with reanalysis background series and neighboring stations. *Journal of Climate*, *25*(23), 8108–8131. https://doi.org/10.1175/jcli-d-11-00668.1

Hegerl, G. C., Zwiers, F. W., Braconnot, P., Gillett, N. P., Luo, Y., Marengo Orsini, J. A., et al. (2007). Understanding and attributing climate change. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, et al. (Eds.), *Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change* (pp. 663–745). Cambridge, UK and New York: Cambridge University Press.

Lewis, N., & Curry, J. (2014). The implications for climate sensitivity of AR5 forcing and heat uptake estimates. *Climate Dynamics*, *45*(3–4), 1009–1023. https://doi.org/10.1007/s00382-014-2342-y

McKitrick, R. R., & Vogelsang, T. (2014). HAC-robust trend comparisons among climate series with possible level shifts. *Environmetrics*, *25*(7), 528–547. https://doi.org/10.1002/env.2294

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., et al. (2007). Global climate projections. In S. Solomon, et al. (Eds.), *Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change* (pp. 747–845). Cambridge, UK and New York: Cambridge University Press.

National Academy of Sciences (2003). Cloud, water vapor, and lapse rate feedbacks. In *Understanding climate change feedbacks* (pp. 21–40). Washington, DC: National Academy Press. Retrieved from www.nap.edu

van Oldenborg, G. J. (2016). Climate data explorer. Retrieved from http://climexp.knmi.nl/selectfield_co2.cgi?someone@somewhere

Powell, A. M. Jr., & Xu, J. (2011). Abrupt climate regime shifts, their potential forcing and fisheries impacts. *Atmospheric and Climate Sciences*, *01*(02), 33–47. https://doi.org/10.4236/acs.2011.12004

Seidel, D. J., & Lanzante, J. R. (2004). An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *Journal of Geophysical Research*, *109*, D14108. https://doi.org/10.1029/2003JD004414

Sherwood, S. C., Lanzante, J., & Meyer, C. (2005). Radiosonde daytime biases and late 20th century warming. *Science*, *309*, 1556–1559.

Spencer, R. W., & Braswell, W. D. (2014). The role of ENSO in global ocean temperature changes during 1955-2011 simulated with a 1D climate model. *Asia-Pacific Journal of Atmospheric Sciences*, *50*(2), 229–237. https://doi.org/10.1007/s13143-014-0011-z

Swanson, K. L. (2013). Emerging selection bias in large-scale climate change simulations. *Geophysical Research Letters*, *40*, 3184–3188. https://doi.org/10.1002/grl.50562

Tsonis, A., Swanson, K., & Kravtsov, S. (2007). A new dynamical mechanism for major climate shifts. *Geophysical Research Letters*, *34*, L13705. https://doi.org/10.1029/2007GL030288

Varotsos, C. A., Efstathiou, M. N., & Cracknell, A. P. (2013). Plausible reasons for the inconsistencies between the modeled and observed temperatures in the tropical troposphere. *Geophysical Research Letters*, *40*, 4906–4910. https://doi.org/10.1002/grl.50646

Vogelsang, T. J., & Franses, P. H. (2005). Testing for common deterministic trend slopes. *Journal of Econometrics*, *126*(1), 1–24. https://doi.org/10.1016/j.jeconom.2004.02.004