DEBUNKING THE CLIMATE HIATUS

By

Bala Rajaratnam
Joseph Romano
Michael Tsiang
Noah S. Diffenbaugh

Technical Report No. 2015-16
August 2015

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

DEBUNKING THE CLIMATE HIATUS

By

Bala Rajaratnam
Joseph Romano
Michael Tsiang
Noah S. Diffenbaugh
Stanford University

Technical Report No. 2015-16
August 2015

**This research was supported in part by
National Science Foundation grants CMG 1025465,
AGS 1003823, DMS 1106642, DMS 1007732, and DMS 1307973.**

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://statistics.stanford.edu

This paper is to appear in *Climatic Change*.

Debunking the climate hiatus

Bala Rajaratnam[1,2,4], Joseph Romano[1,3], Michael Tsiang[2], and Noah S. Diffenbaugh[2,4]

[1]Department of Statistics, Stanford University
[2]Department of Earth System Science, Stanford University
[3]Department of Economics, Stanford University
[4]Woods Institute for the Environment, Stanford University

July 30, 2015

Abstract

The reported "hiatus" in the warming of the global climate system during this century has been the subject of intense scientific and public debate, with implications ranging from scientific understanding of the global climate sensitivity to the rate in which greenhouse gas emissions would need to be curbed in order to meet the United Nations global warming target. A number of scientific hypotheses have been put forward to explain the hiatus, including both physical climate processes and data artifacts. However, despite the intense focus on the hiatus in both the scientific and public arenas, rigorous statistical assessment of the uniqueness of the recent temperature time-series within the context of the long-term record has been limited. We apply a rigorous, comprehensive statistical analysis of global temperature data that goes beyond simple linear models to account for temporal dependence and selection effects. We use this framework to test whether the recent period has demonstrated i) a hiatus in the trend in global temperatures, ii) a temperature trend that is statistically distinct from trends prior to the hiatus period, iii) a "stalling" of the global mean temperature, and iv) a change in the distribution of the year-to-year temperature increases. We find compelling evidence that recent claims of a "hiatus" in global warming lack sound scientific basis. Our analysis reveals that there is no hiatus in the increase in the global mean temperature, no statistically significant difference in trends, no stalling of the global mean temperature, and no change in year-to-year temperature increases.

Keywords: Global Warming Hiatus Warming Slowdown Internal Variability Hypothesis Testing Uncertainty Quantification

1 Introduction, Motivation and Approach

The international debate on the "hiatus" in the warming of the global climate system over the last 15 years has intensified (e.g., Meehl et al. [2011], IPCC [2013], Otto et al. [2013], Fyfe et al. [2013], Kosaka and Xie [2013], Santer et al. [2014], Trenberth and Fasullo [2013], Smith [2013], Guemas et al. [2013], Chen and Tung [2014], Boykoff [2014], Hawkins et al. [2014], England et al. [2014], Karl et al. [2015], Cowtan et al. [2015]). The implications of the purported hiatus (also referred to as a "pause" or "slowdown") are far reaching. First, contradictory scientific conclusions have emerged regarding the relationship between climate change and anthropogenic global warming, especially during a period of heightened carbon emissions [Kosaka and Xie, 2013]. Second, the discrepancy between climate model projections and observations appear to point to an overestimation of climate sensitivity to anthropogenic forcings [Otto et al., 2013, Fyfe et al., 2013].

The perceived hiatus has led to a myriad of resources being expended on trying to better understand the geophysical mechanisms that lead to a possible hiatus, including, among others, volcanic activity [Santer et al., 2014], Pacific Ocean variability [Kosaka and Xie, 2013, Trenberth and Fasullo, 2013], and increased ocean heat uptake [Smith, 2013, Guemas et al., 2013, Chen and Tung, 2014], as well as spurious artifacts of the global climate observing system [Durack et al., 2014, Cowtan and Way, 2014, Karl et al., 2015, Cowtan et al., 2015]. The purported hiatus has therefore inspired valuable scientific insight into the processes that regulate decadal-scale variations of the climate system. However, the perception of a hiatus has important repercussions for public decision making, as the implications that global warming has paused or slowed down [Boykoff, 2014, Hawkins et al., 2014], and that climate models have overestimated the rate of warming (e.g., Fyfe et al. [2013]), influence the perceived level of mitigation action that is needed to obtain particular policy targets [Otto et al., 2013]. In addition, fundamental to any work on the hiatus is to ascertain whether there is sufficient empirical evidence in support of its existence. Surprisingly, to our knowledge, a rigorous statistical analysis has not been undertaken, at least not one which incorporates temporal dependencies without making strong assumptions about the underlying process. Without empirical evidence in support of the hiatus claims, any further conclusions stemming from the assumption should be called into question.

As a part of our investigation to better understand the hiatus, we develop a comprehensive scientific framework that is intended to systematically test hypotheses that have been implied in statements claiming a hiatus in global warming. We first identify a typology of the scientific assertions that have been put forward, including i) that there has been a hiatus in the trend in global warming, ii) that there is a difference in trends before and during the hiatus, iii) that there has been a hiatus in the change in mean global temperature, and iv) that there is a difference in warming before and during the hiatus (when accounting for possibly non-linear increases without explicit reference to a linear trend). (See Supplementary Section 3.4 for more detail on the typology.) We next connect these scientific claims with four classes of distinct testable statistical hypotheses, with each hypothesis focusing on different aspects of the underlying (unknown) temperature process. We then identify and develop appropriate statistical tools in order to test each of these hypotheses in a principled manner, and under progressively less restrictive - and therefore more generally applicable modeling assumptions, thereby allowing for a deeper understanding of the nuances of the global temperature time series. In particular, we attempt to properly account for temporal dependence, we use less restrictive resampling methods to assess statistical significance, and we employ a flexible nonparametric modeling approach. By applying these progressively more general techniques in a cascading approach, we are able to test the extent to which invalid statistical assumptions can lead to erroneous scientific conclusions.

Our analysis is first undertaken using NASA-GISS global mean land-ocean temperature index. It is subsequently also repeated on the NOAA and HadCRUT4 datasets for comparison purposes (see Supplemental Tables 1 and 2). The analysis is also undertaken on the recently released ERSSTv4 (Karl et al. [2015]) datasets (see Supplemental Tables 11 and 12). Plots of the NASA-GISS raw and smoothed global mean land-ocean temperature index from 1880 to 2013, with the base period 1951-1980, are given in in Figure 1 (top). As there is a clear underlying trend, a moving average is superimposed on the time series. A statistical analysis of the serial correlation in the residuals after fitting a regression line is also given in Figure 1 (top). The autocorrelation in the temperature time series is non-negligible. The presence of autocorrelation motivates the need to use less naive statistical methods to understand the evolution of temperature over time (see also Supplemental Section 2.2).

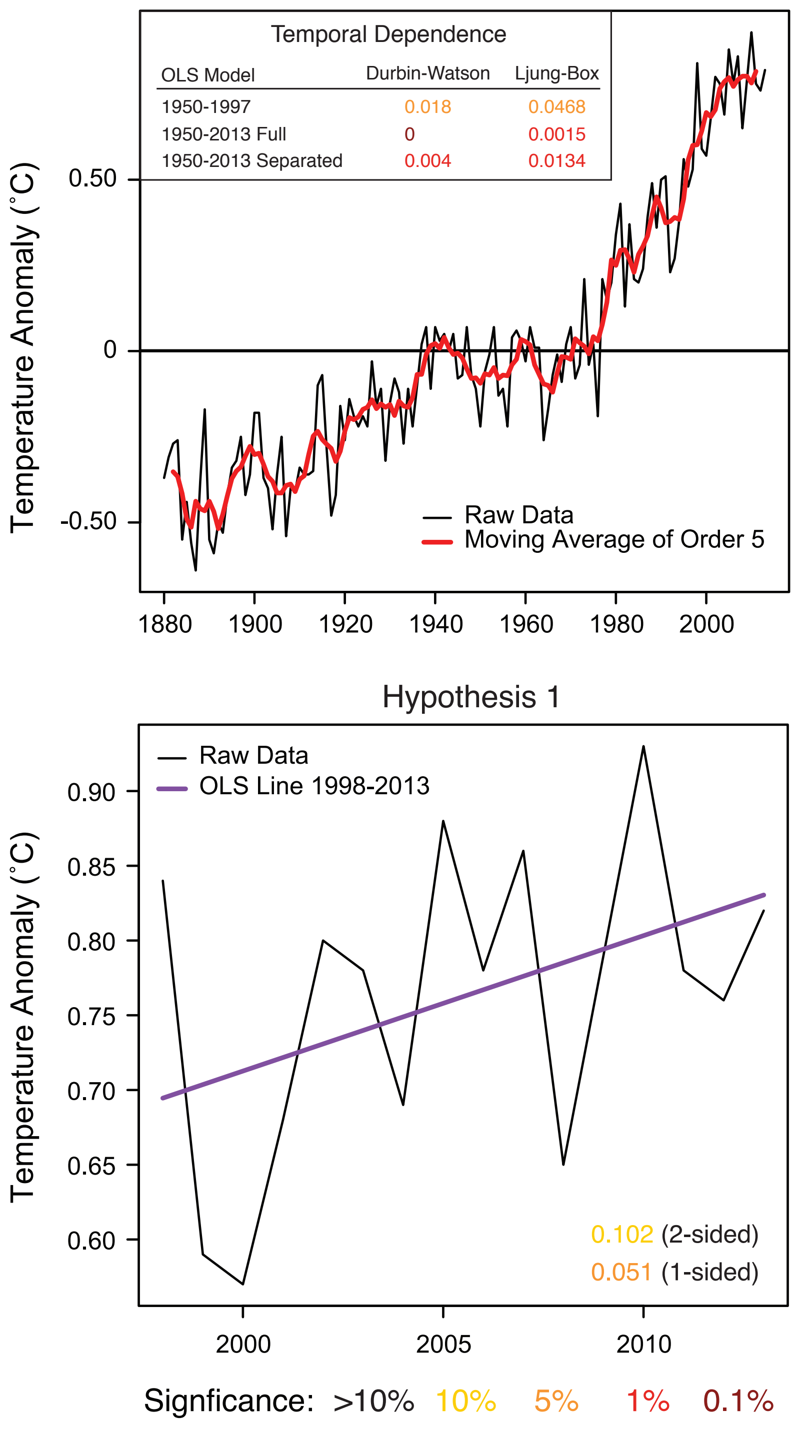


Figure 1: Top panel: Global mean land-ocean temperature index from 1880 to 2013, with base period 1951-1980 and moving average superimposed. The table provides Durbin-Watson and Ljung-Box p-values for the residuals from three OLS fits between 1950-2013. The Ljung-Box test here considers residual autocorrelation in the first 20 lags. The 1950-2013 Full OLS model fits a single regression line to all observations from 1950 to 2013. The 1950-2013 Separated model fits a separate regression line to the 1950-1997 and 1998-2013 periods. Bottom panel: Plot of the global mean land-ocean temperature index, from 1998 to 2013, with the ordinary least squares regression line superimposed.

2 Methods

Datasets: The datasets of global surface temperature anomalies used in our analysis come from three sources: the NASA Goddard Institute for Space Studies (GISS) Surface Temperature Analysis (GISTEMP) Data, the NOAA National Climatic Data Center (NCDC) data, and the HadCRUT4 data, produced from the Met Office Hadley Centre in collaboration with the University of East Anglia Climatic Research Unit (CRU). Each source combines monthly land and sea surface temperature measurements into spatial grids that are then averaged into a single global temperature series. Temperature anomalies are computed from a baseline period, which differs by dataset. The differences in the three datasets largely come from the adjustment/infilling methods for sparse temporal/spatial coverage [Hansen et al., 2010, Morice et al., 2012]. (See the Supplemental Section for more details.). Note that given the global mean temperature data that is available, the main goal of our analysis is to understand the possible mischaracterizing of hiatus claims as compared to understanding the source of observational errors of the temperature process.

Temporal Dependence and Uncertainty Quantification: The global temperature record exhibits temporal correlation. Standard statistical methods tend to ignore this important feature, which in turn can lead to incorrect statistical modeling assumptions and incorrect statistical significance which can in turn lead to erroneous scientific conclusions. For the purposes of uncertainty quantification when testing each of the four statistical hypotheses, we either model the temporal dependence in the global temperature time series explicitly through a parametric autoregressive model, or account for it through the nonparametric circular block bootstrap, stationary block bootstrap or subsampling. (See the Supplemental Section for more details.)

Statistical Hypothesis Testing: The various scientific assertions regarding the climate hiatus are collected into four groups and then formulated as four testable statistical hypotheses. These four hypotheses are specified rigorously, in a principled statistical framework, and are given in Supplemental Sections 3.1, 3.2, 3.3 and 3.4. The Wald test is used to test slope parameters in the linear regression model in Hypotheses I and II leading to Normal or t-distribution based p-values. Moreover, p-values based on the bootstrap and subsampling are also calculated as alternatives to the Wald test whenever appropriate. When comparing two distributions, the Kolmogorov-Smirnov test is used, together with the bootstrap or subsampling, to account for temporal dependence. (See the Supplemental Section for more details.)

Observational Uncertainties: It is important to recognize that the temperature data that is used in our analysis are estimates of an unobserved process and is thus subject to observational errors and the implied uncertainties. Observational uncertainties could arise due to various factors, including instrumental error, changes in the observing network configuration and observing technology, and also due to uncertainties in adjustments made to the data. The HadCRUT4 dataset allows an analysis that incorporates observational uncertainties. The single time-series used for the analysis of the HadCRUT4 data is actually derived from multiple time series which are constructed in order to reflect observational uncertainties. This analysis is provided in Supplemental Section 4 .

3 Results

3.1 Hypothesis I: Hiatus in temperature trend during 1998-2013

A basic assertion regarding the hiatus is that the steady increase in global surface temperature around a linear positive trend has stopped, or "paused" [Guemas et al., 2013]. This sentiment is reflected in statements that "Despite a sustained production of anthropogenic greenhouse gases, the Earth's mean near-surface temperature paused its rise during the 2000-2010 period" [Guemas et al., 2013], and that "climate skeptics have seized on the temperature *trends* as evidence that global warming has ground to a halt" [Tollefson, 2014]. These scientific claims can be turned into a precise statistical null hypothesis: the slope in the regression line of global temperature on time is zero during the hiatus period.

We use three methods with increasing levels of generality to test the above hypothesis. Specific details of the methodology are provided in Supplementary Section 3.1. First, beginning with the 1998-2013 period we fit a standard regression to the response variable global temperature on time during 1998-2013, with errors assumed to be independently and identically distributed (see Figure 1 for the fit). A two-sided hypothesis test yields a p-value of 0.102 (a one-sided test yields a p-value of 0.051). Thus, the claim of a zero warming trend during the hiatus period cannot be rejected at the 5% significance level. The second method fits a linear regression with autocorrelated errors that follow a parametric autoregressive model with lag 1. This model aims to directly address the year-to-year temporal dependency present in the global temperature record. A p-value of 0.075 is obtained for the regression slope coefficient by the bootstrap method (with one-sided p-value less than 5%). Taking temporal dependence into account, there is now more evidence against the null hypothesis of a climate hiatus. The third method is completely nonparametric, and instead of using the parametric AR(1) approach to model the temporal dependency, a block bootstrap is used which allows for quite general forms of temporal dependence, and yields a two-sided p-value of 0.019. There is now compelling evidence to reject the claim of no warming trend during the 1998-2013 period at the 5% significance level (and even at the 1% level for a one-sided test). Moreover, the p-values corresponding to starting years 1999 and 2000 are 0.005 and 0.017 respectively, yielding even lower p-values - and stronger evidence against a hiatus - than when using a starting year of 1998. The sensitivity analysis highlights the fact that choosing the year 1998 had a priori favored the hiatus claim. Moreover, assuming the hiatus as the null makes it harder to conclude otherwise. Regardless, the assertion of a climate hiatus is nevertheless rejected at the 5% level. We therefore conclude that there is "overwhelming evidence" against the claim that there has been no trend in global surface temperature over the past $\approx$ 15 years.

Note also that, in applying progressively more general statistical techniques, the scientific conclusions have progressively strengthened from "not significant," to "significant at the 10% level," and then to "significant at the 5% level." It is therefore clear that naive statistical approaches can possibly lead to erroneous scientific conclusions. Methods that rely upon a strong modeling assumption of no temporal dependence, or that of a specific form, are less reliable than methods that capture dependence without assuming structural knowledge of the type of dependence.

3.2 Hypothesis II: Difference in temperature trends

Otto et al. [2013] state that: "the rate of mean global warming has been lower over the past decade than previously." This statement encompasses a second interpretation of the purported hiatus: that the hiatus represents a "slowdown" of global warming [Chen and Tung, 2014], in which the rate of warming is less during the hiatus compared with the warming prior to the hiatus [Chen and Tung, 2014, Otto et al., 2013, Smith, 2013]. This claim can be formulated as a testable statistical hypothesis, where the null hypothesis is that the regression slope before the hiatus period minus the regression slope during the hiatus period is zero or negative, versus the alternative hypothesis that this difference is positive.

We employ three different methods with increasing levels of statistical sophistication to test this hypothesis. Specific details of the methodology are provided in Supplementary Section 3.2. First, a standard regression of global temperature on time is fitted to both the 1998-2013 hiatus period and the period 1950-1997, with errors assumed to be independently and identically distributed (see Figure 2 top left panel). The first method yields a p-value of 0.210. Thus, there is no evidence of a difference in warming trends even at the 10% significance level. The second method accounts for the temporal dependency in the global temperature record by using a block bootstrap approach, yielding a p-value of 0.323. The evidence for a difference in trends is further weakened when temporal dependency is accounted for. The third approach uses the method of subsampling [Politis et al., 1999] to determine how the current 16-year trend during 1998-2013 compares against all the previous 16-year trends observed between 1950 and 1997. A p-value of 0.3939 is obtained and evidence for the hiatus is further weakened. From the plots in Figure 2 (bottom panel), observe that during the 1950-1997 period, there are several 16-year periods with both higher and lower linear trends. Therefore the observed trend during 1998-2013 does not appear to be anomalous in a historical context.

See Figure 2 (top right panel) for a summary of results of hypothesis II. Varying the cut-off year from 1998 to either 1999 or 2000 yields p-values of 0.214 and 0.348, respectively, for the bootstrap method. Even after properly accounting for temporal dependence, and undertaking a sensitivity analysis, there is no compelling evidence to suggest that the slopes are significantly different. We therefore conclude that the rate of warming over the past ≈ 15 years is not appreciably different from the rate of warming prior to the recent period.

3.3 Hypothesis III: Hiatus in the mean global temperature

Some claims have simply asserted that the annual *mean* global temperature has remained constant since 1998 (versus slowing of the *trend* in global warming). For example, Kosaka and Xie [2013] state that "Despite the continued increase in atmospheric greenhouse gas concentrations, the annual-mean global temperature has not risen in the twenty-first century", while Tollefson [2014] states that "Average global temperatures hit a record high in 1998 – and then the *warming stalled.*" This claim can also be precisely formulated as a testable statistical hypothesis. The statistical model can be written as $x_t = \mu_t + \varepsilon_t$, where t denotes time (in years), x_t is the 1998-2013 global mean temperature anomalies series, μ_t is the mean parameter and ε_t is the random noise component(with $\mathbb{E}(\varepsilon_t) = 0, \mathbb{V}\mathrm{ar}(\varepsilon_t) = \sigma^2$). The corresponding null hypothesis and alternative are given as $H_0 : \mathbb{E}(x_{1998}) = \mathbb{E}(x_{1998+t})$ for $t = 1, 2, \cdots, 15$ versus $H_A : \mathbb{E}(x_{1998}) \neq \mathbb{E}(x_{1998+t})$.

Specific details of the methodology are provided in Supplementary Section 3.3. Hypothesis III is tested in four different ways. There are two options for determining the value of $\mathbb{E}[x_{1998}] = \mu_{1998}$: to directly use the observed 1998 temperature record x_{1998} as a substitute for μ_{1998}, or to alternatively estimate μ_{1998} from the regression line from the period 1950-1997. Figure 3 (top panel) illustrates this concept. As the two approaches for specifying μ_{1998} yield fixed values, the inherent variability therein can be explicitly accounted for by using the bootstrap. Doing so propagates the variability in a rigorous manner. The table in Figure 3 (bottom panel) summarizes the results of testing hypothesis III.

For Method A, when x_{1998} is used as a substitute for μ_{1998}, the statistical test concludes that the mean has decreased during the hiatus, and thus strongly favors the hiatus claim. However, since this one single observed value is not a consistent estimate of μ_{1998}, the conclusion is not reliable. In Method B when μ_{1998} is estimated from the 1950-1997 regression line, the null hypothesis is rejected in the opposite direction, suggesting that the mean temperature has actually increased during the hiatus period. Thus, the selection effect from choosing 1998 as the reference cut-off year has a tremendous impact on the statistical conclusion. Method C, which specifically incorporates the variability inherent in estimating μ_{1998} as x_{1998} leads to a different conclusion than in Method A. In particular, as soon as the variability in estimating μ_{1998} to be x_{1998} is incorporated, one can

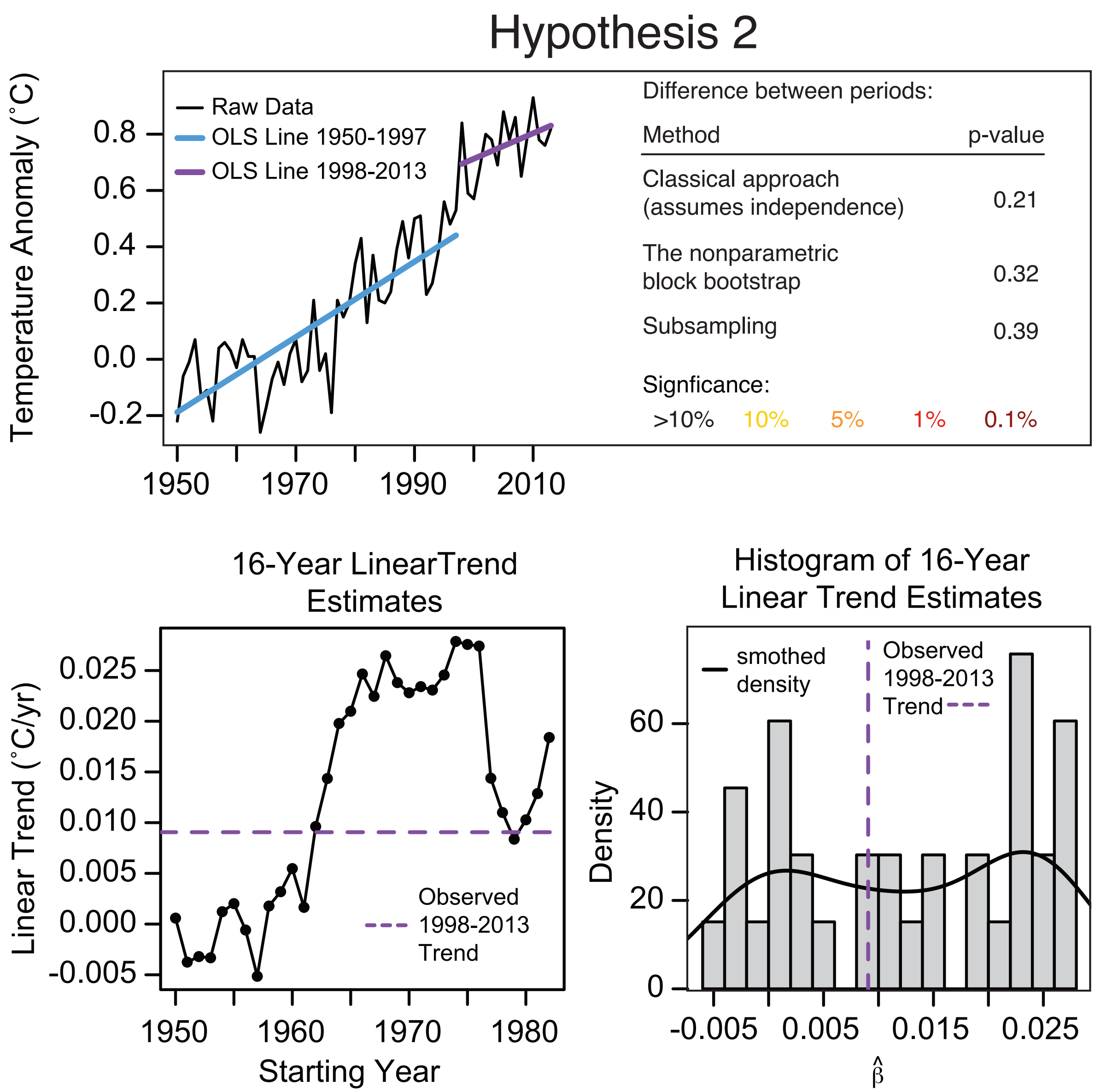
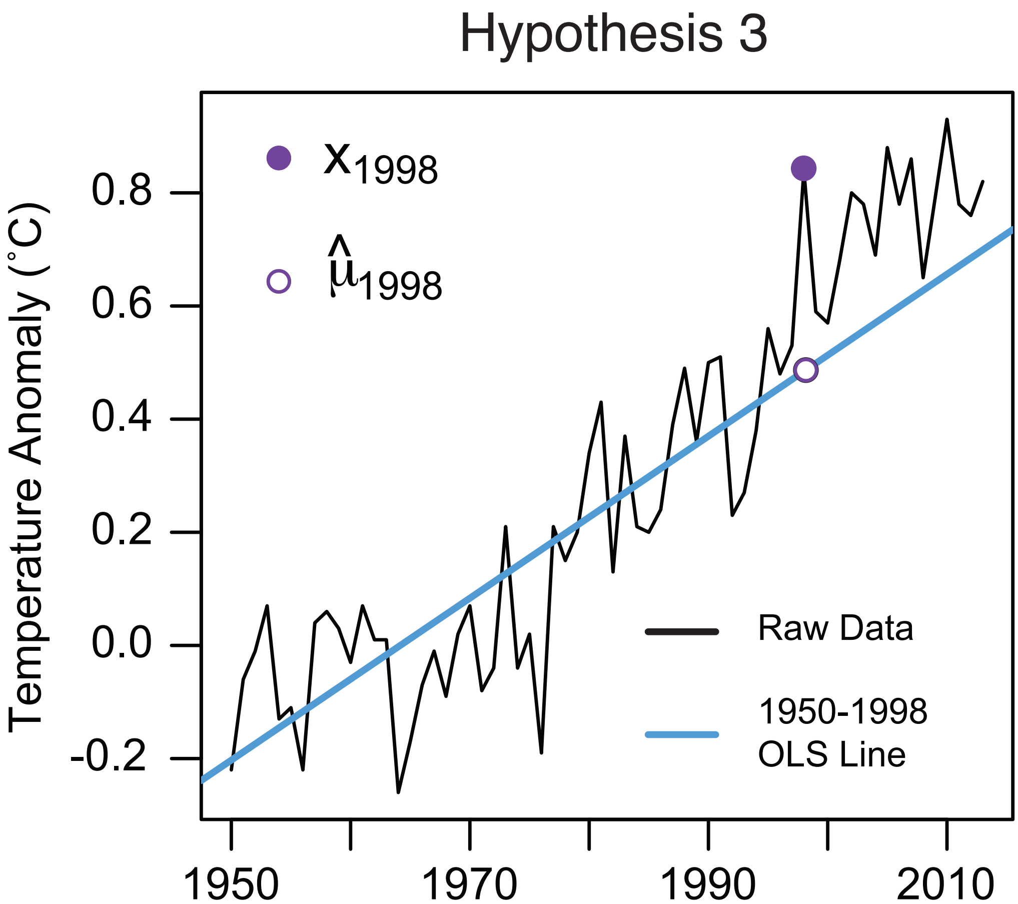


Figure 2: Top panel (left): Plot of the global mean land-ocean temperature index, from 1950 to 2013, with the base period of 1951-1980. The regression fits for the two time periods (1950-1997 and 1998-2013) are superimposed. Top panel (right): Summary table of results for Hypothesis II Bottom panel (left): Time series plot of 16-year observed trends. Bottom panel (right): Histogram of 16-year observed trends.



Hypothesis 3 with 1998 as start of "hiatus" period:

	$E(x_{1998})$	Variability of x_{1998}	Result	Remark
A	x_{1998}	Assume fixed	Reject H_0	decrease in mean
B	$\hat{\mu}_{1998}$	Assume fixed	Reject H_0	increase in mean
C	x_{1998}	Simulate by bootstrap	Retain H_0	no change in mean
D	$\hat{\mu}_{1998}$	Simulate by bootstrap	Reject H_0	increase in mean

Figure 3: Top panel: Figure illustrating how the mean μ_{1998} can be estimated. Bottom panel: Summary table of results for Hypothesis III with 1998 as start of hiatus period.

no longer reject the null hypothesis that the mean has remained constant - even when the high value x_{1998} is used. Method D uses a value for μ_{1998} which is estimated from the 1950-1997 regression and also incorporates the variability of this estimate. Here the assertion that the mean is either zero, or has decreased, is rejected.

Given the results of this nuanced analysis, we conclude that claims that the global mean temperature has not changed in recent decades are not supported by evidence. In addition, our nuanced analysis gives much needed rigor to the claim that using 1998 as a reference year amounts to "cherry picking" [Leber, 2014, Stover, 2014], see also Supplemental Section for detailed discussions). The results are further validated when the analysis is repeated with 1999 and 2000 as the starts of the hiatus period (see Supplemental Section 3.3). Note furthermore that since 2014 was the warmest year on record, the results of our analysis can be viewed as being even more conservative, similar to using 1998 as the starting point.

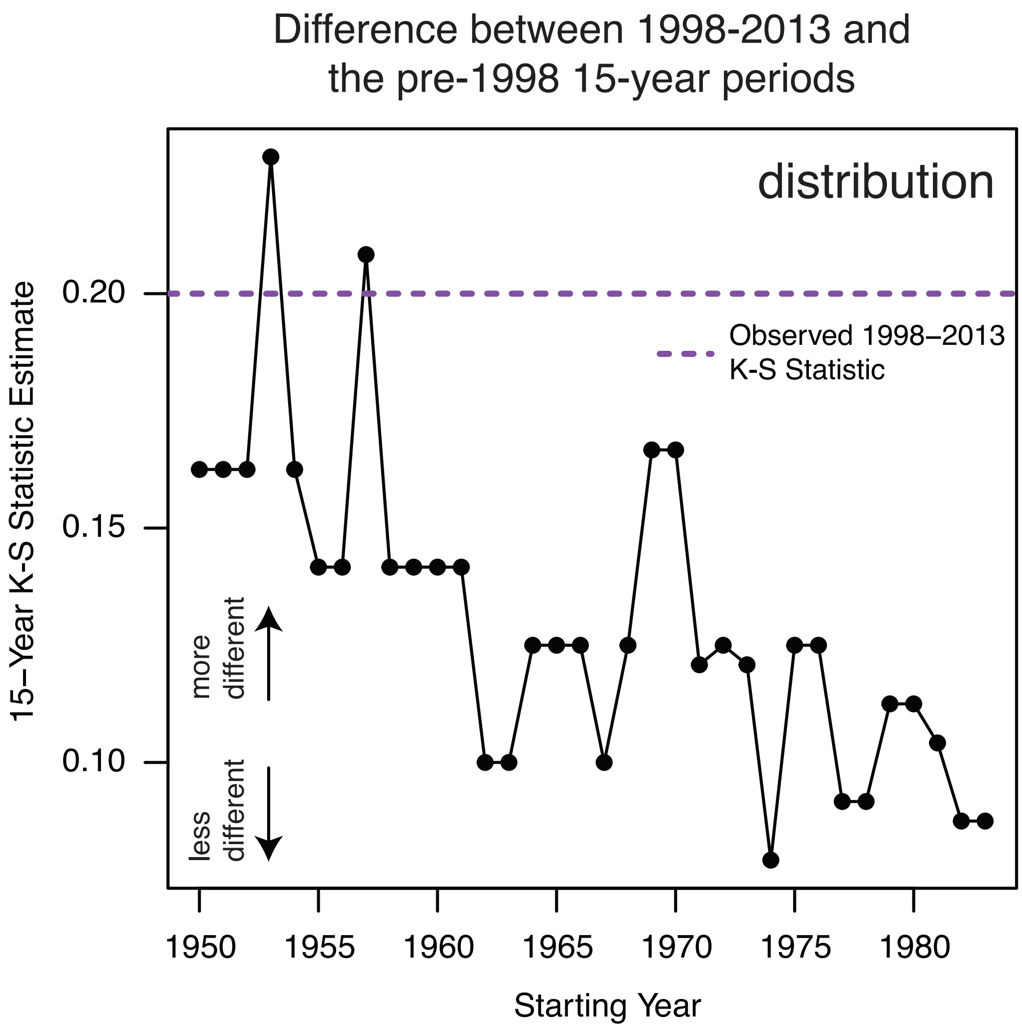
3.4 Hypothesis IV: Difference in year-to-year temperature changes

It is also instructive to extend the analysis above without relying on a linear model to understand trends or means. One such approach is to assess whether the distribution of year-to-year temperature *changes* is markedly different between the hiatus period and the prior periods. Such analysis is inherently less reliant on a statistical model of temperature on time, and hence makes fewer assumptions. The scientific assertion here is that year-to-year changes in global mean temperature during 1998-2013 are different from those during 1950-1997. Under the null hypothesis, these year-to-year changes are assumed to come from a common underlying distribution, though we do not assume that the observations of differences are independent. This framework also allows for testing of specific features of the distribution, including changes in the mean, median and variance. The empirical distribution of annual changes in the global temperature can be constructed by taking first differences: the global mean temperature during a given year is subtracted from the global mean temperature in the previous year. The first differences during 1998-2013 give rise to a 15-year times series of temperature changes. Differences in distribution (using the Kolmogorov-Smirnov (K-S) statistic), in means, medians and variances are tested using the block bootstrap and subsampling, thus taking temporal dependency fully into account. Specific details of the methodology are provided in Supplementary Section 3.4.

The results of this analysis are given in Figure 4. Using either bootstrap or subsampling there is no evidence at the 5% significance level to suggest that the distribution of changes during the hiatus period is different from the previous period 1950-1997. The same applies to the mean and variance of the distributions. The difference in medians is not statistically significant at the 5% level using the block bootstrap approach, but is significant when using subsampling. However this difference in medians completely disappears when the starting year of the hiatus is changed to either 1999 or 2000, hence the result is not robust (see Table 8). Given these results, we conclude that the distribution of annual changes in global temperature has not been different in the past 15 years than earlier in the global temperature record.

3.5 Re-analyzing recently-updated global temperature observations

We have also implemented our methodology on the recently released ERSSTv4 dataset to compare our results to the results obtained in a recent paper by Karl et al. [2015]. Unlike the study by Karl et al. [2015], we do not directly or indirectly impose Gaussianity on the temperature data (in the most general approach that we propose for each hypothesis). We also do not impose an autoregressive structure for modeling the temporal dependence. Instead we account for the



Hypothesis 4 using bootstrap and subsampling:

Test	Bootstrap p-value	Subsampling p-value
Difference in distribution	0.295	0.059
Difference in mean	0.362	0.118
Difference in median	0.058	< 0.029
Difference in variance	0.496	0.265
Difference in log variance	0.483	——

Signficance: >10% 10% 5% 1% 0.1%

Figure 4: Top panel: Time series plot of 15-year observed KS differences. Bottom panel: Summary table of results for Hypothesis IV using bootstrap and subsampling.

temporal dependency more flexibly and non-parametrically using the circular block bootstrap and related methods. The increased sophistication allows one to have more confidence in the results' general validity as our approach makes fewer assumptions. The end result is also compelling. First, the results in Karl et al. [2015] show a positive slope during the hiatus period (Hypothesis I) only at the 10% significance level. Our analysis shows however that removing the arbitrary and parametric autoregressive structure on the residuals and using the block bootstrap yields significance at the 0.1% level. The p-value stemming from our approach is less than 0.0005. The implication of the much stronger conclusion is that the warming trend observed during 1998-2014 [1] arising from a model of no warming is less than 1 in 2000 (as compared to less than 1 in 20 from Karl et al. (2015)). Thus the conclusion is made stronger by a factor of 100 using the methodology we have developed.

Now consider hypothesis II which compares the warming trend during the hiatus period to that in the previous period (1950-1997). Karl et al. [2015] asserts that the analysis on the corrected NOAA global temperature shows that the 90% confidence interval for the trend in the hiatus period encompasses that of the previous period. Note that this confidence interval is based on the period 1998-2012 and is thus calculated on only 15 years of data. Since the theoretical justification of such confidence intervals is valid for large sample sizes, it is not clear how reliable the conclusion really is. On the other hand, our subsampling methodology for comparing the trends in the two periods is applicable even when the sample size in the hiatus period is small. In particular, the validity of the subsampling approach here does not rely on asymptotic arguments (i.e., increasing sample sizes) during the hiatus period. Details of the analysis are given in Tables 11 and 12 in Supplementary Section 6.

Recall that the analysis by Karl et al. [2015] requires the use of the corrected NOAA dataset to reject the claim of a hiatus. We note that our analysis rejects the hiatus claim even when using the older NOAA temperature dataset (that is even without correcting for the data biases). The use of methodology with far fewer restrictive assumptions appears to be more robust to errors in the data. This may not be unexpected since biases in the data tend to violate basic parametric assumptions, whereas the less restrictive techniques, such as the ones we develop, can handle a variety of data generating mechanisms simply by their very non-parametric nature. Note that, by and large, the conclusions reached by Karl et al. [2015] and our conclusions agree. However, it is important to mention that an approach based on stringent or unrealistic assumptions which agrees with our conclusions for this dataset may fail to do so on another dataset.

3.6 Summary

We summarize the overall results from all four hypothesis tests I, II, III and IV in Tables 5 and 6 in Supplementary Section 4. These two tables also analyze the sensitivity of the results to two important factors: first when the cut-off year is changed from 1998 to either 1999 or 2000; and second when the NOAA or HadCRUT4 datasets are used instead of the NASA-GISS dataset. As there are four hypotheses being tested, using a battery of rigorous test procedures, the number of hypothesis being tested are numerous. Hence the issue of multiple hypothesis testing surfaces. In particular, a certain number of these hypotheses are expected to be falsely rejected by chance alone, casting further doubt on any of the hiatus claims.

Our rigorous statistical framework yields strong evidence against the presence of a global warming hiatus. Accounting for temporal dependence and selection effects rejects - with overwhelming evidence - the hypothesis that there has been no trend in global surface temperature over the past

[1] Though we consider the hiatus period as 1998-2013 elsewhere in the analysis, we consider the hiatus period as 1998-2014 here in order to compare directly with Karl et al. [2015]

≈15 years. This analysis also highlights the potential for improper statistical assumptions to yield improper scientific conclusions. Our statistical framework also clearly rejects the hypothesis that the trend in global surface temperature has been smaller over the recent ≈ 15 year period than over the prior period. Further, our framework also rejects the hypothesis that there has been no change in global mean surface temperature over the recent ≈15 years, and the hypothesis that the distribution of annual changes in global surface temperature has been different in the past ≈15 years than earlier in the record. Taken together, these results clearly reject the presence of a hiatus, pause, or slowdown in global warming. In rejecting all four hiatus hypotheses, our results instead demonstrate that the evolution of global surface temperature over the past 1-2 decades is not abnormal or unexpected within the context of the long-term record of variability and change.

Without empirical evidence in support of the hiatus claims, the assumption that there has been a hiatus/pause/slow-down in global warming should be called into question. That being said, recent work investigating the geophysical causes of the recent temperature time series have provided valuable insights into the processes that create decadal-scale variability in global temperature within a long-term trend of global warming. Moreover, it is also useful that errors in data aggregation have been corrected in the recent work of Karl et al. [2015].

Funding Acknowledgments: B.R. was partially funded by the US National Science Foundation under grants DMS-CMG 1025465, AGS-1003823, DMS-1106642, DMS-CAREER-1352656 and the US Air Force Office of Scientific Research grant award FA9550-13-1-0043. J.R. was partially funded by the US National Science Foundation under grants DMS-1307973 and DMS-1007732.

References

Maxwell T. Boykoff. Media discourse on the climate slowdown. *Nature Climate Change*, 4:156–158, 2014. doi:10.1038/nclimate2156.

Xianyao Chen and Ka-Kit Tung. Varying planetary heat sink led to global-warming slowdown and acceleration. *Science*, 345(6199):897–903, 2014. DOI: 10.1126/science.1254937.

Donald Cochrane and Guy Orcutt. Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245):32–61, 1949.

K. Cowtan, Z. Hausfather, E. Hawkins, P. Jacobs, M. E. Mann, S. K. Miller, B. A. Steinman, M. B. Stolpe, and R. G. Way. Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophysical Research Letter (in print)*, 42, 2015. doi:10.1002/2015GL064888.

Kevin Cowtan and Robert G. Way. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, 2014. doi:10.1002/qj.2297.

Paul J. Durack, Peter J. Gleckler, Felix W. Landerer, and Karl E. Taylor. Quantifying underestimates of long-term upper-ocean warming. *Nature Climate Change*, 4:999–1005, 2014. doi:10.1038/nclimate2389.

Matthew H. England, Shayne McGregor, Paul Spence, Gerald A. Meehl, Axel Timmermann, Wenju Cai, Alex Sen Gupta, Michael J. McPhaden, Ariaan Purich, and Agus Santoso. Recent intensification of wind-driven circulation in the pacific and the ongoing warming hiatus. *Nature Climate Change*, 4:222–227, 2014. doi:10.1038/nclimate2106.

John C. Fyfe, Nathan P. Gillett, and Francis W. Zwiers. Overestimated global warming over the past 20 years. *Nature Climate Change*, 3:767–769, 2013.

Virginie Guemas, Francisco J. Doblas-Reyes, Isabel Andreu-Burillo, and Muhammad Asif. Retrospective prediction of the global warming slowdown in the past decade. *Nature Climate Change*, 2013. doi:10.1038/nclimate1863.

J. Hansen, R. Ruedy, Mki. Sato, and K. Lo. Global surface temperature change. *Reviews of Geophysics*, 48(RG4004), 2010. doi:10.1029/2010RG000345.

Ed Hawkins, Tamsin Edwards, and Doug McNeall. Pause for thought. *Nature Climate Change*, 4: 154–156, 2014. doi:10.1038/nclimate2150.

IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)].* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Thomas R. Karl, Anthony Arguez, Boyin Huang, Jay H. Lawrimore, James R. McMahon, Matthew J. Menne, Thomas C. Peterson, Russell S. Vose, and Huai-Min Zhang. Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, 348(6242):1469–1472, 2015. DOI: 10.1126/science.aaa5632.

Yu Kosaka and Shang-Ping Xie. Recent global-warming hiatus tied to equatorial pacific surface cooling. *Nature*, 2013. doi:10.1038/nature12534.

Rebecca Leber. We're stuck with the global warming "hiatus" myth for years to come. *New Republic*, August 2014. `http://www.newrepublic.com/article/119212/study-atlantic-ocean-role-global-warming-hiatus`.

Gerald A. Meehl, Julie M. Arblaster, John T. Fasullo, Aixue Hu, and Kevin E. Trenberth. Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods. *Nature Climate Change*, 1:360–364, 2011. doi:10.1038/nclimate1229.

Colin P. Morice, John J. Kennedy, Nick A. Rayner, and Phil D. Jones. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research*, 117:D08101, 2012. doi:10.1029/2011JD017187.

Alexander Otto, Friederike E. L. Otto, Olivier Boucher, John Church, Gabi Hegerl, Piers M. Forster, Nathan P. Gillett, Jonathan Gregory, Gregory C. Johnson, Reto Knutti, Nicholas Lewis, Ulrike Lohmann, Jochem Marotzke, Gunnar Myhre, Drew Shindell, Bjorn Stevens, and Myles R. Allen. Energy budget constraints on climate response. *Nature Geoscience*, 6:415–416, 2013. doi:10.1038/ngeo1836.

Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling.* Springer-Verlag, New York, 1999.

B. Rajaratnam, J. Romano, and M. Tsiang. Fixed b subsampling with application to testing for a change. *Technical Report, Department of Statistics, Stanford University*, 2014.

Benjamin D. Santer, Céline Bonfils, Jeffrey F. Painter, Mark D. Zelinka, Carl Mears, Susan Solomon, Gavin A. Schmidt, John C. Fyfe, Jason N. S. Cole, Larissa Nazarenko, Karl E. Taylor, and Frank. J. Wentz. Volcanic contribution to decadal changes in tropospheric temperature. *Nature Geoscience*, 2014. doi:10.1038/ngeo2098.

Doug Smith. Oceanography: Has global warming stalled? *Nature Climate Change*, 3:618–619, 2013.

Dawn Stover. The global warming "hiatus". *Bulletin of the Atomic Scientists*, September 2014. `http://thebulletin.org/global-warming-%E2%80%9Chiatus%E2%80%9D7639`.

Jeff Tollefson. The case of the missing heat. *Nature*, 505:276–278, 2014.

Kevin E. Trenberth and John T. Fasullo. An apparent hiatus in global warming? *Earth's Future*, 1:19–32, 2013. doi:10.1002/2013EF000165.

1 Supplementary Section: Summary Tables

Table 1: Significance table comparing the results across the three data sets with hiatus start year of 1998/1999/2000. The p-values provided in the table are determined as follows: First the p-values are observed as the block size increases and until these p-values stabilize. The highest of the stabilized p-values is chosen in order to ensure Type I error control.

	NASA	NOAA	HadCRUT4
Hypothesis I: Bootstrap $H_0 : \beta_{\text{post}} = 0$ vs. $H_A : \beta_{\text{post}} \neq 0$	1998 | 1999 | 2000 0.019 | 0.005 | 0.017 Does not vary with cutoff year	1998 | 1999 | 2000 0.172 | 0.090 | 0.244 Does not vary with cutoff year	1998 | 1999 | 2000 0.194 | 0.124 | 0.332 Does not vary with cutoff year
Hypothesis II: Bootstrap $H_0 : \beta_{\text{pre}} = \beta_{\text{post}}$ vs. $H_A : \beta_{\text{pre}} \neq \beta_{\text{post}}$	1998 | 1999 | 2000 0.323 | 0.214 | 0.348 Does not vary with cutoff year	1998 | 1999 | 2000 0.237 | 0.188 | 0.332 Does not vary with cutoff year	1998 | 1999 | 2000 0.393 | 0.284 | 0.469 Does not vary with cutoff year
Hypothesis III $H_0 : \mathbb{E}(x_{\text{cutoff}}) = \mathbb{E}(x_{\text{cutoff}+t})$ vs. $H_A : \mathbb{E}(\text{cutoff}) \neq \mathbb{E}(x_{\text{cutoff}+t})$			
$\mathbb{E}(x_{\text{cutoff}}) = x_{\text{cutoff}}$ Variance assumed fixed	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year
$\mathbb{E}(x_{\text{cutoff}}) = \widehat{\mu}_{\text{cutoff}}$ Variance assumed fixed	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year
$\mathbb{E}(x_{\text{cutoff}}) = x_{\text{cutoff}}$ Variance simulated by bootstrap	1998 | 1999 | 2000 0.462 | 0.618 | 0.541 Does not vary with cutoff year	1998 | 1999 | 2000 0.423 | 0.623 | 0.571 Does not vary with cutoff year	1998 | 1999 | 2000 0.420 | 0.631 | 0.626 Does not vary with cutoff year
$\mathbb{E}(x_{\text{cutoff}}) = \widehat{\mu}_{\text{cutoff}}$ Variance simulated by bootstrap	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.001 | < 0.001 | < 0.001 Does not vary with cutoff year

Table 2: Significance table comparing the results across the three data sets with hiatus start year of 1998/1999/2000. The p-values provided in the table are determined as follows: First the p-values are observed as the block size increases and until these p-values stabilize. The highest of the stabilized p-values is chosen in order to ensure Type I error control.

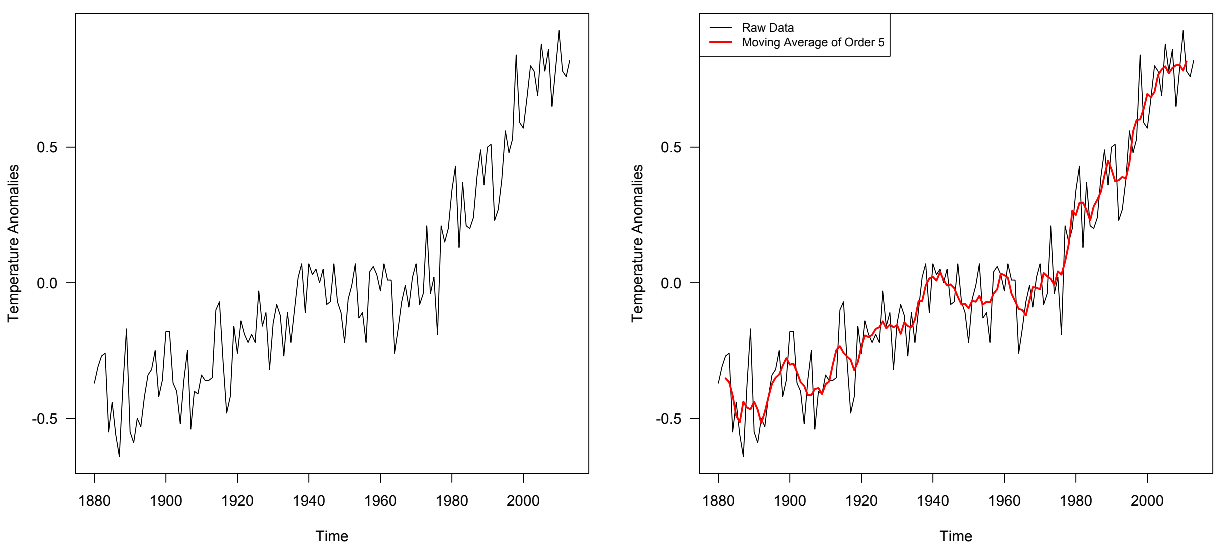
<table>
<tr><th></th><th>NASA</th><th>NOAA</th><th>HadCRUT4</th></tr>
<tr><td colspan="4">Hypothesis IV: Bootstrap</td></tr>
<tr><td>Kolmogorov-Smirnov Test
$H_0 : F_{\Delta X} = F_{\Delta Y}$ vs. $H_A : F_{\Delta X} \neq F_{\Delta Y}$</td><td>1998 | 1999 | 2000
0.295 | 0.662 | 0.632
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.218 | 0.261 | 0.383
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.131 | 0.225 | 0.449
Does not vary with cutoff year</td></tr>
<tr><td>Difference in mean
$H_0 : \overline{\Delta X} = \overline{\Delta Y}$ vs. $H_A : \overline{\Delta X} \neq \overline{\Delta Y}$</td><td>1998 | 1999 | 2000
0.362 | 0.996 | 0.906
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.331 | 0.962 | 0.887
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.385 | 0.874 | 0.798
Does not vary with cutoff year</td></tr>
<tr><td>Difference in median
$H_0 : \mathrm{med}(\Delta X) = \mathrm{med}(\Delta Y)$
vs. $H_A : \mathrm{med}(\Delta X) \neq \mathrm{med}(\Delta Y)$</td><td>1998 | 1999 | 2000
0.058 | 0.434 | 0.515
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.411 | 0.928 | 0.595
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.412 | 0.724 | 0.849
Does not vary with cutoff year</td></tr>
<tr><td>Difference in variance
$H_0 : \mathbb{V}\mathrm{ar}(\Delta X) = \mathbb{V}\mathrm{ar}(\Delta Y)$
vs. $H_A : \mathbb{V}\mathrm{ar}(\Delta X) \neq \mathbb{V}\mathrm{ar}(\Delta Y)$</td><td>1998 | 1999 | 2000
0.568 | 0.251 | 0.328
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.103 | 0.025 | 0.051
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.181 | 0.019 | 0.055
Varies with cutoff year</td></tr>
<tr><td>Difference in log variance
$H_0 : \log \mathbb{V}\mathrm{ar}(\Delta X) = \log \mathbb{V}\mathrm{ar}(\Delta Y)$
vs. $H_A : \log \mathbb{V}\mathrm{ar}(\Delta X) \neq \log \mathbb{V}\mathrm{ar}(\Delta Y)$</td><td>1998 | 1999 | 2000
0.483 | 0.175 | 0.311
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.067 | 0.004 | 0.024
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.112 | 0.003 | 0.026
Varies with cutoff year</td></tr>
<tr><td colspan="4">Hypothesis IV: Subsampling</td></tr>
<tr><td>Kolmogorov-Smirnov Test</td><td>1998 | 1999 | 2000
0.059 | 0.250 | 0.237
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.029 | 0.083 | 0.132
Varies with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td></tr>
<tr><td>Difference in mean</td><td>1998 | 1999 | 2000
0.118 | 0.583 | 0.579
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.147 | 0.556 | 0.605
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.206 | 0.694 | 0.632
Does not vary with cutoff year</td></tr>
<tr><td>Difference in median</td><td>1998 | 1999 | 2000
< 0.029 | 0.194 | 0.974
Varies with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | 0.583 | 0.816
Varies with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td></tr>
<tr><td>Difference in variance</td><td>1998 | 1999 | 2000
0.265 | 0.056 | 0.132
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td></tr>
</table>

2 Supplementary Section: Datasets and temporal dependence

2.1 Datasets used in analysis

The NASA GISTEMP dataset uses the 1951-1980 average as the baseline period and estimates anomalies up to 1200 km from the nearest measurement station, allowing for broad spatial coverage. The NOAA data reconstructs land data for unobserved regions using a method called "empirical orthogonal teleconnections." The HadCRUT4 data does not use any spatial infilling and thus has gaps in grid squares with very sparse (or no) data. The HadCRUT4 data therefore does not account for warming in the Arctic and Antarctic regions, leading to documented coverage bias [Cowtan and Way, 2014].

We primarily present in the main text the results from analysis of the NASA GISS data set, as it provides the largest spatial coverage of the three datasets. The NASA GISS data is also plotted in Figure 5. However, the NOAA and HadCRU datasets are also thoroughly analyzed to ensure that our results are not biased by any particular dataset. See Summary Tables 1 and 2. It is clear from these summary tables that analyzing a restricted spatial domain can lead to different scientific conclusions.



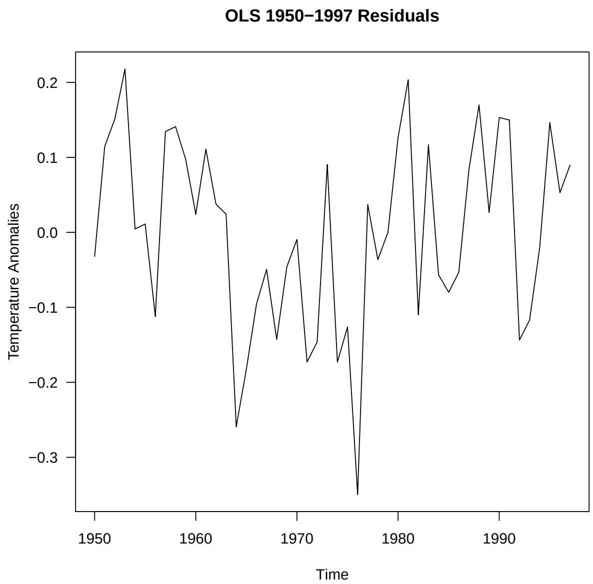
(a) NASA GISS data

(b) NASA GISS data with 5-year moving average

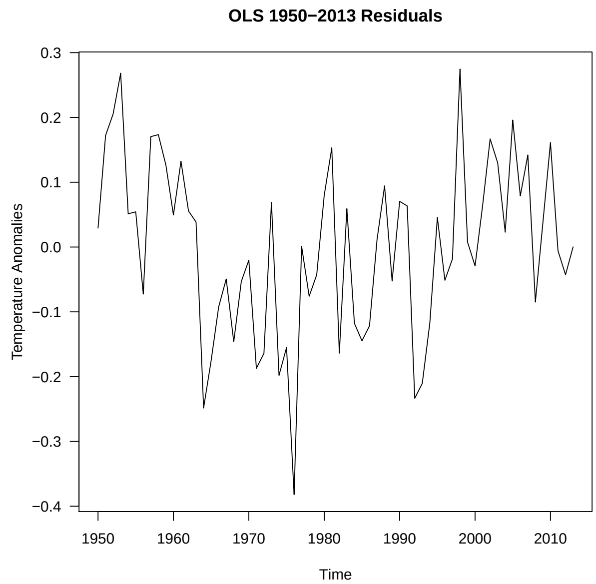
Figure 5: Plots of (a) the global mean land-ocean temperature index, from 1880 to 2013, with the base period 1951-1980 and (b) with a 5-year simple moving average superimposed.

2.2 Serial Dependence in the global temperature record

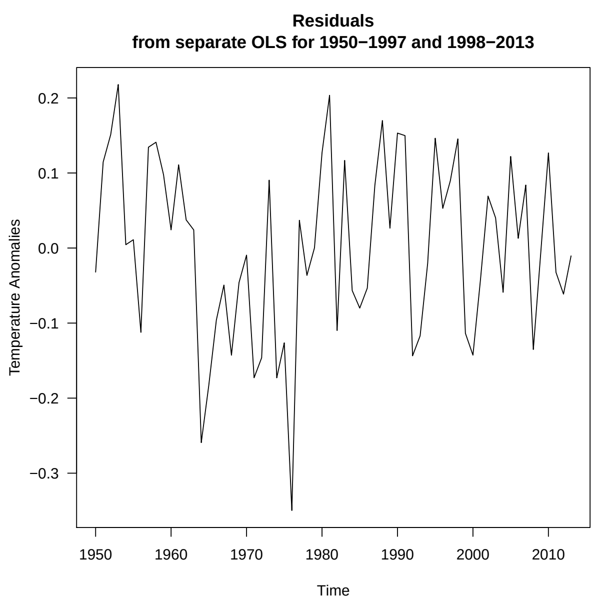
Residual plots from a standard least squares fit and corresponding PACF and ACF plots are given below. These clearly illustrate the presence of serial correlation in the global temperature record, and thus the need to properly account for it.



(a) Plot of the 1950-1997 OLS residuals

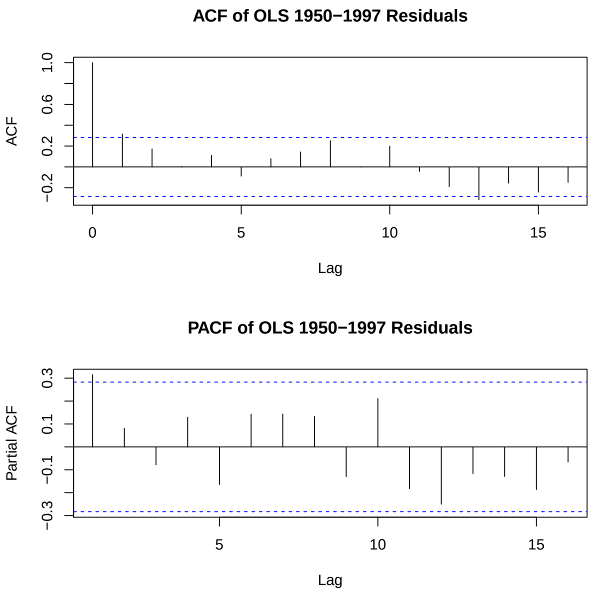


(b) Plot of the 1950-2013 OLS residuals

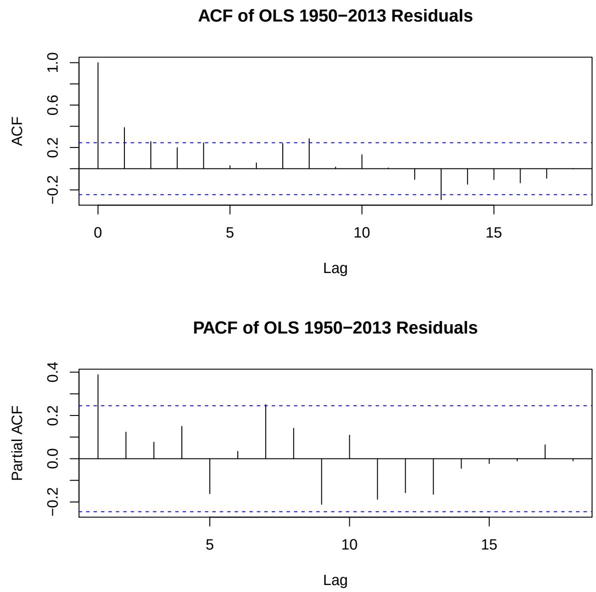


(c) Plot of the residuals from separate 1950-1997 and 1998-2013 OLS fits

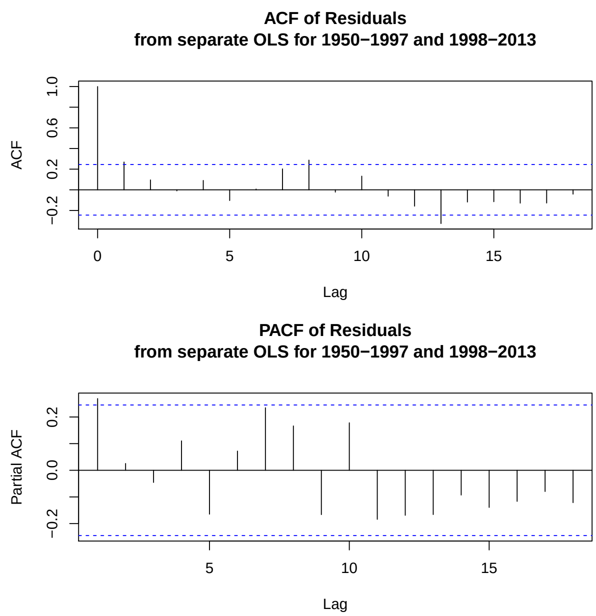
Figure 6: Plots of the residuals from 1950-2013.



(a) ACF and PACF plots for the 1950-1997 OLS residuals



(b) ACF and PACF plots for the 1950-2013 OLS residuals



(c) ACF and PACF plots for the residuals from separate 1950-1997 and 1998-2013 OLS fits

Figure 7: ACF and PACF plots for residuals from 1950-2013.

3 Supplementary Section: Details of Methodology and Additional Results

3.1 Hypothesis I

Consider the model where the global temperature series x_t for the 1998-2013 period follows a linear model, given by

$$x_t = \alpha_1 + \beta_1 t + \varepsilon_t,$$

where $\mathbb{E}(\varepsilon_t) = 0$ and $\mathbb{V}\mathrm{ar}(\varepsilon_t) = \sigma^2$. The claim that the linear rate of change in global temperature has stalled can be restated as saying there is no linear trend in global temperature during the period 1998-2013. The corresponding statistical hypothesis can be stated as

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0.$$

Three methods with increasing levels of generality and sophistication are employed in order to test Hypothesis I: Three methods (with increasing levels of generality/sophistication) are used to test this hypothesis:

- Method IA: No temporal dependence
- Method IB: Temporal dependence: using an AR(1) model
- Method IC: Temporal dependence: using the bootstrap only

3.1.1 Method IA: No temporal dependence

Under the assumption of independently and identically distributed errors, ordinary least squares is used to estimate the slope β_1, and is given as $\widehat{\beta}_1 = 0.0090$ with the standard error $\mathrm{se}(\widehat{\beta}_1) = 0.0052$. The Wald statistic is constructed as

$$W = \frac{\widehat{\beta}_1 - 0}{\mathrm{se}(\widehat{\beta}_1)} = 1.7510.$$

Under the null hypothesis that $\beta_1 = 0$, W approximately follows a t_{n-2} distribution, where n is the number of observations. We compute the p-value to be

$$p = P\left[|t_{n-2}| > |W|\right] = 2F(-|W|) = 0.1018,$$

where F is the cdf of $t_{n-2} = t_{14}$.

It is important to recognize that the observed temperature time series are potentially subject to errors due instrumental errors and other reasons. A more sophisticated formulation of the standard regression model could also be formulated. A key assumption that has been made in our analysis in this regard is that the observational errors can be absorbed into the residuals of the regression model.

3.1.2 Method IB: Temporal dependence: Autoregressive structure in the residuals

Assume that the global temperature series x_t for the 1998-2013 period follows a linear model, given by

$$x_t = \alpha_1 + \beta_1 t + \varepsilon_t,$$

where ε_t follows an AR(1) model, namely

$$\varepsilon_t = \phi\varepsilon_{t-1} + \delta_t,$$

where δ_t are *iid* innovations with $\mathbb{E}(\delta_t) = 0$ and $\mathbb{V}\text{ar}(\delta_t) = \sigma^2$.

The $\widehat{\beta}_1$ now denotes the estimate of β_1 using the iterative Cochrane-Orcutt procedure [Cochrane and Orcutt, 1949]. A semiparametric block bootstrap is implemented in order to approximate $\mathbb{V}\text{ar}(\widehat{\beta}_1)$. The algorithm is given below:

1. Fit the model $\widehat{x}_t = \widehat{\alpha}_1 + \widehat{\beta}_1 t + \widehat{\phi}\varepsilon_{t-1}$ using the iterative Cochrane-Orcutt procedure and compute the sample innovations $\widehat{\delta}_t = x_t - \widehat{x}_t$.

2. Use the circular block bootstrap with block size b to generate a bootstrap series of innovations δ_t^* of length equal to the original data series.

3. Construct bootstrap observations $x_t^* = \widehat{x}_t + \delta_t^*$, on which we rerun the regression analysis to yield a bootstrap replication $\widehat{\beta}_1^*$.

4. To approximate the sampling distribution of $\widehat{\beta}_1$, repeat Steps 2 and 3, B times, to get $\widehat{\beta}_{1,1}^*, \ldots, \widehat{\beta}_{1,B}^*$.

Approximate two-sided p-values are calculated in three ways. First, we compute the bootstrap estimate of $\mathbb{V}\text{ar}(\widehat{\beta}_1)$ by

$$\widehat{\mathbb{V}\text{ar}}_b(\widehat{\beta}_1) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\widehat{\beta}_{1,j}^* - \frac{1}{B}\sum_{k=1}^{B}\widehat{\beta}_{1,k}^*\right)^2$$

and construct the Wald statistic

$$W = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{\mathbb{V}\text{ar}}_b(\widehat{\beta}_1)}}.$$

Under the null hypothesis that $\beta_1 = 0$, W approximately has a t_{n-3} distribution, where n is the number of observations. We thus compute the p-value by

$$\hat{p} \approx P[|t_{n-3}| > |W|] = 2F(-|W|),$$

where F is the cdf of t_{n-3}.

Asymptotically, W converges in distribution to $\mathcal{N}(0,1)$, so we can also approximate the p-value by

$$\hat{p} \approx P[|Z| > |W|] = 2\Phi(-|W|),$$

where $Z \sim \mathcal{N}(0,1)$ and Φ is the CDF of Z.

We also compute bootstrap p-values by computing

$$\hat{p} = \frac{1}{B}\sum_{k=1}^{B} I\left(\left|\widehat{\beta}_{1,k}^* - \widehat{\beta}_1\right| > \left|\widehat{\beta}_1\right|\right).$$

We report the bootstrap standard errors and p-values below for various block sizes b and $B = 1000$.

The p-values become significant (at the 10% level) for block sizes of $b = 3$ or larger. Accounting for the temporal dependence in the data, there is sufficient evidence at the 10% significance level to reject the hiatus claim.

Table 3: Bootstrap standard errors for $\widehat{\beta}$ at various bootstrap block sizes and the corresponding p-values computed using the t_{13} and $\mathcal{N}(0,1)$ distributions and the bootstrap approximation for the 1998-2013 time period when assuming an AR(1) model in the residuals.

Block Size	Std.Error	t_{13}	$\mathcal{N}(0,1)$	Bootstrap
$b=1$	0.0111	0.2704	0.2497	0.241
$b=2$	0.0083	0.1463	0.1223	0.116
$b=3$	0.0069	0.0853	0.0626	0.071
$b=4$	0.0046	0.0161	0.0057	0.005
$b=5$	0.0067	0.0778	0.0555	0.075
$b=6$	0.0037	0.0045	0.0006	0.000
No AR(1)/Bootstrap	0.0052	0.1018	0.0799	

A counterintuitive result emerges from the analysis above since we reject the null hypothesis when accounting for the influence of temporal dependence (using a simple AR(1) dependence model), but cannot reject the null hypothesis when assuming independence. This is unexpected given the greater uncertainty in the slope estimates given the weak persistence in the global mean temperature. In order to understand this issue better, the ACF and PACF plots of the residuals for the period 1998-2013 were calculated (see Figure 8). It is clear from the PACF plot that there is non-negligible negative autocorrelation in the 1998-2013 residual time series. This negative autocorrelation explains the apparent contradiction. It is important to note that the counterintuitive PACF estimate could be due to sampling variability. The PACF plot for the 1999-2013 residual time series, that is without the 1998 temperature data point, reveals interesting points. The negative lag 1 autocorrelation and partial autocorrelation in the 1998-2013 series is no longer present when year 1998 is removed from the analysis, underscoring the effect of this one time point on the entire analysis. Furthermore, non-negligible positive partial autocorrelation starts to emerge when year 1998 is removed.

3.1.3 Method IC: Temporal dependence: The nonparametric block bootstrap

A very general method to assess the uncertainty in the estimates of β_1 is to use the block bootstrap. Let $\widehat{\beta}_1$ denote the ordinary least squares estimate of β_1. To approximate $\mathbb{V}\text{ar}(\widehat{\beta}_1)$, consider the following algorithm:

1. Fit the model $\widehat{x}_t = \widehat{\alpha}_1 + \widehat{\beta}_1 t$ using ordinary least squares and compute the sample residuals $\widehat{\varepsilon}_t = x_t - \widehat{x}_t$.

2. Use the circular block bootstrap with block size b to generate a bootstrap series of residuals ε_t^* of length equal to the original data series.

3. Construct bootstrap observations $x_t^* = \widehat{\alpha}_1 + \widehat{\beta}_1 t + \varepsilon_t^*$ on which the regression analysis is repeated to yield a bootstrap replication $\widehat{\beta}_1^*$.

4. To approximate the sampling distribution of $\widehat{\beta}_1$, repeat Steps 2 and 3, B times, to get $\widehat{\beta}_{1,1}^*, \ldots, \widehat{\beta}_{1,B}^*$.

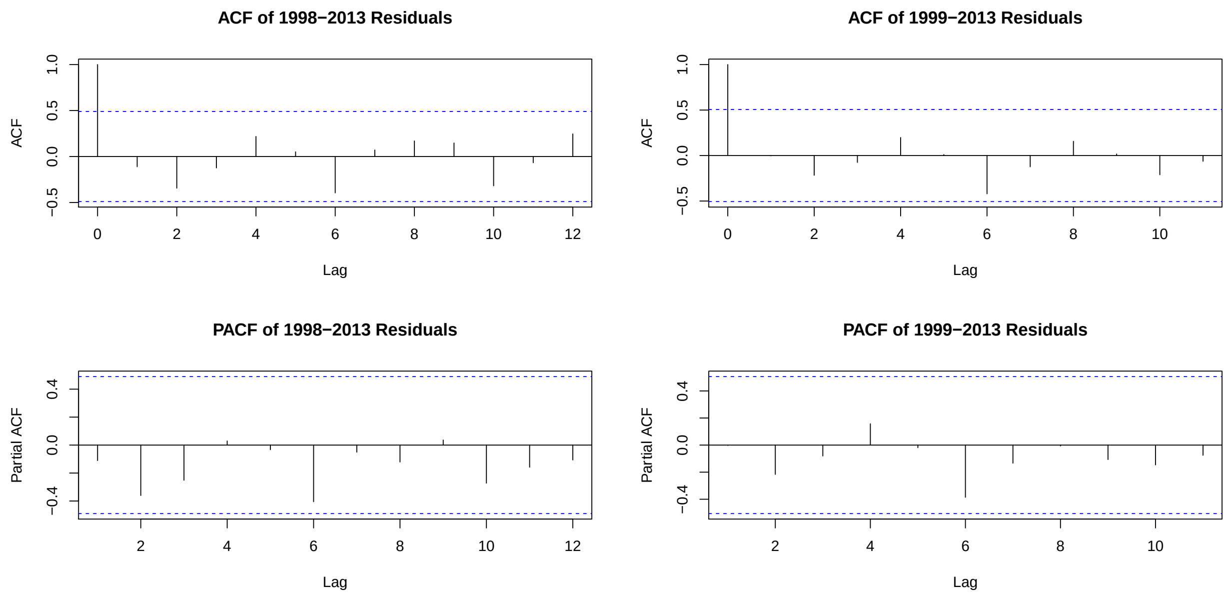


Figure 8: Top: ACF plots of residuals time series 1998-2013 (left) and 1999-2013 (right). Bottom: PACF plots of residuals time series 1998-2013 (left) and 1999-2013 (right).

As in Method IB, approximate p-values are calculated in three ways. First, the bootstrap estimate of $\mathbb{V}\text{ar}(\widehat{\beta}_1)$ is computed by

$$\widehat{\mathbb{V}\text{ar}}_b(\widehat{\beta}_1) = \frac{1}{B-1}\sum_{j=1}^{B}\left(\widehat{\beta}^*_{1,j} - \frac{1}{B}\sum_{k=1}^{B}\widehat{\beta}^*_{1,k}\right)^2$$

and construct the Wald statistic

$$W = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{\mathbb{V}\text{ar}}_b(\widehat{\beta}_1)}}.$$

Under the null hypothesis that $\beta_1 = 0$, W approximately has a t_{n-2} distribution, where n is the number of observations. The corresponding p-value is computed by

$$\hat{p} = P[|t_{n-2}| > |W|] = 2F(-|W|),$$

where F is the CDF of t_{n-2}.

Asymptotically, W converges in distribution to $\mathcal{N}(0,1)$, so one can also approximate the p-value by

$$\hat{p} = P[|Z| > |W|] = 2\Phi(-|W|),$$

where $Z \sim \mathcal{N}(0,1)$ and Φ is the CDF of Z.

The bootstrap p-values can be computed by evaluating

$$\hat{p} = \frac{1}{B}\sum_{k=1}^{B} I\left(\left|\widehat{\beta}^*_{1,k} - \widehat{\beta}_1\right| > \left|\widehat{\beta}_1\right|\right).$$

The bootstrap standard errors and p-values for various block sizes b and $B = 1000$ are reported in Table 4. For the 1998-2013 period, $n = 16$, so $W \sim t_{16-2} = t_{14}$. Since n is small, the p-values computed using the t_{14} distribution are more reliable.

Table 4: Bootstrap standard errors for $\widehat{\beta}$ at various bootstrap block sizes and the corresponding p-values computed using the t_{14} and $\mathcal{N}(0,1)$ distributions and the bootstrap approximation for the 1998-2013 time period.

Block Size	Std.Error	t_{14}	$\mathcal{N}(0,1)$	Bootstrap
$b=1$	0.0048	0.0818	0.0607	0.046
$b=2$	0.0044	0.0610	0.0416	0.05
$b=3$	0.0037	0.0288	0.0149	0.019
$b=4$	0.0032	0.0145	0.0053	0.001
$b=5$	0.0035	0.0224	0.0102	0.006
$b=6$	0.0032	0.0127	0.0043	0.001
No Bootstrap	0.0052	0.1018	0.0799	

From Table 4 the p-values are both significant (at the 5% level) and stable from block size $b=3$ and larger.

3.2 Hypothesis II

The second hypothesis test is set up in the context of two linear models, and is given by

$$\begin{aligned} x_t &= \alpha_0 + \beta_0 t + \varepsilon_t \\ y_s &= \alpha_1 + \beta_1 s + \varepsilon_s, \end{aligned}$$

where x_t and y_s are the 1950-1997 and 1998-2013 global mean temperature anomalies series respectively, and ε_t is random noise, so $\mathbb{E}(\varepsilon_t) = 0, \mathbb{V}\mathrm{ar}(\varepsilon_t) = \sigma^2$. The claim is that the linear trend during the 1998-2013 hiatus period is lower than the trend during the previous period 1950-1997 [2]. The corresponding statistical hypothesis is then given as

$$H_0 : \beta_0 - \beta_1 \leq 0 \quad \text{versus} \quad H_A : \beta_0 - \beta_1 > 0.$$

Three methods (with increasing levels of generality/sophistication) are used to test this hypothesis:

- Method IIA: No temporal dependence
- Method IIB: Temporal dependence: using the nonparametric block bootstrap
- Method IIC: Temporal dependence: using subsampling

3.2.1 Method IIA: No temporal dependence

First, temporal dependence in the observations is ignored and errors are assumed to be independent. The hypothesis test is based on the standard Wald statistic

$$W = \frac{\widehat{\beta}_0 - \widehat{\beta}_1}{\sqrt{\frac{\widehat{\sigma}_0^2}{\sum_{j=1}^{n_0} t_j^2} + \frac{\widehat{\sigma}_1^2}{\sum_{k=1}^{n_1} s_k^2}}},$$

[2]Changing the reference period from 1950-1997 to 1880-1997 only strengthens the null hypothesis of no difference between the hiatus period and before. This follows from the fact that the trend during 1880-1997 is more similar to the trend in the hiatus period. Thus the selected period 1950-1997 can be regarded as a lower bound on p-values for tests of difference in slopes.

where $\widehat{\beta}_0, \widehat{\beta}_1$ are the respective ordinary least squares estimates for β_0 and β_1, $\widehat{\sigma}_0^2, \widehat{\sigma}_1^2$ are the estimates for the residual variances, and t_j and s_k denote standardized time units within each time interval.

The estimates obtained are $\widehat{\beta}_0 = 0.0134$, $\widehat{\beta}_1 = 0.0090$, yielding the Wald statistic $W = 0.8063$. Assuming independent observations, the distribution of W can be approximated by $\mathcal{N}(0,1)$. The one-sided p-value is given by

$$p = P[Z > W] = P[Z > 0.8063] = 0.2100,$$

where $Z \sim \mathcal{N}(0,1)$. The observed difference in slopes is not statistically significant at the 5% significance level. Hence there is no compelling evidence to suggest that the slopes are significantly different.

3.2.2 Method IIB: Temporal dependence: The nonparametric block bootstrap

Method IIB tests the hypotheses while accounting for the temporal dependence in the observations. Specifically, the block bootstrap regression method is employed in order to approximate $\mathbb{V}\mathrm{ar}(\widehat{\beta}_0 - \widehat{\beta}_1)$. The implementation of the block bootstrap is described below.

1. Fit the models $\widehat{x}_t = \widehat{\alpha}_0 + \widehat{\beta}_0 t$ and $\widehat{y}_s = \widehat{\alpha}_1 + \widehat{\beta}_1 s$ using ordinary least squares and compute the sample residuals series

$$\widehat{\varepsilon}_t = \begin{cases} x_t - \widehat{x}_t & \text{if } 1950 \leq t \leq 1997 \\ y_t - \widehat{y}_t & \text{if } 1998 \leq t \leq 2013. \end{cases}$$

2. The circular block bootstrap is used with block size b to generate a bootstrap series of residuals ε_t^* of length equal to the original data series.

3. The bootstrap observations are constructed as follows:

$$\begin{aligned} x_t^* &= \widehat{\alpha}_0 + \widehat{\beta}_0 t + \varepsilon_t^*, \quad && \text{if } 1950 \leq t \leq 1997 \\ y_t^* &= \widehat{\alpha}_1 + \widehat{\beta}_1 t + \varepsilon_t^*, \quad && \text{if } 1998 \leq t \leq 2013 \end{aligned}$$

   on which the regression analysis is rerun to yield bootstrap replications $\widehat{\beta}_0^*$ and $\widehat{\beta}_1^*$.

4. To approximate the sampling distribution of $\widehat{\beta}_0 - \widehat{\beta}_1$, Steps 2 and 3 are repeated, B times, to get $\widehat{\beta}_{0,1}^*, \ldots, \widehat{\beta}_{0,B}^*$, $\widehat{\beta}_{1,1}^*, \ldots, \widehat{\beta}_{1,B}^*$ and compute $\widehat{\beta}_{0,1}^* - \widehat{\beta}_{1,1}^*, \ldots, \widehat{\beta}_{0,B}^* - \widehat{\beta}_{1,B}^*$.

The approximate p-values are calculated in two ways. First, the bootstrap estimate of $\mathbb{V}\mathrm{ar}(\widehat{\beta}_0 - \widehat{\beta}_1)$ is computed by

$$\widehat{\mathbb{V}\mathrm{ar}}_b(\widehat{\beta}_0 - \widehat{\beta}_1) = \frac{1}{B-1} \sum_{j=1}^{B} \left[\left(\widehat{\beta}_{0,j}^* - \widehat{\beta}_{1,j}^* \right) - \frac{1}{B} \sum_{k=1}^{B} \left(\widehat{\beta}_{0,k}^* - \widehat{\beta}_k^* \right) \right]^2$$

as an ingredient in the Wald statistic

$$W = \frac{(\widehat{\beta}_0 - \widehat{\beta}_1) - 0}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}_b(\widehat{\beta}_0 - \widehat{\beta}_1)}}.$$

Under the null hypothesis that $\beta_0 - \beta_1 = 0$, W converges in distribution to $\mathcal{N}(0,1)$, so the one-sided p-value is approximated by

$$\hat{p} = P[Z > W] = 1 - \Phi(W),$$

where $Z \sim \mathcal{N}(0,1)$ and Φ is the CDF of Z.

The one-sided bootstrap p-value is obtained by computing

$$\hat{p} = \frac{1}{B} \sum_{k=1}^{B} I\left[\left(\widehat{\beta}^*_{0,k} - \widehat{\beta}^*_{1,k}\right) - \left(\widehat{\beta}_0 - \widehat{\beta}_1\right) > \left(\widehat{\beta}_0 - \widehat{\beta}_1\right)\right].$$

The bootstrap standard errors and p-values are reported for various block sizes b and $B = 1000$ in Table 5. Even after taking temporal dependence into account, the observed difference is not statistically significant. The p-values are fairly stable, since they do not vary much by block size. Note that after accounting for temporal dependence, the p-values change from 0.2 to around 0.3.

Table 5: Bootstrap standard errors for $\widehat{\beta}_0 - \widehat{\beta}_1$ at various bootstrap block sizes and the corresponding one-sided p-values computing using the $\mathcal{N}(0,1)$ distribution and the bootstrap approximation to test for a difference in slopes between the global temperatures in the 1950-1997 and 1998-2013 periods.

Block Size	Std.Error	$\mathcal{N}(0,1)$	Bootstrap
$b = 1$	0.0075	0.2838	0.257
$b = 2$	0.0078	0.2908	0.297
$b = 3$	0.0086	0.3085	0.306
$b = 4$	0.0085	0.3054	0.323
$b = 5$	0.0085	0.3068	0.299
$b = 6$	0.0086	0.3079	0.323
No bootstrap	0.0053	0.2100	

3.2.3 Method IIC: Temporal dependence using subsampling

The third method employs the technique of subsampling [Politis et al., 1999] as a means to quantify the uncertainty around the difference in the two observed regression slopes. Here the 16-year regression slope obtained when fitting a regression analysis on data from 1998-2013 is compared against all contiguous 16-year trends during the period 1950-1997. Note that this distribution does not overlap with the 1998-2013 period. A p-value is computed based on the quantile of the distribution of 16-year trends, that corresponds to the observed (1998-2013) trend. This approach yields a valid statistical method that approximates the null distribution of the test statistic $\widehat{\beta}_1$, and falls within the overall framework of subsampling - see Rajaratnam et al. [2014] for theoretical details.

The observed trend during the hiatus period, $\widehat{\beta}_1 = 0.0091$, yields a p-value of 0.3939. As in the previous two methods, the p-value is not significant at the nominal 5% level. Having said this, from Figure 2 there is a clear pattern in the distribution of 16 year linear trends over time: all 16 year trends starting at 1950 all the way to 1961 are lower than the trend during hiatus period, and all 16 year linear trends starting at years 1962 all the way to 1982 are higher than the trend during the hiatus period, with the exception of the 1979-1994 trend.

3.3 Hypothesis III

As mentioned in the main text of the paper, Hypothesis III is tested in four ways.

1. **Method IIIA: $\mathbb{E}(x_{1998}) = x_{1998}$ and variability of x_{1998} is not modeled.** First $\mathbb{E}(x_{1998})$ is estimated by the observed value $x_{1998} = 0.84$. This value is assumed to be fixed as the variability associated with this estimate is not modeled. The stationary and circular block bootstraps are used to sample from the 1999-2013 series to generate a sampling distribution for $\widehat{\mu}^*_{\text{after}} - x_{1998}$. It turns out that the entire bootstrap sampling distribution, regardless of block bootstrap method and block size, is negative. Hence one can reject the null hypothesis and conclude that the post-1998 mean is statistically significantly different from the 1998 mean.

2. **Method IIIB: $\mathbb{E}(x_{1998}) = \widehat{\mu}_{1998}$ and variability of $\widehat{\mu}_{1998}$ is not modeled.** The value for $\mathbb{E}(x_{1998})$ is estimated by the fitted value $\widehat{\mu}_{1998} = \widehat{\alpha} + \widehat{\beta}(1998) = 0.4844$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are estimated using ordinary least squares on the 1950-1998 series. This estimated value is once more assumed to be fixed, as the variability associated with this estimate is not modeled.Note that using the observed 1998 as a substitute for the true underlying mean μ_{1998} can be viewed as "cherry picking" a reference year which favors the hiatus claim. Thus estimating μ_{1998} from the regression line from the period 1950-1997 provides a statistically rigorous way to avoid this pitfall. The stationary and circular block bootstraps are used to sample from the 1999-2013 series and generate a sampling distribution for $\widehat{\mu}^*_{\text{after}} - \widehat{\mu}_{1998}$. In this case, the entire bootstrap sampling distribution, regardless of block bootstrap method and block size, is positive. That is, the sign of the difference is reversed. One can thus reject the null hypothesis and conclude that the post-1998 mean is statistically significantly different from the 1998 mean.

3. **Method IIIC: $\mathbb{E}(x_{1998}) = x_{1998}$ and variability of x_{1998} is modeled.** The value for $\mathbb{E}(x_{1998})$ is estimated by the observed value $x_{1998} = 0.84$. We now explicitly model both the variability of the $\mathbb{E}(x_{1998})$ estimate and the $\mathbb{E}(x_{1998+t})$ estimate of $\widehat{\mu}_{\text{after}} = \frac{1}{15}\sum_{t=1}^{15} x_{1998+t} = 0.7573$ by using the circular block bootstrap. The 1950-1998 and 1999-2013 series are sampled separately. For each bootstrap series, we again estimate $\mathbb{E}(x_{1998})$ by the 1998 observation x^*_{1998} in the bootstrap series. We estimate $\mathbb{E}(x_{1998+t})$ by $\widehat{\mu}^*_{\text{after}} = \frac{1}{15}\sum_{t=1}^{15} x^*_{1998+t}$. The bootstrap sampling distributions of $(\widehat{\mu}^*_{\text{after}} - x^*_{1998}) - (\widehat{\mu}_{\text{after}} - x_{1998})$ for various block sizes are obtained. The observed difference in mean estimates $\widehat{\mu}_{\text{after}} - x_{1998} = -0.0827$ is in the far left tail of the bootstrap sampling distributions, regardless of block size. However when using a two-sided test, unlike in Method IIIA, we retain the null hypothesis and conclude that sufficient evidence is not available to deduce that the post-1998 mean is statistically significantly different from the 1998 mean.

4. **Method IIID: $\mathbb{E}(x_{1998}) = \widehat{\mu}_{1998}$ and variability of $\widehat{\mu}_{1998}$ is modeled.** The value for $\mathbb{E}(x_{1998})$ is estimated by the fitted value $\widehat{\mu}_{1998} = \widehat{\alpha} + \widehat{\beta}(1998) = 0.4844$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are estimated using ordinary least squares on the 1950-1998 series. Both the variability of the $\mathbb{E}(x_{1998})$ estimate and the $\mathbb{E}(x_{1998+t})$ estimate of $\widehat{\mu}_{\text{after}} = 0.7573$ are explicitly modeled by using the circular block bootstrap. For each bootstrap series, we estimate $\mathbb{E}(x_{1998})$ by the fitted value $\widehat{\mu}^*_{1998} = \widehat{\alpha}^* + \widehat{\beta}^*(1998)$, where $\widehat{\alpha}^*$ and $\widehat{\beta}^*$ are estimated using ordinary least squares on the 1950-1998 bootstrap series. We estimate $\mathbb{E}(x_{1998+t})$ by $\widehat{\mu}^*_{\text{after}} = \frac{1}{15}\sum_{t=1}^{15} x^*_{1998+t}$. The bootstrap sampling distributions of $(\widehat{\mu}^*_{\text{after}} - \widehat{\mu}^*_{1998}) - (\widehat{\mu}_{\text{after}} - \widehat{\mu}_{1998})$ for various block sizes are obtained. The observed difference in mean estimates $\widehat{\mu}_{\text{after}} - \widehat{\mu}_{1998} = 0.2729$ is in the far right tail of the bootstrap sampling distributions, regardless of block size. Thus, as in Method

Table 6: Summary table of results for Hypothesis III with 1999 cutoff

Method	$\mathbb{E}(x_{1999})$	Variability of x_{1999}	Result	Remark
IIIA	x_{1999}	Assume fixed	Reject H_0	increase in mean
IIIB	$\widehat{\mu}_{1999}$	Assume fixed	Reject H_0	increase in mean
IIIC	x_{1999}	Simulate by bootstrap	Retain H_0	no change in mean
IIID	$\widehat{\mu}_{1999}$	Simulate by bootstrap	Reject H_0	increase in mean

Table 7: Summary table of results for Hypothesis III with 2000 cutoff

Method	$\mathbb{E}(x_{2000})$	Variability of x_{2000}	Result	Remark
IIIA	x_{2000}	Assume fixed	Reject H_0	increase in mean
IIIB	$\widehat{\mu}_{2000}$	Assume fixed	Reject H_0	increase in mean
IIIC	x_{2000}	Simulate by bootstrap	Retain H_0	no change in mean
IIID	$\widehat{\mu}_{2000}$	Simulate by bootstrap	Reject H_0	increase in mean

IIIB, one can reject the null hypothesis and conclude that the post-1998 mean is statistically significantly different from the 1998 mean.

Effect of varying the start of the hiatus period to 1999 and 2000:

Explanation of the differences between Hypotheses I and III: Recall that in the linear model in hypothesis I, setting the population slope coefficient to zero corresponds to a constant mean global temperature during the hiatus period. In this sense hypothesis I and III coincide. There are however some not-so-subtle differences. First, the class of linear models considered in hypothesis I is a sub-model of the more general model considered in hypothesis III. This difference leads to different test statistics that guard against Type I error in the context of that particular model. Second, the class of alternatives are also different between hypothesis I and hypothesis III. Thus the statistical power, which guards against Type II error, associated with the two tests is also different.

Further discussion of the choice of 1998 as the start of the hiatus period:

The results of the detailed statistical analysis presented above is quite nuanced. The four statistical tests together give compelling evidence to refute the assertion that global mean temperature has stalled during the hiatus period. In fact the increase in global mean temperature appears to continue unabated during the hiatus period. This warming trend appears to be masked by the global mean temperature record for 1998. The analysis also suggests that the observed global mean temperature record for 1998 is extreme in the sense that if the associate variability is accounted for, then ensuing years do not reflect a significant decrease in temperature.

The robustness of the above results are examined by varying the starting year of the hiatus from 1998 to 1999 and 2000. The results from this sensitivity analysis are given in Tables 6 and 7. The decrease in mean suggested by method IIIA when x_{1998} is used as a substitute for μ_{1998} no longer holds true when the cut-off year 1999 or 2000 is used. In fact, the recorded temperature for 1999 is relatively lower than those recorded in the period 2000-2013 and leads to the conclusion that global mean temperatures have actually *increased* during the purported hiatus period. This sensitivity analysis once more underscores our earlier point that a selection effect occurs when picking 1998 as the start of the hiatus period. The result above can also be interpreted against the backdrop of hypothesis I which (essentially) tested for the slope after 1998. In that analysis the slope was found to be positive and significantly different from zero.

There are two additional ways to see how the conclusion in the testing of the global mean temperature is sensitive to the single observed value in year 1998. One approach is to pick the year 2000 as the start of the hiatus period. Just as 1998 is hand-picked, one can also hand-pick 2000. When year 2000 is picked however, the conclusion of a stalling in mean global warming is completely reversed, and the opposite conclusion from the 1998 case is reached. So it is clear from just this experiment that data snooping can strongly influence the final conclusion. A second approach is to look at the largest temperature value before 1998. Note that the 1998 value (anomaly) is 0.84 vs. the previous high of only 0.56 in 1995. A simple analysis shows than any value for 1998 which is lower than 0.81 would not lead to a conclusion that would suggest that the mean has stalled. Hence it is clear that the 1998 value is very high compared to any of the previous values by a large margin. Thus, any hiatus claims after the fact of observing an exceptionally warm year as a means of comparing amounts to cherry-picking.

3.4 Hypothesis IV

Formally, consider the hypotheses

$$H_0 : F_{\Delta X}(\cdot) = F_{\Delta Y}(\cdot) \quad \text{versus} \quad H_A : F_{\Delta X}(\cdot) \neq F_{\Delta Y}(\cdot),$$

where ΔX denotes the year-to-year increase in annual temperatures during 1950-1998 and ΔY denote the year-on-year increase between 1998-2013. This corresponds to a null hypothesis that within a long term period of increases (as witnessed by the general increase between 1950-1998), shorter periods of zero or negative trends (as observed in the period 1998-2013) are not unusual.

The empirical changes in annual temperatures are computed by taking the first differences of the observed global mean annual temperatures series. That is,

$$\begin{aligned} \Delta X_t &= x_t - x_{t-1} \quad \text{for } 1881 \leq t \leq 1998 \\ \Delta Y_s &= y_s - y_{s-1} \quad \text{for } 1999 \leq s \leq 2013. \end{aligned}$$

The following five tests are implemented in the above framework:

- Hypothesis IVA: Test for a difference in distributions
- Hypothesis IVB: Test for a difference in means
- Hypothesis IVC: Test for a difference in medians
- Hypothesis IVD: Test for a difference in variances
- Hypothesis IVE: Test for a difference in log variances

The implementation of hypothesis IVA using the Kolmogorov-Smirnov test in conjunction with the block bootstrap is outlined below. Implementation of hypotheses IVB, C, D, E follow similarly.

Consider now using the Kolmogorov-Smirnov test for hypothesis IVA:

$$H_0 : F_{\Delta X}(\cdot) = F_{\Delta Y}(\cdot) \quad \text{versus} \quad H_A : F_{\Delta X}(\cdot) \neq F_{\Delta Y}(\cdot),$$

where ΔX denotes the change in annual temperatures between 1894-1998 and ΔY denote the change between 1998-2013. The Kolmogorov-Smirnov statistic is given by

$$D = D_{m,n} = \sup_x |F_{\Delta X,m}(x) - F_{\Delta Y,n}(x)|,$$

where $F_{\Delta X,m}$ and $F_{\Delta Y,n}(x)$ are the empirical distribution functions of ΔX and ΔY, respectively.

The block bootstrap is used to approximate the sampling distribution of the usual test statistic $D_{m,n}$. Details of the algorithm are given below:

1. Use the stationary block bootstrap with block sizes drawn from a geometric distribution with probability p of success (i.e., expected block size of $1/p$) to ΔX_t and ΔY_s to generate bootstrap series ΔX_t^* and ΔY_s^* for $1950 \leq t \leq 1998$ and $1999 \leq s \leq 2013$.

2. Compute the bootstrap Kolmogorov-Smirnov statistic

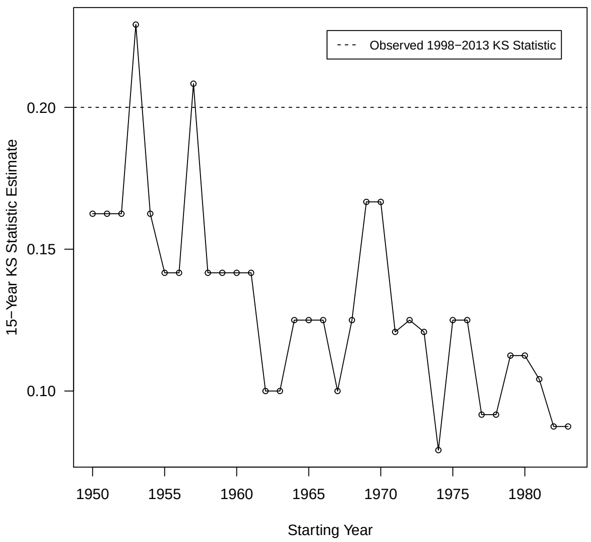
$$D^* = \sup_x |F_{\Delta X^*,m}(x) - F_{\Delta Y^*,n}(x)|.$$

3. To approximate the sampling distribution of D, repeat Steps 1 and 2 above, B times, to get $D_1^*, D_2^*, \ldots, D_B^*$.

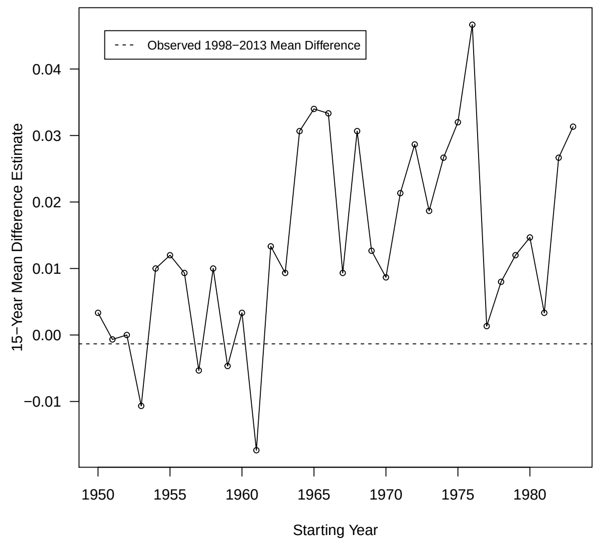
The entire 1950-2013 series is used to generate the bootstrapped series, and the bootstrap p-values are computed by

$$\hat{p} = \frac{1}{B} \sum_{i=1}^{B} I(D^* > D).$$

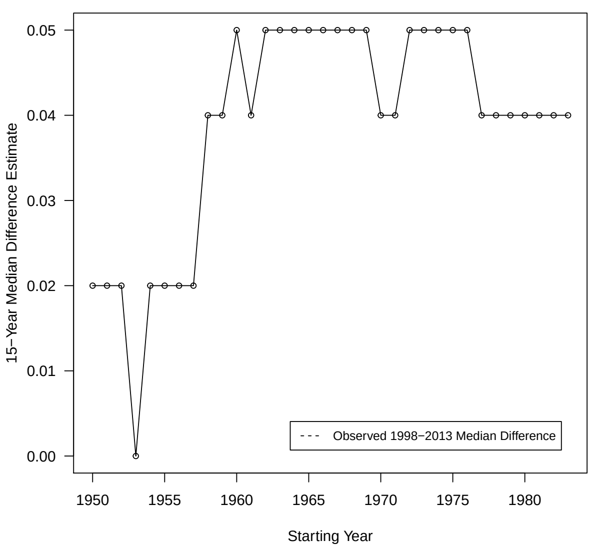
The subsampling results are also illustrated in Figures 9, 10, 11, 12 and 13. These figures illustrate how the Kolmogorov-Smirnov statistic, mean, median and variance of the year-to-year temperature increases recorded during the hiatus compare to the distribution of these quantities during the 1950-1997 period. It is clear that the recorded statistics during the hiatus period are rendered non-significant in the subsampling context because of differences observed further back in the past, and not in the recent past.



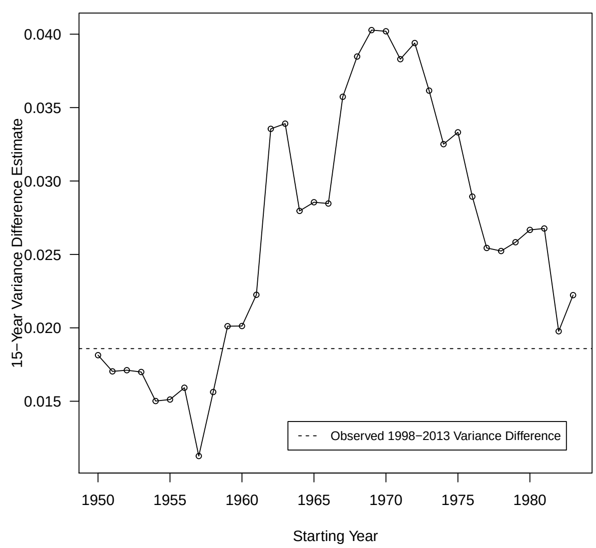
(a) Time series plot of 15-year observed KS differences



(b) Time series plot of 15-year observed mean differences

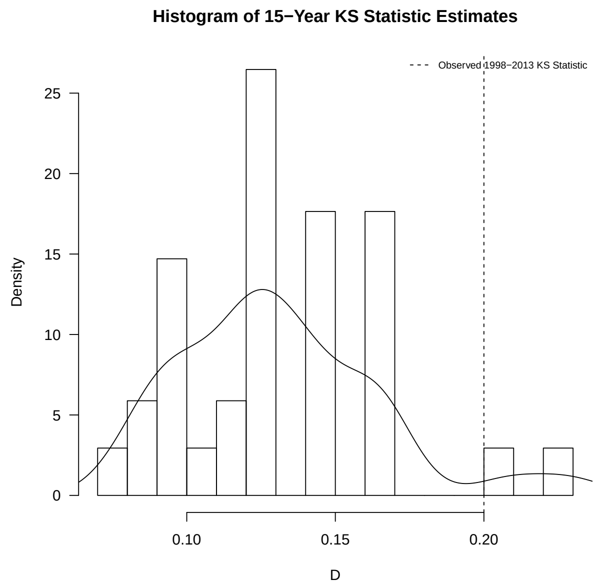


(c) Time series plot of 15-year observed median differences

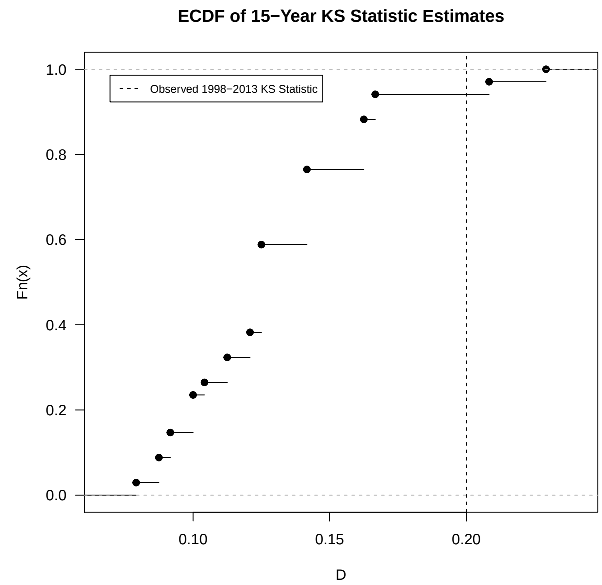


(d) Time series plot of 15-year observed variance in the differences

Figure 9: Time series plots of 15-year difference estimates observed between 1950 to 2013.

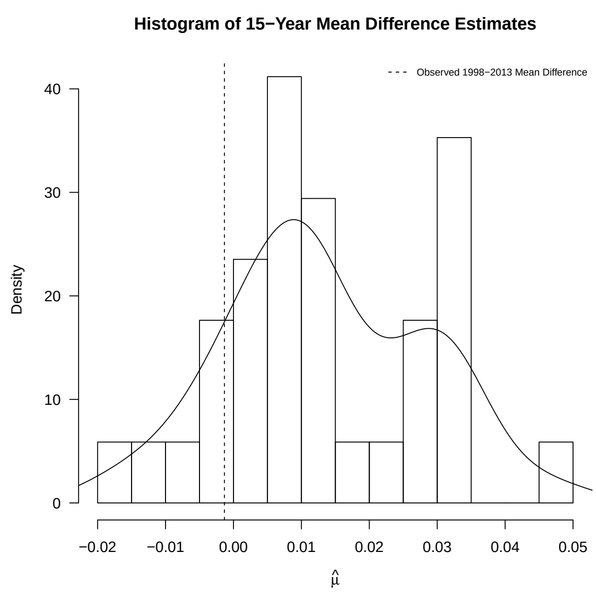


(a) Histogram of 15-year observed KS differences

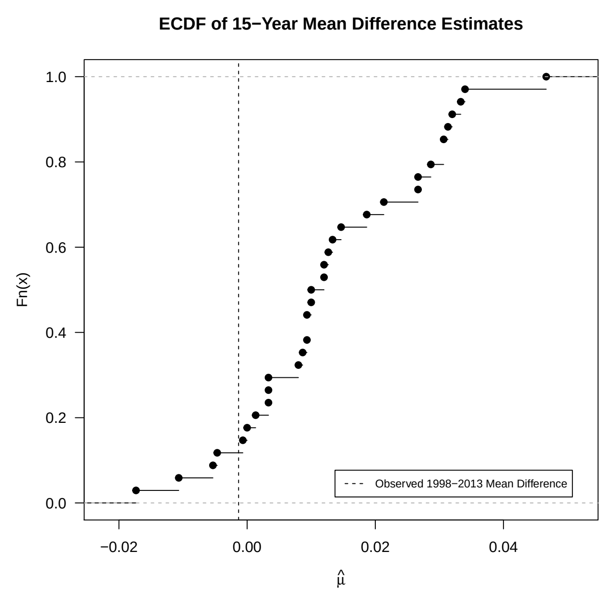


(b) ECDF of 15-year observed KS differences

Figure 10: Plots of 15-year KS difference estimates observed between 1950 to 2013. The dashed lines in (a) and (b) indicate the observed 15-year KS value for the period 1998–2013.

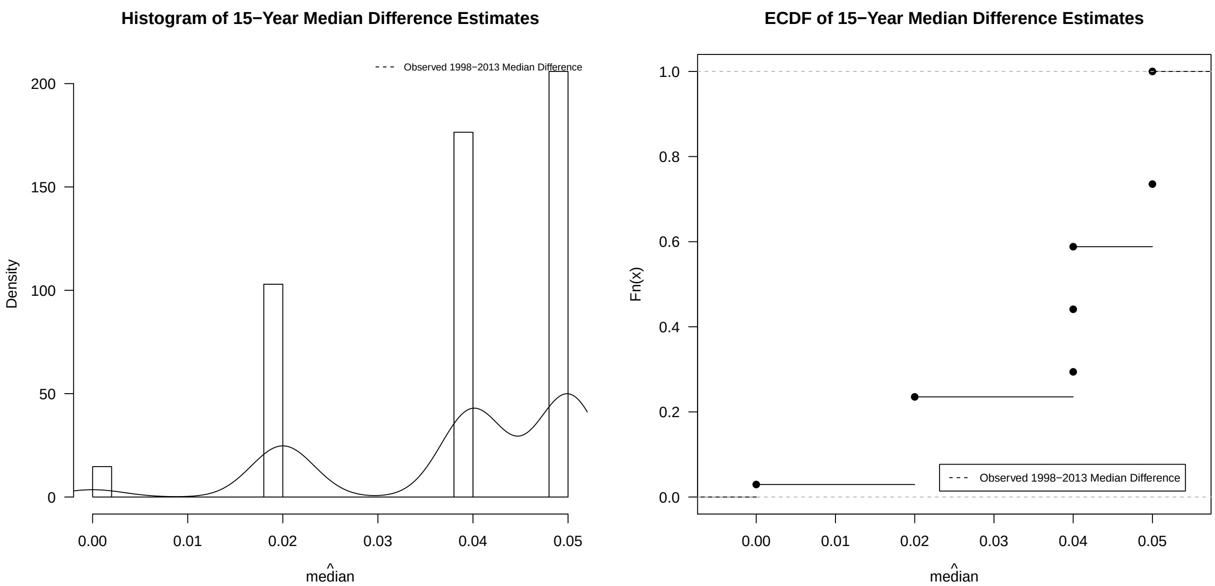


(a) Histogram of 15-year observed mean differences



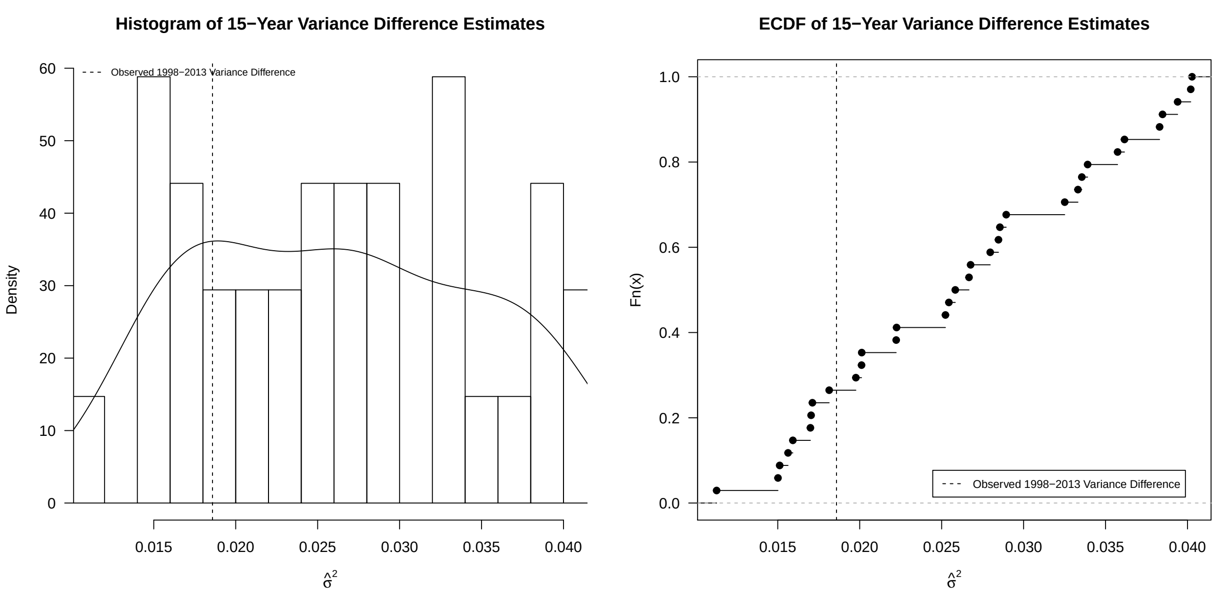
(b) ECDF of 15-year observed mean differences

Figure 11: Plots of 15-year mean difference estimates observed between 1950 to 2013. The dashed lines in (a) and (b) indicate the observed 15-year mean value for the period 1998–2013.



(a) Histogram of 15-year observed median differences (b) ECDF of 15-year observed median differences

Figure 12: Plots of 15-year median difference estimates observed between 1950 to 2013. The dashed lines in (a) and (b) indicate the observed 15-year median value for the period 1998–2013.



(a) Histogram of 15-year observed variance in the differences (b) ECDF of 15-year observed variance in the differences

Figure 13: Plots of 15-year variance estimates in the differences observed between 1950 to 2013. The dashed lines in (a) and (b) indicate the observed variance for the period 1998–2013.

The analysis described above was also repeated when the starting year of 1998 was varied to 1999 or 2000 - see Table 8. The K-S bootstrap based test for difference in distributions is not significant at the nominal 5% level when the later cut-off years of 1999 or 2000 are used. In fact the p-value which was lower than the 5% level in the 1998 analysis is no longer less than 0.05. The sensitivity analysis once more reveals that hiatus claims can be linked to the reference year of 1998.

Table 8: Summary Table of results for Hypothesis IV using starting years 1999 and 2000

Test	Bootstrap		Subsampling	
	1999	2000	1999	2000
Difference in distribution	$= 0.611$	$= 0.578$	$= 0.250$	$= 0.237$
Difference in mean	$= 0.995$	≈ 0.906	$= 0.583$	$= 0.579$
Difference in median	$= 0.434$	≈ 0.515	$= 0.194$	$= 0.974$
Difference in variance	$= 0.251$	≈ 0.378	$= 0.056$	$= 0.132$
Difference in log variance	$= 0.175$	≈ 0.335	–	–

4 Supplementary Section: Incorporating Observational Uncertainties

Note that there are two distinct questions that can be asked regarding the "trend" in the temperature series. The first is whether the observed temperature record exhibits an upward or downward trend that is greater than variations which can be attributed to observational error alone. This questions aims to characterize the observed record. The second question aims to understand if there is a deterministic component in the underlying stochastic process which generates the data. These two questions are distinct and understanding if a trend is significant requires examining both the observational uncertainties and also assessing the variability that is inherent in the underlying model.

One of the datasets of global surface temperature anomalies used in our analysis is the HadCRUT4 data, produced from the Met Office Hadley Centre in collaboration with the University of East Anglia Climatic Research Unit (CRU). The HadCRUT4 data is "an ensemble data set in which the 100 constituent ensemble members sample the distribution of likely surface temperature anomalies given our current understanding of these uncertainties." [Morice et al., 2012]

The primary HadCRUT4 series that was analyzed is the median of the 100 ensemble member time series. In order to account for the observational uncertainties in the data, all of the analyses was rerun on "the lower and upper bounds of the 95% confidence interval of the combined effects of all the uncertainties described in the HadCRUT4 error model (measurement and sampling, bias and coverage uncertainties)." [Morice et al., 2012]

The results from all three analyses are shown in Tables 9 and 10. Most of the conclusions are robust to the choice of HadCRUT4 series we use. The main difference is appears in the results for Hypothesis I. There does not appear to be a significant linear trend during the hiatus period in the median series for the HadCRUT4 dataset, whereas there is a significant linear trend at the 5% significance level in the lower and upper series. The conclusion of a significant linear trend during the hiatus period is consistent with the results that was found using the NASA GISS temperature anomalies dataset.

Table 9: Significance table comparing the results across the three data sets with hiatus start year of 1998/1999/2000. The p-values provided in the table are determined as follows: First the p-values are observed as the block size increases and until these p-values stabilize. The highest of the stabilized p-values is chosen in order to ensure Type I error control.

<table>
<tr><th></th><th>Lower</th><th>Median</th><th>Upper</th></tr>
<tr><td>Hypothesis I: Bootstrap
$H_0 : \beta_{\text{post}} = 0$ vs. $H_A : \beta_{\text{post}} \neq 0$</td><td>1998 | 1999 | 2000
0.031 | 0.024 | 0.033
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.194 | 0.081 | 0.235
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.030 | 0.025 | 0.036
Does not vary with cutoff year</td></tr>
<tr><td>Hypothesis II: Bootstrap
$H_0 : \beta_{\text{pre}} = \beta_{\text{post}}$ vs. $H_A : \beta_{\text{pre}} \neq \beta_{\text{post}}$</td><td>1998 | 1999 | 2000
0.439 | 0.322 | 0.425
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.393 | 0.284 | 0.469
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.481 | 0.357 | 0.469
Does not vary with cutoff year</td></tr>
<tr><td>Hypothesis III
$H_0 : \mathbb{E}(x_{\text{cutoff}}) = \mathbb{E}(x_{\text{cutoff}+t})$
vs. $H_A : \mathbb{E}(\text{cutoff}) \neq \mathbb{E}(x_{\text{cutoff}+t})$</td><td></td><td></td><td></td></tr>
<tr><td>$\mathbb{E}(x_{\text{cutoff}}) = x_{\text{cutoff}}$
Variance assumed fixed</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td></tr>
<tr><td>$\mathbb{E}(x_{\text{cutoff}}) = \widehat{\mu}_{\text{cutoff}}$
Variance assumed fixed</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td></tr>
<tr><td>$\mathbb{E}(x_{\text{cutoff}}) = x_{\text{cutoff}}$
Variance simulated by bootstrap</td><td>1998 | 1999 | 2000
0.419 | 0.622 | 0.602
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.420 | 0.631 | 0.607
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.417 | 0.622 | 0.615
Does not vary with cutoff year</td></tr>
<tr><td>$\mathbb{E}(x_{\text{cutoff}}) = \widehat{\mu}_{\text{cutoff}}$
Variance simulated by bootstrap</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.001 | < 0.001 | < 0.001
Does not vary with cutoff year</td></tr>
</table>

Table 10: Significance table comparing the results across the three data sets with hiatus start year of 1998/1999/2000. The p-values provided in the table are determined as follows: First the p-values are observed as the block size increases and until these p-values stabilize. The highest of the stabilized p-values is chosen in order to ensure Type I error control.

<table>
<tr><th></th><th>Lower</th><th>Medium</th><th>Upper4</th></tr>
<tr><td colspan="4">Hypothesis IV: Bootstrap</td></tr>
<tr><td>Kolmogorov-Smirnov Test
$H_0 : F_{\Delta X} = F_{\Delta Y}$ vs. $H_A : F_{\Delta X} \neq F_{\Delta Y}$</td><td>1998 | 1999 | 2000
0.207 | 0.302 | 0.370
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.218 | 0.334 | 0.435
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.319 | 0.425 | 0.497
Does not vary with cutoff year</td></tr>
<tr><td>Difference in mean
$H_0 : \overline{\Delta X} = \overline{\Delta Y}$ vs. $H_A : \overline{\Delta X} \neq \overline{\Delta Y}$</td><td>1998 | 1999 | 2000
0.488 | 0.732 | 0.692
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.385 | 0.886 | 0.799
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.525 | 0.703 | 0.673
Does not vary with cutoff year</td></tr>
<tr><td>Difference in median
$H_0 : \mathrm{med}(\Delta X) = \mathrm{med}(\Delta Y)$
vs. $H_A : \mathrm{med}(\Delta X) \neq \mathrm{med}(\Delta Y)$</td><td>1998 | 1999 | 2000
0.548 | 0.810 | 0.943
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.432 | 0.559 | 0.751
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.525 | 0.725 | 0.985
Does not vary with cutoff year</td></tr>
<tr><td>Difference in variance
$H_0 : \mathbb{V}\mathrm{ar}(\Delta X) = \mathbb{V}\mathrm{ar}(\Delta Y)$
vs. $H_A : \mathbb{V}\mathrm{ar}(\Delta X) \neq \mathbb{V}\mathrm{ar}(\Delta Y)$</td><td>1998 | 1999 | 2000
0.141 | 0.027 | 0.053
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.174 | 0.019 | 0.054
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.135 | 0.026 | 0.052
Varies with cutoff year</td></tr>
<tr><td>Difference in log variance
$H_0 : \log \mathbb{V}\mathrm{ar}(\Delta X) = \log \mathbb{V}\mathrm{ar}(\Delta Y)$
vs. $H_A : \log \mathbb{V}\mathrm{ar}(\Delta X) \neq \log \mathbb{V}\mathrm{ar}(\Delta Y)$</td><td>1998 | 1999 | 2000
0.097 | 0.007 | 0.016
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.123 | 0.003 | 0.024
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.099 | 0.009 | 0.032
Varies with cutoff year</td></tr>
<tr><td colspan="4">Hypothesis IV: Subsampling</td></tr>
<tr><td>Kolmogorov-Smirnov Test</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | 0.027
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | 0.056 | 0.026
Varies with cutoff year</td><td>1998 | 1999 | 2000
0.061 | 0.086 | 0.189
Does not vary with cutoff year</td></tr>
<tr><td>Difference in mean</td><td>1998 | 1999 | 2000
0.333 | 0.771 | 0.892
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.235 | 0.694 | 0.632
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.333 | 0.771 | 0.892
Does not vary with cutoff year</td></tr>
<tr><td>Difference in median</td><td>1998 | 1999 | 2000
0.424 | 0.343 | 0.703
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.294 | 0.389 | 0.447
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
0.424 | 0.343 | 0.703
Does not vary with cutoff year</td></tr>
<tr><td>Difference in variance</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td><td>1998 | 1999 | 2000
< 0.029 | < 0.028 | < 0.026
Does not vary with cutoff year</td></tr>
</table>

5 Supplementary Section: Scientific Claims & corresponding Statistical Hypotheses

Group I: The rate of change (linear increase) from 1998 onwards

- "Global mean surface temperature over the past 20 years (1993-2012) rose at a rate of $0.14 \pm 0.06\,^{\circ}$C per decade (95% confidence interval). This rate of warming is significantly slower than that simulated by the climate models..." (Fyfe, J. et al., Nature Climate Change, 2013)
- "...many governments are demanding a clearer explanation of the slowdown in temperature increases since 1998." (McGrath, M., BBC News, 2013)
- "Climate sceptics have seized on the temperature trends as evidence that global warming has ground to a halt." (Tollefson, J., Nature, 2014)
- "...despite a marked warming over the course of the 20th century, temperatures have not really risen over the past ten years." (The Economist, 2013)

Group II: Comparing the rates of change between the 1998-present period and before

- "The rate of global mean warming has been lower over the past decade than previously." (Otto et. al, Nature Geoscience, 2013)
- "The rise in the surface temperature of earth has been markedly slower over the last 15 years than in the 20 years before that." (Gillis, J., The New York Times, 2013)
- "...it is now clear that the rate of warming has slowed substantially over the past 15 years or so..." (Smith, D., Nature Climate Change, 2013)

Group III: Stalling of the global mean from 1998 onwards

- "Global warming first became evident beyond the bounds of natural variability in the 1970s, but increase in global mean surface temperatures have stalled in the 2000s." (Trenberth, K. and Fasullo, J., Earth's Future, 2013)
- "...average atmospheric temperatures have risen little since 1998..." (Tollefson, J., Nature, 2014)
- "Despite the continued increase in atmospheric greenhouse gas concentrations, the annual-mean global temperature has not risen in the twenty-first century..." (Kosaka and Xie, Nature, 2013)
- "...the Earth's mean near-surface temperature paused its rise during the 2000-2010 period." (Guemas et al., Nature Climate Change, 2013)
- "Average global temperatures hit a record high in 1998 – and then the warming stalled." (Tollefson, J., Nature, 2014)

Group IV: Difference in year-to-year temperature increases

- "...many governments are demanding a clearer explanation of the slowdown in temperature increases since 1998." (McGrath, M., BBC News, 2013)
- "...despite a marked warming over the course of the 20th century, temperatures have not really risen over the past ten years." (The Economist, 2013)

6 Supplementary Section: Comparisons with the recent study by Karl et al. (2015)

	NOAA without 2014 (using ERSSTv3)	NOAA without 2014 (using ERSSTv4)	NOAA with 2014 (using ERSSTv4)
Hypothesis I: Bootstrap $H_0 : \beta_{\text{post}} = 0$ vs. $H_A : \beta_{\text{post}} \neq 0$	1998 $\mid$ 1999 $\mid$ 2000 0.172 $\mid$ 0.090 $\mid$ 0.244 Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 0.001 $\mid$ 0.000 $\mid$ 0.002 Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 0.000 $\mid$ 0.000 $\mid$ 0.000 Does not vary with cutoff year
Hypothesis II: Bootstrap $H_0 : \beta_{\text{pre}} = \beta_{\text{post}}$ vs. $H_A : \beta_{\text{pre}} \neq \beta_{\text{post}}$	1998 $\mid$ 1999 $\mid$ 2000 0.237 $\mid$ 0.188 $\mid$ 0.332 Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 0.425 $\mid$ 0.292 $\mid$ 0.425 Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 0.452 $\mid$ 0.324 $\mid$ 0.435 Does not vary with cutoff year
Hypothesis III $H_0 : \mathbb{E}(x_{\text{cutoff}}) = \mathbb{E}(x_{\text{cutoff}+t})$ vs. $H_A : \mathbb{E}(\text{cutoff}) \neq \mathbb{E}(x_{\text{cutoff}+t})$			
$\mathbb{E}(x_{\text{cutoff}}) = x_{\text{cutoff}}$ Variance assumed fixed	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year
$\mathbb{E}(x_{\text{cutoff}}) = \widehat{\mu}_{\text{cutoff}}$ Variance assumed fixed	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year
$\mathbb{E}(x_{\text{cutoff}}) = x_{\text{cutoff}}$ Variance simulated by bootstrap	1998 $\mid$ 1999 $\mid$ 2000 0.423 $\mid$ 0.623 $\mid$ 0.571 Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 0.565 $\mid$ 0.557 $\mid$ 0.493 Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 0.671 $\mid$ 0.539 $\mid$ 0.437 Does not vary with cutoff year
$\mathbb{E}(x_{\text{cutoff}}) = \widehat{\mu}_{\text{cutoff}}$ Variance simulated by bootstrap	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year	1998 $\mid$ 1999 $\mid$ 2000 $< 0.001 \mid < 0.001 \mid < 0.001$ Does not vary with cutoff year

Table 11: Significance table comparing the results across the three data sets. The p-values that are reflected in the table are determined as follows: First the p-values are observed as the block size increases until these values stabilize. The highest of the stabilized p-values is chosen in order to reflect conservative Type I error control.

	NOAA without 2014 (using ERSSTv3)	NOAA without 2014 (using ERSSTv4)	NOAA with 2014 (using ERSSTv4)
Hypothesis IV: Bootstrap			
Kolmogorov-Smirnov Test $H_0 : F_{\Delta X} = F_{\Delta Y}$ vs. $H_A : F_{\Delta X} \neq F_{\Delta Y}$	1998 | 1999 | 2000 0.218 | 0.261 | 0.383 Does not vary with cutoff year	1998 | 1999 | 2000 0.206 | 0.221 | 0.357 Does not vary with cutoff year	1998 | 1999 | 2000 0.187 | 0.214 | 0.298 Does not vary with cutoff year
Difference in mean $H_0 : \overline{\Delta X} = \overline{\Delta Y}$ vs. $H_A : \overline{\Delta X} \neq \overline{\Delta Y}$	1998 | 1999 | 2000 0.331 | 0.962 | 0.887 Does not vary with cutoff year	1998 | 1999 | 2000 0.432 | 0.857 | 0.762 Does not vary with cutoff year	1998 | 1999 | 2000 0.587 | 0.677 | 0.638 Does not vary with cutoff year
Difference in median $H_0 : \text{med}(\Delta X) = \text{med}(\Delta Y)$ vs. $H_A : \text{med}(\Delta X) \neq \text{med}(\Delta Y)$	1998 | 1999 | 2000 0.411 | 0.928 | 0.595 Does not vary with cutoff year	1998 | 1999 | 2000 0.595 | 0.825 | 0.351 Does not vary with cutoff year	1998 | 1999 | 2000 0.947 | 0.431 | 0.317 Does not vary with cutoff year
Difference in variance $H_0 : \mathbb{V}\text{ar}(\Delta X) = \mathbb{V}\text{ar}(\Delta Y)$ vs. $H_A : \mathbb{V}\text{ar}(\Delta X) \neq \mathbb{V}\text{ar}(\Delta Y)$	1998 | 1999 | 2000 0.103 | 0.025 | 0.051 Varies with cutoff year	1998 | 1999 | 2000 0.155 | 0.034 | 0.064 Varies with cutoff year	1998 | 1999 | 2000 0.091 | 0.024 | 0.029 Varies with cutoff year
Difference in log variance $H_0 : \log \mathbb{V}\text{ar}(\Delta X) = \log \mathbb{V}\text{ar}(\Delta Y)$ vs. $H_A : \log \mathbb{V}\text{ar}(\Delta X) \neq \log \mathbb{V}\text{ar}(\Delta Y)$	1998 | 1999 | 2000 0.067 | 0.004 | 0.024 Varies with cutoff year	1998 | 1999 | 2000 0.068 | 0.012 | 0.023 Varies with cutoff year	1998 | 1999 | 2000 0.062 | 0.008 | 0.017 Varies with cutoff year
Hypothesis IV: Subsampling			
Kolmogorov-Smirnov Test	1998 | 1999 | 2000 0.029 | 0.083 | 0.132 Varies with cutoff year	1998 | 1999 | 2000 0.029 | 0.083 | 0.132 Varies with cutoff year	1998 | 1999 | 2000 < 0.030 | 0.029 | 0.081 Varies with cutoff year
Difference in mean	1998 | 1999 | 2000 0.147 | 0.556 | 0.605 Does not vary with cutoff year	1998 | 1999 | 2000 0.265 | 0.667 | 0.632 Does not vary with cutoff year	1998 | 1999 | 2000 0.333 | 0.714 | 0.865 Does not vary with cutoff year
Difference in median	1998 | 1999 | 2000 < 0.029 | 0.583 | 0.816 Varies with cutoff year	1998 | 1999 | 2000 0.471 | 0.611 | 0.947 Does not vary with cutoff year	1998 | 1999 | 2000 0.697 | 0.971 | 1.000 Does not vary with cutoff year
Difference in variance	1998 | 1999 | 2000 < 0.029 | < 0.028 | < 0.026 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.029 | < 0.028 | < 0.026 Does not vary with cutoff year	1998 | 1999 | 2000 < 0.030 | < 0.029 | < 0.027 Does not vary with cutoff year

Table 12: Significance table comparing the results across the three data sets. The p-values that are reflected in the table are determined as follows: First the p-values are observed as the block size increases until these values stabilize. The highest of the stabilized p-values is chosen in order to reflect conservative Type I error control.