20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Using genetic programming for symbolic regression to detect climate change signatures

J.H. Ricketts[a]

[a] *IEEE Computational Intelligence Society*
Email: jim.ricketts@ieee.org

Abstract: Most often, climate change signals are slow moving (in human terms), low amplitude changes embedded in high amplitude, noisy data. There are many techniques for extracting such signals. This paper introduces a technique which in contrast with empirical methods produces a decomposition of a time series into a set of equations plus a "residual". This is referred to as stepwise symbolic secomposition (SSD). The extraction of symbolic equations in lieu of empirical functions assists with characterization of time series.

This paper takes two examples of climate data, and applies two different techniques for characterising the low amplitude, slow change embedded therein. The record of CO_2 levels at Mauna-Loa since March 1958, is used to demonstrate and contrast empirical mode decomposition (EMD), and SSD.

The mean monthly tidal gauge records from the small number of gauges which have more than 120 years of data are analysed in more detail by SSD, then three techniques are used to characterize the residual. The techniques are (a) LOESS smoothing, (b) EMD, and (c) high order polynomial regression.

SSD uses a genetic programming system called Eureqa from Cornell Creative Machines Lab guided by an information metric, to extract the most compact informative function it can at each step; the process is repeated on the residuals until no sufficiently informative function is found. EMD and SSD are in stark contrast in the order in which signals are decomposed. EMD extracts high frequency components first. SSD extracts components based on a mixture of parsimonious representation and variance explained. EMD leaves a low frequency filtrate of the signal in its residue, SSD tends to operate as a broad band filter, leaving high frequency noise plus a possible low frequency signal. EMD, LOESS and polynomial fitting all serve to extract low frequency components of the signal from this SSD residual.

For the set of tidal gauge data, in the absence of a change in global sea level rise, the SSD procedure should randomize all segments of residual signals equally – the residual should be whitened with respect to the initial signal. It is shown however that late 20th century portions of the residuals show behaviour that is consistent with accelerating sea level rise in keeping with the bulk of the literature.

Keywords: *Genetic programming, empirical mode decomposition (EMD), stepwise symbolic decomposition (SSD), sea level rise.*

1. INTRODUCTION

The aim of this paper is to introduce the use of genetic programming to implement symbolic decomposition of time series data into sets of equations plus a residue. This is contrasted with empirical methods. The extracted equations have potential to be easily related to physical properties underlying the time series.

1.1. Signatures

Climate change attribution studies often use the term "signature" without formal definition. A signature, in this paper, is an observation or result of analysis which is characteristic of an hypothesis and discriminatory from competing hypotheses. It need not have predictive power as such, that is, the signature of an event would not be a necessarily be measure of some quantity, nor a prediction of its future values; rather it may be the general trajectory of change or the "shape" of the change.

Implicit in the idea of signatures, is the notion that processes of interest consist of superimposed signal, or perhaps many signals, and noise. In a complex situation where there are superpositions of signals over-laid by noise it may be impossible to fully separate any of the components.

1.2. Climate Change

Climate change is the result of what are in human experience slow, low frequency processes. Natural variability, occurring at all time-scales, certainly shows effects from variations in solar radiation, orbital cycles, vulcanism, and a number of quasi oscillatory processes. Anthropogenic effects are attributed processes which are rapid compared to many natural "cycles" but still slow in human terms, e.g. especially the slow accumulation of greenhouse gases in the atmosphere, and physical consequences of that. From first principles (1) anthropogenic climate change then would be expected show as changes occurring after the presumptive causal anthropogenic changes occurred; (2) such changes may be simply additive on other variations; modulate them, or possibly even be modulated by them. In all cases we expect standard physics to apply, and we would expect such signatures to be expressed as changes recent in time, probably monotonic, and non-stationary. The so-called "hockey stick" curve (see Figure 3 in Mann et al. (1999)) is one such example.

1.3. Non-local and locally adaptive functions

Consider ordinary regression by least-squares (OLS). When regression parameters are all influenced by every point, the relationship is described as "non-local". The estimated shape of the curve at a particular point in time is influenced by points even in the far future. In a non-stationary situation (i.e. one where the statistics vary over time) this is not correct, and can be quite pathological.

When applied to time series, "local functions of time" are those functions in which values of the regression parameters are estimated from only a small set of proximal values, and thus can vary in time. Similarly, many splining methods use composites of curves (typically low order polynomials, constrained to join smoothly) estimated only from a small proximal set of values. Of course such splines provide no basis for extrapolation.

1.4. Symbolic Regression

Three general approaches to finding relationships between variables for predictive purposes are considered. Conventional regression approaches (e.g. linear, polynomial, multiple, and non-linear), seek to estimate parameters to some sort of equation, which could be represented in the form in (1).

$$y_t = f(Ps, Xs) + E_t \tag{1}$$

Where *f* is a predetermined function, *Ps* is a set of parameters to be estimated, *Xs* is a set of predictor variables and *E* is usually presumed to be a normally distributed random error term.

Empirical methods (e.g. empirical orthogonal functions, empirical orthogonal teleconnections, or empirical mode decomposition) use some heuristic to decompose the relationships into un-parameterized sets which could be shown as (2).

$$y_t = \sum_m [Xs_t^m] + E_t \tag{2}$$

Where functions are not given in explicit form, *m* is the mode (roughly, the order in which the component is extracted), and *Xs* are linearly decomposed into *m* factors.

In symbolic regression one seeks to find functions in the form of equations (and parameters) which explain the data. If we subsume the task of estimating parameters into the task of estimating functions we can represent this as shown in (3).

$$y_t = f(Xs_t) + f'(Xs_t) + \cdots + E_t \tag{3}$$

It is often introduced as an alternative approach to supervised learning (Schmidt & Lipson, 2010). A major advantage of symbolic regression over empirical methods is that the outcomes of the analyses are standard mathematical expressions. Symbolic regression is achieved by use of a genetic programming system from Cornell Creative Machines Lab called "Eureqa" (CORNELL 2011).

1.5. Empirical Mode Decomposition

Empirical mode decomposition (EMD) was introduced by Huang et al. (1998). It is a true empirical method, designed to work with non-stationary, non-linear data, which partitions the time series into a finite, usually small number of components known as intrinsic mode functions (IMFs) plus a residue. IMFs are extracted via a filtering process which extracts higher frequency components first, thus for our purposes the item of interest is likely to be the residue. The essential point is that the functions are (a) non-symbolic and (b) local. The consequence of locality is that signature events such as accelerations may be blurred due to the filtering. A known problem is that the discontinuities associated with the boundary data and missing data are prone to propagating artifacts into the IMFs (Wu and Riemenschneider, 2010). This is undesirable in this analysis.

2. METHODS

2.1. Stepwise Symbolic Decomposition

Stepwise symbolic decomposition (SSD) is defined here as a recursive procedure whereby symbolic regression is used to fit an equation to data, leaving a residual, and then successively equations are sought to fit the residuals from each prior step, until the best function found is y=constant. Each symbolic equation corresponds to a mode in an empirical method and for convenience the term "modal function" is used. Functions are created from a small suite of component operators and functions, selected for the purpose. Any procedure may be used to create these interim functions, but in this paper Eureqa was used to find the most parsimonious equation at each step.

At this point the final residual time series is relatively whitened with respect to the initial time series. Since it is the residual of a finite number of function applications drawn from a restricted set of component functions, is will not necessarily be fully whitened – hence it may carry a signal from some processes which cannot be compactly characterised by the components. The sum of the modal functions will constitute the "explained" components of the signal; the residual will consist of noise, plus a potential "unexplained" component. Attribution of a signature is the detection of just such a component.

Consider a time series *S(t).* Write it as a process model with y_t being for example sea-level at some date, *F(t)* a pure function of time and E_t an error or residual term.

$$y_t = F(t) + E_t \text{ which can be rearranged, } E_t = y_t - F(t) \tag{4}$$

Step 1. We will minimise the estimated mean error. Estimate a function *F(t)* and apply it to obtain an estimated residual.

$$E_t^1 = y_t - F^1(t) \tag{5}$$

Step 2. Now successively estimate functions which minimize the residuals between and then compute those secondary residuals,

$$E_t^{n+1} = E_t^n - F^{n+1}(t), \text{ which is equivalently, } E_t^{n+1} = y_t - F^1(t) \ldots - F^{n+1}(t) \tag{6}$$

and in which, since each function explains less variation, the error terms behave asymptotically,

$$E_t^{n+1} \to E_t^{\infty} \tag{7}$$

Step 3. Stop if the best function fitted is constant.

Step 4. The residual term E_t^n is now whitened relative to y_t but may contain components of functions which have not been admitted by the function compositions used in obtaining it. Therefore it is fitted to some indicative function. The result of this final fitting step will constitute the "signature". It is this concept of "relative whitening" that is depended on in the final analysis.

The functions obtained on the way through will give information about underlying time dependent processes in a convenient form.

SSD was performed using Eureqa. Target expression building blocks were; constants, addition, subtraction, multiplication, division, sine, cosine and a logistic function. The error metric was set to mean squared error. Trial solutions were ranked using Eureqa's standard fitness criterion (fitness/complexity), and the run was halted 30 seconds after the first ranked trial solution was more complex than $y=k$; or after five minutes in which case the SSD was complete. There are two reasons for these halting criteria. (1) It was considered desirable to decompose the signal into simple expressions and in general the first proposed solutions are simpler; (2) it was expected that subsequently discovered functions would cover unexplained residuals. If the SSD was not complete the selected function and a new residual were computed, and the best function for explaining the new residual was found the same way.

2.2. Post processing of SSD

The remaining step in this process is the extraction of a slow moving signal from E_t^{n+1}. For this purpose any standard method can be applied. A LOESS smooth or even a polynomial may be used. It is important to note that the residual signal being processed is one from which much real information has been removed and therefore one should beware of drawing over-strong conclusions from a single analysis. In this paper two different climate variables are analysed. Now are considered in order, the CO_2 curve from Mauna Loa, and the monthly mean tide gauge measurements from a small number of gauges with more than 100 years of data. CO_2 data are decomposed by EMD and SSD to contrast the approaches but further analysis is out of the scope of this paper. Tide gauge measurements are subjected to a further, illustrative analysis.

2.3. Mauna-Loa CO_2 Curve

Data were downloaded from NOAA (2012). Several estimates are available in this file including interpolation of missing months, but the decision was made to use the un-interpolated monthly average as these data are less complete, giving a chance to assess how each method deals with missing data. The EMD R package was downloaded from CRAN (2013). Using R script, EMD was fitted using the "wave" stopping parameter after detrending and LOESS was fitted with a span parameter of 1/3. SSD was performed on the un-detrended data as described above. Both EMD and SSD produce a "residual" and a set of extracted functions.

2.4. Tide gauge analysis

A complete set of monthly tide gauge RLR records was downloaded from PSMSL (2011). Since Jevrejeva et al. (2008) suggest that there may be a "60 to 70 year cycle", individual gauges were selected on the basis that they consisted of at least 120 years of collection, be complete over the first 30 and last 30 years, and be least 90% complete. Twenty seven records met these criteria, the bulk of which are either Baltic Sea or North Atlantic and many of these are discontinuous during the period of the second world war.

A simple physical model was used as a basis for determining a signature of climate change. The oceans are viewed as interconnected basins within which various influences can cause bulk movement either from edge to edge of a basin or between basins, to which there is an accepted near constant background increase, and into which and from which there may be annual to sub-decadal scale movement to and from land. Isostatic rebound which is known to affect tide gauge measurements, presumed to be approximately linear, was not compensated for as it was expected to be removed in pre-whitening. Similarly the annual cycle was not removed as it was also expected to be identified and removed in pre-whitening, and it was interesting to determine whether or not this step was able to deal with the changing annual cycle evident in a number of gauge data. Local changes in the water table are not directly accounted for in this work and are potentially confounding factors. Since the selected data are representative mostly of the North Atlantic and Baltic the data could not be used to determine inter-basin effects, nor even intra-basin movements.

A signature of climate change would be a tendency for the unexplained component remaining in the residuals to be relatively random in the year 1900, reflecting mostly inter and intra-basin effects, and more coherent with a positive bias in year 2000. Water table changes may show locally and induce the same signature if water table withdrawal post-war causes subsidence under tide gauges. It is assumed that such effects random over the gauges used, however this assumption has not been tested nor compensated for.

The work of both Jevrejeva et al. (2006) and Watson (2011) suggests there might be a pause in SLR mid-century which they demonstrated with a quadratic fit to the level. Such a curve assumes and imposes a constant acceleration, and thus cannot be used to locate changes in acceleration; so it is important to use a less constrained method.

The data were pre-whitened using SSD as described above. The residuals were then fitted using three different smoothing techniques with different characteristics. Firstly they were fitted to a sixth order polynomial. The sixth order polynomial is acknowledged to be at first sight a severely under-constrained curve, however it was chosen since (a) it is even, (b) it allows first and second derivatives to be computed and for the second derivative to be assigned up to three inflexion points, (c) the curve was not being extrapolated beyond the data, (d) analysis of the values at years 1900 and 2000 were to be compared to lessen influences from the higher order terms.

The same analysis was performed using LOESS smoothing of the residuals followed by extraction of estimates of the first and second derivative of the smooth using the R "splinefun" function. Lastly the residuals were also smoothed using EMD and derivatives extracted using the same R function.

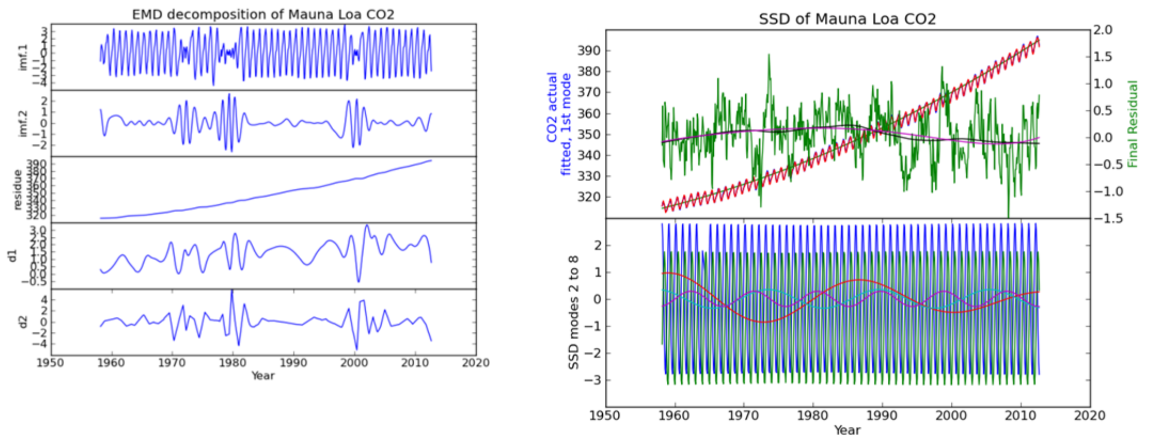

Figure 1: ***Left panel*** shows the EMD results. The top two panes are IMF 1 and IMF 2, the middle pane is the residue, and the bottom panes are the 1st and 2nd derivatives of the residue. ***Right panel*** All essentials of the SSD. Top pane shows; the raw data in blue, first extracted modal function as a smooth line beneath it, the sum of all modal functions (the "explained" portion) in red, mostly obscuring the raw data. The unexplained residual is shown in green over the top. A 6th order polynomial fit to the residual is shown in magenta and a LOESS smooth of the residuals is shown in black. The bottom pane shows other extracted modal functions, consecutively having lower amplitudes, illustrating SSD preferentially extracting informative functions first.

3. RESULTS

3.1. Mauna-Loa CO_2 Curve

The essential EMD results are shown in the left panel of Figure 1. Two IMFs are sufficient to extract a residue which is monotonic. Close examination of the residue in shows a curvilinearity and some minor accelerations. The IMFs show how EMD deals with discontinuities and missing values since there appears to be three regions where "energy" is transferred between IMF 1 and IMF 2, and these correspond to missing data. This transfer is also seen in the derivatives.

The essentials of SSD are shown in the right panel. Points to note are that the original data are closely modeled by the sum of the modal functions, but the residual probably has some structure, which can be further analysed.

3.2. Tide gauge analysis

The values of the first derivatives of the "unexplained" components of tide gauge level were extracted at years 1900 and 2000 for each SSD tide gauge analysis by each of the smoothing techniques (see Figure 2). These were compared using a two factor ANOVA where the 27 gauges are regarded as replicates (see Table 1). Broadly similar conclusions can be drawn by each of the analyses, however it is clear that the choice of analytic method makes a difference. The mean derivative in year 2000 is significantly greater than in year

1900 regardless of smoothing method used. The variance and the mean also vary by choice of smoothing method with polynomial fit showing more variance than the other methods (see Table 2). Two tailed t-tests comparing years 1900 and 2000, show P values of 0.006 for polynomial, 0.054 for EMD and 0.017 for LOESS smoothing.

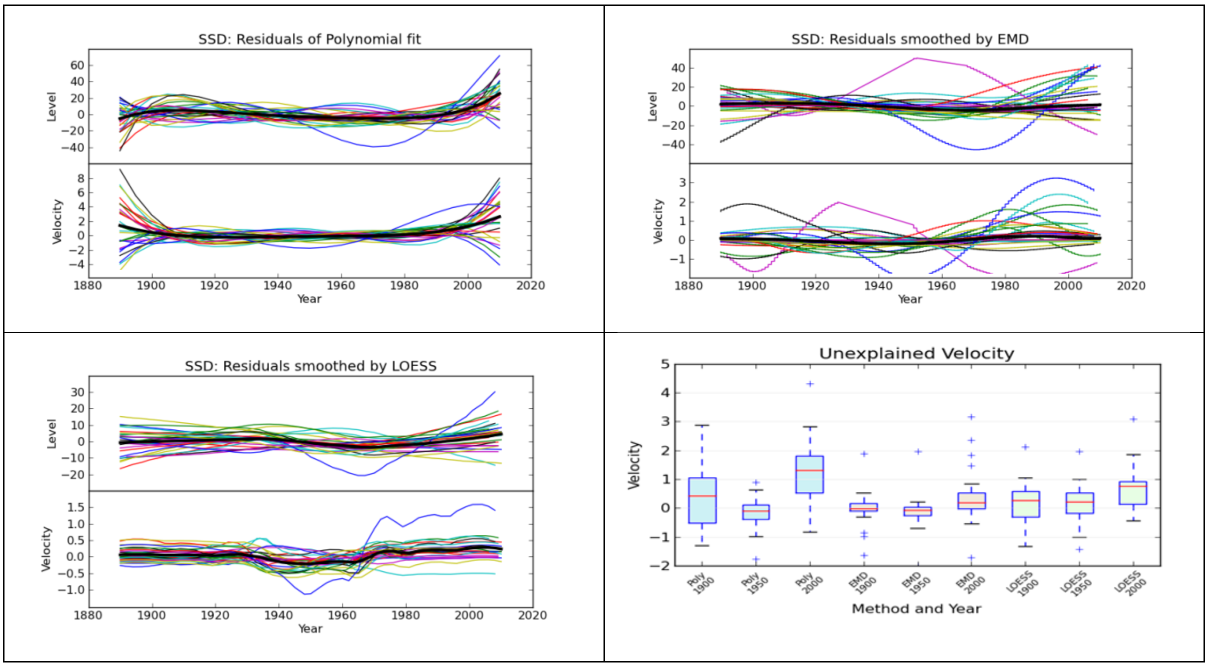

Figure 2: Plots of the smoothed unexplained SSD components showing the three smoothing methods and the first derivative of these ("velocity") as well. The number of tidal records shown is 27. The thick black line is the mean curve in each case. The bottom right pane shows boxplots for each of the smoothing methods and years 1900, 1950 and 2000.

4. DISCUSSION AND CONCLUSIONS

This paper has introduced the use of genetic programming as a tool for performing symbolic regression and shown one simple method by which complex signals can be analysed. The main advantage of SSD is that extracted functions are symbolic and corresponding underlying physical processes may be more easily discerned. For example, in the analysis of tidal gauge data, in almost all cases the first extracted modal function was $F_t^1 = a + bt + c\sin(2\pi t + d)$ where t is in years, and this serves to simultaneously remove most of the linear trend and most (not necessarily all) the annual cycle. Since this removes most of the known autocorrelation, many methods would proceed directly to an analytic method. However the structure of the extracted functions is related to the choices of building blocks, and the error function selected for Eureqa; and conclusions about signals remaining in the residuals must take into account the building blocks used. In frequency analyses a non-sinusoidal oscillator is represented as a family of frequencies. Similarly in SSD a single physical process may be represented as a family of functions.

These tide gauge records are not a uniform sampling of the ocean basins, being approximately evenly split between Baltic Sea and North Atlantic sites with San Francisco added. Whilst these results are consistent with more sophisticated techniques, they are data driven results and must be regarded as illustrative. There is evidence that there is a component of SLR at the year 2000 which is greater than at 1900 and that this appears in the latter part of the 20th Century after a minimum mid-century (see Table 2); this is the presumed signature of climate change. There is also evidence of a mid 20th century reduction in SLR followed by an acceleration, consistent with previous work.

According to Church and White (2011) the mean rate of global SLR from 1900-2009 is 1.7+/- 0.4 mm/yr, and since 1961, 1.9 +/-0.4 mm/yr. From the SSD analysis, the estimated change in "unexplained" SLR between 1900 and 2000 is between 0.42 and 0.84 mm/yr; the estimated change from 1950 to 2000 varies between 0.52 and 1.40 mm/yr, so that the increase in SLR between the two intervals is broadly compatible with their figures. However this comparison is acknowledged to be one between an estimate of global SLR and a data driven estimate (using SSD) of an attributed rate of rather localized change. It would be interesting to extend this approach using a larger selection of tide gauges with shorter records, and to compare results once ground water and isostatic effects are taken into account.

It would also be interesting to analyse the CO_2 curve in more depth.

Table 1: Two-way ANOVA. Three smoothing methods, two years (1900 and 2000), 27 replicates of the first derivatives of the residuals (mm/yr).

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Method	11.48	2	5.74	7.31	0.000926	3.05
Year	13.88	1	13.88	17.67	4.41E-05	3.90
Interaction	1.30	2	0.65	0.83	0.437938	3.05
Within	122.50	156	0.79			
Total	149.15	161				

Table 2: The means and variances of the first derivatives (mm/yr) of the unexplained tide gauge levels by each smoothing method.

Method	*Poly*			*EMD*			*LOESS*		
Year	*1900*	*1950*	*2000*	*1900*	*1950*	*2000*	*1900*	*1950*	*2000*
Average	0.42	-0.14	1.26	-0.02	-0.12	0.40	0.20	0.17	0.70
Variance	1.17	0.28	1.20	0.34	0.35	0.89	0.54	0.43	0.57

ACKNOWLEDGMENTS

Thanks to Dr Ian Smith for an early hearing of the ideas herein, and Dr Murray Peel for introducing me to EMD. (Limits in my use and understanding of EMD are mine alone). I thank Mr John Carter for introduction to Eureqa and for a review of an early version of this paper.

REFERENCES

Church, J. A., & White, N. J. (2011). Sea-level rise from the late 19th to the early 21st century. *Surveys in Geophysics, 32*(4-5), 585-602.

CORNELL. (2011) Retrieved from http://creativemachines.cornell.edu/eureqa

CRAN. (2103) Retrieved August, 2013, from http://cran.r-project.org/web/packages/EMD/index.html

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q, Liu, H. H. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London .Series A: Mathematical, Physical and Engineering Sciences, 454*(1971), 903-995.

Jevrejeva, S., Grinsted, A., Moore, J. C., & Holgate, S. (2006). Nonlinear trends and multiyear cycles in sea level records. *Journal of Geophysical Research: Oceans, 111*(C09012), 1-11.

Jevrejeva, S., Moore, J. C., Grinsted, A., & Woodworth, P. L. (2008). Recent global sea level acceleration started over 200 years ago? *Geophysical Research Letters, 35*(L08715), 1-4.

Mann, M. E., Bradley, R. S., & Hughes, M. K. (1999). Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophysical Research Letters, 26*(6), 759-762.

NOAA (2012).Mauna-loa monthly CO2. Retrieved August, 2012, from ftp://ftp.cmdl.noaa.gov/ccg/co2/trends/co2_mm_mlo.txt

PSMSL. (2011) Retrieved 07/27, 2011, from http://www.psmsl.org/data/obtaining/complete.php

Schmidt, M., & Lipson, H. (2010). Symbolic regression of implicit equations. In R. Riolo, U. O'Reilly & T. McConaghy (Eds.), *Genetic programming theory and practice* (pp. 73-85) Springer US.

Watson, P. J. (2011). Is there evidence yet of acceleration in mean sea level rise around mainland Australia? *Journal of Coastal Research,* , 368-377.

Wu, Q., & Riemenschneider, S. D. (2010). Bounadry extension and stop criteria for empirical mode decomposition. *Advances in Adaptive Data Analysis, 02*(02), 157-169.