# A proposed mildly explosive/self-exciting threshold autoregressive model applied to climatic time series

**J. M. Whyte** [a] and **A. V. Metcalfe** [a]

[a]*School of Mathematical Sciences, The University of Adelaide,*
*North Terrace 5005, South Australia, Australia*
*Email: jason.whyte@adelaide.edu.au*

**Abstract:** When selecting a time series model for a particular application it is appropriate to consider properties of the series under consideration. For example, it is not appropriate to apply standard autoregressive models to time series showing irreversibility, a feature observed in environmental series showing feedbacks. Models considered more suitable for modelling irreversible time series include Self-Exciting Threshold AutoRegressive (SETAR) models; a class of model composed of two or more regimes where the prediction equation used at a particular time is determined by some combination of previous time series values.

In certain applications SETAR models are not able to reproduce rapid changes observed in series, as observed in the Burgundy Spring-Summer temperature anomaly series reconstructed from grape harvest dates [Chuine *et al.*, *Nature* **432** (2004)]. To address this limitation, SETAR modified to have a regime capable of unstable behaviour flanked by two stable regimes are considered. One potentially suitable model class for the unstable middle regime is a mildly explosive model (MEM), an AR(1) model with a parameter greater than one which decays monotonically over time permitting unstable behaviour. MEM have found use in modelling financial exuberance and informed "collapsing bubble models" used to model rapid changes in the NASDAQ index. The SETAR class modified to have a MEM middle regime is termed here a MEMTAR.

As an aid to evaluating the properties of a series, either historical or model derived, some descriptors of time series behaviour are proposed. These descriptors aim to detect a lack of directional symmetry in a series and inform the choice of model for a given application. A simulation study is used to obtain descriptors of values from a SETAR and a MEMTAR which each had an unstable order one autoregressive middle regime; a random walk for the SETAR and a less explosive version of a MEM for the MEMTAR. The simulation exercise results show that the MEMTAR tended to take extreme values more readily than the SETAR, suggesting the MEMTAR is worth further consideration.

SETAR and MEMTAR models composed of autoregressive processes of orders one to three are applied to annual sunspot data from 1700 to 1988 and the reconstructed Burgundy temperature series of Chuine *et al.* from 1370 to 1977. Results are compared with those obtained from fitting standard autoregressive models. Results show that SETAR and MEMTAR models improve upon the ability of autoregressive models to model peak values.

*Keywords:* Mildly explosive model, self-exciting threshold autoregressive model, directional time series, Burgundy reconstructed Spring-Summer temperature anomaly, Wolf's annual sunspot data.

## 1   INTRODUCTION

When modelling a phenological series, it is appropriate to assess properties of the series to ensure that an appropriate model is used. Of particular importance is directionality (or irreversibility), a property ascribed to a time series by Lawrance [1991] if "... it possesses probabilistic properties which depend on the direction of time." A time series that does not have the property of irreversibility is termed reversibile. Plots of a time series may reveal directionality by highlighting what Lawrance calls "a lack of directional symmetry in certain characteristic behaviour sequences, such as rises and falls." A noted example is Wolf's annual sunspot data (obtainable from Hyndman [n d]) in which the series rises to peak values at a greater rate than the rate of decay from the peak.

Lawrance [1991] asserted that when historical data show directionality it is not appropriate to make predictions with a model unable to reproduce this feature, such as a Gaussian autoregressive moving average model. A variety of more appropriate models were suggested, including threshold models (Tong [1995]). This category includes Self-Exciting Threshold AutoRegressive (SETAR) models; a class of model composed of two or more autoregressive regimes where the prediction equation used at a given time is determined by the value of some combination of previous values in the time series. These models have the appeal that the model is capable of adaptation to prevailing conditions and hence appears better suited to modeling systems showing feedback behaviour. An extension of the SETAR class is Threshold AutoRegressive models with eXogenous input (TARX) which have found use in particular use financial applications and recently in rainfall-runoff modelling through a study of daily runoff in Australia's Murray–Darling Basin using daily rainfall as the exogenous input (Whyte et al. [2011]). The authors' use of TARX models in rainfall-runoff modelling and SETAR models in modelling daily maximum temperature data in South Australia (unpublished study in progress) has demonstrated that frequently the models cannot adequately reproduce extreme values in the time series. This paper proposes a modification to SETAR models intended to address this limitation.

The SETAR model class has fixed coefficients, ensuring that a regime exhibits a particular behaviour for all time. Relaxing this condition allows time-varying behaviour, such as a model capable of a degree of unstable behaviour which is able to change over time. This paper considers variants on the class of three regime SETAR models in which an autoregressive model displaying unstable behaviour is flanked by two stable models, ensuring stability overall. This unstable regime is realized by adapting a mildly explosive model (MEM) (Phillips and Magdalinos [2004]), a model class which has found use in modelling financial exuberance and the ensuing collapse of the NASDAQ bubble. The SETAR with MEM middle regime is termed here a MEMTAR; its design is intended to improve upon the SETAR's ability to model the rapid rises and falls seen in phenological series such as the reconstructed Burgundy Spring-Summer temperature data (Chuine et al. [2004]).

This paper is structured as follows. Section 2 defines the models under consideration, defines some measures of time series behaviour which are used subsequently to compare model predictions and introduces the data used for case studies. Section 3 describes the conditions of the simulation study used to collect information on the behaviour of the SETAR and MEMTAR models. Various orders of autoregressive models, SETAR and MEMTAR are fit to time series data and results are discussed in Section 4. Conclusions are drawn in Section 5.

## 2   PRELIMINARIES

### 2.1   Notation

$\tau_L$ and $\tau_U$: lower and upper thresholds respectively with $-\infty < \tau_L < \tau_U < \infty$.
$E[X]$ and $Var[X]$: the expected value of random variable $X$ and the variance of $X$ respectively.
$N(\mu, \sigma^2)$: the univariate normal distribution with mean $\mu$ and variance $\sigma^2$.
$Lap(l, s)$: the Laplace distribution with location parameter $l$ and scale parameter $s$.
$Exp(\lambda)$: the exponential distribution with rate $\lambda$, having cumulative distribution function $F(\cdot)$, where $F(x)$ has support for $x \geq 0$.
$TExp(\lambda, c)$: the translated exponential distribution with rate $\lambda$, having cumulative distribution function $\acute{F}(\cdot)$ defined by $\acute{F}(x - c) = F(x)$, giving $\acute{F}(x)$ support for $x \geq -c$.

## 2.2 Mathematical models

**AR($p$) models.** A scalar sequence $\{y_t\}$ is an autoregressive process of order $p$ (AR($p$) process) if successive values are related by

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon_t, \quad t \geq p+1, \ \mathrm{E}[\varepsilon_t] = 0, \ \mathrm{Var}[\varepsilon_t] = \sigma^2, \ \alpha_p \neq 0, \tag{1}$$

where the noise terms $\{\varepsilon_t\}_{t=p+1}^n$ are white noise, that is, independently and identically distributed.

If an AR(1) process (defined by (1) with $p = 1$) has $|\alpha_1| \geq 1$ the resulting process is termed unstable, otherwise it is stable. An AR(1) process with $\alpha_1 = 1$ is unstable and is termed a random walk.

**Variations on AR($p$) models.** We consider variations on AR(1) processes in which $\alpha_1$ is replaced by a time varying term $\alpha_{1_t}$ which allows periods of unstable behaviour. Unstable behaviour is promoted by taking $\alpha_{1_t}|_{t=1} > 1$ and allowing $\alpha_{1_t}$ to decay towards one as time progresses. This behaviour is captured by the model

$$y_t = \alpha_{1_t} y_{t-1}, \quad t \geq 2, \ \text{where } \alpha_{1_t} = 1 + \frac{c}{k_t}, \ c > 0, \tag{2}$$

where $\{k_t\}_{t=1}^\infty$ is a sequence of positive values that tends to infinity more slowly than $t$ does.

Taking $k_t \equiv t^\delta$ for $\delta \in (0,1)$ leads to the form of (2) known as a mildly explosive model (MEM) (Phillips and Magdalinos [2004]). The case where $\delta = 1$ is a less explosive version of a MEM as the coefficient decays more quickly.

**Self-exciting threshold autoregressive (SETAR) models.** Our model is similar to the open-loop threshold autoregressive systems given in Tong [1995, Page 101]. For a SETAR of three regimes used to model $y_t$, three prediction equations are available. Define $\mathbf{y}_t^{[p]} \equiv \{y_{t-1}, \ldots, y_{t-p}\}$ for $t > p$. The three regime SETAR of interest here predicts $y_t$ depending on the value of $y_{t-1}$ according to

$$y_t = \begin{cases} f^{[L]}(\mathbf{y}_t^{[p_U]}) + \varepsilon_t, & y_{t-1} < \tau_L, \\ f^{[M]}(\mathbf{y}_t^{[p_M]}) + \varepsilon_t, & \tau_L \leq y_{t-1} < \tau_U, \quad t > \max(p_L, p_M, p_U), \\ f^{[U]}(\mathbf{y}_t^{[p_L]}) + \varepsilon_t, & y_{t-1} \geq \tau_U, \end{cases} \tag{3}$$

where $L$, $M$ and $U$ denote the lower, middle and upper regimes consisting of autoregressive processes as in (1) having order $p_L$, $p_M$ and $p_U$ respectively. These regimes are defined by $\tau_L$ and $\tau_U$. A random noise term at time $t$ is denoted by $\varepsilon_t$. The case of (3) having $p_L = p_M = p_U = p$ is henceforth referred to as SETAR($p$) and has $3p$ parameters.

The SETAR template (3) is used subsequently to define some specific cases and variants having a middle regime capable of unstable behaviour.

**SETAR models with unstable middle regime.** Consider data $\{y_t\}_{t=1}^n$. For a SETAR(1) defined by (3) with lower and upper regimes described by stable AR(1) models

$$f^{[i]}(\mathbf{y}_t^{[1]}) = \beta_1^{[i]} y_{t-1}, \quad \left|\beta_1^{[i]}\right| < 1, \ i = L, U, \tag{4}$$

unstable behaviour of the middle regime is achieved in two ways. The first choice is the random walk, ((1) with $p = 1$ and $\alpha_1 = 1$),

$$f^{[M]}(\mathbf{y}_t^{[1]}) = y_{t-1}. \tag{5}$$

The second choice uses the mildly explosive model (2) as the middle regime giving a model termed here a MEMTAR. The mildly explosive model of (2) is adapted such that the time-varying coefficient returns to its initial value after each entry of the process into the region between the thresholds and decays if the process remains in this region. Defining the set $\mathcal{M} \equiv \{x \in \mathbb{R} | \tau_L \leq x < \tau_U\}$, then the set of time points when the middle regime is first used for predictions is

$$T^* \equiv \{t \in \{2, \ldots, n\} \, | y_t \in \mathcal{M} \text{ and } y_{t-1} \notin \mathcal{M}\}, \tag{6}$$

and the last time point spent by the process between the two thresholds following entry time $t^*$ is

J. M. Whyte and A.V. Metcalfe, A mildly explosive/SETAR model applied to climatic time series

$$t^\dagger(t^*) \equiv \min_t \left\{ t \in \{t^*, \dots, n\} \, | \, y_t \in \mathcal{M} \text{ and } y_{t+1} \notin \mathcal{M} \right\}. \tag{7}$$

Hence given $t^* \in T^*$, the middle regime of (3) is used to make predictions for $t^* + 1 \leq t \leq t^\dagger(t^*) + 1$. Recalling (2), we use $k_t = t^\delta$, $1 \leq t \leq n$, $\delta \in (0, 1]$, and hence the middle regime of a MEMTAR models $y_t$ for some period beyond $t^* \in T^*$ by

$$f^{[M]}(\mathbf{y}_t^{[1]}) = \beta_{1_t}^{[M]} y_{t-1}, \quad \text{where } \beta_{1_t}^{[M]} = 1 + \frac{c}{k_{(t-t^*)}}, \quad c > 0, \ t^* + 1 \leq t \leq t^\dagger(t^*) + 1. \tag{8}$$

A MEMTAR which is a SETAR with $p_L = p_U = p$ and middle regime given by (8) is henceforth termed MEMTAR($p; \delta$) and has $2p + 1$ parameters.

**Outline of model fitting process.** All model fitting is implemented in R [The R Project for Statistical Computing]. The setar() function (Di Narzo [2010]) fits a SETAR to data given various features of the model. This paper considers the SETAR($p$) version of (3). Model parameters are estimated by solving a series of conditional least squares problems. For fixed values of the upper and lower thresholds, model parameters are estimated by independently minimizing the squared residuals that arise for predictions made for the upper, lower and middle regimes. A grid search over feasible values for the upper and lower thresholds finds the thresholds and associated parameters that minimize the total residual sum of squares.

The R program devised for fitting a MEMTAR to data is similar to the process used for fitting a SETAR as it is necessary to search over all feasible combinations of upper and lower thresholds. At this stage the fitting routine finds the best fit of a MEMTAR($p, \delta$) to data given the $p$ and $\delta$. Parameters for the upper and lower regimes are estimated by determining the least squares estimate of the linear model used however it is necessary to treat the middle regime differently.

Expressing (8) as

$$f^{[M]}(\mathbf{y}_t^{[1]}) = y_{t-1} + c \frac{y_{t-1}}{k_{(t-t^*)}}, \quad t^* + 1 \leq t \leq t^\dagger(t^*) + 1 \tag{9}$$

suggests fitting (9) to data by performing a linear regression of $y_t$ against explanatory variables $y_{t-1}$ and $y_{t-1}/k_{(t-t^*)}$. This is achieved in R through using the lm() function to relate $y_t$ to $y_{t-1}/k_{(t-t^*)}$ with an "offset" term $y_{t-1}$.

Fitting of AR($p$) models (1) to data is by the ar() function with model fitting by minimizing ordinary least squares.

### 2.3 Some descriptors of time series behaviour

**Definition 1.** *Consider a sequence $\{y_t\}_{t=1}^n$ and fixed values $\tau_L < \tau_U$ termed the lower and upper thresholds respectively. An upper (lower) sojourn of a process is defined from the time the process first crosses above the upper threshold (below the lower threshold), $t_F$, up to and including the last time point for which the process remains above (below) the threshold, $t_L$. The sojourn is either for a single time point or for a time interval. The largest peak (smallest trough) that occurs for $t_F \leq t \leq t_L$ is called extreme value E, defined to occur at time $t_E$. Some descriptors of this sojourn are:*
**Upper (Lower) sojourn time from Entry to Extreme value (UEE/LEE):** $t_E - t_F \geq 0$.
**Upper (Lower) sojourn time After an Extreme value (UAE/LAE):** $t_L - t_E \geq 0$.
**Upper (Lower) Sojourn Time (UST/LST):** $(t_L - t_F) + 1 \geq 1$.

Consider the collections of upper sojourns and lower sojourns of the time series independently to obtain the distributions of the descriptors described in Definition 1 in each case. We propose that differences between the distributions of descriptors provide measures of directionality in the time series. For example, differences in UEE and UAE distributions suggest that a series has different behaviour when climbing to peaks compared to when descending from peaks.

### 2.4 Data

Wolf's annual sunspot numbers for 1700 to 1988 (file SUNSPOT.DAT obtained from The Time Series Library Hyndman [n d].) provides a well-known example of a directional time series and a test for the

descriptors proposed in Definition 1. Also considered is the series of temperature anomalies for the period 1370 to 2003, defined relative to the April to August mean temperature of Dijon between 1960 and 1989. This series was reconstructed using grape harvest dates in Burgundy Chuine et al. [2004]. This study used the series from 1370 to 1977 to avoid a missing value towards the end of the series.

**Initial consideration of the data.** The autocorrelation function (ACF) of the sunspot data showed periodic behaviour. To reduce this feature of the time series $\{x_t\}_{t=1}^n$, a linear model featuring harmonic terms of period $T = 11$ years given by

$$x_t = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{T}\right) + \beta_2 \sin\left(\frac{2\pi t}{T}\right). \tag{10}$$

was fitted to the time series. Inspection of the ACF for the residuals obtained showed that the periodic behaviour was substantially reduced and this series is subsequently referred to as the deseasonalized sunspot series. The ACF of the Burgundy temperature data did not show substantial periodic behaviour.

Directionality of each time series was investigated by taking the lower and upper thresholds as the first and third quartiles of the data respectively and considering the descriptors proposed in Section 2.3. For the deseasonalized sunspot data the UEE distribution showed lower values than the UAE distribution, agreeing with asymmetry reported for the series in the introduction. The Burgundy temperature data had a relatively small number of upper and lower sojourns, but given this it appeared that the time spent on lower excursions (LST) was generally shorter than that spent on upper excursions (UST) and that the time spent climbing to a peak (UEE) was shorter than that spent descending from a peak (UAE).

## 3 A SIMULATION STUDY OF THRESHOLD MODELS WITH UNSTABLE MIDDLE REGIME

Values from the threshold models with an unstable middle regime introduced in Section 2.2 are simulated so that descriptors of the simulated series described in Section 2.3 are computed and compared.

### 3.1 Overview of simulation conditions and models

Output is simulated from each model with stable AR(1) upper and lower regimes described by (4) with $\beta_1^{[L]} = \beta_1^{[U]} = 0.9$. The simulation uses symmetric thresholds $\tau_L = -4$ and $\tau_U = 4$ and starts from $y_1 = 0$. Distributions used to generate noise are N(0, 1), Lap(0, 1/$\sqrt{2}$) and TExp(1, −1), which all have mean zero and standard deviation one. For each source of noise, $10^4$ model values are simulated and the same random seed is used.

The SETAR(1) with random walk middle regime defined by (5) has $\beta_1^{[M]} = 1$. For the MEMTAR(1;1) the middle regime is described by (8) with $c = 0.01$ and $k_t = t/5$.

### 3.2 Simulation results and comparison of model behaviour

For all three sources of noise, the MEMTAR(1;1) showed a larger number of both lower and upper sojourns than the SETAR(1), the number of extra upper and lower sojourns depended on the noise distribution. It was decided to pair a particular upper descriptor and its corresponding lower descriptor (eg. UEE and LEE) from the SETAR(1) with those of the MEMTAR(1;1) and consider the results by noise type using boxplots of the distributions. For most descriptors and a given noise type, the comparison did not show any difference between the distributions from the first to the third quartile. The notable exception to this finding is in the case of Laplace noise where the MEMTAR(1;1) UEE distribution has a smaller median and third quartile than that of the SETAR(1), suggesting that the MEMTAR(1;1) generally spends less time above the upper threshold than does the SETAR(1). This behaviour is shown in Figure 1.

The simulation study has shown that the MEMTAR(1;1) is able to reach values further from zero more frequently than the SETAR(1) with a random walk middle regime. We propose that MEMTAR with $\delta \in (0, 1)$ is capable of an even greater degree of unstable behaviour and hence is worth investigating as an alternative to the existing SETAR class.
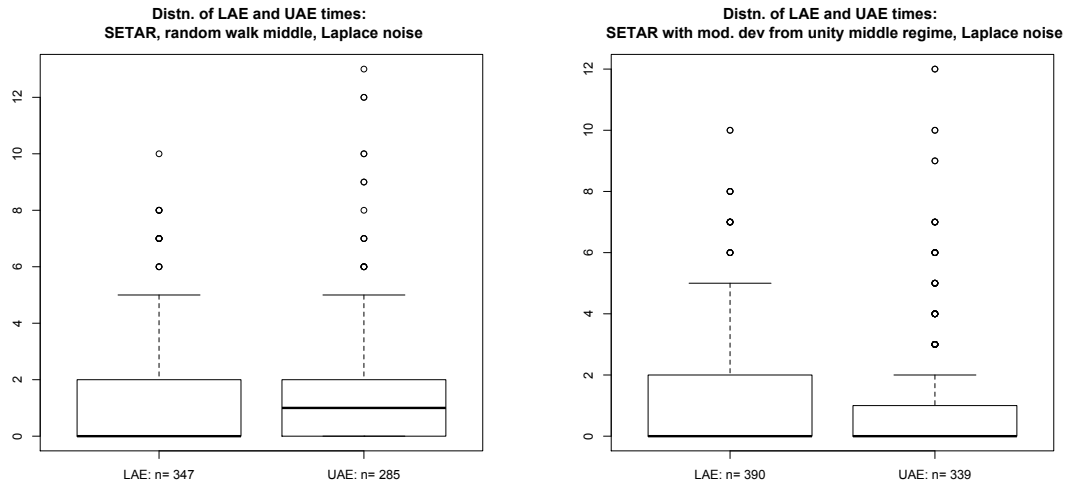
**Figure 1**. Boxplots of UAE and LAE obtained by simulations using $\mathrm{Lap}(0, 1/\sqrt{2})$ noise for the SETAR with random walk middle regime (left) and the SETAR with a moderate deviation from unity middle regime with $c = 0.01$ and $k_t = t/5$ (right). The value $n$ is the number of lower or upper sojourns as appropriate.

## 4  MODELLING OF PHENOLOGICAL TIME SERIES

Autoregressive models, AR($p$), SETAR($p$) and MEMTAR($p; \delta$) of orders one to three were fit to the deseasonalized sunspot series (Table 1) and the Burgundy temperature series (Table 2) noting the variance of residuals ($\sigma_R^2$). For the sunspot data, the $\sigma_R^2$ determined for AR($p$) models with $p = 3, 4, 5$ did not improve on $\sigma_R^2$ obtained for the AR(2) model. For the temperature data the $\sigma_R^2$ observed for fitting AR($p$) models with $p = 3, 4, 5, 6$ did not improve on $\sigma_R^2$ obtained for the AR(2) model. The results for the temperature series show that a MEMTAR(1,1) produces a fit to data that is comparable to the SETAR(1) and better than the AR(3) for the same number of parameters.

Fitting of the SETAR($p$) and MEMTAR($p, \delta$) models for $p = 2, 3$ resulted in very few points modelled by the middle regime. This may in part explain why the variance of residuals does not decrease substantially in going from $p = 2$ to $p = 3$.

Inspection of plots of fitted values superimposed on the data indicates that MEMTAR and SETAR models show an improvement in modelling of peak values compared to the autoregressive models (1) of the same or greater order.

For $p = 2, 3$ the SETAR model shows better performance than the MEMTAR of the same order. This is possibly because the SETAR(2,$\delta$) has one more parameter and the SETAR(3,$\delta$) two more parameters than the MEMTAR of the same order. It may also be the case that given the more extreme values produced the the MEMTAR($p; \delta$) for $p = 2, 3$ compared to the MEMTAR(1;$\delta$) that the middle regime is too explosive, overpredicting values in the time series. This would inflate the residual sum of squares determined for the middle regime and cause the model fitting routine to choose values of the upper and lower thresholds that are as close together as possible to minimize this effect.

## 5  CONCLUSIONS AND RECOMMENDATIONS

The MEMTAR(1,1) model improves upon the performance of AR(3) and SETAR(1) models in modelling the series considered as it produces a smaller variance of residuals. At this point R code for fitting MEMTAR models does not search over $\delta$ values and hence it is possible some improvement in MEMTAR performance will result from implementing this feature.

**Table 1**. Variance of residuals ($\sigma_R^2$) arising from fitting selected autoregressive models to sunspot data.

| $p$ | AR($p$) | MEMTAR($p; \delta$) | | SETAR($p$) |
|---|---|---|---|---|
| | | $\delta = 0.5$ | $\delta = 1$ | |
| 1 | 388.7 | 370.2 | 367.2 | 380.4 |
| 2 | 258.3 | 249.0 | 249.0 | 231.9 |
| 3 | 258.3 | 247.8 | 247.8 | 225.0 |

**Table 2**. Variance of residuals ($\sigma_R^2$) to three decimal places arising from fitting selected autoregressive models to Burgundy reconstructed temperature data.

| $p$ | AR($p$) | MEMTAR($p; \delta$) | | SETAR($p$) |
|---|---|---|---|---|
| | | $\delta = 0.5$ | $\delta = 1$ | |
| 1 | 0.886 | 0.858 | 0.856 | 0.866 |
| 2 | 0.882 | 0.859 | 0.859 | 0.852 |
| 3 | 0.882 | 0.840 | 0.835 | 0.826 |

The current MEMTAR class has the same middle regime behaviour regardless of whether the process enters the region between the two regimes from above the upper threshold or below the lower threshold. As processes may exhibit differing behaviour depending on whether they have an increasing or decreasing trend, some refinement to the MEMTAR may be appropriate to allow the middle regime to make predictions from time $t^*$ that depends on whether the $y_{t^*-1}$ is greater than the upper threshold or less than the lower threshold.

**REFERENCES**

Chuine, I., P. Yiou, N. Viovy, B. Seguin, V. Daux, and E. Le Roy Ladurie (2004). Grape ripening as a past climate indicator. *Nature 432*, 289–290.

Di Narzo, A. F. (2010). setar. Self-exciting threshold model fitting in R, accessed in package tsDyn, version 0.7-60.

Hyndman, R. J. (n. d.). Time series data library, http://robjhyndman.com/tsdl. accessed July 15, 2011.

Lawrance, A. J. (1991). Directionality and reversibility in time series. *International Statistical Review 59*(1), 67–79.

Phillips, P. C. B. and T. Magdalinos (2004, July). Limit theory for moderate deviations from a unit root. Cowles foundation discussion paper, number 1417.

The R Project for Statistical Computing (n. d.). R. http://www.r-project.org/.

Tong, H. (1995). *Non-linear Time Series: A Dynamical System Approach*. New York: Oxford University Press.

Whyte, J. M., A. Plumridge, and A. V. Metcalfe (2011). Comparison of predictions of rainfall-runoff models for changes in rainfall in the Murray-Darling Basin. *Hydrology and Earth System Sciences Discussions 8*(1), 917–955.