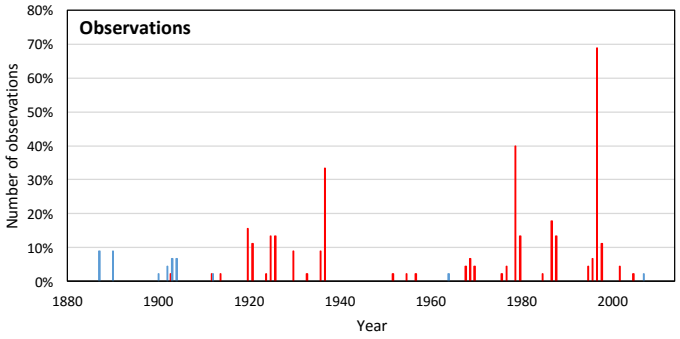
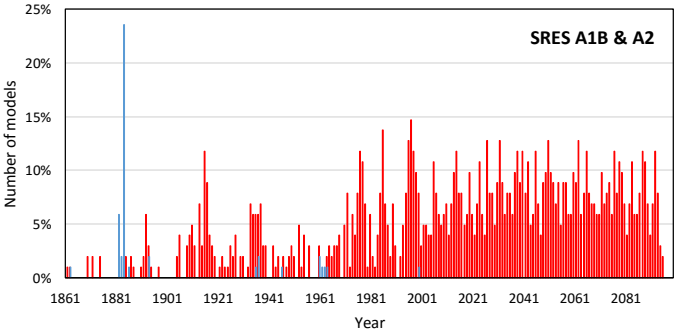
Reconciling the signal and noise of atmospheric warming on decadal timescales

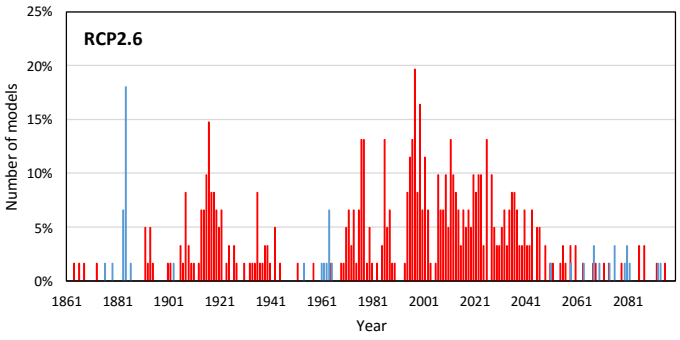
Roger N Jones and James H Ricketts

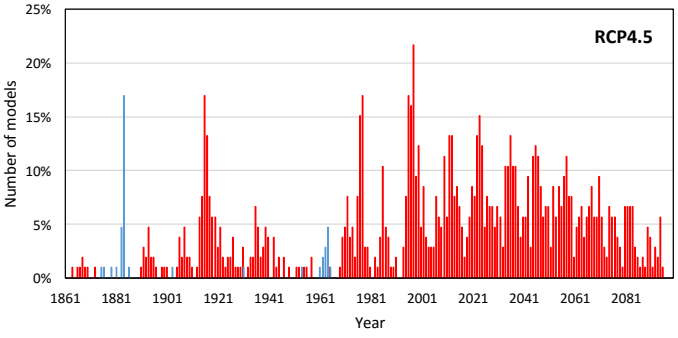
Climate Change Working Paper 38

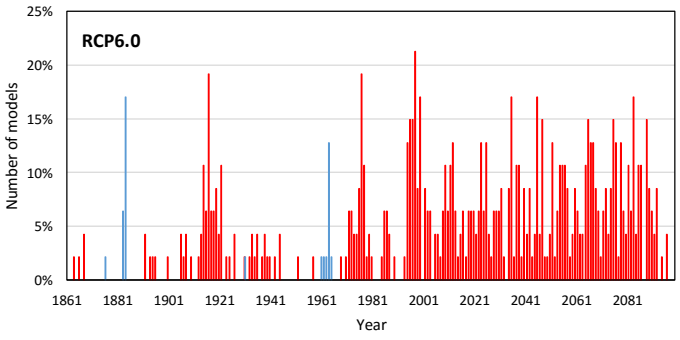
Victoria Institute of Strategic Economic Studies

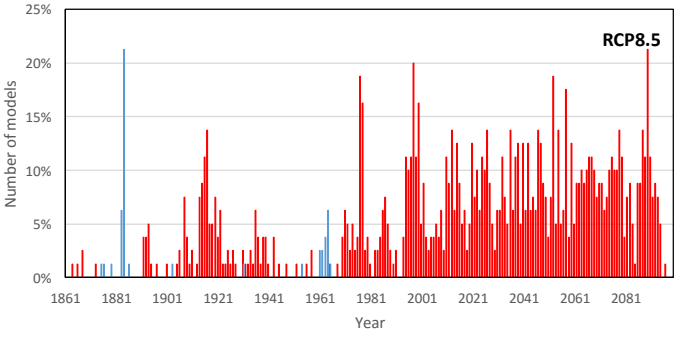












Key words: global warming, climate change, decadal variability, step change, severe testing, statistical induction, signal to noise, complex trends

Acknowledgements

JR is the holder of a Victoria University postgraduate research scholarship. Data sources include the Met Office Hadley Centre, National Aeronautics and Space Administration Goddard Institute for Space Studies and United States National Climatic Data Center, Berkeley Earth, Cowtan and Way and the Australian Bureau of Meteorology. CMIP3 and CMIP5 archives are made available by the modelling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM). The U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Fruitful discussions were held with Celeste Young, Peter Sheehan, Roger Bodman, Penny Whetton, Kevin Hennessy, Kris Ebi, Ben Preston, Jim Bowler, Rod Marsh and the big shift group. D Kelly O'Day provided the macro templates, which have been adapted to provide the step and trend charts.

Suggested citation

Jones, R.N. and Ricketts, J.H. (2016) Reconciling the signal and noise of atmospheric warming on decadal timescales. Climate Change Working Paper No. 38, Victoria Institute of Strategic Economic Studies, Victoria University, Melbourne, Australia.

Publishing Notes

This is a presubmission working paper and a companion paper to The climate wars and "the pause" – are both sides wrong? Climate Change Working Paper. No. 37.

ISBN: 978-1-86272-726-7

Victoria Institute of Strategic Economic Studies
Victoria University
Po Box 14428 Melbourne
Victoria Australia 8001
Phone +613 9919 1340
Fax +613 9919 1350
vises@vu.edu.au

Abstract

Interactions between externally-forced and internally-generated climate variations on decadal timescales is a major determinant of changing climate risk. Severe testing is applied to observed global and regional surface and satellite temperatures and modelled surface temperatures to determine whether these interactions are independent, as in the traditional signal-to-noise model, or whether they interact, resulting in steplike warming. The multi-step bivariate test is used to detect step changes in temperature data. The resulting data are then subject to six tests designed to show strong differences between the two statistical hypotheses, h_{step} and h_{trend}: (1) Since the mid-20th century, most of the observed warming has taken place in four events: in 1979/80 and 1997/98 at the global scale, 1988/89 in the northern hemisphere and 1968/70 in the southern hemisphere. Temperature is more steplike than trend-like on a regional basis. Satellite temperature is more steplike than surface temperature. Warming from internal trends is less than 40% of the total for four of five global records tested (1880–2013/14). (2) Correlations between step-change frequency in models and observations (1880–2005), are 0.32 (CMIP3) and 0.34 (CMIP5). For the period 1950–2005, grouping selected events (1963/64, 1968–70, 1976/77, 1979/80, 1987/88 and 1996–98), correlation increases to 0.78. (3) Steps and shifts (steps minus internal trends) from a 107-member climate model ensemble 2006–2095 explain total warming and equilibrium climate sensitivity better than internal trends. (4) In three regions tested, the change between stationary and non-stationary temperatures is steplike and attributable to external forcing. (5) Steplike changes are also present in tide gauge observations, rainfall, ocean heat content, forest fire danger index and related variables. (6) Across a selection of tests, a simple stepladder model better represents the internal structures of warming than a simple trend – strong evidence that the climate system is exhibiting complex system behaviour on decadal timescales. This model indicates that *in situ* warming of the atmosphere does not occur; instead, a store-and-release mechanism from the ocean to the atmosphere is proposed. It is physically plausible and theoretically sound. The presence of steplike – rather than gradual – warming is important information for characterising and managing future climate risk.

Introduction

The climate research community conducts two separate narratives describing how the atmosphere warms under the influence of increasing greenhouse gases: one focused on methods and the other on theory (Jones, 2015b, a). The method-focused narrative describes how model and observational data should be analysed and used in detection and attribution, projection and forecasting. The theory-focused narrative describes how the climate system changes over multiple timescales. These narratives are usually articulated separately and are often at cross purposes with each other. Although they both recognise the climate as having linear and nonlinear components, one treats them as being separate, whereas the other explores the possibility that the two interact.

The two main hypotheses that describe the interactions between climate change and variability over decadal timescales are (Corti et al., 1999;Hasselmann, 2002):

1. Externally-forced climate change and internally-generated natural variability change independently of each other.
2. They interact, where patterns of the response project principally onto modes of climate variability (Corti et al., 1999) or form a two-way relationship (Branstator and Selten, 2009).

These two hypotheses have very different outcomes for the characterisation of climate-related risk (Jones et al., 2013). The methods-focused narrative centres on the use of a signal-to-noise model using ordinary least-squares trend analysis (e.g., North et al., 1995;Hegerl and Zwiers, 2011;Santer et al., 2011). This dominates climate practice and has led to the construction of the gradualist adaptation narrative, which describes adaptation as an incremental series of adjustments over time (Jones et al., 2013). However, if the internally and externally forced components of climate interact, producing steps, shifts or jumps, adaptation planning based on gradual change will lead to risk being underdetermined. *H1* is the default assumption, but according

to the latest Intergovernmental Panel on Climate Change report, the choice between the two remains unresolved (Kirtman et al., 2013).

This paper explores the potential for *H2* to be true by applying severe testing principles (Mayo and Spanos, 2010) to detect and analyse step changes in temperature data. The bivariate test of Maronna and Yohai (1978) is used to test regional and global surface air temperature, global satellite temperature of the lower troposphere and global mean temperature from the CMIP3 and CMIP5 climate model archives.

Analytic Framework

Reasoning by statistical inference

The gradualist narrative is a product of induction – if warming is accepted as following a trend, induction leads to the assumption that warming is gradual. Induction also suggests that gradual increases in radiative forcing will lead to gradual warming. As an analytic process, induction is reasoning by inference, which can be by analogy, statistical inference (often using probabilities) or induction to a particular (if all A have been B, then the next A will be B, Curd and Cover, 1998). The application of linear trend analysis to atmospheric warming is invariably justified as inference to the best explanation. However, the latter stance has been criticised, because using parsimony and likelihood tests to select a structural model is inferior to tests that apply experimental reasoning and are statistically suited for the problem in question (Mayo, 1996;Mayo and Spanos, 2010;Spanos, 2010). This is particularly relevant for complex systems exhibiting intrinsic nonlinear behaviour.

These authors argue that conditions for severe testing should be probative, rather than relying on a particular probability threshold. If test *T* has no likelihood of finding flaws in *H*, then it is not a good test. Mayo (1996) calls this the fallacy of acceptance: no evidence against the null is interpreted as evidence for it, and evidence against the null is interpreted as evidence for an alternative.

Consistent with this, Mayo and Spanos (2011) advise care in distinguishing between the error statistic and the probability of confirmation – likelihood tests passing criteria such as *H* if pH_0<0.05 run the risk of being interpreted as addressing scientific hypotheses with the same level of confidence. This is a weakness of ordinary trend analysis; although a trend may register a low probability of meeting the null hypothesis, it does not infer that the data forms a smooth trend, only that the residuals are normally distributed when the trend is removed. Trend statistics assume no history, whereas in physical systems, process is often one way, which should be considered in any testing environment.

Inference to the best explanation can be rescued if it passes the test: “no available competing hypothesis explains a fact as well as *H* does” (Musgrave, 2010). Mayo (2005) provides criteria for severe testing that is even stricter: Data *x* in test *T* provide good evidence for inferring *H* to the extent that hypothesis *H* has passed a severe test with *x*. The severe testing of not-*H* is part of this severity, meaning that all other possibilities must be exhausted before *H* can be accepted.

Cox and Mayo (2010) distinguish between probabilistic and behavioural reasoning – the first explores which levels of probability are appropriate for making a specific inference – when do data provide good evidence for *H*? Behavioural reasoning will prompt a particular decision based on a given probabilistic position being met. For example, if a test achieves pH_0<0.05 a hypothesis may be considered ‘proven’. Probabilistic reasoning is suitable for complex situations where there is no single cause or effect and there is no easy way to distinguish different types of error when analysing these. For example, a test might provide a particular *p* value but might not represent all of the phenomena of interest. This is relevant to the analysis of temperature records.

Development of severe testing

A hypothesis *H* passes a severe test *T* with data *x* if (Mayo and Spanos, 2010):

1. *x* agrees with *H* and,
2. with very high probability, test *T* would have produced a result that accords less well with *H* than does *x*, if *H* were false or incorrect.

Here, we are using the substantive null of model adequacy approach described by Mayo and Cox (2010) – specific results are used to provide theoretical evidence that can relate statistical tests with a scientific null H_0. Rather than *H* and H_0, rival hypotheses *H1* and *H2* are represented by statistically distinct models. This structure implies two stages of data and hypothesis development. The first stage considers the original temperature data from observations or models and its link with hypotheses *H1* and *H2*. and the second stage is that produced by various statistical tests. The two hypothesis stages are scientific hypotheses (*H*; Mayo calls these theories, but here they are applied as hypotheses) and statistical hypotheses (*h*). The latter are subject to Type I and Type II errors. Sometimes these error types are used to justify scientific hypotheses but one is linked to statistical error and the other is probative, so they should be kept distinct.

Scientific hypotheses need to be examined closely in order to identify the probative conditions required to be built into one or more statistical tests for severe testing. These are then subject to error-based statistical testing to ensure that both *H* and *-H* and *h* and *-h* have sound and testable bases. This very quickly becomes complicated:

- Type I statistical errors – incorrectly rejecting the null (i.e., wrongly accepting a statistical hypothesis).
- Type II statistical errors – failing to reject the null (i.e., failing to accept a statistical hypothesis).
- *H1* hypothesis – The climate signal is gradual and variability is noise, therefore the signal is gradual and follows a (probably monotonic) trend. Non-linear changes in climate data are due to internal climate variability and are random. The hypothesis is confirmed once a trend is judged as significant.
- *H1* null hypothesis – a trend has not emerged from the variability.
- *H2* hypothesis – externally forced and internally generated climate processes interact with each other producing a nonlinear climate signal on decadal timescales.
- *H2* null hypothesis – the climate change signal is gradual and any nonlinearity is of internal origin (*H1*).

This structure is asymmetric and points to the fact that *H1* has never really been severely tested – the usual *H1* null hypothesis is that a trend has not emerged beyond a level of significance. Although alternative structures have been explored including steps and segmented trends (Seidel and Lanzante, 2004), the results have been inconclusive, and the exact nature of change on decadal timescales is still an open question (Trenberth, 2015).

Slingo (2013) describes it thus:

> *Statistical model comparisons … do not allow the identification of a 'true' model, only the computation of the relative likelihood of the models considered. Additionally, these statistical models do not take into account the physics of the climate or any of the known external influences that affect global temperatures. In an analysis … if all the tested models are poorly specified then even the most likely of the tested set of models will be a poor representation of the behaviour of the real climate. In such cases the relative likelihood of the models considered is of little scientific value.*

Another complication is that nonlinear change has been used to challenge global warming theory on the basis that if observed change is not gradual, therefore *H1* is false, and climate change is either disproven or overstated (Jones and Ricketts, 2016). Evidence of nonlinear change, such as step change, is therefore associated with challenges to global warming theory (e.g., see Skeptical Science, 2015).
This asymmetry in null hypotheses means that if *H2* was to be accepted or rejected on the basis of severe testing, severe testing is needed for both *H1* and *H2*.

Development of key statistical tests

Understanding how the global atmosphere warms over decadal timescales is an inverse, ill-posed mathematical problem, similar to that for borehole or satellite temperatures (e.g., Aires et al., 2001;Şerban and Jacobsen, 2001). Atmospheric warming is the product of external forcing and internal variability subject to some undetermined delay between forcing and response. Global mean temperature also integrates regional signals, so is spatially complex – if responses to forcing are regional occurring in different places at different times, then the global signal will be a product of these changes. Global average warming is a complex timeseries that exhibits both linear and nonlinear characteristics.

External forcing is boundary-limited which implies long-term linear behaviour (Lorenz, 1975), and internal variability is nonlinear – both points are agreed to under *H1* and *H2*. The major question concerns how these combine. The long-term signal is invariably measured using trend analysis, however, the cut-off between what constitutes short- and long-term timescales is subjective, being a product of the test method and the question being posed (Jones and Ricketts, 2016).

The accumulation of heat energy through radiative forcing is additive, whereas its entrainment into the hydrothermal system is nonlinear (Ozawa et al., 2003;Lucarini and Ragone, 2011;Tsonis and Swanson, 2012), favouring *H2*. The presence and nature of change points in the warming record and in climate model output will indicate whether this could be the case. Change points in climate data are increasingly being located by a variety of different techniques (Rodionov, 2005;Reeves et al., 2007;Tsonis et al., 2007;Overland et al., 2008;Hope et al., 2010;Killick et al., 2010;Beaulieu et al., 2012;Fischer et al., 2012;Jones, 2012;Jandhyala et al., 2013;Ruggieri, 2013;Cahill et al., 2015).

The hypotheses *H1* and *H2*, imply at least two structural models. For *H1*, close adherence to a warming trend implies that the atmosphere warms gradually. If so, this must occur via either of the following two processes or a combination:

1. Radiatively-forced warming occurs in the atmosphere and the heat generated is retained there (*in situ* warming). Statistically, this would manifest as gradually increasing temperatures, especially over land. It would also imply a trend in lower troposphere satellite temperatures as the airmass warms gradually.
2. Most of the heat generated by added greenhouse gas forcing goes into the ocean and is gradually released into the atmosphere. Again, this would imply gradual warming, especially over the oceans, with the land following suit, but with greater variation if decadal changes in shallow and deep-ocean mixing of heat are taken into account. Discussions in the literature are not clear as to whether the oceanic component is due to varied take-up or release of heat from the ocean.
3. If both 1 and 2 are operating, then the warming rate in the atmospheric component would be monotonic (proportional to forcing) and the contribution from the ocean governed by interannual and decadal variability. This would be best represented by a segmented trend if decadal-scale deep and shallow ocean mixing of heat is a key factor.

For *H2*, we build on the hypothesis that patterns of forced response may either project onto modes of climate variability or that they interact. These manifest as step changes in temperature data. Currently, if such changes are detected, they are routinely attributed to climate variability. The possibility that they may be a response to external forcing has not been severely tested (Jones and Ricketts, 2016).

If warming is mediated by the hydrothermal ocean-atmosphere system, it could be entrained by the nonlinear processes involved (Lucarini and Ragone, 2011). Lucarini and Ragone (2011) describe this as the generation of entropy, as moist static energy is transformed into mechanical energy like a heat engine. This could flip between different states, modulated by Lorenzian 'strange attractors' as described by Palmer (1993). Increased forcing may increase entropy, making the climate system more energetic and less efficient, speeding up these processes over time (Lucarini and Ragone, 2011), much like a bubbling pot.

Given that this entrainment would be into nonlinear modes of climate variability, we characterise it as rapid changes or steps often associated with regime change (see also Tsonis and Swanson, 2012). This is essentially a store and release process, where heat stored in the ocean is released when the ocean-atmosphere relationship becomes unstable, precipitating regime change. These ideas are discussed in more detail in Jones and Ricketts (2016). To date, we have identified widespread step changes in a range of climate variables, most notably temperature (Jones et al., 2013), and carried out attribution studies for one region, south-east Australia (Jones, 2012). The exploration of *H2* is to therefore detect step changes more broadly and to contrast this with trend-like behaviour.

Given that this entrainment would be into nonlinear modes of climate variability, we characterise it as rapid changes or steps often associated with regime change (see also Tsonis and Swanson, 2012). This is essentially a store and release process, where heat stored in the ocean is released when the ocean-atmosphere relationship becomes unstable, precipitating regime change. These ideas are discussed in more detail in Jones and Ricketts (2016). To date, we have identified widespread step changes in a range of climate variables, most notably temperature (Jones et al., 2013), and carried out attribution studies for one region, south-east Australia (Jones, 2012). The exploration of *H2* is to therefore detect step changes more broadly and to contrast this with trend-like behaviour.

The challenge for severe testing is to create tests that can distinguish between *H1*, *-H1*, *H2* and *-H2* with sufficient confidence to select one over the other. The following six probative tests have been identified to test the relationship between linear and nonlinear behaviour and their responses to external forcing:

1. Stratified analysis of change points: the timing and distribution of change points and their relationship with known regime changes and with each other. Change points aligning with known nonlinear processes indicate a causal link.
2. Identification of similar patterns of steps between observations and physical models indicates a physically coherent origin, rather than random stochasticity.
3. Using internal trends and shifts (steps minus internal trends) to estimate the gradual and rapid warming components in a record, and testing each of these against criteria such as total warming and equilibrium climate sensitivity (ECS) in observations and models separately.
4. Tests to diagnose stationarity – using a simple linear inverse model to measure the emergence of an anomalous signal from the background noise of variability, and whether it is gradual or steplike.
5. Testing of other variables including rainfall, sea surface temperatures, sea level rise, and air pressure, to see whether they undergo similar changes.
6. Statistical models applied to test underlying step- and trend-like structures in the data.

Method

A multi-step and rule-based application of the Maronna-Yohai bivariate test (MYBT, Maronna and Yohai, 1978) has been developed to expand the original test beyond the detection of single steps (Ricketts and Jones, 2016, see Supplementary Information for details). Previously, the bivariate test has been used to detect inhomogeneities in climate variables (Potter, 1981a;Bücher and Dessens, 1991a;Kirono and Jones, 2007;Sahin and Cigizoglu, 2010), decadal regime shifts in climate-related data and step changes in a wide range of climatic timeseries (Buishand, 1984;Gan, 1995;Vivès and Jones, 2005;Boucharel et al., 2011;Jones, 2012;Jones et al., 2013). The main purpose of automating the test is to improve its objectivity and robustness by using a predefined set of rules.

The test adapts the formulation of Bücher and Dessens (1991a) and tests a single serially-independent variate (x_i) against a reference variate (y_i) using a random timeseries following Vivès and Jones (2005). The important outputs of the test in a timeseries of length N are, (1) The T_i statistic which is defined for times $i < N$, (2) the T_{i0} value which is the maximum T_i value, (3) i_0, the time associated with T_{i0}, (4) shift at that time, and (5) p, the probability of zero shift. Note that i_0 is the last year prior to the change. In this paper, we routinely give the year of change.

A single timeseries analysis consists of a *screening pass*, followed by a *convergent pass*. In both passes, we apply a *resampling test* to each segment being examined, where the test is repeated 100 times, resampling the random number reference series. The screening pass starts from the most significant shift in a timeseries, determined using the resampling test and, if $p<0.01$, the series is divided into shorter timeseries either side of the step and these are tested until all steps have been detected. As such, it is a recursive procedure whereby the first steps detected may be influenced by as-yet-unlocated steps. The convergent pass then serially refines these segments to provide a causal sequence. The convergent process is repeated until a stable set of step changes is produced.

The above analysis is run 100 times. Ricketts and Jones (2016) show that this procedure may produce several different but related solutions (solution = set of change dates); the most common solution is returned as the best estimate. Alternatives often indicate the presence of localised events embedded in larger scale areally-averaged data. The majority solution is selected for further analysis. Most historical temperature records analysed contain one or two stable configurations for surface temperature and zero or one for satellite temperature. Climate model data may produce a larger number of stable solutions, especially the higher forcing scenarios.

Mean annual data for observations is considered serially independent – and in most cases applied in the paper, the MYBT is reliable. Deseasonalised quarterly and monthly data can be used to locate a shift within a year, but is not serially independent, so is used here in combination with the t-test either side of the change date to assess significance. A resampling test that shuffles data either side of a shift will also indicate whether a change point is abrupt, or the timeseries is trend-like. Twenty-first century model data is not serially independent under high rates of forcing, an issue discussed in Sect 4.3. For error testing using statistical hypotheses, we routinely use behavioural reasoning at levels of $p<0.01$ for the bivariate test (exceptions are noted), and non-significant (NS, $p>0.05$), $p<0.05$ and $p<0.01$ for trend analysis and the t-test.

Local attribution of step changes uses a technique detailed in Jones (2012). The basic methodology is suitable for continental mid-latitude areas where annual average maximum temperature (*Tmax*) is correlated with total rainfall (*P*), and minimum temperature (*Tmin*) is correlated with *Tmax* (Power et al., 1998;Nicholls et al., 2004;Karoly and Braganza, 2005). For Central England Temperature, a largely maritime climate, diurnal temperature is assessed against precipitation instead of *Tmax*. The method uses the following steps:

1. Homogenous regional average data is obtained for *Tmax*, *Tmin* and *P*.
2. A period of stationary climate is calculated by testing when the relationship between *Tmin* and *Tmax* undergoes a statistically significant step change. The relationship between *Tmax* and *P* will change at the same, or later date.
3. Linear regressions are calculated between each pair (*Tmax/P* and *Tmin/Tmax*) for the stationary period.
4. Externally forced warming is estimated for the non-stationary period using these regressions.
5. The results are tested for step changes.

Timeseries tested here are mean annual global air temperature anomalies from five groups (NCDC, Peterson and Vose, 1997;GISS, Hansen et al., 2010;HadCRU, Morice et al., 2012a;BEST, Rohde et al., 2012;C&W, Cowtan and Way, 2014a), hemispheric temperatures from three groups (HadCRU, NCDC and GISS) and zonal temperatures from two groups (NCDC and GISS) to see how prevalent step changes are, whether they coincide across different records and to investigate the relationship between step changes and trends. Tropospheric satellite temperatures from two groups (UAH, Christy et al., 2003;Christy et al., 2007; RSS, Mears and Wentz, 2009a) are also tested. The specific records used are detailed in the Supplementary Information.

Simulated mean global surface temperature from the CMIP3 and CMIP5 climate model archives is also tested. The analysis is carried out in two parts. The first part investigates simulated 20th century temperatures to determine how well the models reproduce the pattern of step changes in the observed data. The second part analyses how step changes evolve over the 21st century under the different Radiative Concentration Pathways (RCPs).

Comparing steps and trends creates two statistical hypotheses, h_{step} and h_{trend}. Conventional trend analysis will extract a monotonic signal from the data. Measurement of change where nonlinear behaviour is present is not an exact process. The bivariate test measures total change between segments of a timeseries, ignoring any trend that may be present. These we refer to as steps. However, trends may also be present within timeseries registering steps, so internal trends are calculated between steps. The distance between the end of one trend and the start of the next is referred to as a shift. The process of calculating steps then trends, we call the step and trend model. Steps, internal trends and shifts all provide data for severe testing.

Shifts and internal trends are not strictly additive – summed over a number of steps they can add up to more or less than the change in temperature measured between the beginning and end of a series. These differences are largest in records containing reversals and negative trends.
The main phenomena analysed are:

- Steps – measurement of the whole change between two timeseries assuming stationarity as produced by the bivariate test. This assumes no trend either side of the step.
- Internal trends – measurement of the trends between steps.
- Shifts – measure of the internal step between the end of a preceding trend and the beginning of the next trend.
- Trend/step ratio – the ratio between total internal trends and total steps in a multi-step timeseries.
- Trend/shift ratio – the ratio between total internal trends and internal shifts (steps minus trends).

Results – observations

Global and zonal temperatures

This section undertakes global, hemispheric and zonal analysis to determine temporal and spatial patterns of step changes in observed temperature, consistent with tests 1 and 2.

Step changes meeting the $p<0.01$ threshold in global and zonal temperatures show a great deal of structure over the 1880–2014 time period. All series were tested from their earliest recorded date (1850 and 1880) and results from 1880–2014 are shown. Downward steps occur in the late 19th and early 20th century, upward steps between 1912 and 1938 with one downward step in 1964. From 1968, upward steps dominate, with one exception in the high southern hemisphere (SH) latitudes in 2007 (Fig. 1).

The 1997 step change is global, with some regional steps occurring in 1996 and 1998. A global step change occurs 1979/80; also registering in many regions, except the northern hemisphere mid and high latitudes. All other step changes occur across more limited regions, with some being confined solely to land or to ocean. The 1997 step is the largest at 0.31±0.01 °C. The 1979/80 step is the next largest at 0.22±0.03 °C. The greater variation in size of 1979/80 is affected by the timing and size of previous steps and trends. In the first half of the 20th century, three global records show positive steps in 1920/21 and in 1937, and two in 1930 (Fig. 1). The GISS record also shows a downward step in 1902, coinciding with the northern hemisphere (NH) ocean, tropics and southern hemisphere. The two groups are based on the early 20th century differences: GISS, BEST, C&W in one group and HadCRU and NCDC in the other. The anomaly averaged from all five records shows upward step changes in 1930, 1979 and 1997, coinciding with the HadCRU and NCDC records.

Differences emerge between ocean and land records. The global HadSST (HadCRU) record shifts in 1937, 1979 and 1997, whereas the ERSST (NCDC) record shifts in 1890, 1930, 1977, 1987 and 1997. Global land records from both CRU and NCDC shift in 1920/21, 1980 and 1997. Northern hemisphere land and ocean step changes are consistent across three records: in 1924/25, 1987 and 1997. The NH ocean shows a downward step in 1902/03 and is less consistent between the two records tested for subsequent upward steps. The SH is consistent across 1937, 1979 and 1997, with two records showing a downward step in 1890 and an upward step in 1969.

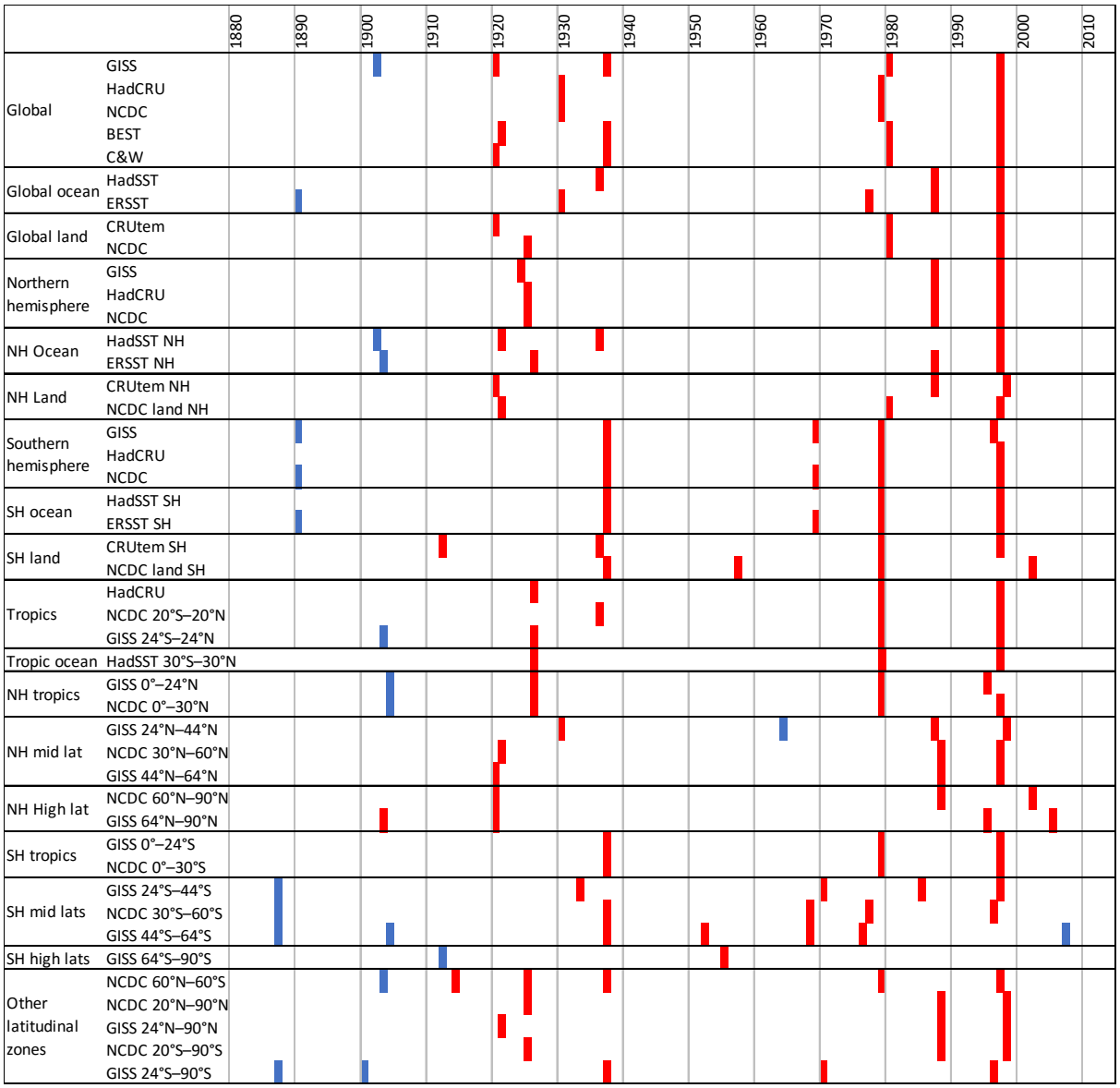


Figure 1: Dates of statistically significant step changes (p<0.01) 1880–2014, for a range of mean annual temperature records. Downward steps are blue and upward red. Records are sourced from Goddard Institute of Space Studies (GISS), the Hadley Centre and Climate Research Unit: HadCRU (land and ocean), HadSST (ocean), CRUtem (land), National Climatic Data Center: NCDC (land, land and ocean), ERSST (ocean), Berkeley Earth Surface Temperature (BEST) and Cowtan and Way (C&W). See Supplementary Information for details.

The tropics show a downward step in 1902/03, and upward steps in 1926, 1979 and 1997. Three NH mid-latitude records step upwards in 1920, 1921 or 1930, in 1987/88 and 1997/98. One zonal record also shows a downward step in 1964. The two NH high latitude records show a single downward step in 1902 and in 2005, both step upwards in 1921 and 1994 and a single step upwards in 2005. The three SH mid-latitude records show a downward step in 1887 and one in 1902, and upward steps in 1933 or 1937, 1968 or 1970, 1977/1978 or 1984, and 1997 or 1998. SH high latitude data is not very reliable, being absent for NCDC 60°S–90°S. The GISS 64°S–90°S average anomaly steps downward in 1912 and an upward in 1955.

Fig. 2 shows the internal trends and their error significance for the five global mean temperature records. Steps and trends are consistent for the last two periods 1979/80 to 1996 and 1997 to 2013/14, but diverge in the middle of the record, due to differences in the timing and magnitude of steps and accompanying internal trends. Data quality may be an issue in the earlier parts of the record. For example, the version of GISS data

used here shows five steps in 1902, 1920, 1937, 1980 and 1997, whereas a previous version to 2013 stabilised on steps in 1930, 1979 and 1997, consistent with the average anomaly of all five records. This indicates that the timing and magnitude of steps in the early 20th century can be influenced by adjustments made to improve data quality. However, all global step change dates coincide with regional steps, showing that while the relative importance of dates associated with step changes may be different, the dates themselves are quite stable. This gives us added confidence we are not detecting false positives.

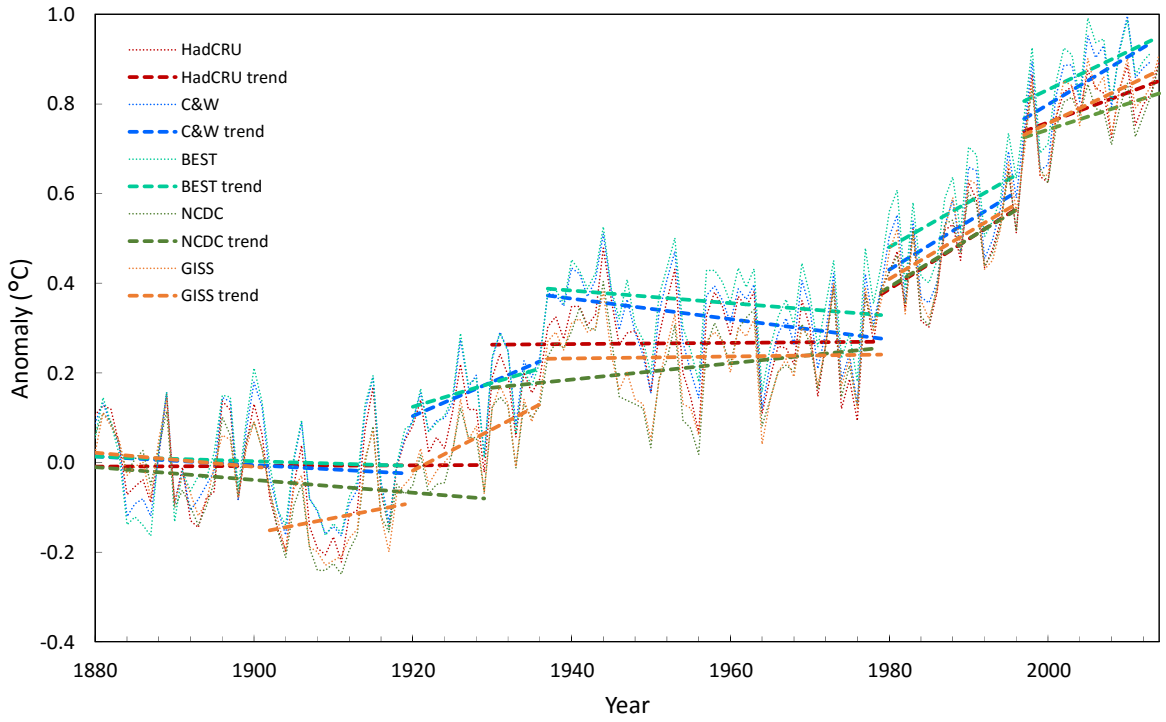
Internal trends are mainly non-significant in the early record, the exception being the GISS 1920–37 period. The 1979/80 to 1996 trend is significant at the $p<0.01$ level in two records (HadCRU and NCDC) and $p<0.05$ in the other three records. The NH step change in 1987 seen in all three records tested strongly influences this trend, which is examined further in the next section. The post-1997 period is non-significant in two records and trends at $p<0.05$ in three records.

There is no objective way to partition shifts and internal trends. Giving the first preference to internal trends in calculating ratios provides the criteria for the severe testing of non-linear responses. As some of the internal trends show $p>0.05$, this is a conservative stance preferencing the methodological status quo. Expressed as a ratio between internal trends and steps, four global records range between 0.32 and 0.38 with the GISS record yielding a ratio of 0.62 due to the cool reversal in the early 20th century. For trends and shifts, the ratio ranges between 0.44 and 0.58 with the GISS record an outlier at 1.38.

Test 2 aims to determine whether at the regional level, trends or steps are more prominent than at the global scale. The global trend/step ratio for the HadCRU record, for example, is 0.55 (0.30 °C/0.55 °C), for the NH is 0.31, the SH 0.28 and the tropics (30°N–30°S) is 0.33; close to the average of the two hemispheres. When divided into land and ocean, the HadCRU and NCDC records, show 0.90 and 1.15 for land, and 0.16 and 0.26 for ocean, respectively, showing the oceans to be more steplike and the land having roughly equal measure. SH ocean is very steplike (0.16) and SH land, less so (0.39). The mid-latitudes are also very steplike as is the tropical ocean. High ratios (>1) often involve a temporary cool reversal around the early 20th century.

This is also the case for single steps on a regional basis. In 1997/87 the global shift was 0.16±0.01 °C, a ratio of about 50% compared to the step change of 0.32 °C. For the northern hemisphere, this ratio varied between 57% and 68% for three land and three ocean data sets. For the northern hemisphere mid-latitudes, land and ocean from two data sets (NCDC 30°N–60°N, GISS 24°N–44°N), steps/shifts measure 0.43 °C/0.44 °C, close to a 1:1 ratio, indicating no trend.

The more steplike character of both the oceans and the mid-latitudes is consistent with those areas being the loci of change in terms of decadal regimes and nonlinear equator-to-pole transport. This is inconsistent with the hypothesis of gradual warming Varying shift dates and rates of change will contribute to the global record being more trend-like than individual regions.



HadCRU			C&W			BEST			NCDC			GISS		
Period	Change	Trend Sig.	Period	Change	Trend Sig.	Period	Change	Trend Sig.	Period	Change	Trend Sig.	Period	Change	Trend Sig.
												1880–1901	-0.03	NS
1850–1929	-0.01	NS	1850–1919	-0.06	NS	1850–1920	-0.05	NS	1880–1929	-0.07	NS	1902	-0.13	
1930	0.26		1920	0.16		1921	0.15		1930	0.26		1902–1919	0.06	NS
												1920	0.18	
			1920–1936	0.12	NS	1920–1936	0.09	NS				1920–1936	0.15	**
			1937	0.16		1937	0.19					1937	0.18	
1930–1979	0.01	NS	1937–1979	-0.10	*	1937–1979	-0.06	NS	1930–1978	0.09	NS	1937–1979	0.01	NS
1980	0.20		1980	0.19		1980	0.20		1979	0.26		1980	0.26	
1980–1996	0.19	**	1980–1996	0.17	*	1980–1996	0.16	*	1979–1996	0.18	**	1980–1996	0.17	*
1997	0.32		1997	0.33		1997	0.31		1997	0.30		1997	0.31	
1997–2014	0.11	NS	1997–2013	0.17	*	1997–2013	0.14	*	1997–2014	0.10	NS	1997–2014	0.14	*
Trend	0.30			0.31			0.27			0.30			0.49	
Step	**0.78**			**0.85**			**0.86**			**0.82**			**0.80**	
Shift	0.55			0.59			0.63			0.54			0.36	
Trend/step	0.38			0.37			0.32			0.36			0.62	
Trend/shift	0.55			0.52			0.44			0.56			1.38	

Figure 2: Mean global anomalies of surface temperature with internal trends. The annual anomalies (dotted lines) from five records (HadCRU, C&W, BEST, NCDC, GISS) are taken from a 1880–1899 baseline. Internal trends (dashed lines) are separated by step changes detected by the bivariate test at the $p<0.01$ error level. The size of each step (in red) and change in temperature of each internal trend (in black) is shown in the figure table along with its significance, where NS is $p>0.05$, * is $p>0.01<0.05$, ** is $p<0.01$. Totals of trends, steps, shifts (change from one trend to the next) and ratios are also shown.

Satellite era records

A comparison of surface and satellite temperatures stratifies records according to altitude and source of measurement. Satellite records of lower troposphere temperatures sourced from the RSS and UAH records beginning in December 1978, were analysed for step changes (1979–2014). Anomalies were investigated annual and seasonally. Annual mean global and zonal temperatures show 1995 and 1998 as the two main shift dates, with 1995 more prominent at the global scale (Table 1). For individual seasons, steps in 1995 are dominated by the NH JJA and SON periods, especially on land. This can be traced back to warm El Niño conditions in 1994/5. For the quarterly timeseries (4 seasons x 36 years), the JJA and SON quarters of 1997 dominate the UAH global record, less so for the RSS record.

Quarterly anomalies for the RSS and UAH satellite and HadCRU and GISS surface mean global temperature were compared for similarities and differences. Quarterly timeseries are affected by autocorrelation due to the

El Niño-Southern Oscillation (ENSO), for the bivariate test making results robust for timing but not significance. Student's t-test (two sided, unequal variance), was used as a back-up.

RSS shifts in DJF 1987/88 by 0.11 °C (p<0.05 MYBT and p<0.1 t-test) and UAH shifts in DJF 1987/88 and 0.09 °C (NS MYBT and p<0.05 t-test). For surface temperature, HadCRU and GISS shift in JJA 1987 by 0.14 °C and 0.15 °C, respectively (p<0.01, both tests). On an annual basis, the bivariate test registers 1987/88 at the p<0.05 level. The lower significance in the satellite records is due to the slightly lower shift size and higher variance. RSS shifts in JJA 1997 by 0.23 °C, UAH shifts in DJF 1997/98 by 0.26 °C, HadCRU in JJA 1997 by 0.26 °C and GISS in SON 1997 by 0.25 °C (all p<0.01, both tests). These four data sets show consistent shift dates in 1997 and similar shift dates in 1986/7, showing that the significant step change in the NH is present at the global scale. This suggests that the period of accelerated trend noted by many for 1976–1998 (e.g., Trenberth, 2015) is actually a period containing two step changes, one global (1979/80) and one regional (1987/88).

Table 1. Dates of step changes for lower tropospheric satellite temperature anomalies, with annual timeseries and quarterly breakdowns in parentheses (DJF, MAM, JJA, SON), and quarterly timeseries. Data sources are Remote Sensing Systems (RSS) and University of Alabama, Huntsville (UAH).

Region	Annual timeseries (quarterly breakdown)		Quarterly timeseries	
	RSS	UAH	RSS	UAH
Global land & ocean	1995 (98,98,95,95)	1995 (97,98,94,95)	JJA 1997	SON 1997
Global land	1995 (95,98,95,95)	1998 (98,98,94,95)	SON 1994	SON 1997
Global ocean	1998 (98, - ,97,95)	1995 (97, - , - ,95)	JJA 1997	SON 1997
NH land & ocean	1995 (98,98,94,94)	1998 (98,98,94,94)	JJA1997	SON 1997
NH land	N/A	1998 (98,98,98,98)	N/A	JJA 1997
NH ocean	N/A	1994 (- , - , - ,94)	N/A	JJA1997
SH land & ocean	1995 (98, - , - ,95)	1995 (97, - , 87,95)	SON 1997	SON 1997
SH land	N/A	1995 (95, - , 91,95)	N/A	MAM 2002
SH ocean	N/A	1995 (97, - , - ,95)	N/A	DJF 1998
Tropics land & ocean	1995 (- , - , - ,93)	- (- , - , - ,95)	JJA1997	JJA1997
Tropics land	1995 (- , - , - ,87)	1995 (98, - ,95,95)	SON 1997	JJA1997
Tropics ocean	1995 (- , - , - ,95)	- (- , - , - , -)	JJA 1997	-
NH ex-trop land & ocean	1998 (95,98,98,94)	1998 (98,98,98,94)	SON 1997	DJF 1998
NH ex-trop land	1998 (- ,98,94,94)	1998 (- ,98,98,98)	MAM 1994	DJF 1998
NH ex-trop ocean	1998 (99,98,98,94)	1994 (02,98, - ,94)	SON 1997	MAM 1998
SH ex-trop land & ocean	1998 (96, - , - ,95)	1996 (97, - , - ,95)	DJF 1998	DJF 2001
SH ex-trop land	1995 (- , - , - , -)	2001 (03, - , - ,02)	JJA 1995	MAM 2002
SH ex-trop ocean	1998 (96, - , - , -)	1996 (97, - , - ,95)	DJF 1998	DJF 1998
N polar land & ocean	1995 (03,95,98,95)	1995 (05,95,98,95)	DJF 2000	MAM 1998
N polar land	1995 (- ,94,98,95)	1995 (- ,89,98, -)	DJF 2005	MAM 2000
N polar ocean	1995 (03,05,98,95)	1995 (05,95,98,95)	MAM 2002	MAM 1998
S polar land & ocean	-	-	-	-
S polar land	-	-	-	-

When all four records are plotted on a common baseline of 1979–1998, the surface and satellite temperatures display similar shifts but different internal trends (Fig. 3). Shown this way, the supposed differences between surface and satellite trends are largely removed. The satellite data contain 'significant' negative internal trends over 1979–1986 (RSS p<0.01, UAH p<0.05), otherwise are p>0.05. The surface data show significant positive internal trends over 1997–2014 (GISS p<0.01, HadCRU p<0.05), otherwise are p>0.05. The decline post 1981 and lower trends in the early 1990s in the satellite data are likely due to volcanic eruptions, which amplify cooling at altitude (Free and Lanzante, 2009). The differences in internal trends post 1996 may be due to orbital decay that has not been fully allowed for in the satellite record, cooling from above affecting the satellite data and heating from below affecting the surface data, or a combination of these.

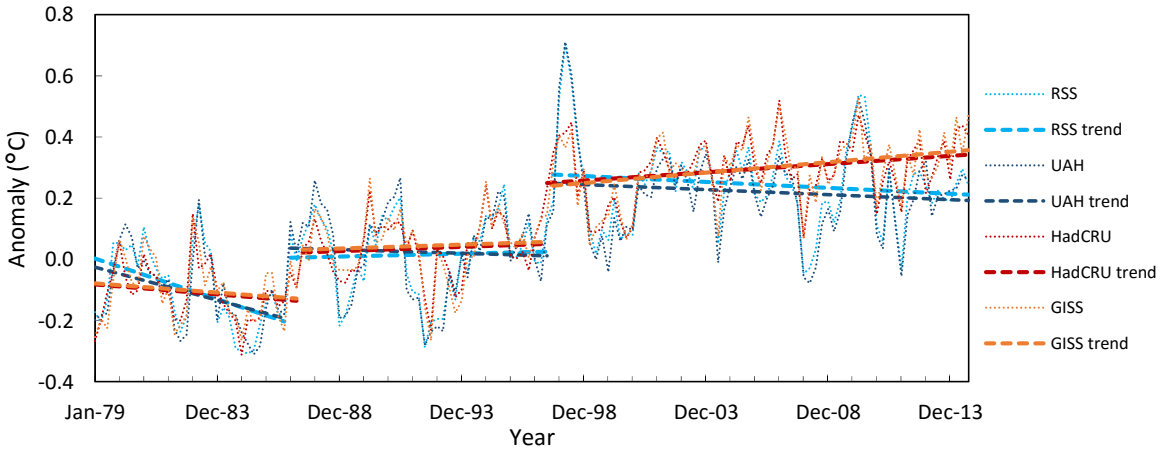


Figure 3: Quarterly mean satellite (RSS, UAH) and surface (HadCRU, GISS) temperature anomalies on a common baseline 1979–2014. Annual anomalies (dotted lines) and internal trends (dashed lines) are separated by step changes.

Unless substantially contaminated by artefacts, these changes do not reflect gradual warming in the atmosphere, but instead may reflect regime-like change controlled from the surface. The capacity for the oceans to emit sufficient heat during El Niño events and absorb it during La Niña to cause large warming anomalies at this scale events suggests that available heat energy is not a limiting factor to abrupt changes.

At this timescale, both surface and satellite temperature records are very steplike. The trend/shift ratios for the HadCRU and GISS records are 0.19 and 0.27 respectively and for the RSS and UAH records are -0.55 and -0.40, respectively, showing the effect of the negative internal trends. Shifts are consequently higher than steps in the satellite data. These are clearly due to the presence of the ENSO cycle within the data and, given the coincidence of step dates with some El Niño events, there is no clear way to allow for these, so the data is analysed and presented as is.

Regional attribution

This section on regional attribution covers the issue of stationarity and the character of change over regional areas. Regional attribution of step changes in annual temperature has previously been carried for south-eastern Australia (SEA, Jones, 2012) and is repeated here for Texas and central England. The methodology is suitable for continental mid-latitude areas where annual average minimum temperature (*Tmin*) is correlated with maximum temperature (*Tmin/Tmax*), and *Tmax* is correlated with total annual rainfall (*Tmax/P*) (Power et al., 1998;Nicholls et al., 2004;Karoly and Braganza, 2005). For maritime areas such as central England, diurnal temperature range (*DTR*) is used (*DTR/P*) instead of *Tmax/P*. The method uses the bivariate method to test the dependent variable against the reference variable. A shift in the dependent variable denotes a regime change.

SEA climate was stationary until 1967 when a step change increased *Tmin* by 0.6 °C with respect to *Tmax* (Jones, 2012). Six independent climate model simulations for the same region become non-stationary by the same means between 1964 and 2003, showing steps of 0.4 to 0.7 °C (Jones, 2012). Texas becomes non-stationary in 1990 with an increase in *Tmin/Tmax* of 0.5 °C. *Tmax* increases by 0.8 °C against *P* in 1998. For Central England, *Tmin* increases against *DTR* by 0.3 °C and *Tmax* against P by 0.9 °C in 1989. *Tmax* also increases against *P* in 1911 by 0.5 °C (Table 2).

The stationary period is used to established regression relationships that calculate *Tmax* and *Tmin* from *P* and *Tmax*, respectively. These regressions are used to estimate how *Tmax* and *Tmin* would have evolved during the non-stationary period. The residual is then attributed to anthropogenic regional warming and is tested using the bivariate test. Here the residuals for *Tmax* and *Tmin* are averaged to estimate externally-forced warming (Tav_{ARW}).

Table 2. Year of non-stationarity in regional temperature for south-eastern Australia, Texas and Central England. Data source, year of first change greater than one standard deviation for *Tmax* against *P* and *Tmin* against *Tmax, or DTR/P* using the bivariate test. The stationary period is also shown.

Data source	*Tmax*/*P*		*Tmin*/*Tmax*		*DTR*/*P*		Stationary Period
	Year	Change	Year	Change	Year	Change	(SEA)
SE Australia	1999	0.7	1968	0.6			1910–1967
Texas	1998	0.8	1990	0.5			1895–1990
Central UK	1989	0.9	N/S		1989	0.3	1878–1988
	1911	0.5					

In SEA, Tav_{ARW} shifts up by 0.5 °C in 1973 (Fig. 4). Similar patterns were found for 11 climate model simulations for SEA, undergoing a series of step changes to 2100 (Jones, 2012). For Texas, Tav_{ARW} shifts by 0.8 °C in 1990. Central England temperature shifts up by 0.7 °C in 1989 and by 0.5 °C in 1911. Using the full record for Central England average temperature from 1659, a significant step change was found in 1920, whereas using a starting date of 1878 identifies 1911. Given that the second mode identified in the longer test is 1911, we conclude the 1911 date is an artefact of the starting date in 1878 and a step change in 1920, consistent with NH data, would register if earlier data were available.

None of the internal trends in Fig. 4 exceed the $p<0.05$ threshold. The trend/shift ratios for *Tav* (not shown in Fig. 4) and attributed to external forcing (Tav_{ARW}) are 0.23 and 0.88, respectively for SEA, 0.45 and -0.53 for Texas and -0.01 and 0.33 for Central England (1878–2014). The lower ratio in SEA Tav_{ARW} is because reduced rainfall post 1997 produces lower attributed $Tmax_{ARW}$ but if that rainfall reduction is also a response to external forcing (Timbal et al., 2010), $Tmax_{ARW}$ will be underestimated. The negative ratio for Texas is because Tav_{ARW} contains negative internal trends, mostly after 1990 (largely a rainfall effect on *Tmax*). For Central England, the ratio for *Tav* has been calculated from the long-term record from 1659, which shows no step changes or trends between 1701 and 1920. Late 20th century warming in both Central England and continental US elsewhere has also been analysed as nonlinear (Franzke, 2012;Capparelli et al., 2013).

These results show that the transition from stationarity to non-stationarity is abrupt for regional temperature at three locations on three continents, and for six independent climate model simulations for one of those locations (SE Australia). The close association of the observed transition in SEA in 1968 with the widespread shift date over the southern hemisphere mid-latitudes indicates that the onset of the warming signal in these broader regions is abrupt (Jones, 2012). The changes in central England in 1989 and Texas in 1990 may also be associated with a widespread step change in the northern hemisphere mid latitudes in 1987/88 (Overland et al., 2008;Boucharel et al., 2009;Lo and Hsu, 2010;Reid and Beaugrand, 2012;North et al., 2013;Menberg et al., 2014;Reid et al., 2015).

The low trend/shift ratios shown for ocean and some zonal areas also occur over the three land areas analysed. This suggests that shifts may be more distinct at regional scales, integrating into a more trend-like global average. This is the case for sea level rise data, where individual tide gauge records exhibit step ladder-like behaviour at individual locations and global mean sea level follows a curve (Jones et al., 2013).

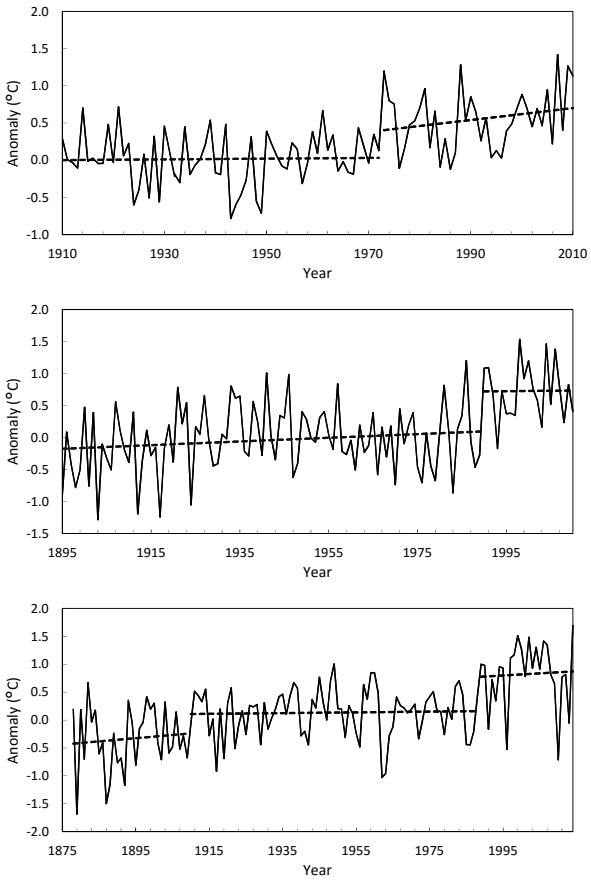


Figure 4: Anomalies of annual mean temperature attributed to nonlinear changes where the influences of interannual variability have been removed for (a) Central England, (b) Texas, and (c) south-eastern Australia. Internal trends (dashed lines) are separated by step changes (p<0.01).

Results – models

20th century simulations (1861–2014)

These sections report on the multi-step analysis of 102 simulations of global mean surface warming from the CMIP3 archive, and 295 simulations from the CMIP5 archive. Further information on the archives is in the SI. The relevant test for models is to identify similar phenomena to observations. Here we describe analyses of the timing of change points and their relationship with known regime changes and the measurement of the relative contributions of steps, shifts and internal trends in the temperature record (part of tests 1 and 3 listed above).

Starting with observations, the percentage of annual steps (p<0.01) in the 45 timeseries of mean annual surface temperature from Fig. 1, are shown in Fig. 5a. Two-thirds of all historical records shift in 1997 and one-third in 1980 and 1937. Lesser peaks of 10–15% occur in 1920, 1921, 1926, 1930, 1968–69, 1987 and 1988. The three shifts in 1979/80, 1987/88 and 1997/98 are the main contributors to the higher rate of trend noted from around 1970. Because these peaks measure how strongly steps occur globally and regionally, percentages denote how pervasive a step is. The models register a significant step at the global scale only, so will only pick up the most extensive step changes – any steps occurring below the assigned level of probability (p<0.01) will show up as part of a trend, as is the case for 1987/88 in the observations.

Fig. 5b shows step changes from the CMIP3 combined SRES scenarios A1B and A2 simulations for the 20th and 21st century: 84 are independent and 18 are ensemble averages. The CMIP3 models were driven by observed forcing including sulphate aerosols to 1999–2000 and not all contain natural forcings (see Table S2). They do a reasonable job of capturing the three main post-1950 peaks. Figs 5c–f show the CMIP5 RCP2.6, RCP4.5, RCP

6.0 and RCP 8.5 ensemble results, respectively. The models were driven by observed forcing, including natural volcanic and solar forcing, to 2005. Visually, the CMIP5 results illustrate the observed peaks and troughs better than CMIP3. This is presumably due to the improved representation of forcing factors and physical processes, and to improved model resolution (Table S3).

The RCP4.5 result (Fig. 5d) with 107 independent members, is the largest multi-model ensemble (MME). The three major post-1950 step changes are reproduced as follows: 55% (58 of 107) of the runs undergo a step change in 1996–98 (17% step in 1996, 16% in 1997 and 22% in 1998), 40% of the runs peak in 1976–78, just missing the observed peak in 1979/80 and 19% peak in 1986–88. In the mid-1970s, the models may be picking up the observed regime shift 1976–77 in the Pacific Ocean (Ebbesmeyer et al., 1991;Miller et al., 1994;Mantua et al., 1997;Hare and Mantua, 2000) as a contemporaneous increase in warming. With weak El Niños affecting observations during 1977–1980 (Wolter and Timlin, 2011), this step change may have been delayed in the observed temperature record until 1979–80.

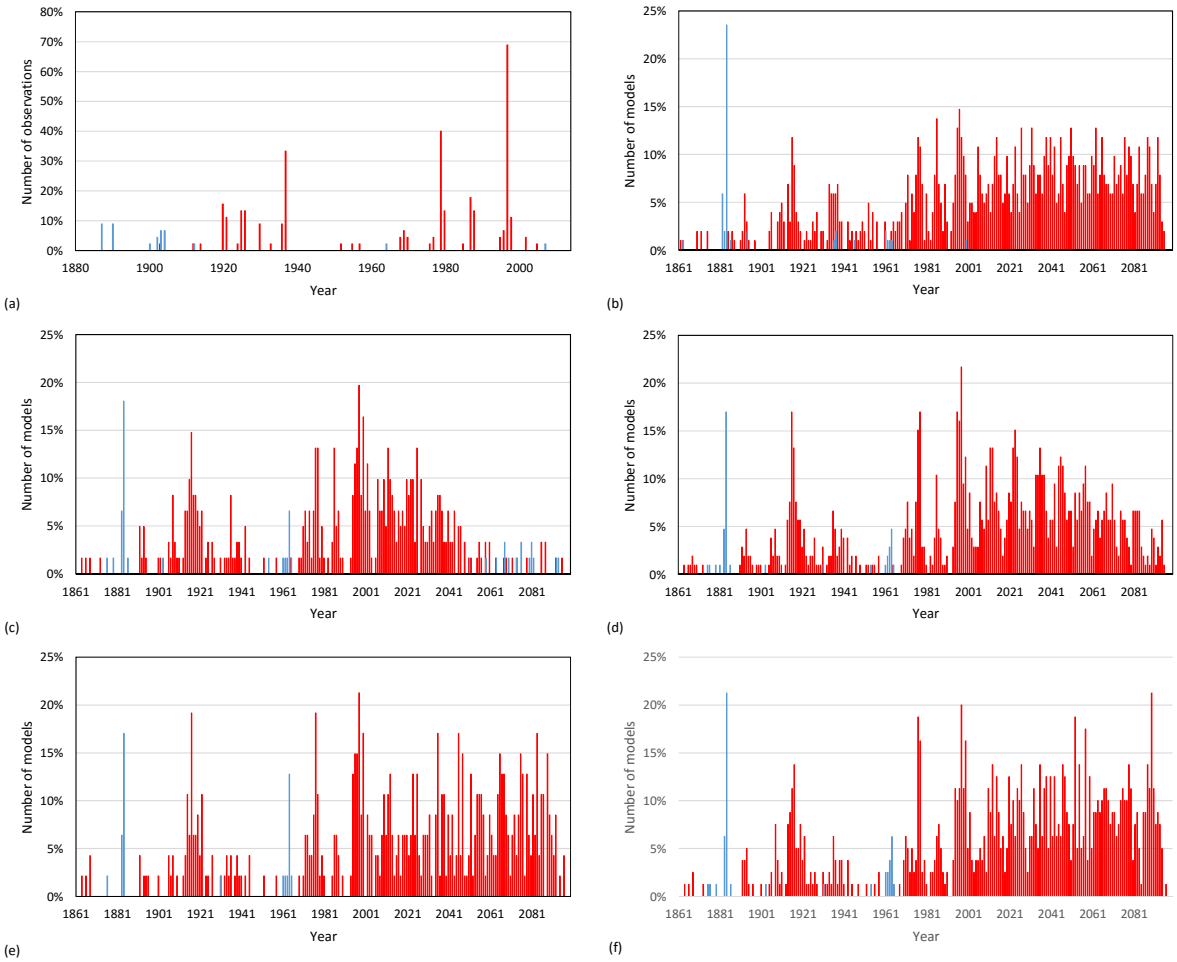


Figure 5: Step changes in observed and simulated surface air temperatures. Frequency in percent of statistically significant step changes from (a) global, hemispheric and zonal averages (45, 1880–2014); (b) global mean warming from 102 model simulations from the CMIP3 archive for SRESA1b and A2 emission scenarios; (c–f) global mean warming 1961–2100 from the CMIP5 archive for the (c) RCP2.6 pathway (61), (d) RCP4.5 pathway (107), (e) RCP6.0 pathway (47) and (f) RCP8.5 pathway (80).

Of the pre-1950 peaks, the models peak around 1916, rather than 1920, and 1936–37 forms a minor peak, less prominent than in the observations. The volcanic eruptions of Krakatoa (1883) and Mt Agung (1963) both feature in the model simulations but less so in the observations. The mid-20th century period of little change is also reasonably well reproduced.

Correlations over the full period 1880–2005 between observations and the CMIP3 and CMIP5 models, are 0.32 and 0.34, respectively (p<0.01). For the period 1950–2005, the correlations rise to 0.45 and 0.40, respectively. If specific events: 1963/64, 1968–70, 1976/77, 1979/80, 1987/88 and 1996–98 are grouped, and all other years analysed individually, then the correlation increases to 0.78 for both CMIP3 and CMIP5 records (note that this treats the simulated and observed peaks in the 1970s separately). We consider this a reasonable test, because all these dates have been linked to regime changes or break points in temperature in the literature. Finessing the exact years involved around these events makes little difference to the result, so the correlation is robust.

Although collectively, the model ensembles reproduce the observed peaks, single models do not fare as well. We experimented with a skill score that worked on scoring matched steps between models and observations, but the resulting scores did not correlate with any other factor. The only event reproduced widely by the models was the 1996–8 step change, peaking in 1997, where 58 of the 107 MME (55%) undergo a step change, although 40% of the MME produces a step in 1976–78.

Relationship between steps and trends over time

Here, we report on the relationships between steps, shifts and trends, the magnitude of warming and ECS to estimate the proportion of signal in each warming component. Total warming over time can be represented by straightforward differencing, change measured from a simple trend and the sum of various components, such as the sum of steps, and of shifts and trends. All come up with slightly different answers, but describe a process that over many decades largely conforms to a trend.

Warming components measured here are steps, the internal trends between steps, and the shifts from one trend to the next. Counting shifts as the remainder between internal trends, preferences trends over shifts. When each is contrasted with an independent variable, such as ECS, this poses a strong test for shifts because internal trends estimate $-H_{step}$ in each timeseries. The hindcast (1861–2005) and projection (2006–2095) components of the RCP4.5 107-member ensemble were analysed separately.

For the hindcasts (1861–2005), total warming (the 2000–05 average minus the 1861–99 average) is positively correlated with total steps (0.93, p<0.01). Their means are 0.97 °C and 0.94 °C, respectively. The correlation between total warming and internal trends is 0.36 (p<0.01) and shifts is 0.58 (p<0.01). Shifts therefore explain 2.5 times the variance explained by internal trends in estimating total warming (Fig. 6a). A simple linear trend measured over the entire period has the same correlation with steps (0.93, p<0.01) but averages 0.76 °C, so underestimates total warming by 0.18 °C. Total warming, total steps, total shifts and total internal trends correlate poorly with ECS (-0.01, -0.01, 0.07 and -0.09, all NS, Table 4, Fig. 6b).

The ratio of total internal trends to total steps slightly favours shifts (mean 0.44), ranging between -0.09 and 1.22. A low ratio means that trends either cancel each other out or are negligible. A high ratio usually indicates the timeseries contains one or more negative shifts and/or a number of positive trends. Observations fit comfortably within this distribution with ratios of 0.32 to 0.38, except the GISS timeseries, which has a ratio of 0.62 because of a downward shift and upward trends in the early part of the record (Fig. 6c). The MME ratios are slightly negative with respect to total warming (-0.14, NS), suggesting that the mix of shifts and trends is largely unrelated to the amount of hindcast warming (1861–2005).

For the historical period, total warming and its various components – steps, shifts or trends – are unrelated to ECS. The relationship between total shifts and total internal trends is negative (0.47, p<0.01), which is to be expected, but the lack of a relationship between the shift\trend ratios and warming or ECS, suggests that this uncertainty is stochastic.

For the projection period, total warming over 2006–95 is based on the difference between five-year averages centred on 2006 and 2095. Total warming averages 1.55 °C, total steps average 1.57 °C and they are highly correlated (0.98, p<0.01). The correlation between shifts and internal trends with total warming is 0.70 and 0.74, respectively, trends having a slightly higher correlation (Fig. 6d). However, correlations between ECS, and

total steps, shifts and trends, are 0.81, 0.72 and 0.43, respectively (all $p<0.01$, Fig. 6e). This shows that the timeseries are becoming more trend-like at higher rates of forcing, when compared to the hindcast period.

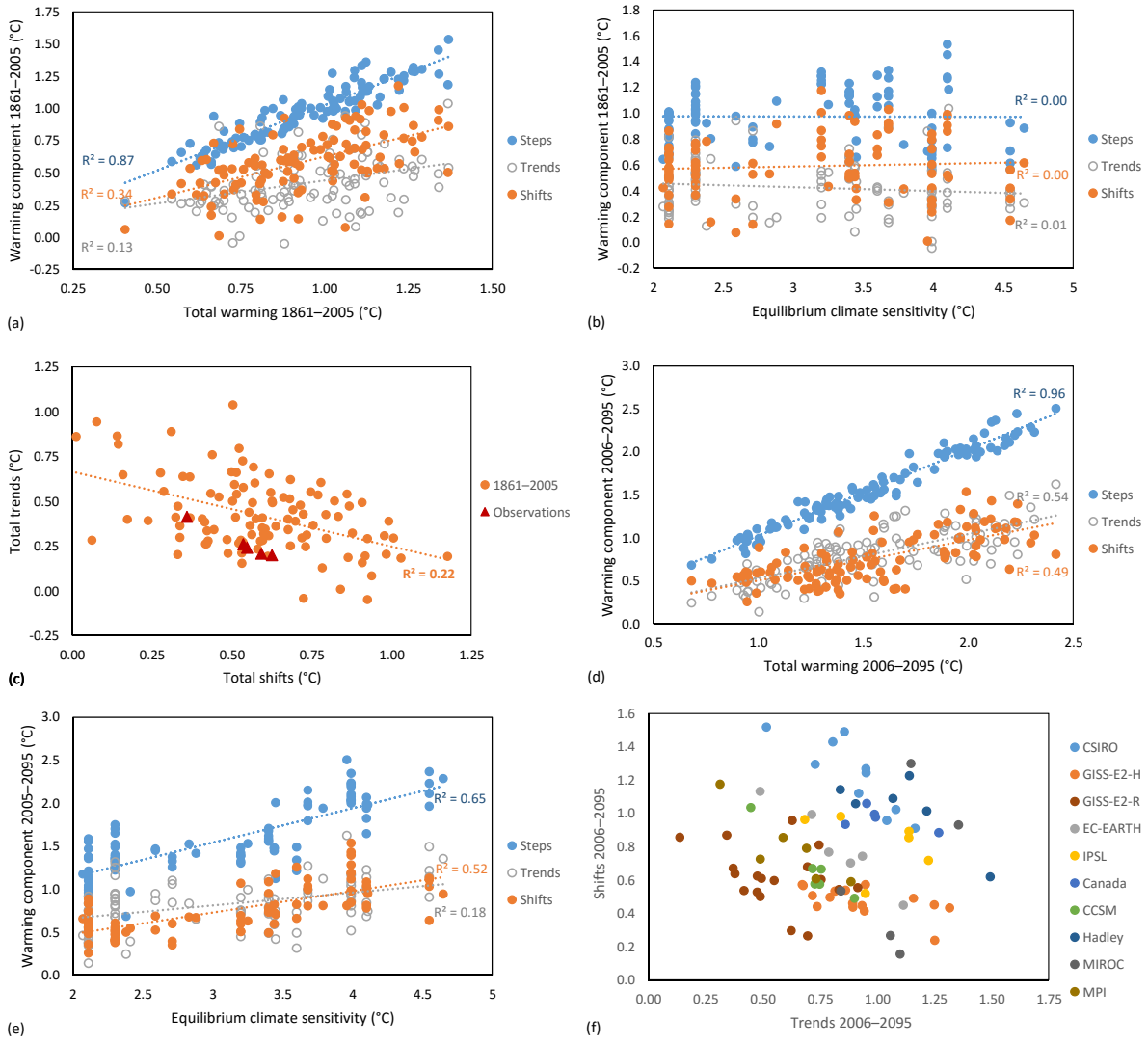


Figure 6: Multi-model ensemble (RCP4.5, 107 members) characteristics of hindcast (1861–2005) and projected (2006–2095) periods. (a) relationship between total warming and steps, trends and shifts (1861–2005); (b) relationship between ECS and steps, trends and shifts (1861–2005); (c) total shifts and total trends 1961–2005 with observed points from five warming records; (d) relationship between total warming and steps, trends and shifts (2006–2095); (e) relationship between ECS and steps, trends and shifts (2006–2095); (f) total shifts and total trends 2005–2095 from individual climate models.

Shifts have 2.9 times more explanatory power than trends with respect to ECS, but 0.9 times the explanatory power with respect to total warming over 2006–2095. We take this as meaning that shifts (steps minus internal trends) carry most of the signal and that trends are more random, affected by short-term (interannual) stochastic behaviour. Some of the signal embedded in trends could also be due to shifts occurring at regional scales, which are too small to register statistically as steps at the global scale.

The ratio of trends to steps is 0.51, ranging from 0.14 to 0.88. The ratio of trends to shifts favours trend (1.22) but has a large range (3.25 to 0.15). The correlations of both ratios with warming are very low (0.07, 0.03, respectively, NS). This seeming paradox where there is no correlation with the amount of warming but there is with ECS, when both ECS and warming are correlated, can be viewed by plotting the different modelling groups according to the relationship between shifts and trends. Individual models plot along linear pathways as was the case for the hindcast ensemble (Fig. 6c). The high sensitivity models plot towards the upper right

and lower sensitivity models to the lower left. The ratios for these individual groups vary widely – the CSIRO eight-model ensemble has trend/shift ratios of 0.25 to 0.56 and the GISS-E2-R seventeen-member ensemble ranges from 0.17 to 0.72. The potential for the same model to produce very different shift/trend ratios shows high stochastic uncertainty, probably generated by ocean-atmosphere interactions. The timing of these interactions appears to be largely unrelated to climate sensitivity, although the warming response to shifts when they do occur is related to sensitivity.

Interestingly, the GISS models form two groups, the main difference being the ocean configuration (see Schmidt et al., 2014a), where the Russell ocean model produces more steplike outcomes and the HYCOM ocean model produces more trend-like outcomes.

For each individual decade from 1876–1875 to 2086–2095, correlations were performed between step size and ECS (Table 3). The late 19th century produces downward steps in response to the Mt Krakatoa eruption in 1883 and is negatively correlated with ECS. Positive steps dominate from 1886 through to 1945 and are positively correlated at levels of low or no significance. The period 1946 to 1965 is negatively correlated with ECS; in 1956–65, corresponding with the 1963 Mt Agung eruption, downward steps result in a negative correlation of -0.52 (p<0.05). Correlations between ECS and step size become positive after 1965, being 0.41 for 1976–85 and 0.49 for 1986–95 (both p<0.01). For the decade 1996–2005, 101 of the 107 member MME undergo an upward step, but the correlation with ECS is only 0.19 (NS). This low correlation may partly be due to a rebound from the negative forcing of the 1991 Mt Pinatubo eruption in the models, which has been over-estimated by about one third (Schmidt et al., 2014b). Correlations for the forcing period (2006–2095) rise to 0.68 in 2006–15 and vary between 0.57 and 0.82 for subsequent decades to 2095.

The lack of predictability in the hindcasts is a result of negative aerosol forcing due to volcanic eruptions and anthropogenic sources occurring after 1950. The more sensitive models produce strong positive and negative responses depending on the direction of forcing, whereas in the less sensitive models this effect is reduced. This effect cancels out any consistent relationship between ECS and step size over the historical period. The implication of this finding is that the magnitude of 20th century warming in the models has little predictive skill and is not a reliable guide to potential future risk.

The hindcast results are also uncorrelated with the 21st-century projections. Total warming (1861–2005) is negatively correlated with 21st century warming (2006–95, -0.25, p~0.01) and uncorrelated with respect to ECS (-0.01). Total steps from the hindcast and forecast periods show similar negative correlations. Internal trends 1861–2005 are also uncorrelated with future total warming, steps or trends. This strongly indicates that 20th century warming may not be a good guide to future warming, if observations are being affected in a similar way.

A final analysis looks at the explanatory power of different change models with respect to ECS over time. Linear and quadratic trends, steps and warming to date are calculated for successive decades for each ensemble member and the results correlated with ECS. Both trends and warming difference respond to negative forcing in the first part of the record. Step changes are less volatile, remaining close to zero until increasing from 1995 and remain higher than the other models until the end of the century (Fig. 7a). The standard error measured from total accrued warming was also least out of the three statistical models. Although it would be possible to derive a closer fit for some of those models with a greater number of factors, step changes clearly carry the greatest signal with respect to ECS over time. The analysis repeated from 1965 produces a similar result (Fig. 7b).

Table 3. Steps collated for each decade from 1876 to 2195 from the RCP4.5 MME, showing total steps up and down and the correlation between step size and ECS. The second part of the table shows the correlations between total warming, steps and trends over the observed and simulated periods and ECS. Correlations are classified as not significant (NS, $p>0.05$), $p<0.05$ (*) and $p<0.01$ (**). Total correlations with the MME are n=107 and with ECS are n=92.

Change and period	Steps up	Steps down	Correlation with ECS	Significance
Steps 1876–1885	0	26	-0.40	*
Steps 1886–1895	13	1	-0.32	NS
Steps 1896–1905	7	1	-0.09	NS
Steps 1906–1915	31	0	0.27	NS
Steps 1916–1925	65	0	0.27	*
Steps 1926–1935	17	1	0.09	NS
Steps 1936–1945	33	0	0.20	NS
Steps 1946–1955	6	1	-0.85	*
Steps 1956–1965	4	12	-0.52	*
Steps 1966–1975	29	0	0.33	NS
Steps 1976–1985	56	0	0.41	**
Steps 1986–1995	34	0	0.49	**
Steps 1996–2005	101	0	0.19	NS
Steps 2006–2015	83	0	0.68	**
Steps 2016–2025	82	0	0.65	**
Steps 2026–2035	70	0	0.74	**
Steps 2036–2045	82	0	0.66	**
Steps 2045–2055	75	0	0.57	**
Steps 2056–2065	65	0	0.67	**
Steps 2066–2075	61	0	0.60	**
Steps 2076–2085	51	0	0.66	**
Steps 2086–2095	27	0	0.82	**
	Mean (°C)	Range (°C)		
Warming 1861–2005	0.9	0.4–1.4	-0.01	NS
Warming 2006–2095	1.5	0.7–2.4	0.81	**
Steps 1861–2005	1.0	0.3–1.5	-0.01	NS
Steps 2006–2095	1.6	0.7–2.5	0.81	**
Shifts 1861–2005	0.6	0.0–1.2	0.07	NS
Shifts 2006–2095	0.8	0.3–1.5	0.72	**
Trends 1861–2005	0.4	0.0–1.0	-0.09	NS
Trends 2006–2095	0.8	0.1–1.6	0.43	**

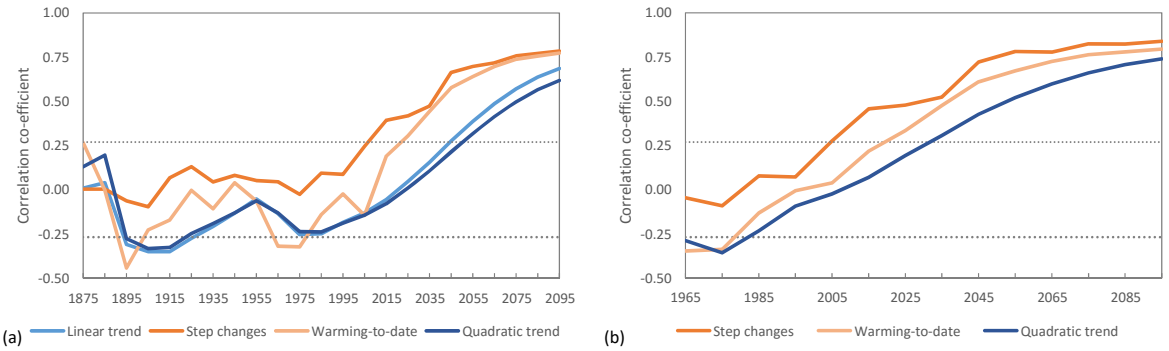


Figure 7: Correlation with successive calculations of change over time using a linear trend, step changes, warming-to-date and quadratic trend; (a) to the nominated date or decadal average subtracted from 1861–99 for warming-to-date and (b) as for (a) and decadal average subtracted from 1861–1959. Dotted lines mark p<0.01.

This result is further evidence for step changes carrying the signal. Warming-to-date assesses any warming irrespective of its cause, whereas if step changes are part of a direct response to forcing they would be a direct part of the response. This certainly seems to be case for climate models, so seems a realistic assumption to also link to observations. The advantage for using warming to date as a measure is that it has roughly a decade's advantage over statistical tests, which require hindsight, so unless the physical mechanism(s) become known, both have roughly equivalent skill at the present time.

21st Century forcing profiles

If increased forcing raises the rate of entropy production, we would expect to see steplike behaviour becoming more trend-like over time. Such behaviour would involve either:

- increase the frequency and distribution of regional step changes that integrate to become more trend-like at the global scale, or
- see an increase in the rate of diffuse warming, producing widespread trend-like behaviour.

If either is the case, then simulations for the four different emissions pathways, RCP2.6, 4.5, 6.0 and 8.5, should show this.

Figs 5c–f shows the percentage of step changes in any given year for the multi-model ensemble for each of these pathways. For RCP2.6, peaks occur to about 2050, after which the ensemble stabilises. Some models step downward, the earliest of these in 2051. Individual members stabilise between 2018 and 2092, with 48 of the final shifts being positive and 13 negative. This timing is weakly correlated with ECS (0.18, NS). ECS is uncorrelated with the size of the final shift, or to the gradient of the following trend. The RCP4.5 ensemble produces frequent steps that peak around 2025 and decline towards the end of the century. RCP6 produces a fairly constant rate of steps and RCP8.5 produces sustained steps throughout the century, peaking in the 2080s at a higher rate than 1996–98.

This evolution shows a step-ladder like process in the 20th century that changes in to an elevator-like process in the 21st becoming more trend-like with increasing forcing. Depending on the subsequent rate of forcing trend-like processes can either recede back to a steplike process or even stabilise. The HadGEM2-ES single model ensemble is used to illustrate this (Fig. 8a).

This ensemble shares the same historical forcing to 2005. It warms by less than observations to 2010, with a reversal 1964–1980, then warms substantially in a series of steps over the next few decades. It undergoes a step change of 0.37 °C and shift of 0.18 °C in 1998, one year after the observed shift. The next step occurs in 2012, 2013, 2014 and 2015 in the four simulations, ranging from 0.40 °C to 0.49 °C in absolute terms and 0.19 °C to 0.27 °C as the shift from the pre-step trend to the post-step trend. The first half of the 21st century shows the influence of decadal variability on mediating step changes. In 2021, the RCP2.6 simulation undergoes a step change and is higher than the others for most of that decade. The RCP6.0 simulation is lower than the others from 2025–45 before accelerating under a sustained step-and-trend process. The relative proportion of internal trends to total warming under the four scenarios is 0.34, 0.60, 0.57 and 0.79, for warming of 1.9 °C, 2.9 °C, 3.7 °C and 5.3 °C, respectively. The RCP4.5 has a higher trend ratio, showing the stochastic uncertainty inherent in the simulations.

Like most statistical tests that detect change points, the bivariate test is considerably weakened under autocorrelated data, where its timing is fairly robust but $p(H_0)$ is sensitive. Such autocorrelations may be caused by simple trends, lag-1 or longer lag processes influencing the complex nature of warming. Removing these without assuming an underlying process is difficult, so one way of assessing the influence is to pass a moving window through a timeseries. If the data is steplike and largely free of autocorrelation, a distinct step will produce a line of horizontal Ti_0 statistics on a single date as it passes through the window. If there are no steps within a window period and autocorrelation is low, background Ti_0 values will return to low values (single digits). With autocorrelation, background Ti_0 values remain above the $p<0.01$ threshold and form a 'cloud', rather than steps producing horizontal lines.

In Fig. 8b–e, successive horizontal lines extending right from low Ti_0 values indicate step-ladder-like behaviour in the 20th century. Horizontal lines that stay on the right without returning to low T_{i0} values indicate both steplike and trending behaviour. A cloud to the far right, as in Fig. 8e, shows a trend-dominated process. Summarising 21st century behaviour under increasing emissions, RCP2.6 shows a return to steplike changes,

stabilising around 2050, RCP4.5 shows a return to steplike change late century, RCP6.0 shows increasing trend-like behaviour over the century and RCP8.5 shows a consistent trend to the end of the century, with few steps.

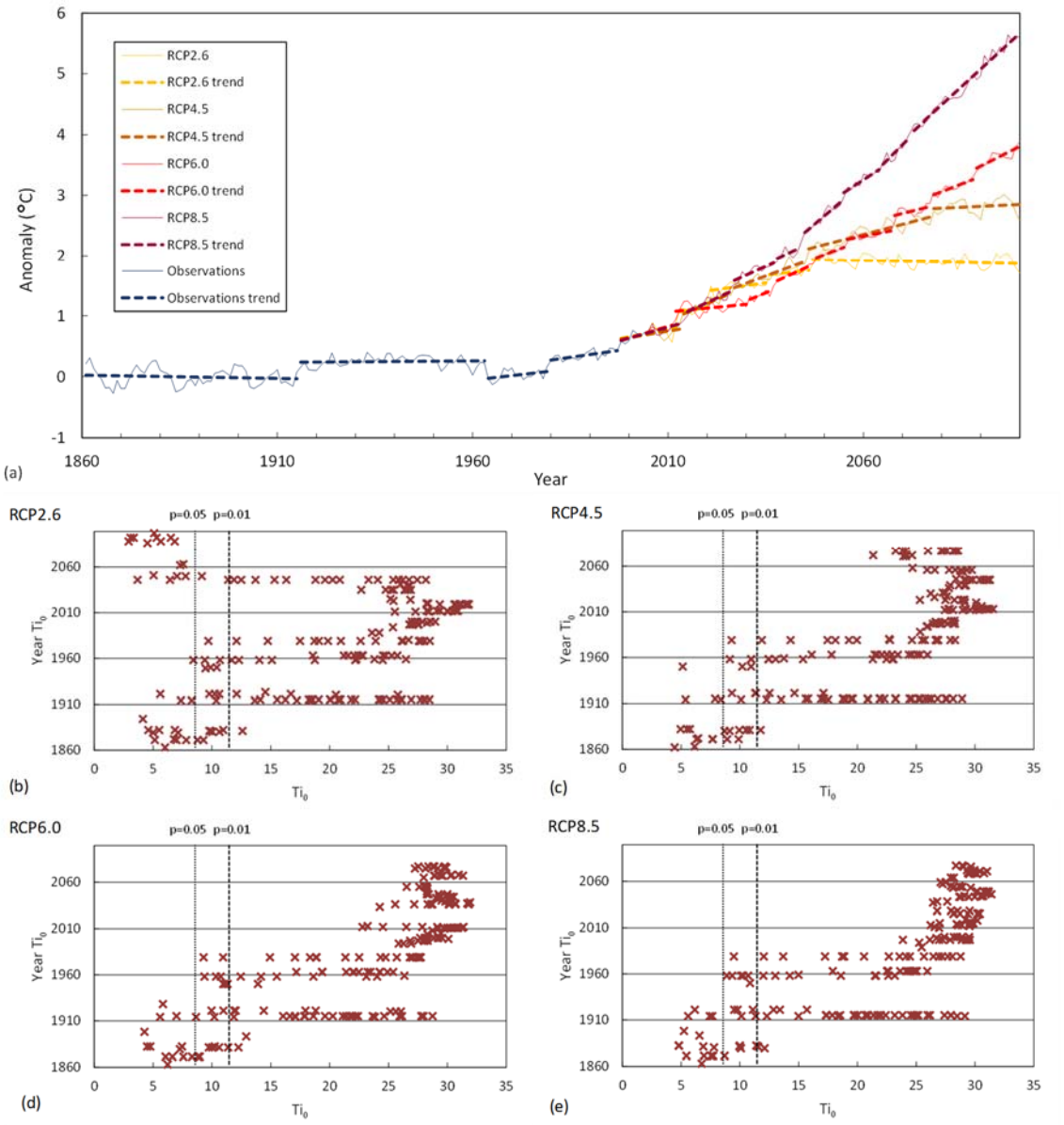


Figure 8: Global mean surface temperature as analysed by the multi-step bivariate test; (a) Step and trend breakdown of global means surface temperature in the RCP2.6, 4.5, 6.5 and 8.0 simulations from the HadGEM-ES model, run 3; (b–e) Ti0 results from a 40-year moving window for the RCP2.6, 4.5, 6.5 and 8.0 simulations, respectively.

An indication of change at the regional scale and how it may relate to global change is illustrated by using selected CMIP3 models for SE Australia as described in Jones (2012). For example, for the CSIRO Mark3.5 A1B simulation, for global mean warming, internal trends comprise 52% of total warming 2006–2095, whereas for SEA *Tmax* the ratio is 13% and *Tmin* 47%. These were consistent for A1B and A2 forced simulations, which are roughly equivalent to RCP4.5 and 6.0. The number of step changes is also notable: four and five at the local scale and twelve at the global scale (Fig. 9). The higher ratio for *Tmin* compared to *Tmax* may be due to *Tmin* being related to large-scale sea surface temperature patterns and *Tmax* being related to more local soil moisture patterns as is the case for the central and western United States (Alfaro et al., 2006). Jones et al. (2013) showed that such changes at the local scale produce significant increases in impact risks.

These analyses do not support increasing trend-like behaviour at the local scale, and therefore favours the first alternative above, but further work across more regions is required to confirm this.

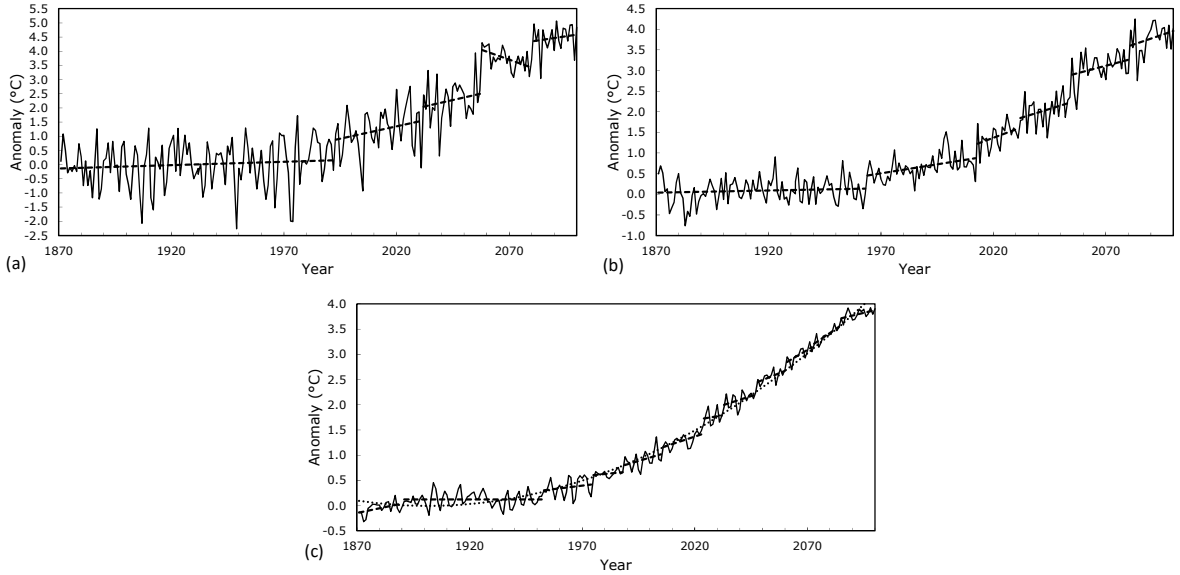


Figure 9: Anomalies of annual mean temperature showing internal trends separated by step changes from the CSIRO Mk3.5 A1B simulation; (a) maximum temperature southeastern Australia; (b) maximum temperature southeastern Australia; (c) global mean surface temperature Internal trends (dashed lines) are separated by step changes (p<0.01).

Severe testing of steps and trends

Earlier sections have identified steps and trends in temperature and tested how trend, step and trend-shift relationships relate to total warming the independent variable ECS. This section examines how well trend, step and step-trend models reproduce the temperature records examined throughout the paper. This tests h_{trend} against h_{step}. The error value assigning $p<h_0$ is not the principal measure being sought. Instead, the statistical model that combines low error with unstructured residuals while sustaining physically plausible assumptions is preferred. Another aim is, if possible, to provide likelihoods for severe testing.

Four statistical models are tested: ordinary least squares trend, LOWESS, step, and step and trend. The LOWESS model was applied with a bandwidth of 0.5 to assess sensitivity to fluctuations in the data, contrasting those with both the trend and step model. It is not considered a valid statistical rival because it is fitted without regard to physical process. Likewise, although the step and trend model will fit well to the data, the step model is the one used for severe testing, being a straightforward measure of h_{step}. The trend model represents h_{trend}.

With the data produced, we look at goodness of fit (r^2), the residual sum of squares (RSS), cumulative ($\sum R$) residuals and cumulative residuals squared ($\sum R^2$). Residuals (R) show how much variance is explained by the model, cumulative residuals will show whether residuals are showing structure not explained by the model and cumulative residuals squared show accumulating error, including rapid changes not accounted for. To these have been added four more tests: F-tests for autocorrelation (F-auto) and heteroscedasticity (F-hetero) of the residuals over the whole record and percentage of exceedance over moving 40-year windows. White's test (White, 1980) is used for heteroscedasticity. The first four of these tests use absolute error, or the amount of a timeseries not explained by the statistical models and the second four show patterns, working on accuracy and precision, respectively. The statistical models that fail a combination of both are therefore the weakest.

Results are shown in Fig. 10 and Tables 4 and 5. The data and statistical models for HadCRU record 1880–2014 are shown in Fig. 10a. Cumulative residuals that track close to zero (Fig. 10b) show the model mimicking the data closely and sustained departures show significant deviation. Here, the trend model deviates substantially and the LOWESS model less so, while the step and step and trend models deviate least. This follows through to

the cumulative residuals squared. The less change the better; whereas upward kinks show rapid changes or large outliers (positive or negative) not incorporated into the model (Fig. 10c). Trend analysis produces an r^2 value of 0.76 and residual sum of squares of 0.87, and the other three statistical models have an r2 of 0.87 and RSS of 0.8. For $\sum R^2$ the trend model behaves more poorly than the other three.

With the autocorrelation and heteroscedasticity tests, the LOWESS test performs less well than the other two for the 40-year tests. Although the LOWESS model performs well over the whole record, it is subject to deviations within the record that cancel each other out – akin to cutting corners. The step and trend model does worst for F-hetero over the whole record, but the best over 40-year windows. This is due to high variance within the early part of the record and is an issue of precision, as standard error of this relationship is almost half that of the trend model (not shown, but is similar to the $\sum R^2$ relationship). The step model is clearly superior to the trend model for the moving window tests. The results for the other four long-term global warming records: BEST, C&W, GISS and NCDC, are not shown but have similar results.

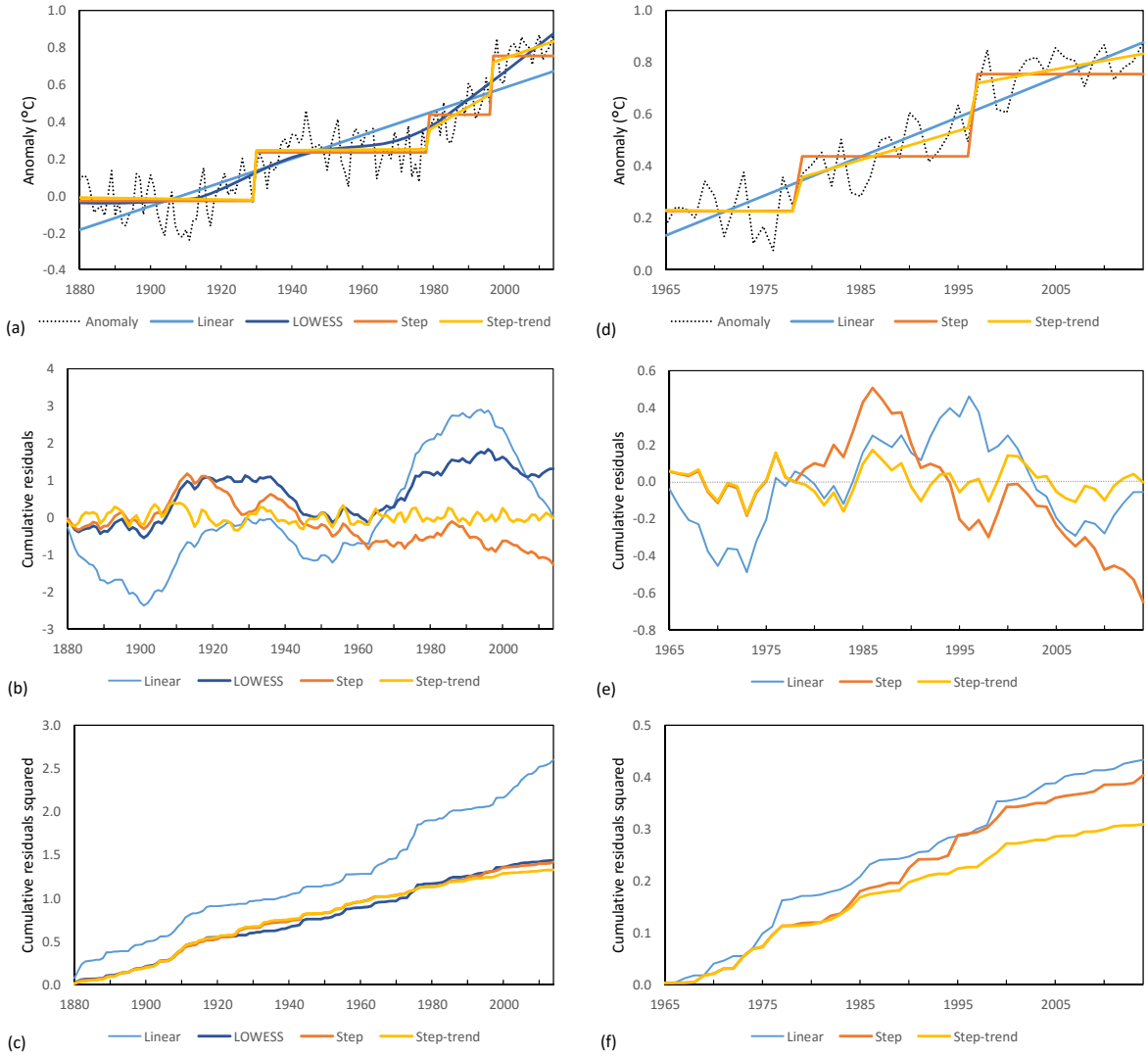


Figure 10: Testing three models to mean global anomalies of surface temperature from the HadCRU record, 1880–2014 (a–c) and 1965–2014 (d–f); (a) and (d) mean annual anomalies and linear, step change and shift and trend models; (b) and (e) cumulative residuals for each model, where success is measured as tracking close to zero; (c) and (f) cumulative sum of residuals squared, where upward steps show nonlinearity not explained by each model.

These tests, omitting LOWESS, were carried out for HadCRU 1965–2014, a period with a sustained radiative forcing signal (Fig. 10d). The results for the different statistical models are similar, with r^2 values of 0.85, 0.86 and 0.89, respectively. The step and trend model is still the best performed, but the step model only slightly

better than the trend model – this is due to the northern hemisphere shift in 1987/88 being incorporated into the into the trend at global scale, where each of the models are statistically identical., Dividing this timeseries into quarters will bring 1987/88 into the picture but also make both the MYBT and t-test test more sensitive.

Table 4. Results of eight tests on four statistical models for selected observed global temperature data (except where noted). The statistical models tested are trends (power shown), LOWESS (0.5 total series smoothing), steps and steps and trends. Result include the adjusted r^2 value, the residual sum of squares (SS), cumulative residuals and squared cumulative residuals. F-tests for the whole series are shown, with $p<0.05$, $p<0.01$ noted if registered, otherwise $p>0.05$. F-test failure for 40-year period autocorrelation and heteroscedasticity is measured at $p<0.01$.

Model	r^2	Residual SS	Cumulative residuals ($\sum R\ y^{-1}$)	Cumulative residuals^2 ($\sum R^2\ y^{-1}$)	F-test auto-correlation (F, pH_0)	F-test hetero-scedasticity (F, pH_0)	40-y periods fail F-test auto-correlation	40-y periods fail F-test hetero-scedasticity
HadCRU 1861–2014								
Trend	0.76	2.6	1.2	1.3	0.0	3.7	58%	13%
LOWESS	0.87	1.4	0.7	0.8	0.3	1.0	28%	13%
Step	0.87	1.4	0.5	0.8	0.7	3.2	0%	0%
Step-trend	0.87	1.3	0.1	0.8	0.2	5.8, 0.05	0%	0%
HadCRU 1965–2014								
Trend	0.85	0.43	0.20	0.24	0.0	1.2	0%	0%
Step	0.86	0.40	0.20	0.21	0.4	0.7	0%	0%
Step-trend	0.89	0.31	0.06	0.18	0.0	1.4	0%	0%
NCDC 30°N–60°N 1880–2014								
Trend	0.64	6.3	1.8	2.3	0.0	10.2, 0.01	51%	9%
LOWESS	0.79	3.7	0.9	1.6	0.2	3.0	19%	0%
Step	0.83	2.9	0.3	1.4	0.0	3.0	0%	1%
Step-trend	0.83	2.9	0.2	1.4	0.0	3.2, 0.05	1%	0%
HadCRU quarterly 1979–2014								
Trend	0.69	1.7	2.0	3.5	0.0	1.1	20%	3%
LOWESS	0.72	1.6	0.5	3.3	0.2	2.8	3%	5%
Step	0.75	1.4	0.7	2.8	0.0	0.2	0%	0%
Step-trend	0.76	1.3	0.2	2.7	0.0	0.4	0%	4%
GISS quarterly 1979–2014								
Trend	0.67	1.9	1.6	4.1	0.0	1.1	20%	0%
LOWESS	0.69	1.8	0.5	3.9	0.1	2.2	6%	2%
Step	0.71	1.6	0.9	3.4	0.0	0.0	4%	0%
Step-trend	0.72	1.6	0.3	3.3	0.0	0.6	0%	0%
RSS quarterly 1979–2014								
Trend	0.40	3.4	4.4	6.9	0.0	1.2	11%	6%
LOWESS	0.46	3.1	1.1	6.4	0.3	2.3	4%	14%
Step	0.52	2.7	0.9	5.5	0.0	0.3	4%	8%
Step-trend	0.53	2.6	0.7	5.1	0.0	1.3	0%	37%
UAH quarterly 1979–2014								
Trend	0.35	3.6	3.1	7.4	0.0	1.8	6%	9%
LOWESS	0.39	3.4	1.0	7.2	0.1	3.3, 0.05	4%	20%
Step	0.46	3.0	1.5	6.1	0.0	0.7	7%	12%
Step-trend	0.46	2.9	0.8	5.8	0.0	1.5	4%	42%

Also shown in Table 4, are the zonal temperatures from NCDC 30°N–60°N (1880–2014) where total internal trends are slightly negative (-0.04 °C) and shifts are positive (1.13 °C or 106% of steps). The pattern of results is similar to those for the global HadCRU record but the residuals are slightly more than double and the cumulative residuals almost double. The step model is clearly superior to the trend model, which fails White's test for the whole record, fails the 40-year F-auto at a level of 51% and has an RSS double that for steps. This record is entirely made up of steps, showing the lack of trend occurring within some regions.

Table 5. Results of eight tests on four statistical models for representing global mean warming from HadGEM-ES climate model run3 RCP2.6, 4.5, 6.0 and 8.5, showing the amount of warming for different measures. The statistical models tested are trends (power shown), LOWESS (0.5 total series smoothing), steps and steps and trends. Results include the adjusted r^2 value, the residual sum of squares (SS), cumulative residuals and squared cumulative residuals. F-tests for the whole series are shown, with $p<0.05$, $p<0.01$ noted if registered, otherwise $p>0.05$. F-test failure for 40-year period autocorrelation and heteroscedasticity is measured at $p<0.01$.

Pathway	Warming (°C)	Steps (°C)	Trends (°C)	Shifts (°C)
RCP2.6	1.93	2.29	0.65	1.24
RCP4.5	2.93	3.30	1.76	1.07
RCP6.0	3.65	3.86	2.09	1.75
RCP8.5	5.34	5.35	4.24	1.41

Model	r^2	Residual SS	Cumulative residual ($\sum R/y$)	Cumulative residual^2 ($\sum R^2/y$)	F-test auto-correlation (F, pH_0)	F-test hetero-scedasticity (F, pH_0)	40-y periods fail F-test auto-correlation	40-y periods fail F-test hetero-scedasticity
RCP2.6								
Trend (x^4)	0.95	3.9	4.7	3.6	0.4	8.9, 0.01	75%	18%
LOWESS	0.96	4.7	7.7	2.8	6.9, 0.01	0.4	64%	31%
Step	0.98	1.1	0.04	1.2	0.1	10.7, 0.01	1%	3%
Step-trend	0.98	0.9	0.01	1.1	0.0	12.1, 0.01	0%	4%
RCP4.5								
Trend (x^2)	0.95	8.8	16.6	4.8	0.8	2.1	77%	73%
LOWESS	0.99	3.9	13.3	2.5	2.3	4.1, 0.05	61%	45%
Step	0.98	2.4	0.5	1.4	0.0	5.7, 0.05	19%	14%
Step-trend	0.99	1.0	0.02	1.1	0.0	13.4, 0.01	0%	2%
RCP6.0								
Trend (x^2)	0.97	4.5	51.1	5.2	3.7	23.5, 0.01	63%	56%
LOWESS	0.98	2.9	24.6	2.4	0.9	8.3, 0.01	52%	31%
Step	0.99	1.2	0.06	1.2	0.1	9.7, 0.01	2%	5%
Step-trend	0.99	0.6	0.01	1.1	0.0	17.9, 0.01	0%	20%
RCP8.5								
Trend (x^3)	0.99	4.3	4.5	3.1	0.0	11.8, 0.01	62%	39%
LOWESS	0.992	3.1	66.6	2.8	2.0	4.5, 0.05	45%	22%
Step	0.99	8.1	2.0	1.7	0.2	106.7, 0.01	13%	18%
Step-trend	0.997	0.7	0.01	1.1	0.0	12.0, 0.01	0%	3%

The quarterly record of HadCRU from Fig. 3 (1965–2014) is more fine-grained, incorporating the 1987/88 shift (Table 4). If warming is gradual, the results for trends should be scalable, however, they perform less well at this timescale. The respective r^2 results are 0.69, 0.72, 0.75 and 0.76, whereas the differences in the cumulative residuals are 2.0, 0.5, 0.7 and 0.2, where zero is a perfect score. Here, the LOWESS model performs similarly to the step model because it closely follows the data. The step model performs better than the trend model for HadCRU quarterly data, and almost as well as the step and trend model. For the GISS quarterly data, the results are similar.

The satellite records are more steplike than surface temperature when measured using cumulative residuals. The step and trend model for the 40-step window heteroscedasticity tests for satellite data fails for both RSS and UAH. This is due to two instances of short-term departures on an otherwise stable background that measures heteroscedasticity as significant with the F-test: 1) a warm period during 1998, which is represented as a single step but lasts four quarters and 2) a small warming event associated with an El Niño event in 2010 lasting two quarters. Removing this short-term warming from these sequences removes the heteroscedasticity. So although not all deviations are removed by representing the satellite record as being stepwise, it still provides a better explanation of change than the trend model.

Simulated global annual mean surface temperatures from climate models show results consistent with observations (Table 5). The data from Fig. 7 were analysed in the same way, except that quadratic (RCP4.5, RCP6.0), cubic (RCP8.5) and quartic (RCP2.6) polynomial functions were used instead of a linear trend. The LOWESS model used here at 0.5 record length is relatively low resolution providing 120-year smoothing. The step model outperforms both the trend and the LOWESS model in all simulations, with the exception of the RSS in the RCP8.5 simulation. The RCP2.6 simulation is the most steplike. In the RCP4.5 simulation, the step model does slightly worse than in the RCP6.0 simulation, which is actually more steplike. This shows the role of stochastic uncertainty in the warming process as portrayed in Fig. 6f. The RCP8.5 simulation is the most trend-like; the step model fails in the final decades of the 2st century because the bivariate test detects no steps, but the climate continues to warm. This is what we would expect if shifts became more local and more frequent, integrating into a curve at the global level, much like sea level rise does today.

Severe testing summary

A range of statistical tests have been used to examine h_{step} and h_{trend} as representatives of scientific hypotheses *H1* and *H2*. This is consistent with the substantive null described by Mayo and Cox (2010) where we have applied tests designed to provide a clear choice between *H1* and *H2*. The focus is on whether atmospheric warming is gradual, forming a monotonic trend, or is stepwise and periodic, forming a complex trend over time. To paraphrase Mayo and Cox (2010): hypothesis *H1* predicts that h_{trend} is at least a very close approximation to the true situation; rival hypothesis *H2* predicts a specified discrepancy from h_{trend}, and the test has a high probability of detecting such a discrepancy from *H1* were *H2* correct. Detecting no discrepancy is evidence for its absence.

As stated in the introductory sections, no single test can undertake that task. We rely on the multi-step Maronna-Yohai bivariate test to identify step changes in the input data but beyond that make as few assumptions as possible. A total of six probative tests with links to the two substantive hypotheses were proposed earlier in the paper – these are designed to pinpoint discrepancies between *H1* and *H2* by analysing the global warming data they seek to explain. The data generated consists of steps, trends and shifts applying the multi-step MYBT model and trends applying least squares trend analysis. The use of models such as LOWESS are for sensitivity testing and not part of the probative assessment.

The results of the probative tests are summarised through the following findings:

Stratified analysis of change points

- Global and regional analyses of steps show a highly coherent pattern of change points, where warming in the second half of the 20th century aligns with known regime changes associated with changes in decadal variability (Table 6). These events comprise the major proportion of historical warming to 2014.
- Analysis of steps, internal trends and shifts in observations attributes higher proportions of warming to shifts at the zonal scale (up to 100%), moving to lower proportions at the global scale. Three regional assessments also contain high shift/step ratios, with trends playing a lesser role.
- This effect is larger in the mid-latitude regions and with SST, indicating the role of equator-to-pole hydrothermal transport of energy in the ocean-atmosphere system. Their timing shows a strong role is being played by decadal variability.
- Surface and satellite temperatures undergo contemporaneous shifts at the global scale, largely removing the discrepancy between trends within the two data sets. Both surface and satellite temperature records are very steplike, with surface trend/shift ratios of 0.19 and 0.27 and satellite ratios of -0.55 and -0.40 showing the effect of internal trends. Shifts are consequently higher than steps in the satellite data.

Similar patterns of change in observations and physical models

- Correlations between step change frequency in the observed 44-member group of global and regional data and the CMIP3 and CMIP5 MMEs analysed (1880–2005), are 0.32 and 0.34, respectively (p<0.01). For the period 1950–2005, correlations rise to 0.45 and 0.40, respectively. Grouping specific events (1963/64, 1968–70, 1976/77, 1979/80, 1987/88 and 1996–98) and analysing other years individually, correlation increases to 0.78 for both CMIP3 and CMIP5 records. Variations in forcing, especially

volcanoes may affect the timing and direction of step changes, but they are not their sole cause, given that 21st century simulations produce step changes from smoothly varying changes in forcing.

- Fifty-eight members of a 107-member MME (CMIP5 RCP4.5) show a step change in 1996–98 reproducing the observed change in 1997 within ±1 year.

Nonlinear components of warming carry more of the signal than linear components

- Analysis of steps and trends in observed and model data shows that steps explain change better than trends when the structure of the residuals area assessed using goodness of fit, residual sum of squares, cumulative residuals, cumulative residuals squared, autoregressive residuals testing and White's test for heteroscedasticity.
- For simulated historical warming 1861–2005, the r^2 values for steps, shifts and trends in explaining total warming are 0.87, 0.43 and 0.13, respectively. Simulated warming for this period is not correlated with ECS.
- For the 21st century (2006–2095) the r^2 values for steps, shifts and trends in explaining total warming are 0.96, 0.54 and 0.49, respectively. The r^2 values for steps, shifts and trends in explaining ECS are 0.65, 0.52 and 0.18, respectively.

Stationary and non-stationary periods are separated by step changes

- In all three locations on three continents tested, and for six independent climate model simulations for SE Australia, warming commenced with a step change in *Tmin* and sometimes *Tmax*. Warming is not slowly emergent in any of this data as would be expected if warming is gradual. The coincident timing of shifts in SE Australia with southern hemisphere step changes and those in the UK and USA with northern hemisphere changes, suggest that warming has commenced abruptly in different areas of the globe at different times, and that the separation between stationarity and non-stationarity in the temperature record is abrupt.

Other variables show similar step changes

- Step changes exhibiting similar timing have been shown for tide gauge observations, rainfall, ocean heat content, forest fire danger index and a range of other climate variables, in addition to many impact variables. These are overwhelmingly attributed to random climate variability, including abrupt changes identified as part of decadal regime change.

The best representations of underlying step- and trend-like structures in the data.

- For observations and selected model data the simple step-ladder model performs better than the monotonic trend model for goodness of fit (r^2), the residual sum of squares (RSS), cumulative ($\sum R$) residuals and cumulative residuals squared ($\sum R^2$), White's test for heteroscedasticity, a moving 40-year window regression of the residuals and a moving 40-year window White's test.

Table 6 summarises the major tests undertaken with expected outcomes for h_{trend} and h_{step}. While objections could be made to each of these on an individual basis, collectively they show that for externally-forced warming on decadal scales, h_{step} is better supported than h_{trend}. However, long-term warming (greater than ~50 years) is largely trend-like, and is proportional to the amount of forcing.

In summary, these tests show that h_{step} is a close approximation of the data when analysing decadal-scale warming. Over the long term, this warming conforms to a complex trend that can be simplified as a monotonic curve, but the actual pathway is steplike. As outlined in Section 3.3, this rules out gradual warming, either in situ in the atmosphere or as gradual release from the ocean, in favour of a more abrupt process of storage and release. The precise mechanisms by which this occurs remain to be determined. This conclusion supports the substantive hypothesis *H2* over *H1*, where the climate change and variability interact, rather than vary independently.

Table 6. Selected test results that distinguish between h_{trend} and h_{step}. The null positions for each are generally not considered diametric. There is no generally accepted null with respect to h_{trend} that references nonlinear change whereas for H_{step} the null is no significant step-wise change points, or if there are they are completely random and do not contain and external forcing signal.

Test	Evidence	h_{trend}	h_{step}	Supporting literature
Global warming 1895–2014	Trend/step ratio 0.32–0.38 (4 records), 0.58 (1 record) Trend shift ratio 0.44–0.58 (4 records), 1.38 (1 record)	Gradual change, fluctuations but no steps	Substantial fraction of record contains steps	(Varotsos et al., 2014;Belolipetsky et al., 2015;Bartsev et al., 2016)
Regime changes	1997 29 in 1997, 37 in 1996–98 of 45 global & regional records	Extreme El Niño 1997/98, stochastic event	Step-wise change points identified in temp and physically-related records	(Overland et al., 2008;Chikamoto et al., 2012a;Chikamoto et al., 2012b;Reid and Beaugrand, 2012;Menberg et al., 2014)
	1987/88 6 in 1987, 4 in 1988 of 44 regional records. Global ocean NH, NH mid-lat	El Niño, stochastic event	Step-wise change points identified in temp and physically-related records	(Overland et al., 2008;Boucharel et al., 2009;Lo and Hsu, 2010;Reid and Beaugrand, 2012;North et al., 2013;Menberg et al., 2014;Reid et al., 2015)
	1979 15 in 1979, 7 in 1980, 5 in 1977, 1 in 1976 of 44 global and regional records. Global, tropics, SH	N Pacific regime shift 1976–77, El Niño 1978/79	Step-wise change points identified in temp and physically-related records	(Hare and Mantua, 2000;Overland et al., 2008;Meehl et al., 2009;Fischer et al., 2012;Reid and Beaugrand, 2012;Menberg et al., 2014)
	1969 4 in 1969, 8 in 1968–70, southern hemisphere	El Niño, stochastic event	Step-wise change points identified in temp and physically-related records	(Li et al., 2005;Hope et al., 2010;Jones, 2012)
Scalability of regional records	Records more steplike at zonal and regional scales and over the oceans.	Regional records would be trend-like if warming is diffuse and gradual	Regional records more steplike, large-scale records more trend-like.	None located
Attribution	Step-wise attribution for SE Australia (obs and models), Texas (obs), Central England (obs)	Gradual emergence of signal	Abrupt emergence of signal	(Jones, 2012)
Quarterly surface and satellite temperature 1979–2014	Surface and satellite records share similar shifts but not trends	Significant trend for periods >30 years	Contemporaneous step-wise change points in independently measured records	None located
Simulated temperature patterns 1861–2005	Clustering on runs test highly non-random (p~0.0˙ runs test) Significant correlations between timing of steps in models and obs CMIP3 0.32, CMIP5 0.34 1880–2005.	No matching patterns, randomicity	Matching step-wise changes between models and observations	None located
Simulated temperature quantities 1861–2005	Trends/steps ratio 0.44±0.22	Gradual change, deviations but no steps	Substantial fraction of record contains shifts	None located
Simulated temperature relationships with independent variable ECS RCP4.5 2006–2095	Correlation and r^2 between ECS and total warming 0.81 & 0.65, steps 0.81 & 0.65, shifts 0.72 & 0.52 and internal trends 0.43 & 0.18	Shifts random with respect to forcing	Shifts and steps more highly correlated with ECS and warming than trends	None located
Autocorrelation and heteroscedasticity observations 1880–2014	Steps better performer than simple trends (Failure rate Trends 58±1% autoc, 10±4% heterosc.; Steps 2±4%	Trends serially independent data, variations due to independent processes	Steps perform better than trends to explain autocorrelation and heteroscedasticity	None located

	autoc, 0% heterosc. 40y window)			
Autocorrelation and heteroscedasticity observations 1965–2014	Trends and steps pass all tests for annual data, steps slightly better correlation than trends (0.86, 0.85 HadCRU)	Trends serially independent data, variations due to independent processes	Steps perform better than trends to explain autocorrelation and heteroscedasticity	None located
Autocorrelation and heteroscedasticity quarterly observations surface temp 1979–2014	Trends fail 40-y autocorr 20%, steps 0%, accumulated error trends/steps 2.9 Little difference heterosc.	Trends serially independent data, variations due to independent processes data	Steps perform better than trends to explain autocorrelation and heteroscedasticity	None located
Autocorrelation and heteroscedasticity quarterly observations satellite temp 1979–2014	Accumulated error trends/steps 4.4, 0.9 and 3.1, 2.1 RSS & UAH Trends and steps little difference autocorr. and heterosc. (except steps 24% v 8% heterosc.)	Trends serially independent data, variations due to independent processes data	Steps perform better than trends to explain cumulative error, little difference autocorrelation and heteroscedasticity	None located

Discussion

It would be reasonable to ask the question – if shifts in temperature and other variables are ubiquitous within the climate system, why have they not been recognised earlier? This question has been explored at length in related papers that cover the following points:

- The history and philosophy of gradualisms and trend analysis as its key tool for understanding how the world works has its origins in the scientific enlightenment and since then has defined H1 as the dominant paradigm of climate change (Jones, 2015b). This has been reinforced by the success of methods for long-term trend analysis (Jones, 2015a). Phenomena that do not fit this model are labelled as noise and considered to be random.
- The value-laden framing of the signal-to-noise model in defining what information is useful for decision making and what is not (Koutsoyiannis, 2010;Jones, 2015b).
- The great success of ordinary least squares and related tests that use linear statistical methods in explaining climate phenomena, covering methods such as timeseries analysis, pattern matching (Santer et al., 1990;Hasselmann, 1993;Mitchell, 2003a), vector analysis and its application to understanding climate processes (Hasselmann, 1979;North et al., 1995), detection and attribution (Hegerl et al., 2007;Stott et al., 2010) and development of climate projections (Hulme and Mearns, 2001;Mitchell, 2003b;Whetton et al., 2005).
- The difficulty in analysing change points in complex data and achieving a clear error judgement using Neyman-Pearson testing (Type I and II errors using pH_0<threshold), and linking that to specific hypotheses (see Mayo, 2010).
- The climate wars, specifically the role of steps and trends, where trends are associated with climate change theory and steps with opposition to the theory (Cahill et al., 2015;Foster and Abraham, 2015;Lewandowsky et al., 2015;Skeptical Science, 2015;Lewandowsky et al., 2016). This has become a situation where methods are being held as representative of particular theoretical positions (Jones and Ricketts, 2016).
- The cognitive values attached to parsimony or Occam's razor, where a phenomenon should be described in the simplest terms possible. Applied to statistics, its main aim is to avoid over-fitting. However, in a complex physical system, a statistically simple relationship may be energetically complex (Jones and Ricketts, 2016). This has not been a factor under consideration to date.
- The link between model skill and predictive capacity is defined by the analytic framework applied. For example, seamless links between weather and climate forecasting over a range of timescales are a key scientific target (Palmer et al., 2008;Hoskins, 2013). The Global Framework for Climate Services (World Meteorological Organization, 2011), reflects this: *Weather and climate research are closely*

intertwined; progress in our understanding of climate processes and their numerical representation is common to both. Seamless prediction (on timescales from a few hours to centuries) needs to be further developed and extended to aspects across multiple disciplines relevant to climate processes (World Meteorological Organization, 2010). Solomon et al. (2011) state that *"Long experience in weather and climate forecasting has shown that forecasts are of little utility without a priori assessment of forecast skill and reliability"*. The assumption that the processes involved are timescale invariant indicate that the meaning of seamless has not really been thought through. For the moment, seamless means a concentration on mean change and other variables that show skill in climate models. However, skill is measured according to the H1 signal to noise construct and would like quite different if analysed in H2 mode (Jones, 2015b). This framing also overlooks the considerable literature on scenarios that has arisen because long-term predictions under considerable uncertainty tend to fail (Wack, 1985a, b;Börjeson et al., 2006).

- The different areas of scientific knowledge and expertise required to understand the climate system. In particular, the relative roles of radiative physics largely understood as being linear and hydrometeorology, with its substantial nonlinear behaviour, remain largely unreconciled.

If mean global atmospheric warming is accepted as an ill-posed mathematical problem, a single test passing a pH_0 threshold cannot adequately represent the various influences present, and a more applied approach is required. This involves undertaking severe statistical testing informed by a process-based understanding of how the climate may change.

Technically, trend analysis and the Maronna-Yohai bivariate test face similar limitations, with respect to the serial independence and normal distribution of the input data, but the former has widespread acceptance whereas the latter is unfamiliar, creating different degrees of trust. Objectively, if the data they analyse is subject to lagging or unit-root processes then the likelihoods expressed by either test will be compromised (e.g., Cohn and Lins, 2005;Koutsoyiannis, 2010). We have been quite open about this with respect to the bivariate test, and it has informed how these tests are applied here.

Most challenges to trend-like behaviour in surface warming are associated with contrarian positions that either seek to repudiate the theory of greenhouse gas driven climate change entirely, or maintain that the risk described by groups such as the IPCC is overstated (McKitrick, 2014, 2015;Tisdale, 2015). In particular, this controversy has surrounded the question of whether warming paused or entered a hiatus from about 1998 or continued unabated. Much of the recent statistical analysis of global atmospheric warming has concerned this issue.

For example, Cahill et al. (2015) recently published a segmented trend model for global mean surface temperature with change points around 1912 (1907–1920, p<0.05 limits), 1940 (1934–1948) and 1970 (1963–1979). If this model is subject to the same tests as in Tables 4 and 5 for the five mean global surface warming records in Fig. 2, the results are similar to the step and trend model used here, so it does produce very low residual error. Using likelihood ratios or any similar measure does not distinguish between the segmented trend and step and trend models, therefore a focus on the probative aspects of severe testing is required, linking h_{trend} with *H1*.

Cahill et al. (2015) used a Bayesian belief approach, stating that step changes are physically implausible: "Isolated pieces of trend line with sudden temperature changes between them (i.e. a 'stairway model') would not provide a physically plausible model for global temperature given the thermal inertia of the system". Although they do not specify the exact physical process, this presumably refers to the thermal inertia of the ocean. This conforms with the description of the ocean-driven component of *H1* described in Sect 2.3: most of the heat generated by added greenhouse gas forcing goes into the ocean and is gradually released into the atmosphere, mediated by the rate of shallow and deep-ocean mixing. It is not clear whether Cahill et al. (2015) refer to a process whereby heat absorbed into the ocean at varying rates, or is released by the ocean at varying rates, both alternatives potentially being mediated by the relationship between shallow and deep-ocean mixing.

In any case, Cahill et al. (2015) explicitly reject h_{step}. In doing so they are implicitly claiming to meet Slingo's (2013) caution that a statistical model be well specified. In their conclusions they restate *H1* – "recent variations in short-term trends are fully consistent with an ongoing steady global warming trend superimposed by short-term stochastic variations" (Cahill et al., 2015). Foster and Abraham (2015) reject discontinuous change for the same reason and extensively test trend-related models to reach a similar conclusion. Such claims also implicitly reject a host of studies that have detected step changes in temperature data (Table 6). Presumably if those studies have all committed Type I errors, the models involved in detecting such changes – the bivariate test, Rodionov's (2006) STARS test and others, would also be invalidated for homogeneity testing of climate data, their other main use. If step changes cannot physically exist in the data, the tests that have detected them are invalid and the homogeneity adjustments to climate records made on the basis of such tests are likewise invalid. This suggests that such arguments have a very narrow focus and are inconsistent with the bigger picture.

Many studies have applied statistical techniques to extract the noise from temperature data to diagnose the signal. For example, Foster and Rahmstorf (2011) remove solar, volcanic and ENSO influences through multilinear regression on a monthly basis from 1979–2010, concluding that the remaining data more closely follow a single trend. However, if ENSO is coupled with regime changes and steplike warming, the regression relationships for ENSO will contain part of the signal. Due to constraints limiting those observations to the satellite-observation period, the fitting data is also the test data. Zhou and Tung (2013) undertook a similar analysis (1856–2010) using non-satellite data for ENSO, adding the Atlantic Meridional Oscillation (AMO), which results in a lower trend for the latter period analysed by Foster and Rahmstorf (2011). However, the AMO is likewise potentially involved in nonlinear changes on decadal timescales, involving rapid shifts in temperature.

A series of studies has explicitly examined climate shifts in oscillatory modes of climate variability on decadal time scales (Swanson et al., 2009;Swanson and Tsonis, 2009;Wang et al., 2009;Tsonis and Swanson, 2011;Tsonis and Swanson, 2012;Wang et al., 2012). If these signals are extracted, a monotonic accelerating curve during the 20th century remains (Swanson et al., 2009). Although they describe these as climate shifts that have the same timing as those in this paper, according to Tsonis and Swanson (2012), these shifts manifest as a change in global temperature trend. This frames shifts as modulating trends, whereas our analysis suggests that the shifts are primary and trends are secondary.

As we discuss in a related paper where *H2* is examined in greater detail, the *H1* hypothesis comes from a radiative-centric view of the climate system, which if it considers hydrodynamics at all, regards it as an independent process (Jones and Ricketts, 2016). In a coupled climate system, this makes little sense. Radiative processes are additive (Ozawa et al., 2003), so cannot supply heat energy in bursts unless directly forced.

However, hydrodynamic processes are nonlinear and are quite capable of supplying the energy required (Ozawa et al., 2003;Lucarini and Ragone, 2011;Ghil, 2012). To suggest the step-wise release of that heat energy is physically implausible overlooks the energetics of the ocean-atmosphere system. The atmosphere contains as much heat energy as the top 3.2 m of ocean (Bureau of Meteorology, 2003). About 93% of historically added heat currently resides in the ocean (Roemmich et al., 2015), thirty times that of the atmosphere. Between 1955 and 2010, the amount of heat added to the atmosphere was about 0.8 x 10^{22} Joules, compared with the 24.0 x 10^{22} Joules added to the top 2000 m of the ocean (Levitus et al., 2012). A physical re-organisation of the ocean-atmosphere system, as part of a regime shift, is large enough to provide the relatively small amount of energy required to cause abrupt sea surface and atmospheric warming.

For example, Reid et al. (2015) in describing the late 1980s regime shift, show it was associated with a large scale shift in temperature and multiple impacts across terrestrial and marine systems, mainly in the northern hemisphere. Changes in the North Pacific in 1977 were considered even more extensive (Hare and Mantua, 2000) as were those in 1997–98 that involved both the Pacific and Atlantic Oceans (Chikamoto et al., 2012a;Chikamoto et al., 2012b).

One important test for a hypothesis is whether it can offer explanations and/or novel predictions not contained within the original scope. This has not been the aim of much recent work, which is focused on whether the period after 1998 was a hiatus, pause or uninterrupted trend. The aim of papers like Cahill et al. (2015) and Foster and Abraham (2015) was to show that the year 1998 was unexceptional and that the so-called 'pause' was part of a longer term trend. Underpinning this claim is that there are no steplike changes within the last part of the record from 1970–2014. The most extraordinary claim of this type was made by Rajaratnam et al. (2015) who examined the segmented trends either side of 1998, concluding they were statistically identical. In doing so, they illustrated those trends as being separated by a gap of almost 0.2 °C as shown in Fig. 2, which they completely ignored in their analysis. Elsewhere, we examine the 1997/98 paper in greater detail, asking whether the climate system is currently undergoing another shift in warming (Jones and Ricketts, 2016).

Meehl (2015) and Slingo (2013) emphasise the importance of having a process-based understanding of how the temperature changes. Temperature change on periods of less than fifty years has not been severely tested in this regard, much of the recent work being defensive rather than innovative.

Conclusions

Here, we have adapted and applied severe testing principles proposed by Mayo and Spanos (2010) to determine the role step changes play in decadal-scale warming. This involves the linking of scientific hypotheses *H* with statistical hypotheses *h*, in order to test *h/H* and *-h/-H* to the point where the test agrees/disagrees with the hypothesis/null with a very high probability of distinguishing between the two. Specifically, the scientific hypothesis – that externally-forced and internally-generated climate processes interact with each other (*H2*) instead of acting independently (*H1*) – is shown by the statistical hypothesis h_{step} passing a series of steps in better shape than h_{trend}.

This finding does not invalidate the huge literature that assesses long-term (>50 years) climate change as a relatively linear process, and the warming response as being broadly additive with respect to forcing (e.g., Lucarini et al., 2010;Marvel et al., 2015). However, on decadal scales, this is not the case – warming appears to be largely governed by a storage and release process, where heat is stored in the ocean and released in bursts projecting onto modes of climate variability as suggested by Corti et al. (1999). We discuss this further in another paper (Jones and Ricketts, 2016).
This has serious implications for how climate change is understood and applied in a whole range of decision-making contexts. The characterisation of changing climate risk as a smooth process will leave climate risk as being seriously underdetermined, affecting how adaptation is perceived, planned and undertaken (Jones et al., 2013).

The interaction of change and variability is typical of a complex, rather than mechanistic, system. The possibility of Lorenzian attractors in the ocean-atmosphere acting on decadal time scales was raised by Palmer (1993) and, despite later discussions about the potential for nonlinear responses on those timescales (e.g., Lucarini and Ragone, 2011;Tsonis and Swanson, 2012), very little progress has been made in translating this into applied research that can portray a better understanding of changing climate risk. This may be due in part to science asking the wrong questions.

The signal to noise model of a gradually changing mean surrounded by random climate variability poorly represents warming on decadal timescales. The separation of signal and noise into 'good' and 'bad', likewise, is poor framing for the purposes of understanding and managing risk in fundamentally nonlinear systems (Koutsoyiannis, 2010;Jones, 2015b). However, as we show, the presence of such changes within climate models shows their current potential for investigating nonlinearly changing climate risks. Investigating step changes in temperature and related variables does not indicate a need to fundamentally change how climate modelling is carried out. It does, however, indicate a need to change how the results are analysed.

Supplementary Information

Statistical method

The principal method used in this paper is the Maronna-Yohai (1978) bivariate test using a modified formulation of Bücher and Dessens (1991b) with reference to Potter (1981b).

This formulation is from Bücher and Dessens (1991b) based on normalised data with no trend. It takes advantage of the normalisation step to allow simpler structure. According to the authors, it was obtained from Potter, although the method published in Potter (1981b) is essentially that of Maronna and Yohai (1978).
In the following, primes reference un-normalised data and functions, normalised data in step 2 is denoted by removal of primes. This usage has been slightly modified from Bücher and Dessens (1991b). Additionally, the second part of (Eqn. S4) corrects an inconsistency in that paper.

Let $x_i',\ i = 1 \ldots n$ be a stationary reference time series and $y_i',\ i = 1 \ldots n$ be a test time-series which is assumed to correlate to x' except for a single shift at some time i_0.
Step 1. Standardize series.

$$\overline{X'} = \frac{\sum_{j=1}^{n} x_j'}{n},\ \overline{Y'} = \frac{\sum_{j=1}^{n} y_j'}{n},\ S_x' = \left(\frac{\sum_{j=1}^{n}\left(x_j' - \overline{X'}\right)^2}{n}\right)^{1/2},\ S_y' = \left(\frac{\sum_{j=1}^{n}\left(y_j' - \overline{Y'}\right)^2}{n}\right)^{1/2} \quad \text{(S1)}$$

$$x_j = \frac{\left(x_j' - \overline{X'}\right)}{S_x'},\ y_j = \frac{\left(y_j' - \overline{Y'}\right)}{S_y'} \text{ for all } j \le n. \quad \text{(S2)}$$

Step 2. Compute test statistics.

$$S_{xy} = \sum_{j=1}^{n} x_j y_j \quad \text{(S3)}$$

$$X_i = \frac{\sum_{j=1}^{i} x_j}{i},\ Y_i = \frac{\sum_{j=1}^{i} y_j}{i} \text{ for all } i < n \quad \text{(S4)}$$

$$F_i = n - \frac{{X_i}^2 ni}{(n-i)} \text{ for all } i < n \quad \text{(S5)}$$

$$D_i = \frac{\left(S_{xy} X_i - nY_i\right)n}{(n-i)F_i} \text{ for all } i < n \quad \text{(S6)}$$

$$T_i = \frac{\left[i(n-i)D_i^2 F_i\right]}{\left(n^2 - S_{xy}^2\right)} \text{ for all } i < n. \quad \text{(S7)}$$

$$T_{i0} = \max(T_i),\ i_0 = i \text{ when } T_{i0} = \max(T_i)\ . \quad \text{(S8)}$$

The time associated with T_{i0} represents the time at which a change occurs and its successor is the first time of the new regime. A mean shift can be computed. For the null trend case, critical values of T_i are provided by Maronna and Yohai (1978) for two-tailed, alpha levels of (0.1, 0.05, and 0.01) for the null hypothesis of no change, given time series lengths of 15, 20, 75 and Potter (1981b) provides these for 100. An interpolating function is used to generalize these results for time series of varying length.

Multi-step bivariate test: application

The purpose of constructing a rule-based process for analysing multiple step changes in a time series, is to remove the need to make individual decisions that utilise the experimenter's judgement, as was the case in earlier papers. It also allows multiple sampling and the addition of randomness, which increases the robustness of the results. The model and its testing is further described in (Ricketts and Jones, 2016).

The bivariate test, rule-based process and diagnostics are coded in Python 2.7.6, developed with the Spyder environment (© 2009-2012 Pierre Raybaut), running in 32-bit Windows 7 and 64-bit Windows 8.1 environments. Moba-Xterm PE v7.2, a windows based Unix emulator, was used to support some collation of results, and data acquisition. Output data was compiled into *.csv files for further testing.

Description

The method is a technique for segmenting time series with zero to many step changes in the mean. The test returns a list of break-points that divide a time series into segments bounded by statistically significant step changes, except for the start and the end. The routine consists of a *screening pass*, which produces a first approximation break-list. This break-list is iteratively refined by a *convergent pass*. Both passes are described below. Each application of the test is subject to a *resampling test*, which determines the resilience of a step-point determination to noise. One hundred iterations sample a test series against a randomly selected reference time series.

The method is probabilistic. Each iteration returns a list comprising a set of shift points, their timing and magnitude and null probability against a serially independent reference, along with a variety of diagnostic variables. A time series with distinct step changes will return the same list for a set of iterations, whereas others may yield several variations. This is especially the case for areal averages that integrate local changes from two or more regions, data with quality issues, or where autocorrelation due to trending behaviour or other processes is present.

Resampling test. The bivariate test is repeated 100 times using different random sequences and the i_0 values and associated T_{i0}, and shifts are collated by mode.

1. On the screening pass only the modal value is examined.
2. On the convergent pass additional selection rules apply. The mode and second mode are examined. There may be a single mode (e.g., 100% selection), or the two modes may be close together or well separated.

The i_0 (time i preceding the shift) returned by the resampling test is the modal value (i.e., most frequent) of each test. Similarly T_{i0}, and shift magnitude are the mean of those values associated with i_0. A segment contains a breakpoint in position i if T_{i0} exceeds the critical T_i value for segment length with a given probability.

Screening pass. This is a binary segmentation technique, similar to that used in similar applications (Scott and Knott, 1974;Killick et al., 2012). The entire time series is analysed for a single breakpoint using the resampling test (100 iterations). If T_{i0} is significant ($p<0.01$), then the segment up to and including i_0 is analysed for an earlier break, and the segment after i_0 is analysed for a later break. This process is repeated for the sub-segments so formed until no significant breaks are found. The result is a series of breakpoints which are then refined on the convergent pass. Because breakpoints found on this pass are returned on the basis of a recursive process, end point effects caused by sampling time series of different lengths may influence the results.

The role of the convergent pass is to combine segments to determine whether the screening pass has oversampled for steps and also to ensure that the selected break points are robust within the selected segmentation.

Convergent pass. The list of n breakpoints from the screening pass breaks the original time series into $s=n+1$ segments. The algorithm then works its way from earliest to latest segments combining consecutive segments into one, and then searching within that segment using binary segmentation to produce a *candidate list* from which two most frequent i_0 values are retained (in practice there are rarely more than two and usually just one). There are two special cases, segments 1 and s which are analysed individually at either end of this process to cover the impact of end point adjustments. This procedure will sometimes reduce the number of step changes from the screening pass.

The convergent pass is reiterated until it produces the same list for a second time and this is returned as the final result.

This pass incorporates **some decision rules**.

1. A prohibition period of seven years is applied at the start and the end of the time series, and after a break point before another point will be accepted. This is because the bivariate test is sensitive to end effects.
2. If the modal year is within the prohibition period from the previous break point, then the two are compared. A resample test is conducted by extending the segment backwards to the start date of the previous segment. If it is a valid break, this then replaces the previous break, otherwise the previous break is retained, and a small "safety margin" is added for one iteration to the low bound of the first segment next time round to prevent the point being re-selected.
3. If the mode is equal to or >90%, or the first mode is >50% and the second mode is >20% then the modal year is accepted, else it is dropped.
4. If after this, a segment contains a single breakpoint, it is retained.
5. If the candidate list is empty, that is, a segment no longer contains a breakpoint then the two segments are merged and treated as a single segment on the next iteration.
6. If the candidate list contains more than one point then the earliest two are retained and the rest discarded. The two points are then trialled using a resampling test to determine if the interval up to the later of the two still contains a break, and if this is still present, it is retained. If not, then the second candidate is similarly tested.

Considerations about the multi-step bivariate test

The role of time

This analysis treats time unidirectional. The bivariate test itself selects the last time *before* a change so is asymmetric. The convergent pass operates from earliest to latest, revising provisional breakpoints and then using them to delineate later breakpoints. On each pass, as soon as a breakpoint is provisionally established, it is used, and no other information is preserved. Therefore, every breakpoint is complete unto itself, and the segment within which it is embedded has its own statistics. *This means that the final set of segments, broken up by the final set of breakpoints,* consists of a series of segments, each of which has its own variance, mean and shape parameters, and embedded trend. The analysis treats each segment as independent, but whether physical dependence (including memory) or otherwise, can be assumed, remains to be assessed. Information theory approaches to assessing statistical best fit may also break down because the system may not deliver the same information between break-points. This is certainly the case for other complex systems covering economics and ecology.

End point effects

The determination of a breakpoint in a time series is sensitive to all of the data including the first and last elements, but is less reliable near the start and end of that series (Vivès and Jones, 2005). This affects two aspects of the analysis:

1. Previously determined shift points become end points when a subsequent segment is tested, making nearby observations more sensitive to end effects. This is the principal reason for the 7-year restriction on break-points. However, temperature data can also produce peaks/troughs due to interannual variability giving a multi-modal distribution of potential break points clustered around an underlying shift. Altering segment lengths can potentially alter the distribution of these modes, therefore leading to the application of decision rules 1, 2 and 6 above, to determine the most robust outcome.
2. The choice of starting year.
    i. The quality of long-term climate data characteristically degrades backwards in time. This may produce artificial break-points, move existing ones, or simplify 'natural' variability. Autocorrelation may also be introduced by some infilling methods. Early shift points should not be regarded as being as reliable as more recent dates.

ii. If the data record starts just before a true shift point or is influenced by a truncated sequence of low or high years then the next (and to a lesser extent, consequent) dates may be affected.

On balance it is much better to start an analysis from the earliest date available, unless data quality is clearly compromised.

Assumptions

The bivariate test itself assumes two variates that are both stationary, except that the test variate may have a single point change in the mean. A time series that has multiple shift points will register only one value of T_{i0} and other points cannot be relied upon. Sometimes the original value of T_{i0} will be removed once the time series is segmented, because it is an 'average' of several others.

The assumption of serial independence is very important. A number of studies have concluded that observed annual temperature and rainfall records fulfil that stricture. Such climate data may also contain components of autocorrelation and variable trends. Some of the variability of trend may simply be redness (drift due to persistence of previous values), some may represent transient processes. However, autocorrelated data such as regional or global mean sea level, sea surface temperature or climate data divided into a monthly or quarterly time series may lead to statistical significance being over-estimated, although the timing of a shift will remain accurate. In the paper, we use the t-test to address this, but when autocorrelation is due to a sustained trend in a time series that contains both steps and trends, the t-test also will give misleading results.

Here we treat annual temperature data as a signal composed of a small but arbitrary number of linear segments delineated by step changes, and embedded in Gaussian noise. The impact of trend is on the assigned significance of the shift returned from the bivariate test, rather than its timing. Additionally, a change of trend when there is no step may cause the bivariate test to allocate a step some years after the change of trend. These can all be determined in post-processing.

A time series that contains nothing but a general trend and variation, will have two properties when analysed by our method:

1. The sum of steps will converge on zero.
2. The probabilistic test will be dominated by random sampling of the reference variate and the number of different break-point lists will increase – that is, it will return unstable sets.

Diagnostics

Every iteration of the 100 break-list runs that comprise an analysis produces a csv file of results, plus a trace of the decision process. The trace file contains the initial data as well as a summary of the break dates with some QA diagnostics. All 100 trace files are collated and the diagnostics are given for each analysis. This includes Ti_0, Shift, Modal Year, Modal Frequency, The Second Modal Year and its frequency.

Terminology

The language of non-linear change is nowhere near as established as is the language for trend analysis. Here we use the following terms in the ways described:

- Break, break-point, break-year: a break denotes an abrupt change in statistical characteristics of any kind (e.g., change in trend, variance).
- Shift: in the paper a shift is the distance between the end of one internal trend and the beginning of the next across a step change.
- Step: an abrupt step-like change as measured by the test.

Calibration of the method.

The method has been calibrated against synthetic data composed with variable lag one/seven autocorrelation, variable number of shift points, varying trends and changes of trend. Its performance has also been tested for its ability to locate a randomly timed shift point in a random series to which is added varying shifts, varying

trends up to those well in excess of any climate model run, and simulating a random shift month within the simulated shift year.

Data sources

Global mean surface temperature

Time series tested are mean annual global air temperature anomalies from five groups (GISS, HadCRU, NCDC, C&W and BEST), hemispheric temperatures from three groups (HadCRU, NCDC and GISS) and zonal temperatures from two groups (NCDC and GISS). Tropospheric satellite temperatures from two groups (RSS and UAH) are also tested (Table S1).

Table S1: Source groups for 20th century observations, surface and satellite.

Name	Version	Download date	Base Period	Global	Hemi-spheric	Zonal	Land-Ocean	References
BEST		15 Jan 2015	1951–1980	Y	N	N	N	(Rohde et al., 2013a;Rohde et al., 2013b)
COWTAN & WAY	2.0	15 Jan 2015	1961–1990	Y	N	N	N	(Cowtan and Way, 2014b)
GISSTEMP3	V3	15 Apr 2015	1951–1980	Y	Y	N	N	(Hansen et al., 1988;GISSTemp Team, 2015)
HadCRUT4 HadSST3 CRUT4	4.3.0.0 3.1.1.0 4v	25 May 2015	1961–1990	Y	Y	Y	Y	(Jones et al., 1999;Jones et al., 2001;Brohan et al., 2006;Rayner et al., 2006;Kennedy et al., 2011;Jones et al., 2012;Morice et al., 2012b;Osborn and Jones, 2014)
NCDC	v3.5.4. 201504	18 Mar 2015	"20th C Average"	Y	Y	Y	Y	(Smith et al., 2008;Vose et al., 2012)
Satellite based atmospheric temperature estimates, Lower Troposphere to Lower Stratosphere								
RSS	V03.3	7 May 2015	1979-1998	Y	Y	Y	Y	(Mears et al., 2003;Mears and Wentz, 2009c, b)
UAH	6.0.beta	5 May 2015	1981-2010	Y	Y	Y	Y	(Christy et al., 2000)

NDCD zonal data version v3.5.4.201504.

Annual and monthly files in ASCII format covering land, ocean, and combined land and ocean were downloaded on 29 May 2015 from ftp://ftp.ncdc.noaa.gov/pub/data/mlost/operational/products/ using wget in recursive mode. Each file contains data for one zonal average and for one of land, ocean and combined land and ocean.

The zonal averages were over: 90°S–90°N (Global), 90°S–0°S (°Southern hemisphere), 0°N–90°N (Northern hemisphere), 90°S–20°S, 60°S–30°S, 60°S–60°N, 30°S–0°N, 0°N–30°N, 20°S–20°N, 20°N–90°N, and 60°N–90°N. Data in the files labelled as 90°S–60°S for all three subsets was clearly corrupted on receipt and was not used. The data format is documented on-line in the file ftp://ftp.ncdc.noaa.gov/pub/data/mlost/operational/products/readme.timeseries, Annual averages are as provided, rather than simple averages of monthly values.

GISSTEMP_3

Data was downloaded on 15 April 2015 in ASCII from http://data.giss.nasa.gov/gistemp/tabledata_v3/ZonAnn.Ts+dSST.txt and format converted to CSV for use. All values are multiplied by 0.01 to produce degrees C, as per the metadata in the file.

Cowtan and Way

Data representing annually averaged was downloaded in ASCII format on 15 Jan 2015, from http://www-users.york.ac.uk/~kdc3/papers/coverage2013/had4_krig_annual_v2_0_0.txt
Both annual and monthly data were downloaded but this initial analysis was of the annual data only.
Data is described at http://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html.

Berkeley Earth

Data representing annual averaged mean global temperature was downloaded in ASCII format on 15 Jan 2015 from http://berkeleyearth.lbl.gov/auto/Global/Land_and_Ocean_summary.txt
Two versions are present in the file. The data used in this study is from column 1, 'Annual Anomaly' computed by extrapolation of temperature in the presence of sea ice by using land-air temperature surface anomalies.

NCDC Land, Ocean, and combined Land and Ocean data

Seasonal analysis was based on data downloaded on 18 Mar 2015, as individual csv files, one per month, using the wget utility from http://www.ncdc.noaa.gov/cag/time-series/global/$extent/$set/1/1/*.csv where $extent is replaced by one of ["global","nhem", "shem"] and $set is one of ["land","ocean","land_ocean"]. The year 2015 is not complete and corresponding values were ignored.

Seasonal averages were computed as simple averages of the monthly values.
Annual averaged data was downloaded interactively from http://www.ncdc.noaa.gov/cag/ on 26 May 2015 (the same site) using 12 Month time scales to December for global and hemispheric extents giving a total of nine files.

Hadley/CRU Land, Ocean, Land and Ocean data

Data reported here was downloaded on 25 May 2015 as ASCII text files from
http://www.metoffice.gov.uk/hadobs/ File formats are described algorithmically at
http://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/series_format.html
Monthly and seasonal analyses were performed using the appropriate monthly values, corresponding annual averages were drawn from the last column.

Satellite derived lower tropospheric temperature data, RSS and UAH

RSS

The front page for this organisation is at http://www.remss.com/. Information on upper air temperatures is at http://www.remss.com/measurements/upper-air-temperature.
One complex data set is provided, Temperature of Lower Troposphere (TLT), "constructed by calculating a weighted difference between measurements made at different Earth incidence angles to extrapolate MSU channel 2 and AMSU channel 5 measurements lower in the atmosphere."

Data for Land, Ocean, and Land and Ocean were downloaded in a simpler ASCII format, all bands on one line per month per year, on 7 May 2015 from ftp://ftp.remss.com/msu/data/uah_compatible_format
Data files are from Jan 1979 to present.

Anomalies are computed by subtracting the mean monthly value determined by averaging 1979 through 1998 data for each channel from the average brightness temperature for each month. The set of 12 month means for 1979 to 1998 are included in the netCDF files available on the ftp server (ftp.remss.com/msu)

UAH

These data are version 6.0.beta.
UAH: Data were downloaded on 5 May 2015 from http://vortex.nsstc.uah.edu/data/msu/v6.0beta/.
http://vortex.nsstc.uah.edu/data/msu/v6.0beta/tlt/uahncdc_lt_6.0beta1
A readme file is at http://vortex.nsstc.uah.edu/data/msu/docs/readme.msu.

Anomaly base period, as per metadata and confirmed by NCDC, 1981–2010.

Model Data

Data used are simulated annual mean surface temperature from the Climate Model Intercomparison Project (CMIP)3 and CMIP5 archives.

CMIP3/AR4

Data were downloaded under script control 17 July 2014. Data were also reloaded and cross checked from the KNMI data explorer web site on 25 Feb 2015 as per the CMIP5 data below. In all, 102 model runs were downloaded, with 14 being ensembles, and the rest being independent runs.

Within the metadata for each file are model name and identifiers, which are either run<N> or E<L> where L is the list of run numbers in an ensemble average. The models, their forcing, and run and ensemble numbers are listed in Supplementary Table 1. Models were forced by observed natural and anthropogenic factors to 2000 or 2001, and by SRES scenarios A1b or A2 through to 2099 or 2100. The BCC model is an exception, being forced by the SRESA2 scenario from 1871.

Table S2: List of modelling groups and global climate models used for simulations of 20th and 21st century climate, available from the CMIP3 database managed by PCMDI http://www-pcmdi.llnl.gov/ipcc/info_for_analysts.php. The forcing factors for 20th century climate are: G – Well-mixed greenhouse gases, O – Ozone, SD – Sulfate direct, SI – Sulfate indirect, BC – Black carbon, OC – Organic carbon, MD – Mineral dust, SS – Sea salt, LU – Land use, SO – Solar irradiance and V – Volcanic aerosol. Updated from (CSIRO and BoM, 2007).

Originating Group(s), Country	**Model**	**Forcings used in model simulations**	**Scenarios**	**Runs & (E)nsembles**	**Start date**
Bjerknes Centre for Climate Research, Norway	BCCR	G, SD	SRESA1b	1	1850
Beijing Climate Center, China	BCC	G, SD*	SRESA2	1	1871
Canadian Climate Centre, Canada	CCCMA T47	G, SD	SRESA1b	1–5, E1–3	1850
			SRESA2	1	1850
Canadian Climate Centre, Canada	CCCMA T63	G, SD	SRESA1b	1	1850
Meteo-France, France	CNRM	G, O, SD, BC	SRESA1b	1	1860
			SRESA2	1	1860
CSIRO, Australia	CSIRO-MK3.0	G, O, SD	SRESA1b	1	1871
			SRESA2	1	1871
CSIRO, Australia	CSIRO-MK3.5	G, O, SD	SRESA1b	1	1871
			SRESA2	1	1871
Geophysical Fluid Dynamics Lab, USA	GFDL 2.0	G, O, SD, BC, OC, LU, SO, V	SRESA1b	1	1861
			SRESA2	1	1861
Geophysical Fluid Dynamics Lab, USA	GFDL 2.1	G, O, SD, BC, OC, LU, SO, V	SRESA1b	1	1861
			SRESA2	1	1861
NASA/Goddard Institute for Space Studies, USA	GISS-AOM	G, SD, SS	SRESA1b	1–2, E1–2	1850
NASA/Goddard Institute for Space Studies, USA	GISS-E-H	G, O, SD, SI, BC, OC, MD, SS, LU, SO, V	SRESA1b	1–3, E1–3	1880
NASA/Goddard Institute for Space Studies, USA	GISS-E-R	G, O, SD, SI, BC, OC, MD, SS, LU, SO, V	SRESA1b	1–4, E1–4	1880
			SRESA2	1	1880
Instituto Nazionale di Geofisica e Vulcanologia, Italy	INGV	G, SD	SRESA1b	1	1870
			SRESA2	1	1870
LASG/Institute of Atmospheric Physics, China	IAP	G, SD	SRESA1b	1–3, E1–3	1850
Institute of Numerical Mathematics, Russia	INMCM	G, SD, SO	SRESA1b	1	1871
			SRESA2	1	1871
Institut Pierre Simon Laplace, France	IPSL	G, SD, SI	SRESA1b	1	1860
			SRESA2	1	1860
Centre for Climate Research, Japan	MIROC-H	G, O, SD, BC, OC, MD, SS, LU, SO, V	SRESA1b	1–3, E1–3	1850
			SRESA2	1–3, E1–3	1850

Centre for Climate Research, Japan	MIROC-M	G, O, SD, BC, OC, MD, SS, LU, SO, V	SRESA1b	1	1900
Meteorological Institute University of Bonn, Meteorological Research Institute KMA, Germany/Korea	MIUB	G, SD, SI	SRESA1b	1–3, E1–3	1860
			SRESA2	1–3, E1–3	1860
Max Planck Institute for Meteorology DKRZ, Germany	MPI-ECHAM5	G, O, SD, SI	SRESA1b	1–4, E1–3	1860
			SRESA2	1–3, E1–3	1860
Meteorological Research Institute, Japan	MRI	G, SD ,SO	SRESA1b	1–5, E1–5	1851
			SRESA2	1–5, E1–5	1851
National Center for Atmospheric Research, USA	NCAR-CCSM	G, O, SD, BC, OC, SO, U	SRESA1b	1–4,9, E1–2	1870
			SRESA2	1–4	1870
National Center for Atmospheric Research, USA	NCAR-PCM1	G, O, SD, SO, V	SRESA1b	1,4	1890
			SRESA2	1–4, E2–4	1890
Hadley Centre, UK	HADCM3	G, O, SD, SI	SRESA1b	1	1860
			SRESA2	1	1860
Hadley Centre, UK	HADGEM1	G, O, SD, SI, BC, OC, LU, SO, V	SRESA1b	1	1860
			SRESA2	1	1860

CMIP5/AR5

Data were downloaded from the KNMI data explorer web site http://climexp.knmi.nl/ RCP4.5, RCP6.0 and RCP8.5 (7 Jan 2015), RCP2.6 (19 Feb 2015). Files were renamed under script control using metadata within the files to simplify. Each line contains a year and one entry per month. Annual averages are calculated as simple averages of model months.

Notes:

1. The second line of metadata specifies the variable (tas: Temperature at Surface), the climate model the driving emissions prescription/RCP and an *ensemble member identifier* composed of three parts (r9i1p1), as described in the CMIP5 reference syntax (http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax_v0-25_clean.pdf).

   The ensemble member template is (r<N>i<M>p<L>). Identifiers relevant here are the run number (r(N)) and physics perturbation (p(L)).
2. Model calendars in general do not reflect real world ones. So some models assume 30 days per month for a 360 day year, some assume 365 day years with no leap years and a 28 day February. For simplicity, here annual averages are simple averages of 12 monthly values.

Four multi-model ensembles were analysed: RCP2.6 (61 members), RCP4.5 (107 members), RCP6.0 (47 members) and RCP8.5 (80 members). Details are listed in Table SI3.

Table S3: List of modelling groups and global climate models used for simulations of 20th and 21st century climate, available from the CMIP5 database http://cmip-pcmdi.llnl.gov/cmip5/availability.html, with run numbers (r(N)) and physics perturbations (p(L)), and equilibrium climate sensitivity (ECS). ECS is taken from Sherwood et al. (2014) unless otherwise noted. If not allocated otherwise, runs have the physical perturbation p1.

Centre	Model	RCP2.6	RCP4.5	RCP6.0	RCP8.5	ECS
BoM/CSIRO, Australia	ACCESS1-0		r1		r1	3.79
BoM/CSIRO, Australia	ACCESS1-3		r1		r1	3.45
Beijing Climate Center, China	BCC-CSM1-1	r1	r1	r1	r1	2.88
Beijing Climate Center, China	BCC-CSM1-1-M	r1	r1	r1		
Beijing Normal University, China	BNU-ESM	r1	r1		r1	4.11
Canadian Climate Centre, Canada	CanESM2	r1–5	r1–5		r1–5	3.68
National Center for Atmospheric Research, USA	CCSM4	r1,3–6	r1–6	r1–6	r1–6	3.20[1]
National Center for Atmospheric Research, USA	CESM1-BGC		r1		r1	
National Center for Atmospheric Research, USA	CESM1-CAM5	r1–3	r1–3	r1–3	r1–2	4.10[2]

Euro-Mediterranean Center on Climate Change, Italy	CMCC-CM		r1		r1	
Euro-Mediterranean Center on Climate Change, Italy	CMCC-CMS		r1		r1	
Meteo-France, France	CNRM-CM5	r1	r1		r1,2,4,6,10	3.25
CSIRO/QCCCE, Australia	CSIRO-Mk3-6-0	r1–10	r1–10	r1–10	r1–10	3.99
EC-Earth Consortium	EC-EARTH	r8,12	r1,2,6,8,9,12		r1,2,8,9,11,12,13	3.4[3]
LASG/Institute of Atmospheric Physics, China	FGOALS-g2	r1	r1		r1	3.45
The First Institute of Oceanography, SOA, China	FIO-ESM		r1–3	r1–3	r1–3	
Geophysical Fluid Dynamics Lab, USA	GFDL-CM3		r1	r1	r1	3.96
Geophysical Fluid Dynamics Lab, USA	GFDL-ESM2G		r1	r1	r1	2.38
Geophysical Fluid Dynamics Lab, USA	GFDL-ESM2M		r1		r1	2.41
NASA/Goddard Institute for Space Studies, USA	GISS-E2-H	r1p1–r1p3	r1p1–r5p3	r1p1–r1p3	r1p1–r1p3	2.30
NASA/Goddard Institute for Space Studies, USA	GISS-E2-H-CC		r1			
NASA/Goddard Institute for Space Studies, USA	GISS-E2-R	r1p1–r1p3	r1p1–r5p3	r1p2,r1p3	r1p1–r1p3	2.11
NASA/Goddard Institute for Space Studies, USA	GISS-E2-R-CC		r1			
National Institute of Meteorological Research, South Korea	HadGEM2-AO	r1	r1	r1	r1	
Met Office Hadley Centre, UK	HadGEM2-CC		r1		r1	
Met Office Hadley Centre, UK	HadGEM2-ES	r1–4	r1–4	R2–4	r1–4	4.55
Institute of Numerical Mathematics, Russia	INM-CM4		r1		r1	2.07
Institut Pierre Simon Laplace, France	IPSL-CM5A-LR	r1–4	r1–4	r1	r1–4	4.1
Institut Pierre Simon Laplace, France	IPSL-CM5A-MR	r1	r1	r1	r1	
Institut Pierre Simon Laplace, France	IPSL-CM5B-LR		r1		r1	2.59
Centre for Climate Research, Japan	MIROC5	r1–3	r1–3	r1–3	r1–3	2.71
Centre for Climate Research, Japan	MIROC-ESM	r1	r1	r1	r1	4.65
Centre for Climate Research, Japan	MIROC-ESM-CHEM	r1	r1	r1	r1	
Max Planck Institute for Meteorology DKRZ, Germany	MPI-ESM-LR	r1–3	r1–3		r1–3	3.60
Max Planck Institute for Meteorology DKRZ, Germany	MPI-ESM-MR	r1	r1–3		r1	3.44
Meteorological Research Institute, Japan	MRI-CGCM3	r1	r1	r1	r1	2.59
Norwegian Climate Center, Norway	NorESM1-M	r1	r1	r1	r1	2.83
Norwegian Climate Center, Norway	NorESM1-ME	r1	r1	r1	r1	

[1.] The estimate from the model developers (Meehl et al., 2011)
[2.] Estimate from the model developers (Meehl et al., 2013)
[3.] Estimate from the model developers (Lacagnina et al., 2014)

Discussion of results

The bivariate test is one of the most robust tests available for testing serially independent time series data for step, or abrupt, changes. However, climate data fulfils this condition only some of the time. The evidence presented in Ricketts (2015), supports previous conclusions that annual time series of observed temperature can be regarded as serially independent, especially where it shows little or limited sign of intervening trends that are statistically significant. Qualitatively, this is the step ladder-like behaviour where large step changes occur in a time series with limited internal trends. For the 20th century simulations to 2005 analysed here, these same conditions are considered to be met. A longer discussion on the reliability of the test under these conditions can be found in Ricketts (2015);Ricketts and Jones (2016).

Where there is the potential for steps and trends to be present in the same time series, then the bivariate test, and all other tests used in assessing step changes, become less robust. These conditions are present in most simulations after 2005. This is the principal reason for developing the rule-based test with multiple iterations to assess stable configurations.

Some testing was carried out with artificial time series containing red noise (autocorrelation 0.1 with a one-year lag, 0.25 with a seven-year lag) combined with random step changes and trends. By itself, red noise will produce step changes at a higher rate than serially independent data, thereby overstating the probability of exceedance. However, in using the test for detection, we are mainly interested in using the test to detect the timing and magnitude of steps as accurately as possible.

Our major assumption about a warming climate is that regime shifts (an organised and abrupt change in the structure and function of a system), red-noise driven shifts in the variable under analysis, random shifts and trending behaviour are all possible. In such a system, abrupt changes will become more common, therefore increase relative risk if those changes are driving impacts. This is the main purpose for the bivariate test in this paper, where it is being used to detect large shifts in mean temperature.

When all these phenomena are combined in artificial data, the combination of steps, red noise, random noise and trends will detect step changes that:

1. May not be serially independent, therefore overstating the probability of being a clear step change but not its timing or magnitude,
2. May produce a step change that averages two underlying step changes,
3. May variously suppress or amplify potential step changes, thus affecting the drivers of risk,
4. May detect a step change in a trending variable, where the internal steps by themselves may be insignificant.

The latter possibility, we consider as the only real false positive, but all the others warrant caution. Points one and two will reveal step changes, but not necessarily their case, point three suggests that not all underlying changes in a system may manifest and point four illustrates where the test will falsely identify steps and trends. The latter we identify in the paper by using shift-step and trend-step ratios, where the former will be small in a trending timeseries. This situation is associated with high radiative forcing.

Statistical testing environment

Most of the severe testing using statistical tools was carried out using MS Excel 2013 worksheets. Although Excel is widely frowned upon for statistical testing, it provides a highly flexible testing environment where templates can be constructed for rapid analysis of quantitative and graphic output. As methods are stabilised, they are brought into the Python modelling environment. This two-stage environment is useful because of the experimental nature of this work. All of the tests utilised within the Excel environment have been tested in other computing environments to ensure their reliability. The built in randomisation algorithms in Excel, are highly autocorrelated so are not used in this work, except in a diagnostic capacity.

Additional tools include the Loess utility (Peltier, 2009) and the multiple trend calculation and charting program (originally from D Kelly O'Day but no longer available online), which has been modified to conduct the moving window bivariate test and nonlinear regression, in addition to plotting up to 15 steps in a time series.

References

Aires, F., C. Prigent, W. Rossow and M. Rothstein, 2001: A new neural network approach including first-guess for retrieval of atmospheric water vapor, cloud liquid water path, surface temperature and emissivities over land from satellite microwave observations. *Journal of Geophysical Research: Atmospheres*, **106**, 14887-14907.

Alfaro, E.J., A. Gershunov and D. Cayan, 2006: Prediction of Summer Maximum and Minimum Temperature over the Central and Western United States: The Roles of Soil Moisture and Sea Surface Temperature. *Journal of Climate*, **19**, 1407-1421.

Bartsev, S.I., P.V. Belolipetskii, A.G. Degermendzhi, Y.D. Ivanova, A.A. Pochekutov and M.Y. Saltykov, 2016: Refocusing on the dynamics of the Earth's climate. *Herald of the Russian Academy of Sciences*, **86**, 135-142.

Beaulieu, C., J. Chen and J.L. Sarmiento, 2012: Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **370**, 1228-1249.

Belolipetsky, P., S. Bartsev, Y. Ivanova and M. Saltykov, 2015: Hidden staircase signal in recent climate dynamic. *Asia-Pacific Journal of Atmospheric Sciences*, **51**, 323-330.

Börjeson, L., M. Höjer, K.-H. Dreborg, T. Ekvall and G. Finnveden, 2006: Scenario types and techniques: towards a user's guide. *Futures*, **38**, 723-739.

Boucharel, J., B. Dewitte, B. Garel and Y. Du Penhoat, 2009: ENSO's non-stationary and non-Gaussian character: The role of climate shifts. *Nonlinear Processes in Geophysics*, **16**, 453-473.

Boucharel, J., B. Dewitte, Y. Penhoat, B. Garel, S.-W. Yeh and J.-S. Kug, 2011: ENSO nonlinearity in a warming climate. *Climate Dynamics*, **37**, 2045-2065.

Branstator, G. and F. Selten, 2009: "Modes of Variability" and Climate Change. *Journal of Climate*, **22**, 2639-2658.

Brohan, P., J.J. Kennedy, I. Harris, S.F.B. Tett and P.D. Jones, 2006: Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research: Atmospheres*, **111**, n/a-n/a.

Bücher, A. and J. Dessens, 1991a: Secular trend of surface temperature at an elevated observatory in the Pyrenees. *Journal of Climate*, **4**, 859-868.

Bücher, A. and J. Dessens, 1991b: Secular trend of surface temperature at an elevated observatory in the Pyrenees. *Journal of Climate*, **4**, 859-868.

Buishand, T., 1984: Tests for detecting a shift in the mean of hydrological time series. *Journal of Hydrology*, **73**, 51-69.

Bureau of Meteorology, 2003: *The greenhouse effect and climate change*. Bureau of Meteorology, Melbourne, 74 pp.

Cahill, N., S. Rahmstorf and A.C. Parnell, 2015: Change points of global temperature. *Environmental Research Letters*, **10**, 084002.

Capparelli, V., C. Franzke, A. Vecchio, M.P. Freeman, N.W. Watkins and V. Carbone, 2013: A spatiotemporal analysis of US station temperature trends over the last century. *Journal of Geophysical Research: Atmospheres*, **118**, 7427-7434.

Chikamoto, Y., M. Kimoto, M. Ishii, M. Watanabe, T. Nozawa, T. Mochizuki, H. Tatebe, T.T. Sakamoto, Y. Komuro and H. Shiogama, 2012a: Predictability of a stepwise shift in Pacific climate during the late 1990s in hindcast experiments using MIROC. *Journal of the Meteorological Society of Japan*, **90**, 1-21.

Chikamoto, Y., M. Kimoto, M. Watanabe, M. Ishii and T. Mochizuki, 2012b: Relationship between the Pacific and Atlantic stepwise climate change during the 1990s. *Geophysical Research Letters*, **39**.

Christy, J.R., R.W. Spencer and W.D. Braswell, 2000: MSU tropospheric temperatures: Dataset construction and radiosonde comparisons. *Journal of Atmospheric and Oceanic Technology*, **17**, 1153-1170.
Christy, J.R., R.W. Spencer, W.B. Norris, W.D. Braswell and D.E. Parker, 2003: Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures. *Journal of Atmospheric and Oceanic Technology*, **20**, 613-629.
Christy, J.R., W.B. Norris, R.W. Spencer and J.J. Hnilo, 2007: Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements. *Journal of Geophysical Research: Atmospheres*, **112**.
Cohn, T.A. and H.F. Lins, 2005: Nature's style: Naturally trendy. *Geophysical Research Letters*, **32**, L23402.
Corti, S., F. Molteni and T.N. Palmer, 1999: Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature*, **398**, 799-802.
Cowtan, K. and R.G. Way, 2014a: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, **140**, 1935-1944.
Cowtan, K. and R.G. Way, 2014b: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, n/a-n/a.
Cox, D. and D.G. Mayo, 2010: Objectivity and Conditionality in Frequentist Inference. In: *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* [Mayo, D.G. and A. Spanos (eds.)] Cambridge University Press, Cambridge UK and New York USA, 276-304.
CSIRO and BoM, 2007: *Climate Change in Australia: technical report 2007*. CSIRO, Melbournepp.
Curd, M. and J.A. Cover, 1998: *Philosophy of Science: The Central Issues*. WW Norton New York and London, 1379 pp.
Ebbesmeyer, C.C., D.R. Cayan, D.R. McLain, F.H. Nichols, D.H. Peterson and K.T. Redmond, 1991: 1976 step in the Pacific climate: forty environmental changes between 1968-1975 and 1977-1984. *Proceedings of the Seventh annual Pacific Climate (PACLIM) Workshop*, *Asilomar, California, April, 1990*, Betancourt, J.L. and V.L. Tharp, Eds., Vol. Technical Report 26, California Department of Water Resources. Interagency Ecology Studies Program, City, pp 115-126.
Fischer, T., M. Gemmer, L. Liu and B. Su, 2012: Change-points in climate extremes in the Zhujiang River Basin, South China, 1961–2007. *Climatic Change*, **110**, 783-799.
Foster, G. and S. Rahmstorf, 2011: Global temperature evolution 1979-2010. *Environmental Research Letters*, **6**, 044022.
Foster, G. and J. Abraham, 2015: Lack of evidence for a slowdown in global temperature. *US CLIVAR*, **13**, 6-9.
Franzke, C., 2012: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of Climate*, **25**, 4172-4183.
Free, M. and J. Lanzante, 2009: Effect of volcanic eruptions on the vertical temperature profile in radiosonde data and climate models. *Journal of Climate*, **22**, 2925-2939.
Gan, T.Y., 1995: Trends in air temperature and precipitation for Canada and north-eastern USA. *International Journal of Climatology*, **15**, 1115-1134.
Gao, J. and K. Hawthorne, 2006: Semiparametric estimation and testing of the trend of temperature series. *Econometrics Journal*, **9**, 332-355.
Ghil, M., 2012: Climate variability: nonlinear and random effects. *Encyclopedia of Atmospheric Sciences. Elsevier*, 1-6.
GISSTemp Team, 2015: *GISS Surface Temperature Analysis (GISTEMP)*. NASA Goddard Institute for Space Studies. pp.
Hansen, J., I. Fung, A. Lacis, D. Rind, S. Lebedeff, R. Ruedy, G. Russell and P. Stone, 1988: Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model. *Journal of Geophysical Research: Atmospheres (1984–2012)*, **93**, 9341-9364.

Hansen, J., R. Ruedy, M. Sato and K. Lo, 2010: Global surface temperature change. *Reviews of Geophysics*, **48**.
Hare, S.R. and N.J. Mantua, 2000: Empirical evidence for North Pacific regime shifts in 1977 and 1989. *Progress In Oceanography*, **47**, 103-145.
Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies. In: *Meteorology of Tropical Oceans* [Shaw, D.B. (ed.) Royal Meteorological Society, London, UK, 251-259.
Hasselmann, K., 1993: Optimal fingerprints for the detection of time-dependent climate change. *Journal of Climate*, **6**, 1957-1971.
Hasselmann, K., 2002: Is Climate Predictable? In: *The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes* [Bunde, A., J. Kropp and H.J. Schellnhuber (eds.)] Springer, Berlin Heidelberg, 141-188.
Hegerl, G. and F. Zwiers, 2011: Use of models in detection and attribution of climate change. *Wiley Interdisciplinary Reviews: Climate Change*, **2**, 570-591.
Hegerl, G.C., F.W. Zwiers, P. Braconnot, N.P. Gillett, Y. Luo, J.A.M. Orsini, N. Nicholls, J.E. Penner and P.A. Stott, 2007: Understanding and Attributing Climate Change. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)] Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 663-745.
Hope, P., B. Timbal and R. Fawcett, 2010: Associations between rainfall variability in the southwest and southeast of Australia and their evolution through time. *International Journal of Climatology*, **30**, 1360-1372.
Hoskins, B., 2013: The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Quarterly Journal of the Royal Meteorological Society*, **139**, 573-584.
Hulme, M. and L. Mearns, 2001: Climate Scenario Development. In: *Climate Change 2001: The Scientific Basis 2001* [Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. Van Der Linden and D. Xiaosu (eds.)] Cambridge University Press, Cambridge and New York, 739-768.
Jandhyala, V., S. Fotopoulos, I. MacNeill and P. Liu, 2013: Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, n/a-n/a.
Ji, F., Z. Wu, J. Huang and E.P. Chassignet, 2014: Evolution of land surface air temperature trend. *Nature Clim. Change*, **4**, 462-466.
Jones, P.D., M. New, D.E. Parker, S. Martin and I.G. Rigor, 1999: Surface air temperature and its changes over the past 150 years. *Reviews of Geophysics*, **37**, 173-199.
Jones, P.D., T.J. Osborn, K.R. Briffa, C.K. Folland, E.B. Horton, L.V. Alexander, D.E. Parker and N.A. Rayner, 2001: Adjusting for sampling density in grid box land and ocean surface temperature time series. *Journal of Geophysical Research: Atmospheres*, **106**, 3371-3380.
Jones, P.D., D.H. Lister, T.J. Osborn, C. Harpham, M. Salmon and C.P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres*, **117**, D05127.
Jones, R.N., 2012: Detecting and attributing nonlinear anthropogenic regional warming in southeastern Australia. *Journal of Geophysical Research*, **117**, D04105.
Jones, R.N., C.K. Young, J. Handmer, A. Keating, G.D. Mekala and P. Sheehan, 2013: *Valuing Adaptation under Rapid Change*. National Climate Change Adaptation Research Facility, Gold Coast, Australia, 182 pp.
Jones, R.N., 2015a: *Reconciling anthropogenic climate change and variability on decadal timescales: hypotheses and scientific narratives*. Climate Change Working Paper No. 32, Victoria Institute of Strategic Economic Studies, Victoria University, Melbourne, 16 pp.

Jones, R.N., 2015b: *Reconciling anthropogenic climate change and variability on decadal timescales: history and philosophy*. Climate Change Working Paper No. 33, Victoria Institute of Strategic Economic Studies, Victoria University, Melbourne, 18 pp.
Jones, R.N. and J.H. Ricketts, 2016: *The climate wars and "the pause" – are both sides wrong?* Climate Change Working Paper No. 37, Victoria Institute of Strategic Economic Studies, Victoria University, Melbourne, 23 pp.
Karoly, D.J. and K. Braganza, 2005: A new approach to detection of anthropogenic temperature changes in the Australian region. *Meteorology and Atmospheric Physics*, **89**, 57-67.
Kennedy, J.J., N.A. Rayner, R.O. Smith, D.E. Parker and M. Saunby, 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research: Atmospheres*, **116**, n/a-n/a.
Killick, R., I.A. Eckley, K. Ewans and P. Jonathan, 2010: Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, **37**, 1120-1126.
Killick, R., P. Fearnhead and I. Eckley, 2012: Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107**, 1590-1598.
Kirono, D. and R. Jones, 2007: A bivariate test for detecting inhomogeneities in pan evaporation time series. *Australian Meteorological Magazine*, **56**, 93-103.
Kirtman, B., S. Power, A.J. Adedoyin, G. Boer, R. Bojariu, I. Camilloni, F. Doblas-Reyes, A. Fiore, M. Kimoto, G. Meehl, M. Prather, A. Sarr, C. Schär, R. Sutton, G.J.v. Oldenborgh, G. Vecchi and H.-J. Wang, 2013: Near-term Climate Change: Projections and Predictability. In: *Climate Change 2013: The Physical Science Basis. Working Group I contribution to the IPCC 5th Assessment Report* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)] Cambridge University Press, Cambridge and New York, 121.
Koutsoyiannis, D., 2010: HESS Opinions "A random walk on water". *Hydrology and Earth System Sciences*, **14**, 585-601.
Lacagnina, C., F. Selten and A.P. Siebesma, 2014: Impact of changes in the formulation of cloud-related processes on model biases and climate feedbacks. *Journal of Advances in Modeling Earth Systems*, **6**, 1224-1243.
Levitus, S., J.I. Antonov, T.P. Boyer, O.K. Baranova, H.E. Garcia, R.A. Locarnini, A.V. Mishonov, J.R. Reagan, D. Seidov, E.S. Yarosh and M.M. Zweng, 2012: World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010. *Geophysical Research Letters*, **39**, L10603.
Lewandowsky, S., N. Oreskes, J.S. Risbey, B.R. Newell and M. Smithson, 2015: Seepage: Climate change denial and its effect on the scientific community. *Global Environmental Change*, **33**, 1-13.
Lewandowsky, S., J.S. Risbey and N. Oreskes, 2016: The "Pause" in Global Warming: Turning a Routine Fluctuation into a Problem for Science. *Bulletin of the American Meteorological Society*, **97**, 723–733.
Li, F., L. Chambers and N. Nicholls, 2005: Relationships between rainfall in the southwest of Western Australia
and near global patterns of sea-surface temperature and mean sea-level pressure variability. *Australian Meteorological Magazine*, **54**, 23-33.
Lo, T.T. and H.H. Hsu, 2010: Change in the dominant decadal patterns and the late 1980s abrupt warming in the extratropical Northern Hemisphere. *Atmospheric Science Letters*, **11**, 210-215.
Lorenz, E.N., 1975: Climate Predictability. In: *The Physical Bases of Climate and Climate Modelling.*, Vol. GARP Publication Series, Vol. 16 World Meteorological Organisation, Geneva, 132-136.
Lucarini, V., K. Fraedrich and F. Lunkeit, 2010: Thermodynamics of climate change: generalized sensitivities. *Atmos. Chem. Phys.*, **10**, 9729-9737.
Lucarini, V. and F. Ragone, 2011: Energetics of climate models: net energy balance and meridional enthalpy transport. *Reviews of Geophysics*, **49**, RG1001.
Mantua, N.J., S.R. Hare, Y. Zhang, J.M. Wallace and R.C. Francis, 1997: A Pacific interdecadal climate oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, **78**, 1069-1079.

Maronna, R. and V.J. Yohai, 1978: A bivariate test for the detection of a systematic change in mean. *Journal of the American Statistical Association*, **73**, 640-645.
Marvel, K., G.A. Schmidt, D. Shindell, C. Bonfils, A.N. LeGrande, L. Nazarenko and K. Tsigaridis, 2015: Do responses to different anthropogenic forcings add linearly in climate models? *Environmental Research Letters*, **10**, 104010.
Mayo, D.G., 1996: *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago, 509 pp.
Mayo, D.G., 2005: Evidence as passing severe tests: highly probable versus highly probed hypotheses. In: *Scientific Evidence: Philosophical Theories and Applications* [Achinstein, P. (ed.) John Hopkins University Press, Baltimore and London, 95-127.
Mayo, D.G., 2010: Towards critical progressive rationalism: Exchanges with Alan Musgrave. In: *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* [Mayo, D.G. and A. Spanos (eds.)] Cambridge University Press, Cambridge UK and New York USA, 113-124.
Mayo, D.G. and D. Cox, 2010: Frequentist Statistics as a Theory of Inductive Inference. In: *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* [Mayo, D.G. and A. Spanos (eds.)] Cambridge University Press, Cambridge UK and New York USA, 247-275.
Mayo, D.G. and A. Spanos, 2010: *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of science*. Cambridge University Press, Cambridge UK, 419 pp.
Mayo, D.G. and A. Spanos, 2011: Error Statistics. In: *Philosophy of Statistics* [Bandyopadhyay, P.S. and M.R. Forster (eds.)], Vol. 7 North-Holland, Amsterdam, 153-198.
McKitrick, R.R., 2014: HAC-Robust Measurement of the Duration of a Trendless Subsample in a Global Climate Time Series. *Open Journal of Statistics*, **4**, 527-535.
McKitrick, R.R., 2015: Energy Policy and Environmental Stewardship: Risk Management not Risk Avoidance. *Greer-Heard Point-Counterpoint Forum New Orleans Baptist Theological Seminary, April 10-11 2015*, City, pp 18.
Mears, C.A., M.C. Schabel and F.J. Wentz, 2003: A Reanalysis of the MSU Channel 2 Tropospheric Temperature Record. *Journal of Climate*, **16**, 3650-3664.
Mears, C.A. and F.J. Wentz, 2009a: Construction of the RSS V3. 2 lower-tropospheric temperature dataset from the MSU and AMSU microwave sounders. *Journal of Atmospheric and Oceanic Technology*, **26**, 1493-1509.
Mears, C.A. and F.J. Wentz, 2009b: Construction of the Remote Sensing Systems V3.2 Atmospheric Temperature Records from the MSU and AMSU Microwave Sounders. *Journal of Atmospheric and Oceanic Technology*, **26**, 1040-1056.
Mears, C.A. and F.J. Wentz, 2009c: Construction of the RSS V3.2 Lower-Tropospheric Temperature Dataset from the MSU and AMSU Microwave Sounders. *Journal of Atmospheric and Oceanic Technology*, **26**, 1493-1509.
Meehl, G.A., A. Hu and B.D. Santer, 2009: The Mid-1970s Climate Shift in the Pacific and the Relative Roles of Forced versus Inherent Decadal Variability. *Journal of Climate*, **22**, 780-792.
Meehl, G.A., W.M. Washington, J.M. Arblaster, A. Hu, H. Teng, C. Tebaldi, B.N. Sanderson, J.-F. Lamarque, A. Conley, W.G. Strand and J.B. White, 2011: Climate System Response to External Forcings and Climate Change Projections in CCSM4. *Journal of Climate*, **25**, 3661-3683.
Meehl, G.A., W.M. Washington, J.M. Arblaster, A. Hu, H. Teng, J.E. Kay, A. Gettelman, D.M. Lawrence, B.M. Sanderson and W.G. Strand, 2013: Climate Change Projections in CESM1(CAM5) Compared to CCSM4. *Journal of Climate*, **26**, 6287-6308.
Meehl, G.A., 2015: Decadal climate variability and the early-2000s hiatus. *US CLIVAR*, **13**, 1-6.
Menberg, K., P. Blum, B.L. Kurylyk and P. Bayer, 2014: Observed groundwater temperature response to recent climate change. *Hydrology and Earth System Sciences*, **18**, 4453-4466.

Miller, A.J., D.R. Cayan, T.P. Barnett, N.E. Graham and J.M. Oberhuber, 1994: The 1976–77 climate shift of the Pacific Ocean. *Oceanography*, **7**, 21-26.
Mitchell, T.D., 2003a: Pattern scaling - An examination of the accuracy of the technique for describing future climates. *Climatic Change*, **60**, 217-242.
Mitchell, T.D., 2003b: Pattern scaling: an examination of the accuracy of the technique for describing future climates. *Climatic Change*, **60**, 217-242.
Morice, C.P., J.J. Kennedy, N.A. Rayner and P.D. Jones, 2012a: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, **117**, D08101.
Morice, C.P., J.J. Kennedy, N.A. Rayner and P.D. Jones, 2012b: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, **117**, n/a-n/a.
Musgrave, A., 2010: Critical Rationalism, Explanation and Severe Tests. In: *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* [Mayo, D.G. and A. Spanos (eds.)] Cambridge University Press, Cambridge UK, 88-112.
Nicholls, N., P. Dellamarta and D. Collins, 2004: 20th century changes in temperature and rainfall in New South Wales. *Australian Meteorological Magazine*, **53**, 263-268.
North, G.R., K.-Y. Kim, S.S.P. Shen and J.W. Hardin, 1995: Detection of Forced Climate Signals. Part 1: Filter Theory. *Journal of Climate*, **8**, 401-408.
North, R.P., D.M. Livingstone, R.E. Hari, O. Köster, P. Niederhauser and R. Kipfer, 2013: The physical impact of the late 1980s climate regime shift on Swiss rivers and lakes. *Inland Waters*, **3**, 341-350.
Osborn, T.J. and P.D. Jones, 2014: The CRUTEM4 land-surface air temperature data set: construction, previous versions and dissemination via Google Earth. *Earth Syst. Sci. Data*, **6**, 61-68.
Overland, J., S. Rodionov, S. Minobe and N. Bond, 2008: North Pacific regime shifts: Definitions, issues and recent transitions. *Progress In Oceanography*, **77**, 92-102.
Ozawa, H., A. Ohmura, R.D. Lorenz and T. Pujol, 2003: The second law of thermodynamics and the global climate system: A review of the maximum entropy production principle. *Reviews of Geophysics*, **41**, 1018.
Palmer, T.N., 1993: A nonlinear dynamical perspective on climate change. *Weather*, **48**, 314-326.
Palmer, T.N., F.J. Doblas-Reyes, A. Weisheimer and M.J. Rodwell, 2008: Toward Seamless Prediction: Calibration of Climate Change Projections Using Seasonal Forecasts. *Bulletin of the American Meteorological Society*, **89**, 459-470.
Peltier, J., 2009: *Loess Utility*. Peltier Technical Services, Inc., http://peltiertech.com/pp.
Peterson, T.C. and R.S. Vose, 1997: An overview of the Global Historical Climatology Network temperature database. *Bulletin of the American Meteorological Society*, **78**, 2837-2849.
Potter, K., 1981a: Illustration of a new test for detecting a shift in mean in precipitation series. *Monthly Weather Review*, **109**, 2040-2045.
Potter, K.W., 1981b: Illustration of a New Test for Detecting a Shift in Mean in Precipitation Series. *Mon. Wea. Rev.*, **109**, 2040-2045.
Power, S., F. Tseitkin, S. Torok, B. Lavery and B. McAvaney, 1998: Australian temperature, Australian rainfall, and the Southern Oscillation, 1910-1996: coherent variability and recent changes. *Australian Meteorological Magazine*, **47**, 85-101.
Rajaratnam, B., J. Romano, M. Tsiang and N. Diffenbaugh, 2015: Debunking the climate hiatus. *Climatic Change*, **133**, 129–140.
Rayner, N.A., P. Brohan, D.E. Parker, C.K. Folland, J.J. Kennedy, M. Vanicek, T.J. Ansell and S.F.B. Tett, 2006: Improved Analyses of Changes and Uncertainties in Sea Surface Temperature Measured In Situ since the Mid-Nineteenth Century: The HadSST2 Dataset. *Journal of Climate*, **19**, 446-469.
Reeves, J., J. Chen, X.L. Wang, R. Lund and Q.Q. Lu, 2007: A Review and Comparison of Changepoint Detection Techniques for Climate Data. *Journal of Applied Meteorology and Climatology*, **46**, 900-915.

Reid, P.C. and G. Beaugrand, 2012: Global synchrony of an accelerating rise in sea surface temperature. *Journal of the Marine Biological Association of the United Kingdom*, **92**, 1435-1450.
Reid, P.C., R.E. Hari, G. Beaugrand, D.M. Livingstone, C. Marty, D. Straile, J. Barichivich, E. Goberville, R. Adrian and Y. Aono, 2015: Global impacts of the 1980s regime shift. *Global Change Biology*.
Ricketts, J.H., 2015: A probabilistic approach to climate regime shift detection based on Maronna's bivariate test. *The 21st International Congress on Modelling and Simulation (MODSIM2015)*, *Gold Coast, Queensland, Australia*, Weber, T., M.J. McPhee and R.S. Anderssen, Eds., The Modelling and Simulation Society of Australia and New Zealand, City, pp 1310-1316.
Ricketts, J.H. and R.N. Jones, 2016: The multi-step Maronna-Yohai bivariate test for detecting multiple step changes in climate data. *Manuscript in preparation*.
Rodionov, S.N., 2005: A brief overview of the regime shift detection methods. *Large-Scale Disturbances (Regime Shifts) and Recovery in Aquatic Ecosystems: Challenges for Management Toward Sustainability. UNESCO-ROSTE/BAS Workshop on Regime Shifts*, *Varna, Bulgaria*, Velikova, V. and N. Chipev, Eds., City, pp 17-24.
Rodionov, S.N., 2006: Use of prewhitening in climate regime shift detection. *Geophysical Research Letters*, **33**, L12707.
Roemmich, D., J. Church, J. Gilson, D. Monselesan, P. Sutton and S. Wijffels, 2015: Unabated planetary warming and its ocean structure since 2006. *Nature Climate Change*, **5**, 240-245.
Rohde, R., R.A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom and C. Wickham, 2012: A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, **1**, 1000101.
Rohde, R., R. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom and C. Wickham, 2013a: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geoinfor Geostat Overview 1: 1. *of*, **7**, 2.
Rohde, R., R. Muller, R. Jacobsen, S. Perlmutter, A. Rosenfeld, J. Wurtele, J. Curry, C. Wickham and S. Mosher, 2013b: Berkeley earth temperature averaging process. *Geoinfor. Geostat.: An Overview*, **1**, 1-13.
Ruggieri, E., 2013: A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, **33**, 520-528.
Sahin, S. and H.K. Cigizoglu, 2010: Homogeneity analysis of Turkish meteorological data set. *Hydrological Processes*, **24**, 981-992.
Santer, B.D., T.M.L. Wigley, M.E. Schlesinger and J.F.B. Mitchell, 1990: *Developing climate scenarios from equilibrium GCM results*. Research Report No. 47, Max Planck Institut für Meteorologie, Hamburg, 79 pp.
Santer, B.D., C. Mears, C. Doutriaux, P. Caldwell, P.J. Gleckler, T.M.L. Wigley, S. Solomon, N.P. Gillett, D. Ivanova, T.R. Karl, J.R. Lanzante, G.A. Meehl, P.A. Stott, K.E. Taylor, P.W. Thorne, M.F. Wehner and F.J. Wentz, 2011: Separating signal and noise in atmospheric temperature changes: The importance of timescale. *Journal of Geophysical Research*, **116**, D22105.
Schmidt, G.A., M. Kelley, L. Nazarenko, R. Ruedy, G.L. Russell, I. Aleinov, M. Bauer, S.E. Bauer, M.K. Bhat and R. Bleck, 2014a: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, **6**, 141-184.
Schmidt, G.A., D.T. Shindell and K. Tsigaridis, 2014b: Reconciling warming trends. *Nature Geoscience*, **7**, 158-160.
Scott, A.J. and M. Knott, 1974: A Cluster Analysis Method for Grouping Means in the Analysis of Variance. *Biometrics*, **30**, 507-512.
Seidel, D.J. and J.R. Lanzante, 2004: An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *Journal of Geophysical Research*, **109**, D14108.
Şerban, D.Z. and B.H. Jacobsen, 2001: The use of broad-band prior covariance for inverse palaeoclimate estimation. *Geophysical Journal International*, **147**, 29-40.

Skeptical Science, 2015: *The Escalator*. John Cook, Brisbane, http://www.skepticalscience.com/graphics.php?g=47, (Last accessed February 23, 2016).
Slingo, J., 2013: *Statistical models and the global temperature record*, UK Met Office, Exeter, Devon, 15 pp.
Smith, T.M., R.W. Reynolds, T.C. Peterson and J. Lawrimore, 2008: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *Journal of Climate*, **21**, 2283-2296.
Solomon, A., L. Goddard, A. Kumar, J. Carton, C. Deser, I. Fukumori, A.M. Greene, G. Hegerl, B. Kirtman, Y. Kushnir, M. Newman, D. Smith, D. Vimont, T. Delworth, G.A. Meehl and T. Stockdale, 2011: Distinguishing the Roles of Natural and Anthropogenically Forced Decadal Climate Variability. *Bulletin of the American Meteorological Society*, **92**, 141-156.
Spanos, A., 2010: Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. *Journal of Econometrics*, **158**, 204-220.
Stern, D.I. and R.K. Kaufmann, 2000: Detecting a global warming signal in hemispheric temperature series: Astructural time series analysis. *Climatic Change*, **47**, 411-438.
Stott, P.A., N.P. Gillett, G.C. Hegerl, D.J. Karoly, D.A. Stone, X. Zhang and F. Zwiers, 2010: Detection and attribution of climate change: a regional perspective. *Wiley Interdisciplinary Reviews: Climate Change*, **1**, 192-211.
Swanson, K.L., G. Sugihara and A.A. Tsonis, 2009: Long-term natural variability and 20th century climate change. *Proceedings of the National Academy of Sciences*, **106**, 16120-16123.
Swanson, K.L. and A.A. Tsonis, 2009: Has the climate recently shifted? *Geophysical Research Letters*, **36**, L06711.
Timbal, B., J. Arblaster, K. Braganza, E. Fernandez, H. Hendon, B. Murphy, M. Raupach, C. Rakich, I. Smith, K. Whan and M. Wheeler, 2010: *Understanding the anthropogenic nature of the observed rainfall decline across south-eastern Australia*. CAWCR Technical Report, The Centre for Australian Weather and Climate Research, Melbourne, 180 pp.
Tisdale, B., 2015: *On Global Warming and the Illusion of Control – Part 1. A Comprehensive Illustrated Introduction to the Hypothesis of Human-Induced Global Warming*. Bob Tisdale, bobtisdale.wordpress.com, 733 pp.
Trenberth, K.E., 2015: Has there been a hiatus? *Science*, **349**, 691-692.
Tsonis, A. and K. Swanson, 2012: Review article" On the origins of decadal climate variability: a network perspective". *Nonlinear Processes in Geophysics*, **19**, 559-568.
Tsonis, A.A., K. Swanson and S. Kravtsov, 2007: A new dynamical mechanism for major climate shifts. *Geophysical Research Letters*, **34**, GL030288.
Tsonis, A.A. and K.L. Swanson, 2011: Climate mode covariability and climate shifts. *International Journal of Bifurcation and Chaos*, **21**, 3549-3556.
Varotsos, C.A., C.L. Franzke, M.N. Efstathiou and A.G. Degermendzhi, 2014: Evidence for two abrupt warming events of SST in the last century. *Theoretical and Applied Climatology*, **116**, 51-60.
Vivès, B. and R.N. Jones, 2005: *Detection of Abrupt Changes in Australian Decadal Rainfall (1890-1989)*. CSIRO Atmospheric Research Technical Paper, CSIRO Atmospheric Research, Melbourne, 54 pp.
Vose, R.S., D. Arndt, V.F. Banzon, D.R. Easterling, B. Gleason, B. Huang, E. Kearns, J.H. Lawrimore, M.J. Menne, T.C. Peterson, R.W. Reynolds, T.M. Smith, C.N. Williams and D.B. Wuertz, 2012: NOAA's Merged Land–Ocean Surface Temperature Analysis. *Bulletin of the American Meteorological Society*, **93**, 1677-1685.
Wack, P., 1985a: The Gentle Art of Reperceiving - Scenarios: Shooting the Rapids (part 2 of a two-part article). *Harvard Business Review*, 2-14.
Wack, P., 1985b: The Gentle Art of Reperceiving - Scenarios: Uncharted Waters Ahead (part 1 of a two-part article). *Harvard Business Review*, 73-89.
Wang, G., K.L. Swanson and A.A. Tsonis, 2009: The pacemaker of major climate shifts. *Geophysical Research Letters*, **36**, L07708.

Wang, G., P. Yang, X. Zhou, K.L. Swanson and A.A. Tsonis, 2012: Directional influences on global temperature prediction. *Geophysical Research Letters*, **39**, L13704.
Whetton, P.H., R.N. Jones, K.L. McInnes, K.J. Hennessy, R. Suppiah, C.M. Page, J. Bathols, P.J. Durack and CSIRO Marine and Atmospheric Research., 2005: *Australian Climate Change Projections for Impact Assessment and Policy Application: A review*. CSIRO Marine and Atmospheric Research, Aspendale, Vic., 34 pp.
White, H., 1980: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, **48**, 817-838.
Wolter, K. and M.S. Timlin, 2011: El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, **31**, 1074-1087.
World Meteorological Organization, 2010: *Position Paper on Global Framework for Climate Services*. World Meteorological Organisation, Geneva, 52 pp.
World Meteorological Organization, 2011: *Climate Knowledge for Action: A Global Framework for Climate Services - Empowering the Most Vulnerable*. Report No. 1065 World Meteorological Organization, Geneva, 247 pp.
Zhou, J. and K.-K. Tung, 2013: Deducing Multidecadal Anthropogenic Global Warming Trends Using Multiple Regression Analysis. *Journal of the Atmospheric Sciences*, **70**, 3-8.