21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Stepwise symbolic regression compared to a probabilistic bivariate test for step-change detection

J.H. Ricketts [a]

[a] *Victoria Institute of Strategic Economic Studies, Victoria University, 300 Flinders St, Melbourne, Victoria, Australia*
Email: james.ricketts@live.vu.edu.au

Abstract: The idea of stepwise symbolic regression (SSD) using genetic programming was introduced at MODSIM 2013 and illustrated with climate data. SSD cleanly separates signals into linearly combined symbolic modal functions (the "explained" components) and a residual (the "unexplained"). The idea of examining the residuals for signatures of slowly developing processes, for example sea level change, was demonstrated.

SSD can be used for exploration of multivariate series, and modal functions can be composed of smooth as well as non-differentiable functions – for example IF-THEN-ELSE, and hence has potential for change-point detection in either univariate or multivariate time series.

Climate change is inherently a multivariate problem suite, and regime shifts (for example the Pacific Decadal Oscillation) may show as change points in either univariate or multivariate time series. Jones in 2012, proposed that periodic step-like regime shifts occur in the recent climate record. He used a bivariate test to examine possible shifts in SE Australian regional climate records. A working hypothesis is proposed, that climate change can be usefully modelled as a series of step shifts which comprise a large component of recent change. This is further hypothesized to be more observable at finer scale and so this and related work looks at global and then hemispheric and then zonal data sets and, to date, has shown that the working hypothesis holds.

A major feature of interest in recent temperature records, certainly in the context of risk, is the trend and the changes of trend. Both debate and climate risk are often framed in a context of an expectation that climate changes, especially anthropogenic ones, are expected to reflect smooth change. It is of interest to be able to identify times of regime shifts in composite records such as mean annual temperature records, as these may indicate increased near term risk.

The Maronna bivariate test used assumes no trend or change of trend and at most a single step shift. Testing has shown its ability to locate a time of shift is not sensitive to trend when even small shifts are present although its imputed significance could be. A probabilistic bivariate test (*PBV*), which builds on the Maronna test, incorporates some decision rules, and can be applied to data which contains zero to many shifts, is the subject of another paper submitted to this conference.

This gives the opportunity to compare PBV to the very general framework of SSD. Two different SSD modal function types and selection procedures were tested. One, *SSDstep* with early termination at each iteration, although quite general, is shown to produce similar results to the bivariate test. The second, *SSDtrend*, with delayed termination at each iteration, specifically models a break as a piecewise regression with two unrelated linear segments. It shows different behavior, especially during the early records, but converges on the *SSDstep* and the bivariate test after the 1960s where all methods showed good consensus on timing of shifts. This is consistent with regime shifts in temperature records being augmented by anthropogenic heat retention.

This paper will demonstrate the potential use of non-differentiable SSD modal functions in explorations of data sets which may be imprinted by regime shifts, and contrast the results with the PBV test which is under development.

Keywords: *Symbolic regression, stepwise symbolic decomposition, Maronna bivariate, regime shift*

1. INTRODUCTION

Jones (2012) has proposed that periodic step-like regime shifts occur in the recent climate record. He used a bivariate test (Maronna and Yohai, 1978) to examine possible shifts in SE Australian regional climate records. Additionally a working hypothesis is proposed, that climate change can be usefully modelled as a series of step shifts which comprise a large component of recent change. This is further hypothesized to be more observable at finer scale and so this and related work looks at global and then hemispheric and then zonal data sets and, to date, has shown that the working hypothesis holds.

The Maronna Bivariate test is an homogeneity test which was proposed as a means of identifying inhomogeneities in the relationship between two time series. Although the original paper proposed a general framework, the cases analysed consisted of single shifts in the mean between two series. Both type I and type II cases (fully correlated and randomly correlated relationship between the variates) were analysed. Based on this work Vives and Jones (2005) proposed a method of using a random control and iteratively selecting multiple break points. The automation of this procedure and illustrative results are the topic of a companion paper at this conference, the probabilistic bivariate test (PBV) (Ricketts, 2015).

The idea of stepwise symbolic regression (SSD) using genetic programming was introduced at MODSIM 2013 by Ricketts (2013), and illustrated with climate data. SSD cleanly separates signals into linearly combined symbolic modal functions (the "explained" components) and a residual (the "unexplained"). The idea of examining the residuals for signatures of slowly developing processes, for example sea level change, was demonstrated. A major advantage of symbolic regression over empirical methods is that the outcomes of the analyses are standard mathematical expressions. Symbolic regression is achieved, non-deterministically, by use of a genetic programming system from Cornell Creative Machines Lab called "Eureqa" (Schmidt and Lipson, 2013).

The decomposition was shown as

$$y_t = f(Xs_t) + f'(Xs_t) + \cdots + E_t \quad (1)$$

Using Eureqa, building blocks are selected from a menu of simple functions and operators such as Addition, Divide, Sin, Log, Exp, and arbitrary complex formulations of these. Each building block is assigned a complexity metric. Each step in the process attempts to find, by some decision rule, the "best balance" of complexity and some selected objective, for instance RMS error in order to explain the residual from the previous iteration. A modal function was deemed to be selected once a function was discovered that was "better" than *y=constant.*

Subsequent analysis of $E_t$was used to detect residual but coherent signals, across ensembles of similar data (tide gauge measurements), in keeping with anthropogenic climate warming. This paper, by contrast, demonstrates the use of the modal functions $(f(Xs_t) + f'(Xs_t) + \cdots)$ composed of carefully selected building blocks to locate abrupt or episodic change points.

Climate change is inherently a multivariate problem suite, and regime shifts. For example the Pacific Decadal Oscillation (PDO) (Trenberth and Hurrell, 1994) may show as change points in either univariate or multivariate time series.

1.1. Mean Annual Global Temperature

An essential feature of the mean annual global temperature is that it is a composite signal. It is also an estimate, that is, whilst it is derived from measurements these are not taken in a way that admits direct computation of a mean. The sampling density tends to have increased over time, be biased towards inhabited areas, and to have limited ocean representation, especially in the early record. Further the nominal temperature incorporated is the air temperature at two metres, a measure that is also not generally taken over the ocean and inferred instead from sea-surface temperatures.

Ocean temperatures and gradients partly determine wind flows, and also influence moisture fluxes into the air. Thus they are strongly causal of the land temperatures, evaporation and rainfall. Much attention has focused of late on the broad characteristics of the mean annual global temperature. Implicit in some risk analyses is an assumption that the signal consists of noise plus a warming signal which is slow moving and – in effect – smoothly differentiable.

On the other hand a number of more local phenomena show as abrupt changes of regional temperatures and linked variables, for example see Jones (2012). Therefore the global mean signal is composed from regional

signals with their discontinuities, detectable over most of the Earth (Ricketts, 2015), each of which would be expected to show at least traces, even inhomogeneities, in the mean annual global signal. Traces due to transient extrinsic excursions such as El Niño Southern Oscillation (ENSO) and cooling after volcanic eruptions are not to be confused with those due to climate.

1.2. Deterministic breakpoint methods.

There is a growing number of methods for determining the structure of inhomogeneities in a time series, and the reader is referred to http://www.changepoint.info/ which attempts to give a portal for discussion surrounding these methods. Other methods have been tried and given good results. These include Structural Change and Change Point methods, both of which can be found via the above web site. Assessment of these methods is out of scope for this paper.

1.3. The search problem

The problem of detecting multiple steps in a time series which may contain zero or more steps breaks is essentially a search problem. We are looking for a series of times with specific characteristics – viz that they are times of abrupt shift which may also be associated with changes of trend. It may also be represented as an optimization problem where the time series is segmented to minimize some error metric. A step change is one of a class of breakpoints where some element of the probability distributions changes persistently. Transient changes are not to be identified as step changes although they may coincide with a step.

When a break is detected the issues of (a) how the statistics either side are affected and accounted for, and (b) how the order of determination affects the composition of the solution sets, are important.

All search methods have their own biases towards type 1 and type 2 error. This paper concerns the selection of building block operations and analyses of the composite modal functions for detection of changes in mean, where data may contain confounding autocorrelation and changes in trend, and variance. As such it is specifically biased to type 2 errors because the role of the search method is seen to be to focus attention on possible events, not to prove their existence.

The remainder of the paper is structured as follows. In section 2, the selection of building blocks and construction of modal functions is discussed. In section 3 an experimental method is briefly described and two SSD variants and the probabilistic bivariate are tested against known climate data; the results are tabulated and discussed in section 4. A brief discussion and conclusion is in Section 5.

2. SETTING UP EUREQA

2.1. Building blocks

In change point analysis in the climate domain one is most often interested in finding evidence of small changes embedded in a noisy domain. The selection of building blocks is critical since it reflects the assumptions of the researcher as well as affecting the evolution of the search.

For example, Eureqa provides two building blocks in particular, “step” and “if-then-else” which can segment data.

$$\text{“Step” is numeric, } step(x_t) = \begin{cases} 1, x_t > 0 \\ 0, x_t \leq 0 \end{cases}\text{,} \tag{2}$$

which allows expressions such as $y_t = k \cdot step(x_t)$ which will have the values $k \cdot step(x_t) = \begin{cases} k \cdot x_{t,} > 0 \\ 0, x_{t,} \leq 0 \end{cases}$

$$\text{“if-then-else”} = \begin{cases} g(x_t), f(x_t) > 0 \\ h(x_t), f(x_t) \leq 0 \end{cases}. \tag{3}$$

2.2. Suitable SSD models for breaks in temperature time series

SSD was proposed as a data driven method which allows the system to trial quite arbitrary functions composed of regular building blocks Early trials with temperature data showed that if *sin* and *cos* or other periodic functions were included in the building blocks then the modal functions would converge rapidly on a

polynomial plus from 2 to 5 sinusoids with a featureless residual. Whilst it is tempting to conclude this disproves the necessity to postulate abrupt shifts, it is also the case that summed sinusoids can sufficiently mimic step shifts that they fall below detectability.

Therefore the SSD modal functions have been restricted using Eureqa's ability to specify skeleton target functions.

A first general model for a modal function finding breaks is …

$$y = if\left(step\left(f(x_{1..n})\right), g(x_{1..n}), h(x_{1..n})\right), \text{ where } step \text{ is defined above.} \quad (4)$$

This (referred to hereafter as *SSDstep*) requires Eureqa to separately determine a partitioning function and a function for each partition. Whilst this is highly under-determined, one would expect that, should the data contain a significant component of abrupt shifts then the initial stages of the search will be guided primarily by the ability of *f* to locate the strongest shift, with *h* and *g* less influential. Modal functions were chosen as the least complex found with a single discontinuity which remained present in the Pareto set for one minute once they appeared. Overall termination was deemed once the method was plainly refining previous results and no new break times were found after two iterations.

In the Eureqa framework the search expression is closely mirrored. To model a temperature series ("Temp") as a time series by date ("Date") , *f1*, *f2*and *f3* are used to represent three separate functions. One should note that the step function is largely redundant since this is equivalent to …

$$Temp = if\left(f_1(Date), f_2(Date), f_3(Date)\right) \text{ with an implicit coercion of } f_1. \quad (5)$$

Building blocks were, Constant, Input-variable, Addition, Subtraction, Multiplication. Step-Function, If-Then-Else were not tested directly, although they were used for comparative work.

A second model (referred to hereafter as *SSDtrend*) was tried, based on a general two part piece-wise regression, since I am interested in locating series of steps separating approximately linear segments.

$$\begin{aligned} Temp &= if(Date > f_0(), f_1() + Date \cdot f_2(), f_3() + Date \cdot f_4()) \\ &+ if\left(Date < 1880, log(-1), if(Date > 2014, log(-1), 0)\right) \end{aligned} \quad (6)$$

In (6), the Eureqa convention is that empty parameter lists such as $f_0()$ represent constants to be determined and the second term takes advantage of Euequa's weighting to implement a constraint on *Date* to be in range.

The error metric was selected as "Squared error (AIC)". This was selected because it is one presented by Nutonian Inc. as "analog[s] to Akaike Information Criteria", and incorporates a complexity criterion. Modal functions were the least complex non-linear equations found after 10 minutes and the search for further modal functions terminated once none were found in that time.

2.3. Shift points as modal functions in the Stepwise Symbolic Regression framework

SSD produces a decomposition into a sequence of modal functions and a residual. Previously I concentrated on the structure of the residual and demonstrated that restriction of the composition of the modal functions meant that the residual would retain "signatures composed of noise plus any components which cannot be built out of the building blocks. Due to the nature of the building block functions any signal remaining in the residual would be irregular, asymmetric or non-polynomial.

Here we are interested in analyzing the modal functions for clues as to the location in time of shift points which would be discontinuities. The modal functions in equation (5) are much less constrained than in equation (6). This is controlled by a variation in the halting rules. For (5) a modal function was deemed selected if it was the least complex step function with a single root, stable after 30 seconds. For (6) only a single solution is available at any time so halting was at 10 minutes and that rule was the modal function. Re-iterated selection of modal function is guaranteed to be convergent for (5) and in practice proved to be for (6). The overall decomposition into modes was deemed to be over when the candidate modal function was a flat line or the process returned duplicate step points (representing pointless refinements).

The PBV includes, as a tunable parameter, a prohibition rule (defaulting to seven years), implemented to minimize false positives associated with sub-decadal variation, that prevents a step change from being returned within seven years. The SSD approach does not implement this, and so when two steps are found within seven years the earlier one is retained.

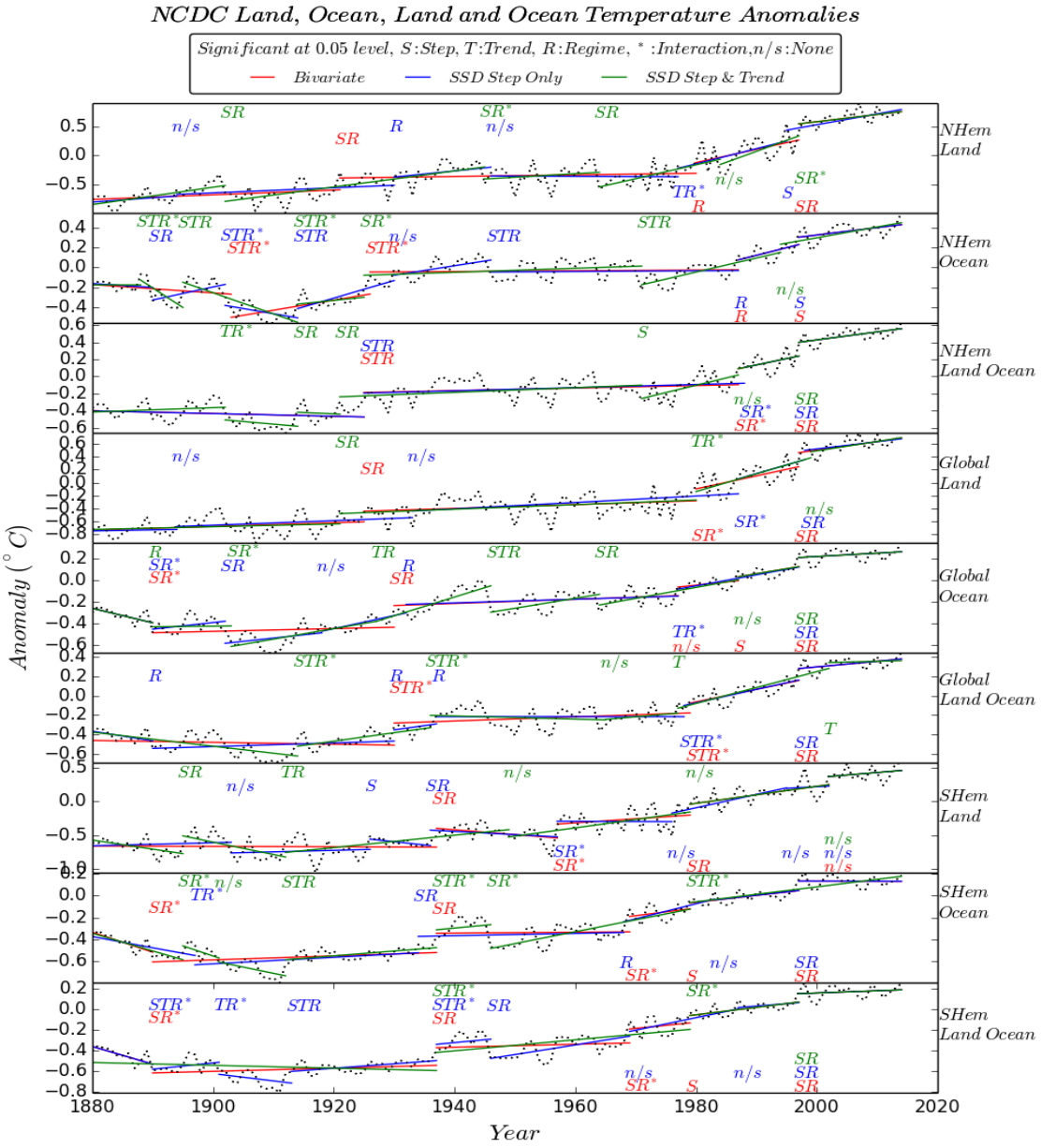

Figure 1. Comparative plots of the three methods applied to nine data sets. ANCOVA has been used to test the statistical significance at P<0.05of shifts (*S*), trend changes (*T*), and two segments preferred to one (*R*). *significance of *R* is less than either *S* or *T*. *n/s* indicates not significant when tested this way.

Also, the modal functions are not used to estimate shifts at these times, rather a conventional ANCOVA is used as reported in Ricketts (2015).

3. GLOBAL AND HEMISPHERIC LAND, OCEAN, AND LAND AND OCEAN DATA

On 29 May 2015, annual files were downloaded from the National Climatic Data Centre (NCDC) reference site at ftp://ftp.ncdc.noaa.gov/pub/data/mlost/operational/products/. These covered Land, Ocean, and combined Land and Ocean, in ASCII format.

Monthly Southern Oscillation Index values (SOI) were downloaded on 22 September 2015, from the Bureau of Meteorology ftp://ftp.bom.gov.au/anon/home/ncc/www/sco/soi/soiplaintext.html, in ASCII format and converted to annual (May to April) means. SOI is a lead index of ENSO.

The global, northern and southern hemispheric data were then analysed by SSD using equations (5) and (6) separately and then compared to results from PBV. For each shift date in each determination, diagnostics were available by ANCOVA (using R script). These are size of shift, change of trend, and probabilities associated with these, plus the probability that a two segment piecewise regression explains more error than a single regression over the combined segments (see Figure 1 and Table 1), plus the joint independent probability that neither the shift nor the change of trend are significant (Table 1 only).

4. RESULTS

All methods produced an assortment of nominal shift dates. Both the SSD methods had a tendency to place more breaks than PBV, and this is seen especially between the 1880s and 1920s, and between the 1920s and late 1970s. These were always placed after the late 20th century dates and probably represent in part an artifact of the method. *SSDstep* did a reasonable job of finding PBV's shift dates, whereas both the SSD approaches may have identified a trend change associated with the start of a slow decline in temperatures from around that time. An expected shift in the northern hemisphere at 1986 may have been found in southern ocean temperatures by *SSDstep* but it does not reach statistical significance in isolation. The event round 1976/8, associated with a step change in Pacific winds, the great Pacific reorganization, is generally found to be a mixture of step change and change of trend.

Table 1. Years found by each method shown by decade. These years are the year during which a change occurred, and the next year is the start of a new regime. Exponents on the years indicate whether or not the point remained significant under an ANCOVA and why. S–step, T–trend, *–two segment preferred, A–all of these three, J–Only by a Joint Independent Probability of Step and Trend.

Zone	Method	1880s	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1990s	21C
Global land	SSD Step		93				32					86^{S*}	97^{S}	
	SSD Trend					20^{S}					79^{A}		98	
	PBV					$\mathbf{24^{S}}$					79^{T*}		$\mathbf{96^{S}}$	
Global Ocean	SSD Step	89^{A}		01^{S*}	17		31^{T*}				76^{T*}		96^{A}	
	SSD Trend	89^{A}		02^{S*}		26^{T*}		45^{A}		$\mathbf{63^{S}}$		86	96^{S}	
	PBV	$\mathbf{89^{S*}}$				$\mathbf{29^{S*}}$					76	86	$\mathbf{96^{S}}$	
Gl Land/Ocean	SSD Step	89^{*}				29^{*}	36^{*}				77^{A}		96^{S*}	
	SSD Trend				13^{A}		35^{A}			64	76			01^{S*}
	PBV					$\mathbf{29^{A}}$					78^{A}		$\mathbf{96^{S*}}$	
NH Land	SSD Step	93				29^{*}		45^{ST}			76^{T}		94^{S}	
	SSD Trend			01^{S*}				44^{T*}		$\mathbf{63^{S*}}$		83	96^{S}	
	PBV				$\mathbf{20^{S}}$						79^{T}		$\mathbf{96^{S}}$	
NH Ocean	SSD Step	89^{S}		01^{A}	13^{A}	29		45^{A}				86^{S*}	96^{S*}	
	SSD Trend	87^{A}	94^{S}		13^{A}	24^{S*}				$\mathbf{70^{A}}$			93^{S}	
	PBV			$\mathbf{02^{A}}$		$\mathbf{25^{A}}$						86^{*}	$\mathbf{96^{SJ}}$	
NH Land/Ocean	SSD Step					24^{A}						$\mathbf{87^{S*}}$	96^{S*}	
	SSD Trend			01^{T}	$13^{S}/20^{S}$					$\mathbf{70^{A}}$		86	96^{S*}	
	PBV					$\mathbf{24^{A}}$						$\mathbf{86^{S*}}$	$\mathbf{96^{S*}}$	
SH Land	SSD Step			02^{*}		25^{S}	35^{S}		56^{S*}		75		94	01
	SSD Trend		94^{S}		11^{T}			48			78			01
	PBV						$\mathbf{36^{S}}$		$\mathbf{56^{S}}$		78			**01**
SH Ocean	SSD Step	86^{T}		00			33^{S}			67		$\mathbf{82^{T}}$	96^{S}	
	SSD Trend		94^{*}	00	11^{A}		36^{T*}	45^{A}			78^{S}			
	PBV	$\mathbf{89^{S*}}$					$\mathbf{36^{S}}$			$\mathbf{68^{S}}$	78		$\mathbf{96^{S*}}$	
SH Land/Ocean	SSD Step	89^{S}		00^{T}	12^{A}		36^{A}	45^{A}		$\mathbf{68^{*}}$		86	96^{S}	
	SSD Trend						36^{A}				$\mathbf{78^{S*}}$		96^{S}	
	PBV	$\mathbf{89^{S*}}$					$\mathbf{36^{S}}$			$\mathbf{68^{S}}$	78		$\mathbf{96^{S}}$	

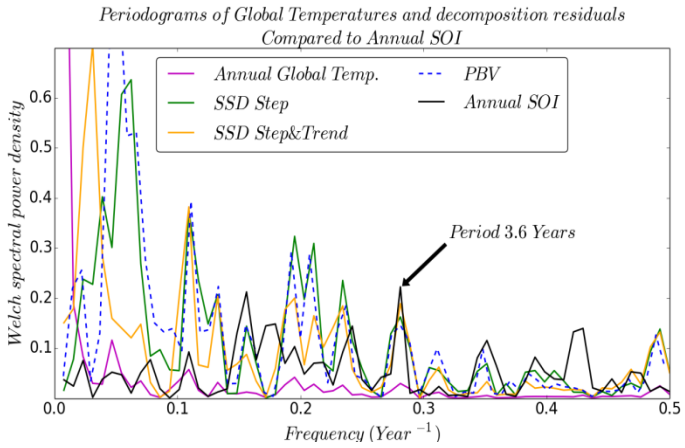

Figure 2. Welch periodograms of the Global Mean Land/Ocean temperature is compared to residuals from PB,V and both SSD analyses, and the SOI. Note the coincident peaks at a frequency of 0.28 Year^{-1} (a period of 3.6 Years).

The post 1996 event however stands out as a step change rather than a trend change.

Although analysis of the residuals is not complete, it can be shown that a frequency structure remains. Piecewise modification of time series is expected to impose some artifacts but encouragingly, a principal ENSO frequency exists. Figure 2 shows as an example, a Welsh periodogram in which a prominent SOI peak of 0.28year^{-1} consistent with An and Wang (2000) is present in the global mean temperature and enhanced in the residuals.

5. DISCUSSION

There are major differences between the PBV and the two SSD methods. One major difference, and a point of interest in this work, how they retain or discard error estimates from prior iterations. The Bivariate test, at each iteration, examines only the data within a bounded segment; the SSD approach demonstrated here successively refines the error for all of the data, and does not segment the dataset. A second major difference is that SSD is not restricted to a step and linear trend model, rather, linearity emerges naturally and in fact many times it produced only steps without trends; in the cases where a trend was imposed it was detected early and retained. A third is that the Bivariate test has an imposed tunable seven year refractory period after a break, SSD had none such, and this shows in that some clustering occurred, possibly indicating refinements of the "true" break dates due to retention of temporally distant error terms.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the valuable advice of two anonymous reviewers.

REFERENCES

An, S.-I. & Wang, B. 2000. Interdecadal Change of the Structure of the ENSO Mode and Its Impact on the ENSO Frequency*. *Journal of Climate,* 13(12)**,** pp 2044-2055.

Jones, R. N. 2012. Detecting and attributing nonlinear anthropogenic regional warming in southeastern Australia. *Journal of Geophysical Research: Atmospheres (1984--2012),* 117(D4)**,** pp.

Maronna, R. & Yohai, V. J. 1978. A bivariate test for the detection of a systematic change in mean. *Journal of the American Statistical Association,* 73(363)**,** pp 640-645.

Ricketts, J. H. 2013. Using genetic programming for symbolic regression to detect climate change signatures *In:* Piantadosi, J., Anderssen, R. S. & Boland, J. E. (eds.) *MODSIM2013, 20th International Congress on Modelling and Simulation.* *www.mssanz.org.au/modsim2013*. Adelaide, Australia Modelling and Simulation Society of Australia and New Zealand.

Ricketts, J. H. 2015. A probabilistic approach to climate regime shift detection based on Maronna's bivariate test. (Submitted). *The 21st International Congress on Modelling and Simulation (MODSIM2015).* Gold Coast, Queensland, Australia

Schmidt, M. & Lipson, H. 2013. Eureqa 1.12.1 Beta. Available from http://www.eureqa.com/.

Trenberth, K. E. & Hurrell, J. W. 1994. Decadal atmosphere-ocean variations in the Pacific. *Climate Dynamics,* 9(6)**,** pp 303-319.

Vives, B. & Jones, R. N. 2005. *Detection of abrupt changes in Australian decadal rainfall (1890-1989)*: CSIRO Atmospheric Research.