Nicole A Brown

Registration number 100082177

2019

Change-point Analysis: Did Global Warming Stop in 1998?

Supervised by Dr Gavin Cawley

University of East Anglia
Faculty of Science
School of Computing Sciences

Abstract

The anthropogenic impact since industrialisation is evident in accelerated global warming. Contrarily, the academic discussion has moved towards a "pause" or substantially slower warming trend in global temperatures. This concept has spread to public and political opinion. Many scientific articles indicate a change-point around 1998 but lack an understanding of natural variability and fail to address the limitations of change-point tests under autocorrelation. This report will investigate the effect of autocorrelation on change-point testing and further explore the concerns when performing statistical analysis on global temperature data.

Acknowledgements

I want to thank my supervisor Dr Gavin Cawley and the module organiser Dr Pierre Chardaire for their advice and support.

Contents

List of Figures

List of Tables

1. Introduction

Scientific articles refer to a slowdown or pause in global warming from 1998 onwards despite increased levels of Greenhouse Gasses (GG), Kosaka and Xie (2013) Santer and others (2014) H. England et al. (2014). This notion is not consistent with the expected Global Mean Surface Temperature (GMST) response to increases in GG. There has been an increase in the number of articles of this nature since 2013, with little consensus on the definition of this slowdown or pause and failure to justify the methods used in the analysis. Risbey et al. (2018) reviews many of these articles and finds no statistical evidence of this slowdown. Despite this, 1998 has been pointed out as the start of a pause period. Respected scientific sources, Nature (2017) and Stocker et al. (2013), formalise this "pause" by reporting it as fact. This misuse of terms feeds the public narrative of global warming, Mooney (2013). Given the high-risk impact of these articles, statistically robust testing and acknowledging the drawbacks of methods used has never been more critical.

1.1. Natural Variation

Noise from natural variation can overshadow the greenhouse signal; Michell et al. (2001) worked to identify the signal of the forced response to GG from this "noise". It is essential to understand the behaviour of this variation and the factors that contribute to it. Dominant influences are El Niño–Southern Oscillation (ENSO), the aerosol loading from volcanic eruptions and to a lesser extent the solar cycle.

ENSO is the fluctuation in sea surface temperature and air pressure of the overlying atmosphere in the Pacific Ocean around the equator. It has the most substantial impact of the factors, Trenberth et al. (2002). Figure 1 shows the cold and warm events from 1950-1997. There is a shift to fewer cold spikes since 1976 and an exceptionally long period of warm spikes between 1990-1995, this is consistent with a Pacific equatorial upswelling in 1976 and a resumption in 1998. Stockwell and Cox (2009), using the Chow test for structural breaks, found a change-point in global temperature in 1978. ENSO and GMST appear highly correlated over short periods; with a lag between a spike and when the influence can be detected. The regime shift may have contributed to

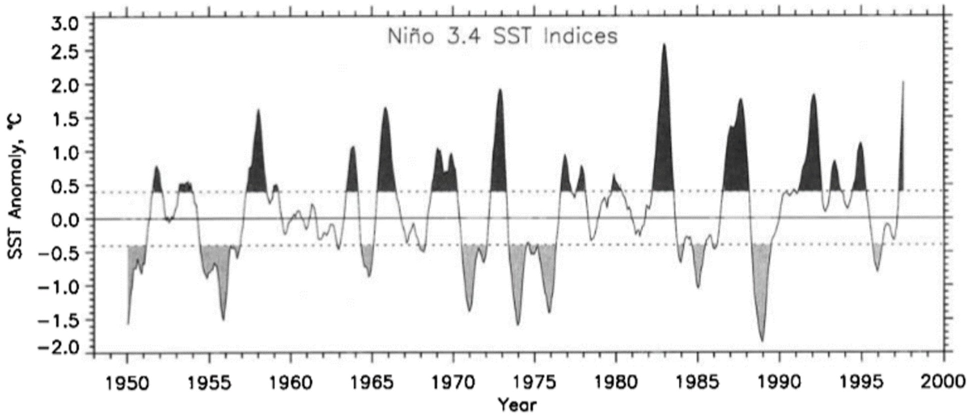


Fig. 1: Trenberth and Hoar (1997) Time series plots of the Niño 3.4 SST indices as five month running means using data from NOAA and relative to a base period climatology from 1950-79. Values exceeding threshold ±0.4°C for Niño 3.4 are stippled to indicate ENSO events.

the 1978 change-point, and the unusual spikes between 1990-1995 may have influenced the research that claims 1998 as another change-point.

Volcanic activity, specifically the build-up of particles in the atmosphere after an eruption, also has a significant effect on GMST. Abnormally high levels of volcanic activity lead to cooling. Rampino and Self (1982) found intensely active cooling after Krakatoa erupted in 1883 meaning sizeable volcanic forcing in the years following this event.

Foster and Rahmstorf (2011) model these main exogenous factors using: the multivariate el Nino index (MEI) with reference to Wolter and Timlin (1998), the aerosol optical thickness (AOD) data from Sato et al. (1993) and the total solar irradiance (TSI) data from Fröhlich (2006). Using multiple linear regression, they adjust several datasets removing the estimated effect of these factors. Figure 1 shows steady warming since 1998 beneath the "noise" suggesting that any deviation from an unchanging linear trend is due to natural variation. The trend of ENSO, aerosol loading and solar activity in most datasets (all except UAH) is negative, so the adjusted sets have slightly higher warming than observations in unadjusted sets. Lean and Rind (2008) performed a multivariate correlation analysis for the period 1889–2006 using the CRU temperature data, and found that they could explain 76% of the temperature variance over this period

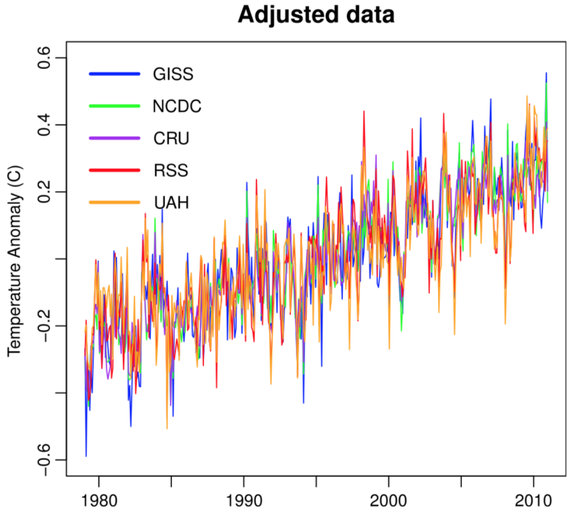


Fig. 2: Foster and Rahmstorf (2011) Five different global temperature data sets adjusted removing short-term effects.

from anthropogenic forcing, El Niño, volcanic aerosols and solar variability. After removing the effects of the exogenous factors, we can see the warming is mostly down to anthropogenic forcing approximately 0.163°C per decade which was rising steadily 1979-2010. Throughout this period the supposed slowdown or pause should be visible; this is not the case beyond natural variability. The critical thing to remember is that the natural forces mentioned have very short-term cycles and will not continue in the same direction to fix the warming caused by increased GG, as shown in Figure 3.

The Earth's climate varies on a vast range of temporal scales. Since the industrial revolution, a long-term change has occurred showing increase effects from these natural forces, Houghton et al. (2001). Modelling these natural forces is not straight forward and the models used are often over-simplified. O'Kane et al. (2013) uses non-stationary cluster analysis to identify ENSO dynamical regimes. This analysis reveals significant trends, indicating fundamental changes to the meta-stability of the ocean dynamics in response to changes in atmospheric forcing. Models are advancing as new computing techniques develop; these more advanced techniques allow the actual warming trend to show beneath the noise.

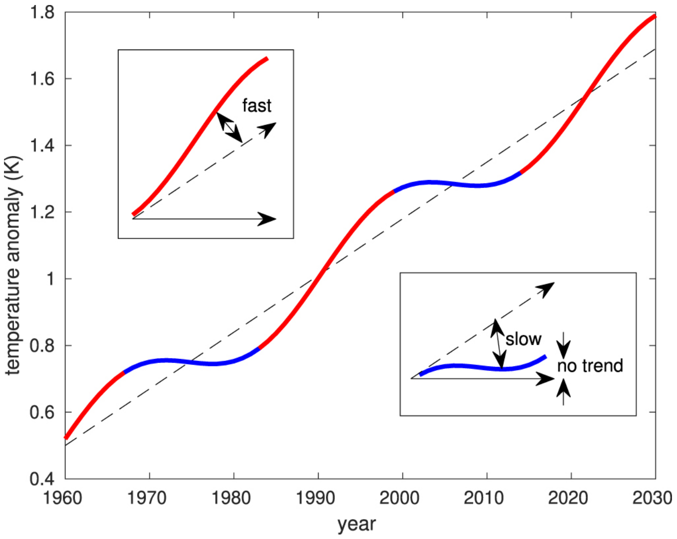


Fig. 3: Risbey et al. (2018) Smoothed global mean surface temperature series (blue/red line). The series is a linear trend plus sinusoidal variation to mimic multidecadal fluctuations.

1.2. Autocorrelation

The climate is an interconnected system of different forces. This absence of independence means that environmental variance is often positively autocorrelated; each random error is likely to be similar to the previous random error. Autocorrelation needs considering when modelling global temperature data. White noise could be used to model the variation if the random errors were independent of one another but when autocorrelation is present more complex series are needed.

Before statistical analysis, it is vital to recognise the serial correlation of residuals. Autocorrelation in the residual series suggests an inherent defect in most models and violates error homoskedasticity, which is an assumption of most change-point tests. Time series analysis can be used to detect the serial correlation of residuals from linear regression. The most straightforward approach is the Durbin-Watson Test statistic (1950,1951,1971).

1.3. The Chow Test

In time series analysis the Chow test is commonly used to test for the presence of a structural break. The test looks at whether the coefficients of two linear regressions are equal. Below is a simplified version of the Chow statistic given that the linear model $y = a + bx_1 + cx_2 + \varepsilon$, is split into two models $y_1 = a_1 + b_1x_1 + c_1x_2 + \varepsilon$ and $y_2 = a_2 + b_2x_1 + c_2x_2 + \varepsilon$:

$$F = \frac{(SS - SS_1 - SS_2)/k}{(SS_1 + SS_2)(\tau - 2k)}. \tag{1}$$

Where SS is the sum of squared residuals from the full model, SS_1 is the sum of squared residuals from the first split model, and SS_2 is the sum of squared residuals from the second split model. τ is the total number of observations, and k is the number of parameters. The test statistic follows the F distribution with k and $\tau - 2k$ degrees of freedom. The null hypothesis of the Chow test in this case would be $H_o : a_1 = a_2$, $b_1 = b_2$, and $c_1 = c_2$. In other words, there is no structural break.

Looking at the rejection probabilities of the Chow test under autocorrelation, the test is exceptionally non-robust Krämer (1989). Krämer focused on auto-aggressive disturbances based on an AR(1) model. Giles and Scott (1992) tested the robustness with the use of a moving average model, MA(1), and found similar distortion. The conclusion suggests that structural breaks may be falsely detected under autocorrelation as the Chow test is biased towards rejecting the null hypothesis.

Newey and West (1987) attempted to overcome the non-robustness of the Chow test under autocorrelation using a weighted estimator of the covariance matrix. Unfortunately, they concluded that the test is only trustworthy when the disturbances are independent. There is a need for further exploration into a resolution for this problem.

2. Related Work

2.1. Durbin-Watson Test

Consider the ordinary linear regression model:

$$y_i = a + \sum_{j=1}^{m} b_j x_{ij} + \varepsilon \tag{2}$$

Where m is the number of variables $(j = 1, 2..., m)$, n is the sample size $(i = 1, 2..., n)$, x_i represents the explanatory variables and y_i represents a dependent variable. The constant a is the intercept, b_j is the regression coefficients and ε_i is the predicted error. The predicted error is, $\varepsilon_i \sim WN(0, \sigma^2)$, a white noise series with a variance of σ^2. The residual series must be 0 for this predicted error. If errors come from random disturbances outside the model, they should fit this series.

The Durban-Watson statistic can be used to test if the residual series fits the white noise series:

$$DW = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2} = \frac{\sum_{i=1}^{n-1}(\Delta\varepsilon_i)^2}{\sum_{i=1}^{n}\varepsilon_i^2} = 2(1-p) \tag{3}$$

Where $\Delta\varepsilon_i = \varepsilon_{i-1}$, and

$$p = \frac{\sum_{i=2}^{n}\varepsilon_i\varepsilon_{i-1}}{\sum_{i=1}^{n}\varepsilon_i^2} \tag{4}$$

The example above presumes a lag of 1. The autocorrelation coefficient of the residual series is denoted by p and has the possible values $-1 \leq p \leq 1$. There is presumed to be no serial correlation when $p = 0$. The test statistic (DW) varies from 0 to 4, values close to 2 represent no autocorrelation at a specified significance level α.

The Durbin-Watson test does have limitations; the test cannot accurately measure the level of autocorrelation but can only reject the null hypothesis of non-autocorrelated errors. Also, the test is only appropriate for serial correlation of residuals from the least squares regression of a times series, Chen (2015).

With regards to a split regression model, as seen in the chow test (1.3), there are further complications. It is challenging to detect autocorrelation concerning misspecified models due to coefficient stability and unknown reliability, Consigliere (1981). To prevent least squares estimates developing from a misspecified model Consigliere suggests looking at the two regressions separately and estimating the true model errors, under stability and non-stability. In this case, the regression model above (2) divides at a change-point. The null hypothesis would be whether the autocorrelation coefficient is equal for both models, $H_o : p_1 = p_2$. However, this involves prior knowledge of the change-point.

2.2. Limitations

The main problems of past papers on this topic revolve around bias. This bias is usually due to prior knowledge before testing and underpins many of the statistical methods used. The change-point 1998 is often already presumed before testing. Risbey et al. (2018) suggest how this point may be chosen down to subsequent low warming and not a random selection. Looking back to Figure 3, this is a period where natural variation causes unusually lower warming rates. This selection bias gives implications for the interpretation of the p values as more false positives are likely, Lewandowsky et al. (2019). Easterling and Wehner (2009) refer to this bias as "cherry Picking" as it gives the researcher influence over the outcome. At the least, results should have an allowance for this bias. Methods such as multiple testing, for different start points, would overcome this issue.

The time interval used to test for a change-point seems to influence results. Multi-decadal variation means the length of data must outreach the short turn fluctuations. At least a 17-year interval is required for sufficient power to detect the signal; otherwise, the data will be too noisy, Santer and others (2011). Research that focuses on intervals less than 17 years often misinterpret natural variation for a slower warming trend.

Throughout different research papers, there is a lack of agreement on precisely how to define a change-point. This change is sometimes described as a switch to no warming trend (either zero or negative) or as a substantially lower trend. Without a clear definition, it is difficult to know how to correctly test if a change-point is present.

Methods to assess change-points vary. The simple Chow test gives a broken trend without regard for the previous trend. This instantaneous jump introduces another degree of freedom, which affects the statistical significance of the change-point. An instant spike in temperature is not scientifically plausible and would need to be justified. A continuous trend test, such as that used by Stockwell and Cox (2009), would take into account the previous trend. A continuation also means an extra degree of freedom does not come into the analysis. A continuous and broken trend example is shown in Figure 4:

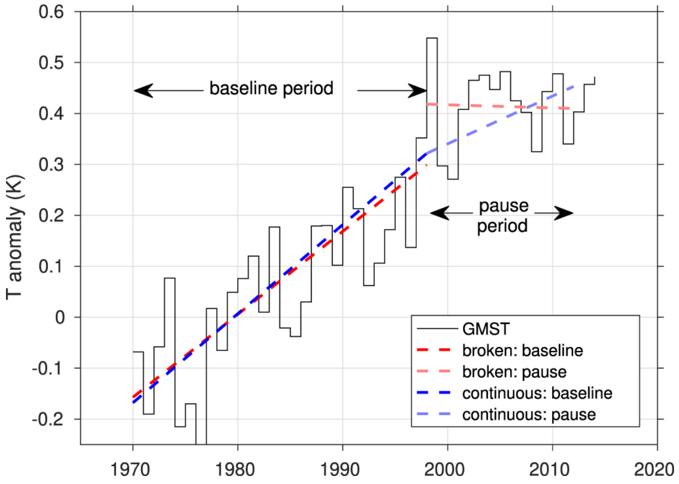


Fig. 4: Risbey et al. (2018) GMST annual mean series for HadCRUT3 (black line). The best fit broken trends pre and post 1998 are shown as red lines and the best fit continuous trends about 1998 are shown as blue lines.

3. Design

3.1. Datasets

There are several sources of GMST data. Each dataset varies on coverage and reliability. Observations from stations across the surface of the globe have inadequate coverage in areas such as the Arctic; interpolation is used to overcome this. Satellite datasets have better coverage in general and get 100% coverage every few days, but records are only available from 1979. Satellite data has calibration delays and unknown effects of orbital decay, Foster and Rahmstorf (2011). The technology has also evolved, most notably the switch from MSU to AMSU technology with the launch of NOAA-15 in 1998. GMST datasets, surface and satellite, will always come with limitations and uncertainties.

With historical datasets, there are more biases in the measurements. HadCRUT has lower covariance with the overall data but with each update moves closer to other datasets. Before these updates, HadCRUT3 underestimated warming due to inadequate coverage. The switch from HadCRUT3 to HadCRUT4 partially corrects for ship-buoy bias with the increased use of buoy records, Risbey et al. (2018). These updates can lead to the calculation of different trend magnitudes. HadCRUT, NOAA and GISTEMP data all show significant changes after bias reductions. Unfortunately, there are delays in resolving these biases to ensure transparency in the procedure. These delays cause miscommunication between data providers and researchers; with failure to highlight the limitations of the data. Significant differences in datasets and old versions allow researchers to select data based on bias and the expected effect on the outcome of their analysis.

Throughout this investigation the datasets GISSTEMP and HadCRUT will be used for modelling and analysis. They have different coverage, with HadCRUT having the lowest warming estimate of short-term trends. The data interval will start in 1979 as satellite data is only available from this point. Comparing the results from each dataset will help determine if there is a meaningful difference in long-term responses to GG.

3.2. Testing for a Change-point

This investigation will look at the reliability of both a continuous and broken trend change-point test. The original code for both tests is in appendix A. The method incorporates the Chow Test, 1.3, with a 95% confidence interval and uses the statistic:

$$F = \frac{(RSS - ESS)}{ESS * (n - 2k)}. \tag{5}$$

Where EES is the explained sum of squares compared to RSS, the residual sum of squares, with a change-point splitting sub-sections. The initial test plots for HadCRUT and GISSTEMP are in appendix B.

3.3. Analysis of Time series

Before testing the robustness of the change-point tests, it is required to analyse both GMST time series to see how they compare. Figures 5 and 6 show plots of the raw data converted to time series objects. Linear regression models, in the form of equation 6, overlap the time series.

$$temperature \sim time(temperature) \tag{6}$$

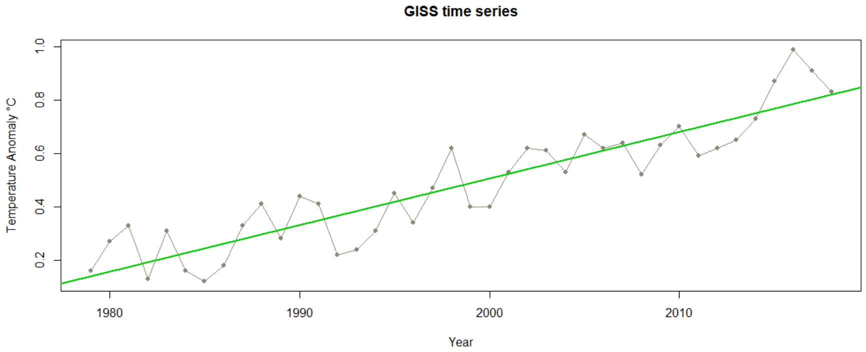


Fig. 5: Time series with linear trend line from GISTEMP annual averages of GMST, from 1979-2018.

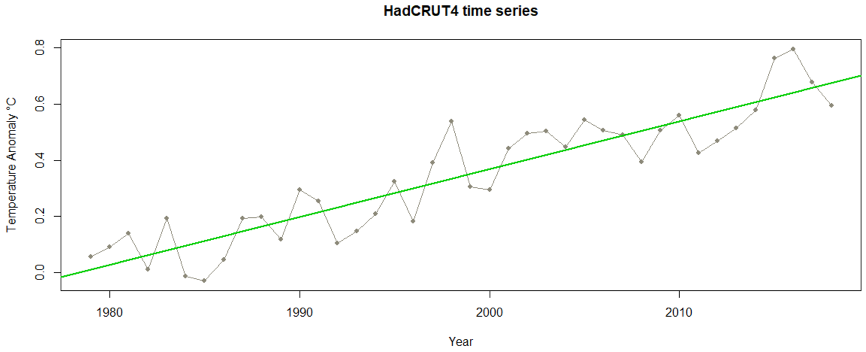


Fig. 6: Time series with linear trend line from HadCRUT4 annual averages of GMST, from 1979-2018.

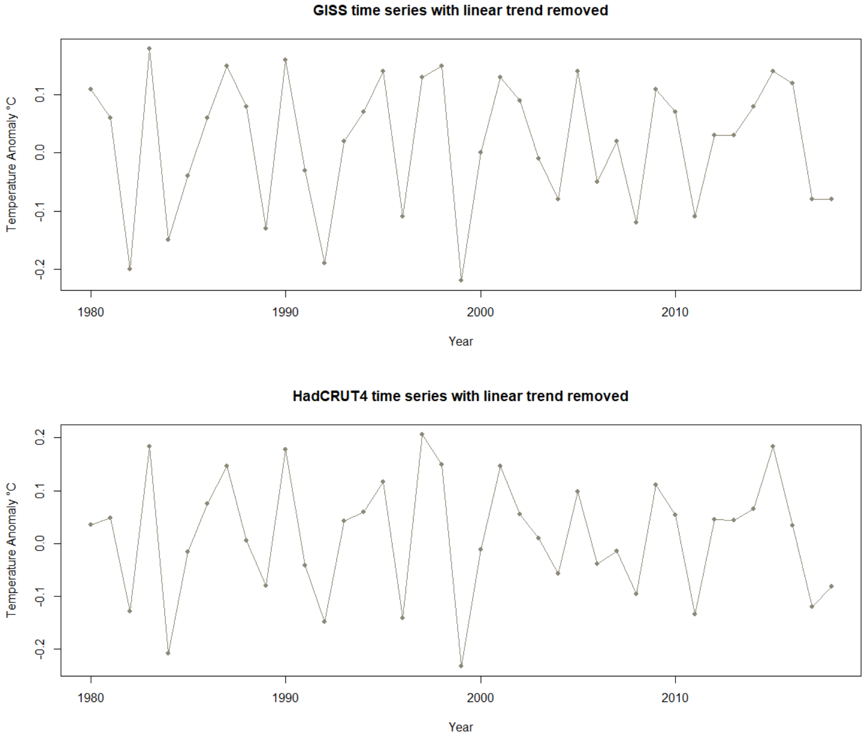


Fig. 7: Time series of GISTEMP and HadCRUT after the linear trend is removed.

Although, Figures 5 and 6 have noticeable differences the variances are similar. Removing the linear trend from both time series suggests similar variance, Figure 7. The square root of the residuals of the GISTEMP model is 0.09319 and for the HadCRUT model 0.08993. The Durbin-Watson test is used to look more specifically at autocorrelation in Table 1. Although the test is not exact, the similar results confirm both have autocorrelation present.

Table 1: Durbin-Watson Test[a]

	DW	p
GISTEMP	1.4469	0.02428
HadCRUT	1.5367	0.04745

[a] For Durbin-watson test description refer to section 2.1.

In both time series, true autocorrelation is greater than 0. When the ACF plots of both time series are side by side, they are nearly identical. As the error variance and autocorrelation are similar, further analysis will only focus on HadCRUT4.

3.3.1. Seasonality

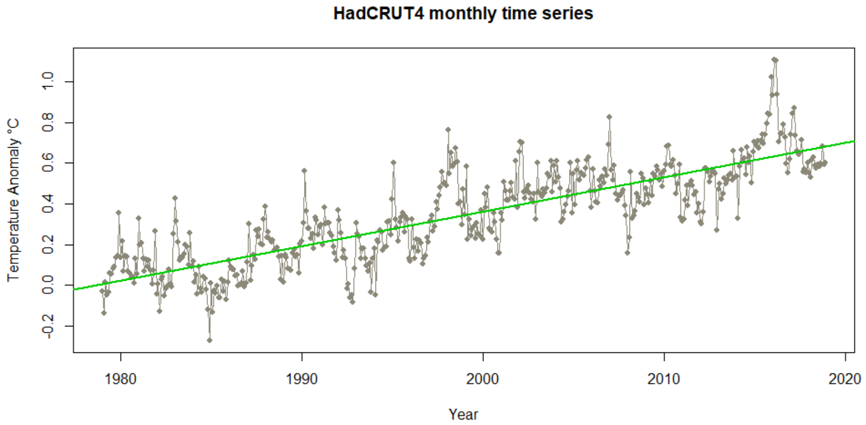


Fig. 8: Time series with linear trend line from HadCRUT4 monthly averages of GMST, from 1979-2018.

For monthly data, the time series frequency changes from 1 to 12. Figure 8 shows the monthly time series with a linear trend. To better understand the effect of seasonality on this time series a boxplot of each monthly cycle and decomposition of the time series are shown in Figure 9. The seasonality plot shows a significant influence on the overall trend.

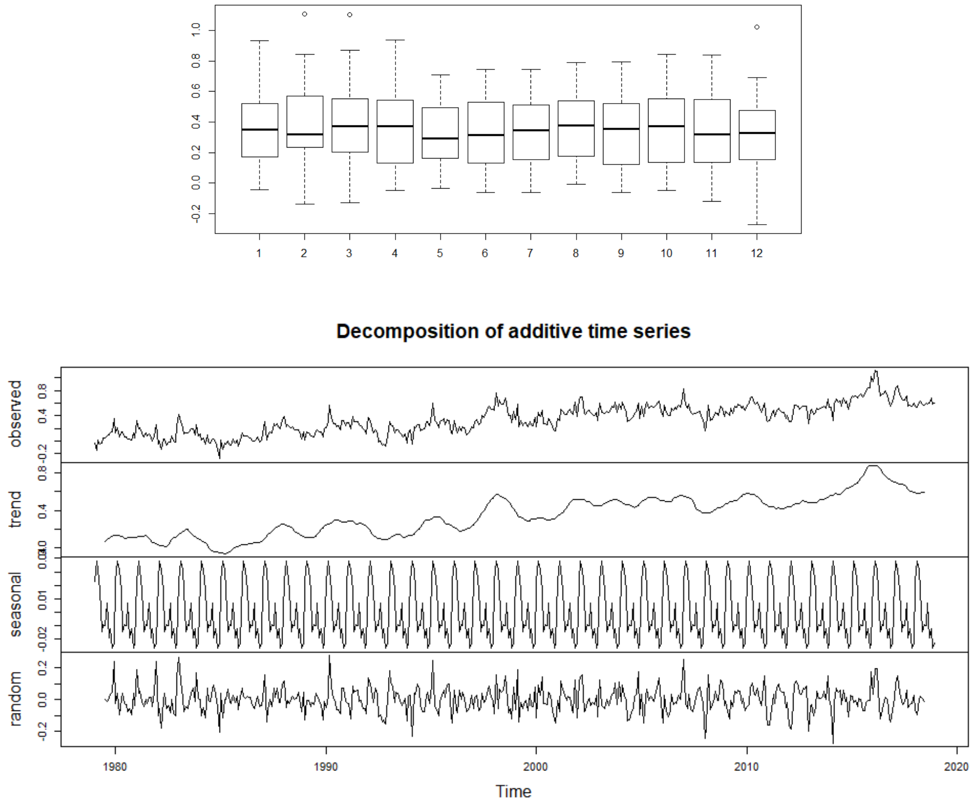


Fig. 9: Boxplot of time series monthly cycle and decomposition.

The seasonality is subtracted to overcome this influence. Looking at Figure 11, with the linear trend removed the variation is slightly higher with seasonality. The standard deviation of the residuals of the seasonally adjusted time series is 0.1282.

Using the Durbin-Watson test, in Table 2, the alternative hypothesis of true autocorrelation is met both before and after seasonality is removed.

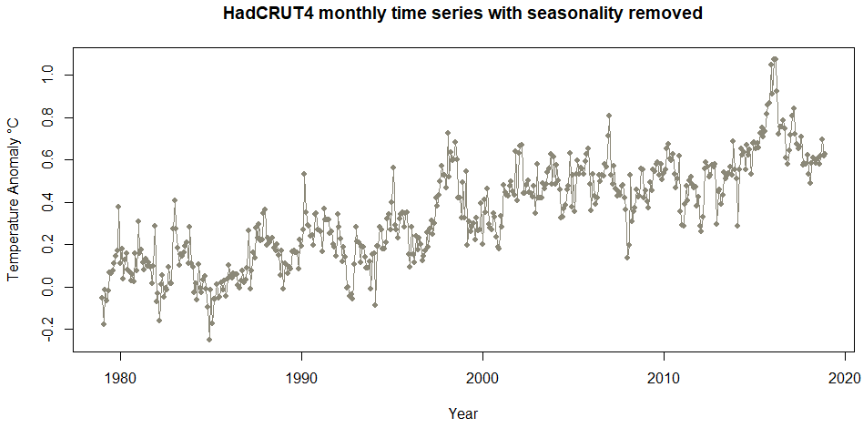
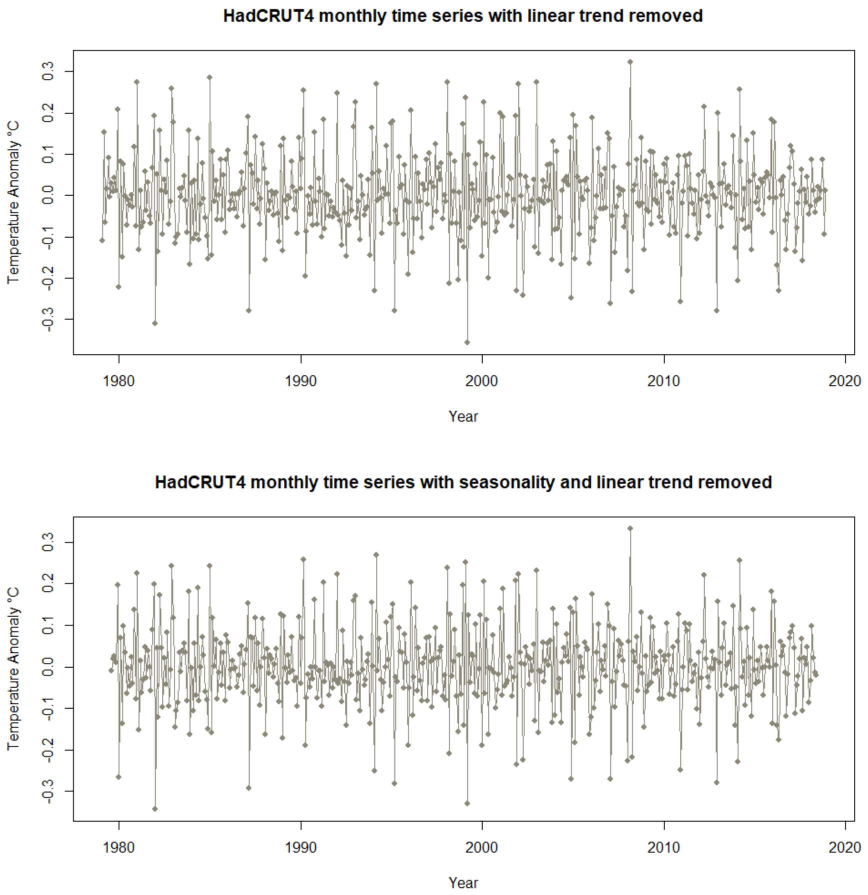


Fig. 10: Monthly HadCRUT time series with seasonality removed.

Table 2: Durbin-Watson Test after removing seasonality[a]

	DW	**p**
Monthly Time series	0.63213	2.2e-16
Seasonality removed	0.6247	2.2e-16

[a] For Durbin-watson test description refer to section 2.1.



HadCRUT4 monthly time series with linear trend removed

HadCRUT4 monthly time series with seasonality and linear trend removed

Fig. 11: Monthly HadCRUT time series with linear trend removed and seasonality removed.

3.4. The Robustness of the Chow Test

Running the change-point tests in simulations demonstrates their robustness to autocorrelation. Synthetic data based on a linear regression model with error variance from a white noise series, such as that found in equation 2, will give a false positive rate under no autocorrelation. This benchmark rate is then compared to the false positive rate when using an AR series, in this case an AR(1) model, for the error variance. Analysis of the GMST datasets enables the modelling of the synthetic data in both cases. For both the white noise simulations and AR simulations 1000 synthetic realisations run, repeated 1000 times. Any unknown significant change-points detected are recorded as false positives.

3.4.1. White Noise Generation

A model based on the summary of the HadCRUT3 annual time series is below:

$$temperature = -33.705978 + 0.017038 * time + \varepsilon \tag{7}$$

In this case, $\varepsilon \sim WN(0, \sigma^2)$. For every simulation, the white noise error is randomly generated. The square root of the residuals, from the linear model in equation 6, is used to model the white noise. Figure 12 is an example of annual synthetic white noise data. Comparing the synthetic example to Figure 6, the model matches up well with the real observations.

Figure 13 is an example of the monthly white noise simulations. The error variances are based on seasonally adjusted data rather than the original monthly time series.

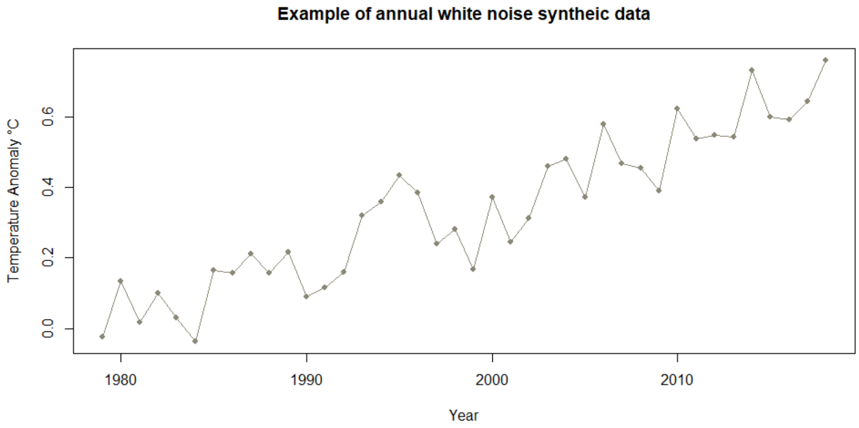


Fig. 12: Synthetic annual white noise data example.

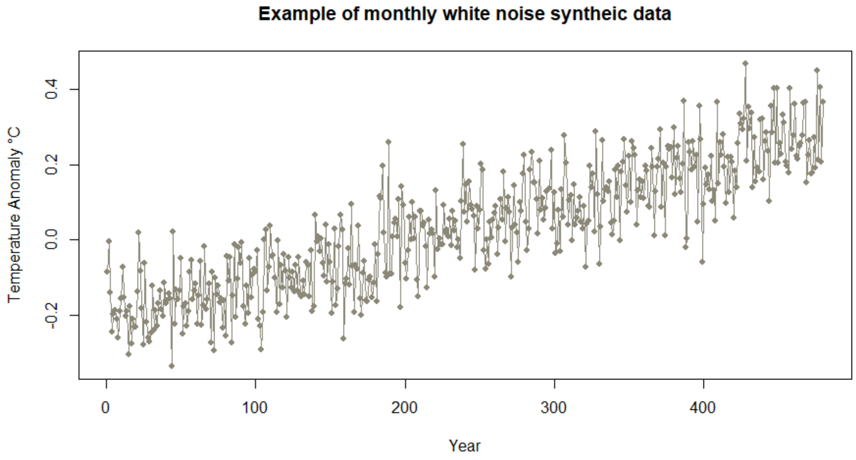


Fig. 13: Synthetic monthly white noise data example.

3.4.2. Simulating Autocorrelation

A simple expression of a AR(1) model:

$$x_t = \delta + \Phi_1 x_{t-1} + w_t. \tag{8}$$

Where x at time t is a linear function of the value of x at $t-1$ and w is normally distributed $N(0, \sigma_w^2)$ and independent of x.

The residuals from the GMST time series exhibit strong autocorrelation and the AR synthetic data must replicate this. Mudelsee (2010) was useful in the design of the synthetic AR model. Realisations of autocorrelated noise are based on an AR(1) model with the use of the arima.sim function in $\mathbb{R}$. Other AR models, such as an autoregressive moving average model ARMA(1,1), did not generate realistic synthetic data.

Figures 14 and 15 are examples of annual and monthly AR synthetic data. ACF plots and the augmented Dickey-Fuller Test suggest that the synthetic data matches well with the autocorrelation in real datasets.

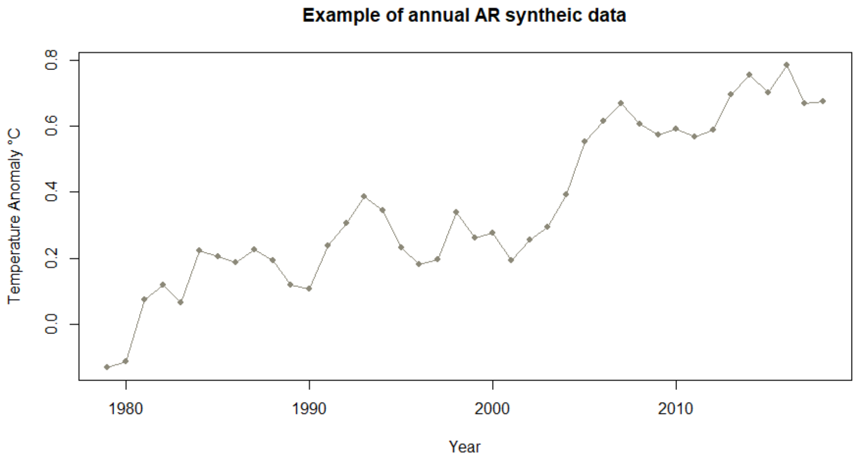


Fig. 14: Synthetic annual AR data example.

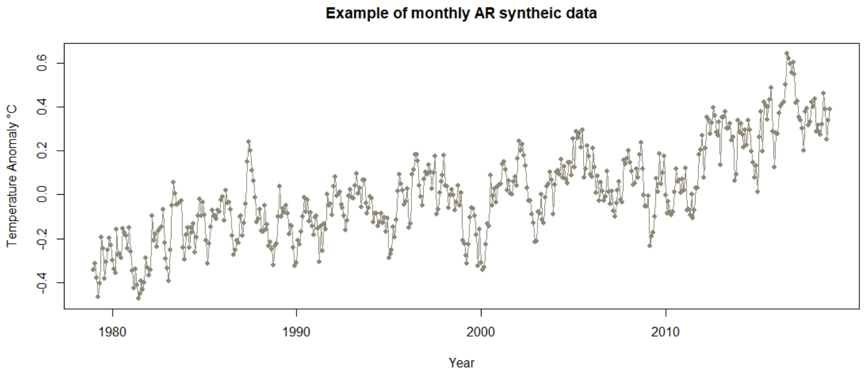


Fig. 15: Synthetic monthly AR data example.

4. Results and Analysis

The results from the simulations are below; a simple false positive rate considers the p values of each falsely identified change-point. The tests run each time with a 95% confidence level. Tables 3 and 4 separate results into annual and monthly synthetic data as the effect of autocorrelation is likely higher in monthly data. It is crucial to consider that these results are not an exact measure of the power of these tests but a way of comparing their robustness with and without the presence of autocorrelation.

4.1. White Noise Simulations

False positive results are low for annual data in the white noise simulations; 5% is the expected level. However, the rates are higher than expected for monthly simulations. The broken trend false positive rates are higher for both yearly and monthly data. Considering these results incorporated both underestimating and overestimating the trend the false positive rate could be divided by two to find the false positives in just one direction.

Table 3: False positive rates under White Noise Simulations

	Broken Trend*	**Unbroken Trend***
Annual[a]	14%	5%
Monthly[b]	29%	9%
Rate[c]	21.5%	7%

[a] Example of simulations in Appendix C.

[b] Example of simulations in Appendix D.

[c] An average of the false positive rate for monthly and annual simulations.

*Reference appendix A.

4.2. AR Model Simulations

Looking at Table 4 the false positive rates substantially increased under autocorrelation. As with white noise simulations, the broken trend test produces false positives at a rate much higher than the continuous trend.

Table 4: False positive rates under AR Simulations

	Broken Trend*	**Unbroken Trend***
Annual[a]	40%	22%
Monthly[b]	42%	28%
Rate[c]	41%	25%

[a] Example of simulations in Appendix E.

[b] Example of simulations in Appendix F.

[c] An average of the false positive rate for monthly and annual simulations.

*Reference appendix A.

4.3. Analysis

The results of the simulation do confirm both tests are non-robust under autocorrelation. This is especially true for the simple chow test, which has the highest false positive rates.

The false positive rates, other than the annual white noise simulations, are all higher than would be expected. When using real-world data false positives may not be as common. These exceptionally higher rates may be down to the methods used in the error modelling or in the simulation method. There is need for further investigation into the cause of this.

5. Further Work

This paper gives evidence the non-robustness of change-point testing under autocorrelation but does not manage to overcome the problem. For a statistical resolution further work is required. With more time, this paper would have explored possible mathematical transformations and the robustness of adjusted data.

Using the Durbin-Watson test, a transformation of data could increase the reliability of these change-point tests. This requires a known change-point, however, using methods of multiple testing could overcome this issue. Consigliere (1981) suggests starting with whether the auto-correlation coefficient should be equal for the two sub-samples; under a null hypothesis of $\beta_1 = \beta_2$, and $p_1 = p_2$.:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} .\beta + \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} . \tag{9}$$

Reviews of past research allow improvements moving forwards. From these reviews, it is clear there has been a lack of understanding of how to correctly use datasets, and better communication of the limitations is required. At present most papers do not withstand statistical scrutiny but make conclusions despite this. Going forward, a clear definition of a change-point, with a null and alternative hypothesis, must be agreed. In the context of GMST, a physical description is essential, Risbey et al. (2018). There must be a focus on reducing bias and having a standard in testing. At the least, the sensitivities of tests employed and significant biases must be addressed. As research into this topic increases, more robust testing techniques are arising.

As computing technology develops, there are more avenues to explore when looking at GMST. There is now an increased understanding of climate systems, Lewandowsky et al. (2019), with new modelling capabilities. The advances in computing can also be applied to change-point analysis.

6. Conclusion

In conclusion, there is no significant evidence of a pause in 1998, Risbey et al. (2018). Many scientific articles that have claimed a change-point in 1998 fail to address: the limitations of data, the drawbacks of statistical methods and the effect of bias.

A critical issue is the responsibility of researchers to avoid bias. The unreliability of the change-point tests under autocorrelation is commonly known but often ignored. When researchers use the power of choice in statistical methods, datasets or timeframes they can alter results. There is a powerful influence from climate change contrarians which has directed conversation in research towards substantially slower warming. Given the high risk of climate change, there is a huge moral obligation to prevent this bias. When phrases like "pause" or "slowdown" arise in scientific research, this spreads to public and political opinion. The confusion over the urgency of global warming has meant, up until recently, it has not been a priority of global organisations and governments. The implications of this loss of emergency cannot be known, but are likely significant. The reduced momentum for government implementation on reducing GG has been disastrous. Increased natural disasters, weather anomalies and significant warming are all obvious effects of anthropogenic GG increases, Hansen et al. (2016). There is more recent progress with the announcement of Britain's net-zero carbon target of 2050.

Increased research can help resolve any future global warming contradictions that arise. These advancements include measuring uncertainty, quantifying trends and detecting change-points. There is now more pressure on researchers to deconstruct biases and asses the costs.

A. Code for Chow Test

```
# broken
t1 = t[1:(break-1)]
x1 = x[1:(break-1)]
t2 = t[break:length(t)]
x2 = x[break:length(t)]
xfit1 = lm(x1~t1)
xfit2 = lm(x2~t2)
S1 = sum(xfit1$res^2)
S2 = sum(xfit2$res^2)
this.buff = this.buff+1
f = (St-S1-S2)*(length(t)-4)/(S1+S2)/2
p = pf(ff,2,length(t)-4,lower.tail=FALSE)
Ft[this.buff] = f
pv[this.buff] = p
# unbroken
phi = t-t[break]
phi[phi < 0] = 0
xfit = lm(x~t+phi)
Sub = sum(xfit$res^2)
ff1 = (St-Sub)*(length(t)-3)/Sub
pp1 = pf(f1,1,length(t)-3,lower.tail=FALSE)
F1[this.buff] = ff1
p1[this.buff] = pp1
```

B. Initial Plots

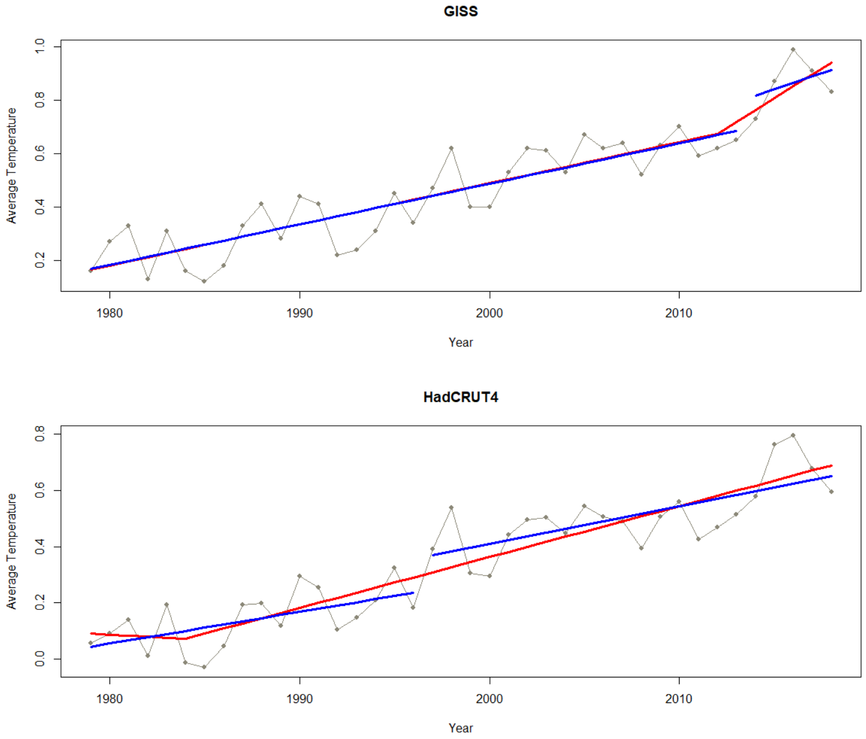


Fig. 16: Initial Plots - data sources: GISS - NASA's Goddard Institute for Space Studies. HadCRUT - Climatic Research Unit (University of East Anglia) in conjunction with the Hadley Centre (UK Met Office).

C. Annual White noise simulation example

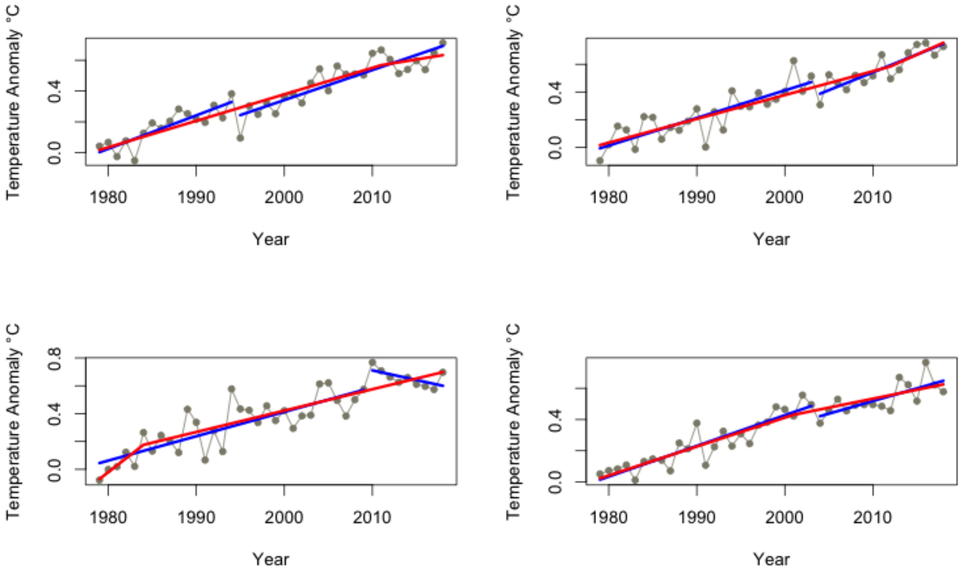


Fig. 17: Sample of annual white noise simulations.

D. Monthly White noise simulation example

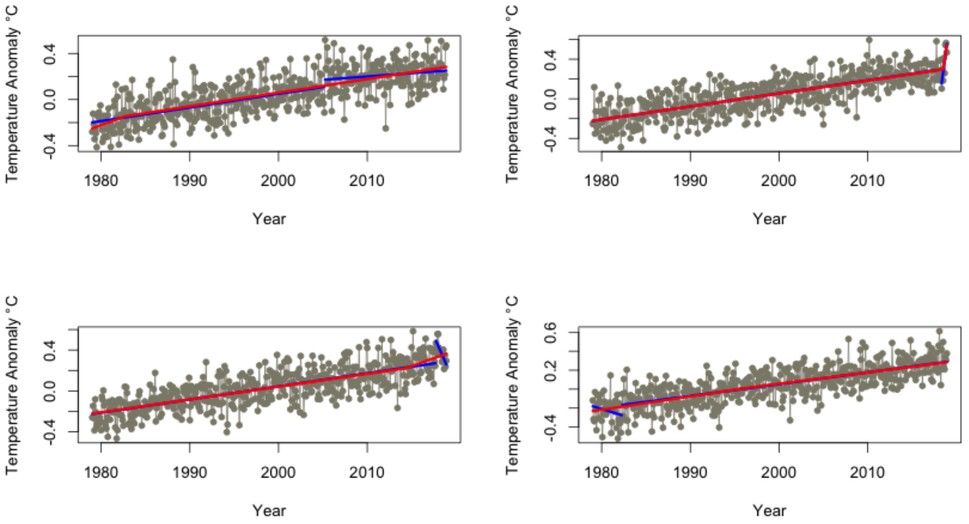


Fig. 18: Sample of monthly white noise simulations.

E. Annual AR model simulation example

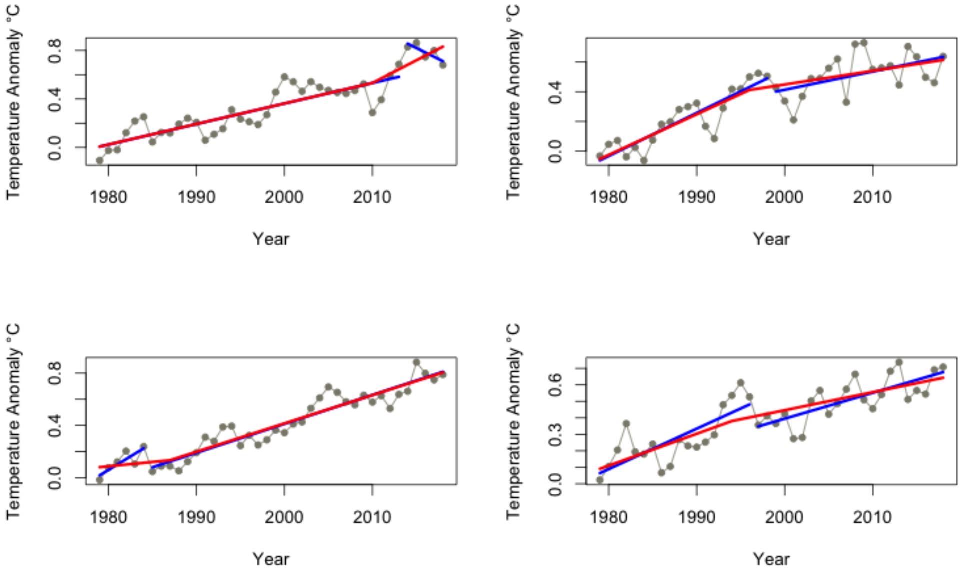


Fig. 19: Sample of annual AR simulations.

F. Monthly AR model simulation example

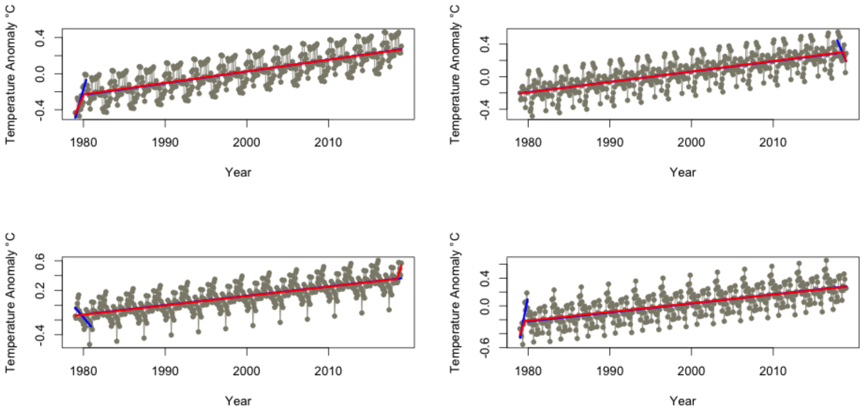


Fig. 20: Sample of monthly AR simulations.

References

Chen, Y. (2015). Spatial autocorrelation approaches to testing residuals from least squares regression. *PLOS ONE*, 11.

Consigliere, I. (1981). The chow test with serially correlated errors. *Rivista Internazionale di Scienze Sociali*, 89(2):125–137.

Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression.I. *Biometrika*, 37(3-4):409–428.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression.II. *Biometrika*, 38(1-2):159–178.

Durbin, J. and Watson, G. S. (1971). Testing for serial correlation in least squares regression.III. *Biometrika*, 58(1):1–19.

Easterling, D. R. and Wehner, M. F. (2009). Is the climate warming or cooling? *Geophysical Research Letters*, 36(8).

Foster, G. and Rahmstorf, S. (2011). Global temperature evolution 1979–2010. *Environmental Research Letters*, 6(4):044022.

Fröhlich, C. (2006). Solar irradiance variability since 1978: Revision of the pmod composite during solar cycle 21. *Space Science Reviews*, 125:53–65.

Giles, D. and Scott, M. (1992). Some consequences of using the chow test in the context of autocorrelated disturbances. *Economics Letters*, 38(2):145 – 150.

H. England, M., Mcgregor, S., Spence, P., Meehl, G., Timmermann, A., Cai, W., Sen Gupta, A., J. McPhaden, M., Purich, A., and Santoso, A. (2014). Recent intensification of wind-driven circulation in the pacific and the ongoing warming hiatus. *Nature Climate Change*, 4:222–227.

Hansen, J., Sato, M., Hearty, P., Ruedy, R., Kelley, M., Masson-Delmotte, V., Russell, G., Tselioudis, G., Cao, J., Rignot, E., Velicogna, I., Tormey, B., Donovan, B., Kandiano, E., von Schuckmann, K., Kharecha, P., LeGrande, A. N., Bauer, M., and Lo, K.-W. (2016). Ice melt, sea level rise and superstorms: Evidence from paleoclimate data, climate modeling, and modern observations that 2c global warming could be dangerous. *Atmos. Chem. Phys.*, 16:3761–3812.

Houghton, J., H. Ding, Y., Griggs, D., Noguer, M., van der Linden, P., Dai, X., Maskell, M., and A. Johnson, C. (2001). *Climate Change 2001: The Scientific Basis*, page 881.

Kosaka, Y. and Xie, S.-P. (2013). Recent global-warming hiatus tied to equatorial pacific surface cooling. *Nature*, 501.

Krämer, W. (1989). The robustness of the chow test to autocorrelation among disturbances. *Statistical Analysis and Forecasting of Economic Structural Change*.

Lean, J. L. and Rind, D. H. (2008). How natural and anthropogenic influences alter global and regional surface temperatures: 1889 to 2006. *Geophys. Res. Lett.*, 35:L18701.

Lewandowsky, S., Cowtan, K., Risbey, J. S., Mann, M. E., Steinman, B. A., Oreskes, N., and Rahmstorf, S. (2019). Erratum: The ‘pause’ in global warming in historical context: II. comparing models to observations (2018 environ. res. lett. 13 123007). *Environmental Research Letters*, 14(4):049601.

Michell, J., Karoly, D., Folland, C., and others, m. (2001). *Detection of Climate Change and Attribution of Causes*, pages 697–738.

Mooney, C. (2013). Who created the global warming “pause”?

Mudelsee, M. (2010). *Climate time series analysis. Classical statistical and bootstrap methods. 2nd ed*, volume 51.

Nature (2017). What pause? *Nature*, 545(7652).

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

O'Kane, T., Matear, R., A. Chamberlain, M., S. Risbey, J., Sloyan, B., and Horenko, I. (2013). Decadal variability in an ogcm southern ocean: Intrinsic modes, forced modes and metastable states. *Ocean Modelling*, 69:1–21.

Rampino, M. and Self, S. (1982). Historic eruptions of tambora (1815), krakatau (1883), and agung (1963), their stratospheric aerosols, and climatic impact. *Quaternary Research*, 18:127–143.

Risbey, J., Lewandowsky, S., Cowtan, K., Oreskes, N., Rahmstorf, S., Jokimäki, A., and Foster, G. (2018). A fluctuation in surface temperature in historical context: Reassessment and retrospective on the evidence. *Environmental Research Letters*, 13:123008.

Santer, B. and others, m. (2011). Separating signal and noise in atmospheric temperature changes: The importance of timescale. *Journal of Geophysical Research*, 116(D22105).

Santer, B. and others, m. (2014). Volcanic contribution to decadal changes in tropospheric temperature. *Nature Geoscience*, 7.

Sato, M., Hansen, J. E., McCormick, M. P., and Pollack, J. B. (1993). Stratospheric aerosol optical depths, 1850-1990. *J. Geophys. Res.*, 98:22987–22994.

Stocker, T. F., Q. D. et al. (2013). *Technical summary*, pages 33–115. Cambridge University Press, Cambridge, UK.

Stockwell, D. R. B. and Cox, A. (2009). Structural break models of climatic regime-shifts: claims and forecasts. *ArXiv e-prints*.

Trenberth, K. E., Caron, J. M., Stepaniak, D. P., and Worley, S. (2002). Evolution of el nino southern oscillation and global atmospheric surface temperatures. *J. Geophys. Research*.

Trenberth, K. E. and Hoar, T. J. (1997). El niño and climate change. *Geophysical Research Letters*, 24(23):3057–3060.

Wolter, K. and Timlin, M. S. (1998). Measuring the strength of ENSO events: How does 1997/98 rank? *Weather*, 53:315–324.