22nd International Congress on Modelling and Simulation, Hobart, Tasmania, Australia, 3 to 8 December 2017
mssanz.org.au/modsim2017

# Characterizing change-points in climate series with a severe approach

**J.H. Ricketts** and **R.N Jones**

*Victoria Institute of Strategic Economic Studies (VISES), Victoria University, 300 Flinders St, Melbourne*

*Email: james.ricketts@live.vu.edu.au*

**Abstract:** An increasing number of papers have been published which analyse regime shifts within climate, in part or whole, seeking an embedded signal composed of abrupt changes, usually within the temperature record. Simultaneously a separate, related series of papers has addressed the reality of the so-called hiatus circa 1996/7, explaining it variously as an artifact of correctable data deficiencies, boreal cooling, or a statistical mis-identification. Previously, abrupt changes in 1976 and 1986 have been shown to be significant regimes shifts in the whole Earth system. We recently published a paper (Jones & Ricketts, 2017) henceforth JH2017, demonstrating the existence of abrupt level changes in global and zonal temperature records, observed and modelled, arguing that a substantial portion of the progression of temperature records can be validly attributed to regime state changes which vary in their regionality and occur on decadal time-scales.

Very few papers employ a severe testing approach, such as proposed by Deborah G. Mayo and Spanos (2006) to provide more nuanced information about such changes, leaving open questions about interpretation of findings. We have developed a method which extends the Maronna-Yohai (MY) Bivariate Test for inhomegeneities in measurement series, to detection of multiple step changes (the MSBV) in temperature series and reported it at MODSIM2015, and used it as a basis in JR2017. To support and further this work we have also developed a suite of tests which allow us to more severely examine inferences about the nature of the processes at the time of change, and so far to support our view that most abrupt change in global records is the result of rapid regional state changes.

Inference of the existence of abrupt shifts embedded within a complex times series requires detection methods sensitive to level changes and tolerant of simultaneous changes of trend, variance, autocorrelation, and red-drift, given that many of these parameters may shift together. In our work, the timing of events is key. Detection of abrupt level change precludes any form of low-pass filtering, hence we prefer to err on the side of false positives, using a simple detection method and re-assessing possible shift-points using methods grounded against a variety of null hypotheses. The so-called multiple testing problem is primarily a problem of multiply testing the same null against varied alternatives (an “accept if any” approach); whereas our approach tests separate aspects of the data to strengthen inference (an “accept if all” approach). In using the MY test (assuming stationarity) we trade precision in timing against uncertainty in level change, and elect to re-assess the significance against a disjoint segmented model using ANCOVA (in this application equivalent to a Chow test). We utilise three econometric tests to test for data which may show unit root-like behaviour (loosely, for a time-series, progression independent of time). These are the KPSS test for either trend stationarity or level stationarity against an alternative of unit root; the ADF (augmented Dickey-Fuller test) testing a null hypothesis of unit root against an alternative of stationarity after compensation for auto-correlation and trend; and the Zivot-Andrews (ZA) test which tests for a unit root with drift against stationarity plus a single change. We include a test under development (the disputed residuals t-test or DRT), that tests the residuals of a disjoint segmented model against the residuals of the non-disjoint model with the same trends, since it has been claimed that these cases cannot be sufficiently resolved. We use the studentized Breusch-Pagan test to assess the impact of the derived multiple change-point models on heteroskedacidicty since homoskedacidicity of residuals is expected.

These tests are applied to MSBV analyses of synthetic data then, global and zonal observed and modelled annual mean temperature records. Major results reported are that (1) determination of the timing of a trend change is much more precise than of a change of trend for numerical reasons, (2) global and zonal records appear to have the statistics of composites of a limited number of local to regional quasi-oscillatory processes, potentially interacting with forced warming, and (3) where unit root or red-noise behaviour influence on change-points is reported, it is likely to be an artefact of composition of the signal since spatial segmentation of zones markedly reduces the relevant indicators, (4) land based analyses show much less redness and unit root behaviour than ocean records in the same zones, supporting a view of rapid regime change over land following sustained ocean changes.

***Keywords:*** *Change point, unit root, ADF, Zivot-Andrews, KPSS, Maronna-Yohai*

## 1. INTRODUCTION

We recently published a paper (Jones & Ricketts, 2017), henceforth JR2017, analyzing occurrences of step-like changes in global and zonal temperature records, observed and modelled, arguing that most of the warming in observed temperature since 1950 can be validly attributed to external forcing interacting with internally generated climate variability. These interactions produce stepladder-like warming over decadal time-scales. It presents a competing view ($H_{step}$) to the standard, signal/noise position ($H_{trend}$), that the warming response to external forcing is gradual and that any non-gradual changes are due to independently occurring internal climate variability. This latter position is by far the dominant view – the recent controversy on the nature of global mean surface warming 1998–2014, both pro and con, was argued on these terms. For example, whether or not reduced warming over that period was consistent with theory or represented a non-standard signal or lack of a signal, which challenged standard theory. We argued that both positions, whether internal and external processes are independent or interact, have theoretical support with the former preferred for cognitive reasons such as Occam's razor. But the standard signal-to-noise model has never passed a severe test, given that other statistical hypotheses perform similarly well and have never been fully ruled out. Severe testing is based on the intuition that "*Data $x_0$ in test T provide good evidence for inferring H (just) to the extent that H passes severely with $x_0$, i.e., to the extent that H would (very probably) not have survived the test so well were H false.*" (Deborah G. Mayo & Spanos, 2006). They propose that a severity criterion supplies a meta-statistical principle for evaluating statistical inferences (their page 328), where the *severity* of testing is not assigned to hypothesis *H,* but to the testing procedure.

Inferences about the physical world turn on observed data about that world interpreted with the aid of statistical models. Generally it is presumed that consideration of a physical model has led to an adequate inferential statistical model. Misspecification testing (M-S) was proposed as an approach to determining whether the assumptions needed to reliably model the statistical variables are met (Deborah G Mayo & Spanos, 2004). The authors differentiate between model specification and model selection. An adequate model specification licenses primary statistical inference, and with it statistical model selection from the specified family. Serial feature detection in any time series is a form of model selection from a family of related models, reliant on model-specification. It must be noted that a series of tests are performed, a single detection test and multiple probative tests, but that as each is against an independent null, and increases the overall power of the testing regime, this does not involve a multiple-testing issue (ibid.).

A physical model has specified relationships and associated statistics, which together allow statistical inferences that license physical inference. But statistical tests are framed against often implicit statistical assumptions rather than physical ones. The implicit assumptions must be considered, together with the linkage between the physical process and statistical models. Where competing physical models cannot be correctly distinguished by the tests given specific data, the statistical models or the model selection processes are mis-specified.

In JR2017, we tested alternative hypotheses that warming was step-like and trend-like using severe testing. The testing was probative rather than probabilistic, so rather than depending on *p* values from statistical tests, each of six test clearly laid out theoretical and/or physically distinct alternatives tied to the test outcomes. The analysis showed that $H_{step}$ clearly passed all six tests in preference to $H_{trend}$. We argued that the physical processes underpinning these two alternatives were very different. $H_{trend}$ relies on in situ atmospheric warming with gradual exchange with the ocean, potentially mediated by decadal variability. $H_{step}$ maintains there is little or no direct atmospheric warming and that all available added heat not taken up by heat sinks such as land or snow and ice melt is absorbed by the ocean, as is the case for all such heat trapped by natural greenhouse gases. This heat is released in a storage and release process associated with decadal regime change. Under this hypothesis, climate change is not an independent process to climate variability but is essentially enhanced climate variability, producing nonlinear responses on decadal timescales. Our application of statistics is the detection and analyses of these responses, ultimately for the better characterization of climate risk.

For the purposes of this paper we propose two process models and two matched statistical models (1) Our H1 process hypothesis of non-interaction requires that heat due to warming follow an essentially diffusive pathway, or be partitioned from interacting components within the whole Earth system until it is transported back to space. (2) Our H2 process model: Complex feedback – where multiple self-organised entities, themselves capable of forming higher levels structures, have distinct discernable states and regime changes. (3) Our H1 statistical model: No support for discontinuous piece-wise regressions; or where piece-wise regressions are accepted, no monotonic trends in either extent of timings of those detected. (4) Our H2 statistical model: Piecewise regressions representing periods of stationarity and reflecting regime like stability, separated by change-points representing regime changes; distinct spatial organisation.

## 2. CHANGE POINTS AND PARAMETER INTERACTION

The assumptions of the detection method used must be considered. Our method is less usual since it detects shifts, and it is easier to be precise about timing of level changes than trend changes. The removal of a trend-change is the simplest high-pass filter, and detection requires a high-stop or low-pass filter. A step change however is not easily framed this way because a step-change and high frequency noise overlap in the Fourier basis so that no filter can be specified to detect one and remove the other (Little & Jones, 2011). Further, for a change of trend to be detected with the same precision in time as a change of level, the standard deviation of the residual of the time series about the statistical model has to be half of that required to detect a shift to the same precision (Ricketts, 2017). For this reason we use a method biased towards level-detection.

## 3. PROBATIVE TESTS

### 3.1. Testing for adequacy of model.

A form of misspecification is the unjustified restriction to families of models with reduced degrees of freedom. If the "True" family is derived from a family of disjoint linear segments, and the selected family is either our disjoint step-wise model ("constant trend"), or a non-disjoint segmented ("broken-stick") model, then the free parameters will be modified to minimize the error induced by exclusion of the required parameter, sometimes referred to as "cannibalization". The misspecification can be tested for; statistically by examining residual structure; physically by deducing an improbable consequential structure. For reasons alluded to above, the misspecification of a broken-stick model has more impact on timing of change-points than the misspecification of a step-change in our domain.

If a statistical model is not well specified for the data then the residual of the model fit may show signs of systematic variation, detectable by the studentized Breusch-Pagan test (Breusch & Pagan, 1979). This might be taken as a signal of a model misspecification. Additionally for our purposes, a dataset itself can be characterized by for example specifying a deterministic model (e.g linear or quadratic).

### 3.2. Testing segments containing individual change-points.

Analysis of covariance (ANCOVA) assesses the impact of the assignment of specific change-point relative to no change, for a single change, essentially replicating the Chow test. We zero the time parameter on the change-point, and we implement it via a pair of ANOVA tests for a change of intercept and a change of trend. Steps and trend-changes can either jointly support the notion of persistent state change if they are of the same sign or, if of different sign, they can be at least partial evidence of transient impulsive change. Thus ANCOVA also partially probes a specific model misspecification – the omission of transient change.

Unit root tests probe the data for features that can imitate deterministic structural changes due to stochastic behaviour of an unrecognised red or near red progression. However the each of tests can be deceived by other data compositions. In particular time-wise averaging of spatio-temporally lagged deterministic data may appear to particular tests to contain a unit-root indicative of un-time-correlated progression, especially of it is composed of multiple complex processes.

### 3.3. Disputed Residuals t-test.

The disputed residuals test maneuver is defeated by the misspecification of a non-disjoint model for detection, but if a disjoint model is used, attempts to aggressively probe the possibility of a coincident trend and step change of the same sign mis-locating the time of change.

It is further documented in (Ricketts, 2017). It is a two tailed t-test that tests the proposition that a change-point separating two regression lines and including a step is a mis-identification of a trend-only change commencing on the intersection of the two regression lines.

### 3.4. Unit root

Unit root behaviour is well described in the econometric literature. Stock (1994) identifies two categories of unit roots, moving average (MA) and autoregressive (AR) and identifies them as integrated lag zero or I(0), and integrated lag one or I(1) respectively. These can lead to the identification of a chance period of rapid unidirectional random change as a step change in the mean, and a drift like behaviour as a change in trend.

Here we use three methods sourced from the econometric literature to assess the evidence for various manifestations of unit root behaviour. If an interval shows evidence of unit-root behaviour it is possible that the detection methods or our subsequent analysis have been misled. On the other hand, it is sensible to assess these post-hoc, as it is always possible that the redness is transitory, and other shifts in the same series remain valid.

**Table 1.** Unit root tests used and their main assumptions.

| Test | Null hypothesis | Contrast hypothesis | Not resolved |
|---|---|---|---|
| **ADF (trend and drift)** | Unit Root after allowing for autocorrelation and trend. | Trend stationary with inferred lack of UR after allowing for auto-correlation. | Possible exogenous change. |
| **KPSS (Level/Trend)** | Level/Trend Stationarity | Non-stationary with a presumption of I(0)/I(1) Unit Root | Possible exogenous change.<br>Autocorrelation |
| **ZA** | I(1) UR with drift. | Exogenous change at a date (any of a trend change, shift or transient change) | Meaning of date if $H_0$ not rejected is not defined.<br>> 1 structural break |

The Augmented Dickey Fuller test (ADF ), tests a null hypothesis of unit root against an alternative of stationarity after compensation for trend and auto-correlation  (Dickey & Fuller, 1981). It has been used to assess stationarity in climate series, in remote sensing of vegetation indices, and a study of Granger causality and comparative hemispheric response.

The KPSS test (Kwiatkowski, Phillips, Schmidt, & Shin, 1992) is a test for either trend stationarity or level stationarity against an alternative of unit root. The sense of the test is inverted compared to the ADF.

The Zivot-Andrews test (ZA) (Zivot & Andrews, 2002) tests for the presence of unit root against an alternative of exogenous change, assuming exactly one structural break, also giving a time of break. It has been shown to be susceptible to the rejection of the null of a unit root in the presence of a structural break. However we are using complementary tests and are interested in both step and trends, so comparing the estimate of time is also of interest.

The KPSS and the ADF tests do not consider exogenous changes in their null or alternate hypotheses. Consequently the presence of such in the data would constitute a misspecification which must be accounted for in interpretation. Here, the tests may repeated on the residual after removal of the shift and segment trends (which constitute the statistical model for our physical model). This effectively allows the stationarity tests to test the proposition that the data are stationary after removal of trend and one exogenous change. Because the ZA test already allows for a single exogenous change it allows this test to detect a second exogenous change. A ZA test that favours a unit root with drift prior to the removal of a specified exogenous change but not after may indicate a degree of interaction, perhaps transient behavior. This is not implausible for climate processes (see Table 3).

## 4. SYNTHETIC DATA EXPERIMENTS

Specific data was used to calibrate each of the tests as now described. Two methods for constructing synthetic tests data were used. The first data set (DS1) is complex and provides combinations of data with and without UR and with and without step changes and with zero to 3 sub-detectable steps added. The second (DS2) tests the effect of accelerating trends in combination with step changes of varying size, on average about 50% of these being below the nominal detection threshold. There is no UR behaviour in this set.

Test data for DS1 was constructed using a flat time series and with zero or one step, of either one or two std dev, and zero to three randomly located steps of +/- 0.3 std dev were also added to simulate transience or sub-detectable shifts. There was no autocorrelation. A second similar data set was also produced in which after a random location the flat random data was replaced by its cumulative sum, to simulate onset of UR behaviour. In total 48 separate formulations were produced and 100 data sets derived from each of these were tested. It must be noted that the presence of variable onsets of periods of integrated white noise makes a data set in which it is expected to be difficult to separate those effects from step and trend changes. The ZA test proved good for distinguishing sets with either a unit root or a step change, but once a step change and a transient UR were combined it gave essentially random results as expected from its design. The ADF tended to give elevated levels of false negatives for UR if a deterministic step was also present (as follows from its specification). Testing for stationarity with KPSS shows that a step change may be mistaken for a UR.

DS2 is constructed from curvilinear trend and assorted steps, many below the nominal thresholds of detectability and included data expected to be deceptive to the MSBV (Table 2). The results were subjected to a complex analysis which led us to conclude that the DRT remains useful within its design. Where the MSBV misidentified steps due to the presence of multiple sub-detectable events the ANCOVA test classified most of these as not significant and passed as significant at p(<0.5) 95% of the correct determinations. The ADF test tended to favour UR over trend stationarity once the underlying trend increased, and the ZA identified less than 1% of changes to be due to UR. The studentized Breusch-Pagan test was run with the derived steps models and a linear and a quadratic model and supported the step model in all cases.

**Table 2.** DS2. Synthetic Data Timing and extent of Shifts. Total Rise is shown both as anomaly and as standard deviations. Shifts of < 0.5 (are not guaranteed to be found by MEBV.

| A | | B | | C | | D | | E | |
|---|---|---|---|---|---|---|---|---|---|
| Year | Shifts | Year | Shifts | Year | Shifts | Year | Shifts | Year | Shifts |
| 1955 | 0.57 | 1974 | 0.96 | 1970 | 0.80 | 1951 | **0.42** | 1955 | **0.40** |
| 1983 | **0.34** | 1975 | 0.97 | 1987 | 0.80 | 1980 | 0.83 | 1981 | 0.59 |
| 1999 | 0.72 | 2010 | **0.46** | 1996 | 0.68 | 2010 | 0.99 | 2001 | 0.80 |
| 2030 | 0.85 | 2027 | 0.79 | 2028 | **0.41** | 2018 | 0.77 | 2029 | **0.38** |
| 2036 | 1.00 | 2032 | **0.43** | 2036 | 0.57 | 2047 | 0.60 | 2039 | 0.59 |
| 2055 | 0.61 | 2055 | 1.00 | 2050 | 0.94 | 2058 | 0.60 | 2057 | 0.84 |
| 2071 | 0.94 | 2074 | 0.93 | 2076 | 0.54 | 2068 | 0.87 | 2071 | **0.40** |
| 2097 | **0.31** | 2085 | **0.39** | 2095 | 0.54 | 2085 | **0.42** | 2080 | **0.42** |

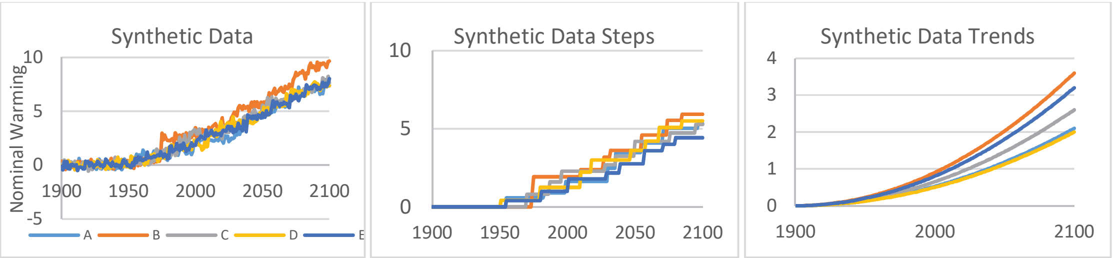

**Figure 1.** Construction of DS2, curvilinear tends and shifts (See Table 2).

## 5. CASE STUDY

We briefly illustrate the heterogeneity of the surface record. GISS two degree gridded, 1200Km smoothed gridded monthly combined land and ocean temperatures were downloaded from KNMI at https://climexp.knmi.nl/ 6 March 2017. This was regridded to 5 degrees as part of a larger study, and for this work area weighted zonal averages at the same latitudes as for GISSTEM3 reported in JR2017 were produced. Note that this is differently treated and later than in JR2017. The southern mid-latitude zone, 60S.30S land/ocean data, and eight 45 degree sectors, selected to not overlap Drake's passage, were analysed by MSBV and the above test suite. All of the change-points in the 60S.30S zone were significant by ANCOVA (p<0.05). See Tables 3, and Figure 2.

Although not further shown here, the land shifts in all observed zones, and mostly in climate models tend to show little residual UR behavior whereas ocean zones and also southern mid-latitude combine land/ocean show that breaks probably aggregate sub-zonally. The analysis in Table 3 shows the differences between zonal and sector analyses of the annual weighted mean temperatures of the oceans in the zone from 60S to 30S. The segments of data within which change-points were detected by the MSBV test were tested for stationarity and exogenous change against alternative explanations of unit-root type drift, and significance of the change-point as a disjoint change in linear progression was checked by ANCOVA.

The DRT was performed to probe the possibility of a misspecification of a trend change against a level change by the MSBV. The unit root tests were repeated against the residual of the disjoint linear model. The studentized Breusch-Pagan test detected no unexplained heteroskedacidicty, for the zonal, and for six of eight sectors, with the break-models. There is a strong contrast between the results for the zonal mean on one hand and most of the

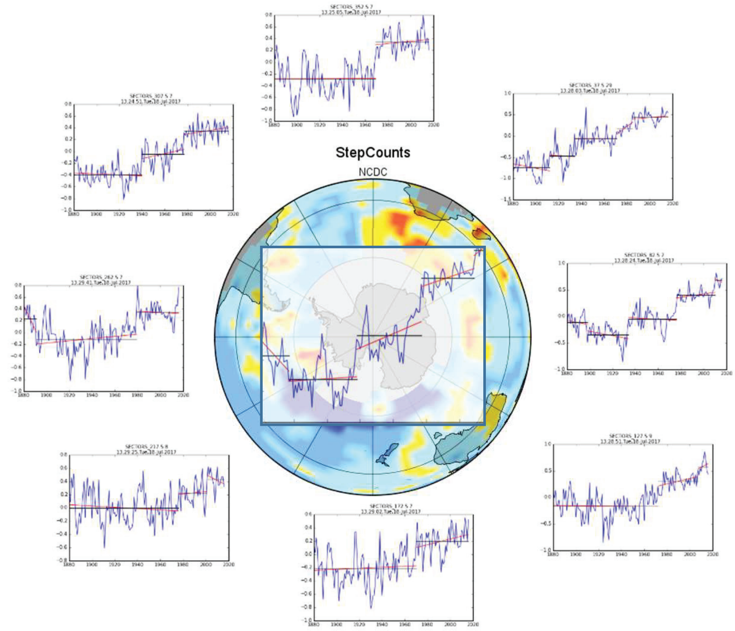

**Figure 2.** Change points over combined land/ocean for 60S.30S (centre) and eight sectors (surrounds). The reference projection is the pattern of step counts from 1880 to 2016 on a five degree grid (dark blue is zero, red is five).

weighted sector means on the other. Zonally the KPSS tests show the data to be non-stationary even after removal of the detected exogenous changes, and in fact for trend-stationarity to become *less* likely afterwards. The ADF test shows that even after allowing for autocorrelation, trend, and even after removal of a proposed exogenous change, the data are not stationary. But the ZA tests suggest that this may be due to other events (either below detection thresholds or unmasking of other exogenous types). However in the sectors, the KPSS results largely show initial non-stationarity giving way to stationarity once the exogenous change constituted by the change point is removed. This is supported by the ZA tests. The ADF tests also support this although there is a distinct longitudinal difference in the sectors in that those sectors influenced by the southern Indian Ocean are the only ones in which the ADF initially does not support trend stationarity. In the sector results where the alternative exogenous year produced by the ZA test was close the best-estimate year of shift of the MSBV for the data, the alternative chosen for the ZA on the residuals is markedly different – a consistent result. That sector scale UR tests show little evidence of UR in data, and none in residuals is consistent with multiple loci of change, smaller than zonal scale of the order of sector scale.

## 6. SUMMARY AND CONCLUSIONS

We have outlined a program of assessing presumptive shift dates in climate data. In doing so we have attempted specify testing that fits within a formal framework of severe testing. We have also utilised the idea of model mis-specification. It is telling that the MSBV, based on maximized likelihood of step changes rather than minimised residual in a segmented model, none the less compares favourably to other methods producing generally smaller models. We separate event detection from assessment, using the appearance of abrupt shifts to detect potential rapid and persistent changes, since level changes are more localizable in time. Regardless of detection method, validation is necessary. The tests outlined here assist a probative analysis, selected because they are automatable. The tests for heteroskedacidicity are conventional. Although genuine unit root behavior in these temperature records would indicate unusual physical processes, econometric methods using it as a considered basis for analysis are useful. The meaning of transient unit roots in climate data is an open question. The tests do help to locate an important potentially deceptive condition in the data, likewise assist with building a more complete picture of events. The ANCOVA approach is conventional but serves as much in attribution of change to two types of change, shift (LC) and trend-change (IO). Despite the tests themselves being automatable, the decision rules for their interpretation remain under-determined. Zonal organization may reflect aspects of atmospheric organization, but ocean organization is complex and shows scale dependence. The shift circa 1996/8 previously reported in this zone from earlier data supplied at zonal scale (JR2017) is no longer significant and 2008/9 is preferred given a sustained level shift. Spatial analyses on 2 degree and 5 degree grids in show that ocean shifts circa 1996 were limited to north of 45 degrees south in the Western South Pacific and later in the Tasman sea (Ricketts, 2017) in any case. The results we present here are entirely compatible with our H2 – complex interaction process model, and not with the H1 – non-interaction model. Drawing strong conclusions about climate change progression on the basis of global temperature records alone is a fraught endeavor at best – especially absent a physical inference model to inform statistical inference.

## ACKNOWLEDGEMENTS

Much included material forms part of the PhD thesis of J Ricketts, funded by a Victoria University scholarship.

**Table 3.** Excerpted from (Ricketts, 2017), illustrative rows only. Zonal land/ocean 60S.30S as first analysed and then subdivided into eight blocks of 45 degrees and area averaged. The blocks are identified by the centre longitude of the block. The values were chosen so that Drake's passage cleanly divides blocks in the East of South-Pacific and West of South-Atlantic. For the DRT the one shift point that the DRT prefers as trend change is marked in orange. The exogenous change-year in each case marks a disjoint separation between a prior linear segment and a posterior one – the implicit model. The KPSS, ADF, and ZA tests were also performed against the residuals after removal of the implicit model. These results are tabulated for the raw data segments and the residuals in each case in the form *raw/residual* to allow assessment of the impact of an exogenous change on the assessment of stationarity. Significant residual heteroskedacidicity denoted by * (p<0.01), † (p<0.05). (See Figure 2).

| Location | Bivariate | | DRT | KPSS | | ADF | Zivot-Andrews | | ANCOVA |
|---|---|---|---|---|---|---|---|---|---|
| 60S.30S Whole zone or Sector | Break | Shift at Break | Change not classified | Level Stationary | Trend Stationary | Inferred Unit Root | Inferred Unit Root | Difference | Persistent Regime |
| Centroid $45^0$x$30^0$ | Year | $^0$C | Pr | Pr | Pr | Pr | Pr | Years | Pr |
| 60S.30S | 1897 | -0.09 | 0.586 | 0.01/0.01 | 0.03/0.01 | >0.10/>0.1 | >0.10/0.01 | -16/0 | 0.0042 |
| 60S.30S | 1938 | 0.21 | <0.001 | 0.01/0.01 | 0.10/0.014 | >0.10/>0.1 | >0.10/0.01 | 17/0 | 0.00001 |
| 60S.30S | 1977 | 0.25 | <0.001 | 0.01/0.01 | 0.01/0.01 | >0.10/>0.1 | >0.10/0.01 | 31/0 | <0.00001 |
| 60S.30S | 2009 | 0.08 | 0.134 | 0.01/0.05 | 0.05/0.01 | 0.10/>0.1 | >0.10/0.01 | -2/0 | 0.03153 |
| **37.5** | 1912 | 0.47 | <0.001 | 0.06/0.1 | 0.03/0.1 | >0.10/0.05 | 0.10/0.05 | 5/-12 | 0.000704 |
| … | | | | | | | | | |
| **37.5** | 1985 | 0.15 | 0.29 | 0.01/0.1 | 0.01/0.1 | 0.05/0.05 | 0.05/0.01 | -18/-18 | 0.008191 |
| **82.5** | 1898 | -0.17 | 0.037 | 0.01/0.1 | 0.10/0.1 | 0.10/0.01 | 0.01/0.01 | -17/-17 | 0.017399 |
| **82.5** | 1934 | 0.45 | <0.001 | 0.01/0.1 | 0.08/0.1 | >0.10/0.01 | 0.01/0.01 | 0/-30 | <0.00001 |
| **82.5** | 2010 | 0.30 | <0.001 | 0.01/0.1 | 0.06/0.1 | >0.10/0.01 | 0.05/0.01 | -1/26 | 0.001456 |
| **217.5†** | 1976 | 0.27 | 0.001 | 0.07/0.1 | 0.01/0.1 | 0.01/0.01 | 0.01/0.01 | 70/40 | 0.009083 |
| **217.5†** | 2001 | 0.29 | 0.003 | 0.02/0.1 | 0.10/0.1 | 0.01/0.01 | 0.05/0.05 | -1/12 | 0.004911 |
| **352.5*** | 1969 | 0.58 | <0.001 | 0.01/0.1 | 0.01/0.1 | 0.01/0.01 | 0.01/0.01 | 0/66 | <0.00001 |

## REFERENCES

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287-1294.

Dickey, D. A., & Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica: Journal of the Econometric Society*, 1057-1072.

Jones, R. N., & Ricketts, J. H. (2017). Reconciling the signal and noise of atmospheric warming on decadal timescales. *Earth Syst. Dynam., 8*(1), 177-210. doi:10.5194/esd-8-177-2017

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics, 54*(1), 159-178.

Little, M. A., & Jones, N. S. (2011). *Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory.* Paper presented at the Proc. R. Soc. A.

Mayo, D. G., & Spanos, A. (2004). Methodology in Practice: Statistical Misspecification Testing. *Philosophy of Science, 71*(5), 1007-1025. doi:10.1086/425064

Mayo, D. G., & Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science, 57*(2), 323-357.

Ricketts, J. H. (2017). *Understanding the nature of abrupt decadal shifts in a changing climate (in prep).* (PhD), Victoria University, Victoria Institute for Strategic Economic Studies.

Stock, J. H. (1994). Unit roots, structural breaks and trends *Handbook of econometrics* (Vol. 4, pp. 2739-2841).

Zivot, E., & Andrews, D. W. K. (2002). Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. *Journal of Business & Economic Statistics,* 20(1), 25-44.