**Authors' response**

We would like to thank both reviewers for their positive and thoughtful comments, most of which we have accommodated in some way. It is a large paper, so we are grateful for the effort and attention the reviewers have given.

The revision outlines a structure that better relates theory to plausible physical mechanisms and statistical tests at the beginning of the paper to make it quite clear what we are testing and why. The different uses of linear and linearity in the manuscript are clarified. The more philosophical passages have been removed.

**Response to Reviewer 1**

POINT 1: *My main concern is that it is not clear enough what exactly the hypotheses are that are tested. Sometimes the authors say that what is tested is whether (i) internally and externally forced components of the climate system are independent or not (page 3, page 29); sometimes they say that what is tested is whether (ii) the development of climate variables follows a trend or is step-like (abstract, Section 5). It is not the case that independence of internally and externally forced components of the climate system implies that there is a trend; this is possible, but there could also be independence and at the same time step-like behaviour. Also, it is not the case that dependence of internally and externally forced components implies that there is necessarily step-wise behavior. This could be the case, but there could also be dependence and a trend at the same time. As a result, it remains unclear what exactly is tested: (A) (Only) whether the internally and externally forced components are independent. (B) (Only) whether the climate variables follow a trend or not. (C) Whether the internally and externally forced components are independent AND whether there is a trend. (D) Whether the internally and externally forced components are dependent AND there is step-like behavior of the climate variables. Throughout the paper, the authors need to be clearer what exactly is tested.*

*"but there could also be independence and at the same time step-like behaviour"*

The exact hypotheses have been clarified, making the testing environment much clearer. In particular, how *H1* and *H2* related to $h_{step}$ and $h_{trend}$.

This has resulted in the rewriting of the introduction and an expansion of the methodology section to relate scientific to statistical hypotheses and described the tests in more detail.

*"it is not the case that dependence of internally and externally forced components implies that there is necessarily step-wise behaviour"*

True, and we have made it clear that we are not claiming this and why we are using steps, given the amount of prior knowledge we have about their occurrence.

POINT 2: *Related to this, if what is tested is (C) and (D) (as is often suggested; cf. in particular the hypotheses on page 5), then it is important to see that (C) and (D) are not exhaustive (because there are also the possibilities that there is independence and a step-wise development; or that there is dependence and a trend). The authors want to test an exhaustive set of hypotheses, but (C) and (D) are not exhaustive.*

We are not testing an exhaustive set of either physical or statistical hypotheses and have made this clear; however, understanding the theoretical background and potential mechanisms that informs the testing environment is important. The existing research identifying step changes as a mechanism for warming change on decadal timescales needs to be tested in an environment that is dominated

by an existing paradigm that says the opposite. Because paradigms are partly a sociological construct (though are informed by theory) only a strong case that can cover the theory along with addressing cognitive values has a chance of being accepted.

The framework we are using is a theoretical-mechanistic – statistical induction framework where theory is used to distinguish plausible mechanisms that allow a clear choice to be made between alternative hypotheses. This will be described more clearly. Statistical testing provides the means to do this, so statistical hypotheses need to be developed that match the alternative mechanisms.

*POINT 3: Throughout the paper the assumption seems to be that "trend-like" and gradual as opposed to step-wise and non-gradual means that there is a linear relationship (e.g. on page 7). It is unclear why gradual implies that there is a linear relationship. There can be gradual behavior with various kinds of relationships (a quadratic relationship etc).*

Here linear is referring to whether the temperature signal follows a secular trend therefore can be defined by a line and refers to the linear transfer relationship between forcing and temperature. Nonlinear response implies the transfer is nonlinear and is modified by nonlinear physical processes. This has been clarified in Section 2 para 2. The usage in the paper is physical and when statistically linear is used is will be specific, e.g., linear trends.

*POINT 4: On page 7 the hypothesis states that there is a "(probably monotonic)" trend. The brackets are confusing. Is it now tested that the trend is monotonic or is it allowed that the trend is not monotonic?*

Related to the previous point – we have clarified that the underlying signal of atmospheric warming is assumed to be monotonic (while we accept that total global warming is monotonic). Section 2 has been rewritten as per the overall response to both reviewers 1 and 2.

*POINT 5: on page 5 six tests are described. It should be clearly stated which tests test which hypotheses (becomes clear later, but should be stated clearly early on).*

The tests have been clarified. They all test the hypotheses but Tests 1–4 are mainly probative, Test 5 a mixture and Test 6 is largely error testing to determine whether steps or trend best describe the data.

*POINT 6: The beginning of Section 2: here it is argued that the gradualist thesis is derived from induction. Yet, as the paper later argues, the data actually do not support the gradualist thesis and the gradualist thesis rather seems to be often adopted for no empirical reasons (convenience, simplicity). Hence it seems that the gradualist thesis is not justified by induction after all.*

It is probably more accurate to say the gradualist thesis is sustained by statistical inference (as a form of induction). The paper has been made clearer in this regard and focuses on severe testing.

Below are some notes but this discussion is omitted from the paper.

We have done some more work around this to inform other papers on the same theme. In the 1970s, various theoretical arguments were put to suggest if a system received a small internal forcing the response would be proportional (e.g., Leith 1973, 1975). Much of the subsequent work was based on vectorizing the forcing-response relationship, which will produce a linear outcome. This was based on statistical physics.

There are two types of induction therefore: one is largely analogical, based on statistical physics, proposed by scholars like Leith. This is now generally accepted as an approximation that only holds if

the residuals are Gaussian (e.g., Palmer, 1999) and they are not in every case. There have been two styles of statistical induction. The first in the 19th century based on, which is truth-centred. The other is modern statistical induction, which is often reduced to a mechanised form based on establishing a sufficiently small probability for the null hypothesis. For some, the trend has again become truth-centred, but this can is only be the case if warming is gradual. Otherwise, it is a statistically-derived approximation of the relationship between forcing and response. This issue is at the core of the paper.

Other cognitive values such as convenience and simplicity are applied, because they are mentioned in the literature frequently, often as an escape clause to bypass some acknowledged but unknown complexity. They are also used as a defence against over-parameterisation and overfitting. Their use is mainly sociological rather than scientific. A full discussion of all of these is not appropriate for this paper, so we will pare it back to the basic issues – we acknowledged these other issues because readers will realise that the story is incomplete but a more comprehensive exploration is not feasible.

This issue has been complicated by the climate wars, where trend analysis has been used as a defence, so if not strictly truth-centred, it had to be defended as scientifically correct. This has hampered the consideration of alternatives. This comment is largely background because most of the philosophical content would be removed in a revision leaving the severe testing component.

*POINT 7: The beginning of Section 2: "The application of linear trend analysis to atmospheric warming is invariably justified as inference to the best explanation". I am puzzled by this sentence. Why is there suddenly a reference to the inference to the best explanation (previously the matter of concern was induction).*

This point has been omitted from the revision.

Technical corrections have been addressed.

**Response to Reviewer 2**

**General comments**

The single paper option has been chosen as recommended by the editor and the introduction and Section 2 rewritten to focus on the statistical testing. However, it is important to preface the statistical tests with the probative conditions as to why those tests were chosen. The relationship between theoretical propositions of independence and interacting externally-forced and internally-generated change have been simplified (as per the response to reviewer 1), the mechanisms reflecting those theoretical propositions and why they were selected and the statistical hypotheses are better described. The most important aspect of this is to test between sustained incremental change as would be manifest in gradual warming and episodic change. For that reason, retaining Page 5 lines 6–19 is important because it explains what the statistical hypotheses represent.

**Specific comments**

Abstract – test numbers bracketed

Introduction – see above

2.3 line 23 – ill-posed inverse problem will be given a plain-language description and term has been removed. This relates to part of test 1: regional stratification – the test description has been revised.

External forcing – linear – the explanation referred to in the response to Reviewer 1's comments has clarified the relationship between physical and statistical linearity.

Decadal scale – the most widespread usage is around timescales, not means. We don't see how this can be confusing. CMIP time series are all annual – we're not sure where the 5-year figure comes from. The only place where five-year averages are used is for estimating total warming over a given time period. We do compare steps within set decades to ECS but that is a special case.

Page 6 – *I do not see why entrainment of heat energy into the various heat reservoirs of the Earth and especially the hydrothermal system need always be nonlinear*

Perhaps this is a short-hand argument, but the transport of heat from the equator to the poles is fundamentally nonlinear at the global scale. Our substantial argument has been moved to the discussion and the linearity/nonlinearity argument much better focussed.

Page 6 – *Lines 2 to 6 outline a number of alternative approaches to determine 'shifts', 'change points', 'step changes', but there is no discussion of the advantages/disadvantages of these different approaches and why they were not used in this study. See also: Drijfhout et al. (2015) and Reid et al. (2016) Reid PC, Hari RE, Beaugrand G et al. (2016) Global impacts of the 1980s regime shift. Global Change Biology, 22, 682-703.*

An objective rule-based version of the bivariate test was the best tool we could use, based on our previous use of the test and the effort that has been put into developing the multi-step rule-based model. This has been discussed in past papers and we will be a little more emphatic about it. This discussion will be held in another paper currently being prepared by Ricketts, who does test some of the alternatives. There is a good summary in Rodionov (2005) and the bivariate test is on a par with the Alexanderssen test (Rodionov does not mention the bivariate test but colleagues at the Australian Bureau of Meteorology tested both in the 1990s when developing homogenization strategies and judged them to have similar performance). The bivariate test has the advantage of being able to use different reference time series. The multi-step procedure was developed to overcome the problems with multiple steps, where the test results do not hold – whereas they will for sequential testing.

This has been clarified in the text

*Page 7/8 – acronyms*

We have put a Table for acronyms in the Supplementary Information (Table S4) – ECS is defined on Page 6.

Page 7 – ECS

Its use is explained in more detail later, as an independent variable against which to measure timeseries components through correlation to determine which carries more signal and which carries more noise. We have slightly expanded the explanation of the tests as also requested by Reviewer 1

Page 9 lines 22 to 28. *A diagrammatic representation of the different terms used for the analyses is needed. A descriptive expansion of what is meant by each of the terms would be helpful. The word 'shift' has been used in a different way in previous papers and a different word would be more appropriate here for this characteristic. Is the text in brackets at the end of the last bullet correct?*

Finding terminology to go with nonlinear change and defining and measuring it is difficult, whereas there is so much language associated with trend analysis and framing around this that we are used to. We retain steps for the bivariate test because that is what it measures (as is the case for the STARS (Rodionov) test). After much consideration, shift was chosen as the most representative measure of visually what can be seen and measured across the gap produced by displaced trends. A diagram will be produced. Note that the term regime is still being debated (e.g., Overland et al., 2008) – a scientific language for nonlinearity needs to be developed.

Some words to this effect and a figure have been added to a new Section 2.3.4 on metrics.

Page 23 Section 5.1. *This section would be better drafted as the conclusions of the paper rather than as a summary of severe testing.*

We would prefer to leave this here and focus on this summary in the discussion – it is a long paper and we see a summary and the conclusions as being slightly different. We also think this summary is needed to help frame the discussion

Page 25 line 2. *Again I do not like the use of the word decadal here. Table 6 does not show that $h_{step}$ is better at a decadal scale the steps are occurring within a year, but may continue at a new level or develop a trend afterwards for more than a decade*.

Can be changed to decadal timescales, but the use of decadal scale to signify timescales of decades is almost ubiquitous in climatology and if decadal means are signified, decadal mean is the term usually used (Google scholar confirms this if "decadal scales" and climate are search terms). We are quite puzzled with these objections to using decadal in this sense. The only exception might be for ocean sediment coring, where sampling horizons can be decades and centuries.

Page 27 line 1-2. *The hiatus is now thought to be due to an increased storage of heat deeper in the ocean and is not a continuing event considering the warming of the last few years. See Reid 2016 and references included.*

We disagree with this interpretation because we view the hiatus as a regime, and the steady state in between step changes a normal part of climate. The discussion has been expanded to describe our views on the role of regimes.

The so-called hiatus and previous mid-century pause (Wally Broecker coined that term) was clearly related to a La Niña phase of the Interdecadal Pacific Oscillation, however, for the 1977 – 1996 period of the El Niño phase, we suggest there were two steps rather than a trend. We are likely to be in the next step change and Peyser et al. (2016) have identified the trigger for this in dynamic changes in sea level in the warm pool, leading to an outburst of heat (They interpret this as variability, but our view is that it is a nonlinear expression of the climate signal).

In the discussion, we reinterpret two recent publications (Peyser et al., 2016; Meehl et al., 2016) that came out after the initial submission to explain the trigger for the recent warming and that which occurred in 1996–98 (Peyser et al., 2016). Meehl et al. (2016) suggest that the IPO may be switching from negative to positive that they interpret as the resumption of an increased warming trend, largely similar to the comment above.

Page 18 line 4. *Lack of predictability. How can the authors be so definite that this might be due to aerosols?*

Because of the negative correlations associated with decades of negative steps responding to volcanic eruptions and sulphate aerosols of the immediate post WWII period and the positive correlations cancelling each other out – as discussed on page 17 – not sure why this is contentious – text slightly edited to make this more obvious.

Page 14 lines 23-25. *Sea level steps are said to be ubiquitous in local tide gauge time series, Table 3 in Jones et al. 2013, but were not checked or analysed by Jones et al.*

Misreading of the table therein – model sea level not checked – that statement was for observations. Tide gauge records are illustrated in Figure 36 (Fremantle, San Francisco) of that reference. A new Figure, Figure 5 contains a panel of four non-temperature step changes in the revised version to better illustrate Test 5: two tide gauge record (San Francisco, Fremantle), rainfall (northern Australia and south-west Western Australia) and shallow ocean heat content.

Technical corrections – all corrections addressed unless otherwise indicated

Page 2 Line 2. Abstract. *Change to: 'variations that extend over decadal scales of time'. See later comment on use of decadal.*

See responses above

Page 2 Line 13. *'the correlation'*

2.2 Line 21. *First mention of H1 and H2 together. They were used separately in the introduction.*

These have been rewritten as per both reviewers' comments.

Page 6 Line 7. *Start 'For H1: : :.'.on a different line to make it comparable to H2 below. There are no citations to back up the statements made in the H1 section.*

The H1 case is cited in paragraph 1 and the Corti et al and Hasselmann references describe both.

Line 18. *Decadal again. The transfer from one regime to another is evident at an annual level and not decadal.*

See comments above

Line 17. *Should not be numbered 3 or indented.*

No. We think there a three distinct points, rather than two choices. Wording amended to clarify this.

Lines 25 to 30. *This text should be part of a discussion and not here.*

Incorporated into the discussion

Line 33. *At the end it is important to note that regime change is precipitated, but to a new level or a trend.*

Incorporated into the discussion – Peyser et al. and Meehl et al. (2016) allow us to better identify the mechanism.

Line 31 to Page 7 line 4 *repeated below.* – will remove

Page 7 line 12. *–H1 and –H2 mentioned for the first time. Define what they mean in general language.*

Section rewritten and *-H1* and *-H2* have been removed to simplify.

Page 7 line 13 onwards – *Six tests are identified. It is not clear if the first two are the same as the two tests mentioned on page 9 lines 31-32. Please make this section clearer.*

They are the same – have expanded this section to say what the tests do in more detail.

Page 8 line 19/20 *"MYBT is considered reliable". Is this remark necessary without some backup? You could refer to page 13 line 2 in the Supplementary Information*

We are be more direct about why the bivariate test is being used in Section 2.3. Twelve months' development went into the rule-based multi-step component adapting the original test with a great deal of testing to ensure that the results were robust, consistent with known steps and the test could be as reliable as possible with data that contained real trends and lagged autocorrelations (ENSO-like). There is no doubt that redness itself will produce shifts in the data, which is why theory is so important when trying to interpret the results. Some of this is documented in Ricketts (2015), but unfortunately the final paper describing the model has not yet been finalised.

Page 9 line 2. *Put in a heading Data and distinguish between the observed and modelled time series by putting them in different paragraphs. It would have been helpful to leave a line space between each paragraph.*

Done

Page 9 line 13 *Again provide a new sub-heading*

Ok

Page 9 line 32. *Again, does the reference to Test 1 and 2 refer to the first two tests of the six mentioned earlier?*

Yes – made clear

Page 11 *below line 13 put in a heading: Shift/Trend Ratios*

Ok - added

Page 12 lines 1 and 2. *An important result. Missing full stop after warming.*

Ok

Page 12 line 7. *Suggest change to "Annual and seasonal anomalies were investigated". And edit next sentence so not starting with Annual.*

ok

Page 12 line 12. *Why are quarterly anomalies only examined for the satellite temperatures?*

*This needs explaining.*

Sentence added to say that seasonal anomalies were examined to distinguish between 1995 and 1998 step dates. Quarterly (3-monthly) time series were assessed for satellite and two surface temperature records.

Page 12 line 23. *Confirmation of the results from Reid et al. 2016 that the 1987 regime shift is evident at a global scale and yet on the next line it is said to be only evident at a regional scale.*

Have added that these results were mainly due to using different area averages – however, not confident that all the step changes/shifts identified by Reid et al., (2016) should be allocated to 1987/88 (e.g., Australia).

Page 12 line 25 and 26. *An important result. "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)".*

Page 12 lines 333-34. *An important observation. "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". As is the subsequent comment on heat release from the ocean during El Niño. See commentary in Reid 2016 on this issue.*

We will go into this into the discussion, especially through Peyser et al. (2016)

Page 13 line 5. *Which timescale?* – clarified

Page 13 line 14. *Insert 'out' after carried.* – ok

Page 14 line 18 *An important observation. "indicates that the onset of the warming signal in these broader regions is abrupt (Jones, 2012)".*

Page 14 line 21 *Use year (2016) of hard copy publication for Reid et al. (2015).* – ok

Page 14 lines 23-25. *Sea level steps are said to be ubiquitous in local tide gauge time series, Table 3 in Jones et al. 2013, but were not checked or analysed by Jones et al.*

Misreading of the table therein – model sea level not checked – that statement was for observations. Tide gauge records are illustrated in Figure 36 (Fremantle, San Francisco) of that reference and are updated here.

Page 17 line 4. *Why are 5 year averages used here, the first mention that the data has been treated in this way.*

This is a simple difference using 5-year averages to avoid single-year variations. The IPCC often uses 10-year averages in a non-stationary system for simple differences, we use five to minimise the sampling errors with one year but wanted to keep this interval as short as possible. Because it has been done consistently for an ensemble it will provide consistent results.

Page 18 line 33. *Lable bullet A1 and at top of next page the bullet A2*.

Not sure what A1 and A2 signify – left unchanged

Page 19 lines8-9. *"peaking in the 2080s: : :.." does not fit with the figure 5f. What should be Fig. 5f is a duplicate of Fig. 5d.*

Not sure how that happened – have fixed

Page 20 line 23. Is the first part of Section 5 essential to the paper? Would it be better to label it 'Sensitivity testing'.

Yes, it is essential and is a comprehensive way of testing whether the time series are steplike or trend-like on the timescales of interest. The section has been edited slightly to sharpen it and we are making it clear this is largely error testing.

Page 20 line 25. *Insert 'and' after 'warming'?*
Page 21. Line 30 change to: *'performs the best'*
Page21 line 31-33. *Duplication 'into the' and 'test'. Change to: 'at a global scale when each model is'*
Page 22 line 30. *'21st'*
Page 24 line 9. *Spelling 'are' not 'area'*
Page 24 line 18-19. *Edit sentence beginning: 'Warming is not: : :.'*
Page 24 lines 24-25. *Make sure this statement is backed up by appropriate citations in the results section*

All ok

Page 24 lines 9-10 and 30-32 *repetition. Is this necessary.*

Section edited and this has been removed

Page 25 lines 5 to 10. *Delete 'In summary' and draft as the final paragraph of the conclusions.*

Will move into the conclusions and edit

Page 25 line 7 *in situ in italics. And, 'or as a gradual'.*

Page 25 lines 9-10. *Edit to: 'where climate change and variability interact rather than varying independently.'*

Page 25 line 13 *Discussion. Include a discussion of how the results of Drijfhout et al. (2015) compare to those presented in this paper.*

Drijfhout S, Bathiany S, Beaulieu C et al. (2015) Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models. Proceedings of the National Academy of Sciences, 112, E5777-E5786.

Done

Page 25 line 15 *change 'earlier' to 'before'?*
Page 25 line 17. *'gradualism' and 'as a key tool to understand how'?*
Page 25 line 23 *'to explain climate'*
Page 25 line 24. *Change' 'covering methods' to 'applying procedures'?*
Page 25 line 25. *Delete 'and its application to understanding climate processes'*.
Page 26 line 5. *'analytical'.*
Page 26 line 12 *a priori Italics*
Page 26 lines13-14. *Important observation that needs to be included in the conclusions. 'the processes involved are timescale invariant indicate that the meaning of seamless has not really been thought through'.*
Page 26 line16. *'would likely be'. And change 'considerable' to 'sizeable' as repeated on the next line.*
Page 26 line17. *Change 'under' to 'that have'.*

Page 26 line 19. *First sentence of bullet. Something is missing.*
Page 26 line 20-21. *'physics, understood as being primarily linear and hydrometeorology with its substantial nonlinear behaviour; both remain largely unreconciled.*

All the above removed except for the bullet point on decadal prediction, retained in discussion

Page 27 line 9. *'stated'*

Removed

Page 27 line 20. *Somewhere in the text above it is worth stating that both Cahill and Foster consider that the hiatus was a non-event.*

This passage removed from the discussion

Page 28 lines 7-9. *Delete: 'As we discussed in a related paper where H2 is examined in greater detail' and the reference to Jones and Ricketts, 2016 as this paper is only 'in preparation'. Edit the sentence without the above text except for H2.*

Text removed and discussion rewritten

Page 28 lines 11-18. *An important paragraph. You might also cite Roemmich's recent papers and Reid 2016 to back up this paragraph*.

Cited

Page 28 lines 19- 22 *repeated below on lines 23-26.*

Removed

Page 28 line 31, *The word 'extraordinary' is perhaps a bit too strong*.

Edited

Page 28 line 32. *'to either side'*

Page 29 lines 1-2. *Leave out the sentence: 'Elsewhere : : :..'., but, raise the possibility that we are undergoing another shift at present.*

Substantially rewritten with new research to suggest that we are undergoing another shift at present

Page 29 lines 3-5. *Poor ending to this section. Edit and improve as a statement to round off the discussion.*

Completely rewritten – this point has not been retained.

Page 29 line 13. *See earlier comment on >50 year climate change*.

The context has been made much clearer – we state that it is a complex trend over the long term – this is physically important as it relates to changing boundary conditions. The theoretical background for why this is important has been expanded

Page 29 line 17. *Delete sentence beginning: 'We discuss this : : :..'*

Done

Page 46. *Figure 4. I don't know what the journal policy is for sub-figures, but I prefer the lettering, a, b, c to be in the top left hand corner, inside the enclosing border of each sub-plot. It would also help if*

*the respective sub-plots were labelled: England, Texas and Australia within the enclosing border. Insert at the beginning of the legend: 'Regional temperature change'.*

Page 47. *Same comment as for Figure 4. Label a, b, c, d, e, f in the top left corner of the subplots and in the top right in order: Add in sequence in the top right corner of a: 'observed', of b: simulated, of c: '2.6', of d: '4.5', of e: '6.0' and of f: '8.5'. In the legend add downward blue and upward red as for Figure 1.*

Figures edited as per suggestions

**References**


Broecker, W. S.: Global warming: Take action or wait?, Jokull, 1-16, 2005.

Drijfhout, S., Bathiany, S., Beaulieu, C., Brovkin, V., Claussen, M., Huntingford, C., Scheffer, M., Sgubin, G., and Swingedouw, D.: Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models, Proceedings of the National Academy of Sciences, 112, E5777-E5786, 2015.

Jones, R. N., Young, C. K., Handmer, J., Keating, A., Mekala, G. D., and Sheehan, P.: Valuing Adaptation under Rapid Change, National Climate Change Adaptation Research Facility, Gold Coast, Australia, 182 pp., 2013.

Leith, C.: The standard error of time-average estimates of climatic means, Journal of Applied Meteorology, 12, 1066-1069, 1973.

Leith, C.: The design of a statistical-dynamical climate model and statistical constraints on the predictability of climate, in: The Physical Basis of Climate and Climate Modelling, World Meteorological Organisation, Geneva, 137-141, 1975.

Meehl, G. A., Hu, A., and Teng, H.: Initialized decadal prediction for transition to positive phase of the Interdecadal Pacific Oscillation, Nature Communications, 7, 11718, 10.1038/ncomms11718, 2016.

Overland, J., Rodionov, S., Minobe, S., and Bond, N.: North Pacific regime shifts: Definitions, issues and recent transitions, Progress In Oceanography, 77, 92-102, 2008.

Palmer, T. N.: A nonlinear dynamical perspective on climate prediction, Journal of Climate, 12, 575-591, 1999.

Peyser, C. E., Yin, J., Landerer, F. W., and Cole, J. E.: Pacific sea level rise patterns and global surface temperature variability, Geophysical Research Letters, n/a-n/a, 10.1002/2016GL069401, 2016.

Reid, P. C., Hari, R. E., Beaugrand, G., Livingstone, D. M., Marty, C., Straile, D., Barichivich, J., Goberville, E., Adrian, R., and Aono, Y.: Global impacts of the 1980s regime shift, Global Change Biology, 22, 703, 10.1111/gcb.13106, 2016.

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.

Rodionov, S. N.: 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, 14-16 June 2005, 2005.

# Reconciling the signal and noise of atmospheric warming on decadal timescales

Roger N Jones* and James H Ricketts

Victoria Institute of Strategic Economic Studies, Victoria University, Melbourne, Victoria 8001, Australia

*Correspondence to*: Roger N. Jones (roger.jones@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~~ distinguish between the two statistical hypotheses, $h_{step}$ and $h_{trend}$. Test ~~(~~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)~~Test 2: Correlations between step-change frequency in observations and 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), the correlation increases to 0.78. ~~(3)~~Test 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)~~Test 4: In three regions tested, the change between stationary and non-stationary temperatures is steplike and attributable to external forcing. ~~(5)~~Test 5: Steplike changes are also present in tide gauge observations, rainfall, ocean heat content~~, forest fire danger index~~ and related variables. ~~(6)~~Test 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.

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

## 1 Introduction

The dominant paradigm for how the climate changes over decadal timescales is based on the standard signal-to-noise model, where the externally-driven signal of climate change forms a trend surrounded by the internally-generated noise of climate variability. Here, the external driver of interest is radiative forcing produced by anthropogenic greenhouse gas emissions, mediated by other anthropogenic emissions such as sulphate aerosols and black carbon. This paradigm is widely represented by trend analysis, which extracts a monotonic signal from a noisy time series (e.g., North et al., 1995;Hegerl and Zwiers,

2011;Santer et al., 2011). The resulting methodology dominates climate practice, forming the basis for detection and attribution, projection, prediction and characterisation of climate risk.

However, it is not the only theoretically plausible representation of a changing climate (Palmer, 1999;Branstator and Selten, 2009;Solomon et al., 2011;Kirtman et al., 2013). ~~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.~~

~~1.~~ The two main hypotheses that describe ~~the interactions between~~ how externally-driven and internally-generated climate may ~~be related change and variability~~ over decadal timescales are (Corti et al., 1999;Hasselmann, 2002):

*H1.* Externally-forced climate change and internally-generated natural variability change independently of each other.

*H2.* They interact, for example, 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 interactions can lead to a range of different outcomes. For global mean surface temperature, the signal is generally portrayed as following a linear pathway that conforms to the relationship $\delta T = \lambda \delta F$, where $T$ is temperature, $F$ is forcing and $\lambda$ is a constant related to feedback processes (Ramaswamy et al., 2001;Andrews et al., 2015). This is widely accepted for both *H1* and *H2* over longer timescales (e.g., >50 years), but how boundary-limited and initial conditions uncertainties combine over shorter time scales remains uncertain~~{Hansen, 2011 #5287}~~.

For *H1*, if the response to external forcing is considered to be independent of variability over shorter timescales (<50 years), the trend model will hold, despite often being obscured by variability. Such variability is generally represented as stochastic behaviour in annual to decadal phenomena, where teleconnections, lagged effects and regime changes all potentially interact (Solomon et al., 2011;Kirtman et al., 2013). Alternatively, instead of a gradual line or curve, a segmented trend is sometimes proposed, where the signal of atmospheric warming is modified by varying decadal regimes governing oceanic sources and sinks of heat (Meehl et al., 2013;Cahill et al., 2015;Trenberth, 2015).

The potential behaviour of warming under *H2* has many possible permutations because the signal may project onto the regime-like structures of decadal climate variability, or may dynamically modify those structures. Here, we deal with one such type of response, ~~in~~ manifesting as step changes. Step changes have been detected in warming and related climatic variables by several different methods (Jones, 2010;Reid and Beaugrand, 2012;Jones et al., 2013;Belolipetsky, 2014;Belolipetsky et al., 2015;Bartsev et al., 2016;Reid et al., 2016); in one case, step-like warming over SE Australia has been attributed to

anthropogenic forcing (Jones, 2012). The purpose of this paper is to detect step changes in a range of temperature records and to apply severe testing to steps and trends to determine which carries the greater part of the warming signal. The results are used to determine whether *H1* or *H2* is the more viable hypothesis and, if the signal is shown to be nonlinear, to explore the nature of the interaction between external forcing and internal variability.

~~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~~ We apply a methodology combining theoretical-mechanistic and statistical-inductive reasoning to test which model, step or trend, better represents the warming signal on decadal timescales. It is applied to the substantive null of model adequacy approach described by Mayo and Cox (2010) as part of ~~explores the potential for *H2* to be true by applying~~ severe testing principles articulated by Mayo and Spanos (2010) ~~to detect and analyse step changes in temperature data~~. The theoretical-mechanistic component is used to outline plausible, alternative physical processes required to sustain steps or trends. Step changes are measured using an objective rule-based multi-step adaptation of ~~T~~the bivariate test of Maronna and Yohai (1978) ~~is used~~ to analyse 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. The data produced by those analyses is then subject to six tests designed to distinguish between steps and trends as the main driver of the anthropogenic climate signal over decadal timescales.

## 2 ~~Analytic Framework~~Methodology

The process of theoretical-mechanistic and statistical-inductive reasoning requires matching scientific hypotheses (*H*) with statistical hypotheses (*h*) in order to distinguish between alternative hypotheses. The next few sections detail how this has been carried out. This employs a hierarchy of models between theory and data consistent with that suggested by Suppes (1962) and articulated by Haig (2016). This separates underlying theory, from experimental models that apply statistical induction, from statistical models to prepare primary data for testing. By and large, the statistical models are used to undertake error testing and the experimental models, probative testing of theoretically plausible hypotheses.

~~3~~ We define linear and nonlinear from a physical rather than statistical perspective. Physically linearity will produce a proportional response to a stress, whereas for statistical linearity, the parameters of a relationship are linear. For temperature, a physically linear response to forcing will produce a straight line or curve, whereas a nonlinear response will result in a discontinuity, such as a step change. ;W~~w~~here the terms linear and nonlinear are used, they refer to the nature of physical responses within the climate system. When linear trends are used, they are referred to specifically.

### ~~3.1~~2.1 ~~Reasoning by statistical inference~~Development of physical mechanisms for probative testing

Application of a theoretical-mechanistic process starts from well-agreed theoretical positions (core theory), then builds on that theory to explores alternative mechanisms required to support competing hypotheses. The e~~E~~xplor~~a~~tion of~~ing~~ plausible mechanisms produces probative criteria for severe testing. This paper cannot undertake a full survey of the theory behind anthropogenic global warming, but the trapping of heat by added greenhouse gases, creating an imbalance between the surface and the top of the atmosphere, and between the equator and the poles is widely agreed as the foundational theory. However, in between the time when heat is trapped in the atmosphere and when it is measured as a change in temperature there is a gap in understanding, which has competing explanations. These explanations focus on how that trapped heat is stored within the climate system and subsequently distributed. Most techniques analyse observed or modelled climate and use inductive reasoning to infer the pathways involved. ~~-~~Because *H1* implies a gradual signal and *H2* ~~has~~ a nonlinear signal, ~~here~~ represented here as steplike change, these pathways will be distinctly different.

For *H1*, close adherence to a warming trend implies that the atmosphere warms gradually. If so, this must occur via either of the following processes or a combination:

1. A measurable proportion of radiatively-forced anthropogenic warming trapped in the atmosphere is retained *in situ*, given that it is well understood that most of the trapped heat is absorbed by the ocean. 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 from the surface.
2. Most of the heat trapped by anthropogenic greenhouse gas forcing is absorbed by the ocean, ~~parhaps~~perhaps at varying rates of take-up 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 may be due to varied take-up or release of heat from the ocean, or both.
3. If both 1 and 2 are operating, then the warming rate in the atmospheric component would be gradual and the contribution from the ocean governed by interannual and decadal variability. This would be best represented by a segmented trend if decadal-scale regimes of deep and shallow ocean mixing of heat are a key factor.

Nonlinear warming requires mechanisms such as regime change combining with storage and release processes. On decadal timescales, ocean-atmosphere interaction is the only possible source for such changes. If warming is mediated by the hydrothermal ocean-atmosphere system, it could be entrained by the nonlinear processes involved in the distribution of energy skywards and polewards from the equator through quasi-oscillatory systems (Ozawa et al., 2003;Lucarini and Ragone, 2011). Lucarini and Ragone (2011) describe the overall process of distribution 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). One important distinguishing characteristic for nonlinear behaviour in a changing climate is whether it is internally-generated and essentially random, whereas if it is forced, the response will be related to changing boundary conditions (Lorenz, 1975;Hasselmann, 2002). Distinguishing between these possibilities is the focus of the testing regime: whether gradual or step-like changes provide the better explanation for the response to external forcing.

1. 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.

2. 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.

3. 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 $p H_0 < 0.05$ 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.

4. 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.

5. 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.

### 3.22.2 Development of severe testing

The aim of severe testing is to produce highly probed rather than highly probable results (Mayo, 2005). 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.

Two sets of data are produced representing competing statistical hypotheses $h_{step}$ and $h_{trend}$. These are statistically distinct models linked to rival hypotheses *H1* and *H2*. The aim of testing is to produce highly probed rather than highly probable results (Mayo, 2005). Previously, statistical testing of alternative structures for warming has been inconclusive. For example, when Seidel and Lanzante (2004) tested trends, steps, segmented trends and step and trend statistical models, no single model stood out. They concluded that detection and attribution studies should consider abrupt changes. Studies that extract short-term components of climate variability from time series to producinge a more trend-like result (Foster and Rahmstorf, 2011;Werner et al., 2015) or decompose temperature timeseries into separate signal and noise components (Wu et al., 2011;Yao et al., 2016) all implicitly assume *H1*. Consequently, The exact nature of change on decadal timescales remains an open question (Trenberth, 2015). If warming conforms to a long-term complex trend and is additive Marvel et al. (2015) such studies will only produce a trend-like output because they are not configured to detect alternative structures. However, because they are framed on *H1*, these tests do not show that such structures do not exist.

Therefore, $h_{trend}$ has never been severely tested to the point where its other alternatives have been eliminated. {Marvel, 2015 #5133@@author-year}While the null hypothesis for *H2* is *H1* and for $h_{step}$ is $h_{trend}$, the null hypothesis for $h_{trend}$ is 'no trend has emerged from background variability'. Another complication is that nonlinear change on decadal timescales has been used

to challenge global warming theory on the basis that if observed change is not gradual, climate change is either disproven or overstated (e.g., Legates et al., 2015). 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 severe testing needs to cover both *H1* and *H2*.

The following six ~~probative~~ tests are used to test the relationship between linear and nonlinear behaviour and their responses to external forcing:

Test 1 Comprehensive and ~~S~~stratified analysis of change points: the timing and distribution of change points and their relationship with known regime changes and with each other. Global, hemispheric and zonal analyses of observed temperature allow~~s~~ global and regional changes and their timing to be identified. Change points aligning with known ~~nonlinear~~ climate processes indicate a causal link.

Test 2 Identification of similar patterns of steps between observations and historical climate simulated by physical models indicates a physically coherent origin, rather than random stochasticity.

Test 3 Partitioning effects for independent testing: using internal trends and shifts (steps minus internal trends) to estimate the gradual and rapid warming components in a single time series, and testing each of these against criteria such as total warming and equilibrium climate sensitivity (ECS) in observations and models separately.

Test 4 Detection and attribution: testing stationarity and change – 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.

Test 5 Testing of other variables including rainfall, sea surface temperatures, sea level rise, and ~~air pressure~~ocean heat content, to see whether they undergo similar changes.

Test 6 Direct testing for step- and trend-like structures in time series.

The first four tests can be considered largely probative, where $h_{step}$ and $h_{trend}$ are tested to determine whether *H1* or *H2* provides the better explanation for the relationship between external forcing and internal variability. The last two focus mainly on error testing to see how well $h_{step}$ and $h_{trend}$ explain the climate data. The combination of different tests means that deriving a single probability through an objective process is not possible. The procedure we follow here ~~turns severe testing into a~~ uses a two-sided test between $h_{step}$ and $h_{trend}$ as representatives of *H1* and *H2*. Paraphrasing Mayo and Spanos (2010) to address the results: with very high probability, T~~t~~ests 1–6 would have produced a result that accords less well with *H2/H1* than does *H1/H2*, if *H2/H1* were false or incorrect.

### ~~3.3~~2.3 Statistical testing

The Maronna-Yohai bivariate test (MYBT, Maronna and Yohai, 1978) is used to detect step changes in temperature data. ~~The MYBT~~ This test has been widely used to detect inhomogeneities in climate variables (Potter, 1981;Bücher and Dessens,

1991;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;Vivès and Jones, 2005;Boucharel et al., 2011;Jones, 2012;Jones et al., 2013). ~~The principal author~~One of us (Jones) has been using it for 25 years, both for adjusting inhomogeneous data (Jones, 1995;Kirono and Jones, 2007) and also for detecting abrupt changes in climate variables. Surprisingly, the MYBT is rarely included in reviews of change point analysis techniques (Rodionov, 2005;Reeves et al., 2007) despite being on a par or better than other techniques (Vivès and Jones, 2005). For example, it performed similarly to the STARS test in Jones et al. (2013) but has the advantage of not needing ~~to be~~ tuninged and being able to accommodate a reference data set~~,~~. ~~providing a degree of flexibility that few other tests have.~~ That made it our testing model of choice, especially because all six tests use here compare step changes in time series to a null reference and Test 4 assesses ~~nonlinear responses~~ step changes ~~in~~ between correlated variables.

The major advance required was to ~~take the~~ adapt the test from assessing single to multiple points ~~that uses~~ by developing an objective set of rules ~~to~~ that would detect a minimal and stable configuration ~~of~~ of multiple step changes. Previously~~,~~ ~~it was this involved~~ a trial~~-~~ and~~-~~ error process of detecting a robust set of step changes one at a time. A multi-step, rule-based application of the MYBT was developed to ~~expand the original test~~carry this out (Ricketts, 2015, see Supplementary Information for details).

The test adapts the formulation of Bücher and Dessens (1991) testing 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. This 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. This procedure may produce several different but related solutions (~~solution =~~ sets 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~~ thresholds 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.

### ~~3.3.1~~2.3.1 ~~Local~~ Regional attribution

~~Local~~ Regional attribution of step changes (Test 4) 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.

### ~~3.3.2~~2.3.2 Observed data

Time series 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., 2012;BEST, Rohde et al., 2012;C&W, Cowtan and Way, 2014), 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. Lower t~~T~~ropospheric satellite temperatures from two groups (UAH, Christy et al., 2003;Christy et al., 2007; RSS, Mears and Wentz, 2009) are also tested.

For the regional data, Australian data was sourced from the Australian Bureau of Meteorology, Texas data from the National Climate Data Center and central England temperatures from the Met Office Hadley Climate Centre. Tide gauge records were sourced from the Permanent Service for Mean Sea Level and the ocean heat content records from the KNMI Climate Explorer. The specific records used ~~are~~ are ~~detailed~~ described in the Supplementary Information.

### ~~3.3.3~~2.3.3 Model data

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). The output data are provided in the Supplementary Information.

### ~~3.3.4~~2.3.4 Metrics

Measurement of change where nonlinear behaviour is present is not an exact process, and there is not any established terminology that carries commonly understood technical meanings, so here we define a limited number of terms used in the paper. The ~~bivariate test~~ MYBT measures total change between segments of a timeseries, ignoring any trend that may be present. These we refer to as steps. Internal trends are calculated between steps and 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 (Fig. 1):

- Steps – measurement of the whole change across a discontinuity 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 using ordinary least squares trend analysis.
- Shifts – measurement 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).

**Figure 1: Record of mean annual surface temperature anomalies 1880–2014 from the Hadley Centre and Climate Research Unit (HadCRU), showing step changes ($p$<0.01), internal trends and shifts, taken from the end of one internal trend to the start of the next across a step.**

## 43 Results – observations

### 4.13.1 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 Ttest~~s 1 and 2~~. All series were tested from their earliest recorded date (1850 and 1880) and results from 1880–2014 are shown.

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. 2~~1~~).

**Figure 2~~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 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. 2~~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.

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. 3~~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 $p>0.05$ ~~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~~ is 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 $p>0.05$ ~~non-significant~~ in two records and ~~trends at~~ $p<0.05$ in three records.

**Figure 3: 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.**

### 3.1.1 Step/trend and shift/trend ratios

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 because it gives first preference to gradual change. ~~As some of the internal trends show $p>0.05$, t~~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.

~~**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.**~~

This ~~is~~ also ~~the case~~ holds 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 at regional scales will contribute to the global record being more trend-like than individual regions.

### ~~5.1.~~3.2 Satellite~~.~~ era records

A comparison of surface and lower tropospheric satellite temperatures stratifies records according to altitude and source of measurement, also consistent with Test 2. Satellite records of annual and seasonal lower troposphere ~~temperatures~~ anomalies

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~~ M~~m~~ean annual global and zonal temperatures show 1995 and 1998 as the two main ~~shift~~ step dates, with 1995 more prominent at the global scale (Table 1). Seasonal temperatures were assessed to distinguish between these dates. 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~~to provide more precision on dates of step changes. 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~~for probabilities for false positive (Type I) errors. Student's t-test (two sided, unequal variance), which is insensitive to serial correlation, was used as a back-up.

**Table 1 about here**

For the quarterly results, 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* $p$>0.05~~ 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~~ error probabilities in the satellite records ~~is~~ are 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~~ largely northern hemisphere (1987/88).

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. 4~~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.

Unless substantially contaminated by artefacts, these changes do not ~~reflect~~ represent gradual warming in the atmosphere, but ~~instead may reflect~~ may represent 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~~ the global scale ~~events~~ suggests that available heat energy is not a limiting factor ~~to~~ for abrupt changes.

**Figure ~~4~~3: Quarterly mean lower tropospheric 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.**

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 where La Niña events precede shifts and El Niño events accompany them. If they are not assumed to be a 'contaminating influence' of noise affecting the signal, ~~and, given the coincidence of step dates with some El Niño events, t~~there is no clear way to allow for the~~m~~s~~e~~, so the data is analysed and presented as is. As we discuss later in the paper, it appears that El Niño has an active role in step-like warming.

### ~~5.2.~~3.3 Regional attribution

This section on regional attribution covers the issue of stationarity and the character of change over regional areas and addresses Test 4. 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).

**Table 2 about here**

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 ($Tav_{ARW}$).

In SEA, $Tav_{ARW}$ shifts up by 0.5 °C in 1973 (Fig. 54). 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.

**Figure 54: 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).**

None of the internal trends in Fig. 54 ~~exceed the~~ achieve $p$<0.05 ~~threshold~~. The trend/shift ratios for *Tav* (not shown in Fig. 54) 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 20$^{th}$ 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., 2016).

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).

### 3.4 Other climate variables

If climate changes in a step-wise manner, it would be expected that other variables would show signs of this (Test 5). Instances of step changes in the literature are widespread, and are mentioned elsewhere in this paper (e.g., Table 6). For rainfall, notable examples are a step change in the Sahel in 1970 (L'Hôte et al., 2002;Mahé and Paturel, 2009), south-west Western Australia in the late 1960s/early 1970s (Li et al., 2005;Power et al., 2005;Hope et al., 2010) and the western US in 1930s (Narisma et al., 2007). Similar changes have been detected in streamflow records worldwide, showing that regime changes in moisture have been a long-standing aspect of climate variability (Whetton et al., 1990). Few more recent changes have been directly attributed to increasing gases, although south-west WA is an exception (Cai and Cowan, 2006;Timbal et al., 2006;Delworth and Zeng, 2014), with large-scale shifts in synoptic types accompanying a rapid decrease in rainfall (Hope et al., 2006). The bivariate test identifies a step change in south-west WA winter rainfall in 1969, shown in Fig. 6a with an upward step in summer rainfall in northern Australia one year later.

Ocean heat content of the upper ocean also shows step changes occurring in 1977, 1996 and 2003 (Fig. 6b). Changes in long-run tide gauge records also show a step-ladder-like process of sea level rise, with the San Francisco record, quality controlled and dating back to 1855, being a good example, showing step changes in 1866, 1935, 1957 and 1982 (Fig. 6c). Step changes in the Fremantle tide gauge data records, one of the longest in southern hemisphere, shows that most of the decline in the average return intervals of extreme events noted by Church et al. (2006) before and after 1950, occurred in two events (Fig. 6d) in the late 1940s and the late 1990s. This variation in rise has been noted by White et al. (2014). None of the internal trends in Fig. 6a–d attain $p<0.05$, showing the dynamic nature of change and limited trend-like behaviour in these examples.

**Figure 6: Records showing internal trends separated by step changes of (a) total rainfall for south-west Western Australia (winter) and northern Australia (summer, 1900–2015); (b) global ocean heat content of the top 700 m (1955–2014); (c) tide gauge data for San Francisco, USA (1855–2015) and (d) Fremantle, Australia (1912–1925, 1927–2015). Step changes ($p<0.01$) identified by the bivariate test.**

## ~~6.~~4 Results – models

### ~~6.1.~~4.1 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~~covering T~~t~~ests 1, ~~1~~2 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. 2~~1~~, are shown in Fig. 7~~5~~a. 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. 7~~5~~b shows step changes from the CMIP3 combined SRES ~~scenarios~~ A1B and A2 simulations ~~{Nakicenovic, 2000 #1323}~~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 7~~5~~c–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. 7~~5~~d) 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.

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.

**Figure 7~~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).**

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.

### ~~6.2.~~4.2 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, addressing Test 3. 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. 8~~6~~a). 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. 8~~6~~b).

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. 8~~6~~c). 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.

**Figure 8~~6~~: Multi-model ensemble (RCP4.5, 107 members) characteristics of hindcast (1861–2005) and projection~~ed~~ (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.**

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. 8~~6~~d). However, correlations between ECS, and total steps, shifts and trends, are 0.81, 0.72 and 0.43, respectively (all $p<0.01$, Fig. 8~~6~~e). This shows that the timeseries are becoming more trend-like at higher rates of forcing, when compared to the hindcast period. 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. 8~~6~~f~~e~~). The high sensitivity models plot towards the upper right and lower sensitivity models to the lower left. The trend/step 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~~ steps 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. 9~~7~~a). 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. 97b).

**Figure 97: Correlations between ECS and linear trends, total step changes, warming-to-date and quadratic trends; (a) from 1861 to the current decade (warming to date: 1861–99 average subtracted from current decadal average) and (b) from 1961. Dotted lines mark** ~~**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~~be a better predictor. This ~~certainly seems to~~ is ~~be~~ the case for climate models, so may apply ~~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) for steps become known, both have roughly equivalent predictive skill at the present time.

### ~~6.3.~~4.3 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 75c–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. 108a).

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.

**Figure 108: 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) $Ti_0$ results from a 40-year moving window for the RCP2.6, 4.5, 6.5 and 8.0 simulations, respectively.**

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~~ becomes increasingly 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~~ its 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. 108b–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 $Ti_0$ values indicate both steplike and trending behaviour. A cloud to the far right, as in Fig. 108e, 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.

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. 119). 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.

**Figure 119: Anomalies of annual mean temperature showing internal trends separated by step changes from the CSIRO Mk3.5 A1B simulation; (a) maximum temperature south-eastern Australia; (b) minimum temperature south-eastern Australia; (c) global mean surface temperature. Internal trends (dashed lines) are separated by step changes ~~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).**

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.

## ~~7.~~5 ~~Severe t~~Testing of steps ~~and~~ versus trends

Earlier sections have identified steps and trends in temperature and tested how trend, step and trend-shift relationships relate to total warming and 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 (ResSS), cumulative ($\sum\sum$R) residuals and cumulative residuals squared ($\sum\sum R^2$). 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. 120 and Tables 4 and 5. The data and statistical models for HadCRU record 1880–2014 are shown in Fig. 120a. Cumulative residuals that track close to zero (Fig. 120b) 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. 120c). 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 $r^2$ of 0.87 and ResSS of 0.8. For $\sum\sum R^2$ the trend model behaves more poorly than the other three.

**Figure 120: 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.**

TWith the autocorrelation and heteroscedasticity tests, the LOWESS test performs less well than the autocorrelation and heteroscedasticity tests other two for the 40-year testswindows. 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 performs 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 $\sum\sum R^2$ 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.

These tests, omitting LOWESS, were carried out for HadCRU 1965–2014, a period with a sustained radiative forcing signal (Fig. 120d). 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 is only slightly better than the trend model – this is due to

the northern hemisphere shift in 1987/88 being incorporated into the global mean ~~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 about here**

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, showing the steplike structure of this record. Here, ~~T~~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 ResSS double that for steps. This record is entirely made up of steps, showing the lack of trend occurring within some regions.

The quarterly record of HadCRU from Fig. 4~~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. 10~~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 ResSS 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. 8~~6~~f. The RCP8.5 simulation is the most trend-like; the step model fails in the final decades of the 21st 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.

**Table 5 about here**

### ~~7.1.~~6 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 or even segmented 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~~ temperature data they seek to explain. The data generated consists of steps, trends and shifts ~~applying~~ calculated using the multi-step MYBT model and ~~and trends applying~~ least squares trend analysis. The use of statistical models such as LOWESS are for sensitivity testing and not part of the probative assessment.

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

**Test 1 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 downward internal trends. Shifts are consequently higher than steps in the satellite data.

**Test 2 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.

**Test 3 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.

**Test 4 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~~ it was 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.

**Test 5 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 (Jones et al., 2013). These are overwhelmingly attributed to random climate variability, including abrupt changes identified as part of decadal regime change.

**Test 6 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 (ResSS), 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 a 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 varying independently.

**Table 6 about here**

## ~~8.~~7 Discussion

There are many reasons as to why *H1* – where climate change and variability are considered to be independent of each other – has dominated climate research despite the lack of a conclusive theoretical or statistical case. They include historical, social, theoretical and political considerations too broad to cover here. The following discussion briefly covers three areas addressing *H2*: theoretical support, potential mechanisms and why these findings matter for climate practice. One important point is that if both climate models and observations demonstrate step-like warming, the appropriate physical relationships are represented in the models and are being misdiagnosed.

The greenhouse effect can be considered as the product of two coupled subsystems within the larger climate system. The radiative subsystem traps longwave radiation reflected from the surface. The hydrodynamic subsystem distributes this trapped heat from the equator to the top of the atmosphere and the poles (Ozawa et al., 2003)~~{Eagly, 2003 #2514}~~. Internally, this subsystem is always out of equilibrium. The presence of oscillatory mechanisms on interannual to decadal timescales (White et al., 2003;Chen et al., 2008) is a major vehicle for this transport, potentially manifesting as climate regimes that exist in steady states punctuated by step-like regime changes. However, when anthropogenic greenhouse gases are added, preventing longwave radiation leaving the atmosphere, the whole climate system moves out of equilibrium. This requires even greater amounts of heat to be transported to the top of the atmosphere and the poles, warming the planet in the process. The $64 million question is whether this occurs independently of ongoing climate variability or enhances it in some way.~~(White et al., 2003;Chen et al., 2008){White, 2003 #5323}{Chen, 2008 #5324}~~

One justification given for favouring trend-like warming is the absence of a plausible mechanism for steplike change (Cahill et al., 2015;Foster and Abraham, 2015). However, hydrodynamic processes 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 (Levitus et al., 2012;Roemmich et al., 2015), whereas the atmosphere contains about 3% of the total. A similar amount of the heat has been stored within the land mass (Balmaseda et al., 2013) and on an annual basis a similar flux is absorbed in melting ice (Hansen et al., 2011). A physical re-organisation of the ocean-atmosphere system, as part of a regime change, is therefore large enough to provide the relatively small amount of energy required to cause abrupt sea surface and atmospheric warming {Roemmich, 2015 #4120}{Reid, 2016 #5142}.

For example, Reid et al. (2016) in describing the late 1980s regime change, show it was associated with large-scale shifts 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 involving both the Pacific and Atlantic Oceans (Chikamoto et al., 2012a;Chikamoto et al., 2012b). Jones (2012) noted two types of regime

change: one where co-dependent variables such as maximum temperature and rainfall undergo a step change but remain in a stationary relationship, and the other, non-stationary change, where warming over land undergoes a step change independent of rainfall change. This suggests that although regime changes are a normal part of internal climate variability, they may be enhanced by some type of heat storage and release mechanism. The dates of step changes summarised in Table 6 coincide with El Niño events but the heat emitted by a normal El Niño is absorbed within months, so an added mechanism is required.

Benestad (2016) reviews models used to build a mental picture of the greenhouse effect, nominating radiative-convective and heat balance models as two types historically used for this purpose. Modern climate models are almost as complex as the climate itself, so need to be understood through simpler models (Held, 2005;Benestad, 2016). ~~(Held, 2005;Benestad, 2016)~~Most, if not all assessments take up the basic assumptions of those simpler models, forming a consensus mental picture that operates as the dominant paradigm of how climate changes. In this case, the main assumption is that a proportion of trapped heat remains in the atmosphere, immediately becoming available for convective and radiative transfer, and the bulk of the remainder is absorbed by the ocean with some heat being taken up by the land and by ice melt. Statistical assessments are also conducted on this basis, so that variations between rates of absorption and deep and shallow ocean mixing, are prominent in discussions of decadal variability in atmospheric warming rates (Meehl et al., 2013;Trenberth and Fasullo, 2013;Watanabe et al., 2013;Drijfhout et al., 2014;Meehl, 2015;Steinman et al., 2015). Release of heat from the ocean seems to be regarded similarly, occurring gradually, with the rate of release being modulated by deeper ocean mixing.

Alternatively, the identification of possible steady-state conditions punctuated by regime changes, with declining internal trends occurring over some ocean regions, the region 30 °N–60 °N, and in tropospheric satellite temperatures, suggest that little or none of heat being trapped in the atmosphere by anthropogenic greenhouse gases actually remains there. Most heat is trapped near the ground/ocean surface and much of that is radiated downwards (Trenberth, 2011). *In situ* warming is hardly plausible if the atmosphere has no intrinsic heat memory. This is supported by observations on land where the overpassing airmass takes on the characteristics of the underlying surface, achieving energy balance within a 300 m distance (Morton, 1983). Very little of the heat trapped over land can be absorbed by the land surface, but given that the atmosphere interacts with the top 70 m of ocean over an annual cycle (Hartmann, 1994), there is ample opportunity for the majority of available heat trapped over land not absorbed by land, lakes and ice to circulate and be absorbed by the ocean.

In terms of energy budgets, the additional forcing from anthropogenic greenhouse gases is roughly 2% of the estimated total annual budget of 155 $Wm^{-2}$ trapped mainly by water vapour and $CO_2$ (Schmidt et al., 2010). As >90% of that 2% is accepted as being absorbed by the ocean, it is not clear why some of the rest would remain in the atmosphere if its absorption by the ocean is not energy limited. The lack of positive internal trends in lower tropospheric satellite temperatures also indicates that the air column is not warming in situ but maintains a fairly stable temperature punctuated by step changes (Fig. 4), so is relatively stable unless heated from below.

The dominant mental model concerning this relationship asks the question "How much of the anthropogenic heat being trapped in the atmosphere is being absorbed by the ocean?", whereas if we accept that the ocean absorbs all the additional heat, this question changes to "How much anthropogenic heat is the ocean discharging?" By following the same pathway as naturally trapped heat, anthropogenic heat will become entrained into the normal processes of climate variability. The two types of predictability discussed by Lorenz (1975) and Hasselmann (2002) therefore need to be re-examined, rather than being separated into signal and noise on the assumption they are independent (e.g., Corti et al., 1999).

The first type of predictability is initial conditions sensitive and boundary limited, so will vary within a certain amplitude. Lorenz (1975) considered this to be transitive, the outcome depending on the pathway taken. The second type depends on changing boundary conditions and is intransitive, with the outcome being insensitive to initial conditions and to the pathway taken. The first can clearly be related to the weather and the second to climate, but Lorenz (1968) suggested this model would be appropriate for days to millions of years, calling the outcome on century timescales almost intransitive. The almost-intransitive model (Lorenz, 1968) has proved to be robust for concepts such as effective radiative forcing (Hansen et al., 2005) and effective climate sensitivity (Andrews et al., 2015) when linear assumptions are applied, although they would be sensitive to bifurcations if they were to occur (Hasselmann, 2002).

On decadal timescales a changing climate would be subject to a combination of free and forced variations (c. f., Lorenz, 1979), affected by spatial and temporal nonlinearity. ~~{Shine, 1999 #5348}{Hansen, 2005 #5349}~~Andrews et al. (2015) investigated the nonlinearity aspect of warming over shorter timescales by examining the early part of an instantaneous quadrupling of $CO_2$ in a coupled model, finding that feedbacks were negative for the first 20 years becoming more positive over time. However, when their model was subject to gradually increased sea surface temperatures, circumventing the greenhouse effect, the early response became linear. This is consistent with our findings that trapped heat is initially stored, only being released after a time delay and atmospheric feedbacks are responding to that release. Spatial analysis of climate feedbacks also shows nonlinearity, with the tropics and subtropics showing positive feedbacks and compensating heat loss, and negative feedbacks with accelerated polar warming, the whole process serving to dampen climate sensitivity (Feldl and Roe, 2013). The coincident timing of step changes in both observations and models (Fig. 7) suggests that other factors, such as short-term volcanic forcing, can also influence the timing of step changes.

So what mechanisms are driving step changes in warming? ~~(Hansen et al., 2011){Hansen, 2011 #5287}(Hartmann, 1994)~~Recently, Peyser et al. (2016) linked dynamic sea level in the Pacific Ocean, measured using an east-west seesaw index, to rapid changes in global mean surface temperature. In 1996/1997, that index underwent a west-to-east seesaw movement of 149 mm. Based on a linear regression between the seesaw index and surface temperature calculated from control runs of 38 CMIP5 climate models, they estimate a jump in surface temperature of 0.29 ± 0.10 °C in 1997/1998, close to our estimate of 0.32 °C or 0.25 °C if 1987/88 is taken into account. Another seesaw change of 111 mm in 2014/15 they estimated as contributing to a rapid warming of 0.21 ± 0.07 °C in 2016. We interpret their observations of rapid sea level rise in the western

Pacific region as representing the sustained storage of heat in the Indo-Pacific warm pool. Heat absorbed in the tropical Pacific is blown westward into the warm pool where it accumulates, maintaining the tropical Pacific as a region of generally low warming (Power et al., 2016). As the warm pool reaches critical limits, it becomes unstable, releasing surplus heat as a tongue of warm water from the west to eastern Pacific during an El Niño event.

Meehl et al. (2016) have also suggested that the negative phase of the Interdecadal Pacific Oscillation that commenced in 1997/98 (Overland et al., 2008;Meehl et al., 2013), could change to positive during 2015–2019 as part of oscillatory mechanisms associated the build-up of heat in the western Pacific. This would provide the accompanying regime change required to sustain higher temperatures after the initial outburst – and is consistent with widespread coral bleaching in 2014–2016 (Normile, 2016) rivalling that of 1998. Note that both Peyser et al. (2016) and Meehl et al. (2013) interpret their results as variability acting on a long-term trend; however, we reinterpret their findings as supporting a heat pulse and regime change, producing steplike warming.

In storing heat for redistribution, the Indo-Pacific warm pool acts a global heat engine (Bosc et al., 2009), a function it has fulfilled for millions of years over a wide range of climatic changes (Gagan et al., 2004;de Garidel-Thoron et al., 2005;Abram et al., 2009). The storage and release mechanism identified by Peyser et al. (2016) may therefore be an additional response to a build-up of heat over and above oscillations associated with ongoing decadal regime change. Storage and release mechanisms may exist in other ocean basins but would need to be identified.

None of these findings challenge the core theory of anthropogenic global warming – rather they offer an explanation for how climate change and variability may be interacting within a fully-coupled climate system. The ocean has a homeostatic relationship with the atmosphere, taking up available heat and maintaining steady state conditions within an oscillatory system of climate regimes. Here homeostasis is a result rather than a postulate (Kleidon, 2004). Heat will accumulate in the shallow ocean until such time as it becomes unstable and is released as part of a step-wise regime change. Sustained forcing would produce a series of regime changes becoming successively warmer, forming a step-ladder – elevator-like record of change. Note that these changes are quite different to those catalogued by Drijfhout et al. (2015) who screened the CMIP5 model ensemble for abrupt shifts that could be considered as singularities, locating 37 ocean, sea ice, snow cover, permafrost, and terrestrial biosphere changes. Their methodology rejects the changes analysed in this paper so constitutes a different type of risk to those surveyed here.

A climate behaving as a coupled system of interacting internal and external forcing, oscillating around stable states and warming through a storage and release process is substantially different to one that is warming gradually via diffusion. Step changes will lead to rapid changes in both mean climate and extremes, leading to nonlinear changes in impacts, compounding with changes in exposure (Jones et al., 2013). As detection and attribution, climate forecasting and characterisation of future

climate risk are almost totally focused on being scaled to gradual changes in mean variables, a step and trend process will require a substantial re-think as to how these activities can be conceptualised.

~~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 look~~like~~ quite different if analysed in *H2* mode ~~(Jones, 2015b)~~. ~~{Lorenz, 1975 #2756@@author-year}{Hasselmann, 2002 #2616@@author-year}~~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 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.~~

- ~~(Leith, 1973, 1975;Lorenz, 1975;Hasselmann, 1990;Palmer, 1999;Hasselmann, 2002)~~

- ~~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 pH0 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 htrend 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 hstep. 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.

**Commented [RJ1]:** Also mention that the hiatus does not exist

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 . 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 1022 Joules, compared with the 24.0 x 1022 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.

**Commented [RJ2]:** Also Roemmich and Reid et al., 2016

For example, Reid et al. (2016) 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).

For example, Reid et al. (2016) 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.~~

Commented [RJ3]: Stronger finish

~~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.~~

## ~~9.~~8 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 *H1* and *H2* with statistical hypotheses $h_{trend}$ and $h_{step}$, ~~in order to test~~ and subjecting them to severe testing. Paraphrasing the severity principle of Mayo (2010) the results of Tests 1–6 provides evidence for hypothesis *H2* if and only if $h_{step}$ passes a severe test with very high probability, where $h_{trend}$ would have uncovered the falsity of *H2*, and yet no such error is detected. Error and probative testing of steps against trends lends little support for the proposition that the climate warms gradually. This is despite trends being given preference when measuring nonlinearity through shifts (steps minus trends). If trend-like behaviour was dominating warming or was on an even footing with steplike change, these tests would have identified it. *H1* is only suitable for intransitive estimates of change, where the initial conditions, pathway, and nonlinear components of forcing are unimportant.

Surface and tropospheric warming on decadal timescales is dominated by stepwise changes in temperature ~~(~~(This paper;Reid and Beaugrand, 2012;Jones et al., 2013;Belolipetsky et al., 2015;Bartsev et al., 2016;Reid et al., 2016). The basic physical mechanism for moving from *H1* to *H2* is deceptively simple: instead of warming occurring *in situ* in the atmosphere and/or being released gradually from the ocean, all available heat from additional greenhouse gases not absorbed by the land surface, snow and ice and in lakes is absorbed by the ocean. There, it is entrained into the nonlinear processes of climate variability, where the added forcing interacts with those processes. The most plausible explanation for steplike behaviour is that steady-state decadal regimes are punctuated by step-like bursts of warming that are subsequently maintained by higher sea surface temperature emplaced by ocean-atmosphere regime changes.

~~*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~~ conclusion 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)Branstator and Selten (2009). We discuss this further in another paper (Jones and Ricketts, 2016).~~ However,

~~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.~~

~~T~~the signal ~~–~~ to ~~–~~ noise model of a gradually changing mean surrounded by random climate variability ~~poorly~~ poorly represent~~s~~ ~~s~~ 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). 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.

Climate conceptualised as a mechanistic system and described using classical statistical methods is substantially different to climate conceptualised as a complex system. With record atmospheric and surface ocean temperatures in 2015–16 variously being described as a singular event, a reinvigoration of trend-like warming or a wholesale shift to a new climate regime, this issue is too important to be left unresolved.

## ~~5~~9 Code availability

With Supplementary Information as a zip file (Python and R modules)

## ~~6~~10 Data availability

With Supplementary Information as Excel files

## ~~7~~11 Team list

Roger N. Jones, Victoria University

James H. Ricketts, Victoria University

## ~~8~~12 Author contributions

RJ conceived the study, JR coded and tested the multi-step model, RJ developed the severe testing regime for the results and with JR undertook analyses, JR put together the SI and maintained quality control, RJ led the paper with contributions from JR.

## ~~9~~13 Competing interests

The authors declare that they have no conflict of interests.

~~10~~14 **Acknowledgements**

JR is the holder of a Victoria University posgraduate research scholarship. Data sources include the Met Office Hadley Centre, National Aeronautics and Space Administration Goddard Institute for Space Studies, National Oceanographic Data Center and United States National Climatic Data Center, Berkely Earth, Cowtan and Way, Permanent Service for Mean Sea Level and the Australian Bureau of Meteorology. CMIP3 and CMIP5 archives are made available by the modeling 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. D Kelly O'Day provided the macro templates, which has been adapted to provide the step and trend charts. Two anaonymous reviews and the editor made valuable comments that heled focus the paper.

## ~~11~~15 References


Abram, N. J., McGregor, H. V., Gagan, M. K., Hantoro, W. S., and Suwargadi, B. W.: Oscillations in the southern extent of the Indo-Pacific Warm Pool during the mid-Holocene, Quaternary Science Reviews, 28, 2794-2803, 2009.

Alfaro, E. J., Gershunov, A., and Cayan, D.: 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, doi:10.1175/JCLI3665.1, 2006.

Andrews, T., Gregory, J. M., and Webb, M. J.: The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models, Journal of Climate, 28, 1630-1648, 2015.

Balmaseda, M. A., Trenberth, K. E., and Källén, E.: Distinctive climate signals in reanalysis of global ocean heat content, Geophysical Research Letters, 40, 1754-1759, 10.1002/grl.50382, 2013.

Bartsev, S. I., Belolipetskii, P. V., Degermendzhi, A. G., Ivanova, Y. D., Pochekutov, A. A., and Saltykov, M. Y.: Refocusing on the dynamics of the Earth's climate, Herald of the Russian Academy of Sciences, 86, 135-142, 10.1134/s1019331616020015, 2016.

Belolipetsky, P. V.: The Shifts Hypothesis - an alternative view of global climate change, arXiv preprint arXiv:1406.5805, arXiv:1406.5805, 2014.

Belolipetsky, P. V., Bartsev, S., Ivanova, Y., and Saltykov, M.: Hidden staircase signal in recent climate dynamic, Asia-Pacific J Atmos Sci, 51, 323-330, 10.1007/s13143-015-0081-6, 2015.

Benestad, R. E.: A mental picture of the greenhouse effect, Theoretical and Applied Climatology, 1-10, 10.1007/s00704-016-1732-y, 2016.

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

Bosc, C., Delcroix, T., and Maes, C.: Barrier layer variability in the western Pacific warm pool from 2000 to 2007, Journal of Geophysical Research: Oceans, 114, C06023, 10.1029/2008JC005187, 2009.

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

Boucharel, J., Dewitte, B., Penhoat, Y., Garel, B., Yeh, S.-W., and Kug, J.-S.: ENSO nonlinearity in a warming climate, Climate Dynamics, 37, 2045-2065, 10.1007/s00382-011-1119-9, 2011.

Branstator, G., and Selten, F.: "Modes of Variability" and Climate Change, Journal of Climate, 22, 2639-2658, doi:10.1175/2008JCLI2517.1, 2009.

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

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

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

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

Cai, W., and Cowan, T.: SAM and regional rainfall in IPCC AR4 models: Can anthropogenic forcing account for southwest Western Australian winter rainfall reduction?, Geophysical Research Letters, 33, n/a-n/a, 10.1029/2006GL028037, 2006.

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

Chen, J., Genio, A. D. D., Carlson, B. E., and Bosilovich, M. G.: The Spatiotemporal Structure of Twentieth-Century Climate Variations in Observations and Reanalyses. Part II: Pacific Pan-Decadal Variability, Journal of Climate, 21, 2634-2650, 10.1175/2007jcli2012.1, 2008.

Chikamoto, Y., Kimoto, M., Ishii, M., Watanabe, M., Nozawa, T., Mochizuki, T., Tatebe, H., Sakamoto, T. T., Komuro, Y., and Shiogama, H.: 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, 2012a.

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

Christy, J. R., Spencer, R. W., Norris, W. B., Braswell, W. D., and Parker, D. E.: Error estimates of version 5.0 of MSU-AMSU bulk atmospheric temperatures, Journal of Atmospheric and Oceanic Technology, 20, 613-629, 2003.

Christy, J. R., Norris, W. B., Spencer, R. W., and Hnilo, J. J.: Tropospheric temperature change since 1979 from tropical radiosonde and satellite measurements, Journal of Geophysical Research: Atmospheres, 112, 10.1029/2005JD006881, 2007.
Church, J. A., Hunter, J. R., McInnes, K. L., and White, N. J.: Sea-level rise around the Australian coastline and the changing frequency of extreme sea-level events, Australian Meteorological Magazine, 55, 253-260, 2006.
Corti, S., Molteni, F., and Palmer, T. N.: Signature of recent climate change in frequencies of natural atmospheric circulation regimes, Nature, 398, 799-802, 1999.
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, 140, 1935-1944, 10.1002/qj.2297, 2014.
de Garidel-Thoron, T., Rosenthal, Y., Bassinot, F., and Beaufort, L.: Stable sea surface temperatures in the western Pacific warm pool over the past 1.75 million years, Nature, 433, 294-298, 2005.
Delworth, T. L., and Zeng, F.: Regional rainfall decline in Australia attributed to anthropogenic greenhouse gases and ozone levels, Nature Geosci, 7, 583-587, 10.1038/ngeo2201
http://www.nature.com/ngeo/journal/v7/n8/abs/ngeo2201.html#supplementary-information, 2014.
Drijfhout, S., Bathiany, S., Beaulieu, C., Brovkin, V., Claussen, M., Huntingford, C., Scheffer, M., Sgubin, G., and Swingedouw, D.: Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models, Proceedings of the National Academy of Sciences, 112, E5777-E5786, 2015.
Drijfhout, S. S., Blaker, A. T., Josey, S. A., Nurser, A. J. G., Sinha, B., and Balmaseda, M. A.: Surface warming hiatus caused by increased heat uptake across multiple ocean basins, Geophysical Research Letters, 41, 7868-7874, 10.1002/2014GL061456, 2014.
Ebbesmeyer, C. C., Cayan, D. R., McLain, D. R., Nichols, F. H., Peterson, D. H., and Redmond, K. T.: 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, 1991.
Feldl, N., and Roe, G. H.: The Nonlinear and Nonlocal Nature of Climate Feedbacks, Journal of Climate, 26, 8289-8304, 10.1175/jcli-d-12-00631.1, 2013.
Fischer, T., Gemmer, M., Liu, L., and Su, B.: Change-points in climate extremes in the Zhujiang River Basin, South China, 1961–2007, Climatic Change, 110, 783-799, 10.1007/s10584-011-0123-8, 2012.
Foster, G., and Rahmstorf, S.: Global temperature evolution 1979-2010, Environmental Research Letters, 6, 044022, 2011.
Foster, G., and Abraham, J.: Lack of evidence for a slowdown in global temperature, US CLIVAR, 13, 6-9, 2015.
Franzke, C.: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature, Journal of Climate, 25, 4172-4183, 2012.
Free, M., and Lanzante, J.: Effect of volcanic eruptions on the vertical temperature profile in radiosonde data and climate models, Journal of Climate, 22, 2925-2939, 2009.
Gagan, M. K., Hendy, E. J., Haberle, S. G., and Hantoro, W. S.: Post-glacial evolution of the Indo-Pacific warm pool and El Nino-Southern Oscillation, Quaternary International, 118, 127-143, 2004.
Ghil, M.: Climate variability: nonlinear and random effects, Encyclopedia of Atmospheric Sciences. Elsevier, 1-6, 2012.
Haig, B. D.: Tests of statistical significance made sound, Educational and Psychological Measurement, 0013164416667981, 10.1177/0013164416667981, 2016.
Hansen, J., Sato, M., Ruedy, R., Nazarenko, L., Lacis, A., Schmidt, G. A., Russell, G., Aleinov, I., Bauer, M., Bauer, S., Bell, N., Cairns, B., Canuto, V., Chandler, M., Cheng, Y., Del Genio, A., Faluvegi, G., Fleming, E., Friend, A., Hall, T., Jackman, C., Kelley, M., Kiang, N., Koch, D., Lean, J., Lerner, J., Lo, K., Menon, S., Miller, R., Minnis, P., Novakov, T., Oinas, V., Perlwitz, J., Perlwitz, J., Rind, D., Romanou, A., Shindell, D., Stone, P., Sun, S., Tausnev, N., Thresher, D., Wielicki, B., Wong, T., Yao, M., and Zhang, S.: Efficacy of climate forcings, Journal of Geophysical Research: Atmospheres, 110, n/a-n/a, 10.1029/2005JD005776, 2005.
Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Reviews of Geophysics, 48, 2010.
Hansen, J., Sato, M., Kharecha, P., and Schuckmann, K. v.: Earth's energy imbalance and implications, Atmospheric Chemistry and Physics, 11, 13421-13449, 2011.
Hare, S. R., and Mantua, N. J.: Empirical evidence for North Pacific regime shifts in 1977 and 1989, Progress In Oceanography, 47, 103-145, 2000.
Hartmann, D. L.: Global Physical Climatology, International Geophysics, Academic Press, San Diego USA and London UK, 411 pp., 1994.

Hasselmann, K.: Is Climate Predictable?, in: The Science of Disasters: Climate Disruptions, Heart Attacks, and Market Crashes, edited by: Bunde, A., Kropp, J., and Schellnhuber, H. J., v. 2, Springer, Berlin Heidelberg, 141-188, 2002.
Hegerl, G., and Zwiers, F.: Use of models in detection and attribution of climate change, Wiley Interdisciplinary Reviews: Climate Change, 2, 570-591, 2011.
Held, I. M.: The Gap between Simulation and Understanding in Climate Modeling, Bulletin of the American Meteorological Society, 86, 1609-1614, 10.1175/BAMS-86-11-1609, 2005.
Hope, P., Timbal, B., and Fawcett, R.: Associations between rainfall variability in the southwest and southeast of Australia and their evolution through time, International Journal of Climatology, 30, 1360-1372, 10.1002/joc.1964, 2010.
Hope, P. K., Drosdowsky, W., and Nicholls, N.: Shifts in the synoptic systems influencing southwest Western Australia, Climate Dynamics, 26, 751-764, 10.1007/s00382-006-0115-y, 2006.
Hoskins, B.: 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, 2013.
Jones, R. N.: Modelling the Hydrologic and Climatic Controls of Closed Lakes, Western Victoria, Ph D, Civil and Environmental Engineering, University of Melbourne, 283 pp., 1995.
Jones, R. N.: North central Victorian climate: past, present and future, Proceedings of the Royal Society of Victoria, 122, 147-160, 2010.
Jones, R. N.: Detecting and attributing nonlinear anthropogenic regional warming in southeastern Australia, Journal of Geophysical Research, 117, D04105, 10.1029/2011jd016328, 2012.
Jones, R. N., Young, C. K., Handmer, J., Keating, A., Mekala, G. D., and Sheehan, P.: Valuing Adaptation under Rapid Change, National Climate Change Adaptation Research Facility, Gold Coast, Australia, 182 pp., 2013.
Karoly, D. J., and Braganza, K.: A new approach to detection of anthropogenic temperature changes in the Australian region, Meteorology and Atmospheric Physics, 89, 57-67, 10.1007/s00703-005-0121-3, 2005.
Kirono, D., and Jones, R.: A bivariate test for detecting inhomogeneities in pan evaporation time series, Australian Meteorological Magazine, 56, 93-103, 2007.
Kirtman, B., Power, S., Adedoyin, A. J., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schär, C., Sutton, R., Oldenborgh, G. J. v., Vecchi, G., and Wang, H.-J.: 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, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge University Press, Cambridge and New York, 121, 2013.
Kleidon, A.: Beyond Gaia: Thermodynamics of Life and Earth System Functioning, Climatic Change, 66, 271-319, 10.1023/B:CLIM.0000044616.34867.ec, 2004.
Koutsoyiannis, D.: HESS Opinions "A random walk on water", Hydrology and Earth System Sciences, 14, 585-601, 10.5194/hess-14-585-2010, 2010.
L'Hôte, Y., Mahé, G. I. L., Somé, B., and Triboulet, J. P.: Analysis of a Sahelian annual rainfall index from 1896 to 2000; the drought continues, Hydrological Sciences Journal, 47, 563-572, 10.1080/02626660209492960, 2002.
Legates, D. R., Soon, W., and Briggs, W. M.: Climate consensus and 'misinformation': A rejoinder to Agnotology, scientific consensus, and the teaching and learning of climate change, Science & Education, 24, 299-318, 2015.
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.
Li, F., Chambers, L., and Nicholls, N.: 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, 2005.
Lo, T. T., and Hsu, H. H.: Change in the dominant decadal patterns and the late 1980s abrupt warming in the extratropical Northern Hemisphere, Atmospheric Science Letters, 11, 210-215, 2010.
Lorenz, E. N.: Climatic determinism, Meteorological Monographs, 8, 1-3, 1968.
Lorenz, E. N.: Climate Predictability, in: The Physical Basis of Climate and Climate Modelling, World Meteorological Organisation, Geneva, 132-136, 1975.
Lorenz, E. N.: Forced and free variations of weather and climate, Journal of the Atmospheric Sciences, 36, 1367-1376, 1979.

Lucarini, V., Fraedrich, K., and Lunkeit, F.: Thermodynamics of climate change: generalized sensitivities, Atmos. Chem. Phys., 10, 9729-9737, 10.5194/acp-10-9729-2010, 2010.
Lucarini, V., and Ragone, F.: Energetics of climate models: net energy balance and meridional enthalpy transport, Reviews of Geophysics, 49, RG1001, 10.1029/2009RG000323, 2011.
Mahé, G., and Paturel, J.-E.: 1896–2006 Sahelian annual rainfall variability and runoff increase of Sahelian Rivers, Comptes Rendus Geoscience, 341, 538-546, http://dx.doi.org/10.1016/j.crte.2009.05.002, 2009.
Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M., and Francis, R. C.: A Pacific interdecadal climate oscillation with impacts on salmon production, Bulletin of the american Meteorological Society, 78, 1069-1079, 1997.
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.
Marvel, K., Schmidt, G. A., Shindell, D., Bonfils, C., LeGrande, A. N., Nazarenko, L., and Tsigaridis, K.: Do responses to different anthropogenic forcings add linearly in climate models?, Environmental Research Letters, 10, 104010, 2015.
Mayo, D. G.: Evidence as passing severe tests: highly probable versus highly probed hypotheses, in: Scientific Evidence: Philosophical Theories and Applications, edited by: Achinstein, P., John Hopkins University Press, Baltimore and London, 95-127, 2005.
Mayo, D. G.: Learning from error, severe testing, and the growth of theoretical knowledge, in: Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science, edited by: Mayo, D. G., and Spanos, A., Cambridge University Press, Cambridge UK and New York USA, 28-57, 2010.
Mayo, D. G., and Cox, D.: 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, edited by: Mayo, D. G., and Spanos, A., Cambridge University Press, Cambridge UK and New York USA, 247-275, 2010.
Mayo, D. G., and Spanos, A.: Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of science, Cambridge University Press, Cambridge UK, 419 pp., 2010.
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, 2009.
Meehl, G. A., Hu, A., and Santer, B. D.: The Mid-1970s Climate Shift in the Pacific and the Relative Roles of Forced versus Inherent Decadal Variability, Journal of Climate, 22, 780-792, w10.1175/2008JCLI2552.1, 2009.
Meehl, G. A., Hu, A., Arblaster, J. M., Fasullo, J., and Trenberth, K. E.: Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation, Journal of Climate, 26, 7298-7310, 10.1175/JCLI-D-12-00548.1, 2013.
Meehl, G. A.: Decadal climate variability and the early-2000s hiatus, US CLIVAR, 13, 1-6, 2015.
Meehl, G. A., Hu, A., and Teng, H.: Initialized decadal prediction for transition to positive phase of the Interdecadal Pacific Oscillation, Nature Communications, 7, 11718, 10.1038/ncomms11718, 2016.
Menberg, K., Blum, P., Kurylyk, B. L., and Bayer, P.: Observed groundwater temperature response to recent climate change, Hydrology and Earth System Sciences, 18, 4453-4466, 10.5194/hess-18-4453-2014, 2014.
Miller, A. J., Cayan, D. R., Barnett, T. P., Graham, N. E., and Oberhuber, J. M.: The 1976–77 climate shift of the Pacific Ocean, Oceanography, 7, 21-26, 1994.
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, D08101, 10.1029/2011JD017187, 2012.
Morton, F. I.: Operational estimates of areal evapotranspiration and their significance to the science and practice of hydrology, Journal of Hydrology, 66, 1-76, 1983.
Narisma, G. T., Foley, J. A., Licker, R., and Ramankutty, N.: Abrupt changes in rainfall during the twentieth century, Geophysical Research Letters, 34, n/a-n/a, 10.1029/2006GL028628, 2007.
Nicholls, N., Dellamarta, P., and Collins, D.: 20th century changes in temperature and rainfall in New South Wales, Australian Meteorological Magazine, 53, 263-268, 2004.
Normile, D.: El Niño's warmth devastating reefs worldwide, Science, 352, 15-16, 10.1126/science.352.6281.15, 2016.
North, G. R., Kim, K.-Y., Shen, S. S. P., and Hardin, J. W.: Detection of Forced Climate Signals. Part 1: Filter Theory, Journal of Climate, 8, 401-408, 10.1175/1520-0442(1995)008<0401:DOFCSP>2.0.CO;2, 1995.

North, R. P., Livingstone, D. M., Hari, R. E., Köster, O., Niederhauser, P., and Kipfer, R.: The physical impact of the late 1980s climate regime shift on Swiss rivers and lakes, Inland Waters, 3, 341-350, 2013.
Overland, J., Rodionov, S., Minobe, S., and Bond, N.: North Pacific regime shifts: Definitions, issues and recent transitions, Progress In Oceanography, 77, 92-102, 2008.
Ozawa, H., Ohmura, A., Lorenz, R. D., and Pujol, T.: The second law of thermodynamics and the global climate system: A review of the maximum entropy production principle, Reviews of Geophysics, 41, 1018, 10.1029/2002RG000113, 2003.
Palmer, T. N.: A nonlinear dynamical perspective on climate change, Weather, 48, 314-326, 10.1002/j.1477-8696.1993.tb05802.x, 1993.
Palmer, T. N.: A nonlinear dynamical perspective on climate prediction, Journal of Climate, 12, 575-591, 1999.
Palmer, T. N., Doblas-Reyes, F. J., Weisheimer, A., and Rodwell, M. J.: Toward Seamless Prediction: Calibration of Climate Change Projections Using Seasonal Forecasts, Bulletin of the American Meteorological Society, 89, 459-470, 10.1175/bams-89-4-459, 2008.
Peterson, T. C., and Vose, R. S.: An overview of the Global Historical Climatology Network temperature database., Bulletin of the American Meteorological Society, 78, 2837-2849, 1997.
Peyser, C. E., Yin, J., Landerer, F. W., and Cole, J. E.: Pacific sea level rise patterns and global surface temperature variability, Geophysical Research Letters, n/a-n/a, 10.1002/2016GL069401, 2016.
Potter, K.: Illustration of a new test for detecting a shift in mean in precipitation series, Monthly Weather Review, 109, 2040-2045, 1981.
Power, S., Tseitkin, F., Torok, S., Lavery, B., and McAvaney, B.: Australian temperature, Australian rainfall, and the Southern Oscillation, 1910-1996: coherent variability and recent changes, Australian Meteorological Magazine, 47, 85-101, 1998.
Power, S., Sadler, B., and Nicholls, N.: The influence of climate science on water management in Western Australia - Lessons for climate scientists, Bulletin of the American Meteorological Society, 86, 839-+, 10.1175/bams-86-6-839, 2005.
Power, S., Delage, F., Wang, G., Smith, I., and Kociuba, G.: Apparent limitations in the ability of CMIP5 climate models to simulate recent multi-decadal change in surface temperature: implications for global temperature projections, Climate Dynamics, 1-17, 10.1007/s00382-016-3326-x, 2016.
Ramaswamy, V., Boucher, O., Haigh, J., Hauglustaine, D., Haywood, J., Myhre, G., Nakajima, T., Shi, G. Y., and Solomon, S.: Radiative forcing of climate change, in: Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., Linden, P. J. v. d., Dai, X., Maskell, K., and Johnson, C. A., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 349-416, 2001.
Reeves, J., Chen, J., Wang, X. L., Lund, R., and Lu, Q. Q.: A Review and Comparison of Changepoint Detection Techniques for Climate Data, Journal of Applied Meteorology and Climatology, 46, 900-915, 10.1175/JAM2493.1, 2007.
Reid, P. C., and Beaugrand, G.: Global synchrony of an accelerating rise in sea surface temperature, Journal of the Marine Biological Association of the United Kingdom, 92, 1435-1450, doi:10.1017/S0025315412000549, 2012.
Reid, P. C., Hari, R. E., Beaugrand, G., Livingstone, D. M., Marty, C., Straile, D., Barichivich, J., Goberville, E., Adrian, R., and Aono, Y.: Global impacts of the 1980s regime shift, Global Change Biology, 22, 703, 10.1111/gcb.13106, 2016.
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.
Rodionov, S. N.: 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, 14-16 June 2005, 2005.
Roemmich, D., Church, J., Gilson, J., Monselesan, D., Sutton, P., and Wijffels, S.: Unabated planetary warming and its ocean structure since 2006, Nature Climate Change, 5, 240-245, 10.1038/nclimate2513 http://www.nature.com/nclimate/journal/v5/n3/abs/nclimate2513.html#supplementary-information, 2015.
Rohde, R., Muller, R. A., 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, Geoinformatics & Geostatistics: An Overview, 1, 1000101, 10.4172/2327-4581.1000101, 2012.
Sahin, S., and Cigizoglu, H. K.: Homogeneity analysis of Turkish meteorological data set, Hydrological Processes, 24, 981-992, 2010.

Santer, B. D., Mears, C., Doutriaux, C., Caldwell, P., Gleckler, P. J., Wigley, T. M. L., Solomon, S., Gillett, N. P., Ivanova, D., Karl, T. R., Lanzante, J. R., Meehl, G. A., Stott, P. A., Taylor, K. E., Thorne, P. W., Wehner, M. F., and Wentz, F. J.: Separating signal and noise in atmospheric temperature changes: The importance of timescale, Journal of Geophysical Research, 116, D22105, 10.1029/2011jd016263, 2011.

Schmidt, G. A., Ruedy, R. A., Miller, R. L., and Lacis, A. A.: Attribution of the present-day total greenhouse effect, Journal of Geophysical Research: Atmospheres, 115, n/a-n/a, 10.1029/2010JD014287, 2010.

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer, S. E., Bhat, M. K., and Bleck, R.: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive, Journal of Advances in Modeling Earth Systems, 6, 141-184, 2014a.

Schmidt, G. A., Shindell, D. T., and Tsigaridis, K.: Reconciling warming trends, Nature Geoscience, 7, 158-160, 10.1038/ngeo2105, 2014b.

Seidel, D. J., and Lanzante, J. R.: An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes, Journal of Geophysical Research, 109, D14108, 10.1029/2003jd004414, 2004.

The Escalator: http://www.skepticalscience.com/graphics.php?g=47, access: February 23, 2016, 2015.

Solomon, A., Goddard, L., Kumar, A., Carton, J., Deser, C., Fukumori, I., Greene, A. M., Hegerl, G., Kirtman, B., Kushnir, Y., Newman, M., Smith, D., Vimont, D., Delworth, T., Meehl, G. A., and Stockdale, T.: Distinguishing the Roles of Natural and Anthropogenically Forced Decadal Climate Variability, Bulletin of the American Meteorological Society, 92, 141-156, 10.1175/2010bams2962.1, 2011.

Steinman, B. A., Mann, M. E., and Miller, S. K.: Atlantic and Pacific multidecadal oscillations and Northern Hemisphere temperatures, Science, 347, 988-991, 10.1126/science.1257856, 2015.

Suppes, P.: Models of data, in: Studies in the Methodology and Foundations of Science, edited by: Nagel, E., Suppes, P., and Tarski, A., Stanford University Press, Stanford CA, 252-261, 1962.

Timbal, B., Arblaster, J. M., and Power, S.: Attribution of the late-twentieth-century rainfall decline in southwest Australia, Journal of Climate, 19, 2046-2062, 2006.

Timbal, B., Arblaster, J., Braganza, K., Fernandez, E., Hendon, H., Murphy, B., Raupach, M., Rakich, C., Smith, I., Whan, K., and Wheeler, M.: Understanding the anthropogenic nature of the observed rainfall decline across south-eastern Australia, The Centre for Australian Weather and Climate Research, Melbourne, 180, 2010.

Trenberth, K. E.: Changes in precipitation with climate change, Climate Research, 47, 123, 2011.

Trenberth, K. E., and Fasullo, J. T.: An apparent hiatus in global warming?, Earth's Future, 1, 19-32, 2013.

Trenberth, K. E.: Has there been a hiatus?, Science, 349, 691-692, 10.1126/science.aac9225, 2015.

Varotsos, C. A., Franzke, C. L., Efstathiou, M. N., and Degermendzhi, A. G.: Evidence for two abrupt warming events of SST in the last century, Theoretical and Applied Climatology, 116, 51-60, 2014.

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

Wack, P.: The Gentle Art of Reperceiving - Scenarios: Shooting the Rapids (part 2 of a two-part article), Harvard Business Review 2-14, 1985a.

Wack, P.: The Gentle Art of Reperceiving - Scenarios: Uncharted Waters Ahead (part 1 of a two-part article), Harvard Business Review 73-89, 1985b.

Watanabe, M., Kamae, Y., Yoshimori, M., Oka, A., Sato, M., Ishii, M., Mochizuki, T., and Kimoto, M.: Strengthening of ocean heat uptake efficiency associated with the recent climate hiatus, Geophysical Research Letters, 40, 3175-3179, 2013.

Werner, R., Valev, D., Danov, D., Guineva, V., and Kirillov, A.: Analysis of global and hemispheric temperature records and prognosis, Advances in Space Research, 55, 2961-2973, http://dx.doi.org/10.1016/j.asr.2015.03.005, 2015.

Whetton, P., Adamson, D., and Williams, M.: Rainfall and river flow variability in Africa, Australia and East Asia linked to El Niño-Southern Oscillation events, Geological Society of Australia Symposium Proceedings, 1990, 71-82,

White, H.: A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica: Journal of the Econometric Society, 48, 817-838, 1980.

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.

White, W. B., Tourre, Y. M., Barlow, M., and Dettinger, M.: A delayed action oscillator shared by biennial, interannual, and decadal signals in the Pacific Basin, Journal of Geophysical Research: Oceans, 108, n/a-n/a, 10.1029/2002JC001490, 2003.
Wolter, K., and Timlin, M. S.: El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext), International Journal of Climatology, 31, 1074-1087, 10.1002/joc.2336, 2011.
World Meteorological Organization: Position Paper on Global Framework for Climate Services, World Meteorological Organisation, Geneva, 2010.
World Meteorological Organization: Climate Knowledge for Action: A Global Framework for Climate Services - Empowering the Most Vulnerable, World Meteorological Organization, Geneva, 247, 2011.
Wu, Z., Huang, N., Wallace, J., Smoliak, B., and Chen, X.: On the time-varying trend in global-mean surface temperature, Climate Dynamics, 37, 759-773, 10.1007/s00382-011-1128-8, 2011.
Yao, S.-L., Huang, G., Wu, R.-G., and Qu, X.: The global warming hiatus—a natural product of interactions of a secular warming trend and a multi-decadal oscillation, Theoretical and Applied Climatology, 123, 349-360, 10.1007/s00704-014-1358-x, 2016.

**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 | - | - | - | - |

**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 | | | | | |

**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 | ** |

**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% |

**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% |

**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)<br>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<br>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<br>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., 2016) |
| | 1979<br>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<br>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),<br>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% autoc, 0% heterosc. 40y window) | Trends serially independent data, variations due to independent processes | Steps perform better than trends to explain autocorrelation and heteroscedasticity | None located |
| 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 |