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.

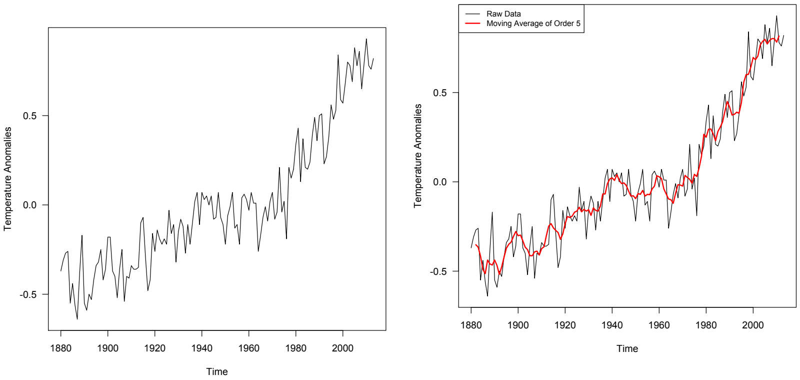
	NASA	NOAA	HadCRUT4
Hypothesis IV: Bootstrap			
Kolmogorov-Smirnov Test $H_0 : F_{\Delta X} = F_{\Delta Y}$ vs. $H_A : F_{\Delta X} \neq F_{\Delta Y}$	1998 | 1999 | 2000 0.295 | 0.662 | 0.632 Does not vary with cutoff year	1998 | 1999 | 2000 0.218 | 0.261 | 0.383 Does not vary with cutoff year	1998 |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

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.



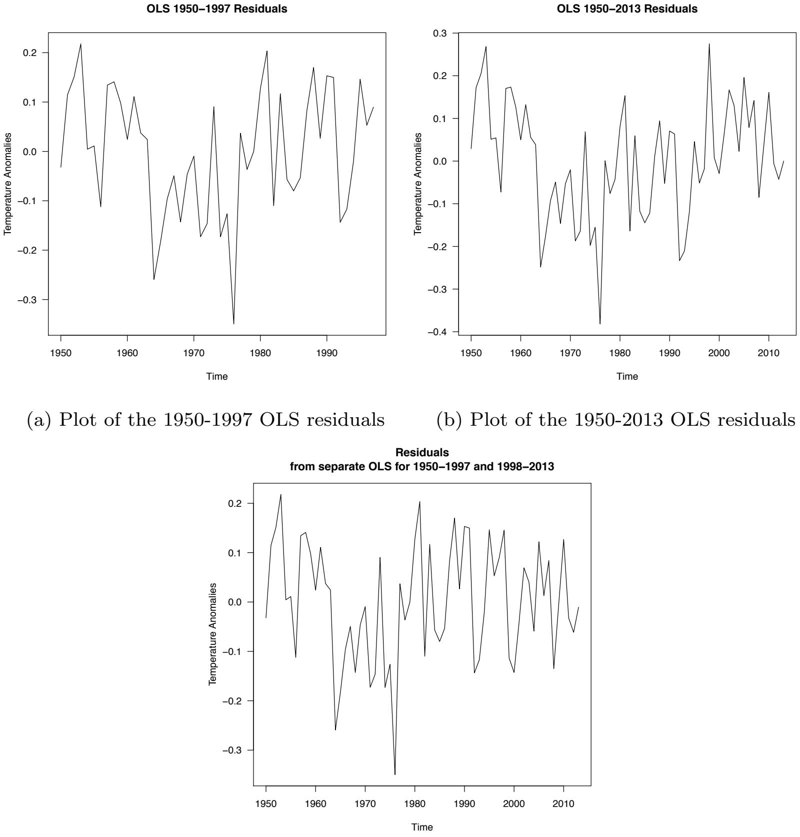
(a) NASA GISS data

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

Fig. 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.

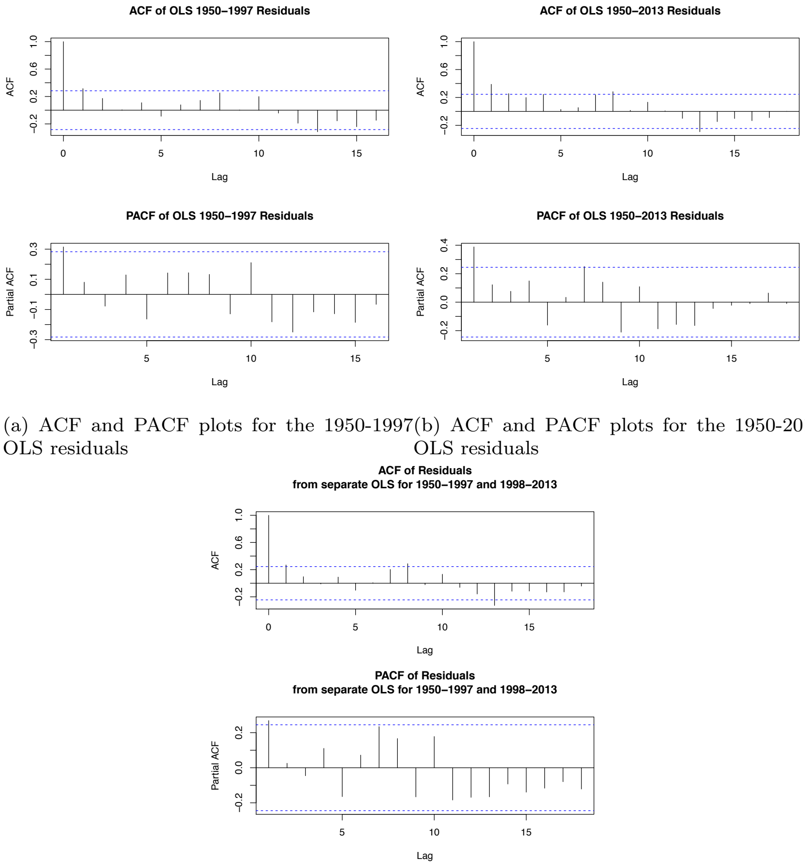


(a) Plot of the 1950-1997 OLS residuals

(b) Plot of the 1950-2013 OLS residuals

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

Fig. 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

Fig. 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 $\mathbb{V}\text{ar}(\varepsilon_t) = \sigma^2$. 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 $\text{se}(\widehat{\beta}_1) = 0.0052$. The Wald statistic is constructed as

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

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 $B = 1000$.

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	

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.

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.

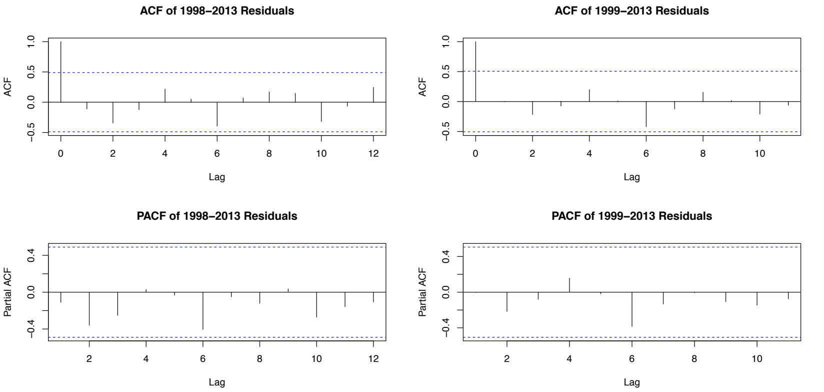


Fig. 8: Top: ACF plots of residuals time series 1998-2013 (left) and 1999-2013 (right). Bottom: PACF plots of residuals time series 1998-2013 (left) and 1999-2013 (right).

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 $\widehat{\beta}_{1,1}^*, \ldots, \widehat{\beta}_{1,B}^*$.

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 $\mathbb{E}(\varepsilon_t) = 0, \mathbb{V}\text{ar}(\varepsilon_t) = \sigma^2$. 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

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

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.

[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.

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

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

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 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

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

where $\varDelta X$ denotes the year-to-year increase in annual temperatures during 1950-1998 and $\varDelta Y$ 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,

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

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:

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

where $\varDelta X$ denotes the change in annual temperatures between 1894-1998 and $\varDelta Y$ denote the change between 1998-2013. The Kolmogorov-Smirnov statistic is given by

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

where $F_{\varDelta X,m}$ and $F_{\varDelta Y,n}(x)$ are the empirical distribution functions of $\varDelta X$ and $\varDelta Y$, 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 $\varDelta X_t$ and $\varDelta Y_s$ to generate bootstrap series $\varDelta X_t^*$ and $\varDelta Y_s^*$ for $1950 \leq t \leq 1998$ and $1999 \leq s \leq 2013$.
2. Compute the bootstrap Kolmogorov-Smirnov statistic

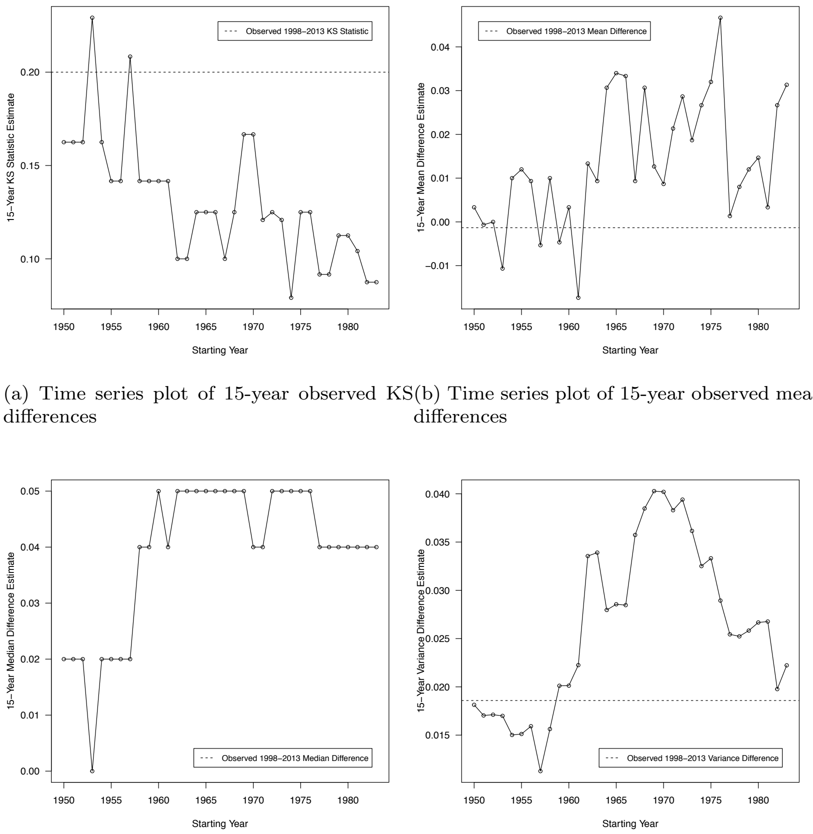
$$D^* = \sup_x \left| F_{\varDelta X^*,m}(x) - F_{\varDelta Y^*,n}(x) \right|.$$

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

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

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.



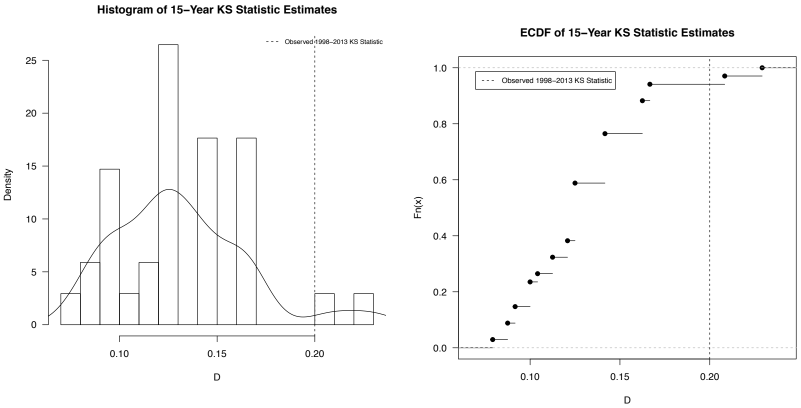
(a) Time series plot of 15-year observed KS differences

(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

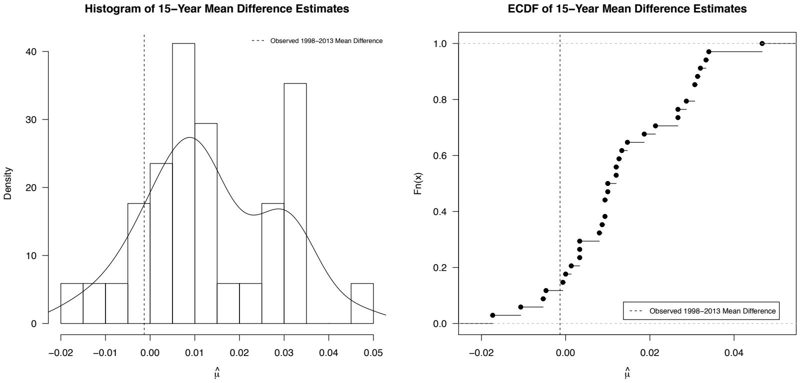
Fig. 9: Time series plots of 15-year difference estimates observed between 1950 to 2013.



(a) Histogram of 15-year observed KS differences

(b) ECDF of 15-year observed KS differences

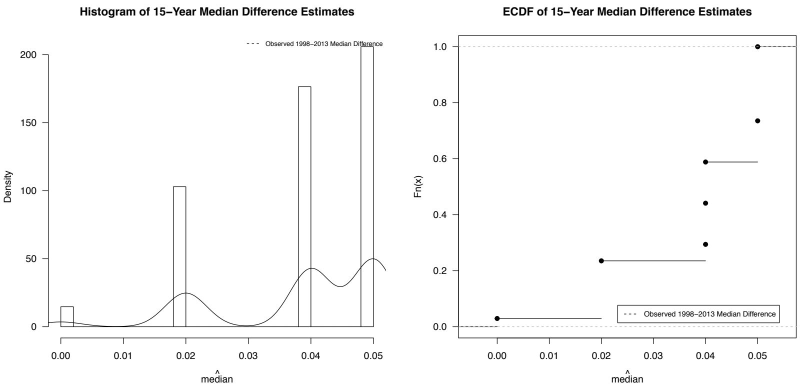
Fig. 10: Plots of 15-year KS difference estimates observed between 1950 to 2013. The dashed lines in (a) and (b) indicate the observed 15-year KS value for the period 1998–2013.



(a) Histogram of 15-year observed mean differences

(b) ECDF of 15-year observed mean differences

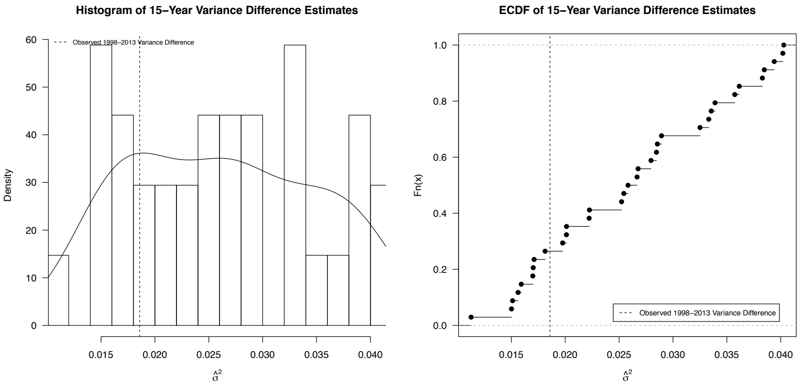
Fig. 11: Plots of 15-year mean difference estimates observed between 1950 to 2013. The dashed lines in (a) and (b) indicate the observed 15-year mean value for the period 1998–2013.



(a) Histogram of 15-year observed median differences

(b) ECDF of 15-year observed median differences

Fig. 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

Fig. 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.

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

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 | 1999 | 2000 0.172 | 0.090 | 0.244 Does not vary with cutoff year	1998 | 1999 | 2000 0.001 | 0.000 | 0.002 Does not vary with cutoff year	1998 | 1999 | 2000 0.000 | 0.000 | 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 | 1999 | 2000 0.237 | 0.188 | 0.332 Does not vary with cutoff year	1998 | 1999 | 2000 0.425 | 0.292 | 0.425 Does not vary with cutoff year	1998 | 1999 | 2000 0.452 | 0.324 | 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 | 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 0.423 | 0.623 | 0.571 Does not vary with cutoff year	1998 | 1999 | 2000 0.565 | 0.557 | 0.493 Does not vary with cutoff year	1998 | 1999 | 2000 0.671 | 0.539 | 0.437 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 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.