S1. Statistical method

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

This formulation is from Bücher and Dessens (1991) 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 (1981) 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 (1991). 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.

$$\overline{X'} = \frac{\sum_{j=1}^{n} x'_j}{n},\ \overline{Y'} = \frac{\sum_{j=1}^{n} y'_j}{n},\ S'_x = \left(\frac{\sum_{j=1}^{n}\left(x'_j - \overline{X\prime}\right)^2}{n}\right)^{1/2},\ S'_y = \left(\frac{\sum_{j=1}^{n}\left(y'_j - \overline{Y\prime}\right)^2}{n}\right)^{1/2} \tag{S1}$$

$$x_j = \frac{\left(x'_j - \overline{X\prime}\right)}{S'_x},\ y_j = \frac{\left(y'_j - \overline{Y\prime}\right)}{S'_y} \text{ for all } j \le n. \tag{S2}$$

Step 2. Compute test statistics.

$$S_{xy} = \sum_{j=1}^{n} x_j y_j \tag{S3}$$

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

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

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

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

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

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 (1981) provides these for 100. An interpolating function is used to generalize these results for time series of varying length.

S1.1 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.

S1.2 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.

S1.3 Considerations about the multi-step bivariate test

S1.3.1 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.

S1.3.2 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.

S1.3.3 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.

S1.3.4 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.

S1.3.5 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.

S1.4 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.

S2. Data sources

S2.1 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, 2014)
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., 2012;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, 2009b, a)
UAH	6.0.beta	5 May 2015	1981-2010	Y	Y	Y	Y	(Christy et al., 2000)

S2.1.1 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.

S2.1.2 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.

S2.1.3 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.

S2.1.4 Berkeley

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.

S2.1.5 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.

S2.1.6 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.

S2.2 Satellite derived lower tropospheric temperature data, RSS and UAH

S2.2.1 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)

S2.2.2 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.

2.2.2.1 Data formats

Data looks like this, from December 1978 onwards.

```
Year Mo Globe  Land Ocean   NH   Land Ocean   SH   Land Ocean Trpcs  …  AUST
1978 12  0.86  0.21  1.11  0.26  …  2.57
```

2.2.2.2 Anomaly base period

As per metadata and confirmed by NCDC, 1981–2010.

S2.3 Model Data

S2.3.1

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

S2.3.2 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 SRESA2	1 1	1861 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 SRESA2	1–4, E1–4 1	1880 1880
Instituto Nazionale di Geofisica e Vulcanologia, Italy	INGV	G, SD	SRESA1b SRESA2	1 1	1870 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 SRESA2	1 1	1871 1871
Institut Pierre Simon Laplace, France	IPSL	G, SD, SI	SRESA1b SRESA2	1 1	1860 1860
Centre for Climate Research, Japan	MIROC-H	G, O, SD, BC, OC, MD, SS, LU, SO, V	SRESA1b SRESA2	1–3, E1–3 1–3, E1–3	1850 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 SRESA2	1–3, E1–3 1–3, E1–3	1860 1860
Max Planck Institute for Meteorology DKRZ, Germany	MPI-ECHAM5	G, O, SD, SI	SRESA1b SRESA2	1–4, E1–3 1–3, E1–3	1860 1860
Meteorological Research Institute, Japan	MRI	G, SD ,SO	SRESA1b SRESA2	1–5, E1–5 1–5, E1–5	1851 1851
National Center for Atmospheric Research, USA	NCAR-CCSM	G, O, SD, BC, OC, SO, U	SRESA1b SRESA2	1–4,9, E1–2 1–4	1870 1870
National Center for Atmospheric Research, USA	NCAR-PCM1	G, O, SD, SO, V	SRESA1b SRESA2	1,4 1–4, E2–4	1890 1890
Hadley Centre, UK	HADCM3	G, O, SD, SI	SRESA1b SRESA2	1 1	1860 1860
Hadley Centre, UK	HADGEM1	G, O, SD, SI, BC, OC, LU, SO, V	SRESA1b SRESA2	1 1	1860 1860

S2.3.3 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)

S3. 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.

S4. Nonlinear attribution methods and other data

Data sources for South-east Australia are detailed in Jones (2012) along with a detailed methodology. They were downloaded from the Australian Bureau of Meteorology climate facility (http://www.bom.gov.au/climate/change/index.shtml). These data have not been updated beyond 2010 because of a change in the method of calculating high quality temperature data (Trewin, 2012), which creates small, but detectable biases in the data (unpublished analyses). Adjustments are linearised, which smooths out nonlinear behaviour.

Central England temperatures come from the HadCET data and Central England rainfall downloaded in May 2015 (Parker et al., 1992;Alexander and Jones, 2000).

Data for Texas was sourced from the USHCN-V2 dataset downloaded in October 2011 (Menne et al., 2009). Station records were subject to basic quality control, dispensing with uncorrelated inhomogeneities and a simple average produced.

Rainfall data for Figure 6a and b were downloaded from the Australian Bureau of Meteorology climate facility in December 2016. The data have been quality controlled and homogenised (Lavery et al., 1997).

Ocean heat content data were downloaded from the NASA National Oceanographic Data Center web site in March 2016 (https://www.nodc.noaa.gov/OC5/3M_HEAT_CONTENT/) (Levitus et al., 2012).

Sea level data were downloaded from the Permanent Service for Mean Sea Level site (http://www.psmsl.org/) in December 2016. San Francisco data (Smith, 1980) and Fremantle data (White et al., 2014).

S5. Statistical testing environment

Much of the probative testing using statistical tools was carried out using MS Excel 2013/2016 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 checked in other computing environments to ensure their reliability. The randomisation techniques in Excel, which are highly autocorrelated, 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.

Table S4: Symbols and acronyms

Symbol	Measure	Unit
CMIP3	Third Climate Model Intercomparison Project	
CMIP5	Fifth Climate Model Intercomparison Project	
δF	Change in atmospheric forcing	W m^{-2}
ECS	Equilibrium climate sensitivity	°C
h	Statistical hypothesis	
H	Scientific hypothesis	
λ	Climate feedback factor	
MME	Multi-model ensemble	
MYBT	Maronna-Yohai bivariate test	
RCP	Representative Concentration Pathway	
ResSS	Residual Sum of Squares	
SEA	South-east Australia	
SRES	Special Report on Emission Scenarios	
Tav	Average annual air temperature	°C
Tmax	Maximum annual air temperature	°C
Tmin	Minimum annual air temperature	°C
P	Precipitation	mm
DTR	Diurnal temperature range	°C
Tmax/P	*Tmax* tested using *P* as a reference	°C
Tmin/Tmax	*Tmin* tested using *Tmax* as a reference	°C
Tav_{ARW}	Anthropogenic component of regional *Tav*	°C
T	Test	
T_i	Test statistic for the bivariate test	
T_{i0}	Maximum test statistic for the bivariate test	
δT	Change in temperature	°C

S6. Archived data and programs

Accompanying this document are the following data sets:

- Obs_output_step_data-Jones&Ricketts.xlsx
- Models_output_step_data-Jones&Ricketts.xlsx

These include much of the output detailed above and data used in figures.

Also incorporated are the programs:

- The multi-step bivariate test program suite
- Interactive_Regression-GMT-models many shifts-archive.xlsm

References

Alexander, L. V., and Jones, P. D.: Updated precipitation series for the U.K. and discussion of recent extremes, Atmospheric Science Letters, 1, 142-150, 10.1006/asle.2000.0016, 2000.

Brohan, P., Kennedy, J. J., Harris, I., Tett, S. F. B., and Jones, P. D.: 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, 10.1029/2005JD006548, 2006.

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

Christy, J. R., Spencer, R. W., and Braswell, W. D.: MSU tropospheric temperatures: Dataset construction and radiosonde comparisons, Journal of Atmospheric and Oceanic Technology, 17, 1153-1170, 2000.

Cowtan, K., and Way, R. G.: 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, 10.1002/qj.2297, 2014.

CSIRO, and BoM: Climate Change in Australia: technical report 2007, CSIRO, Melbourne, 148 pp., 2007.

GISSTemp Team: GISS Surface Temperature Analysis (GISTEMP), NASA Goddard Institute for Space Studies. , 2015.

Hansen, J., Fung, I., Lacis, A., Rind, D., Lebedeff, S., Ruedy, R., Russell, G., and Stone, P.: Global climate changes as forecast by Goddard Institute for Space Studies three-dimensional model, Journal of Geophysical Research: Atmospheres (1984–2012), 93, 9341-9364, 1988.

Jones, P. D., New, M., Parker, D. E., Martin, S., and Rigor, I. G.: Surface air temperature and its changes over the past 150 years, Reviews of Geophysics, 37, 173-199, 10.1029/1999RG900002, 1999.

Jones, P. D., Osborn, T. J., Briffa, K. R., Folland, C. K., Horton, E. B., Alexander, L. V., Parker, D. E., and Rayner, N. A.: Adjusting for sampling density in grid box land and ocean surface temperature time series, Journal of Geophysical Research: Atmospheres, 106, 3371-3380, 10.1029/2000JD900564, 2001.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P.: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, Journal of Geophysical Research: Atmospheres, 117, n/a-n/a, 10.1029/2011JD017139, 2012.

Jones, R. N.: Detecting and attributing nonlinear anthropogenic regional warming in southeastern Australia, Journal of Geophysical Research, 117, D04105, 10.1029/2011jd016328, 2012.

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., and Saunby, M.: 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, 10.1029/2010JD015220, 2011.

Killick, R., Fearnhead, P., and Eckley, I.: Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association, 107, 1590-1598, 2012.

Lacagnina, C., Selten, F., and Siebesma, A. P.: 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, 10.1002/2014MS000341, 2014.

Lavery, B., Joung, G., and Nicholls, N.: An extended high-quality historical rainfall dataset for Australia Australian Meteorological Magazine, 46, 27-38, 1997.

Levitus, S., Antonov, J. I., Boyer, T. P., Baranova, O. K., Garcia, H. E., Locarnini, R. A., Mishonov, A. V., Reagan, J. R., Seidov, D., Yarosh, E. S., and Zweng, M. M.: World ocean heat content and thermosteric sea level change (0–2000 m), 1955–2010, Geophysical Research Letters, 39, L10603, 10.1029/2012GL051106, 2012.

Maronna, R., and Yohai, V. J.: A bivariate test for the detection of a systematic change in mean, Journal of the American Statistical Association, 73, 640-645, 1978.

Mears, C. A., Schabel, M. C., and Wentz, F. J.: A Reanalysis of the MSU Channel 2 Tropospheric Temperature Record, Journal of Climate, 16, 3650-3664, 10.1175/1520-0442(2003)016<3650:AROTMC>2.0.CO;2, 2003.

Mears, C. A., and Wentz, F. J.: 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, 10.1175/2008JTECHA1176.1, 2009a.

Mears, C. A., and Wentz, F. J.: 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, 10.1175/2009JTECHA1237.1, 2009b.

Meehl, G. A., Washington, W. M., Arblaster, J. M., Hu, A., Teng, H., Tebaldi, C., Sanderson, B. N., Lamarque, J.-F., Conley, A., Strand, W. G., and White, J. B.: Climate System Response to External Forcings and Climate Change Projections in CCSM4, Journal of Climate, 25, 3661-3683, 10.1175/JCLI-D-11-00240.1, 2011.

Meehl, G. A., Washington, W. M., Arblaster, J. M., Hu, A., Teng, H., Kay, J. E., Gettelman, A., Lawrence, D. M., Sanderson, B. M., and Strand, W. G.: Climate Change Projections in CESM1(CAM5) Compared to CCSM4, Journal of Climate, 26, 6287-6308, 10.1175/JCLI-D-12-00572.1, 2013.

Menne, M. J., Williams, C. N., and Vose, R. S.: The United States Historical Climatology Network Monthly Temperature Data - Version 2, Bulletin of the American Meteorological Society, 993-1107, 2009.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: 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, 10.1029/2011JD017187, 2012.

Osborn, T. J., and Jones, P. D.: The CRUTEM4 land-surface air temperature data set: construction, previous versions and dissemination via Google Earth, Earth Syst. Sci. Data, 6, 61-68, 10.5194/essd-6-61-2014, 2014.

Parker, D. E., Legg, T. P., and Folland, C. K.: A new daily Central England Temperature Series, 1772-1991, International Journal of Climatology, 12, 3170342, 1992.

Potter, K. W.: Illustration of a New Test for Detecting a Shift in Mean in Precipitation Series, Mon. Wea. Rev., 109, 2040-2045, 1981.

Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., Vanicek, M., Ansell, T. J., and Tett, S. F. B.: 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, 10.1175/JCLI3637.1, 2006.

Ricketts, J. H.: 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, 2015.

Ricketts, J. H., and Jones, R. N.: The multi-step Maronna-Yohai bivariate test for detecting multiple step changes in climate data, Manuscript in preparation, 2016.

Rohde, R., Muller, R., Jacobsen, R., Muller, E., Perlmutter, S., Rosenfeld, A., Wurtele, J., Groom, D., and Wickham, C.: A new estimate of the average Earth surface land temperature spanning 1753 to 2011. Geoinfor Geostat Overview 1: 1, of, 7, 2, 2013a.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C., and Mosher, S.: Berkeley earth temperature averaging process, Geoinfor. Geostat.: An Overview, 1, 1-13, 2013b.

Scott, A. J., and Knott, M.: A Cluster Analysis Method for Grouping Means in the Analysis of Variance, Biometrics, 30, 507-512, 10.2307/2529204, 1974.

Smith, R. A.: Golden Gate tidal measurements: 1854-1978, Journal of the Waterway Port Coastal and Ocean Division, 106, 407-409, 1980.

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006), Journal of Climate, 21, 2283-2296, 2008.

Trewin, B.: Techniques involved in developing the Australian Climate Observations Reference Network – Surface Air Temperature (ACORN-SAT) dataset, CAWCR, Melbourne, 92, 2012.

Vivès, B., and Jones, R. N.: Detection of Abrupt Changes in Australian Decadal Rainfall (1890-1989), CSIRO Atmospheric Research, Melbourne, 54, 2005.

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., Kearns, E., Lawrimore, J. H., Menne, M. J., Peterson, T. C., Reynolds, R. W., Smith, T. M., Williams, C. N., and Wuertz, D. B.: NOAA's Merged Land–Ocean Surface Temperature Analysis, Bulletin of the American Meteorological Society, 93, 1677-1685, 10.1175/BAMS-D-11-00241.1, 2012.

White, N. J., Haigh, I. D., Church, J. A., Koen, T., Watson, C. S., Pritchard, T. R., Watson, P. J., Burgette, R. J., McInnes, K. L., You, Z.-J., Zhang, X., and Tregoning, P.: Australian sea levels—Trends, regional variability and influencing factors, Earth-Science Reviews, 136, 155-174, http://dx.doi.org/10.1016/j.earscirev.2014.05.011, 2014.